Corrected Version of 'A Unifying Variational Perspective on Some Fundamental Information Theoretic Inequalities'

Sangwoo Park, Erchin Serpedin, and Khalid Qaraqe

arXiv:1211.4795v5 [cs.IT] 4 Feb 2016

Abstract

This paper proposes a unifying variational approach for proving some fundamental information theoretic inequalities. Fundamental information theory results such as maximization of differential entropy, minimization of Fisher information (Cramér-Rao inequality), worst additive noise lemma, and extremal entropy inequality (EEI) are interpreted as functional problems and proved within the framework of calculus of variations. Several applications and possible extensions of the proposed results are briefly mentioned.

Index Terms

Maximizing Entropy, Minimizing Fisher Information, Worst Additive Noise, Extremal Entropy Inequality, Calculus of Variations

I. INTRODUCTION

IN the information theory realm, it is well-known that given the second-order moment (or variance), a Gaussian density function maximizes the differential entropy. Similarly, given the second-order moment, the Gaussian density function minimizes the Fisher information, a result which is referred to as the Cramér-Rao inequality in the signal processing literature. Surprisingly, the proofs proposed in literature for these fundamental results are quite diverse, and no unifying feature exists. Since differential entropy or Fisher information is a functional with respect to a probability density function, the most natural way to establish these results is by approaching them from the perspective of functional analysis. This paper presents a unifying variational framework to address these results as well as numerous other fundamental information theoretic results. A challenging information theoretic inequality, referred to as the extremal entropy inequality (EEI) [2], can be dealt with successfully in the proposed functional framework. Furthermore, the proposed variational calculus perspective is useful in establishing other novel results, applications and extensions of the existing information theoretic inequalities.

The main theme of this paper is to illustrate how some tools from calculus of variations can be used successfully to prove some of the fundamental information theoretic inequalities, which have been widely used in information theory and other fields, and to establish some applications. The proposed variational approach provides alternative proofs for some of the fundamental information theoretic inequalities and enables finding novel extensions of the existing results. This statement is strengthened by the fact that the proposed variational framework is quite general and powerful, and it allows easy integration of various linear and inequality constraints into the functional that is to be optimized. Therefore, we believe that a large number of applications could benefit of these tools. The proposed variational approach offers also a potential guideline for finding the optimal solution for many open problems.

Variational calculus techniques have been used with great success in solving important problems in image processing and computer vision [3] such as image reconstruction (denoising, deblurring), inverse problems, and image segmentation. Recently, variational techniques were also advocated for optimization of multiuser communication

systems [4], for deriving analytical wireless channel models using the maximum entropy principle when only limited information about the environment is available [5], and for designing optimal training sequences for radar and sonar applications [6]-[7]. Maximum entropy principle found also applications in spectral estimation (e.g., Burg's maximum entropy spectral density estimator [1]) and Bayesian statistics [8].

The major results of this paper are enumerated as follows. First, using calculus of variations, the maximizing differential entropy and minimizing Fisher information theorems are proved under the classical (standard) assumptions found in the literature as well as under a different set of assumptions. It is shown that a Gaussian density function maximizes the differential entropy but it minimizes the Fisher information, given the second-order moment. It is also shown that a half normal density function maximizes the differential entropy over the set of non-negative random variables, given the second-order moment. Furthermore, it is shown that a half normal density function minimizes the Fisher information over the set of non-negative random variables, provided that the regularity condition[1] is ignored and the second-order moment is given. It is also shown that a chi density function minimizes the Fisher information over the set of non-negative random variables, under the assumption that the regularity condition holds and the second-order moment is given.

Second, a novel proof of the worst additive noise lemma [9] is provided in the proposed functional framework. Previous proofs of the worst additive noise lemma were based on Jensen's inequality or data processing inequality [9], [10]. Unlike the previous proofs, our approach is purely based on calculus of variations techniques, and the vector version of the lemma is treated.

Third, EEI is studied from the perspective of a functional problem. The main advantage of the proposed new proof is that neither the channel enhancement technique and the entropy power inequality (EPI), adopted in [2], nor the equality condition of data processing inequality and the technique based on the moment generating functions, used in [11], are required. Using a technique based on calculus of variations, an alternative proof of EEI is provided. Finally, several applications and extensions of the proposed results are discussed.

The rest of this paper is organized as follows. Some variational calculus preliminary results and their corollaries are first reviewed in Section II. Maximizing differential entropy theorem and minimizing Fisher information theorem (Cramér-Rao inequality) are proved in Section III. In Section IV, the worst additive noise lemma is introduced and proved based on variational arguments. EEI is proved in Section V. In Section VI, some additional applications of the proposed variational techniques are briefly mentioned. Finally, Section VII concludes this paper.

II. Some Preliminary Calculus of Variations Results

In this section, we will review some of the fundamental results from variational calculus, and establish the concepts, notations and results that will be used constantly throughout the rest of the paper. These results are standard and therefore will be described briefly without further details. Additional details can be found in calculus of variations books such as [12]-[14].

Definition 1. *A functional $U[f]$ might be defined as*

$$U[f] = \int_a^b K(x, f(x), f'(x))dx, \tag{1}$$

which is defined on the set of continuous functions ($f(x)$) with continuous first-order derivatives ($f'(x) = df(x)/dx$) on the interval $[a, b]$. The function $f(x)$ is assumed to satisfy the boundary conditions $f(a) = A$ and $f(b) = B$. The functional $K(\cdot, \cdot, \cdot)$ is also assumed to have continuous first-order and second-order (partial) derivatives with respect to (wrt) all of its arguments.

Definition 2. *The increment of a functional $U[f]$ is defined as*

$$\Delta U[t] = U[f + t] - U[f], \tag{2}$$

where the function $t(x)$, that satisfies the boundary conditions $t(a) = t(b) = 0$, represents the admissible increment of $f(x)$, and it is assumed independent of the function $f(x)$ and twice differentiable.

---

[1]The regularity condition is defined in Theorems 6 and 7.

Definition 3. *Suppose that given $f(x)$, the increment in (2) is expressed as*

$$\Delta U[t] = \varphi[t] + \epsilon \|t\|, \tag{3}$$

where $\varphi[t]$ is a linear functional, ϵ goes to zero as $\|t\|$ approaches zero, and $\|\cdot\|$ denotes a norm defined in the case of a function $f(x)$ as:

$$\|f\| = \sum_{i=0}^{n} \max_{a \le x \le b} \left| f^{(i)}(x) \right|, \tag{4}$$

where $f^{(i)}(x) = d^i f(x)/dx^i$ are assumed to exist and be continuous for $i = 0, \ldots, n$ on the interval $[a, b]$, and the summation upper index n might vary depending on the normed linear space considered (e.g., if the normed linear space consists of all continuous functions $f(x)$, which have continuous first-order derivative on the interval $[a, b]$, $\|f\| = \max_{a \le x \le b} |f(x)| + \max_{a \le x \le b} |f'(x)|$, and in this case $n = 1$; see e.g., [12] for further details). Under the above assumptions, the functional $U[f]$ is said to be differentiable, and the major part of the increment $\varphi[t]$ is called the (first-order) variation of the functional $U[f]$ and it is expressed as $\delta U[f]$.

Based on Definitions 1, 2, 3 and Taylor's theorem (see e.g., [12]-[14] for additional justifications), the first-order and the second-order variations of a functional $U[f]$ can be expressed as

$$\delta U[f] = \int \left[K'_f \left(x, f, f' \right) t(x) + K'_{f'} \left(x, f, f' \right) t'(x) \right] dx \tag{5}$$

$$\delta^2 U[f] = \frac{1}{2} \int \left[K''_{ff} \left(x, f, f' \right) t(x)^2 + 2 K''_{ff'} \left(x, f, f' \right) t(x) t'(x) + K''_{f'f'} \left(x, f, f' \right) t'(x)^2 \right] dx$$

$$= \frac{1}{2} \int \left[K''_{f'f'} t'^2 + \left(K''_{ff} - \frac{d}{dx} K''_{ff'} \right) t^2 \right] dx, \tag{6}$$

where K'_f and $K'_{f'}$ stand for the first-order partial derivatives wrt f and f', respectively, $K''_{ff'}$ denotes the second-order partial derivative wrt f and f', K''_{ff} represents the second-order partial derivative wrt f, and $K''_{f'f'}$ is the second-order partial derivative wrt f'. Throughout the paper to simplify the exposition, the arguments of functionals or functions are omitted unless the arguments are ambiguous or confusing. Also, the range of integration in various integrals will not be explicitly marked unless the range is ambiguous.

Theorem 1 ([12]). *A necessary condition for the functional $U[f]$ in (1) to have an extremum (or local optimum) for a given function $f = f^*$ is that its first variation vanishes at $f = f^*$:*

$$\delta U[f^*] = 0, \tag{7}$$

for all admissible increments. This implies

$$K'_{f^*} - \frac{d}{dx} K'_{f'^*} = 0, \tag{8}$$

a result which is known as Euler's equation. When the functional in (1) includes multiple functions (e.g., $f_1, \ldots, f_m$) and multiple integrals wrt $x_1, \ldots, x_n$, i.e.,

$$\int \cdots \int K \left(x_1, \ldots, x_n, f_1, \ldots, f_m, f'_1, \ldots, f'_m \right) dx_1 \cdots dx_n,$$

then Euler's equation in (8) takes the form of the system of equations:

$$K'_{f_i^*} - \sum_{j=1}^{n} \frac{d}{dx_j} K'_{f_i'^*} = 0, \qquad i = 1, \ldots, m. \tag{9}$$

In particular, when the functional does not depend on the first-order derivative of the functions $f_1, \ldots, f_m$, the equations in (9) reduce to

$$K'_{f_i^*} = 0, \qquad i = 1, \ldots, m. \tag{10}$$

Proof: Details of the proof of this theorem can be found, e.g., in [12]-[14]. ∎

Theorem 2 ([12]). *A necessary condition for the functional $U[f]$ in (1) to have a minimum for a given $f = f^*$ is that the second variation of functional $U[f]$ be nonnegative:*

$$\delta^2 U[f^*] \geq 0, \tag{11}$$

for all admissible increments. This implies

$$K''_{f'^* f'^*} \geq 0. \tag{12}$$

In particular, when the functional in (1) does not depend on the first-order derivative of the function f, (12) simplifies to

$$K''_{f^* f^*} \geq 0. \tag{13}$$

When the functional in (1) includes multiple functions (e.g., $f_1, \ldots, f_m$) and multiple integrals wrt $x_1, \ldots, x_n$, i.e.,

$$\int \cdots \int K(x_1, \ldots, x_n, f_1, \ldots, f_m)\, dx_1 \cdots dx_n,$$

then the condition in (13) is expressed in terms of the positive semi-definiteness of the matrix:

$$\begin{bmatrix} K''_{f_1 f_1} & \cdots & K''_{f_1 f_m} \\ \vdots & \ddots & \vdots \\ K''_{f_m f_1} & \cdots & K''_{f_m f_m} \end{bmatrix} \succeq 0. \tag{14}$$

Proof: The inequality in (13) is easily derived from the inequality in (12) since $K''_{f'_x f'_x}$ and $K''_{f_x f'_x}$ are vanishing in (6) when the functional in (1) does not depend on the first-order derivative of the function f_x. The remaining details of the proof can be tracked in [12]. ∎

Theorem 3 ([12]). *Given the functional*

$$U[f_1, f_2] = \int_a^b K(x, f_1, f_2, f'_1, f'_2)\,dx, \tag{15}$$

assume that the admissible functions satisfy the following boundary conditions:

$$f_1(a) = A_1,\ f_1(b) = B_1,\ f_2(a) = A_2,\ f_2(b) = B_2,$$
$$k(x, f_1, f_2) = 0, \tag{16}$$
$$L[f_1, f_2] = \int_a^b \tilde{L}(x, f_1, f_2, f'_1, f'_2)\,dx = l, \tag{17}$$

where a, b, A_1, B_1, A_2, B_2, and l are constants, $k(x, f_1, f_2)$ is a functional wrt f_1 and f_2, and $U[f_1, f_2]$ is assumed to have an extremum for $f_1 = f_1^*$ and $f_2 = f_2^*$.

If f_1^* and f_2^* are not extremals of $L[f_1, f_2]$, or $k'_{f_1^*}$ and $k'_{f_2^*}$ do not vanish simultaneously at any point in (16), there exist a constant λ and a function $\lambda(x)$ such that f_1^* and f_2^* are extremals of the functional

$$\int_a^b (K(x, f_1, f_2, f'_1, f'_2) + \lambda \tilde{L}(x, f_1, f_2, f'_1, f'_2) + \lambda(x) k(x, f_1, f_2))\,dx. \tag{18}$$

Based on Theorem 3, the following corollary is derived.

Corollary 1. *Given the functional*

$$U[f_x, f_y] = \int_a^b \int_a^b K(x, y, f_x, f_y)\,dx\,dy, \tag{19}$$

assume that the admissible functions satisfy the following boundary conditions:

$$f_x(a) = A_x,\ f_x(b) = B_x,\ f_y(a) = A_y,\ f_y(b) = B_y,$$
$$k(y, f_x, f_y) = g(y, f_y) - \int_a^b \tilde{k}(x, y, f_x)\,dx = 0,$$
$$L_i[f_x, f_y] = \int_a^b \int_a^b \tilde{L}_i(x, y, f_x, f_y)\,dx\,dy = l_i,\quad i = 1, 2, \cdots, n, \tag{20}$$

where a, b, A_X, B_X, A_Y, and B_Y stand for some constants, f_X is a function of x, f_Y is a function of y, $g(y, f_Y)$ is a function of f_Y, and $\tilde{k}(x, y, f_X)$ is a function of f_X. The functional $U[f_X, f_Y]$ is assumed to have an extremum at $f_X = f_{X^*}$ and $f_Y = f_{Y^*}$.

Unless f_{X^*} and f_{Y^*} are extremals of $L_i[f_X, f_Y]$, or $k'_{f_{X^*}}$ and $k'_{f_{Y^*}}$ simultaneously vanish at any point of $k(y, f_X, f_Y)$, there exist constants $\lambda_i, i = 1, 2, \cdots, n$, and a function $\lambda(y)$ such that $f_X = f_{X^*}$ and $f_Y = f_{Y^*}$ is an extremal of the functional

$$\int_a^b \left\{ \left[\int_a^b (K(x, y, f_X, f_Y) + \sum_{i=1}^n \lambda_i \tilde{L}_i(x, y, f_X, f_Y) - \lambda(y)\tilde{k}(x, y, f_X))dx \right] + \lambda(y)g(y, f_Y) \right\} dy. \tag{21}$$

Proof: See Appendix A. ∎

Based on Theorems 1, 2 and Corollary 1, we can derive the following corollary, which will be repeatedly used throughout this paper.

