ASPECTS OF INTERFACE BETWEEN INFORMATION THEORY AND SIGNAL PROCESSING WITH APPLICATIONS TO WIRELESS COMMUNICATIONS

A Dissertation

by

SANG WOO PARK

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Erchin Serpedin
	Khalid Qaraqe
Committee Members,	Tie Liu
	Aydin Karsilayan
	Anxiao Jiang
Department Head,	Chanan Singh

December 2012

Major Subject: Electrical Engineering

ABSTRACT

This dissertation studies several aspects of the interface between information theory and signal processing. Several new and existing results in information theory are researched from the perspective of signal processing. Similarly, some fundamental results in signal processing and statistics are studied from the information theoretic viewpoint.

The first part of this dissertation focuses on illustrating the equivalence between Stein's identity and De Bruijn's identity, and providing two extensions of De Bruijn's identity. First, it is shown that Stein's identity is equivalent to De Bruijn's identity in additive noise channels with specific conditions. Second, for arbitrary but fixed input and noise distributions, and an additive noise channel model, the first derivative of the differential entropy is expressed as a function of the posterior mean, and the second derivative of the differential entropy is expressed in terms of a function of Fisher information. Several applications over a number of fields, such as statistical estimation theory, signal processing and information theory, are presented to support the usefulness of the results developed in Section 2.

The second part of this dissertation focuses on three contributions. First, a connection between the result, proposed by Stoica and Babu, and the recent information theoretic results, the worst additive noise lemma and the isoperimetric inequality for entropies, is illustrated. Second, information theoretic and estimation theoretic justifications for the fact that the Gaussian assumption leads to the largest Cramér-Rao lower bound (CRLB) is presented. Third, a slight extension of this result to the more general framework of correlated observations is shown.

The third part of this dissertation concentrates on deriving an alternative proof for an extremal entropy inequality (EEI), originally proposed by Liu and Viswanath. Compared with the proofs, presented by Liu and Viswanath, the proposed alternative proof is simpler, more direct, and more information-theoretic. An additional application for the extremal inequality is also provided. Moreover, this section illustrates not only the usefulness of the EEI but also a novel method to approach applications such as the capacity of the vector Gaussian broadcast channel, the lower bound of the achievable rate for distributed source coding with a single quadratic distortion constraint, and the secrecy capacity of the Gaussian wire-tap channel.

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

To my wife and family

ACKNOWLEDGEMENTS

I express my gratitude to my advisors, Dr. Erchin Serpedin and Dr. Khalid Qaraqe, who enthusiastically guided and supported me while I overcame obstacles during my graduate studies. They have known me well, and have encouraged me academically. I also greatly appreciate my committee members, Dr. Tie Liu, Dr. Aydin Karsilayan, and Dr. Anxio Jiang, for their help and support. In addition, I would like to thank Dr. Radu Stoleru, who temporarily participated in my dissertation defense as a substitute committee member, and Dr. Deepa Kundur, who was a former committee member.

I express my thanks to people in the Telecommunications, Control and Signal Processing (TCSP) group. I am especially thankful to Jaewon, who has discussed some problems in my research. I also thank all the students of our lab, Sabit, Aitzaz, Aminar, Ali, Jaehan, Huseyin, Qasim, and others. I appreciate all of my friends in Korea: my oldest friends, Seunghwan, Byungrok, Hyunsoo, Hunjoo, etc., my friends in college, Yoonho, Hwan, Jaejin, Yongjun, etc., and my colleague, Juhyun. They encouraged me to start this valuable journey and to finish it.

I also thank the friends I met in Qatar, Suyong, Hojoon, Minjae, Dongsuk, Jihoon, Jongil, and Hyuchul. I am also grateful for my friends at Korean Methodist Church and AKOZ, the Korean basketball club. They have made some unforgettable memories for me. I am also very grateful to Sharon Roe, who has improved my English as a tutor.

I owe much to my family: my grandmother, mother, father, sister, and mother-in-law. I am deeply grateful to them. They always love, take care of, and support me. Without their help, I would never have overcome this challenge.

Lastly, I wish to show my final gratitude to my lovely wife, Yoonjeong, who has always been with and supported me. Her endless love, assistance, and interest enabled me to complete my long journey.

TABLE OF CONTENTS

LIST OF FIGURES

Page

1. INTRODUCTION

A prominent recent trend in the information technology (IT) industry is the convergence of technologies from different fields. A smart phone, for example, functioning as voice and data call, video camera, wireless internet access device, and game console, cannot be solely regarded as a calling device, since it is rather an integrated entity where various technologies are confluent through innovation. From this perspective, the convergence of the technologies and knowledge from various fields has drawn the attention of researchers from academia and industry. As an additional illustration, recent genomic studies have saliently displayed the convergence trend of different technologies and expertise, in which knowledge from computational biology, computer science, machine learning, electrical engineering, statistics, and medical sciences are nicely intertwined together to yield outstanding results. Therefore, it appears that without exploiting the tools and knowledge from a wide range of fields, the secret of deciphering the interactions between genes cannot be revealed.

Similar to the smart phones and genomic studies, the convergence and integration of results and knowledge from fields as diverse as wireless communications, information theory, estimation theory, and signal processing have been advocated and studied. For instance, De Bruijn's identity [46], a mathematical equation that expresses the relationship between differential entropy and Fisher information, two fundamental concepts in information theory and signal processing, has been exploited for proving the entropy power inequality (EPI) and establishing channel capacity under several different scenarios [43], [53], [32], [31], [52], [13], [12]. I-MMSE identity [18] is another example of application of De Bruijn's identity. I-MMSE identity is equivalent to De Bruijn's identity. In addition, I-MMSE illustrates an interesting

connection between the input-output mutual information and minimum mean square error. This identity has been also widely used by many researchers [18], [42], [42], [20], [39], [19]. An important common feature of these two identities is that they establish a relationship between entropy and Fisher information, a relationship which helped to solve several important problems, e.g., EPI was established by Rioul in 2011 using De Bruijn's identity and I-MMSE identity.

This dissertation focuses on the connections among fundamental concepts, methods, and inequalities proposed in the fields of information theory, signal processing, optimization theory, and statistics. The focus is not only on establishing theoretic results and proofs but also on finding practical applications of the proposed theoretical results. The summary of the main contributions of this research is as follows.

In Section 2, Stein and De Bruijn identities are studied. Stein's identity (or lemma) was first established in 1956 [47], and it has attracted a lot of interest due to its applications in the James-Stein estimation technique, empirical Bayes methods, and numerous other fields, see e.g., [6], [26], [22], [35], [34], [15]. De Bruijn's identity has recently attracted increased interest due to its applications in statistical estimation theory and turbo (iterative) decoding schemes. De Bruijn's identity shows a link between two fundamental concepts in information theory: entropy and Fisher information [1], [24], [9], [18], [40], [42].

The first major result of Section 2 is the fact that De Bruijn's identity and Stein's identity are equivalent, in the sense that each identity implies the other one. The important fact of this result is that the whole set of applications established via Stein's identity could be transferred and proved into the realm of De Bruijn's identity and vice versa. The second major result of this section are two extensions of De Bruijn identity to non-Gaussian random variables. The third major result deals with establishing two fundamental lower bounds in statistical signal processing,

the Bayesian Cramér-Rao lower bound (BCRLB) and the Cramér-Rao lower bound (CRLB), and a novel lower bound, which is tighter than BCRLB. Finally, several additional applications of the developed results are presented.

Section 3 studies the usage of Gaussian assumption in linear regression problems when the actual distribution of additive noise does not obey the Gaussian distribution. Gaussian distribution is one of the most well-known and widely used distributions in engineering, statistics, and physics. There are several reasons for this widespread usage of Gaussian distribution, such as the Central Limit Theorem (CLT), analytical tractability, easy generation of normal random variables, etc., and this explains why the normal distribution is usually assumed. However, very little information is available in the literature concerning the properties of the resulting estimator which assumes a Gaussian distribution of the observations instead of the actual (true) distribution of the observations. Without information about the actual distribution of observations, Gaussian assumption appears as the most conservative choice due to the fact that the Gaussian distribution minimizes the Fisher information, i.e., the inverse of the Cramér-Rao lower bound (CRLB). Therefore, any optimization of the training data based on the CRLB under the Gaussian assumption can be considered to be min-max optimal in the sense of minimizing the largest CRLB, see e.g., [48], [10], [49], [4].

The main theme of Section 3 is to investigate a relationship between the result reported in [48] and the recent information theoretic results presented in [8], [43], to study from an information and estimation theoretic perspective why the Gaussian assumption leads to the largest CRLB, and to slightly extend this result to the more general framework of correlated observations.

In Section 4, the extremal (entropy) inequality (EEI) is studied. The extremal entropy inequality, a generalized version of EPI, was proposed by Liu and Viswanath

[32], and it was further researched by several authors [30], [41]. The extremal entropy inequality was motivated by the question: "What is the optimal solution for the classical entropy power inequality (EPI) under a covariance matrix constraint?" Even though the expected solution is a Gaussian random vector, it is difficult to come up with the solution based on the classical EPI due to the covariance matrix constraint. Therefore, a novel method, called the channel enhancement technique [53] was adopted in the proofs provided in [32].

The main goal of Section 4 is to prove the EEI without using the channel enhancement technique. Our proof is mainly based on four techniques: data processing inequality, moment generating function (MGF), worst additive noise lemma, and classical EPI. The proposed novel proof brings the following significant contributions. First, our proof is simpler and more direct, compared with the proofs in [32]. Second, a more information-theoretic approach is developed. In our proof, the data processing inequality and MGF enable to not only circumvent the step of using the KKT conditions but also to omit the step of proving the existence of the optimal solution which satisfies the KKT conditions, a step which is very complicated to accomplish. Finally, the proposed novel method in our proof can be adapted for applications such as establishing the Gaussian broadcast channel capacity, secrecy capacity of Gaussian wiretap channel, etc., as well as for establishing EEI. These considerations support the versatility of EEI.

Section 5 provides a unifying variational calculus framework for establishing a large class of fundamental information theoretic inequalities. These inequalities provide a useful theoretical basis for the field of information theory as well as other fields. The proposed innovative variational approach not only offers alternative proofs for information theoretic inequalities, but also enables the existing results to be extended in other directions. Furthermore, it is important to remark that the pro-

posed functional approach represents a potential powerful tool for finding guidelines to determine the optimal solution for many other open problems.

The main contributions of Section 5 are enumerated next. First, using calculus of variations, the maximizing differential entropy and minimizing Fisher information theorems are proved under different sets of assumptions, the classical assumptions found in the literature as well as a different set of assumptions. Second, an alternative proof of the worst additive noise lemma [11], [23] is proposed based on the proposed functional analysis framework. Third, a novel proof of EPI is provided in the proposed functional framework. Finally, EEI is studied and justified again from the perspective of a functional problem.

Finally, Section 6 summarizes the results and the main contributions of this dissertation. Concluding remarks and future research directions are also proposed. Future research directions include solving currently open (unsolved) problems and developing new extensions for the results presented in this dissertation.

1.1 Notations

Throughout this dissertation, unless otherwise mentioned, the following notation rules are adopted: a lower case plain-text alphabet (e.g., x or λ) denotes a scalar deterministic variable or a constant, a lower case bold alphabet (e.g., $\mathbf{x}$ or $\boldsymbol{\lambda}$) represents a deterministic vector, an upper case plain-text English alphabet (e.g., X) is a random variable, an upper case bold English alphabet (e.g., $\mathbf{X}$) stands for a random vector or a matrix, and an upper case bold Greek alphabet (e.g., $\boldsymbol{\Sigma}$) denotes a matrix. The dimensions (sizes) of a vector and a matrix are denoted as n and n-by-n, respectively. All information theoretic quantities are represented by conventional notations. For example, $h(\mathbf{X})$ and $I(\mathbf{X};\mathbf{Y})$ stand for differential entropy of a random vector $\mathbf{X}$ and mutual information between a random vector $\mathbf{X}$ and a random

vector $\mathbf{Y}$, respectively. Conditional entropy and conditional mutual information are denoted as $h(\mathbf{X}|\mathbf{Y})$ and $I(\mathbf{X};\mathbf{Y}|\mathbf{Z})$, respectively. The notation $\preceq$ or $\succeq$ stands for positive (semi)definite partial ordering between matrices, i.e., $\mathbf{\Sigma}_1 \preceq \mathbf{\Sigma}_2$ means $\mathbf{\Sigma}_2 - \mathbf{\Sigma}_1$ is a positive semi-definite matrix [21], [33]. A positive definite matrix means a strictly positive definite matrix, and $\nabla_{\mathbf{\Sigma}}$ stands for the Jacobian matrix with respect to $\mathbf{\Sigma}$. The matrix $\mathbf{I}$ denotes an n-by-n identity matrix, and the matrix $\mathbf{0}$ stands for an n-by-n zero matrix. Notation $\mathbb{E}[\cdot]$ denotes an expectation with respect to all random vectors inside $[\cdot]$, and $M_X(\mathbf{S})$ and $M_{X|Y}(\mathbf{S})$ are moment generating functions of a random vector $\mathbf{X}$, and a random vector $\mathbf{X}$ given (conditioned on) random variable $\mathbf{Y}$, respectively. For simplicity, log denotes the natural logarithm.

2. STEIN'S IDENTITY AND DE BRUIJN'S IDENTITY*

2.1 Introduction

Stein's identity (or lemma) was first established in 1956 [47], and since then it has been widely used by many researchers (e.g., [6], [26], [22]). Due to its applications in the James-Stein estimation technique, empirical Bayes methods, and numerous other fields, Stein's identity has attracted a lot of interest (see e.g., [35], [34], [15]).

Recently, another identity, De Bruijn's identity, has attracted increased interest due to its applications in estimation and turbo (iterative) decoding schemes. De Bruijn's identity shows a link between two fundamental concepts in information theory: entropy and Fisher information [1], [24], [9]. Verdú and his collaborators conducted a series of studies [18], [40], [42] to analyze the relationship between the input-output mutual information and the minimum mean-square error (MMSE), a result referred to as the I-MMSE identity for additive Gaussian noise channels, studies which were later extended to non-Gaussian channels in [20], [39]. Also, the equivalence between De Bruijn's identity and I-MMSE identity was shown in [18].

The main theme of this section is to study how Stein's identity (Theorem 2.2) is related to De Bruijn's identity (Theorem 2.1). To compare Stein's identity with De Bruijn's identity, additive noise channels of the following form are considered in this section:

$$Y = X + \sqrt{a}W, \tag{2.1}$$

*Reprinted with permission from "On the equivalence between Stein and de Bruijn identities," Sangwoo Park, Erchin Serpedin, and Khalid Qaraqe, 2012, IEEE Transactions on Information Theory, vol. 58, no. 12, Copyright 2012 by IEEE.

where input signal X and additive noise W are arbitrary random variables, X and W are independent of each other, and parameter a is assumed nonnegative. First, when additive noise W is Gaussian with zero mean and unit variance, the equivalence between the generalized Stein's identity (Theorem 2.2) and De Bruijn's identity (Theorem 2.1) is proved. Since the standard-form Stein's identity in (2.13) requires both random variables X and W to be Gaussian, instead of the standard-form Stein's identity, the generalized version of Stein's identity in (2.12) is used. If we further assume that input signal X is also Gaussian, then both random variables X and W are Gaussian, and the output signal Y is Gaussian. In this case, not only Stein's and De Bruijn's identities are equivalent, but also they are equivalent to the heat equation identity, proposed in [6].

The second major question that we will address in this section is how De Bruijn's identity could be extended. De Bruijn's identity shows the relationship between the differential entropy and the Fisher information of the output signal Y under additive Gaussian noise channels. Therefore, under additive non-Gaussian noise channels, we cannot use De Bruijn's identity. However, we will derive a similar form of De Bruijn's identity for additive non-Gaussian noise channels. Considering additive arbitrary noise channels, the first derivative of the differential entropy of output signal Y will be expressed by the posterior mean, while the second derivative of the differential entropy of output signal Y will be represented by a function of Fisher information. Even though some of these relationships do not include the Fisher information, they still show relationships among basic concepts in information theory and estimation theory, and these relationships hold for arbitrary noise channels.

Based on the results mentioned above, we introduce several applications dealing with both estimation theoretic and information theoretic aspects. In the estimation theory field, the Fisher information inequality, the Bayesian Cramér-Rao lower bound

(BCRLB), and a new lower bound for the mean square error (MSE) in Bayesian estimation are derived. The surprising result is that the newly derived lower bound for MSE is tighter than the BCRLB. The proposed new bound overcomes the main drawback of BCRLB, i.e., its looseness in the low Signal-to-Noise Ratio (SNR) regime, since it provides a tighter bound than BCRLB especially at low SNRs. Even though some of the proposed applications have already been proved before, in this section we show not only alternative ways to prove them, but also new relationships among them. In the information theory realm, Costa's entropy power inequality - previously proved in [7] - is derived in two different ways based on our results. Both proposed methods show novel, simple, and alternative ways to prove Costa's entropy power inequality. Finally, applications in other areas are briefly mentioned.

The rest of this section is organized as follows. Various relationships between Stein's identity and De Bruijn's identity are established in Section 2.3. Some extensions of De Bruijn's identity are provided in Section 2.4. In Section 2.5, several applications based on the proposed novel results are supplied. Finally, conclusions are mentioned in Section 2.6. All the detailed mathematical derivations for the proposed results are given in appendices.

2.2 Preliminary Results

In this section, several definitions and preliminary theorems are provided. First, the concept of Fisher information is defined as follows.

Fisher information of a deterministic parameter θ is defined as

$$\begin{aligned} J_\theta(Y) &= \int_{-\infty}^{\infty} f_Y(y;\theta) \left(\frac{d}{d\theta} \log f_Y(y;\theta) \right)^2 dy \\ &= \mathbb{E}_Y \left[S_{Y_\theta}(Y)^2 \right], \end{aligned} \tag{2.2}$$

where $S_{Y_\theta}(Y)$ denotes a score function and is defined as $(d/d\theta)\log f_Y(y;\theta)$. Under a regularity condition,

$$\begin{aligned}\mathbb{E}_Y\left[S_{Y_\theta}(Y)\right] &= \int_{-\infty}^{\infty}\frac{d}{d\theta}f_Y(y;\theta)dy \\ &= 0,\end{aligned}$$

the Fisher information in (2.2) is equivalently expressed as

$$\begin{aligned}J_\theta(Y) &= -\int_{-\infty}^{\infty}f_Y(y;\theta)\frac{d^2}{d\theta^2}\log f_Y(y;\theta)dy \\ &= -\mathbb{E}_Y\left[\frac{d}{d\theta}S_{Y_\theta}(Y)\right]. \qquad (2.3)\end{aligned}$$

This is a general definition of Fisher information in signal processing, and Fisher information provides a lower bound, called the Cramér-Rao lower bound, for the mean square error of any unbiased estimator. Like other concepts, such as entropy and mutual information, in information theory, Fisher information also shows information about uncertainty. However, it is difficult to directly adopt the definition of Fisher information in information theory despite the fact that it has been commonly used in statistics. Instead, a more specific definition of Fisher information is proposed as follows.