Corollary 2. *Based on the functional defined in (21), the following necessary conditions are derived for the optimal solutions f_{X^*} and f_{Y^*}:*

$$\int K'_{f_{X^*}}(x, y, f_{X^*}, f_{Y^*}) + \sum_{i=1}^n \lambda_i \tilde{L}_{if_{X^*}}'(x, y, f_{X^*}, f_{Y^*}) - \lambda(y)\tilde{k}'_{f_{X^*}}(x, y, f_{X^*})dy = 0, \tag{22}$$

$$\int K'_{f_{Y^*}}(x, y, f_{X^*}, f_{Y^*}) + \sum_{i=1}^n \lambda_i \tilde{L}_{if_{Y^*}}'(x, y, f_{X^*}, f_{Y^*})dx + \lambda(y)g'_{f_{Y^*}}(y, f_{Y^*}) = 0, \tag{23}$$

and the matrix

$$\begin{bmatrix} G''_{f_{X^*}f_{X^*}} & G''_{f_{X^*}f_{Y^*}} \\ G''_{f_{Y^*}f_{X^*}} & G''_{f_{Y^*}f_{Y^*}} \end{bmatrix}, \tag{24}$$

is positive semi-definite. The functional G is defined as

$$G(x, y, f_{X^*}, f_{Y^*}) = K(x, y, f_{X^*}, f_{Y^*}) + \sum_{i=1}^N \lambda_i \tilde{L}_i(x, y, f_{X^*}, f_{Y^*}) - \lambda(y)\tilde{k}(x, y, f_{X^*}) + \lambda(y)g(y, f_{Y^*})q(x),$$

and $q(x)$ is a (arbitrary but fixed) function which satisfies $\int_a^b q(x)dx = 1$, and it is introduced to homogenize the functional in (21). In particular, if function $g(y, f_Y)$ only involves first order component of f_Y, i.e., $g(y, f_Y) = f_Y$, the necessary condition reduces to check the positive semi-definiteness of the matrix

$$\begin{bmatrix} H''_{f_{X^*}f_{X^*}} & H''_{f_{X^*}f_{Y^*}} \\ H''_{f_{Y^*}f_{X^*}} & H''_{f_{Y^*}f_{Y^*}} \end{bmatrix},$$

where

$$H(x, y, f_{X^*}, f_{Y^*}) = K(x, y, f_{X^*}, f_{Y^*}) + \sum_{i=1}^N \lambda_i \tilde{L}_i(x, y, f_{X^*}, f_{Y^*}) - \lambda(y)\tilde{k}(x, y, f_{X^*}).$$

Proof: See Appendix A. ∎

III. MAX Entropy and MIN Fisher Information

This simple but significant result–given the second-order moment (or variance) of a random vector, a Gaussian random vector maximizes the differential entropy–is well-known. In this section, a completely rigorous and general derivation of the distribution achieving the maximum entropy will be first provided. This proof sets up the variational framework for establishing a second important result in this section, namely the Cramér-Rao bound, which states that for a given mean and correlation matrix, a normally distributed random vector minimizes the Fisher information matrix.

Theorem 4 ([1], [10])**.** *Given (a vector mean $\boldsymbol{\mu}_X$ and) a correlation matrix $\boldsymbol{\Omega}_X$, a Gaussian random vector $\mathbf{X}_G$ with the correlation matrix $\boldsymbol{\Omega}_X$ (and the vector mean $\boldsymbol{\mu}_X$) maximizes the differential entropy, i.e.,*

$$h(\mathbf{X}) \leq h(\mathbf{X}_G), \tag{25}$$

where $h(\cdot)$ denotes differential entropy, $\mathbf{X}$ is an arbitrary (but fixed) random vector with the correlation matrix $\mathbf{\Omega}_x$.

Proof: We first construct a functional, which represents the inequality in (25) and required constraints, as follows:

$$\min_{f_X} \quad \int f_X(\mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x}, \tag{26}$$

$$s.\ t. \quad \int f_X(\mathbf{x}) d\mathbf{x} = 1, \tag{27}$$

$$\int \mathbf{x} f_X(\mathbf{x}) d\mathbf{x} = \boldsymbol{\mu}_x \tag{28}$$

$$\int \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) d\mathbf{x} = \mathbf{\Omega}_x. \tag{29}$$

Using Theorem 3, the functional in (26) is expressed as

$$\min_{f_X} \quad U[f_X], \tag{30}$$

where $U[f_X] = \int K(\mathbf{x}, f_X) d\mathbf{x} = \int f_X(\mathbf{x}) \left(\log f_X(\mathbf{x}) + \alpha + \sum_{i=1}^{n} \zeta_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{ij} x_i x_j \right) d\mathbf{x}$, α is the Lagrange multiplier associated with the constraint (27), and ζ_i and λ_{ij} stand for the Lagrange multipliers corresponding to the constraints (28) and (29), respectively.

Based on Theorem 1, by checking the first-order variation condition, we can find the optimal solution $f_{X^*}(\mathbf{x})$ as follows:

$$K'_{f_X}\Big|_{f_X = f_{X^*}} = 1 + \log f_{X^*}(\mathbf{x}) + \alpha + \boldsymbol{\zeta}^T \mathbf{x} + \mathbf{x}^T \mathbf{\Lambda} \mathbf{x} = 0 \tag{31}$$

with $\boldsymbol{\zeta} = [\zeta_1, \cdots, \zeta_n]^T$ and the matrix $\mathbf{\Lambda} = [\lambda_{ij}]$, $i, j = 1, \ldots, n$. Considering the constraints in (27) - (29), from (31) it turns out that

$$
\begin{aligned}
f_{X^*}(\mathbf{x}) &= \exp\left\{ -\mathbf{x}^T \mathbf{\Lambda} \mathbf{x} - \boldsymbol{\zeta}^T \mathbf{x} - \alpha - 1 \right\} \\
&= (2\pi)^{-\frac{n}{2}} \left| \frac{1}{2} \mathbf{\Lambda}^{-1} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \left(\frac{1}{2} \mathbf{\Lambda}^{-1} \right)^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} (2\pi)^{\frac{n}{2}} \left| \frac{1}{2} \mathbf{\Lambda}^{-1} \right|^{\frac{1}{2}} \exp\left\{ -1 - \alpha + \boldsymbol{\mu}^T \mathbf{\Lambda} \boldsymbol{\mu} \right\} \\
&= (2\pi)^{-\frac{n}{2}} |\mathbf{\Omega}_x|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Omega}_x^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},
\end{aligned} \tag{32}
$$

where

$$
\begin{aligned}
\alpha &= -1 + \boldsymbol{\mu}^T \mathbf{\Lambda} \boldsymbol{\mu} + \frac{1}{2} \log (2\pi)^n |\mathbf{\Omega}_x|, \\
\mathbf{\Lambda} &= \frac{1}{2} \mathbf{\Omega}_x^{-1}, \\
\boldsymbol{\zeta} &= -2\mathbf{\Lambda}\boldsymbol{\mu}.
\end{aligned} \tag{33}
$$

Two remarks are now in order. First, the correlation matrix $\mathbf{\Omega}_x$ is assumed to be invertible. When the correlation matrix is non-invertible, similar to the method shown in [2], we can equivalently re-write the functional problem in (26) as

$$\min_{f_X(\mathbf{x})} -h(\mathbf{X}) \quad \Leftrightarrow \quad \min_{f_{\bar{X}}(\mathbf{x})} -h(\bar{\mathbf{X}}), \tag{34}$$

where $\mathbf{X} = \mathbf{Q}_\Omega \bar{\mathbf{X}}$, and in the spectral factorization $\mathbf{\Omega}_x = \mathbf{Q}_\Omega \mathbf{\Lambda}_\Omega \mathbf{Q}_\Omega^T$, $\mathbf{Q}_\Omega$ is an orthogonal matrix, $\mathbf{\Lambda}_\Omega = diag(\Lambda_1, \ldots, \Lambda_m, 0, \ldots, 0)$ and $diag$ denotes a diagonal matrix.

Let $\bar{\mathbf{X}} = \left[\bar{\mathbf{X}}_a^T, \bar{\mathbf{X}}_b^T \right]^T$, where the dimensions of $\bar{\mathbf{X}}_a$ and $\bar{\mathbf{X}}_b$ are m and $n - m$, respectively. It can be observed that the correlation matrix (or covariance matrix) of $\bar{\mathbf{X}}$, $\mathbf{\Omega}_{\bar{X}}$, is equal to the diagonal matrix $\mathbf{\Lambda}_\Omega$. Furthermore,

the correlation of $\bar{\mathbf{X}}_b$ is a zero matrix and $\bar{\mathbf{X}}_b$ can be considered as a deterministic vector. Thus, $\bar{\mathbf{X}}_a$ and $\bar{\mathbf{X}}_b$ are statistically independent and the equation in (34) and constraints in (27)-(29) are equivalently re-written as

$$\min_{f_{\bar{X}_a}(\mathbf{x})} -h(\bar{\mathbf{X}}_a),$$

$$s.t. \quad \int f_{\bar{X}_a}(\mathbf{x})d\mathbf{x} = 1,$$

$$\int \mathbf{x}f_{\bar{X}_a}(\mathbf{x})d\mathbf{x} = \boldsymbol{\mu},$$

$$\int \mathbf{x}\mathbf{x}^T f_{\bar{X}_a}(\mathbf{x})d\mathbf{x} = \boldsymbol{\Lambda}_{\Omega_a},$$

where $\boldsymbol{\Lambda}_{\Omega_a} = diag(\Lambda_1, \ldots, \Lambda_m) \succ \mathbf{0}$ is a positive-definite matrix. Therefore, without loss of generality, we can assume that the correlation matrix $\boldsymbol{\Omega}_X$ is invertible.

Based on Theorem 2, since

$$K''_{f_X f_X}\Big|_{f_X=f_{X^*}} = \frac{1}{f_{X^*}(\mathbf{x})} > 0,$$

the second-order variation $\delta^2 U\left[f_{X^*}\right]$ is positive, and the optimal solution f_{X^*} is a minimal solution for the variational problem in (26).

Therefore, the negative of differential entropy $-h(\mathbf{X})$ is minimized (or equivalently $h(\mathbf{X})$ is maximized) when $\mathbf{X}$ is a multi-variate Gaussian random vector. Even though Theorems 1, 2 are necessary conditions for the minimum, in this case, a multi-variate Gaussian density function is the actual solution since there is only one solution, namely the multi-variate Gaussian density function, in the feasible set. An alternative justification of global optimality of multi-variate Gaussian pdf can be achieved by exploiting the convexity of $K(\mathbf{x}, f_X)$ wrt f_X.

Remark 1. *The proof in [1] relies on calculus of variations to find the first-order necessary condition, which only represents a necessary (and not sufficient) condition for optimality. Therefore, an additional technique, referred to as the Kullback-Leibler divergence, was used to prove that the necessary solution globally maximizes the differential entropy. Unlike this proof, by confirming the convexity of the variational problem, we show that Gaussian distribution is indeed the global optimal solution solely based on calculus of variations arguments.*

■

The maximum entropy result can be extended in various ways. A simple variation of the maximum entropy considers only non-negative random variables. Then it turns out that Gaussian random variables are no longer the optimal solution that maximizes the differential entropy. The following theorem can be easily established and states that a half-normal random variable maximizes the differential entropy over the set of non-negative random variables.

Theorem 5. *Within the class of non-negative random variables with given second-order moment m_X^2, a half-normal random variable X_{HN} maximizes the differential entropy, i.e.,*

$$h(X) \leq h(X_{HN}), \tag{35}$$

where X is an arbitrary (but fixed) non-negative random variable with the second-order moment m_X^2, and $h(\cdot)$ denotes differential entropy.

Proof: The proof is omitted since it can be established following similar steps to the proof of Theorem 4. ■

Adopting a similar variational approach to the one in Theorem 4, we can also determine the probability density function that minimizes the Fisher information matrix as shown by the following theorem.

Theorem 6 (Cramér-Rao Inequality (a vector version)). *Given a vector mean $\boldsymbol{\mu}_X$ and a correlation matrix $\boldsymbol{\Omega}_X$, the Gaussian density function with the vector mean $\boldsymbol{\mu}_X$ and the correlation matrix $\boldsymbol{\Omega}_X$ minimizes the Fisher information matrix, i.e.,*

$$\mathbb{J}(\mathbf{X}) \succeq \mathbb{J}(\mathbf{X}_G), \tag{36}$$

where $\mathbf{X}$ *and* $\mathbf{X}_G$ *stand for an arbitrary (but fixed) random vector and Gaussian random vector, respectively, with given mean* $\boldsymbol{\mu}_x$ *and correlation matrix* $\boldsymbol{\Omega}_x$, *and* $\mathbb{J}(\cdot)$ *denotes the Fisher information matrix:*

$$\mathbb{J}(\mathbf{X}) = \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{bmatrix}, \tag{37}$$

$$s_{ij} = \int \left(\frac{\frac{d}{dx_i} f_x(\mathbf{x})}{f_x(\mathbf{x})} \right) \left(\frac{\frac{d}{dx_j} f_x(\mathbf{x})}{f_x(\mathbf{x})} \right) f_x(\mathbf{x}) d\mathbf{x}.$$

Proof: We first represent the inequality in (36) as a functional with the required constraints as follows:

$$\min_{f_x} \quad \int \boldsymbol{\xi}^T \nabla f_x(\mathbf{x}) \nabla f_x(\mathbf{x})^T \boldsymbol{\xi} \frac{1}{f_x(\mathbf{x})} d\mathbf{x}, \tag{38}$$

$$s.\ t. \quad \int f_x(\mathbf{x}) d\mathbf{x} = 1,$$

$$\int \mathbf{x} f_x(\mathbf{x}) d\mathbf{x} = \boldsymbol{\mu}_x,$$

$$\int \mathbf{x}\mathbf{x}^T f_x(\mathbf{x}) d\mathbf{x} = \boldsymbol{\Omega}_x, \tag{39}$$

where $\boldsymbol{\xi}$ *is an arbitrary but fixed non-zero vector, defined as* $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n]^T$.

Using Theorem 3, the functional problem in (38) is expressed as

$$\min_{f_x} \quad U[f_x], \tag{40}$$

where $U[f_x] = \int K(\mathbf{x}, f_x, \nabla f_x) d\mathbf{x}$, $K(\mathbf{x}, f_x, \nabla f_x) = (\boldsymbol{\xi}^T \nabla f_x(\mathbf{x}) \nabla f_x(\mathbf{x})^T \boldsymbol{\xi}/f_x(\mathbf{x})) + \alpha f_x(\mathbf{x}) + f_x(\mathbf{x}) \sum_{i=1}^n \zeta_i x_i + f_x(\mathbf{x}) \sum_{i=1}^n \sum_{j=1}^n \lambda_{ij} x_i x_j$, *and* α, ζ_i, *and* λ_{ij} *are the Lagrange multipliers corresponding to the three constraints in (39).*

Based on Theorem 1, by confirming the first-order variation condition, i.e., $\delta U[f_{x^*}] = 0$, *we can find the optimal solution* $f_{x^*}(x)$ *as follows:*

$$K'_{f_x} - \sum_{i=1}^n \frac{\partial}{\partial x_i} K'_{f'_{x_i}} \bigg|_{f_x = f_{X^*}} = 0, \tag{41}$$

where

$$K'_{f_x} = -\frac{\boldsymbol{\xi}^T \nabla f_x(\mathbf{x}) \nabla f_x(\mathbf{x})^T \boldsymbol{\xi}}{f_x(\mathbf{x})^2} + \alpha + \boldsymbol{\zeta}^T \mathbf{x} + \mathbf{x}^T \boldsymbol{\Lambda} \mathbf{x},$$

$$\frac{\partial}{\partial x_i} K'_{f'_{x_i}} = \frac{\partial}{\partial x_i} \left(\frac{2 \sum\limits_{j=1}^n \frac{\partial}{\partial x_j} f_x(\mathbf{x}) \xi_i \xi_j}{f_x(\mathbf{x})} \right)$$

$$= \frac{2 \sum\limits_{j=1}^n \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f_x(\mathbf{x}) \xi_i \xi_j}{f_x(\mathbf{x})} - \frac{2 \sum\limits_{j=1}^n \frac{\partial}{\partial x_j} f_x(\mathbf{x}) \xi_i \xi_j \frac{\partial}{\partial x_i} f_x(\mathbf{x})}{f_x(\mathbf{x})^2}. \tag{42}$$

Therefore, the left-hand side of the equation in (41) is expressed as

$$K'_{f_x} - \sum_{i=1}^n \frac{\partial}{\partial x_i} K'_{f'_{x_i}} = \frac{\sum\limits_{i=1}^n \sum\limits_{j=1}^n \frac{\partial}{\partial x_i} f_x(\mathbf{x}) \frac{\partial}{\partial x_j} f_x(\mathbf{x}) \xi_i \xi_j}{f_x(\mathbf{x})^2} - \frac{2 \sum\limits_{i=1}^n \sum\limits_{j=1}^n \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f_x(\mathbf{x}) \xi_i \xi_j}{f_x(\mathbf{x})} + \alpha + \sum_{i=1}^n \zeta_i x_i + \sum_{i=1}^n \sum_{j=1}^n \lambda_{ij} x_i x_j$$

$$= 0. \tag{43}$$

Unlike Theorem 4, we cannot directly calculate $f_{X^*}(\mathbf{x})$ from (41). Fortunately, the first two parts in equation (43) are expressed as quadratic forms when $f_{X^*}(\mathbf{x})$ is a multi-variate Gaussian density function, and therefore, the multi-variate Gaussian density function satisfies the equality in (43). When $f_{X^*}(\mathbf{x})$ is a multi-variate Gaussian density function:

$$f_{X^*}(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \left|\mathbf{\Sigma}_X\right|^{-\frac{1}{2}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu}_X)^T \mathbf{\Sigma}_X^{-1}(\mathbf{x}-\boldsymbol{\mu}_X)}{2}}$$

with $\mathbf{\Sigma}_X = \mathbf{\Omega}_X - \boldsymbol{\mu}_X \boldsymbol{\mu}_X^T$ and

$$\mathbf{\Sigma}_X^{-1} = \begin{bmatrix} \sigma^2_{X_{11}} & \cdots & \sigma^2_{X_{1n}} \\ \vdots & \ddots & \vdots \\ \sigma^2_{X_{n1}} & \cdots & \sigma^2_{X_{nn}} \end{bmatrix}, \tag{44}$$

its partial derivatives can be expressed as follows:

$$\frac{\partial}{\partial x_i} f_{X^*}(\mathbf{x}) = -\frac{1}{2}\left(\sum_{l=1}^{n} \sigma^2_{X_{il}}\left(x_l - \mu_{X_l}\right) + \sum_{m=1}^{n} \sigma^2_{X_{mi}}\left(x_m - \mu_{X_m}\right)\right) f_{X^*}(\mathbf{x}),$$

$$\frac{\partial}{\partial x_j}\frac{\partial}{\partial x_i} f_{X^*}(\mathbf{x}) = -\frac{1}{2}\left(\sigma^2_{X_{ij}} + \sigma^2_{X_{ji}}\right) f_{X^*}(\mathbf{x}) + \frac{1}{4}\left(\sum_{l=1}^{n} \sigma^2_{X_{il}}\left(x_l - \mu_{X_l}\right) + \sum_{m=1}^{n} \sigma^2_{X_{mi}}\left(x_m - \mu_{X_m}\right)\right)$$

$$\cdot \left(\sum_{l=1}^{n} \sigma^2_{X_{jl}}\left(x_l - \mu_{X_l}\right) + \sum_{m=1}^{n} \sigma^2_{X_{mj}}\left(x_m - \mu_{X_m}\right)\right) f_{X^*}(\mathbf{x}) \tag{45}$$

By substituting (45) into (43), it turns out that

$$K'_{f_{X^*}} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i} K'_{f'_{X_i^*}} = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i \xi_j \left(\sum_{l=1}^{n}\left(\sigma^2_{X_{il}} + \sigma^2_{X_{li}}\right)\left(x_l - \mu_{X_l}\right)\right)\left(\sum_{m=1}^{n}\left(\sigma^2_{X_{jm}} + \sigma^2_{X_{mj}}\right)\left(x_m - \mu_{X_m}\right)\right)$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sigma^2_{X_{ij}} + \sigma^2_{X_{ji}}\right)\xi_i \xi_j + \alpha + \sum_{i=1}^{n} \zeta_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{ij} x_i x_j$$

$$= \sum_{l=1}^{n}\sum_{m=1}^{n} \omega_{lm}\left(x_l - \mu_{X_l}\right)\left(x_m - \mu_{X_m}\right) + \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sigma^2_{X_{ij}}\sigma^2_{X_{ji}}\right)\xi_i \xi_j + \alpha + \sum_{i=1}^{n} \zeta_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{ij} x_i x_j$$

$$= \mathbf{x}^T \mathbf{\Omega} \mathbf{x} + \mathbf{x}^T \mathbf{\Lambda} \mathbf{x} + \boldsymbol{\zeta}^T \mathbf{x} - 2\boldsymbol{\mu}_X^T \mathbf{\Omega} \mathbf{x} + \boldsymbol{\mu}_X^T \mathbf{\Omega} \boldsymbol{\mu}_X + \boldsymbol{\xi}^T \mathbf{\Psi} \boldsymbol{\xi} + \alpha$$

$$= 0, \tag{46}$$

where

$$\mathbf{\Sigma}_{X_{lm}} = \begin{bmatrix} \Sigma^{lm}_{X_{11}} & \cdots & \Sigma^{lm}_{X_{1n}} \\ \vdots & \ddots & \vdots \\ \Sigma^{lm}_{X_{n1}} & \cdots & \Sigma^{lm}_{X_{nn}} \end{bmatrix}, \; \mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1n} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{nn} \end{bmatrix},$$

$$\mathbf{\Psi} = \begin{bmatrix} \psi_{11} & \cdots & \psi_{1n} \\ \vdots & \ddots & \vdots \\ \psi_{n1} & \cdots & \psi_{nn} \end{bmatrix}, \; \mathbf{\Omega} = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{n1} & \cdots & \omega_{nn} \end{bmatrix}$$

$$\Sigma^{lm}_{X_{ij}} = \frac{1}{4}\left(\sigma^2_{X_{il}} + \sigma^2_{X_{li}}\right)\left(\sigma^2_{X_{jm}} + \sigma^2_{X_{mj}}\right)$$

$$= \sigma^2_{X_{li}}\sigma^2_{X_{jm}}, \; i,j = 1,\ldots,n, \quad l,m = 1,\ldots,n$$

$$\psi_{ij} = 2\sigma^2_{X_{ij}}, \; i,j = 1,\ldots,n$$

$$\omega_{lm} = \boldsymbol{\xi}^T \mathbf{\Sigma}_{X_{lm}} \boldsymbol{\xi}, \; l,m = 1,\ldots,n. \tag{47}$$

Therefore, the Lagrange multipliers α and λ_{ij} must be selected as

$$\alpha = -\boldsymbol{\mu}_X^T \mathbf{\Omega} \boldsymbol{\mu}_X - \boldsymbol{\xi}^T \mathbf{\Psi} \boldsymbol{\xi},$$
$$\boldsymbol{\zeta} = 2\mathbf{\Omega} \boldsymbol{\mu}_X,$$
$$\mathbf{\Lambda} = -\mathbf{\Omega}. \tag{48}$$

Since the second-order variation is positive:

$$K''_{\nabla f_X \nabla f_X}\Big|_{f_X=f_{X^*}} \;=\; 2\frac{\boldsymbol{\xi}\boldsymbol{\xi}^T}{f_{X^*}(\mathbf{x})} \succeq 0, \tag{49}$$

based on Theorem 2, the Gaussian distribution $f_{X^*}(\mathbf{x})$ is necessary optimal for the variational problem in (38). Even though Theorems 1 and 2 are necessary conditions for the minimum, in this case, the multi-variate Gaussian density function is sufficiently the global minimum solution since this is a convex optimization problem (the objective function is strictly convex and its constraint set is convex). ∎

Using similar variational arguments, one can show that a half-normal and a chi density function minimize the Fisher information over the set of non-negative random variables as shown by the following two theorems.

Theorem 7. *Within the class of non-negative continuous random variables with fixed second-order moment m_x^2, the Fisher information is minimized by a half-normal random variable X_{HN}:*

$$J(X) \;\succeq\; J(X_{HN}), \tag{50}$$

where X is an arbitrary (but fixed) non-negative random variable with the second-order moment m_x^2, and $J(\cdot)$ denotes the Fisher information.

Remark 2. *Theorem 7 does not assume the following regularity condition:*

$$\int_0^\infty \nabla f(x) dx = 0. \tag{51}$$

for the Fisher information.

The following result establishes the counterpart of Theorem 7 for the class of non-negative random variables with fixed second order moment and whose distribution satisfies the regularity condition in (51).

Theorem 8 ([15])**.** *Within the class of non-negative continuous random variables X with fixed second-order moment and whose distributions satisfy the regularity condition in (51), the Fisher information is minimized by a chi-distributed random variable X_C:*

$$J(X) \;\succeq\; J(X_C), \tag{52}$$

where $J(\cdot)$ stands for the Fisher information.

 Proof: Unlike the proof in [15], by considering the first-order and the second-order moments instead of variance, we construct a variational problem and address the problem using the first-order and second-order necessary conditions, as well as the convexity property of the problem. The details of the proof are omitted because of the similar steps to those encountered in the proof of Theorem 6. ∎

IV. Worst Additive Noise Lemma

Worst additive noise lemma was introduced and exploited in several references [9], [10], [18], and it has been widely used in numerous other applications. One of the main applications of the worst additive noise lemma pertains to the capacity calculation of a wireless communication channel subject to different constraints such as Gaussian MIMO broadcasting, Gaussian MIMO wire-tap, etc. In this section, the worst additive noise lemma for random vectors will be proved solely based on variational arguments.

Theorem 9. *Assume $\mathbf{X}$ is an arbitrary but fixed random vector and $\mathbf{X}_G$ is a Gaussian random vector, whose mean and correlation matrix are identical to those of $\mathbf{X}$, denoted as $\boldsymbol{\mu}_X$ and $\boldsymbol{\Omega}_X$, respectively. Given a Gaussian random vector $\mathbf{W}_G$, assumed independent of both $\mathbf{X}$ and $\mathbf{X}_G$ and with zero mean and the correlation matrix $\boldsymbol{\Omega}_W$, then the following relation holds:*

$$I(\mathbf{X} + \mathbf{W}_G; \mathbf{W}_G) \geq I(\mathbf{X}_G + \mathbf{W}_G; \mathbf{W}_G). \tag{53}$$

 Proof: Our proof is entirely anchored in the variational calculus framework. A summary of our proof runs as follows. First, we construct a variational problem, which represents the inequality in (53) and required constraints in a functional form. Second, using the first-order variation condition, we find the necessary optimal solutions, which

satisfy Euler's equation. Third, using the second-order variation condition, we show that the optimal solutions are necessarily local minima. Finally, we justify that the local minimum is also global.

By setting $\mathbf{Y} = \mathbf{X} + \mathbf{W}_G$, *where* $\mathbf{X}$ *and* $\mathbf{W}_G$ *are independent of each other, in (53), the mutual information* $I(\mathbf{X} + \mathbf{W}_G; \mathbf{W}_G)$ *can be expressed as*

$$I(\mathbf{X} + \mathbf{W}_G; \mathbf{W}_G) = h(\mathbf{Y}) - h(\mathbf{Y}|\mathbf{W}_G) = h(\mathbf{Y}) - h(\mathbf{X}).$$

Then, we consider the functional:

$$\min_{f_X} - \iint f_X(\mathbf{x}) f_{Y|X}(\mathbf{y}|\mathbf{x}) \log \left(\int f_X(\mathbf{x}) f_{Y|X}(\mathbf{y}|\mathbf{x}) d\mathbf{x} \right) d\mathbf{x} d\mathbf{y} + \iint f_X(\mathbf{x}) f_{Y|X}(\mathbf{y}|\mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x} d\mathbf{y} \quad (54)$$

$$s.\ t.\ \int f_X(\mathbf{x}) d\mathbf{x} = 1,$$

$$\int \mathbf{x} f_X(\mathbf{x}) d\mathbf{x} = \boldsymbol{\mu}_X,$$

$$\int \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) d\mathbf{x} = \boldsymbol{\Omega}_X. \quad (55)$$

The density function $f_Y(\mathbf{y})$ *and conditional density function* $f_{Y|X}(\mathbf{y}|\mathbf{x})$ *are expressed as*

$$f_Y(\mathbf{y}) = \int f_X(\mathbf{x}) f_{Y|X}(\mathbf{y}|\mathbf{x}) d\mathbf{x}, \quad (56)$$

$$f_{Y|X}(\mathbf{y}|\mathbf{x}) = f_W(\mathbf{y} - \mathbf{x}), \quad (57)$$

respectively. Therefore, by substituting $f_Y(\mathbf{y})$ *for* $\int f_X(\mathbf{x}) f_{Y|X}(\mathbf{y}|\mathbf{x}) d\mathbf{x}$ *and* $f_W(\mathbf{y} - \mathbf{x})$ *for* $f_{Y|X}(\mathbf{y}|\mathbf{x})$, *respectively, and appropriately changing the constrains in (55), the variational problem in (54) is expressed as*

$$\min_{f_X, f_Y} \iint f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) \left[-\log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) \right] d\mathbf{x} d\mathbf{y} \quad (58)$$

$$s.\ t.\ \iint f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = 1, \quad (59)$$

$$\iint \mathbf{x} f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = \boldsymbol{\mu}_X, \quad (60)$$

$$\iint \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = \boldsymbol{\Omega}_X, \quad (61)$$

$$\iint \mathbf{y} f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = \boldsymbol{\mu}_Y, \quad (62)$$

$$\iint \mathbf{y}\mathbf{y}^T f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = \boldsymbol{\Omega}_Y, \quad (63)$$

$$f_Y(\mathbf{y}) = \int f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) d\mathbf{x}. \quad (64)$$

Based on Corollary 1, the functional problem in (58) can be re-cast into the following equivalent form:

$$\min_{f_X, f_Y} \int \left(\int f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) [-\log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) + \alpha_0 + \sum_{i=1}^n \zeta_i x_i + \sum_{i=1}^n \sum_{j=1}^n \gamma_{ij} x_i x_j + \sum_{i=1}^n \eta_i y_i \right.$$

$$\left. + \sum_{i=1}^n \sum_{j=1}^n \theta_{ij} y_i y_j - \lambda(\mathbf{y})] d\mathbf{x} \right) + f_Y(\mathbf{y}) \lambda(\mathbf{y}) d\mathbf{y}, \quad (65)$$

where $\mathbf{x}^T = [x_1, \ldots, x_n]$, $\mathbf{y}^T = [y_1, \ldots, y_n]$, *and* α_0, ζ_i, γ_{ij}, η_i, θ_{ij}, *and* $\lambda(\mathbf{y})$ *stand for the Lagrange multipliers corresponding to the constraints (59), (60), (61), (62), (63), and (64), respectively.*