If θ is assumed to be a location parameter, then

$$\frac{d}{d\theta}f_Y(y;\theta) = -\frac{d}{dy}f_Y(y-\theta;\theta). \qquad (2.4)$$

Therefore, the definition of Fisher information in (2.2) is changed as follows:

$$\begin{aligned} J_\theta(Y) &= \int_{-\infty}^{\infty} f_Y(y;\theta)\left(\frac{d}{d\theta}\log f_Y(y;\theta)\right)^2 dy \\ &= \int_{-\infty}^{\infty} f_Y(y-\theta;\theta)\left(-\frac{d}{dy}\log f_Y(y-\theta;\theta)\right)^2 dy \\ &= \int_{-\infty}^{\infty} f_{\tilde{Y}}(\tilde{y};\theta)\left(-\frac{d}{d\tilde{y}}\log f_{\tilde{Y}}(\tilde{y};\theta)\right)^2 d\tilde{y} \\ &= \mathbb{E}_{\tilde{Y}}\left[S(\tilde{Y})^2\right], \end{aligned} \tag{2.5}$$

where $S(\tilde{Y})$ denotes a score function, and it is defined as $(d/d\tilde{y})\log f_{\tilde{Y}}(\tilde{y};\theta)$. In equation (2.5), since we only consider a location parameter, we refer to Fisher information in (2.5) as Fisher information with respect to a location (or translation) parameter, and it is denoted as $J(\tilde{Y})$ (even though the definition of Fisher information with respect to a location parameter in (2.5) is derived from the definition of Fisher information in (2.2), the definition in (2.5) is more commonly used in information theory, and we do not distinguish random variable $\tilde{Y} = Y - \theta$ from random variable Y).

Given the channel model in (2.1), by substituting the parameter a for the unknown parameter θ, the expressions of Fisher information in (2.2) and (2.5) are respectively given by

$$\begin{aligned} J(Y) &= \int_{-\infty}^{\infty} f_Y(y;a)\left(\frac{d}{dy}\log f_Y(y;a)\right)^2 dy \\ &= \mathbb{E}_Y\left[S_Y(Y)^2\right], \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} J_a(Y) &= \int_{-\infty}^{\infty} f_Y(y;a)\left(\frac{d}{da}\log f_Y(y;a)\right)^2 dy \\ &= \mathbb{E}_Y\left[S_{Y_a}(Y)^2\right]. \end{aligned} \tag{2.7}$$

Second, two fundamental concepts, differential entropy and entropy power, are defined as follows. Differential entropy of random variable Y, $h(Y)$, is defined as

$$h(Y) = -\int_{-\infty}^{\infty} f_Y(y;a)\log f_Y(y;a)dy, \tag{2.8}$$

where $f_Y(y;a)$ denotes the probability density function (pdf) of random variable Y, log denotes the natural logarithm, and a is a deterministic parameter in the pdf. Similarly, the conditional entropy of random variable Y given random variable X, $h(Y|X)$ is defined as

$$h(Y|X) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y;a)\log f_{Y|X}(y|x;a)dxdy, \tag{2.9}$$

where $f_{X,Y}(x,y;a)$ denotes the joint pdf of random variables X and Y, $f_{Y|X}(y|x;a)$ is the conditional pdf of random variable Y given random variable X.

Entropy power of random variable Y, $N(Y)$, and (conditional) entropy power of random variable Y given random variable X, $N(Y|X)$ are respectively defined as

$$\begin{aligned} N(Y) &= \frac{1}{2\pi e}\exp(2h(Y)), \\ N(Y|X) &= \frac{1}{2\pi e}\exp(2h(Y|X)). \end{aligned} \tag{2.10}$$

Based on the definitions mentioned above, three preliminary theorems- De Bruijn's, Stein's, and heat equation identities- are introduced next.

Theorem 2.1 (De Bruijn's Identity [9], [46])**.** *Given the additive noise channel* $Y = X + \sqrt{a}W$*, let* X *be an arbitrary random variable with a finite second-order moment,*

and W be independent normally distributed with zero mean and unit variance. Then,

$$\frac{d}{da}h(Y) \;=\; \frac{1}{2}J(Y). \tag{2.11}$$

Proof. See [9]. □

Theorem 2.2 (Generalized Stein's Identity [26])**.** *Let Y be an absolutely continuous random variable. If the probability density function $f_Y(y)$ satisfies the following equations,*

$$\lim_{y\to\pm\infty} k(y)f_Y(y) \;=\; 0,$$

and

$$\frac{\frac{d}{dy}f_Y(y)}{f_Y(y)} \;=\; -\frac{\frac{d}{dy}k(y)}{k(y)} + \frac{(\nu - t(y))}{k(y)}$$

for some function $k(y)$, then

$$\mathbb{E}_Y\left[r(Y)\left(t(Y)-\nu\right)\right] \;=\; \mathbb{E}_Y\left[\tfrac{d}{dY}r(Y)k(Y)\right], \tag{2.12}$$

for any function $r(Y)$ which satisfies $\mathbb{E}_Y\left[|r(Y)t(Y)|\right] < \infty$, $\mathbb{E}_Y\left[r(Y)^2\right] < \infty$, and $\mathbb{E}_Y\left[\left|k(Y)\frac{d}{dY}r(Y)\right|\right] < \infty$. $\mathbb{E}_Y[\cdot]$ denotes the expectation with respect to the pdf of random variable Y. In particular, when random variable Y is normally distributed with mean μ_y and variance σ_y^2, equation (2.12) simplifies to

$$\mathbb{E}_Y\left[r(Y)\left(Y-\mu_y\right)\right] \;=\; \sigma_y^2\mathbb{E}_Y\left[\tfrac{d}{dY}r(Y)\right]. \tag{2.13}$$

Equation (2.13) is the well-known classic Stein's identity.

Proof. See [26]. □

Theorem 2.3 (Heat Equation Identity [6])**.** *Let Y be normally distributed with mean μ and variance $1+a$. Assume $g(y)$ is a twice continuously differentiable function, and both $g(y)$ and $|\frac{d}{dy}g(y)|$ are* $O(e^{c|y|})$ for some $0 \leq c < \infty$. Then,*

$$\frac{d}{da}\mathbb{E}_Y\left[g(Y)\right] = \frac{1}{2}\mathbb{E}_Y\left[\frac{d^2}{dY^2}g(Y)\right]. \tag{2.14}$$

Proof. See [6]. □

2.3 Relationships between Stein's Identity and De Bruijn's Identity

In Section 2.2, Theorems 2.1, 2.2, and 2.3 share an analogy: an identity between expectations of functions, which include derivatives. Especially, the heat equation identity admits the same form as De Bruijn's identity by choosing function $g(y)$ as $-\log f_Y(y;a)$. If De Bruijn's identity is equivalent to the heat equation identity, it is also equivalent to Stein's identity, since the equivalence between the heat equation identity and Stein's identity was proved in [6]. However, there are two critical issues that stand in the way of the equivalence between Stein's and De Bruijn's identities: first, the function $g(y)$ in Theorem 2.3 must be independent of the parameter a, which is not true when $g(y) = -\log f_Y(y;a)$. Second, in the heat equation identity, random variable Y must be Gaussian, which may not be true in De Bruijn's identity.

Due to the difficulties mentioned above, we will directly compare De Bruijn's identity (Theorem 2.1) with the generalized Stein's identity (Theorem 2.2).

Theorem 2.4. *Given the channel model (2.1), let X be an arbitrary random variable with a finite second-order moment, and let W be normally distributed with zero mean*

*$O(\cdot)$ denotes the limiting behavior of the function, i.e., $g(y) = O(q(y))$ if and only if there exist positive real numbers K and y^* such that $g(y) \leq K|q(y)|$ for any y which is greater than y^*.

and unit variance. Independence between random variables X and W is also assumed. Then, De Bruijn's identity (2.11) is equivalent to the generalized Stein's identity in (2.12) under specific conditions, i.e.,

$$\begin{aligned} \frac{d}{da}h(Y) &= \frac{1}{2}J(Y) \\ \iff \mathbb{E}_Y\left[r(Y;a)\left(t(Y;a)-\nu\right)\right] &= \mathbb{E}_Y\left[\tfrac{d}{dY}r(Y;a)k(Y;a)\right], \end{aligned}$$

with

$$r(y;a) = -\frac{d}{dy}\log f_Y(y;a), \quad k(y) = 1, t(y;a) = -\frac{\frac{d}{dy}f_Y(y;a)}{f_Y(y;a)}, \quad \textit{and} \quad \nu = 0, \tag{2.15}$$

where $\iff$ denotes the equivalence between before and after the notation.

Proof. See Appendix A.1. □

Now, when random variable Y is Gaussian, i.e., both random variables X and W are Gaussian, we can derive relationships among three identities, De Bruijn, Stein, and heat equation, as a special case of Theorem 2.4.

Theorem 2.5. *Given the channel model (2.1), let random variable X be normally distributed with mean μ and unit variance. Assume W is independent normally distributed with zero mean and unit variance. If we define the functions in (2.12) as follows:*

$$r(y;a) = -\frac{d}{dy}\log f_Y(y;a), \quad k(y;a) = \frac{1}{a}, t(y;a) = y, \quad \textit{and} \quad \nu = \mu,$$

then Stein's identity is equivalent to De Bruijn's identity. Moreover, if we define

$g(y;a)$ as

$$g(y;a) \;=\; -\log f_Y(y;a)$$

in (2.14), then De Bruijn's identity is also equivalent to the heat equation identity.

Proof. In Theorem 2.4, given the channel model (2.1) with an arbitrary but fixed random variable X and a Gaussian random variable W, the equivalence between De Bruijn's identity and the generalized Stein's identity was proved (cf. Appendix A.1). Here, by choosing random variable X as Gaussian, this is a special case of Theorem 2.4. Therefore, the equivalence between the two identities is trivial, and the details of the proof are omitted in this section. The only thing to prove is the second part of this theorem, namely, the equivalence between De Bruijn's identity and the heat equation identity. Since the equivalence between Stein's identity and the heat equation identity is proved in [6], this also proves the second part of the theorem, and the proof is completed. □

The functions $k(y;a)$, $r(y;a)$, $t(y;a)$, and $g(y;a)$ are the same as $k(y)$, $r(y)$, and $t(y)$ in Theorem 2.2 and $g(y)$ in Theorem 2.3, respectively. To show the dependence on parameter a, the functions $k(y;a)$, $r(y;a)$, $t(y;a)$, and $g(y;a)$ are used instead of $k(y)$, $r(y)$, $t(y)$, and $g(y)$, respectively.

2.4 Extension of De Bruijn's Identity

De Bruijn's identity is derived from the attribute of Gaussian density functions, which satisfy the heat equation. However, in general, probability density functions do not satisfy the heat equation. Therefore, to extend De Bruijn's identity to additive non-Gaussian noise channels, a general relationship between differentials of a

probability density function with respect to y and a of the form:

$$\frac{d}{da}f_{Y|X}(y|x;a) = -\frac{1}{2a}\frac{d}{dy}\left((y-x)f_{Y|X}(y|x;a)\right), \tag{2.16}$$

is required, a result that it is obtained in Appendix A.8 by exploiting the assumptions (2.17). The relationship (2.16) represents the key ingredient in establishing the link between the derivative of differential entropy and posterior mean, as described by the following theorem.

Theorem 2.6. *Consider the channel model (2.1), where X and W are arbitrary random variables independent of each other. Given the following assumptions:*

$$\frac{d}{dy}\mathbb{E}_X\left[f_{Y|X}(y|X;a)\right] = \mathbb{E}_X\left[\frac{d}{dy}f_{Y|X}(y|X;a)\right],$$

$$\frac{d}{da}\mathbb{E}_X\left[f_{Y|X}(y|X;a)\right] = \mathbb{E}_X\left[\frac{d}{da}f_{Y|X}(y|X;a)\right], \tag{2.17a}$$

$$\frac{d}{da}\int_{-\infty}^{\infty} f_Y(y;a)\log f_Y(y;a)dy = \int_{-\infty}^{\infty}\frac{d}{da}\Big(f_Y(y;a)\log f_Y(y;a)\Big)dy, \tag{2.17b}$$

$$\lim_{y\to\pm\infty}\mathbb{E}_X\left[Xf_{Y|X}(y|X;a)\right] = \mathbb{E}_X\left[\lim_{y\to\pm\infty}Xf_{Y|X}(y|X;a)\right],$$

$$\lim_{y\to\pm\infty}\mathbb{E}_X\left[f_{Y|X}(y|X;a)\right] = \mathbb{E}_X\left[\lim_{y\to\pm\infty}f_{Y|X}(y|X;a)\right],$$

$$\lim_{y\to\pm\infty} y^2 f_Y(y;a) = 0, \tag{2.17c}$$

$$\left|\frac{\mathbb{E}_X\left[Xf_{Y|X}(y|X;a)\right]}{\sqrt{f_Y(y;a)}}\right| < \infty, \tag{2.17d}$$

where $\mathbb{E}_{X|Y}[\cdot|\cdot]$ denotes the posterior mean, the first derivative of the differential entropy is expressed as

$$\frac{d}{da}h(Y) \;=\; \frac{1}{2a}\left\{1-\mathbb{E}_Y\left[\frac{d}{dY}\mathbb{E}_{X|Y}\left[X|Y\right]\right]\right\}. \tag{2.18}$$

Proof. See Appendix A.2.

Remark 2.1. *This is equivalent to the results in [20].*

□

It can be observed that the conditions (2.17) are required in the dominated convergence theorem and Fubini's theorem to ensure the interchangeability between a limit and an integral, and are not that restrictive. Also, the condition $\lim_{y\to\pm\infty} y^2 f_Y(y;a) = 0$ is not restrictive at all, and it is satisfied by all noise distributions of interest in practice.

Corollary 2.1 (De Bruijn's identity)**.** *Given the channel model in (2.1) with an arbitrary but fixed random variable X with a finite second moment and a Gaussian random variable W with zero mean and unit variance,*

$$\frac{d}{da}h(Y) \;=\; \frac{1}{2}J(Y).$$

Remark 2.2. *This is the well-known De Bruijn's identity [46]. Therefore, De Bruijn's identity is a special case of Theorem 2.6 when random variable W is normally distributed. When random variable W is Gaussian, assumptions in (2.17) are simplified to the existence of a finite second-order moment.*

Corollary 2.2. *Given the channel model in (2.1) with an arbitrary but fixed non-negative random variable X whose moment generating function exists and its pdf is bounded, and an exponential random variable W with unit value of the parameter (i.e., $f_W(w) = exp(-w)\mathrm{U}(w)$, where $\mathrm{U}(\cdot)$ denotes the unit step function),*

$$\frac{d}{da}h(Y) = \frac{1}{2a\sqrt{a}}\left\{\sqrt{a} - \mathbb{E}_X[X] + \mathbb{E}_X\big[\mathbb{E}_{X|Y}[X|Y]\,|Y=X\big]\right\}.$$

When the random variable W is exponentially distributed, assumptions in (2.17) are reduced to the existence of the moment generating function of X, as explained in Appendix A.9. Therefore, the assumptions in (2.17) for an exponential random variable are as simple as the assumptions (2.17) for a Gaussian random variable.

Corollary 2.3. *Given the channel model in (2.1) with an arbitrary but fixed non-negative random variable X whose moment generating function exists and a gamma random variable W with a shape parameter α ($\alpha \geq 2$) and an inverse scale parameter β ($\beta = 1$),*

$$\frac{d}{da}h(Y) \;=\; \frac{1}{2a\sqrt{a}}\Big\{\sqrt{a} - \mathbb{E}_X\left[X\right] + \mathbb{E}_{Y_{\alpha-1}}\left[\mathbb{E}_{X|Y}\left[X|Y\right]|Y = Y_{\alpha-1}\right]\Big\},$$

where $Y_k = X + \sqrt{a}W_k$, and W_k denotes a gamma random variable with shape parameter k. Notation Y_α stands for Y. As explained in Appendix A.9, the assumptions (2.17) are quite simplified in the presence of the moment generating function of random variable X.

For additive non-Gaussian noise channels, the differential entropy cannot be expressed in terms of the Fisher information. Instead, the differential entropy is expressed by the posterior mean as shown in Theorem 2.6. Fortunately, several noise distributions of interest in communication problems satisfy the required assumptions (2.17) in Theorem 2.6 (e.g., Gaussian, gamma, exponential, chi-square with restrictions on parameters, Rayleigh, etc.). Therefore, Theorem 2.6 is quite powerful. If the posterior mean $\mathbb{E}_{X|Y}[X|Y]$ is expressed by a polynomial function of Y, e.g., X and W are independent Gaussian random variables in equation (2.1) or random variables belonging to the natural exponential family of distributions [36], then equation (2.18) can be expressed in simpler forms.

Example 2.1. *Consider an additive white Gaussian noise (AWGN) channel. Given the channel model (2.1), let X and W be normally distributed with zero mean and unit variance. Assume X and W are independent of each other. Then, the posterior mean is expressed as*

$$\mathbb{E}_{X|Y}\left[X|Y=y\right] = \frac{1}{1+a}y,$$

which is linear to y. Therefore, equation (2.18) is expressed as

$$\begin{aligned}\frac{d}{da}h(Y) &= \frac{1}{2a}\left\{1-\mathbb{E}_Y\left[\frac{d}{dY}\mathbb{E}_{X|Y}\left[X|Y\right]\right]\right\} \\ &= \frac{1}{2(1+a)}.\end{aligned}$$

Now, we consider the second derivative of the differential entropy. One interesting property of the second derivative of the differential entropy is that it can always be expressed as a function of the Fisher information (2.7).

Theorem 2.7. *Given the channel model (2.1), let X and W be arbitrary random*

variables, independent of each other. Given the following assumptions:

$$\frac{d^2}{dy^2}\mathbb{E}_X\left[f_{Y|X}(y|X;a)\right]=\mathbb{E}_X\left[\frac{d^2}{dy^2}f_{Y|X}(y|X;a)\right],$$
$$\frac{d^2}{da^2}\mathbb{E}_X\left[f_{Y|X}(y|X;a)\right]=\mathbb{E}_X\left[\frac{d^2}{da^2}f_{Y|X}(y|X;a)\right], \tag{2.19a}$$
$$\frac{d^2}{da^2}\int_{-\infty}^{\infty} f_Y(y;a)\log f_Y(y;a)dy=\int_{-\infty}^{\infty}\frac{d^2}{da^2}\left(f_Y(y;a)\log f_Y(y;a)\right)dy, \tag{2.19b}$$
$$\lim_{y\to\pm\infty}\mathbb{E}_X\left[X^2\frac{\frac{d}{dy}f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}\right]=\mathbb{E}_X\left[\lim_{y\to\pm\infty}X^2\frac{\frac{d}{dy}f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}\right],$$
$$\lim_{y\to\pm\infty}\mathbb{E}_X\left[Xf_{Y|X}(y|X;a)\right]=\mathbb{E}_X\left[\lim_{y\to\pm\infty}Xf_{Y|X}(y|X;a)\right], \tag{2.19c}$$
$$\lim_{y\to\pm\infty}\mathbb{E}_X\left[f_{Y|X}(y|X;a)\right]=\mathbb{E}_X\left[\lim_{y\to\pm\infty}f_{Y|X}(y|X;a)\right],$$
$$\lim_{y\to\pm\infty} y^8 f_Y(y;a)=0, \tag{2.19d}$$
$$\left|\frac{\mathbb{E}_X\left[X^2f_{Y|X}(y|X;a)\right]}{(f_Y(y;a))^{3/4}}\right|<\infty, \tag{2.19e}$$

where $\mathbb{E}_{X|Y}[\cdot|\cdot]$ denotes the posterior mean, the following identity holds:

$$\frac{d^2}{da^2}h(Y) \;=\; -J_a(Y)-\frac{1}{2a}\frac{d}{da}h(Y)-\frac{1}{4a^2}\mathbb{E}_Y\left[\frac{d}{dY}S_Y(Y)\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]\right],$$

or equivalently,

$$\begin{aligned}\frac{d^2}{da^2}h(Y)=&-J_a(Y)-\frac{1}{4a^2}\mathbb{E}_Y\left[\frac{d}{dY}\mathbb{E}_{X|Y}\left[(Y-X)|Y\right]\right]\\&-\frac{1}{4a^2}\mathbb{E}_Y\left[\frac{d}{dY}S(Y)\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]\right].\end{aligned} \tag{2.20}$$

Proof. See Appendix A.3. □

Similar to the corollaries of Theorem 2.6, by specifying a noise distribution and manipulating equation (2.20) in Theorem 2.7, we derive the following corollaries.

Corollary 2.4. *Given the channel (2.1), let X be an arbitrary but fixed random variable with a finite second-order moment, and let W be independent normally distributed with zero mean and unit variance. Then,*

$$\begin{aligned}\frac{d^2}{da^2}h(Y) &= -J_a(Y) - \frac{1}{4a}J(Y) - \frac{1}{4a^2}\mathbb{E}_Y\left[\frac{d}{dY}S_Y(Y)\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]\right] \\ &= -\frac{1}{2}\mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)^2\right].\end{aligned}$$

Remark 2.3. *This result is a scalar version of the result reported in [42]. At the same time, this result is a special case, when X is a Gaussian random variable, of the general result in Theorem 2.7.*

Corollary 2.5. *Under the channel (2.1), let X be an arbitrary but fixed non-negative random variable with a finite moment generating function, and its pdf is bounded. Let W be independent exponentially distributed with unit value as the parameter (λ) of the distribution. Namely, $f_W(w) = exp(-w)\mathrm{U}(w)$, where $\mathrm{U}(\cdot)$ denotes the unit step function. Then,*

$$\begin{aligned}\frac{d^2}{da^2}h(Y) = &-J_a(Y) + \frac{3}{4a^2\sqrt{a}}\mathbb{E}_X\left[\mathbb{E}_{X|Y}\left[Y-X|Y\right]|Y{=}X\right] \\ &+\frac{1}{4a^2} - \frac{1}{4a^3}\mathbb{E}_X\left[\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]|Y{=}X\right].\end{aligned}$$

Corollary 2.6. *Under the channel (2.1), let X be an arbitrary but fixed non-negative random variable with a finite moment generating function, and W be an independent gamma random variable with parameters α ($\alpha \geq 3$) and β ($\beta = 1$), i.e., $f_W(w) = \beta^\alpha w^{\alpha-1}\exp(-\beta w)\mathrm{U}(w)/\Gamma(\alpha)$, where $\mathrm{U}(\cdot)$ denotes the unit step function and $\Gamma(\cdot)$*

stands for the gamma function. Then,

$$\begin{aligned}\frac{d^2}{da^2}h(Y) = &-\frac{1}{4a^3}\mathbb{E}_{Y_{\alpha-2}}\left[\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]|Y=Y_{\alpha-2}\right] \\ &-\frac{1}{4a^2\sqrt{a}}\mathbb{E}_{Y_{\alpha-1}}\left[\mathbb{E}_{X|Y}\left[X|Y\right]|Y=Y_{\alpha-1}\right] \\ &+\frac{(\alpha-1)}{4a^2\sqrt{a}}\mathbb{E}_{Y_{\alpha-1}}\left[\frac{\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]}{\mathbb{E}_{X|Y_{\alpha-1}}\left[Y_{\alpha-1}-X|Y_{\alpha-1}\right]}\middle| Y=Y_{\alpha-1}\right] \\ &-J_a(Y)-\frac{1}{4a^2\sqrt{a}}\left(\sqrt{a}-\mathbb{E}_X\left[X\right]\right),\end{aligned}$$

where $Y_\alpha = X + \sqrt{a}W_\alpha$*, and* W_α *denotes a gamma random variable with a shape parameter* α*.*

Like Corollaries 2.1, 2.2, and 2.3, the assumptions (2.19) reduce to simplified forms in Corollaries 2.4, 2.5, and 2.6. Even though we have not enumerated all possible probability density functions for Theorem 2.6 and Theorem 2.7, many of the probability density functions that present an exponential term satisfy the assumptions (2.17) and (2.19), since such a condition proves to be sufficient for the required interchange between a limit and a integral.

2.5 Applications

As mentioned in [18] and [43], De Bruijn's identity has been widely used in a variety of areas such as information theory, estimation theory, and so on. Similarly, De Bruijn-type identities mentioned in this section can be adopted in many applications. Here, we introduce several applications from the estimation theory realm as well as from the information theory field.

2.5.1 Applications in Estimation Theory

In estimation theory, there exist two fundamental lower bounds: Cramér-Rao lower bound (CRLB) and Bayesian Cramér-Rao lower bound (BCRLB). CRLB is

a lower bound for the estimation error of any unbiased estimator, and it is derived from a frequentist perspective. This lower bound is tight when the output distribution of the channel is Gaussian. CRLB and its tightness can be justified using Cauchy-Schwarz inequality [27]. On the other hand, BCRLB is a lower bound for the estimation error of any estimator, and it is calculated from a Bayesian perspective. BCRLB does not require unbiasedness of estimators unlike CRLB; however, BCRLB requires prior knowledge (i.e., distribution) of random parameters. BCRLB is also tight when all random variables are Gaussian [50].

Surprisingly, assuming a Gaussian additive noise channel, both of these lower bounds can be derived using De Bruijn-type identities, and there exist counterparts both in information theory and estimation theory. Since CRLB and its counterpart, the worst additive noise lemma, are derived in [43], we will only show the derivation of BCRLB and its counterpart in this section.

Lemma 2.1 (Bayesian Cramér-Rao Lower Bound)**.** *Given the channel (2.1), let $\hat{X}$ be an arbitrary estimator of X in a Bayesian estimation framework. Then, the mean square error (MSE) of $\hat{X}$ is lower bounded as follows:*

$$MSE(\hat{X}) \;\geq\; \frac{1}{\mathbb{E}_X\left[J(Y|X)\right] + J(X)},$$

where X is an arbitrary but fixed random variable with a finite second-order moment, W is a Gaussian random variable with zero mean and unit variance, and

$$J(Y|X) \;=\; \int_{-\infty}^{\infty} \left(\frac{d}{dx}\log f_{Y|X}(y|x)\right)^2 f_{Y|X}(y|x)dy. \tag{2.21}$$

Proof. See Appendix A.4. □

Interestingly, there exists a counterpart, based on differential entropies, of BCRLB

in information theory, and this counterpart is a tighter lower bound than BCRLB.

Lemma 2.2. *Under the same conditions as in Lemma 2.1,*

$$MSE(\hat{X}) \;\geq\; N(X|Y), \tag{2.22}$$

where $N(X|Y) = (1/2\pi e)\exp(2h(X|Y))$, $Y = X + \sqrt{a}W$, $a \geq 0$, *and* X *and* W *are independent of each other.*

Proof. See Appendix A.5. □

Remark 2.4. *Lemma 2.2 seems to be similar to the estimation counterpart of Fano's inequality [9, p. 255, Theorem 8.6.6]. However, the current result is completely different than [9, p. 255, Theorem 8.6.6]. In [9], to satisfy the inequality (2.22), the hidden assumption is*

$$Var(X|Y) \;=\; Var(X_G|Y_G), \tag{2.23}$$

where $Var(X|Y)$ *and* $Var(X_G|Y_G)$ *denote posterior variances for random variables* X *and* Y, *and Gaussian random variables* X_G *and* Y_G, *respectively. With the assumption (2.23), the following relations hold:*

$$\begin{aligned}
\mathbb{E}_{X,Y}\left[\left(X - \mathbb{E}_{X|Y}[X|Y]\right)^2\right] &= Var(X|Y) \\
&= Var(X_G|Y_G) \\
&= \frac{1}{2\pi e}\exp(2h(X_G|Y_G)) \\
&\geq \frac{1}{2\pi e}\exp(2h(X|Y)) \\
&= N(X|Y).
\end{aligned}$$

This is nothing but the entropy maximizing theorem, i.e., the Gaussian random variable being the one that maximizes the entropy among all real-valued distributions with fixed mean and variance.

However, under the assumptions $Var(X) = Var(X_G)$ *and* $Var(Y) = Var(Y_G)$*, which are common assumptions in signal processing problems, (2.23) may not be always true due to the following fact. Given the additive Gaussian noise channel,* $Y = X + \sqrt{a}W_G$*, where* X *is an arbitrary non-Gaussian random variable whose variance is identical to that of Gaussian random variable* X_G*, and* W_G *is a Gaussian random variable with zero mean and unit variance,*

$$Var(X|Y) \;\; < \;\; Var(X_G|Y_G), \tag{2.24}$$

where Y_G *is a Gaussian random variable whose variance is identical to that of* Y*. Equation (2.24) violates the assumption (2.23). Therefore, the result in [9, p. 255, Theorem 8.6.6] cannot be adopted under the assumptions,* $Var(X) = Var(X_G)$ *and* $Var(Y) = Var(Y_G)$*, which are common in signal processing problems.*

On the other hand, the inequality in Lemma 2.2 is obtained not by imposing identical posterior variances but by assuming identical second-order moments. Thus, (2.22) represents a lower bound on the mean square error similar to BCRLB. Therefore, Lemma 2.2 illustrates a novel lower bound on the mean square error from an information theoretic perspective.

Surprisingly, this lower bound is tighter than BCRLB as the following lemma indicates.

Lemma 2.3. *Under the same conditions as in Lemma 2.2,*

$$N(X|Y) \;\; \geq \;\; \frac{1}{\mathbb{E}_X\left[J(Y|X)\right] + J(X)}, \tag{2.25}$$

where $Y = X + \sqrt{a}W$*, a is nonnegative,* X *is an arbitrary but fixed random variable with a finite second-order moment,* W *is a Gaussian random variable with zero mean and unit variance, and* $J(Y|X)$ *is defined as equation (2.21). The equality holds if the random variable* X *is Gaussian.*

Proof. See Appendix A.6. □

Figure 2.1 illustrates how tighter the new lower bound (2.22) is compared to BCRLB when X is a student-t random variable, and W is a Gaussian random variable. The degrees of freedom of X is 3, and the variance of W is 1. As shown in Figure 2.1, the new lower bound is much tighter than BCRLB especially in low SNRs where the BCRLB is generally loose. Also, Figure 2.1 shows how tight the new lower bound is with respect to the minimum mean square error.

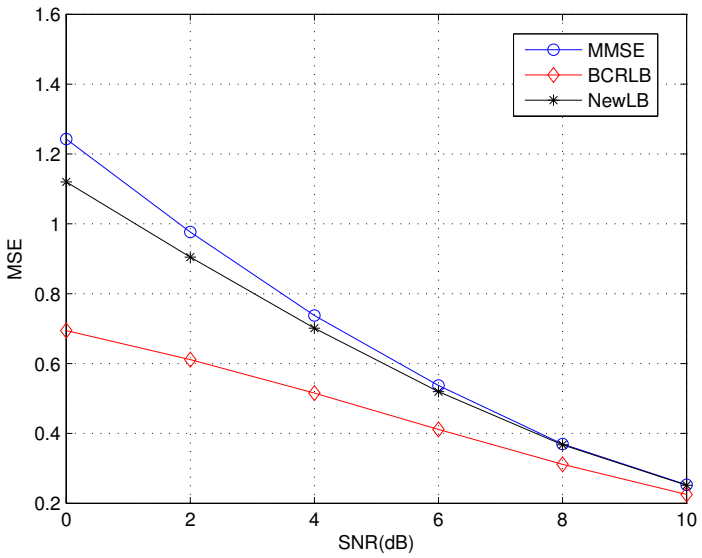

Figure 2.1: Comparison of MMSE, BCRLB, and new lower bound (New LB) in (2.22) with respect to SNR.

2.5.2 Applications in Information Theory

In information theory, the entropy power inequality (EPI) is one of the most important inequalities since it is helps to prove the channel capacity under several different circumstances, e.g., the capacity of scalar Gaussian broadcast channel [3], the capacity of Gaussian MIMO broadcast channel [53], [32], the secrecy capacity of Gaussian wire-tap channel [31], [41] and so on. The channel capacity can be proved not by EPI alone but by EPI in conjunction with Fano's inequality. Depending on the channel model, an additional technique, channel enhancement technique [53], is required. Therefore, various versions of the EPI such as a classical EPI [46], [45], [5], Costa's EPI [7], and an extremal inequality [32] were proposed by several different authors. In this section, we will prove Costa's entropy power inequality, a stronger version of a classical EPI using Theorem 2.7.

Lemma 2.4 (Costa's EPI)**.** *For a Gaussian random variable W with zero mean and unit variance,*

$$N(X+\sqrt{a}W) \;\geq\; (1-a)N(X)+aN(X+W), \tag{2.26}$$

where $0 \leq a \leq 1$, X and W are independent of each other, and the entropy power $N(X)$ is defined as $N(X) = (1/2\pi e)\exp(2h(X))$. Alternatively, the inequality (2.26) is expressed as

$$\frac{d^2}{da^2}N(X+\sqrt{a}W) \;\leq\; 0, \tag{2.27}$$

i.e., $N(X+\sqrt{a}W)$ is a concave function of a [7].

Proof. See Appendix A.7. □

2.5.3 Applications in Other Areas

There are many other applications of the proposed results. First, since Theorem 2.6 is equivalent to Theorem 1 in [20], Theorem 2.6 can be used for applications such as generalized EXIT charts and power allocation in systems with parallel non-Gaussian noise channels as mentioned in [20]. Second, by Theorem 2.4, we showed the equivalence among Stein, De Bruijn, and heat equation identities. Therefore, a broad range of problems (in probability, decision theory, Bayesian statistics and graph theory) as described in [6] could be considered as additional potential applications of Theorems 2.4 and 2.6.

2.6 Conclusions

This section mainly disclosed three information-estimation relationships. First, the equivalence between Stein identity and De Bruijn identity was proved. Second, it was proved that the first derivative of the differential entropy with respect to the parameter a can be expressed in terms of the posterior mean. Second, this section showed that the second derivative of the differential entropy with respect to the parameter a can be expressed in terms of the Fisher information. Finally, several applications based on the three main results listed above were provided. The suggested applications illustrate that the proposed results are useful not only in information theory but also in the estimation theory field and other fields.

3. GAUSSIAN ASSUMPTION: OPTIMAL ESTIMATION*

3.1 Introduction

Gaussian assumption is the most well-known and widely used distribution in many fields such as engineering, statistics and physics. One of the major reasons why the Gaussian distribution has become so prominent is because of the Central Limit Theorem (CLT) and the fact that the distribution of noise in numerous engineering systems is well captured by the Gaussian distribution. Moreover, features such as analytical tractability and easy generation of other distributions from the Gaussian distribution contributed further to the popularity of Gaussian distribution. Especially, when there is no information about the distribution of observations, Gaussian assumption appears as the most conservative choice. This follows from the fact that the Gaussian distribution minimizes the Fisher information, which is the inverse of the Cramér-Rao lower bound (CRLB) (or equivalently stated, the Gaussian distribution maximizes the CRLB). Therefore, any optimization based on the CRLB under the Gaussian assumption can be considered to be min-max optimal in the sense of minimizing the largest CRLB (see [48] and the references cited therein).

Inspired by the early isoperimetric inequality for entropy introduced by Costa and Cover [8] and the more recent results of Rioul [43], Stoica and Babu [48], the goals of this section are threefold: i) to illustrate a connection between [48] and the recent information theoretic results reported in [8], [43], ii) to present information theoretic and estimation theoretic justifications for the fact that the Gaussian assumption

*Reprinted with permission from "Gaussian Assumption: the Least Favorable but the Most Useful," Sangwoo Park, Erchin Serpedin, and Khalid Qaraqe, accepted for the publication in IEEE Signal Processing Magazine, Copyright by IEEE.

leads to the largest CRLB, iii) to show a slight extension of this result to the more general framework of correlated observations. Even though Stoica and Babu provided a simple and quite general proof of result that the largest CRLB is achievable by the Gaussian distribution, the proposed proof is only applicable to the situation when the observations are independent, i.e., the observation noise is white [48]. However, this result can be generalized to arbitrary correlations among samples. In many practical circumstances, the correlation of the noise is inevitable since the observed data comes from a filter, and the filter introduces correlation. Therefore, the importance of this generalization cannot be ignored. This result is also closely related to two well-known results in information theory: first, the fact that a Gaussian random vector maximizes a differential entropy, and second, the worst additive noise lemma (see [43], [11], and the references cited therein). Several researchers have investigated relationships between estimation theoretic (statistical) concepts such as mean-square error and Fisher information and information theoretic concepts such as entropy and mutual information (see e.g., [8], [43] and the references cited therein). However, most of these results are inclined to be rather theoretical than practical. In this section, we show how some of these results can be adopted to a more practical application involving the estimation of a communication channel via a training sequence.

The approach introduced herein section can be adapted to optimally estimate unknown (deterministic or random) parameters in additive noise channels. As presented in the channel model (3.1), the additive noise channel is very general in the sense that the only assumption is the independence between data $\mathbf{x}_{\boldsymbol{\theta}}$ and noise $\mathbf{w}$. Namely, the channel model does not require the Gaussian noise assumption, it admits correlation among noise terms, and it also allows for correlation among data terms. Therefore, the proposed approach can be generally used in signal processing applications involving parameter estimation, spectrum estimation, optimization,

wireless communications and information theory.

3.2 Problem Statement

Consider a random vector $\mathbf{Y}$ which is generated by the following system of equations:

$$\mathbf{Y} \;=\; \mathbf{X}_{\boldsymbol{\theta}} + \mathbf{W}, \tag{3.1}$$

where $\mathbf{Y}$ is an $n \times 1$ observed random vector, $\mathbf{X}_{\boldsymbol{\theta}}$ denotes an $n \times 1$ signal (random) vector which depends on a $k \times 1$ unknown deterministic parameter vector $\boldsymbol{\theta}$, and $\mathbf{W}$ stands for the $n \times 1$ zero-mean noise vector whose covariance matrix is $\boldsymbol{\Sigma}_{\mathbf{w}}$. Random vectors $\mathbf{X}_{\boldsymbol{\theta}}$ and $\mathbf{W}$ are assumed independent of each other. The systems represented by the channel model (3.1) are quite numerous. In particular, the channel model (3.1) might consist of the samples of an arbitrary stochastic process such as ARMA (autoregressive moving average) or ARMAX (ARMA with eXogenous inputs), as mentioned in [48].