Define now the functional U *as*

$$U[f_X, f_Y] = \int \left(\int K(\mathbf{x}, \mathbf{y}, f_X, f_Y) d\mathbf{x} \right) + \tilde{K}(\mathbf{y}, f_Y) d\mathbf{y},$$

where

$$K(\mathbf{x}, \mathbf{y}, f_X, f_Y) = f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x})[-\log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) + \alpha_0 + \sum_{i=1}^{n} \zeta_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} x_i x_j$$

$$+ \sum_{i=1}^{n} \eta_i y_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \theta_{ij} y_i y_j - \lambda(\mathbf{y})],$$

$$\tilde{K}(\mathbf{y}, f_Y) = \lambda(\mathbf{y}) f_Y(\mathbf{y}). \tag{66}$$

Based on Corollary 2, we can find the optimal solution f_{X^*} and f_{Y^*} as follows:

$$\int K'_{f_X}\Big|_{f_X=f_{X^*}, f_Y=f_{Y^*}} d\mathbf{y} = \int f_W(\mathbf{y} - \mathbf{x})(-\log f_{Y^*}(\mathbf{y}) + \log f_{X^*}(\mathbf{x}) + \alpha_0 + \boldsymbol{\zeta}\mathbf{x}^T + \mathbf{x}^T \boldsymbol{\Gamma} \mathbf{x} + \boldsymbol{\eta}^T \mathbf{y}$$

$$+ \mathbf{y}^T \boldsymbol{\Theta} \mathbf{y} + 1 - \lambda(\mathbf{y})) d\mathbf{y}$$

$$= 0 \tag{67}$$

$$\int K'_{f_Y} d\mathbf{x} + \tilde{K}'_{f_Y}\Big|_{f_X=f_{X^*}, f_Y=f_{Y^*}} = -\int \frac{f_{X^*}(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) d\mathbf{x}}{f_{Y^*}(\mathbf{y})} + \lambda(\mathbf{y})$$

$$= 0, \tag{68}$$

where

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nn} \end{bmatrix}, \quad \boldsymbol{\Theta} = \begin{bmatrix} \theta_{11} & \cdots & \theta_{1n} \\ \vdots & \ddots & \vdots \\ \theta_{n1} & \cdots & \theta_{nn} \end{bmatrix} \tag{69}$$

$\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_n]^T$ *and* $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]^T$.

The following relationships satisfy the necessary conditions (67) and (68):

$$0 = -\log f_{Y^*}(\mathbf{y}) + \log f_{X^*}(\mathbf{x}) + \alpha_0 + \boldsymbol{\zeta}\mathbf{x}^T + \mathbf{x}^T \boldsymbol{\Gamma} \mathbf{x} + \boldsymbol{\eta}^T \mathbf{y} + \mathbf{y}^T \boldsymbol{\Theta} \mathbf{y} + 1 - \lambda(\mathbf{y}),$$

$$0 = -1 + \lambda(\mathbf{y}). \tag{70}$$

Considering the constraints in (59)-(64), $f_{X^*}(\mathbf{x})$ and $f_{Y^*}(\mathbf{y})$ in (70) can be expressed as

$$f_{X^*}(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}_X|^{-\frac{1}{2}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu}_X)^T \boldsymbol{\Sigma}_X^{-1}(\mathbf{x}-\boldsymbol{\mu}_X)}{2}}$$

$$f_{Y^*}(\mathbf{y}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}_Y|^{-\frac{1}{2}} e^{-\frac{(\mathbf{y}-\boldsymbol{\mu}_Y)^T \boldsymbol{\Sigma}_Y^{-1}(\mathbf{y}-\boldsymbol{\mu}_Y)}{2}}$$

where $\boldsymbol{\Sigma}_X = \boldsymbol{\Omega}_X - \boldsymbol{\mu}_X \boldsymbol{\mu}_X^T$, $\boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_W$, and $\boldsymbol{\Sigma}_W$ is the covariance matrix of $\mathbf{W}_G$. Based on the equations in (71), it turns out that

$$\alpha_0 = \frac{1}{2}\log(2\pi)^n |\boldsymbol{\Sigma}_X| + \frac{1}{2}\boldsymbol{\mu}_X^T \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\mu}_X - \frac{1}{2}\log(2\pi)^n |\boldsymbol{\Sigma}_Y| - \frac{1}{2}\boldsymbol{\mu}_Y^T \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{\mu}_Y,$$

$$\boldsymbol{\Gamma} = \frac{1}{2}\boldsymbol{\Sigma}_X^{-1},$$

$$\boldsymbol{\zeta} = -\boldsymbol{\Sigma}_X^{-1} \boldsymbol{\mu}_X,$$

$$\boldsymbol{\Theta} = -\frac{1}{2}\boldsymbol{\Sigma}_Y^{-1},$$

$$\boldsymbol{\eta} = -\boldsymbol{\Sigma}_Y^{-1} \boldsymbol{\mu}_Y. \tag{71}$$

Therefore, f_{X^*} and f_{Y^*} are multi-variate Gaussian density functions (without loss of generality, and we can assume that the covariance matrix $\boldsymbol{\Sigma}_X$ is invertible due to the arguments mentioned in Appendix B).

Now, by confirming the second-order variation condition, we will show that the optimal solutions f_{X^*} and f_{Y^*} are necessarily local minima. Using Corollary 2, we will show that the following matrix is positive semi-definite:

$$\begin{bmatrix} K''_{f_X f_X} & K''_{f_X f_Y} \\ K''_{f_Y f_X} & K''_{f_Y f_Y} \end{bmatrix}\Bigg|_{f_X=f_{X^*}, f_Y=f_{Y^*}} \succeq \mathbf{0}. \tag{72}$$

Since the elements of the matrix in (72) are defined as

$$
\begin{aligned}
K''_{f_X f_X}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} &= \frac{f_w(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})}, \\
K''_{f_Y f_Y}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} &= \frac{f_{X^*}(\mathbf{x})f_w(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})^2}, \\
K''_{f_X f_Y}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} &= -\frac{f_w(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})}, \\
K''_{f_Y f_X}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} &= -\frac{f_w(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})},
\end{aligned}
\tag{73}
$$

the matrix is a positive semi-definite matrix, and therefore $\delta^2 U \geq 0$. Because of the convexity of functional $K(\mathbf{x},\mathbf{y},f_X,f_Y)$ wrt variables f_X and f_Y, the optimal solutions f_{X^*} and f_{Y^*} actually globally minimize the variational functional in (58). Even though these optimal solutions are necessarily optimal, there exists only one solution, which is the multi-variate Gaussian density function and it satisfies Euler's equation in (67) and (68). Therefore, f_{X^*} and f_{Y^*} are also sufficient in this case.

An alternative more detailed proof of the fact that f_{X^*} and f_{Y^*} represent global optimal solutions is to show that $U[f_{\hat{X}}, f_{\hat{Y}}] \geq U[f_{X^*}, f_{Y^*}]$, where $f_{\hat{X}}, f_{\hat{Y}}$ denote any arbitrary functions satisfying the boundary conditions and the constraints. First, the following functionals are defined:

$$
\begin{aligned}
F(\mathbf{x},\mathbf{y},f_X,f_Y) &= f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x})[-\log f_Y(\mathbf{y}) + \log f_X(\mathbf{x})], \\
F_0(\mathbf{x},\mathbf{y},f_X) &= f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}), \\
F_1^{(i)}(\mathbf{x},\mathbf{y},f_X) &= x_i f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}), \\
F_2^{(i,j)}(\mathbf{x},\mathbf{y},f_X) &= x_i x_j f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}), \\
F_3^{(i)}(\mathbf{x},\mathbf{y},f_X) &= y_i f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}), \\
F_4^{(i,j)}(\mathbf{x},\mathbf{y},f_X) &= y_i y_j f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}),
\end{aligned}
$$

and thus $K(\mathbf{x},\mathbf{y},f_X,f_Y)$ can be expressed as

$$
K(\mathbf{x},\mathbf{y},f_X,f_Y) = F(\mathbf{x},\mathbf{y},f_X,f_Y) + \alpha_0 F_0(\mathbf{x},\mathbf{y},f_X) + \sum_{i=1}^n \zeta_i F_1^{(i)}(\mathbf{x},\mathbf{y},f_X) + \sum_{i=1}^n \sum_{j=1}^n \gamma_{ij} F_2^{(i,j)}(\mathbf{x},\mathbf{y},f_X)
$$

$$
+ \sum_{i=1}^n \eta_i F_3^{(i)}(\mathbf{x},\mathbf{y},f_X) + \sum_{i=1}^n \sum_{j=1}^n \theta_{ij} F_4^{(i,j)}(\mathbf{x},\mathbf{y},f_X) - \lambda(\mathbf{y})f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}).
$$

Since the Hessian matrix of $K(\mathbf{x},\mathbf{y},f_X,f_Y)$ wrt f_X and f_Y is given by

$$
\begin{bmatrix}
f_w(\mathbf{y}-\mathbf{x})/f_X(\mathbf{x}) & -f_w(\mathbf{y}-\mathbf{x})/f_Y(\mathbf{y}) \\
-f_w(\mathbf{y}-\mathbf{x})/f_Y(\mathbf{y}) & f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x})/f_Y(\mathbf{y})^2
\end{bmatrix},
$$

which is positive semi-definite, $K(\mathbf{x},\mathbf{y},f_X,f_Y)$ is convex wrt f_X and f_Y, and the following inequality holds

$$
K(\mathbf{x},\mathbf{y},f_{\hat{X}},f_{\hat{Y}}) - K(\mathbf{x},\mathbf{y},f_{X^*},f_{Y^*}) \geq \left[(f_{\hat{X}} - f_{X^*})K'_{f_X} + (f_{\hat{Y}} - f_{Y^*})K'_{f_Y} \right]\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}},
\tag{74}
$$

due to the fact that the convex function lies above its tangents. Therefore, it follows that

$$
\begin{aligned}
&U[f_{\hat{X}}, f_{\hat{Y}}] - U[f_{X^*}, f_{Y^*}] \\
&= \iint F(\mathbf{x}, \mathbf{y}, f_{\hat{X}}, f_{\hat{Y}}) - F(\mathbf{x}, \mathbf{y}, f_{X^*}, f_{Y^*}) d\mathbf{x} d\mathbf{y} \\
&= \iint F(\mathbf{x}, \mathbf{y}, f_{\hat{X}}, f_{\hat{Y}}) - F(\mathbf{x}, \mathbf{y}, f_{X^*}, f_{Y^*}) d\mathbf{x} d\mathbf{y} + \alpha_0 \iint F_0(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_0(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \\
&+ \sum_{i=1}^{n} \zeta_i \iint F_1^{(i)}(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_1^{(i)}(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} + \sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} \iint F_2^{(i,j)}(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_2^{(i,j)}(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \\
&+ \sum_{i=1}^{n} \eta_i \iint F_3^{(i)}(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_3^{(i)}(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} + \sum_{i=1}^{n}\sum_{j=1}^{n} \theta_{ij} \iint F_4^{(i,j)}(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_4^{(i,j)}(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \\
&+ \int \lambda(\mathbf{y}) \left[f_{\hat{Y}}(\mathbf{y}) - \int f_{\hat{X}}(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} \right] d\mathbf{y} - \int \lambda(\mathbf{y}) \left[f_{Y^*}(\mathbf{y}) - \int f_{X^*}(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} \right] d\mathbf{y} \\
&= \iint K(\mathbf{x}, \mathbf{y}, f_{\hat{X}}, f_{\hat{Y}}) - K(\mathbf{x}, \mathbf{y}, f_{X^*}, f_{Y^*}) d\mathbf{x} d\mathbf{y} + \int (f_{\hat{Y}} - f_{Y^*}) \lambda(\mathbf{y}) d\mathbf{y}
\end{aligned}
\tag{75}
$$

Based on (74), the righthand side of (75) can be lower bounded as follows:

$$
\begin{aligned}
&U[f_{\hat{X}}, f_{\hat{Y}}] - U[f_{X^*}, f_{Y^*}] \\
&\geq \iint \left[(f_{\hat{X}} - f_{X^*}) K'_{f_X} + (f_{\hat{Y}} - f_{Y^*}) K'_{f_Y} \right] \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} d\mathbf{x} d\mathbf{y} \\
&\quad + \int (f_{\hat{Y}} - f_{Y^*}) \lambda(\mathbf{y}) d\mathbf{y} \\
&\overset{(a)}{=} \int (f_{\hat{X}} - f_{X^*}) \left[\int K'_{f_X} \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} d\mathbf{y} \right] d\mathbf{x} \\
&\quad + \int (f_{\hat{Y}} - f_{Y^*}) \left[\int K'_{f_Y} d\mathbf{x} + \tilde{K}'_{f_Y} \right] \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} d\mathbf{y} \\
&\overset{(b)}{=} 0,
\end{aligned}
\tag{76}
$$

where (a) follows from the fact that

$$
\tilde{K}'_{f_Y} \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} = \lambda(\mathbf{y}),
$$

and (b) is due to (67) and (68). This proves the sufficiency of the Gaussian distributions, and therefore, f_{X^*} and f_{Y^*} minimize the variational problem.

Remark 3. *The constraints related to the vector means in (60) and (62) are unnecessary. Without these constraints, the optimal solutions are still multi-variate Gaussian density functions but the vector means are equal to zero.*

■

V. Extremal Entropy Inequality

Extremal entropy inequality, proposed by Liu and Viswanath [2], was motivated by multi-terminal information theoretic problems such as the vector Gaussian broadcast channel and the distributed source coding with a single quadratic distortion constraint. EEI is an entropy power inequality which includes a covariance constraint. Because of the covariance constraint, the extremal entropy inequality could not be proved directly by using the classical Entropy Power Inequality (EPI). Therefore, new techniques ([16], [11]) were adopted in the proofs reported in [2], [11]. In this section, the extremal entropy inequality will be proved using a variational approach.

Theorem 10. *Assume that $\mu \geq 1$ is an arbitrary but fixed constant and $\boldsymbol{\Sigma}$ is a positive semi-definite matrix. A Gaussian random vector $\mathbf{W}_G$ with positive definite covariance matrix $\boldsymbol{\Sigma}_w$ is assumed to be independent of an*

arbitrary random vector $\mathbf{X}$ *whose covariance matrix* $\boldsymbol{\Sigma}_X$ *satisfies* $\boldsymbol{\Sigma}_X \preceq \boldsymbol{\Sigma}$. *Then, there exists a Gaussian random vector* $\mathbf{X}_G^*$ *with covariance matrix* $\boldsymbol{\Sigma}_{X^*}$ *which satisfies the following inequality:*

$$h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{W}_G) \leq h(\mathbf{X}_G^*) - \mu h(\mathbf{X}_G^* + \mathbf{W}_G), \tag{77}$$

where $\boldsymbol{\Sigma}_{X^*} \preceq \boldsymbol{\Sigma}$.