Based on the channel model (3.1), we define the score function:

$$\mathbf{s}(\boldsymbol{\theta}) \;=\; \nabla_{\boldsymbol{\theta}} \log f_{\mathbf{Y}|\mathbf{X}_{\boldsymbol{\theta}}}(\mathbf{y}|\mathbf{x}_{\boldsymbol{\theta}}), \tag{3.2}$$

where $\nabla_{\boldsymbol{\theta}}$ denotes the gradient with respect to $\boldsymbol{\theta}$, and $f_{\mathbf{Y}|\mathbf{X}_{\boldsymbol{\theta}}}(\mathbf{y}|\mathbf{x}_{\boldsymbol{\theta}})$ is the conditional density function of $\mathbf{Y}$ given $\mathbf{X}_{\boldsymbol{\theta}}$. The Cramér-Rao lower bound (CRLB) is expressed by the diagonal elements of the inverse of the Fisher information matrix (FIM), and the FIM is represented as:

$$\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y}) \;=\; \mathbb{E}_{\mathbf{Y}}[\mathbf{s}(\boldsymbol{\theta})\mathbf{s}(\boldsymbol{\theta})^T], \tag{3.3}$$

where the notation $\mathbb{E}_{\mathbf{Y}}[\cdot]$ stands for the expectation with respect to a random vector $\mathbf{Y}$, and superscript T denotes the operation of transposition for a vector or matrix.

Our goal is to find an optimal estimator for the parameter $\boldsymbol{\theta}$ in the sense that the estimated parameter minimizes the lower bound of the mean square error of the estimator in the worst case scenario.

3.3 Minium Fisher Information-A Statistical Viewpoint

One of the common approaches to estimate unknown parameters is to build estimators that minimize the Cramer-Rao lower bound. Since CRLB is expressed as the inverse of FIM, minimizing the Cramér-Rao lower bound is equivalent to maximizing FIM. Given the channel model (3.1), the score function in (3.2) and the FIM in (3.3) can be re-expressed by the following procedure.

Since $f_{\mathbf{Y}|\mathbf{X}_{\boldsymbol{\theta}}}(\mathbf{y}|\mathbf{x}_{\boldsymbol{\theta}}) = f_{\mathbf{W}}(\mathbf{w})\big|_{\mathbf{w}=\mathbf{y}-\mathbf{x}_{\boldsymbol{\theta}}} = f_{\mathbf{W}}(\mathbf{y}-\mathbf{x}_{\boldsymbol{\theta}})$, where $f_{\mathbf{W}}(\cdot)$ denotes the density function of the noise $\mathbf{W}$, and $\mathbf{X}_{\boldsymbol{\theta}}$ and $\mathbf{W}$ are independent of each other, using the chain rule for computing the derivative of a function, the score function $\mathbf{s}(\boldsymbol{\theta})$ is re-written as:

$$\begin{aligned}
\mathbf{s}(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \log f_{\mathbf{Y}|\mathbf{X}_{\boldsymbol{\theta}}}(\mathbf{y}|\mathbf{x}_{\boldsymbol{\theta}}) \\
&= \nabla_{\boldsymbol{\theta}} \log f_{\mathbf{W}}(\mathbf{y}-\mathbf{x}_{\boldsymbol{\theta}}) \\
&= -\nabla_{\boldsymbol{\theta}}\mathbf{x}_{\boldsymbol{\theta}} \nabla_{\mathbf{w}} \log f_{\mathbf{W}}(\mathbf{w}), \qquad (3.4)
\end{aligned}$$

where the gradient (Jacobian) of the vector $\mathbf{x}_{\boldsymbol{\theta}}$ is defined as the $k \times n$ matrix $\nabla_{\boldsymbol{\theta}}\mathbf{x}_{\boldsymbol{\theta}}$ with its $(i,j)th$ entry equal to $\frac{\partial x_{\theta,j}}{\partial \theta_i}$. Now it turns out that the FIM (3.3) can be

expressed as:

$$\begin{aligned}
\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y}) &= \mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}},\mathbf{W}}\left[\left(\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\nabla_{\mathbf{W}}\log f_{\mathbf{W}}(\mathbf{W})\right)\left(\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\nabla_{\mathbf{W}}\log f_{\mathbf{W}}(\mathbf{W})\right)^T\right] \\
&= \mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}},\mathbf{W}}\left[\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\left(\nabla_{\mathbf{W}}\log f_{\mathbf{W}}(\mathbf{W})\nabla_{\mathbf{W}}\log f_{\mathbf{W}}(\mathbf{W})^T\right)\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}^T\right] \qquad (3.5) \\
&= \mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\mathbf{J}(\mathbf{W})\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}^T\right], \qquad (3.6)
\end{aligned}$$

where the FIM with respect to $\mathbf{W}$ is defined as

$$\mathbf{J}(\mathbf{W}) = \mathbb{E}_{\mathbf{W}}\left[\nabla_{\mathbf{W}}\log f_{\mathbf{W}}(\mathbf{W})\nabla_{\mathbf{W}}\log f_{\mathbf{W}}(\mathbf{W})^T\right]. \qquad (3.7)$$

In equation (3.5), the expectation with respect to both $\mathbf{X}_{\boldsymbol{\theta}}$ and $\mathbf{W}$ can be separated into the outer expectation with respect to $\mathbf{X}_{\boldsymbol{\theta}}$ and the inner expectation with respect to $\mathbf{W}$ since $\mathbf{X}_{\boldsymbol{\theta}}$ and $\mathbf{W}$ are independent of each other. When the vector $\mathbf{X}_{\boldsymbol{\theta}}$ is deterministic, the outer expectation is not required. Therefore, the term related to the random vector $\mathbf{W}$ becomes the FIM, $\mathbf{J}(\mathbf{W})$, defined in equation (3.7), and it is not affected by the outer expectation $\mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}}}[\cdot]$ in equation (3.6).

The following result states that the FIM $\mathbf{J}(\mathbf{W})$, which is a positive semi-definite matrix, is lower-bounded by the FIM $\mathbf{J}(\mathbf{W}_G)$ of a normally distributed random vector $(\mathbf{W}_G)$.

Lemma 3.1 (Cramér-Rao Inequality)**.** *For a random vector* $\mathbf{W}$ *and a Gaussian random vector* $\mathbf{W}_G$ *whose covariance matrix* $\mathbf{\Sigma}_{\mathbf{W}}$ *is identical to the covariance matrix of* $\mathbf{W}$*, the following inequality is satisfied:*

$$\mathbf{J}(\mathbf{W}) \succeq \mathbf{J}(\mathbf{W}_G),$$

where notation $\succeq$ *stands for "greater than or equal to", in the sense of the partial*

ordering of positive semi-definite matrices.

Proof. The proof follows essentially [43]. First, we define the following two score functions:

$$\begin{aligned} \mathbf{s}_{\mathbf{W}}(\mathbf{w}) &= \nabla_{\mathbf{w}} \log f_{\mathbf{W}}(\mathbf{w}), \\ \mathbf{s}_{\mathbf{W}_G}(\mathbf{w}) &= \nabla_{\mathbf{w}} \log f_{\mathbf{W}_G}(\mathbf{w}). \end{aligned} \tag{3.8}$$

The covariance matrix of the difference of the two score functions (3.8) is expressed as

$$\mathbb{E}_{\mathbf{W}} \left[\left(\mathbf{s}_{\mathbf{W}}(\mathbf{W}) - \mathbf{s}_{\mathbf{W}_G}(\mathbf{W}) \right) \left(\mathbf{s}_{\mathbf{W}}(\mathbf{W}) - \mathbf{s}_{\mathbf{W}_G}(\mathbf{W}) \right)^T \right], \tag{3.9}$$

and it is always greater than or equal to the zero matrix $\mathbf{0}$ in terms of the positive semi-definite partial ordering. Notice further that (3.9) can be simplified to

$$\begin{aligned} & \mathbb{E}_{\mathbf{W}} \left[\left(\mathbf{s}_{\mathbf{W}}(\mathbf{W}) - \mathbf{s}_{\mathbf{W}_G}(\mathbf{W}) \right) \left(\mathbf{s}_{\mathbf{W}}(\mathbf{W}) - \mathbf{s}_{\mathbf{W}_G}(\mathbf{W}) \right)^T \right] \\ = \; & \mathbf{J}(\mathbf{W}) - \mathbb{E}_{\mathbf{W}} \left[\mathbf{s}_{\mathbf{W}}(\mathbf{W}) \mathbf{s}_{\mathbf{W}_G}(\mathbf{W})^T \right] - \mathbb{E}_{\mathbf{W}} \left[\mathbf{s}_{\mathbf{W}_G}(\mathbf{W}) \mathbf{s}_{\mathbf{W}}(\mathbf{W}) \right] + \mathbf{J}(\mathbf{W}_G) \\ = \; & \mathbf{J}(\mathbf{W}) - \mathbf{J}(\mathbf{W}_G). \end{aligned} \tag{3.10}$$

Since $\mathbf{W}_G$ is a Gaussian random vector, $\mathbf{s}_{\mathbf{W}_G}(\mathbf{w}) = -\mathbf{\Sigma}_{\mathbf{W}}^{-1}\mathbf{w}$. $\mathbb{E}_{\mathbf{W}}\left[\mathbf{s}_{\mathbf{W}}(\mathbf{W})\mathbf{s}_{\mathbf{W}_G}(\mathbf{W})^T\right] = -\int \left(\nabla_{\mathbf{w}} f_{\mathbf{W}}(\mathbf{w})\right) \mathbf{w}^T d\mathbf{w} \mathbf{\Sigma}_{\mathbf{W}}^{-1} = \int f_{\mathbf{W}}(\mathbf{w}) d\mathbf{w} \mathbf{\Sigma}_{\mathbf{W}}^{-1} = \mathbf{\Sigma}_{\mathbf{W}}^{-1}$ by Green's identity (see e.g., [8] and the references cited therein). Here, Green's identity plays the role of the integration by parts for a vector. Since $\mathbf{J}(\mathbf{W}_G) = \mathbf{\Sigma}_{\mathbf{W}}^{-1}$, the last equality in equation (3.10) is verified. Since the covariance matrix is always positive semi-definite, from

equation (3.10),

$$\mathbb{E}_{\mathbf{W}}\left[\left(\mathbf{s}_{\mathbf{W}}(\mathbf{W})-\mathbf{s}_{\mathbf{W}_G}(\mathbf{W})\right)\left(\mathbf{s}_{\mathbf{W}}(\mathbf{W})-\mathbf{s}_{\mathbf{W}_G}(\mathbf{W})\right)^T\right]=\mathbf{J}(\mathbf{W})-\mathbf{J}(\mathbf{W}_G)\succeq\mathbf{0}. \quad (3.11)$$

Therefore, the proof is completed. □

Due to Lemma 3.1, when $\mathbf{W}$ is a Gaussian random vector, the FIM $\mathbf{J}(\mathbf{W})$ is minimized, and consequently the FIM $\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y})$ is also minimized:

$$\begin{aligned}\mathbf{J}_\theta(\mathbf{Y}) &= \mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\mathbf{J}(\mathbf{W})\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}^T\right]\\ &\succeq \mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\mathbf{J}(\mathbf{W}_G)\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}^T\right]\\ &= \mathbf{J}_\theta(\bar{\mathbf{Y}}), \end{aligned} \quad (3.12)$$

where $\bar{\mathbf{Y}}=\mathbf{X}_{\boldsymbol{\theta}}+\mathbf{W}_G$, and the equality holds if and only if $\mathbf{W}$ is normally distributed. The inequality in equation (3.12) is due to the fact that for an arbitrary matrix $\mathbf{C}$, the inequality $\mathbf{CAC}^T \preceq \mathbf{CBC}^T$ holds whenever positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$ satisfy $\mathbf{A}\preceq\mathbf{B}$.

From equations (3.6) and (3.12), we know that the CRLB depends on the parameter θ only through the FIM, $\mathbf{J}(\mathbf{W})$. In other words, the CRLB only depends on $\mathbf{J}(\mathbf{W})$ when $\mathbf{X}_\theta$ is fixed. Therefore, the Gaussian random vector $\mathbf{W}_G$ maximizes the CRLB (or, equivalently minimizes the FIM, $\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y})$), when $\mathbf{X}_{\boldsymbol{\theta}}$ is fixed. Therefore, any design which optimizes the FIM (3.6) (or equivalently the CRLB) when the random vector $\mathbf{W}$ is Gaussian, can be considered min-max optimal in the light of generating the smallest FIM (or the largest CRLB) in the worst situation.

3.4 Minimum Mutual Information–An Information Theoretic Viewpoint

It is well-known that, given the covariance matrix, a Gaussian random vector minimizes the FIM, a result referred to as the Cramér-Rao inequality (see [48], [43], and the references cited therein). On the other hand, a Gaussian random vector maximizes a differential entropy when the covariance matrix is given (see [43], [9], and the references cited therein). These two results are closely related to each other. First, consider this relationship for random variables. Given a random variable W and a Gaussian random variable W_G, the following inequalities are satisfied:

- $J(W) \geq J(W_G)$ when $N(W) = N(W_G)$,
- $N(W) \geq N(W_G)$ when $J(W) = J(W_G)$,

where $N(\cdot)$ denotes the entropy power of a random variable, and $J(\cdot)$ stands for the Fisher information of a random variable. The above inequalities are easily derived from this general inequality

$$N(W)J(W) \;\; \geq \;\; 1, \tag{3.13}$$

where the equality holds if and only if W is Gaussian. The inequality (3.13) is referred to as the isoperimetric inequality for entropies (see [8], [10], and the references cited therein).

When the variance of W is equal to the variance of W_G, the inequality $J(W) \geq J(W_G)$ can be derived from $N(W) \leq N(W_G)$ using the isoperimetric inequality for entropies. However, we cannot derive the inequality $N(W) \leq N(W_G)$ from $J(W) \geq J(W_G)$ using the isoperimetric inequality. Instead, the worst additive noise lemma (see e.g., [43], [11], [23] and the references cited therein) can be derived from the inequality $J(W) \geq J(W_G)$ when the variances of W and W_G are identical. All

the relationships mentioned above are also valid for random vectors if we substitute either $|\mathbf{J}(\cdot)|^{\frac{1}{n}}$ or $\mathbf{Tr}\{\mathbf{J}(\cdot)\}$ for $J(\cdot)$. The trace and the determinant of a matrix are represented by the notations $\mathbf{Tr}\{\cdot\}$ and $|\cdot|$, respectively. Since the vector generalization is quite direct, these results are not mentioned here except the following lemma.

Lemma 3.2 (Worst Additive Noise Lemma [11], [23])**.** *For a random vector* $\mathbf{W}$ *and a Gaussian random vector* $\mathbf{W}_G$ *whose covariance matrices are identical to each other,*

$$I(\mathbf{W}+\mathbf{Z}_G;\mathbf{Z}_G) \;\geq\; I(\mathbf{W}_G+\mathbf{Z}_G;\mathbf{Z}_G), \tag{3.14}$$

where $I(\cdot;\cdot)$ *stands for mutual information,* $\mathbf{Z}_G$ *is a Gaussian random vector with zero mean and covariance matrix* $\mathbf{\Sigma_Z}$*, and all random vectors are independent of one another.*

Similar to Cramér-Rao inequality (see [48], [43], and the Lemma 3.1), the worst additive noise lemma shows that the mutual information $I(\mathbf{W}+\mathbf{Z}_G;\mathbf{Z}_G)$ is minimized when $\mathbf{W}$ is Gaussian. Consider that notation $h(\cdot)$ stands for differential entropy, and define the function:

$$g(\mathbf{\Sigma_Z}) = h(\mathbf{W}+\mathbf{Z}_G) - h(\mathbf{W}_G+\mathbf{Z}_G) - h(\mathbf{W}) + h(\mathbf{W}_G). \tag{3.15}$$

The function $g(\cdot)$ is non-decreasing with respect to the covariance matrix $\mathbf{\Sigma_Z}$ near the zero matrix $\mathbf{0}$. This is because, due to Lemma 3.2, $g(\mathbf{\Sigma_Z})$ is always non-negative for a covariance matrix $\mathbf{\Sigma_Z}$ which is arbitrarily close to the zero matrix $\mathbf{0}$. Therefore, near the zero matrix, the first derivative of $g(\mathbf{\Sigma_Z})$ with respect to $\mathbf{\Sigma_Z}$ is always positive semi-definite, and using a vector version of De Bruijn's identity [40], the Cramér-Rao

inequality is derived from the Lemma 3.2 as follows:

$$\begin{aligned} & \nabla_{\mathbf{\Sigma_Z}} g(\mathbf{\Sigma_Z})\Big|_{\mathbf{\Sigma_Z}=\mathbf{0}} \succeq \mathbf{0} \\ \iff \quad & \nabla_{\mathbf{\Sigma_Z}} I(\mathbf{W}+\mathbf{Z}_G;\mathbf{Z}_G)\Big|_{\mathbf{\Sigma_Z}=\mathbf{0}} - \nabla_{\mathbf{\Sigma_Z}} I(\mathbf{W}_G+\mathbf{Z}_G;\mathbf{Z}_G)\Big|_{\mathbf{\Sigma_Z}=\mathbf{0}} \succeq \mathbf{0} \\ \iff \quad & \mathbf{J}(\mathbf{W}) - \mathbf{J}(\mathbf{W}_G) \succeq \mathbf{0}, \end{aligned} \tag{3.16}$$

where $\iff$ stands for equivalence.

Therefore, in equation (3.6), the FIM, $\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y})$, is expressed as

$$\begin{aligned} \mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y}) &= \mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\mathbf{J}(\mathbf{W})\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}^T\right] \\ &= 2\mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\left(\nabla_{\mathbf{\Sigma_Z}} I(\mathbf{W}+\mathbf{Z}_G;\mathbf{Z}_G)\Big|_{\mathbf{\Sigma_Z}=\mathbf{0}}\right)\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}^T\right], \end{aligned} \tag{3.17}$$

the smallest FIM, $\mathbf{J}_{\boldsymbol{\theta}}(\bar{\mathbf{Y}})$, in (3.12) is expressed as

$$\mathbf{J}_{\boldsymbol{\theta}}(\bar{\mathbf{Y}}) = 2\mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\left(\nabla_{\mathbf{\Sigma_Z}} I(\mathbf{W}_G+\mathbf{Z}_G;\mathbf{Z}_G)\Big|_{\mathbf{\Sigma_Z}=\mathbf{0}}\right)\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}^T\right], \tag{3.18}$$

and

$$\begin{aligned} & \mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\left(\nabla_{\mathbf{\Sigma_Z}} I(\mathbf{W}+\mathbf{Z}_G;\mathbf{Z}_G)\Big|_{\mathbf{\Sigma_Z}=\mathbf{0}}\right)\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}^T\right] \\ \succeq \; & \mathbb{E}_{\mathbf{X}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}\left(\nabla_{\mathbf{\Sigma_Z}} I(\mathbf{W}_G+\mathbf{Z}_G;\mathbf{Z}_G)\Big|_{\mathbf{\Sigma_Z}=\mathbf{0}}\right)\nabla_{\boldsymbol{\theta}}\mathbf{X}_{\boldsymbol{\theta}}^T\right]. \end{aligned} \tag{3.19}$$

Therefore, one can do the min-max optimal design based on equations (3.17), (3.18), and (3.19).

3.5 Practical Applications

The min-max approach can be adopted to many applications. One of the typical examples is the optimal training sequence design for estimating frequency-selective

fading channels [49], [4]. As a distinctive feature to what was shown in [49], [4], the proposed approach does not require neither the assumption of Gaussian noise nor the white noise assumption.

Assume that a linearly modulated signal filtered through a frequency-selective channel is modeled as follows:

$$\mathbf{Y} \;=\; \mathbf{X}_{\omega_0}\mathbf{S}\mathbf{h} + \mathbf{W}, \tag{3.20}$$

where $\mathbf{Y} = [Y_0, \cdots, Y_{n-1}]^T$, $\mathbf{W} = [W_0, \cdots, W_{n-1}]^T$, $\mathbf{h} = [h_0, \cdots, h_{m-1}]^T$,

$$\mathbf{X}_{\omega_0} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & e^{i\omega_0} & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & e^{i(n-1)\omega_0} \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} s_0 & s_{-1} & \cdots & s_{1-m} \\ s_1 & s_0 & \cdots & s_{2-m} \\ \vdots & \cdots & \ddots & \vdots \\ s_{n-1} & s_{n-2} & \cdots & s_{n-m} \end{bmatrix}, \tag{3.21}$$

$\omega_0 = 2\pi f_0$ is the frequency offset, $\{s_{1-m}, \ldots, s_{n-1}\}$ stands for the training sequence samples, and $\{h_0, \ldots, h_{m-1}\}$ denote the taps of the channel impulse response, assumed of finite length m. The noise $\mathbf{W}$ is an arbitrary random vector with zero mean and noise covariance matrix $\mathbf{\Sigma_W}$.

Since we want to find the optimal training sequences to estimate the channel impulse response and the frequency offset, we first define the unknown parameter vector $\boldsymbol{\theta}$ as $[\omega_0, \mathbf{h}_R, \mathbf{h}_I]^T$, where $\mathbf{h}_R$ and $\mathbf{h}_I$ denote the real and the imaginary parts of the channel $\mathbf{h}$.

Based on equation (3.6),

$$\begin{aligned}
\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y}) &= \mathfrak{Re}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\xi}_{\boldsymbol{\theta}}\mathbf{J}(\mathbf{W})\nabla_{\boldsymbol{\theta}}\boldsymbol{\xi}_{\boldsymbol{\theta}}^{H}\right] && (3.22)\\
&\succeq \mathfrak{Re}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\xi}_{\boldsymbol{\theta}}\mathbf{J}(\mathbf{W}_G)\nabla_{\boldsymbol{\theta}}\boldsymbol{\xi}_{\boldsymbol{\theta}}^{H}\right] && (3.23)\\
&\succeq \mathfrak{Re}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\xi}_{\boldsymbol{\theta}}(\lambda_{\min}\mathbf{I})\nabla_{\boldsymbol{\theta}}\boldsymbol{\xi}_{\boldsymbol{\theta}}^{H}\right] && (3.24)\\
&= \lambda_{\min}\mathfrak{Re}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\xi}_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\boldsymbol{\xi}_{\boldsymbol{\theta}}^{H}\right], && (3.25)
\end{aligned}$$

where $\boldsymbol{\xi}_{\boldsymbol{\theta}} = \mathbf{X}_{\omega_0}\mathbf{Sh}$, $\lambda_{\min}$ represents the minimum eigenvalue of the FIM, $\mathbf{J}(\mathbf{W}_G)$, $\mathfrak{Re}[\cdot]$ denotes the real part of a vector or matrix, and superscript H stands for Hermitian transposition. Since $\boldsymbol{\xi}_{\boldsymbol{\theta}}$ is a complex-valued function which only depends on the unknown deterministic real parameters, in equation (3.22), the equality holds with $\mathfrak{Re}[\cdot]$ and without the expectation. Due to the Lemma 3.1, equation (3.23) is verified, and equation (3.24) is satisfied due to the eigenvalue decomposition.

Equation (3.25) reveals the smallest FIM. It generates the worst CRLB, and it is exactly of the same form as the one shown in [49]. Using the same argument as in [49], the white training sequence is min-max optimal in this case. This min-max approach heavily depends on how much information we have about the unknown parameters. If we know the distribution of the noise vector $\mathbf{W}$, then the min-max approach will be adopted based on equation (3.22), while equation (3.23) will be used when we only know the covariance matrix of the noise vector $\mathbf{W}$. In both cases, the white training sequences are not optimal since the optimal design is affected by the FIM, $\mathbf{J}(\mathbf{W})$, which is related to the correlation of $\mathbf{W}$. The optimal sequences may depend on either the noise distribution or, at least, the noise covariance matrix. However, without any information about the noise vector $\mathbf{W}$, the white training sequences are optimal in the sense of minimizing the worst CRLB.

The presented result, i.e., for a colored noise $\mathbf{W}$ with given correlation matrix,

its FIM $\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y})$ is minimized when the random vector $\mathbf{W}$ is Gaussian, can be also interpreted from a different standpoint as follows. In equation (3.1), assume $\mathbf{Y}$ is passed through a whitening filter, and a new signal $\tilde{\mathbf{Y}}$ is obtained. The noise present in the new output $\tilde{\mathbf{Y}}$ is white since the correlation of the noise is eliminated by the whitening filter. Therefore, we can directly adopt the method proposed in [49]. However, the design of the whitening filter requires the covariance matrix of the noise $\mathbf{W}$. If we have information about the covariance matrix of $\mathbf{W}$, we can construct the optimal training sequences; if we do not have information about $\mathbf{W}$, we have to follow the method proposed in equations (3.24) and (3.25), and use the fact that the covariance matrix is lower-bounded by the minimum eigenvalue of the covariance matrix multiplied by the identity matrix.

3.6 Conclusions

The results provided in previous sections show that, given the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{W}}$, the FIM $\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y})$, (CRLB) is minimized (respectively maximized) by adopting the Gaussian assumption. This fact leads to the min-max optimal approach in the following sense: the FIM $\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y})$ (CRLB) depends on the unknown parameters only through the FIM $\mathbf{J}(\mathbf{W})$. Since the Gaussian noise (not necessarily white) minimizes the FIM $\mathbf{J}(\mathbf{W})$, it also minimizes the FIM $\mathbf{J}_{\boldsymbol{\theta}}(\mathbf{Y})$ (or equivalently, it maximizes the CRLB). Therefore, the optimal design under the Gaussian assumption yields the best CRLB in the worst case. The CRLB is also expressed using the mutual information. In the information theoretic viewpoint, the fact that a Gaussian random vector minimizes the FIM given the covariance matrix is related to the worst additive noise lemma and the fact that a Gaussian random vector maximizes the differential entropy given the covariance matrix.

4. EXTREMAL ENTROPY INEQUALITY

4.1 Introduction

The classical entropy power inequality (EPI) was first established by Shannon [45]. Due to its importance and usefulness, EPI was proved by several different authors using distinct methods. In [46], Stam provided the first rigorous proof, and Stam's proof was further simplified by Blachman [5] and Dembo et al. [10], respectively. Verdú and Guo proposed a new proof of the EPI based on the I-MMSE concept [51]. Most recently, Rioul proved the EPI based only on information theoretic quantities [43]. Before Rioul's proof, most of the reported proofs were based on De Bruijn-type identities and Fisher information inequality, i.e., the previous proofs were performed mainly based on estimation-theoretic techniques rather than information-theoretic techniques.

Due to the significance of the EPI, numerous versions of EPIs such as Costa's EPI [7], the EPI for dependent random variables [25], the extremal entropy inequality [32], etc., have been proposed. Among the EPIs, the extremal entropy inequality is especially prominent since it can be adapted to several important applications investigated recently in the wireless communications area. In [32], Liu and Viswanath proposed the extremal entropy inequality, 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, and suggested several applications for the extremal entropy inequality. The extremal entropy inequality is an entropy power inequality which includes a covariance constraint. Because of the covariance constraint, the extremal inequality could not be proved directly by using the classical EPI. Therefore, a new technique, referred to as the channel enhancement

technique [53], was adopted in the proofs reported in [32].

The proofs proposed in [32] proceed as follows. First, the extremal entropy inequality is cast as an optimization problem. Using the channel enhancement technique, which relies mainly on Karush-Kuhn-Tucker (KKT) conditions, an alternative optimization problem, whose maximum value is larger than the maximum value of the original problem, is proposed, and the alternative problem is solved using the EPI. Finally, the proof is completed by showing that the maximum value of the alternative problem is equal to the maximum value of the original problem. Even though Liu and Viswanath proposed two kinds of proofs, a direct proof and a perturbation proof, both proofs are commonly based on the channel enhancement technique, and they are derived in a similar way except De Bruijn's identity is adapted in the perturbation proof.

The main theme of this section is how to prove the extremal entropy inequality without using the channel enhancement technique. Since the channel enhancement technique is adapted to prove not only the extremal entropy inequality but also the capacity of several different kinds of Gaussian channels, e.g., the capacity of the Gaussian broadcast channel and the secrecy capacity of the Gaussian wire-tap channel, by finding an alternative proof for the extremal entropy inequality, we can also find novel techniques to calculate the capacity of Gaussian broadcast channel, the secrecy capacity of Gaussian wire-tap channel, and so on.

Our proof is mainly based on four techniques: the data processing inequality, the moment generating function, the worst additive noise lemma, and the classical EPI. By using the data processing inequality, the worst additive noise lemma, and the classical EPI, we calculate an upper bound. Then, by applying the equality condition of the data processing inequality, we prove that the upper bound can be achieved. The moment generating functions are implemented to prove the achievement of the

equality condition in the data processing inequality.

The contribution of our proof can be summarized as follows. First, our proof is simpler and more direct compared with the proofs in [32]. Second, we adapt a more information-theoretic approach without using the KKT conditions. The method based on the data processing inequality and the moment generating function enables us to circumvent the step of using the KKT conditions. Moreover, by simply analyzing some properties of positive semi-definite matrices, we can omit the step of proving the existence of the optimal solution which satisfies the KKT conditions, a step which is very complicated to accomplish. In addition, the structure of the covariance matrix of the optimal solution is mentioned in detail by using properties of positive semi-definite matrices. Third, our proof presents a novel investigation method not only for the extremal entropy inequality but also for applications such as the capacity of Gaussian broadcast channel, the secrecy capacity of Gaussian wire-tap channel, and so on. Finally, we show that the extremal entropy inequality can be used for the proof of the secrecy capacity of the Gaussian wire-tap channel. This application supports the versatility of the extremal entropy inequality.

The rest of this section is organized as follows. The extremal entropy inequality without a covariance constraint and its alternative proof are shown in Section 4.2. The extremal entropy inequality and its alternative proof, which are the main results of this section, are provided in Section 4.3. In Section 4.4, an additional application of extremal entropy inequality is introduced, and the importance of our proof is explained. Finally, Section 4.5 concludes this section.

4.2 Entropy Power Inequality

Since the extremal entropy inequality is similar to the classical entropy power inequality (EPI), we first investigate a relationship between the extremal entropy

inequality and the EPI. Without a covariance constraint, the extremal entropy inequality is equivalent to the EPI as shown in Theorem 4.1.

Theorem 4.1. *For an arbitrary random vector* $\mathbf{X}$ *with a covariance matrix* $\mathbf{\Sigma}_X$ *and a Gaussian random vector* $\mathbf{W}_G$ *with a covariance matrix* $\mathbf{\Sigma}_W$*, there exists a Gaussian random vector* $\tilde{\mathbf{X}}_G$ *which satisfies the following inequality:*

$$h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{W}_G) \;\; \leq \;\; h(\tilde{\mathbf{X}}_G) - \mu h(\tilde{\mathbf{X}}_G + \mathbf{W}_G), \tag{4.1}$$

where the constant $\mu \geq 1$*, all random vectors are independent of each other,* $\mathbf{\Sigma}_W$ *is a positive definite matrix, and* $\tilde{\mathbf{X}}_G$ *is a Gaussian random vector which satisfies the following:*

1. *The covariance matrix of* $\tilde{\mathbf{X}}_G$ *is represented by* $\mathbf{\Sigma}_{\tilde{X}}$*, and it is proportional to* $\mathbf{\Sigma}_W$*.*

2. *The differential entropy of* $\tilde{\mathbf{X}}_G$*,* $h(\tilde{\mathbf{X}}_G)$*, is equal to the differential entropy of* X*,* $h(X)$*.*

In addition, the inequality (4.1) is equivalent to the EPI.

Proof.

Lemma 4.1 (Entropy Power Inequality [43], [9])**.** *For independent random vectors* $\mathbf{X}_1$ *and* $\mathbf{X}_2$*,*

$$h(\mathbf{X}_1 + \mathbf{X}_2) \;\; \geq \;\; h(\tilde{\mathbf{X}}_{G_1} + \tilde{\mathbf{X}}_{G_2}), \tag{4.2}$$

where $\tilde{\mathbf{X}}_{G_1}$ *and* $\tilde{\mathbf{X}}_{G_2}$ *are independent Gaussian random vectors,* $h(\tilde{\mathbf{X}}_{G_1}) = h(\mathbf{X}_1)$ *and* $h(\tilde{\mathbf{X}}_{G_2}) = h(\mathbf{X}_2)$*, and the covariance matrices of* $\tilde{\mathbf{X}}_{G_1}$ *and* $\tilde{\mathbf{X}}_{G_2}$ *are proportional.*

Using Lemma 4.1, the following relations are obtained:

$$\begin{aligned} h(\mathbf{X}) &= h(\tilde{\mathbf{X}}_G), \\ h(\mathbf{X}+\mathbf{W}_G) &\geq h(\tilde{\mathbf{X}}_G+\mathbf{W}_G), \end{aligned} \tag{4.3}$$

where $\mathbf{\Sigma}_{\tilde{X}}$ is proportional to $\mathbf{\Sigma}_W$, i.e., $\mathbf{\Sigma}_{\tilde{X}} = \alpha\mathbf{\Sigma}_W$, and α is an appropriate constant which satisfies $h(\mathbf{X}) = h(\tilde{\mathbf{X}}_G)$. Therefore, the inequality (4.1) is derived from Lemma 1, the EPI, and the proof of the inequality (4.1) is completed.

If the inequality (4.1) holds, $h(\mathbf{X}+\mathbf{W}_G) \geq h(\tilde{\mathbf{X}}_G+\mathbf{W}_G)$ since $h(\mathbf{X}) = h(\tilde{\mathbf{X}}_G)$, and $\mathbf{\Sigma}_{\tilde{X}}$ is proportional to $\mathbf{\Sigma}_W$. This is exactly the same as the EPI in Lemma 4.1. Therefore, the inequality (4.1) is equivalent to the EPI. □

While Theorem 4.1 shows a local upper bound, i.e., the upper bound is dependent on a random vector $\mathbf{X}$, since α depends on the random vector $\mathbf{X}$, we can also find a global upper bound as shown in Theorem 4.2 and the reference [32].

Theorem 4.2. *For an arbitrary random vector $\mathbf{X}$ with a covariance matrix $\mathbf{\Sigma}_X$ and a Gaussian random vector $\mathbf{W}_G$ with a covariance matrix $\mathbf{\Sigma}_W$, there exists a Gaussian random vector $\mathbf{X}^*_G$ which satisfies the following inequalities:*

$$h(\mathbf{X}) - \mu h(\mathbf{X}+\mathbf{W}_G) \leq h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G+\mathbf{W}_G), \tag{4.4}$$

$$h(\tilde{\mathbf{X}}_G) - \mu h(\tilde{\mathbf{X}}_G+\mathbf{W}_G) \leq h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G+\mathbf{W}_G), \tag{4.5}$$

where the constant $\mu > 1$, all random vectors are independent of each other, $\mathbf{\Sigma}_W$ is a positive definite matrix, $\tilde{\mathbf{X}}_G$ stands for the Gaussian random vector defined in Theorem 4.1, and $\mathbf{X}^*_G$ is a Gaussian random vector whose covariance matrix $\mathbf{\Sigma}_{X^*}$ is represented by $(\mu-1)^{-1}\mathbf{\Sigma}_W$.

Proof. The proof, here, is a little different from the proof in [32]. In our proof, we deal with both a local upper bound and a global upper bound while a global upper bound is directly calculated in [32].

Define the function $f(\alpha)$ as follows:

$$\begin{aligned} f(\alpha) &= h(\tilde{\mathbf{X}}_G) - \mu h(\tilde{\mathbf{X}}_G + \mathbf{W}_G) \\ &= \frac{n}{2}\log 2\pi e\,|\alpha\boldsymbol{\Sigma}_W|^{\frac{1}{n}} - \frac{\mu n}{2}\log 2\pi e\,|\alpha\boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_W|^{\frac{1}{n}}, \end{aligned} \tag{4.6}$$

where n denotes the dimension of a random vector, and $|\cdot|$ stands for the determinant of a matrix.

Since

$$\begin{aligned} \left.\frac{d}{d\alpha}f(\alpha)\right|_{\alpha=(\mu-1)^{-1}} &= \frac{n}{2(\mu-1)^{-1}} - \frac{\mu n}{2((\mu-1)^{-1}+1)} \\ &= 0, \\ \left.\frac{d^2}{d^2\alpha}f(\alpha)\right|_{\alpha=(\mu-1)^{-1}} &= -\frac{n}{2(\mu-1)^{-2}} + \frac{\mu n}{2((\mu-1)^{-1}+1)^2} \\ &< 0, \end{aligned} \tag{4.7}$$

$f(\alpha)$ is maximized when $\alpha = (\mu-1)^{-1}$.

Therefore, from Theorem 4.1, the following inequality is derived as

$$\begin{aligned} h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{W}_G) &\leq h(\tilde{\mathbf{X}}_G) - \mu h(\tilde{\mathbf{X}}_G + \mathbf{W}_G) \\ &= f(\alpha) \\ &\leq f((\mu-1)^{-1}) \\ &= h(\mathbf{X}_G^*) - \mu h(\mathbf{X}_G^* + \mathbf{W}_G). \end{aligned} \tag{4.8}$$

The inequalities (4.8) include inequalities (4.4) and (4.5), and the validity of

inequalities (4.4) and (4.5) is proved. The upper bound in (4.8) is a global maximum while the upper bound derived in Theorem 4.1 is a local maximum.