Proof: By setting $\mathbf{Y} = \mathbf{X} + \mathbf{W}_G$, *we first consider the following variational problem (without loss of generality, we assume that* $\mathbf{X}$, $\mathbf{W}_G$, *and* $\mathbf{Y}$ *have zero mean):*

$$\min_{f_X, f_Y} \int\int f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x})(-\mu \log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) + \mu(\mu - 1) \log f_w(\mathbf{y} - \mathbf{x})) d\mathbf{x} d\mathbf{y} \tag{78}$$

$$\text{s.t.} \quad \int\int f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = 1,$$

$$\int\int \mathbf{y}\mathbf{y}^T f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = \int\int \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y}$$

$$+ \int\int (\mathbf{y} - \mathbf{x})(\mathbf{y} - \mathbf{x})^T f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y},$$

$$\int\int \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} \preceq \boldsymbol{\Sigma},$$

$$\int\int \mathbf{y}\mathbf{y}^T f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = \boldsymbol{\Sigma}_{Y^*},$$

$$- \int\int f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x} d\mathbf{y} \geq p_x,$$

$$f_Y(\mathbf{y}) = \int f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x}, \tag{79}$$

where p_x *is a constant, and* $\boldsymbol{\Sigma}_{Y^*}$ *stands for the covariance matrix of the optimal solution* $\mathbf{Y}$. *The constraint* $-\int\int f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x} d\mathbf{y} \geq p_x$ *means that the differential entropy of* $\mathbf{X}$ *is greater than a constant* p_x, *i.e.,* $H(\mathbf{X}) \geq p_x$, *and it is introduced because it helps to convexify the problem by enforcing the semi-positive definiteness of the resulting functional second-order variation. This is due to the fact that this constraint introduces an additional Lagrange multiplier* α_1, *which can be selected appropriately to ensure the non-negative definiteness of the second-order variation. Since* p_x *can be any arbitrary small number, we believe that adding this additional constraint is reasonable. In addition, the term* $\mu(\mu - 1) \int\int f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) \log f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = \mu(\mu - 1) h(\mathbf{W}_G)$ *is added to the objective functional (78), and being a constant, it does not affect the optimization problem. Without loss of generality, the matrix* $\boldsymbol{\Sigma}$ *is assumed to be a positive definite matrix due to the same arguments mentioned in [2].*

The optimization problem (78) is re-cast as follows:

$$\min_{f_X, f_Y} \int\int f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x})[-\mu \log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) + \mu(\mu - 1) \log f_w(\mathbf{y} - \mathbf{x})] d\mathbf{x} d\mathbf{y} \tag{80}$$

$$\text{s.t.} \iint f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = 1, \tag{81}$$

$$\iint \left(y_i y_j - x_i x_j - (y - x)_i (y - x)_j \right) f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = 0, \tag{82}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left(\int\int x_i x_j \xi_i \xi_j f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} \right) \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij}^2 \xi_i \xi_j, \tag{83}$$

$$\iint y_i y_j f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} = \sigma_{Y_{ij}^*}^2, \tag{84}$$

$$- \iint f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x} d\mathbf{y} \geq p_x, \tag{85}$$

$$f_Y(\mathbf{y}) = \int f_X(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x}, \tag{86}$$

where the arbitrary deterministic non-zero vector $\boldsymbol{\xi}$ is defined as $[\xi_1, \ldots, \xi_n]^T$, σ_{ij}^2 and $\sigma_{Y_{ij}^*}^2$ denote the ith row and j^{th} column entry of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{Y^*}$ $(i = 1, \ldots, n,$ and $j = 1, \ldots, n)$, respectively.

Using Lagrange multipliers, as shown in Corollary 1, the functional problem in (80) and the constraints in (81)-(86) can be expressed in terms of the Lagrangian:

$$\min_{f_X, f_Y} \quad \int \left(\int K(\mathbf{x}, \mathbf{y}, f_X, f_Y) d\mathbf{x} \right) + \tilde{K}(\mathbf{y}, f_Y) d\mathbf{y}, \tag{87}$$

where

$$K(\mathbf{x}, \mathbf{y}, f_X, f_Y) = f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x})[-\mu \log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) + \mu(\mu - 1) \log f_W(\mathbf{y} - \mathbf{x}) + \alpha_0$$
$$+ \sum_{i=1}^{n} \sum_{j=1}^{n} (\gamma_{ij} y_i y_j - \gamma_{ij} x_i x_j - \gamma_{ij} (y - x)_i (y - x)_j + \theta x_i x_j \xi_i \xi_j + \phi_{ij} y_i y_j) - \alpha_1 \log f_X(\mathbf{x}) - \lambda(\mathbf{y})],$$
$$\tilde{K}(\mathbf{y}, f_Y) = \lambda(\mathbf{y}) f_Y(\mathbf{y}). \tag{88}$$

The Lagrange multipliers α_0, γ_{ij}, θ, ϕ_{ij}, α_1, and $\lambda(\mathbf{y})$ correspond to the constraints in (81), (82), (83), (84), (85), and (86), respectively.

To find the optimal solutions, based on Corollary 2, the first-order variation condition is checked as follows:

$$\int K'_{f_X} \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} d\mathbf{y} = \int f_W(\mathbf{y} - \mathbf{x})[-\mu \log f_{Y^*}(\mathbf{y}) + (1 - \alpha_1) \log f_{X^*}(\mathbf{x}) + \mu(\mu - 1) \log f_W(\mathbf{y} - \mathbf{x}) + \alpha_0$$
$$+ \sum_{i=1}^{n} \sum_{j=1}^{n} (\gamma_{ij} y_i y_j - \gamma_{ij} x_i x_j - \gamma_{ij} (y - x)_i (y - x)_j + \theta x_i x_j \xi_i \xi_j + \phi_{ij} y_i y_j) - \lambda(\mathbf{y}) + 1 - \alpha_1] d\mathbf{y}$$
$$= 0. \tag{89}$$

$$\int K'_{f_Y} d\mathbf{x} + \tilde{K}'_{f_Y} \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} = -\frac{\mu \int f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) d\mathbf{x}}{f_Y(\mathbf{y})} + \lambda(\mathbf{y}) = 0. \tag{90}$$

The following expressions satisfy the equalities in (89) and (90):

$$\lambda(\mathbf{y}) = \mu,$$

$$f_{Y^*}(\mathbf{y}) = (2\pi)^{-\frac{n}{2}} \left| -\frac{\mu}{2} (\boldsymbol{\Gamma} + \boldsymbol{\Phi})^{-1} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{y}^T \left(-\frac{\mu}{2} (\boldsymbol{\Gamma} + \boldsymbol{\Phi})^{-1} \right)^{-1} \mathbf{y} \right\} (2\pi)^{\frac{n}{2}} \left| -\frac{\mu}{2} (\boldsymbol{\Gamma} + \boldsymbol{\Phi})^{-1} \right|^{\frac{1}{2}} \exp\left\{ \frac{c_Y}{\mu} \right\}$$

$$f_W(\mathbf{y} - \mathbf{x}) = (2\pi)^{-\frac{n}{2}} \left| -\frac{\mu(\mu - 1)}{2} \boldsymbol{\Gamma}^{-1} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{x})^T \left(-\frac{\mu(\mu - 1)}{2} \boldsymbol{\Gamma}^{-1} \right)^{-1} (\mathbf{y} - \mathbf{x}) \right\}$$
$$\cdot (2\pi)^{\frac{n}{2}} \left| -\frac{\mu(\mu - 1)}{2} \boldsymbol{\Gamma}^{-1} \right|^{\frac{1}{2}} \exp\left\{ -\frac{c_W}{\mu(\mu - 1)} \right\},$$

$$f_{X^*}(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \left| -\frac{1 - \alpha_1}{2} (\boldsymbol{\Gamma} - \theta \boldsymbol{\Xi})^{-1} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{x}^T \left(-\frac{1 - \alpha_1}{2} (\boldsymbol{\Gamma} - \theta \boldsymbol{\Xi})^{-1} \right)^{-1} \mathbf{x} \right\}$$
$$\cdot (2\pi)^{\frac{n}{2}} \left| -\frac{1 - \alpha_1}{2} (\boldsymbol{\Gamma} - \theta \boldsymbol{\Xi})^{-1} \right|^{\frac{1}{2}} \exp\left\{ \frac{-\alpha_0 + \mu - 1 + \alpha_1 + c_W + c_Y}{1 - \alpha_1} \right\}, \tag{91}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \cdots & \phi_{1n} \\ \vdots & \ddots & \vdots \\ \phi_{n1} & \cdots & \phi_{nn} \end{bmatrix}, \quad \boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nn} \end{bmatrix}$$

$$\boldsymbol{\Xi} = \begin{bmatrix} \xi_1 \xi_1 & \cdots & \xi_1 \xi_n \\ \vdots & \ddots & \vdots \\ \xi_n \xi_1 & \cdots & \xi_n \xi_n \end{bmatrix},$$

$$\mathbf{x} = [x_1, \cdots, x_n]^T,$$

$$\mathbf{y} = [y_1, \cdots, y_n]^T.$$

Now considering the constraints in (81)-(86), the equations in (91) are further processed as follows:

$$f_{Y^*}(y) = (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}_{Y^*}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}_{Y^*}^{-1}\mathbf{y}\right\}$$

$$f_w(y-x) = (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}_w|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{x})^T\mathbf{\Sigma}_w^{-1}(\mathbf{y}-\mathbf{x})\right\}$$

$$f_{X^*}(x) = (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}_{X^*}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_{X^*}^{-1}\mathbf{x}\right\} \tag{92}$$

where

$$
\begin{aligned}
\alpha_0 &= \mu - (1-\alpha_1) + \frac{\mu(\mu-1)}{2}\log(2\pi)^n |\mathbf{\Sigma}_w| \\
&\quad - \frac{\mu}{2}\log(2\pi)^n |\mathbf{\Sigma}_{Y^*}| + \frac{1-\alpha_1}{2}\log(2\pi)^n |\mathbf{\Sigma}_{X^*}|, \\
\mathbf{\Gamma} &= -\frac{\mu(\mu-1)}{2}\mathbf{\Sigma}_w^{-1}, \\
\mathbf{\Phi} &= -\mathbf{\Gamma} - \frac{\mu}{2}\mathbf{\Sigma}_{Y^*}^{-1} \\
&= \frac{\mu(\mu-1)}{2}\mathbf{\Sigma}_w^{-1} - \frac{\mu}{2}(\mathbf{\Sigma}_{X^*}+\mathbf{\Sigma}_w)^{-1}, \\
\mathbf{\Sigma}_{X^*} &= -\frac{1-\alpha_1}{2}(\mathbf{\Gamma}-\theta\mathbf{\Xi})^{-1} \\
&= \frac{1-\alpha_1}{2}\left(\frac{\mu(\mu-1)}{2}\mathbf{\Sigma}_w^{-1}+\theta\mathbf{\Xi}\right)^{-1}, \\
\theta &\geq 0, \tag{93} \\
\alpha_1 &\leq 1-\mu, \tag{94} \\
c_w &= \frac{\mu(\mu-1)}{2}\log(2\pi)^n |\mathbf{\Sigma}_w|, \\
c_Y &= -\frac{\mu}{2}\log(2\pi)^n |\mathbf{\Sigma}_{Y^*}|, \\
|\mathbf{\Sigma}_{X^*}| &= \left(\frac{1}{2\pi e}\exp\left\{\frac{2}{n}p_x\right\}\right)^n.
\end{aligned}
$$

The inequality in (94) is due to the second-order variation condition, which will be presented later in this proof. The inequality (93) is based on the theory of KKT conditions since the multiplier associated with the inequality constraint is nonnegative. Moreover, the complementary slackness condition in the KKT conditions leads to the following relationship:

$$\theta\left[\iint\left(\sum_{i=1}^{n}\sum_{j=1}^{n}x_ix_j\xi_i\xi_j\right)f_{X^*}(\mathbf{x})f_w(\mathbf{y}-\mathbf{x})d\mathbf{x}d\mathbf{y} - \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}^2\xi_i\xi_j\right] = 0. \tag{95}$$

Based on Corollary 2, to make the second variation nonnegative, the positive semi-definiteness of the following matrix is required:

$$\begin{bmatrix} K''_{f_{X^*}f_{X^*}} & K''_{f_{X^*}f_{Y^*}} \\ K''_{f_{Y^*}f_{X^*}} & K''_{f_{Y^*}f_{Y^*}} \end{bmatrix}, \tag{96}$$

which further reduces to the following condition:

$$
\begin{aligned}
&\begin{bmatrix} h_x & h_Y \end{bmatrix}\begin{bmatrix} K''_{f_{X^*}f_{X^*}} & K''_{f_{X^*}f_{Y^*}} \\ K''_{f_{Y^*}f_{X^*}} & K''_{f_{Y^*}f_{Y^*}} \end{bmatrix}\begin{bmatrix} h_x \\ h_Y \end{bmatrix} \\
&= K''_{f_{X^*}f_{X^*}}h_x^2 + K''_{f_{Y^*}f_{Y^*}}h_Y^2 + (K''_{f_{X^*}f_{Y^*}} + K''_{f_{Y^*}f_{X^*}})h_Y h_x \\
&\geq 0, \tag{97}
\end{aligned}
$$

where h_X and h_Y are arbitrary admissible functions. Since $K''_{f_{X^*}f_{X^*}}$, $K''_{f_{X^*}f_{Y^*}}$, $K''_{f_{Y^*}f_{X^*}}$, and $K''_{f_{Y^*}f_{Y^*}}$ are defined as

$$
\begin{aligned}
K''_{f_{X^*}f_{X^*}} &= \frac{(1-\alpha_1)f_w(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})}, \\
K''_{f_{X^*}f_{Y^*}} &= -\frac{\mu f_w(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})}, \\
K''_{f_{Y^*}f_{X^*}} &= -\frac{\mu f_w(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})}, \\
K''_{f_{Y^*}f_{Y^*}} &= \frac{\mu f_{X^*}(\mathbf{x})f_w(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})^2},
\end{aligned}
\tag{98}
$$

the condition in (97) requires

$$
\frac{(1-\alpha_1)f_w(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})}h_X(\mathbf{x})^2 - 2\frac{\mu f_w(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})}h_X(\mathbf{x})h_Y(\mathbf{y}) + \frac{\mu f_{X^*}(\mathbf{x})f_w(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})^2}h_Y(\mathbf{y})^2
$$
$$
\geq \frac{\mu f_w(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})}\left(h_X(\mathbf{x}) - \frac{f_{X^*}(\mathbf{x})}{f_{Y^*}(\mathbf{y})}h_Y(\mathbf{y})\right)^2,
\tag{99}
$$

which holds true if $1-\alpha_1 \geq \mu$ (i.e., $\alpha_1 \leq 1-\mu \leq 0$). Condition $\alpha_1 \leq 0$ is also imposed by the KKT complementary slackness condition corresponding to the constraint (85). Therefore, the optimal solutions f_{X^*} and f_{Y^*} minimize the functional problem in (80), and the proof is completed because of convexity of the functional $K(\mathbf{x},\mathbf{y},f_X,f_Y)$ wrt variables f_X and f_Y.