Remark 4.1. *When $\mu = 1$, the inequalities (4.4) and (4.5) are also satisfied. However, we cannot specify the covariance matrix of $\mathbf{X}_G^*$ since $h(\mathbf{X}_G^*) - \mu h(\mathbf{X}_G^* + \mathbf{W}_G)$ is increasing with respect to $\mathbf{\Sigma}_{X^*}$ and it can be infinitely large as $\mathbf{\Sigma}_{X^*}$ is increased. Therefore, we omit the case when $\mu = 1$ in Theorem 4.2.*

□

As shown in Theorem 4.1 and 4.2, for $\mu \geq 1$, $h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{W}_G)$ is maximized when random vector $\mathbf{X}$ is Gaussian. However, when a covariance constraint is added in the inequalities (4.1), (4.4) and (4.5), we cannot prove whether a Gaussian random vector still maximizes $h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{W}_G)$ or not, based on the same methods as described in the proofs of Theorems 4.1 and 4.2, since the covariance constraint may alter the proportionality relationship between the covariance matrices $\mathbf{\Sigma}_{X^*}$ and $\mathbf{\Sigma}_W$.

4.3 The Extremal Inequality

In [32], Liu and Viswanath proved that a Gaussian random vector still maximizes $h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{W}_G)$ even when a covariance constraint is considered. The inequality (4.4) was formulated as an optimization problem with a covariance constraint as follows:

$$\begin{aligned} \max_{p(\mathbf{X})} \quad & h(\mathbf{X} + \mathbf{W}_G) - \mu h(\mathbf{X} + \mathbf{V}_G), \\ \text{s.t.} \quad & \mathbf{\Sigma}_X \preceq \mathbf{R}, \end{aligned} \tag{4.9}$$

where $\mathbf{W}_G$ and $\mathbf{V}_G$ are independent Gaussian random vectors with positive definite covariance matrices $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_V$, respectively, all random vectors are independent of each other, and the maximization is done over the distribution of random vector

$\mathbf{X}$. Two proofs, a direct proof and a perturbation proof, are provided in [32]. Each proof approaches the problem in a different way but both proofs share an important common approach, namely the channel enhancement technique based on the KKT conditions and proposed originally in [53].

Unlike the original proofs in [32], we will prove Theorems 4.3 and 4.4 without using the channel enhancement technique. Before we deal with the problem (4.9), we first consider a simpler case of it in Theorem 4.3.

Theorem 4.3. *For an arbitrary random vector* $\mathbf{X}$ *with a covariance matrix* $\mathbf{\Sigma}_X$ *and a Gaussian random vector* $\mathbf{W}_G$ *with a covariance matrix* $\mathbf{\Sigma}_W$*, there exists a Gaussian random vector* $\mathbf{X}_G^*$ *with a covariance matrix* $\mathbf{\Sigma}_{X^*}$ *which satisfies the following inequality:*

$$h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{W}_G) \;\; \leq \;\; h(\mathbf{X}_G^*) - \mu h(\mathbf{X}_G^* + \mathbf{W}_G), \tag{4.10}$$

where the constant $\mu \geq 1$*, all random vectors are independent of each other,* $\mathbf{\Sigma}_W$ *is a positive definite matrix,* $\mathbf{\Sigma}_X \preceq \mathbf{R}$*,* $\mathbf{\Sigma}_{X^*} \preceq \mathbf{R}$*, and* $\mathbf{R}$ *is a positive semi-definite matrix.*

Proof. When $\mathbf{R}$ is a positive definite but singular matrix, i.e., $|\mathbf{R}| = 0$, the inequality (4.10) and its covariance constraints are equivalently changed into

$$h(\bar{\mathbf{X}}) - \mu h(\bar{\mathbf{X}} + \bar{\mathbf{W}}_G) \;\; \leq \;\; h(\bar{\mathbf{X}}_G^*) - \mu h(\bar{\mathbf{X}}_G^* + \bar{\mathbf{W}}_G), \tag{4.11}$$

where $\bar{\mathbf{X}}$ is such that $\mathbf{\Sigma}_{\bar{X}} \preceq \bar{\mathbf{R}}$, $\mathbf{\Sigma}_{\bar{X}^*} \preceq \bar{\mathbf{R}}$, and $\bar{\mathbf{R}}$ is a positive definite matrix, as mentioned in [32]. When $\mu = 1$, the inequality (4.10) is easily proved by the Lemma 4.2, which will be presented later.

Therefore, without loss of generality, we assume that $\mu > 1$ and $\mathbf{R}$ is a positive

definite matrix. Then, the right-hand side (RHS) of the equation (4.10) is upper-bounded by means of the following lemma.

Lemma 4.2 (Worst Additive Noise [43], [32], [11])**.** *For random vectors* $\mathbf{X}$*,* $\mathbf{X}_G$*,* $\tilde{\mathbf{W}}_G$*, and* $\mathbf{W}'_G$*,*

$$I(\mathbf{X}+\tilde{\mathbf{W}}_G+\mathbf{W}'_G;\mathbf{W}'_G) \;\geq\; I(\mathbf{X}_G+\tilde{\mathbf{W}}_G+\mathbf{W}'_G;\mathbf{W}'_G), \tag{4.12}$$

where $\mathbf{X}$ *is an arbitrary random vector,* $\mathbf{X}_G$ *is a Gaussian random vector with the covariance matrix identical to that of* $\mathbf{X}$*,* $\tilde{\mathbf{W}}_G$ *and* $\mathbf{W}'_G$ *are Gaussian random vectors, and all random vectors are independent.*

Based on Lemma 4.2, the following inequalities hold:

$$\begin{aligned}
&h(\mathbf{X}+\tilde{\mathbf{W}}_G+\mathbf{W}'_G)-h(\mathbf{X}+\tilde{\mathbf{W}}_G+\mathbf{W}'_G|\mathbf{W}'_G)\\
&\qquad\geq h(\mathbf{X}_G+\tilde{\mathbf{W}}_G+\mathbf{W}'_G)-h(\mathbf{X}_G+\tilde{\mathbf{W}}_G+\mathbf{W}'_G|\mathbf{W}'_G)\\
\iff &h(\mathbf{X}+\tilde{\mathbf{W}}_G+\mathbf{W}'_G)-h(\mathbf{X}+\tilde{\mathbf{W}}_G)\geq h(\mathbf{X}_G+\tilde{\mathbf{W}}_G+\mathbf{W}'_G)-h(\mathbf{X}_G+\tilde{\mathbf{W}}_G) \quad (4.13)\\
\iff &h(\mathbf{X}+\tilde{\mathbf{W}}_G+\mathbf{W}'_G)\geq h(\mathbf{X}+\tilde{\mathbf{W}}_G)+h(\mathbf{X}_G+\tilde{\mathbf{W}}_G+\mathbf{W}'_G)-h(\mathbf{X}_G+\tilde{\mathbf{W}}_G), \quad (4.14)
\end{aligned}$$

where $\iff$ denotes equivalence. Notice that the Gaussian random vector $\mathbf{W}_G$ can be expressed as the sum of two independent Gaussian random vectors $\tilde{\mathbf{W}}_G$ and $\mathbf{W}'_G$ whose covariance matrices satisfy:

$$\mathbf{\Sigma}_W \;=\; \mathbf{\Sigma}_{\tilde{W}}+\mathbf{\Sigma}_{W'}, \tag{4.15}$$

where $\mathbf{\Sigma}_W$, $\mathbf{\Sigma}_{\tilde{W}}$, and $\mathbf{\Sigma}_{W'}$ are the covariance matrices of $\mathbf{W}_G$, $\tilde{\mathbf{W}}_G$, and $\mathbf{W}'_G$, respectively. Henceforth, the Gaussian random vector $\mathbf{W}_G$ is represented as $\mathbf{W}_G = \tilde{\mathbf{W}}_G+\mathbf{W}'_G$.

Based on (4.14) and (4.15), the left-hand side (LHS) of the equation (4.10) is upper-bounded as

$$\begin{aligned} & h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{W}_G) \\ = \; & h(\mathbf{X}) - \mu h(\mathbf{X} + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) && (4.16) \\ \leq \; & h(\mathbf{X}) - \mu \left(h(\mathbf{X} + \tilde{\mathbf{W}}_G) + h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) \right) && (4.17) \\ = \; & h(\mathbf{X}) - \mu h(\mathbf{X} + \tilde{\mathbf{W}}_G) + \mu \left(h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \right). && (4.18) \end{aligned}$$

Using Theorem 4.2, if $(\mu - 1)^{-1}\boldsymbol{\Sigma}_{\tilde{\boldsymbol{W}}} \preceq \mathbf{R}$, the RHS of equation (4.18) is upper-bounded as

$$\begin{aligned} & h(\mathbf{X}) - \mu h(\mathbf{X} + \tilde{\mathbf{W}}_G) + \mu \left(h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \right) && (4.19) \\ \leq \; & h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) + \mu \left(h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \right), && (4.20) \end{aligned}$$

where $\mathbf{X}^*_G$ is a Gaussian random vector whose covariance matrix $\boldsymbol{\Sigma}_{X^*}$ is defined as $(\mu - 1)^{-1}\boldsymbol{\Sigma}_{\tilde{W}}$. Unlike Theorem 4.2, we additionally have to prove that there exists a random vector $\mathbf{X}^*_G$ whose covariance matrix $\boldsymbol{\Sigma}_{X^*}$ satisfies

$$\begin{aligned} \boldsymbol{\Sigma}_{X^*} &= (\mu - 1)^{-1}\boldsymbol{\Sigma}_{\tilde{W}} && (4.21) \\ &\preceq \mathbf{R}, && (4.22) \end{aligned}$$

due to the covariance constraint. Since $\boldsymbol{\Sigma}_X \preceq \mathbf{R}$, we will prove there exists a random vector X^*_G whose covariance matrix $\boldsymbol{\Sigma}_{X^*}$ satisfies

$$\begin{aligned} \boldsymbol{\Sigma}_{X^*} &= (\mu - 1)^{-1}\boldsymbol{\Sigma}_{\tilde{W}} && (4.23) \\ &\preceq \boldsymbol{\Sigma}_X, && (4.24) \end{aligned}$$

instead of proving (4.22).

The equation (4.20) is further proceeded by the following lemma.

Lemma 4.3 (Data Processing Inequality [9]). *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:*

$$I(\mathbf{Y}_1; \mathbf{Y}_3) \leq I(\mathbf{Y}_1; \mathbf{Y}_2). \tag{4.25}$$

The equality holds if and only if random vectors $\mathbf{Y}_1$, $\mathbf{Y}_2$, *and* $\mathbf{Y}_3$ *form the Markov chain:* $\mathbf{Y}_1 \rightarrow \mathbf{Y}_3 \rightarrow \mathbf{Y}_2$.

If the inequality (4.24) is satisfied, then we can form a Markov chain such as

$$\mathbf{X}'_G \rightarrow \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G \rightarrow \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G, \tag{4.26}$$

where all random vectors are independent. Since a Gaussian random vector $\mathbf{X}_G$ can be expressed as the summation of two independent Gaussian random vectors $\mathbf{X}'_G$ and $\mathbf{X}^*_G$ whose covariance matrices satisfy

$$\mathbf{\Sigma}_X \;\; = \;\; \mathbf{\Sigma}_{X'} + \mathbf{\Sigma}_{X^*}, \tag{4.27}$$

where $\mathbf{\Sigma}_X$, $\mathbf{\Sigma}_{X'}$, and $\mathbf{\Sigma}_{X^*}$ stand for covariance matrices of $\mathbf{X}_G$, $\mathbf{X}'_G$, and $\mathbf{X}^*_G$, respectively, the Gaussian random vector $\mathbf{X}_G$ will be represented as $\mathbf{X}_G = \mathbf{X}'_G + \mathbf{X}^*_G$.

Based on Lemma 4.3, we obtain

$$I(\mathbf{X}'_G; \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \leq I(\mathbf{X}'_G; \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G) \tag{4.28}$$

$$\iff h(\mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G)$$
$$\leq h(\mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) \tag{4.29}$$

$$\iff h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G)$$
$$\leq h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) \tag{4.30}$$

$$\iff h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) + h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G)$$
$$\leq h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G). \tag{4.31}$$

The equivalence in (4.30) is due to $\mathbf{X}_G = \mathbf{X}'_G + \mathbf{X}^*_G$.

Even though we need an upper bound of the RHS term in equation (4.20), the equation (4.31) generates a lower bound of the equation (4.20) as follows:

$$h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) + \mu \left(h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \right) \tag{4.32}$$

$$\geq h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \tag{4.33}$$

$$\geq h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G + \mathbf{W}_G). \tag{4.34}$$

However, if we can construct the following Markov chain:

$$\mathbf{X}'_G \to \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G \to \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G, \tag{4.35}$$

and using Lemma 4.3 again, it turns out that

$$I(\mathbf{X}'_G; \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \geq I(\mathbf{X}'_G; \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G), \tag{4.36}$$

and this inequality leads us to a tight upper bound. Indeed,

$$
\begin{aligned}
& I(\mathbf{X}'_G; \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \geq I(\mathbf{X}'_G; \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G) && (4.37)\\
\iff\ & h(\mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \\
& \qquad \geq h(\mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) && (4.38)\\
\iff\ & h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \\
& \qquad \geq h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) && (4.39)\\
\iff\ & h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) + h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \\
& \qquad \geq h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G). && (4.40)
\end{aligned}
$$

The equivalence in (4.39) is due to $\mathbf{X}_G = \mathbf{X}'_G + \mathbf{X}^*_G$.

Now using (4.40), the equations (4.19) and (4.20) are upper-bounded as follows:

$$
\begin{aligned}
& h(\mathbf{X}) - \mu h(\mathbf{X} + \tilde{\mathbf{W}}_G) + \mu \left(h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \right) && (4.41)\\
\leq\ & h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) + \mu \left(h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \right) && (4.42)\\
\leq\ & h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) && (4.43)\\
=\ & h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G + \mathbf{W}_G), && (4.44)
\end{aligned}
$$

and this is exactly the same as the equation (4.34). Therefore, the following equality is satisfied:

$$
\begin{aligned}
& h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) + \mu \left(h(\mathbf{X}_G + \tilde{\mathbf{W}}_G) - h(\mathbf{X}_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G) \right) && (4.45)\\
=\ & h(\mathbf{X}^*_G) - \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G), && (4.46)
\end{aligned}
$$

due to (4.34) and (4.44). Now, we will prove that we can actually construct the

Markov chain (4.35) using the following lemmas.

Lemma 4.4. *For independent random vectors* $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$*, the following equality between moment generating functions (MGFs) is satisfied:*

$$M_{Y_1+Y_2}(\mathbf{S}) \;=\; M_{Y_1}(\mathbf{S})M_{Y_2}(\mathbf{S}), \tag{4.47}$$

where $M_Y(\mathbf{S}) = \mathbb{E}[e^{\mathbf{Y}^T\mathbf{S}}]$*,* $\mathbb{E}[\cdot]$ *is an expectation, and superscript* T *denotes the transpose of a vector. For jointly Gaussian random vectors* $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$*, this equality is a necessary and sufficient condition for the independence between* $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$*.*

Lemma 4.5. *For independent random vectors* $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$ *given a random vector* $\mathbf{Y}_3$*, the following equality is satisfied:*

$$M_{Y_1+Y_2|Y_3}(\mathbf{S}) \;=\; M_{Y_1|Y_3}(\mathbf{S})M_{Y_2|Y_3}(\mathbf{S}). \tag{4.48}$$

Lemma 4.6. *For a Gaussian random vector* $\mathbf{X}$ *with a mean* $\mathbf{U}_X$ *and a covariance matrix* $\mathbf{\Sigma}_X$*, the MGF is expressed as*

$$M_X(\mathbf{S}) \;=\; \exp\left\{\mathbf{S}^T\mathbf{U}_X + \frac{1}{2}\mathbf{S}^T\mathbf{\Sigma}_X\mathbf{S}\right\}. \tag{4.49}$$

In the Markov chain (4.35), since all random vectors are Gaussian (without loss of generality, they are assumed to have zero means), using Lemma 4.6, the following moment generating functions are presented in closed-form expression:

$$\begin{aligned} M_{Y_1|Y_3}(\mathbf{S}) &= \exp\left\{\mathbf{S}^T\mathbf{\Sigma}_{Y_1}\mathbf{\Sigma}_{Y_3}^{-1}Y_3 + \frac{1}{2}\mathbf{S}^T\left(\mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_1}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_1}\right)\mathbf{S}\right\}, \\ M_{Y_2|Y_3}(\mathbf{S}) &= \exp\left\{\mathbf{S}^T\mathbf{\Sigma}_{Y_2}\mathbf{\Sigma}_{Y_3}^{-1}Y_3 + \frac{1}{2}\mathbf{S}^T\left(\mathbf{\Sigma}_{Y_2} - \mathbf{\Sigma}_{Y_2}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_2}\right)\mathbf{S}\right\}, \end{aligned} \tag{4.50}$$

where $\mathbf{Y}_1 = \mathbf{X}'_G$, $\mathbf{Y}_2 = \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G$, $\mathbf{Y}_3 = \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G$, and their covariance matrices are represented by $\mathbf{\Sigma}_{Y_1}$, $\mathbf{\Sigma}_{Y_2}$, and $\mathbf{\Sigma}_{Y_3}$, respectively. Since $\mathbf{\Sigma}_{\tilde{W}} + \mathbf{\Sigma}_{W'}$ is a positive definite matrix, there exists the inverse of $\mathbf{\Sigma}_{Y_3}$.

On the other hand, the MGF of $\mathbf{Y}_1 + \mathbf{Y}_2$ given $\mathbf{Y}_3$ is represented as

$$
\begin{aligned}
&M_{Y_1+Y_2|Y_3}(\mathbf{S}) \\
=&\exp\left\{\mathbf{S}^T\left(\mathbf{\Sigma}_{Y_1}+\mathbf{\Sigma}_{Y_2}\right)\mathbf{\Sigma}_{Y_3}^{-1}Y_3 + \frac{1}{2}\mathbf{S}^T\left(\mathbf{\Sigma}_{Y_1}-\mathbf{\Sigma}_{Y_1}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_1}+\mathbf{\Sigma}_{Y_2}-\mathbf{\Sigma}_{Y_2}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_2}\right)\mathbf{S}\right\} \\
&\times\exp\left\{\mathbf{S}^T\left(\mathbf{\Sigma}_{Y_1}-\mathbf{\Sigma}_{Y_2}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_1}+\mathbf{\Sigma}_{Y_1}-\mathbf{\Sigma}_{Y_1}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_2}\right)\mathbf{S}\right\} \\
=&M_{Y_1|Y_3}(\mathbf{S})M_{Y_2|Y_3}(\mathbf{S})\exp\underbrace{\left\{\mathbf{S}^T\left(\mathbf{\Sigma}_{Y_1}-\mathbf{\Sigma}_{Y_2}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_1}+\mathbf{\Sigma}_{Y_1}-\mathbf{\Sigma}_{Y_1}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_2}\right)\mathbf{S}\right\}}_{(A)}. \quad (4.51)
\end{aligned}
$$

If (A) in the equation (4.51) is vanished, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent given $\mathbf{Y}_3$, and the Markov chain (4.35) is obtained. Using Lemma 11, (1) in [53], we define the covariance matrix $\mathbf{\Sigma}_{\tilde{W}}$ as

$$
\mathbf{\Sigma}_{\tilde{W}} = \left(\left(\mathbf{\Sigma}_X+\mathbf{\Sigma}_W\right)^{-1}+\mathbf{L}\right)^{-1}-\mathbf{\Sigma}_X, \quad (4.52)
$$

where $\mathbf{L} \succeq \mathbf{0}$, and $\mathbf{0}$ denotes an n-by-n zero matrix. The positive semi-definite matrix $\mathbf{L}$ must be chosen to satisfy

$$
\mathbf{\Sigma}_{X^*} \preceq \mathbf{\Sigma}_X, \quad (4.53)
$$

$$
\mathbf{L}\mathbf{\Sigma}_{X'} = \mathbf{\Sigma}_{X'}\mathbf{L} = \mathbf{0}, \quad (4.54)
$$

where $\mathbf{\Sigma}_{X^*} = (\mu-1)^{-1}\mathbf{\Sigma}_{\tilde{W}}$, $\mathbf{\Sigma}_{X'} = \mathbf{\Sigma}_X - \mathbf{\Sigma}_{X^*}$, $\mathbf{L} \succeq \mathbf{0}$. Lemma 4.7 will prove that such a positive semi-definite matrix $\mathbf{L}$ exists.

Lemma 4.7. *There exists a positive semi-definite matrix* $\mathbf{L}$ *which satisfies*

$$\boldsymbol{\Sigma}_{X^*} \preceq \boldsymbol{\Sigma}_X, \quad \mathbf{L}\boldsymbol{\Sigma}_{X'} = \mathbf{0}, \tag{4.55}$$

where $\boldsymbol{\Sigma}_{\tilde{W}} = \left((\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_W)^{-1} + \mathbf{L}\right)^{-1} - \boldsymbol{\Sigma}_X$, $\boldsymbol{\Sigma}_{X^*} = (\mu - 1)^{-1}\boldsymbol{\Sigma}_{\tilde{W}}$, $\boldsymbol{\Sigma}_{X'} = \boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_{X^*}$, *and* $\boldsymbol{\Sigma}_X$ *and* $\boldsymbol{\Sigma}_W$ *stand for a positive semi-definite matrix and a positive definite matrix, respectively.*

Proof. See Appendix B.1. □

The equation (4.52) can be re-written as

$$\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_{\tilde{W}} = \left((\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_W)^{-1} + \mathbf{L}\right)^{-1} \tag{4.56}$$

$$\iff \quad (\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_{\tilde{W}})^{-1} = (\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_W)^{-1} + \mathbf{L}. \tag{4.57}$$

Since $\mathbf{L}\boldsymbol{\Sigma}_{X'} = \boldsymbol{\Sigma}_{X'}\mathbf{L} = \mathbf{0}$, by multiplying $\boldsymbol{\Sigma}_{X'}$ to both sides of the equation (4.57),

$$\begin{aligned} (\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_{\tilde{W}})^{-1} \boldsymbol{\Sigma}_{X'} &= (\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_W)^{-1} \boldsymbol{\Sigma}_{X'} + \mathbf{L}\boldsymbol{\Sigma}_{X'} \\ &= (\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_W)^{-1} \boldsymbol{\Sigma}_{X'}, \end{aligned} \tag{4.58}$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_{X'} (\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_{\tilde{W}})^{-1} &= \boldsymbol{\Sigma}_{X'} (\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_W)^{-1} + \boldsymbol{\Sigma}_{X'}\mathbf{L} \\ &= \boldsymbol{\Sigma}_{X'} (\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_W)^{-1}. \end{aligned} \tag{4.59}$$

Since random vectors $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ are defined as $\mathbf{Y}_1 = \mathbf{X}'_G$, $\mathbf{Y}_2 = \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G$, and $\mathbf{Y}_3 = \mathbf{X}'_G + \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G$, respectively, and they are independent of each other, their covariance matrices are represented as

$$
\begin{aligned}
\boldsymbol{\Sigma}_{Y_1} &= \boldsymbol{\Sigma}_{X'}, \\
\boldsymbol{\Sigma}_{Y_2} &= \boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}, \\
&= \boldsymbol{\Sigma}_{X} + \boldsymbol{\Sigma}_{\tilde{W}}, \\
\boldsymbol{\Sigma}_{Y_3} &= \boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}} + \boldsymbol{\Sigma}_{W'} \\
&= \boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{W} \\
&= \boldsymbol{\Sigma}_{X} + \boldsymbol{\Sigma}_{W}. \qquad (4.60)
\end{aligned}
$$

From the equations (4.58) and (4.60),

$$
\begin{aligned}
& \boldsymbol{\Sigma}_{Y_1} - \boldsymbol{\Sigma}_{Y_2}\boldsymbol{\Sigma}_{Y_3}^{-1}\boldsymbol{\Sigma}_{Y_1} \\
= \; & \boldsymbol{\Sigma}_{X'} - \left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}\right)\left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{W}\right)^{-1}\boldsymbol{\Sigma}_{X'} \\
= \; & \left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}\right)\left(\left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}\right)^{-1}\boldsymbol{\Sigma}_{X'} - \left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{W}\right)^{-1}\boldsymbol{\Sigma}_{X'}\right) \\
= \; & \mathbf{0}, \qquad (4.61)
\end{aligned}
$$

and from the equations (4.59) and (4.60),

$$
\begin{aligned}
& \boldsymbol{\Sigma}_{Y_1} - \boldsymbol{\Sigma}_{Y_1}\boldsymbol{\Sigma}_{Y_3}^{-1}\boldsymbol{\Sigma}_{Y_2} \\
= \; & \boldsymbol{\Sigma}_{X'} - \boldsymbol{\Sigma}_{X'}\left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{W}\right)^{-1}\left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}\right) \\
= \; & \left(\boldsymbol{\Sigma}_{X'}\left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}\right)^{-1} - \boldsymbol{\Sigma}_{X'}\left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{W}\right)^{-1}\right)\left(\boldsymbol{\Sigma}_{X'} + \boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}\right) \\
= \; & \mathbf{0}. \qquad (4.62)
\end{aligned}
$$

□

The more general problem, originally proved in [32], is now considered in Theorem

4.4.

Theorem 4.4. *For an arbitrary random vector* $\mathbf{X}$ *with a covariance matrix* $\mathbf{\Sigma}_X$ *and two independent random vectors* $\mathbf{W}_G$ *and* $\mathbf{V}_G$ *with covariance matrices* $\mathbf{\Sigma}_W$ *and* $\mathbf{\Sigma}_V$*, respectively, there exists a Gaussian random vector* $\mathbf{X}_G^*$ *with a covariance matrix* $\mathbf{\Sigma}_{X^*}$ *which satisfies the following inequality:*

$$h(\mathbf{X}+\mathbf{W}_G)-\mu h(\mathbf{X}+\mathbf{V}_G)\leq h(\mathbf{X}_G^*+\mathbf{W}_G)-\mu h(\mathbf{X}_G^*+\mathbf{V}_G), \tag{4.63}$$

where the constant $\mu\geq 1$*, all random vectors are independent of each other,* $\mathbf{\Sigma}_W$ *is a positive definite matrix,* $\mathbf{\Sigma}_X\preceq\mathbf{R}$*,* $\mathbf{\Sigma}_{X^*}\preceq\mathbf{R}$*, and* $\mathbf{R}$ *is a positive semi-definite matrix.*

Proof. Due to the same reason mentioned in the proof of Theorem 4.3, without loss of generality, we assume $\mu>1$ and $\mathbf{R}$ is a positive definite matrix. The proof is generally similar to the proof of Theorem 4.3. Using Lemma 4.3, the inequality (4.63) can be expressed as

$$\begin{aligned}
& h(\mathbf{X}+\mathbf{W}_G)-\mu h(\mathbf{X}+\mathbf{V}_G) \\
\leq\ & h(\mathbf{X}+\tilde{\mathbf{W}}_G)-\mu h(\mathbf{X}+\mathbf{V}_G)+h(\mathbf{W}_G)-h(\tilde{\mathbf{W}}_G) && (4.64)\\
\leq\ & h(\mathbf{X}_G^*+\tilde{\mathbf{W}}_G)-\mu h(\mathbf{X}_G^*+\mathbf{V}_G)+h(\mathbf{W}_G)-h(\tilde{\mathbf{W}}_G) && (4.65)\\
=\ & h(\mathbf{X}_G^*+\mathbf{W}_G)-\mu h(\mathbf{X}_G^*+\mathbf{V}_G), && (4.66)
\end{aligned}$$

where $\tilde{\mathbf{W}}_G$ is chosen to be a Gaussian random vector whose covariance matrix, $\mathbf{\Sigma}_{\tilde{W}}$, satisfies

$$\begin{aligned}
\mathbf{\Sigma}_{\tilde{W}} &\preceq \mathbf{\Sigma}_W, && (4.67)\\
\mathbf{\Sigma}_{\tilde{W}} &\preceq \mu^{-1}\mathbf{\Sigma}_V. && (4.68)
\end{aligned}$$

The inequality in (4.64) is due to Lemma 4.3, the inequality (4.65) is due to Theorem 4.3, and the equality (4.66) will be proved using the equality condition in Lemma 4.3. We will also prove that there exists a Gaussian random vector $\tilde{\mathbf{W}}_G$ which satisfies the equations (4.67) and (4.68) by proving later Lemma 4.8.

To satisfy the equality in the equation (4.66), the equality condition in Lemma 4.3 must be satisfied, and the following two Markov chains are formed:

1.

$$\mathbf{X}_G^* \rightarrow \mathbf{X}_G^* + \tilde{\mathbf{W}}_G \rightarrow \mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \mathbf{W}'_G, \tag{4.69}$$

2.

$$\mathbf{X}_G^* \rightarrow \mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \mathbf{W}'_G \rightarrow \mathbf{X}_G^* + \tilde{\mathbf{W}}_G, \tag{4.70}$$

where all random vectors are normally distributed, $\tilde{\mathbf{W}}_G$ and $\mathbf{W}'_G$ are independent of each other, $\mathbf{W}_G = \tilde{\mathbf{W}}_G + \mathbf{W}'_G$, and $\mathbf{X}_G^*$ is independent of other random vectors.

The Markov chain (4.69) is naturally formed since $\mathbf{X}_G^*$, $\tilde{\mathbf{W}}_G$, and $\mathbf{W}'_G$ are independent Gaussian random vectors. The validity of the Markov chain (4.70) is proved using the concept of moment generating function. In the Markov chain (4.70), since all random vectors are Gaussian (without loss of generality, they are assumed to have zero means), using Lemma 4.6, the following moment generating functions are expressed in closed-form:

$$\begin{aligned} M_{Y_1|Y_3}(\mathbf{S}) &= \exp\left\{\mathbf{S}^T\boldsymbol{\Sigma}_{Y_1}\boldsymbol{\Sigma}_{Y_3}^{-1}\mathbf{Y}_3 + \frac{1}{2}\mathbf{S}^T\left(\boldsymbol{\Sigma}_{Y_1} - \boldsymbol{\Sigma}_{Y_1}\boldsymbol{\Sigma}_{Y_3}^{-1}\boldsymbol{\Sigma}_{Y_1}\right)\mathbf{S}\right\}, \\ M_{Y_2|Y_3}(\mathbf{S}) &= \exp\left\{\mathbf{S}^T\boldsymbol{\Sigma}_{Y_2}\boldsymbol{\Sigma}_{Y_3}^{-1}Y_3 + \frac{1}{2}\mathbf{S}^T\left(\boldsymbol{\Sigma}_{Y_2} - \boldsymbol{\Sigma}_{Y_2}\boldsymbol{\Sigma}_{Y_3}^{-1}\boldsymbol{\Sigma}_{Y_2}\right)\mathbf{S}\right\}, \end{aligned} \tag{4.71}$$

where $\mathbf{Y}_1 = \mathbf{X}^*_G$, $\mathbf{Y}_2 = \mathbf{X}^*_G + \tilde{\mathbf{W}}_G$, $\mathbf{Y}_3 = \mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \mathbf{W}'_G$, and their covariance matrices are represented by $\mathbf{\Sigma}_{Y_1}$, $\mathbf{\Sigma}_{Y_2}$, and $\mathbf{\Sigma}_{Y_3}$, respectively. Since $\mathbf{\Sigma}_W$ is a positive definite matrix, there always exists the inverse of $\mathbf{\Sigma}_{Y_3}$.

On the other hand, the MGF of $\mathbf{Y}_1 + \mathbf{Y}_2$ given $\mathbf{Y}_3$ is represented as

$$\begin{aligned}
& M_{Y_1+Y_2|Y_3}(\mathbf{S}) \\
= \; & \exp\left\{ \mathbf{S}^T \left(\mathbf{\Sigma}_{Y_1} + \mathbf{\Sigma}_{Y_2}\right) \mathbf{\Sigma}_{Y_3}^{-1} \mathbf{Y}_3 \right. \\
& \qquad \left. + \frac{1}{2} \mathbf{S}^T \left(\mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_1} \mathbf{\Sigma}_{Y_3}^{-1} \mathbf{\Sigma}_{Y_1} + \mathbf{\Sigma}_{Y_2} - \mathbf{\Sigma}_{Y_2} \mathbf{\Sigma}_{Y_3}^{-1} \mathbf{\Sigma}_{Y_2} \right) \mathbf{S} \right\} \\
& \times \exp\left\{ \mathbf{S}^T \left(\mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_2} \mathbf{\Sigma}_{Y_3}^{-1} \mathbf{\Sigma}_{Y_1} + \mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_1} \mathbf{\Sigma}_{Y_3}^{-1} \mathbf{\Sigma}_{Y_2} \right) \mathbf{S} \right\} \\
= \; & M_{Y_1|Y_3}(\mathbf{S}) M_{Y_2|Y_3}(\mathbf{S}) \\
& \times \exp \underbrace{\left\{ \mathbf{S}^T \left(\mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_2} \mathbf{\Sigma}_{Y_3}^{-1} \mathbf{\Sigma}_{Y_1} + \mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_1} \mathbf{\Sigma}_{Y_3}^{-1} \mathbf{\Sigma}_{Y_2} \right) \mathbf{S} \right\}}_{(B)}. \qquad (4.72)
\end{aligned}$$

If (B) in the equation (4.72) is vanished, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent given $\mathbf{Y}_3$, and the Markov chain (4.70) is obtained. Using Lemma 11, (1) in [53], we define a covariance matrix $\mathbf{\Sigma}_{\tilde{W}}$ as follows:

$$\mathbf{\Sigma}_{\tilde{W}} = \left(\mathbf{\Sigma}_W^{-1} + \mathbf{K} \right)^{-1}, \qquad (4.73)$$

where $\mathbf{K} \succeq \mathbf{0}$, $\mathbf{K}\mathbf{\Sigma}_{X^*} = \mathbf{\Sigma}_{X^*}\mathbf{K} = \mathbf{0}$, and $\mathbf{0}$ denotes an n-by-n zero matrix. Then, there exists a positive semi-definite matrix $\mathbf{K}$ which satisfies

$$\mathbf{\Sigma}_{\tilde{W}} \preceq \mu^{-1} \mathbf{\Sigma}_V, \qquad (4.74)$$

$$\mathbf{K}\mathbf{\Sigma}_{X^*} = \mathbf{0}, \qquad (4.75)$$

where $\mathbf{\Sigma}_{X^*} = (\mu - 1)^{-1}(\mathbf{\Sigma}_V - \mu \mathbf{\Sigma}_{\tilde{W}})$. The existence of matrix $\mathbf{K}$ is proved by the

following lemma.

Lemma 4.8. *There always exists a positive semi-definite matrix* $\mathbf{K}$ *which satisfies*

$$\begin{aligned} \boldsymbol{\Sigma}_{\tilde{W}} &\preceq \mu^{-1}\boldsymbol{\Sigma}_V, & (4.76)\\ \mathbf{K}\boldsymbol{\Sigma}_{X^*} &= \boldsymbol{\Sigma}_{X^*}\mathbf{K} = \mathbf{0}, & (4.77) \end{aligned}$$

where $\boldsymbol{\Sigma}_{X^*} = (\mu-1)^{-1}(\boldsymbol{\Sigma}_V - \mu\boldsymbol{\Sigma}_{\tilde{W}})$, *and* $\boldsymbol{\Sigma}_{\tilde{W}} = \left(\boldsymbol{\Sigma}_W^{-1} + \mathbf{K}\right)^{-1}$.

Since $\boldsymbol{\Sigma}_{\tilde{W}}$ is defined as $\left(\boldsymbol{\Sigma}_W^{-1} + \mathbf{K}\right)^{-1}$ in (4.73), $\boldsymbol{\Sigma}_{\tilde{W}}$ satisfies

$$\left(\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}\right)^{-1} = \left(\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_W\right)^{-1} + \mathbf{K}, \quad (4.78)$$

based on Lemma 11, (1) in [53].

Since $\mathbf{K}\boldsymbol{\Sigma}_{X^*} = \boldsymbol{\Sigma}_{X^*}\mathbf{K} = \mathbf{0}$, multiplying $\boldsymbol{\Sigma}_{X^*}$ to both sides of the equation (4.78), the equation (4.78) is expressed as

$$\begin{aligned} \left(\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}\right)^{-1}\boldsymbol{\Sigma}_{X^*} &= \left(\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_W\right)^{-1}\boldsymbol{\Sigma}_{X^*} + \mathbf{K}\boldsymbol{\Sigma}_{X^*} \\ &= \left(\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_W\right)^{-1}\boldsymbol{\Sigma}_{X^*}, & (4.79) \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_{X^*}\left(\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_{\tilde{W}}\right)^{-1} &= \boldsymbol{\Sigma}_{X^*}\left(\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_W\right)^{-1} + \boldsymbol{\Sigma}_{X^*}\mathbf{K} \\ &= \boldsymbol{\Sigma}_{X^*}\left(\boldsymbol{\Sigma}_{X^*} + \boldsymbol{\Sigma}_W\right)^{-1}. & (4.80) \end{aligned}$$

Random vectors $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ are defined as $\mathbf{Y}_1 = \mathbf{X}_G^*$, $\mathbf{Y}_2 = \mathbf{X}_G^* + \tilde{\mathbf{W}}_G$, and $\mathbf{Y}_3 = \mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \mathbf{W}_G'$, respectively, and $\mathbf{X}_G^*$, $\tilde{\mathbf{W}}_G$, and $\mathbf{W}_G'$ are independent of each

other. Therefore, their covariance matrices are represented as

$$\begin{aligned}
\mathbf{\Sigma}_{Y_1} &= \mathbf{\Sigma}_{X^*}, \\
\mathbf{\Sigma}_{Y_2} &= \mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}, \\
\mathbf{\Sigma}_{Y_3} &= \mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}} + \mathbf{\Sigma}_{W'} \\
&= \mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{W}. \qquad (4.81)
\end{aligned}$$

From the equations (4.79) and (4.81),

$$\begin{aligned}
\mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_2}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_1} &= \mathbf{\Sigma}_{X^*} - \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{W}\right)^{-1}\mathbf{\Sigma}_{X^*} \\
&= \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)\left(\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)^{-1}\mathbf{\Sigma}_{X^*} - \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{W}\right)^{-1}\mathbf{\Sigma}_{X^*}\right) \\
&= \mathbf{0}, \qquad (4.82)
\end{aligned}$$

and from the equations (4.80) and (4.81),

$$\begin{aligned}
\mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_1}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_2} &= \mathbf{\Sigma}_{X^*} - \mathbf{\Sigma}_{X^*}\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{W}\right)^{-1}\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right) \\
&= \left(\mathbf{\Sigma}_{X^*}\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)^{-1} - \mathbf{\Sigma}_{X^*}\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{W}\right)^{-1}\right)\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right) \\
&= \mathbf{0}. \qquad (4.83)
\end{aligned}$$

Since the inverse matrix of $\mathbf{\Sigma}_{\tilde{W}}$ exists, $(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}})^{-1}$ also exists.