A more detailed alternative justification of the fact the Gaussian distributions f_{X^*} and f_{Y^*} are global minima is next presented. We will prove the sufficiency of the Gaussian distributions by showing $U[f_{\hat{X}}, f_{\hat{Y}}] \geq U[f_{X^*}, f_{Y^*}]$, where $U[\cdot, \cdot]$ represents the objective functional in the problem and $f_{\hat{X}}, f_{\hat{Y}}$ denote any arbitrary functions satisfying the boundary conditions and the constraints. First, the following functionals are defined:

$$
\begin{aligned}
F(\mathbf{x},\mathbf{y},f_X,f_Y) &= f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x})(-\mu\log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) + \mu(\mu-1)\log f_w(\mathbf{y}-\mathbf{x})), \\
F_0(\mathbf{x},\mathbf{y},f_X) &= f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}), \\
F_1^{(i,j)}(\mathbf{x},\mathbf{y},f_X) &= \left(y_iy_j - x_ix_j - (y-x)_i(y-x)_j\right)f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}), \\
F_2(\mathbf{x},\mathbf{y},f_X) &= \left(\sum_{i=1}^n\sum_{j=1}^n x_ix_j\xi_i\xi_j\right)f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}), \\
F_3^{(i,j)}(\mathbf{x},\mathbf{y},f_X) &= y_iy_j f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}), \\
F_4(\mathbf{x},\mathbf{y},f_X) &= -f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x})\log f_X(\mathbf{x}),
\end{aligned}
$$

and thus

$$
\begin{aligned}
K(\mathbf{x},\mathbf{y},f_X,f_Y) &= F(\mathbf{x},\mathbf{y},f_X,f_Y) + \alpha_0 F_0(\mathbf{x},\mathbf{y},f_X) + \sum_{i=1}^n\sum_{j=1}^n \gamma_{ij}F_1^{(i,j)}(\mathbf{x},\mathbf{y},f_X) + \theta F_2(\mathbf{x},\mathbf{y},f_X) \\
&+ \sum_{i=1}^n\sum_{j=1}^n \phi_{ij}F_3^{(i,j)}(\mathbf{x},\mathbf{y},f_X) + \alpha_1 F_4(\mathbf{x},\mathbf{y},f_X) - \lambda(\mathbf{y})f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x}).
\end{aligned}
$$

It can be verified that the Hessian matrix of $K(\mathbf{x},\mathbf{y},f_X,f_Y)$ w.r.t f_X and f_Y is given by

$$
\begin{bmatrix}
(1-\alpha_1)f_w(\mathbf{y}-\mathbf{x})/f_X(\mathbf{x}) & -\mu f_w(\mathbf{y}-\mathbf{x})/f_Y(\mathbf{y}) \\
-\mu f_w(\mathbf{y}-\mathbf{x})/f_Y(\mathbf{y}) & \mu f_X(\mathbf{x})f_w(\mathbf{y}-\mathbf{x})/f_Y(\mathbf{y})^2
\end{bmatrix},
$$

which is positive semi-definite due to (94). The convexity property of $K(\mathbf{x},\mathbf{y},f_X,f_Y)$ yields that

$$
K(\mathbf{x},\mathbf{y},f_{\hat{X}},f_{\hat{Y}}) - K(\mathbf{x},\mathbf{y},f_{X^*},f_{Y^*}) \geq \left[(f_{\hat{X}}-f_{X^*})K'_{f_X} + (f_{\hat{Y}}-f_{Y^*})K'_{f_Y}\right]\Big|_{f_X=f_{X^*}, f_Y=f_{Y^*}},
\tag{100}
$$

and it follows that

$$U[f_{\hat{X}}, f_{\hat{Y}}] - U[f_{X^*}, f_{Y^*}]$$

$$= \iint F(\mathbf{x}, \mathbf{y}, f_{\hat{X}}, f_{\hat{Y}}) - F(\mathbf{x}, \mathbf{y}, f_{X^*}, f_{Y^*}) d\mathbf{x} d\mathbf{y}$$

$$\overset{(a)}{\geq} \iint F(\mathbf{x}, \mathbf{y}, f_{\hat{X}}, f_{\hat{Y}}) - F(\mathbf{x}, \mathbf{y}, f_{X^*}, f_{Y^*}) d\mathbf{x} d\mathbf{y} + \alpha_0 \left[\iint F_0(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_0(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \right]$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \left[\iint F_1^{(i,j)}(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_1^{(i,j)}(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \right] + \theta \left[\iint F_2(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_2(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \right]$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{n} \phi_{ij} \left[\iint F_3^{(i,j)}(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_3^{(i,j)}(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \right] + \alpha_1 \left[\iint F_4(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_4(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \right]$$

$$+ \int \lambda(\mathbf{y}) \left[f_{\hat{Y}}(\mathbf{y}) - \int f_{\hat{X}}(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} \right] d\mathbf{y} - \int \lambda(\mathbf{y}) \left[f_{Y^*}(\mathbf{y}) - \int f_{X^*}(\mathbf{x}) f_w(\mathbf{y} - \mathbf{x}) d\mathbf{x} \right] d\mathbf{y} \qquad (101)$$

$$= \iint K(\mathbf{x}, \mathbf{y}, f_{\hat{X}}, f_{\hat{Y}}) - K(\mathbf{x}, \mathbf{y}, f_{X^*}, f_{Y^*}) d\mathbf{x} d\mathbf{y} + \int \lambda(\mathbf{y}) \left(f_{\hat{Y}}(\mathbf{y}) - f_{Y^*}(\mathbf{y}) \right) d\mathbf{y}$$

$$\overset{(b)}{\geq} \iint \left[(f_{\hat{X}} - f_{X^*}) K'_{f_X} + (f_{\hat{Y}} - f_{Y^*}) K'_{f_Y} \right] \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} d\mathbf{x} d\mathbf{y} + \int \lambda(\mathbf{y}) \left(f_{\hat{Y}}(\mathbf{y}) - f_{Y^*}(\mathbf{y}) \right) d\mathbf{y}$$

$$= \int (f_{\hat{X}} - f_{X^*}) \left[\int K'_{f_X} \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} d\mathbf{y} \right] d\mathbf{x} + \int (f_{\hat{Y}} - f_{Y^*}) \left[\int K'_{f_Y} \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} d\mathbf{x} + \lambda(\mathbf{y}) \right] d\mathbf{y}$$

$$\overset{(c)}{=} 0, \qquad (102)$$

where the inequality (a) follows from the complementary slackness condition in the KKT conditions (95). Indeed, since $f_{\hat{X}}$ only represents an arbitrary feasible solution and $\theta \geq 0$, it follows that

$$\theta \left[\iint F_2(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} - \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij}^2 \xi_i \xi_j \right] = 0,$$

and

$$\theta \left[\iint F_2(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) d\mathbf{x} d\mathbf{y} - \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij}^2 \xi_i \xi_j \right] \leq 0,$$

and therefore, $\theta \left[\iint F_2(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_2(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \right] \leq 0$. Similarly, the complementary slackness condition associated with (85) leads to $\alpha_1 \left[\iint F_4(\mathbf{x}, \mathbf{y}, f_{\hat{X}}) - F_4(\mathbf{x}, \mathbf{y}, f_{X^*}) d\mathbf{x} d\mathbf{y} \right] \leq 0$. In addition, (b) is due to (100), and (c) follows from (89) and (90). This proves the sufficiency of Gaussian distributions.

Remark 4. *The proposed proof only exploits calculus of variations tools. Unlike the previous proofs, this proof does not adopt neither the channel enhancement technique and EPI as in [2] nor the EPI and data processing inequality as in [11].*

∎

Theorem 11. *Assume that $\mu \geq 1$ is an arbitrary but fixed constant and $\mathbf{\Sigma}$ is a positive semi-definite matrix. Independent Gaussian random vectors $\mathbf{W}_G$ with covariance matrix $\mathbf{\Sigma}_w$ and $\mathbf{V}_G$ with covariance matrix $\mathbf{\Sigma}_v$ are assumed to be independent of an arbitrary random vector $\mathbf{X}$ with covariance matrix $\mathbf{\Sigma}_X \preceq \mathbf{\Sigma}$. Both covariance matrices $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_v$ are assumed to be positive definite. Then, there exists a Gaussian random vector $\mathbf{X}_G^*$ with covariance matrix $\mathbf{\Sigma}_{X^*}$ which satisfies the following inequality:*

$$h(\mathbf{X} + \mathbf{W}_G) - \mu h(\mathbf{X} + \mathbf{V}_G) \leq h(\mathbf{X}_G^* + \mathbf{W}_G) - \mu h(\mathbf{X}_G^* + \mathbf{V}_G), \qquad (103)$$

where $\mathbf{\Sigma}_{X^*} \preceq \mathbf{\Sigma}$.

Proof: See Appendix C.

Remark 5. *The proposed proof does not borrow any techniques from [2]. Even though the proposed proof adopts the equality condition for the data processing inequality, a result which was also exploited in [11], the proposed proof is different from the one in [11] due to the following features. First, the proposed proof uses the equality condition of the data processing inequality only once while the proof in [11] uses it twice. The proof in [2] exploited the channel enhancement technique twice, which is equivalent to using the equality condition in the data processing inequality. Second, the proposed proof does not use the moment generating function technique unlike the proof proposed in [11]; instead the current proof directly exploits a property of the conditional mutual information pertaining to a Markov chain.*

∎

VI. APPLICATIONS

Because of the easiness to incorporate a broad class of constraints, the proposed variational framework finds usage in a large number of applications. Herein section, we will briefly illustrate some potential applications in this regard and state several open research problems which might be also addressed within the considered functional framework.

A. Gaussian Wire-tap Channel

The secrecy capacity of Gaussian wire-tap channel has been studied by many researchers [20], [33]. We will approach the Gaussian wire-tap problem from the estimation viewpoint, rather than considering the secrecy capacity from an information theoretic perspective.

The following scalar Gaussian wire-tap channel is considered:

$$\begin{aligned} Y_1 &= aX + W_G, \\ Y_2 &= aX + W_G + Z_G, \end{aligned} \tag{104}$$

where X is an arbitrary but fixed random variable with zero mean and unit variance, a is a constant, and W_G and Z_G are Gaussian random variables with variances σ_w^2 and σ_z^2, respectively. The random variables W_G and Z_G are independent of each other, and they have zero mean. In the channel model (104), Y_1 and Y_2 are considered as a legitimate receiver and as an eavesdropper, respectively. The goal of this problem is the following. Assume that both receivers use minimum mean square error (MMSE) estimators. Given the value of the mean square error (MSE), which allows to correctly decode the legitimate receiver, what is the optimal distribution which maximizes the difference between the MSE in the legitimate receiver and the MSE in the eavesdropper?

The above mentioned problem adopts both practical and reasonable assumptions due to the following reasons. First, the MMSE estimator is an optimal estimator in the sense that it minimizes the MSE. Therefore, it is reasonable to use such an optimal estimator. Second, to prevent from eavesdropping, finding the signal distribution that maximizes the difference between the MSEs corresponding to the legitimate receiver and the eavesdropper, respectively, represents a legitimate design objective. To find the optimal distribution, the following functional problem is constructed:

$$\begin{aligned} \max_{f_X(x)} \quad & Var(X|Y_2) - Var(X|Y_1), \\ \text{s.t.} \quad & Var(X|Y_1) = R, \end{aligned} \tag{105}$$

where $Var(X|Y) = \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2\right]$, $\mathbb{E}[\cdot]$ denotes the expectation operator, and R is a constant.

The optimization problem in (105) is expressed as

$$\max_{f_X(x)} \quad Var(\mathbb{E}[X|Y_1]|Y_2), \tag{106}$$

$$\text{s.t.} \quad \mathbb{E}\left[\mathbb{E}[X|Y_1]^2\right] = 1 - R. \tag{107}$$

The equation in (106) is due to the total law of variance and the Markov chain $X \to Y_1 \to Y_2$. Since $\mathbb{E}[X^2] = 1$, the equation (107) follows from the constraint in (105).

The objective function in (106) is further expressed as

$$Var\left(\mathbb{E}\left[X|Y_1\right]|Y_2\right) = \mathbb{E}\left[\mathbb{E}\left[X|Y_1\right]^2\right] - \mathbb{E}\left[\mathbb{E}\left[X|Y_2\right]^2\right] \tag{108}$$

and using the equations (107), (108), the optimization problem in (106) is re-formulated in terms of the following variational problem:

$$\min_{f_{Y_2},g} \int \frac{1}{f_{Y_2}(y)} g(y)^2 dy, \tag{109}$$

$$\int y^2 f_{Y_2}(y) dy = m_{Y_2}^2, \tag{110}$$

$$g(y) = \int x f_{Y_2|X}(y|x) f_X(x) dx, \tag{111}$$

where $f_X(x)$ and $f_{Y_2}(y)$ are the probability density functions of X and Y_2, respectively, and $m_{Y_2}^2$ stands for the second-order moment of Y_2.

Since the first term in (108) is given and

$$\mathbb{E}\left[\mathbb{E}\left[X|Y_2\right]^2\right] = \int f_{Y_2}(y) \left(\int x \frac{f_{Y_2|X}(y|x) f_X(x)}{f_{Y_2}(y)} dx\right)^2 dy,$$

the objective function in (109) is derived from the equation (106). Also, the additional constraint in (110) is required to solve this variational problem.

Considering the Lagrange multipliers λ_1 and $\lambda(y)$ to account for the constraints in (110) and (111), respectively, the following variational problem is constructed:

$$\int K(y, f_{Y_2}, g) dy,$$

where

$$K(y, f_{Y_2}, g) = \frac{g(y)^2}{f_{Y_2}(y)} + \lambda_1 y^2 f_{Y_2}(y) + \lambda(y) \left(g(y) - \int x f_{Y_2|X}(y|x) f_X(x) dx\right). \tag{112}$$

In accordance with Theorem 1, we can determine g^* and $f_{Y_2}^*$ to enforce the first-order variation to be zero:

$$K_{f_{Y_2}^*} = -\frac{g^*(y)^2}{f_{Y_2}^*(y)^2} + \lambda_1 y^2 = 0, \tag{113}$$

$$K_{g^*} = \frac{2g^*(y)}{f_{Y_2}^*(y)} + \lambda(y) = 0,$$

Taking into account (113), it follows further that

$$\mathbb{E}\left[X^*|Y_2^*\right] = \frac{g^*(y)}{f_{Y_2}^*(y)} = \sqrt{\lambda_1} y. \tag{114}$$

Since $\mathbb{E}[X^*|Y_2^*]$, the MMSE estimator, is a linear function of y and the channel is corrupted with additive Gaussian noise, it is necessary that X^* is a Gaussian random variable. Based on Theorem 2, it can be verified that the second-order variation is nonnegative. Moreover, due to the convexity of $K(y, f_{Y_2}, g)$ wrt f_{Y_2} and g, we can confirm that the Gaussian solution is optimal, and the proof is completed.

B. Additional Applications

The importance of the variational framework in establishing some fundamental information theoretic inequalities was already illustrated herein paper. At their turn, these information theoretic inequalities played a fundamental role in establishing other important results and applications. For example, the minimum Fisher information theorem (Cramér-Rao inequality) and maximum entropy theorem were used for developing min-max robust estimation techniques [25], results which were recently further extended to the more general framework of noise with arbitrary distribution (and correlation) in [27] and used to explain why the MIMO channel estimation scheme proposed in [26] exhibits a min-max robustness property. Along the same line of potential applications, the extensions of the

maximum entropy and minimum Fisher information results to positive random variables, as stated in Theorems 5, 7 and 8, play a fundamental role in developing robust clock synchronization algorithms for wireless sensor networks and other wireless networks that rely on message exchanges to acquire the timing information. A large class of clock synchronization protocols (see e.g., TPSN, Internet, PBS [28]) rely on the two-way message exchange mechanism and for which the timing synchronization approach reduces to estimating a linear regression model for which the distribution of additive noise has positive support but it is otherwise arbitrary [28]. Designing robust timing synchronization algorithms for such protocols is difficult, because of the variability of delay distributions caused by the variable network traffic. However, this problem can now be resolved at the light of the results brought by Theorems 5, 7 and 8. By optimizing the design of timing messages for the scenario of a chi or log-normal distributed delay, then min-max robust time synchronization algorithms could be developed.

The extremal entropy inequality was used in the vector Gaussian broadcast channel [2], the distributed source coding with a single quadratic distortion constraint problem [2], the Gaussian wire-tap channel [11], and many other problems. Even though these applications were traditionally addressed using the information theoretic inequalities, one can directly approach these applications by means of the proposed variational calculus techniques. One of the benefits of such a variational approach is the fact that it can cope with many types of constraints as opposed to the EEI which is still quite rigid in its formulation. As Prof. Max Costa suggested the authors of this paper in a private communication, in the context of Z Gaussian interference channels, such a variational approach might be helpful to develop novel entropy-power-like inequalities, where the limiting variables are Gaussian and independent but not anymore identically distributed, and to assess the capacity of the Z-Gaussian interference channel.

Additional important extensions of maximum entropy theorem, minimum Fisher information theorem, additive worst noise lemma, and extremal entropy inequality might be envisioned within the proposed variational framework by imposing various restrictions on the range of values assumed by random variables/vectors (e.g., random variables whose support is limited to a finite length interval or finite set of values) or on their second or higher-order moments and correlations. For example, the problem of finding the worst additive noise under a covariance constraint [9] as well as establishing multivariate extensions of Costa's entropy power inequality [30] along the lines mentioned by Liu et al. [21] and Palomar [31], [32] might be also addressed within the proposed variational framework. However, all these challenges together with finding a variational proof of EPI remain open research problems for future study.

VII. Conclusions

In this paper, we derived several fundamental information theoretic inequalities using a functional analysis framework. The main benefit for employing calculus of variations is due to the fact for any information theoretic inequality as long as it can be expressed in terms of a convex functional, the global optimal solution can be obtained from the necessary conditions. A brief summary of this paper contributions is the following. First, the entropy maximizing theorem and Fisher information minimizing theorem were derived under different assumptions. Second, the worst additive noise lemma was proved from the perspective of a functional problem. Third, the extremal entropy inequality was derived using calculus of variations techniques. Finally, applications and possible extensions that could be addressed within the proposed variational framework were briefly presented. Many open research problems were also formulated.

Appendix A
Proof of Corollaries 1 and 2

Even though the functionals in Corollary 1 involve double integrations, they can be regarded as a special case of the functionals in Theorem 3. For example, the functional $U[f_X, f_Y]$ in (19) can be considered as $\int_a^b G(y, f_Y) dy$ where $G(y, f_Y) = \int_a^b K(x, y, f_X, f_Y) dx$. In this way, the augmented functional is given by

$$J[f_X, f_Y] = \int_a^b \left[\int_a^b K(x, y, f_X, f_Y) dx + \sum_{i=1}^n \int_a^b \tilde{L}_i(x, y, f_X, f_Y) dx + \lambda(y) \left(g(y, f_Y) - \int_a^b \tilde{k}(x, y, f_X) dx \right) \right] dy$$

$$= \int_a^b \left\{ \left[\int_a^b (K(x, y, f_X, f_Y) + \sum_{i=1}^n \lambda_i \tilde{L}_i(x, y, f_X, f_Y) - \lambda(y) \tilde{k}(x, y, f_X)) dx \right] + \lambda(y) g(y, f_Y) \right\} dy.$$

This completes the proof of Corollary 1.

Based on the definitions in Section II, the first-order variation of the above augmented functional can be calculated as

$$
\delta J[f_X, f_Y] = \int_a^b \int_a^b \left\{ \frac{\partial K(x, y, f_X, f_Y)}{\partial f_X} \eta(x) + \frac{\partial K(x, y, f_X, f_Y)}{\partial f_Y} \xi(y) + \sum_{i=1}^n \left[\frac{\partial \tilde{L}_i(x, y, f_X, f_Y)}{\partial f_X} \eta(x) + \right. \right.
$$
$$
\left. \frac{\partial \tilde{L}_i(x, y, f_X, f_Y)}{\partial f_Y} \xi(y) \right] - \lambda(y) \frac{\partial \tilde{k}(x, y, f_X)}{\partial f_X} \eta(x) \right\} dx dy + \int_a^b \lambda(y) \frac{\partial g(y, f_Y)}{\partial f_Y} \xi(y) dy
$$
$$
= \int_a^b \left\{ \int_a^b \frac{\partial K(x, y, f_X, f_Y)}{\partial f_X} + \sum_{i=1}^n \frac{\partial \tilde{L}_i(x, y, f_X, f_Y)}{\partial f_X} - \lambda(y) \frac{\partial \tilde{k}(x, y, f_X)}{\partial f_X} dy \right\} \eta(x) dx \quad (115)
$$
$$
+ \int_a^b \left\{ \int_a^b \frac{\partial K(x, y, f_X, f_Y)}{\partial f_Y} + \sum_{i=1}^n \frac{\partial \tilde{L}_i(x, y, f_X, f_Y)}{\partial f_Y} dx + \lambda(y) \frac{\partial g(y, f_Y)}{\partial f_Y} \right\} \xi(y) dy,
$$

where $\eta(x)$ and $\xi(y)$ represent any admissible increments for f_X and f_Y, respectively. Due to Theorem 1, a necessary condition for the function $J[f_X, f_Y]$ to have an extremum for given functions f_{X^*} and f_{Y^*} is that $\delta J[f_X, f_Y]$ vanishes at f_{X^*} and f_{Y^*} for any admissible $\eta(x)$ and $\xi(y)$. This leads to

$$
\int K'_{f_{X^*}}(x, y, f_{X^*}, f_{Y^*}) + \sum_{i=1}^n \lambda_i \tilde{L}_{i f_{X^*}}'(x, y, f_{X^*}, f_{Y^*}) - \lambda(y) \tilde{k}'_{f_{X^*}}(x, y, f_{X^*}) dy = 0,
$$
$$
\int K'_{f_{Y^*}}(x, y, f_{X^*}, f_{Y^*}) + \sum_{i=1}^n \lambda_i \tilde{L}_{i f_{Y^*}}'(x, y, f_{X^*}, f_{Y^*}) dx + \lambda(y) g'_{f_{Y^*}}(y, f_{Y^*}) = 0,
$$

which are exactly (22) and (23).

In order to calculate the second-order variation of $J[f_X, f_Y]$ from the first-order variation (115), we rewrite the term $\lambda(y) \frac{\partial g(y, f_Y)}{\partial f_Y}$ in (115) as $\int_a^b q(x) \lambda(y) \frac{\partial g(y, f_Y)}{\partial f_Y} dx$, where $q(x)$ is an arbitrary but fixed function satisfying $\int_a^b q(x) dx = 1$. Thus, the first-order variation (115) can be rewritten as

$$
\int_a^b \left\{ \int_a^b \frac{\partial K(x, y, f_X, f_Y)}{\partial f_X} + \sum_{i=1}^n \frac{\partial \tilde{L}_i(x, y, f_X, f_Y)}{\partial f_X} - \lambda(y) \frac{\partial \tilde{k}(x, y, f_X)}{\partial f_X} dy \right\} \eta(x) dx
$$
$$
+ \int_a^b \left\{ \int_a^b \frac{\partial K(x, y, f_X, f_Y)}{\partial f_Y} + \sum_{i=1}^n \frac{\partial \tilde{L}_i(x, y, f_X, f_Y)}{\partial f_Y} + q(x) \lambda(y) \frac{\partial g(y, f_Y)}{\partial f_Y} dx \right\} \xi(y) dy \quad (116)
$$

Based on (116), the second-order variation of $J[f_X, f_Y]$ is derived as

$$
\delta^2 J[f_X, f_Y] = \int_a^b \int_a^b \begin{bmatrix} \eta(x) & \xi(y) \end{bmatrix} \begin{bmatrix} G''_{f_X f_X} & G''_{f_X f_Y} \\ G''_{f_Y f_X} & G''_{f_Y f_Y} \end{bmatrix} \begin{bmatrix} \eta(x) \\ \xi(y) \end{bmatrix} dx dy,
$$

where

$$
G(x, y, f_{X^*}, f_{Y^*}) = K(x, y, f_{X^*}, f_{Y^*}) + \sum_{i=1}^N \lambda_i \tilde{L}_i(x, y, f_{X^*}, f_{Y^*}) - \lambda(y) \tilde{k}(x, y, f_{X^*}) + \lambda(y) g(y, f_{Y^*}) q(x),
$$

Since a necessary condition for the functional $J[f_X, f_Y]$ to have a minimum for given functions f_{X^*} and f_{Y^*} is that $\delta^2 J[f_X, f_Y] \geq 0$, this leads to the positive semi-definiteness of

$$
\begin{bmatrix} G''_{f_X f_X} & G''_{f_X f_Y} \\ G''_{f_Y f_X} & G''_{f_Y f_Y} \end{bmatrix}
$$

and completes the proof of Corollary 2.

APPENDIX B
NON-INVERTIBLE CORRELATION (OR COVARIANCE) MATRIX

Let $\mathbf{\Omega}_x = \mathbf{Q}_\Omega \mathbf{\Lambda}_\Omega \mathbf{Q}_\Omega^T$ and $\bar{\mathbf{X}} = \mathbf{Q}_\Omega^T \mathbf{X} = [\bar{\mathbf{X}}_a^T, \bar{\mathbf{X}}_b^T]$, where $\mathbf{\Lambda}_\Omega = diag(\Lambda_1, \ldots, \Lambda_m, 0, \ldots, 0)$, $\mathbf{\Omega}_x$ is a singular matrix, $\mathbf{Q}_\Omega$ is an orthogonal matrix, and $diag(\cdot)$ denotes a diagonal matrix. The correlation matrix of $\bar{\mathbf{X}}_b$ is the zero matrix, and therefore, it is considered as a deterministic vector. Without loss of generality, we can assume $\bar{\mathbf{X}}_b = \mathbf{0}$. The following matrices are also considered:

$$\mathbf{Q}_\Omega^T \mathbf{\Omega}_W \mathbf{Q}_\Omega = \left[\begin{array}{cc} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{C} \end{array} \right],$$
$$\mathbf{D} = \left[\begin{array}{cc} \mathbf{I} & -\mathbf{B}^T \mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{I} \end{array} \right], \tag{117}$$

where the dimensions of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are $m \times m$, $(n-m) \times m$, and $(n-m) \times (n-m)$, respectively. Then,

$$\mathbf{DQ}_\Omega^T \mathbf{X} = \left[\begin{array}{cc} \mathbf{I} & -\mathbf{B}^T \mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{I} \end{array} \right] \left[\begin{array}{c} \bar{\mathbf{X}}_a \\ \mathbf{0} \end{array} \right] = \left[\begin{array}{c} \bar{\mathbf{X}}_a \\ \mathbf{0} \end{array} \right],$$
$$\mathbf{DQ}_\Omega^T \mathbf{W}_G = \left[\begin{array}{c} \bar{\mathbf{W}}_{G_a} \\ \bar{\mathbf{W}}_{G_b} \end{array} \right],$$
$$\mathbb{E} \left[\mathbf{DQ}_\Omega^T \mathbf{W}_G \mathbf{W}_G^T \mathbf{Q}_\Omega \mathbf{D}^T \right] = \left[\begin{array}{cc} \mathbf{A} - \mathbf{B}^T \mathbf{C}^{-1} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{array} \right]. \tag{118}$$

Due to (118), the random vectors $\bar{\mathbf{W}}_{G_a}$ and $\bar{\mathbf{W}}_{G_b}$ are statistically independent of each other.

The left-hand side of the equation in (53) can be re-expressed as

$$\begin{aligned} h(\mathbf{X} + \mathbf{W}_G) - h(\mathbf{X}) &= h(\mathbf{DQ}_\Omega^T \mathbf{X} + \mathbf{DQ}_\Omega^T \mathbf{W}_G) - h(\mathbf{DQ}_\Omega^T \mathbf{X}) \\ &= h(\bar{\mathbf{X}}_a + \bar{\mathbf{W}}_{G_a}, \bar{\mathbf{X}}_b + \bar{\mathbf{W}}_{G_b}) - h(\bar{\mathbf{X}}_a, \bar{\mathbf{X}}_b) \\ &= h(\bar{\mathbf{X}}_a + \bar{\mathbf{W}}_{G_a}) - h(\bar{\mathbf{X}}_a) + \underbrace{h(\bar{\mathbf{X}}_b + \bar{\mathbf{W}}_{G_b}) - h(\bar{\mathbf{X}}_b)}_{(a)}. \end{aligned} \tag{119}$$

In (119), $\bar{\mathbf{X}}_b$ is considered as a deterministic variable, $\bar{\mathbf{W}}_{G_b}$ is given, the term (a) can be ignored in the optimization, and the correlation matrix of $\bar{\mathbf{X}}_a$ is non-singular. Therefore, we can always assume the correlation matrix to be invertible.

APPENDIX C
PROOF OF THEOREM 11

Proof: First, choose a Gaussian random vector $\tilde{\mathbf{W}}_G$ whose covariance matrix $\mathbf{\Sigma}_{\tilde{W}}$ satisfies $\mathbf{\Sigma}_{\tilde{W}} \preceq \mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_{\tilde{W}} \preceq \mathbf{\Sigma}_V$. Since the Gaussian random vectors $\mathbf{V}_G$ and $\mathbf{W}_G$ can be represented as the summation of two independent random vectors $\tilde{\mathbf{W}}_G$ and $\hat{\mathbf{V}}_G$, and the summation of two independent random vectors $\tilde{\mathbf{W}}_G$ and $\hat{\mathbf{W}}_G$, respectively, the left-hand side of the equation in (103) is written as follows:

$$\begin{aligned} &\mu h(\mathbf{X} + \mathbf{V}_G) - h(\mathbf{X} + \mathbf{W}_G) \\ &\geq \mu h(\mathbf{X} + \mathbf{V}_G) - h(\mathbf{X} + \tilde{\mathbf{W}}_G) - h(\mathbf{W}_G) + h(\tilde{\mathbf{W}}_G) \\ &= \mu h(\mathbf{X} + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X} + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G). \end{aligned} \tag{120}$$

Since the expression will be minimized over $f_X(\mathbf{x})$, the last two terms in (120) are ignored, and by substituting $\mathbf{Y}$ and $\hat{\mathbf{X}}$ for $\mathbf{X} + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G$ and $\mathbf{X} + \tilde{\mathbf{W}}_G$, respectively, the inequality in (103) is equivalently expressed as the

following variational problem:

$$\min_{f_{\hat{X}}, f_Y} \quad \mu h(\mathbf{Y}) - h(\hat{\mathbf{X}}) - \mu(\mu - 1) h(\hat{\mathbf{V}}_G)$$

$$\text{s. t.} \quad \iint f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y} - \mathbf{x}) d\mathbf{x} d\mathbf{y} - 1 = 0,$$

$$\iint f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y} - \mathbf{x}) \mathbf{x} \mathbf{x}^T d\mathbf{x} d\mathbf{y} - \boldsymbol{\Sigma}_{\hat{X}} \preceq \mathbf{0},$$

$$\iint f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y} - \mathbf{x}) \mathbf{y} \mathbf{y}^T d\mathbf{x} d\mathbf{y} - \boldsymbol{\Sigma}_{Y^*} = \mathbf{0},$$

$$\iint f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y} - \mathbf{x}) (\mathbf{y} \mathbf{y}^T - \mathbf{x} \mathbf{x}^T - (\mathbf{y} - \mathbf{x})(\mathbf{y} - \mathbf{x})^T) d\mathbf{x} d\mathbf{y} = \mathbf{0},$$

$$-\iint f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y} - \mathbf{x}) \log f_{\hat{X}}(\mathbf{x}) d\mathbf{x} d\mathbf{y} \geq p_{\hat{X}} \tag{121}$$

$$f_Y(\mathbf{y}) = \int f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y} - \mathbf{x}) d\mathbf{x},$$

where $\hat{\mathbf{X}} = \mathbf{X} + \tilde{\mathbf{W}}_G$, $\mathbf{Y} = \hat{\mathbf{X}} + \hat{\mathbf{V}}_G$, $\mathbf{W}_G = \tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G$, $\mathbf{V}_G = \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G$, $\boldsymbol{\Sigma}_{\hat{X}} = \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_{\tilde{W}}$, $\boldsymbol{\Sigma}_{Y^*} = \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_V$, and $\boldsymbol{\Sigma}_{X^*}$ is the covariance matrix of the optimal solution $\mathbf{X}^*$.

The variational problem in (121) is exactly the same as the one in (80). Therefore, using the same method as in the proof of Theorem 10, we obtain the following inequality (see the details in the proof of Theorem 10):

$$\mu h(\mathbf{X} + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X} + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G)$$

$$\geq \mu h(\mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X}_G^* + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G). \tag{122}$$

By appropriately choosing $\mathbf{X}_G^*$ and $\tilde{\mathbf{W}}_G$, the right-hand side of the equation in (122) is expressed as

$$\mu h(\mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X}_G^* + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G)$$

$$= \mu h(\mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X}_G^* + \mathbf{W}_G). \tag{123}$$

The equality in (123) is due to the equality condition of the data processing inequality in [11]. For the completeness of the proof, we introduce a technique, which is slightly different from the one in [11].

To satisfy the equality in the equation (123), the equality condition in the following lemma must be satisfied.

Lemma 1 (Data Processing Inequality [1])**.** *When three random vectors* $\mathbf{Y}_1$, $\mathbf{Y}_2$, *and* $\mathbf{Y}_3$ *represent a Markov chain* $\mathbf{Y}_1 \rightarrow \mathbf{Y}_2 \rightarrow \mathbf{Y}_3$, *the following inequality is satisfied:*

$$I(\mathbf{Y}_1; \mathbf{Y}_3) \leq I(\mathbf{Y}_1; \mathbf{Y}_2). \tag{124}$$

The equality holds if and only if $I(\mathbf{Y}_1; \mathbf{Y}_2 | \mathbf{Y}_3) = 0$.

In Lemma 1, $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ are defined as $\mathbf{X}_G^*$, $\mathbf{X}_G^* + \tilde{\mathbf{W}}_G$, and $\mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G$, respectively. Therefore, the equality condition, $I(\mathbf{Y}_1; \mathbf{Y}_2 | \mathbf{Y}_3) = 0$ is expressed as

$$I(\mathbf{Y}_1; \mathbf{Y}_2 | \mathbf{Y}_3)$$

$$= h(\mathbf{Y}_1 | \mathbf{Y}_3) - h(\mathbf{Y}_1 | \mathbf{Y}_2, \mathbf{Y}_3)$$

$$= \frac{1}{2} \log (2\pi e)^n |\boldsymbol{\Sigma}_{Y_1 | Y_3}| - \frac{1}{2} \log (2\pi e)^n |\boldsymbol{\Sigma}_{Y_1 | Y_2}|$$

$$= \frac{1}{2} \log (2\pi e)^n |\boldsymbol{\Sigma}_{Y_1} - \boldsymbol{\Sigma}_{Y_1} \boldsymbol{\Sigma}_{Y_3}^{-1} \boldsymbol{\Sigma}_{Y_1}| - \frac{1}{2} \log (2\pi e)^n |\boldsymbol{\Sigma}_{Y_1} - \boldsymbol{\Sigma}_{Y_1} \boldsymbol{\Sigma}_{Y_2}^{-1} \boldsymbol{\Sigma}_{Y_1}|$$

$$= \frac{1}{2} \log (2\pi e)^n |\boldsymbol{\Sigma}_{X^*} - \boldsymbol{\Sigma}_{X^*} (\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}} + \boldsymbol{\Sigma}_{\hat{W}})^{-1} \boldsymbol{\Sigma}_{X^*}| - \frac{1}{2} \log (2\pi e)^n |\boldsymbol{\Sigma}_{X^*} - \boldsymbol{\Sigma}_{X^*} (\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}})^{-1} \boldsymbol{\Sigma}_{X^*}|$$

$$= \frac{1}{2} \log (2\pi e)^n |\boldsymbol{\Sigma}_{X^*}| |I - (\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}} + \boldsymbol{\Sigma}_{\hat{W}})^{-1} \boldsymbol{\Sigma}_{X^*}| - \frac{1}{2} \log (2\pi e)^n |\boldsymbol{\Sigma}_{X^*}| |I - (\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}})^{-1} \boldsymbol{\Sigma}_{X^*}|$$

$$= \frac{1}{2} \log (2\pi e)^n |I - (\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}} + \boldsymbol{\Sigma}_{\hat{W}})^{-1} \boldsymbol{\Sigma}_{X^*}| - \frac{1}{2} \log (2\pi e)^n |I - (\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}})^{-1} \boldsymbol{\Sigma}_{X^*}|$$

$$= \frac{1}{2} \log (2\pi e)^n |I - (\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_W)^{-1} \boldsymbol{\Sigma}_{X^*}| - \frac{1}{2} \log (2\pi e)^n |I - (\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}})^{-1} \boldsymbol{\Sigma}_{X^*}|$$

$$= 0. \tag{125}$$

If $(\boldsymbol{\Sigma}_{x^*} + \boldsymbol{\Sigma}_W)^{-1} \boldsymbol{\Sigma}_{x^*} = (\boldsymbol{\Sigma}_{x^*} + \boldsymbol{\Sigma}_{\tilde{W}})^{-1} \boldsymbol{\Sigma}_{x^*}$, the equality in (125) is satisfied, the equality condition in Lemma 1 holds, and therefore, the equality in (123) is proved. The validity of $(\boldsymbol{\Sigma}_{x^*} + \boldsymbol{\Sigma}_W)^{-1} \boldsymbol{\Sigma}_{x^*} = (\boldsymbol{\Sigma}_{x^*} + \boldsymbol{\Sigma}_{\tilde{W}})^{-1} \boldsymbol{\Sigma}_{x^*}$ is proved by Lemma 8 in [11].

Therefore, $I(\mathbf{Y}_1; \mathbf{Y}_2 | \mathbf{Y}_3) = 0$, and from the equations in (120), (122), and (123), we obtain the following extremal entropy inequality:

$$
\begin{aligned}
\mu h(\mathbf{X} + \mathbf{V}_G) - h(\mathbf{X} + \mathbf{W}_G) &\geq \mu h(\mathbf{X} + \mathbf{V}_G) - h(\mathbf{X} + \tilde{\mathbf{W}}_G) - h(\mathbf{W}_G) + h(\tilde{\mathbf{W}}_G) \\
&= \mu h(\mathbf{X} + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X} + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G) \\
&\geq \mu h(\mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X}_G^* + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G) \\
&= \mu h(\mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X}_G^* + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G) \\
&= \mu h(\mathbf{X}_G^* + \mathbf{V}_G) - h(\mathbf{X}_G^* + \mathbf{W}_G),
\end{aligned}
$$

and the proof is completed.

∎

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory (2nd edition)*, New York: Wiley, 2006.

[2] T. Liu and P. Viswanath, "An Extremal Inequality Motivated by Multiterminal Information-Theoretic Problems," *IEEE Trans. Inform. Theory*, vol. 53, no. 5, pp. 1839 - 1851, May 2007.

[3] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations*. Applied Mathematical Sciences vol. 147. Springer Verlag. New York, 2006.

[4] G. Scutari, D. Palomar, F. Facchinei, and J.-S. Pang, "Convex Optimization, Game Theory, and Variational Inequality Theory," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 35-49, May 2010.

[5] M. Debbah and R. Muller, "MIMO Channel Modelling and the Principle of Maximum Entropy," *IEEE Trans. Inform. Theory*, vol. 51, no. 5, pp. 1667-1690, May 2005.

[6] D. F. Delong, Jr., and E. M. Hofstetter, "On the Design of Optimum Radar Waveforms for Clutter Rejection," *IEEE Trans. Inform. Theory,* vol. 13, no. 3, pp. 454-463, Jul. 1967.

[7] L. J. Spafford, "Optimum Radar Signal Processing in Clutter", *IEEE Trans. Inform. Theory,* vol. 14, no. 5, pp. 734-743, Sep. 1968.

[8] E. T. Jaynes, "On the Rationale of Maximum Entropy Methods," *Proc. of the IEEE*, vol. 70, no. 9, pp. 939-952, Sep. 1982.

[9] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 3072 - 3081, Nov. 2001.

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

[11] S. Park, E. Serpedin, and K. Qaraqe "An Alternative Proof of an Extremal Entropy Inequality," arXiv:1201.6681.

[12] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*, New York: Dover, 1991.

[13] J. Gregory, *Constrained Optimization in the Calculus of Variations and Optimal Control Theory*, New York: Van Nostrand Reinhold, 1992.

[14] H. Sagan, *Introduction to the Calculus of Variations*, New York: Dover, 1992.

[15] J. Bercher and C. Vignat, "On minimum Fisher information distributions with restricted support and fixed variance," *Inform. Sci.,* vol. 179, no. 22, pp. 3832-3842, Nov. 2009

[16] H. Weingarten, Y. Steinberg, and S. Shamai, "The Capacity Region of the Gaussian Mutiple-Input Multiple-Output Broadcast Channel," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 3936 - 3964, Sep. 2006.

[17] 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.

[18] S. Ihara, "On the capacity of channels with additive non-Gaussian noise," *Inform. Contr.,* vol. 37, no. 1, pp. 34-39, Apr. 1978.

[19] P. P. Bergmans, "A Simple Converse for Broadcast Channels with Additive White Gaussian Noise," *IEEE Trans. Inform. Theory*, vol. 20, no. 2, pp. 279 - 280, Mar. 1974.

[20] T. Liu and S. Shamai (Shitz), "A Note on the Secrecy Capacity of the Multiple-Antenna Wiretap Channel," *IEEE Trans. Inform. Theory*, vol. 55, no. 6, pp. 2547 - 2553, Jun. 2009.

[21] R. Liu, T. Liu, H. Poor, and S. Shamai, "A Vector Generalization of Costa's Entropy-Power Inequality with Applications," *IEEE Trans. on Inform. Theory,* vol. 56, no. 4, pp. 1865-1879, Apr. 2010.

[22] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Tech. J.*, vol. 27, pp. 623-656, Oct. 1948.

[23] S. Verdu and D. Guo, "A simple proof of the entropy power inequality," *IEEE Trans. Inform. Theory,* vol. 52, no. 5, pp. 2165-2166, May 2006.

[24] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 1057 - 1070, May 1998.

[25] P. Stoica and P. Babu, "The Gaussian Data Assumption Leads to the Largest Cramér-Rao Bound," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 132-133, May 2011.

[26] P. Stoica and O. Besson, "Training Sequence Design for Frequency Offset and Frequency-Selective Channel Estimation," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1910-1917, Nov. 2003.

[27] S. Park, E. Serpedin, and K. Qaraqe, "Gaussian Assumption: The Least Favorable but the Most Useful," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 183-186, May 2013.

[28] E. Serpedin and Q. Chaudhari, *Synchronization of Wireless Sensor Networks: Parameter Estimation, Performance Benchmarks and Protocols,* Cambridge University Press, August 2009.

[29] K. Marton, "A Coding Theorem for the Discrete Memoryless Broadcast Channel," *IEEE Trans. Inform. Theory*, vol. 25, no. 3, pp. 306 ?311, May 1979.

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

[31] M. Payaro, M. Gregori, and D. Palomar, "Yet Another Power Entropy Inequality with an Application," *2011 International Conference on Wireless Communications and Signal Processing (WCSP)*, Nanjing, China, Nov. 2011, pp. 1-5.

[32] M. Payaro and D. Palomar, "A Multivariate Generalization of Costa's Entropy Power Inequality," *IEEE International Symposium in Information Theory 2008 (ISIT 2008)*, Toronto, Canada, Jul. 2008, pp. 1088 - 1092.

[33] S. K. Leung-Yan-Cheong and M. E. Hellman, "The Gaussian wire-tap channel," *IEEE Trans. Inform. Theory*, vol. 24, no. 4, pp. 451 - 456, Jul. 1978.

Sangwoo Park received the B.S. degree in electrical engineering from Chung-Ang University (CAU), Seoul, Korea, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from Texas A&M University, College Station, in 2008 and 2012, respectively. From 2004 to 2005, he worked as a full-time assistant engineer for UMTS/WCDMA projects in Samsung Electronics. Currently, he is a research engineer at KT (Korea Telecom) in Korea. His research interests lie in wireless communications, information theory, and statistical signal processing.

Erchin Serpedin (F'13) received the specialization degree in signal processing and transmission of information from Ecole Superieure ĎElectricite (SUPELEC), Paris, France, in 1992, the M.Sc. degree from the Georgia Institute of Technology, Atlanta, in 1992, and the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, in January 1999. He is currently a professor in the Department of Electrical and Computer Engineering at Texas A&M University, College Station. He is the author of two research monographs, one edited textbook, 100 journal papers and 150 conference papers, and has served as associate editor for about 10 journals such as IEEE Transactions on Information Theory, IEEE Transactions on Communications, Signal Processing (Elsevier), IEEE Transactions on Signal Processing, IEEE Transactions on Wireless Communications, IEEE Communications Letters, IEEE Signal Processing Letters, Phycom, EURASIP Journal on Advances in Signal Processing, and EURASIP Journal on Bioinformatics and Systems Biology. His research interests include signal processing, wireless communications, computational statistics, bioinformatics and systems biology.

Khalid Qaraqe (M'97-S'00) received with honors the B.S. degree in EE from the University of Technology, Baghdad, Irak, in 1986. He received the M.S. degree in EE from the University of Jordan, Jordan, in 1989, and he earned his Ph.D. degree in EE from Texas A&M University, College Station, TX, in 1997. From 1989 to 2004, Dr. Qaraqe held a variety of positions in many companies. He has over 15 years of experience in the telecommunications industry. Dr. Qaraqe has worked for Qualcomm, Enad Design Systems, Cadence Design Systems/Tality Corporation, STC, SBC and Ericsson. He has worked on numerous GSM, CDMA, WCDMA projects and has experience in product development, design, deployment, testing and integration. Dr. Qaraqe joined Texas A&M University at Qatar, in July 2004, where he is now a professor. Dr. Qaraqe research interests include communication theory and its application to design and performance analysis of cellular systems and indoor communication systems. Particular interests are in the development of 3G UMTS, cognitive radio systems, broadband wireless communications and diversity techniques.