Therefore, (B) in the equation (4.72) is zero, and $M_{Y_1+Y_2|Y_3}(\mathbf{S}) = M_{Y_1|Y_3}(\mathbf{S})M_{Y_2|Y_3}(\mathbf{S})$. It means $\mathbf{Y}_1$ and $\mathbf{Y}_3$ are independent given $\mathbf{Y}_2$, i.e., $\mathbf{X}_G^*$ and $\mathbf{X}_G^* + \tilde{\mathbf{W}}_G$ are independent given $\mathbf{X}_G^* + \tilde{\mathbf{W}}_G + \mathbf{W}'_G$, and the Markov chain (4.70) is valid. The equality in the equation (4.66) is achieved by the above procedure, and the proof is completed. □

4.4 Applications

Since the versatility of the extremal entropy inequality was already proved by several applications in [32]. However, the original proofs of the extremal entropy inequality [32] were based on the channel enhancement technique while one of those applications, the capacity of the vector Gaussian broadcast channel, had been already proved by the channel enhancement technique in [53]. Even though the extremal entropy inequality was adapted to prove the capacity of the vector Gaussian broadcast channel in [32], it failed to show a novel perspective since the proof of the extremal entropy inequality was based on the channel enhancement technique, which was already used in [53]. On the other hand, based on our proof, the extremal entropy inequality shows not only its usefulness but also a novel perspective to prove the capacity of the vector Gaussian broadcast channel.

In this section, we propose an additional application for the extremal entropy inequality: the secrecy capacity of the Gaussian wire-tap channel, which was derived by several authors [31], [14], [28], [37], [29]. By adopting the proposed proof of the extremal entropy inequality, we will show a novel simplified proof of the secrecy capacity of the Gaussian wire-tap channel.

We consider the channel defined as

$$\begin{aligned} \mathbf{Y}_R[t] &= \mathbf{X}[t] + \mathbf{Z}_R[t], \\ \mathbf{Y}_E[t] &= \mathbf{X}[t] + \mathbf{Z}_E[t], \end{aligned} \tag{4.84}$$

where $\mathbf{Z}_R[t]$ and $\mathbf{Z}_E[t]$ are additive Gaussian noise vectors with zero means and covariance matrices $\mathbf{\Sigma}_{\mathbf{Z}_R}$ and $\mathbf{\Sigma}_{\mathbf{Z}_E}$, respectively. The covariance matrices, $\mathbf{\Sigma}_{Z_R}$ and $\mathbf{\Sigma}_{Z_E}$, are assumed to be positive definite, and random vectors, $\mathbf{X}[t]$, $\mathbf{Z}_R[t]$, and $\mathbf{Z}_E[t]$,

are independent of each other.

First, we consider a degraded case, i.e., the covariance matrices $\mathbf{\Sigma}_{Z_R}$ and $\mathbf{\Sigma}_{Z_E}$ present the following partial ordering: $\mathbf{\Sigma}_{Z_R} \preceq \mathbf{\Sigma}_{Z_E}$. According to [54], the secrecy capacity of the degraded case, C_D, is expressed as

$$C_D = \max_{\mathbf{\Sigma}_X \preceq \mathbf{R}} \left\{ I(\mathbf{X}; \mathbf{Y}_R) - I(\mathbf{X}; \mathbf{Y}_E) \right\}, \tag{4.85}$$

The difference between the two mutual information is upper bounded as follows:

$$\begin{aligned} & I(\mathbf{X}; \mathbf{Y}_R) - I(\mathbf{X}; \mathbf{Y}_E) \\ = & \underbrace{h(\mathbf{X} + \mathbf{Z}_R) - h(\mathbf{X} + \mathbf{Z}_E)}_{(C_1)} - h(\mathbf{Z}_R) + h(\mathbf{Z}_E) \\ \leq & \underbrace{h(\mathbf{X}_G^* + \mathbf{Z}_R) - h(\mathbf{X}_G^* + \mathbf{Z}_E)}_{(C_2)} - h(\mathbf{Z}_R) + h(\mathbf{Z}_E) \qquad (4.86) \\ \leq & h(\mathbf{X}_G^{(\mathbf{R})} + \mathbf{Z}_R) - h(\mathbf{X}_G^{(\mathbf{R})} + \mathbf{Z}_E) - h(\mathbf{Z}_R) + h(\mathbf{Z}_E), \qquad (4.87) \end{aligned}$$

where $\mathbf{X}_G^*$ is a Gaussian random vector with a covariance matrix $\mathbf{\Sigma}_{X^*}$ which is obtained in Theorem 4.4, and $\mathbf{X}_G^{(\mathbf{R})}$ denotes a Gaussian random vector with zero mean and covariance matrix $\mathbf{R}$.

Since the inequality between (C_1) and (C_2) is a special case of Theorem 4.4 when $\mu = 1$ and $\mathbf{\Sigma}_{Z_R} \preceq \mathbf{\Sigma}_{Z_E}$, the inequality (4.86) is satisfied. The inequality (4.87) also holds because the right-hand side of the inequality (4.87) is an increasing function with respect to a covariance matrix of $\mathbf{X}_G$. Therefore, the secrecy capacity of a degraded vector Gaussian wire-tap channel is expressed as

$$\begin{aligned} C_D &= h(\mathbf{X}_G^{(\mathbf{R})} + \mathbf{Z}_R) - h(\mathbf{X}_G^{(\mathbf{R})} + \mathbf{Z}_E) - h(\mathbf{Z}_R) + h(\mathbf{Z}_E) \\ &= \frac{1}{2} \log \frac{|\mathbf{R} + \mathbf{\Sigma}_{Z_R}|}{|\mathbf{\Sigma}_{Z_R}|} - \frac{1}{2} \log \frac{|\mathbf{R} + \mathbf{\Sigma}_{Z_E}|}{|\mathbf{\Sigma}_{Z_E}|}. \qquad (4.88) \end{aligned}$$

In a general case, i.e., not necessarily a degraded case, the secrecy capacity is more difficult to be calculated since the secrecy capacity cannot be expressed as in the equation (4.85). However, as shown in [31], since the secrecy capacity of a general wire-tap channel can be upper-bounded by the secrecy capacity of a degraded wire-tap channel, the secrecy capacity of a general wire-tap channel is calculated based on the channel enhancement technique. In this section, we will also show that the secrecy capacity of a general wire-tap channel can be upper-bounded by the secrecy capacity of a degraded wire-tap channel by using the following procedure, which is completely different from that of reference [31].

Using Theorem 4.4,

$$\begin{aligned} I(\mathbf{X};\mathbf{X}+\mathbf{Z}_R) - I(\mathbf{X};\mathbf{X}+\mathbf{Z}_E) &= h(\mathbf{X}+\mathbf{Z}_R) - h(\mathbf{X}+\mathbf{Z}_E) - h(\mathbf{Z}_R) + h(\mathbf{Z}_E) \\ &\leq h(\mathbf{X}_G^*+\mathbf{Z}_R) - h(\mathbf{X}_G^*+\mathbf{Z}_E) - h(\mathbf{Z}_R) + h(\mathbf{Z}_E) \\ &= I(\mathbf{X}_G^*;\mathbf{X}_G^*+\mathbf{Z}_R) - I(\mathbf{X}_G^*;\mathbf{X}_G^*+\mathbf{Z}_E), \qquad (4.89) \end{aligned}$$

where $\mu = 1$. Even though a Gaussian random vector $\mathbf{X}_G^*$ maximizes the difference of two mutual information, $I(\mathbf{X};\mathbf{X}+\mathbf{Z}_R) - I(\mathbf{X};\mathbf{X}+\mathbf{Z}_E)$, we cannot consider $I(\mathbf{X}_G^*;\mathbf{X}_G^*+\mathbf{Z}_R) - I(\mathbf{X}_G^*;\mathbf{X}_G^*+\mathbf{Z}_E)$ as the secrecy capacity. Now, we are going to prove the upper bound in (4.89) is actually the secrecy capacity of a general case.

Based on the equations (4.64)-(4.66) in the proof of Theorem 4.4, we know

$$\begin{aligned} & h(\mathbf{X}+\tilde{\mathbf{Z}}_R) - h(\mathbf{X}+\mathbf{Z}_E) - h(\tilde{\mathbf{Z}}_R) + h(\mathbf{Z}_E) \\ \leq\ & h(\mathbf{X}_G^*+\tilde{\mathbf{Z}}_R) - h(\mathbf{X}_G^*+\mathbf{Z}_E) - h(\tilde{\mathbf{Z}}_R) + h(\mathbf{Z}_E) \qquad (4.90) \\ =\ & h(\mathbf{X}_G^*+\tilde{\mathbf{Z}}_R) - h(\mathbf{X}_G^*+\mathbf{Z}_E) + h(\mathbf{Z}_R) - h(\tilde{\mathbf{Z}}_R) - h(\mathbf{Z}_R) + h(\mathbf{Z}_E) \qquad (4.91) \\ =\ & h(\mathbf{X}_G^*+\mathbf{Z}_R) - h(\mathbf{X}_G^*+\mathbf{Z}_E) - h(\mathbf{Z}_R) + h(\mathbf{Z}_E), \qquad (4.92) \end{aligned}$$

where $\tilde{\mathbf{Z}}_R$ is a Gaussian random vector with a covariance matrix $\mathbf{\Sigma}_{\tilde{Z}_R}$ which satisfies $\mathbf{\Sigma}_{\tilde{Z}_R} \preceq \mathbf{\Sigma}_{Z_R}$ and $\mathbf{\Sigma}_{\tilde{Z}_R} \preceq \mathbf{\Sigma}_{Z_E}$. The equation (4.90) denotes the secrecy capacity of a degraded case with noise $\tilde{\mathbf{Z}}_R$ and $\mathbf{Z}_E$, and this secrecy capacity of the degraded case upper-bounds the secrecy capacity of a general case since decreasing the covariance matrix of the noise $\mathbf{Z}_R$ always increases the secrecy capacity. Therefore, the secrecy capacity of a general Gaussian wire-tap channel is upper-bounded as

$$\begin{aligned} C_G &\leq C_D \\ &= h(\mathbf{X}_G^* + \tilde{\mathbf{Z}}_R) - h(\mathbf{X}_G^* + \mathbf{Z}_E) - h(\tilde{\mathbf{Z}}_R) + h(\mathbf{Z}_E) \\ &= h(\mathbf{X}_G^* + \mathbf{Z}_R) - h(\mathbf{X}_G^* + \mathbf{Z}_E) - h(\mathbf{Z}_R) + h(\mathbf{Z}_E), \end{aligned} \tag{4.93}$$

where C_G denotes the secrecy capacity of a general case.

Since we already know that a general case includes a degraded case and $C_D \leq C_G$, using the equations (4.90)-(4.92), we conclude

$$\begin{aligned} C_G &= h(\mathbf{X}_G^* + \tilde{\mathbf{Z}}_R) - h(\mathbf{X}_G^* + \mathbf{Z}_E) - h(\tilde{\mathbf{Z}}_R) + h(\mathbf{Z}_E) \\ &= h(\mathbf{X}_G^* + \mathbf{Z}_R) - h(\mathbf{X}_G^* + \mathbf{Z}_E) - h(\mathbf{Z}_R) + h(\mathbf{Z}_E) \\ &= \max_{0 \preceq \mathbf{\Sigma}_X \preceq \mathbf{R}} \left\{ \frac{1}{2} \log \frac{|\mathbf{\Sigma}_X + \mathbf{\Sigma}_{Z_R}|}{|\mathbf{\Sigma}_{Z_R}|} - \frac{1}{2} \log \frac{|\mathbf{\Sigma}_X + \mathbf{\Sigma}_{Z_E}|}{|\mathbf{\Sigma}_{Z_E}|} \right\}. \end{aligned} \tag{4.94}$$

4.5 Conclusions

The main contributions of this section are summarized as follows. First, an alternative proof of the extremal entropy inequality was provided. The alternative proof is simpler, more direct, and more information-theoretic than the original proofs. The alternative proof is mainly based on the data processing inequality which enables to by-pass the KKT conditions. Moreover, using properties of positive semi-definite

matrices, one can skip the step of proving the existence of the optimal solution which satisfies the KKT conditions, a step which is quite complicated to justify. Second, an additional important application for the extremal entropy inequality was suggested. By showing an additional application, we support how useful the extremal entropy inequality is. Finally, this section proposed a novel method to investigate several applications such as the capacity of the vector Gaussian broadcast channel, the secrecy capacity of the Gaussian wire-tap channel, etc. This novel technique is based on a data processing inequality, and it is very unique and creative in respect that it presents a novel paradigm for lots of applications such as the capacity of the vector Gaussian broadcast channel and the secrecy capacity of the Gaussian wire-tap channel, which were proved commonly based on the channel enhancement technique [32], [53], [31], and [14].

5. INFORMATION THEORETIC INEQUALITIES

5.1 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 the literature for these fundamental results are relatively quite diverse. Since differential entropy or Fisher information is a functional with respect to a probability density function, the most natural way to prove these results is by approaching them from the perspective of functional analysis. However, none of these results have been dealt with fully within the framework of calculus of variations. In addition, a number of challenging information theoretic inequalities such as the entropy power inequality (EPI) and the extremal entropy inequality (EEI) can be dealt with in the proposed framework of functional analysis. We believe that the proposed variational calculus perspective presents usefulness for establishing other novel results and extensions for the existing information theoretic inequalities.

The main theme of this section is to illustrate how some of the 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. This novel approach provides alternative proofs for some of the fundamental information theoretic inequalities and enables establishing extensions of the existing results. However, more importantly is the fact that the proposed approach suggests a potential guideline for finding the optimal solution for many other open problems.

The major results of this section 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 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 is considered and the second-order moment is given.

Second, a novel proof of the worst additive noise lemma [11] 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 [11], [43]. Unlike the previous proofs, our approach is purely based on calculus of variations, and both the scalar and vector versions of the lemma are treated.

Third, EPI is proved based on calculus of variations. We first re-cast EPI into a functional problem. Then, the necessary optimal solutions for the functional problem are found using Euler's equation, which is one of the necessary conditions for the functional problem. In a scalar version of EPI, the necessary optimal solution, which is the Gaussian density function, is actually sufficient since only the Gaussian density function satisfies the Euler's equation. This is one of the main benefits using calculus of variations. In a vector version of EPI, Euler's equation only shows that the Gaussian density functions are necessarily optimal, since the covariance matrices of

the optimal solutions are not determined. However, this information alone–i.e., the optimal solutions are Gaussian–is enough to prove EPI.

Finally, EEI is studied from the perspective of a functional problem. The main advantage of our proof is that neither the channel enhancement technique and EPI, used in [32], nor the equality condition of data processing inequality and the technique based on the moment generating functions, adopted in [41], are required. Using the unified argument based on calculus of variations, EEI is simply proved.

The rest of this section is organized as follows. Some variational calculus preliminary results and their corollaries are first reviewed in Section 5.2. Maximizing differential entropy theorem and minimizing Fisher information theorem (Cramér-Rao inequality) are proved in Section 5.3. In Section 5.4, the worst additive noise lemma is introduced and proved based on calculus of variations. EPI and EEI are proved in Sections 5.5 and 5.6, respectively. In Section 5.7, some applications of addressed information theoretic inequalities are briefly mentioned. Finally, Section 5.8 concludes this section.

5.2 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 section. These results are standard and therefore will be described briefly without further details. For additional details, the readers are suggested to consult any book on calculus of variations such as [16], [17], [44].

Definition 5.1. *A functional $U[f_X]$ is defined as*

$$U[f_X] = \int_a^b K(x, f_X, f'_X)dx, \tag{5.1}$$

which is defined on the set of continuous functions. The function f_X is assumed to have continuous first-order derivative in $[a, b]$ and to satisfy the boundary conditions $f_X(a) = A_X$ and $f_X(b) = B_X$. 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. Also, notation f'_X denotes the first-order derivative wrt x.

Definition 5.2. *The increment of a functional $U[f_X]$ is defined as*

$$\Delta U[h_X] = U[f_X + h_X] - U[f_X], \tag{5.2}$$

where the function h_X is the increment, and it is independent of the function f_X.

Definition 5.3. *Suppose that, given f_X,*

$$\Delta U\left[h_X\right] = \varphi\left[h_X\right] + \epsilon\|h_X\|, \tag{5.3}$$

where $\varphi\left[h_X\right]$ is a linear functional, ϵ goes zero as $\|h_X\|$ goes zero, and $\|\cdot\|$ denotes a norm and it is defined as

$$\|f_X\| = \sum_{i=0}^{n} \max_{a \le x \le b} \left| f_X^{(i)}(x) \right|, \tag{5.4}$$

where $f_X^{(i)}(x) = (d^i/dx^i) f_X(x)$, and summation upper index n varies depending on the normed linear space considered (e.g., if the normed linear space consists of all continuous functions $f_X(x)$–which have continuous first-order derivative–defined on an interval $[a, b]$, $\|f_X\| = \max_{a \le x \le b} |f_X(x)| + \max_{a \le x \le b} |f'_X(x)|$, and in this case $n = 1$). Then, the functional $U\left[f_X\right]$ is said to be differentiable, and the major part of the increment $\varphi\left[h_X\right]$ is called the (first-order) variation of the functional $U\left[f_X\right]$ and it is expressed as $\delta U\left[f_X\right]$.

Based on Definitions 5.1, 5.2, 5.3 and Taylor's theorem (see [16]), the first-order and the second-order variations of a functional $U[f_X]$ are expressed as

$$\begin{aligned}
\delta U[f_X] &= \int \left[K'_{f_X}(x, f_X, f'_X)\, h_X(x) + K'_{f'_X}(x, f_X, f'_X)\, h'_X(x) \right] dx, \qquad (5.5) \\
\delta^2 U[f_X] &= \frac{1}{2} \int \Big[K''_{f_X f_X}(x, f_X, f'_X)\, h_X(x)^2 + 2K''_{f_X f'_X}(x, f_X, f'_X)\, h_X(x) h'_X(x) \\
&\qquad + K''_{f'_X f'_X}(x, f_X, f'_X)\, h'_X(x)^2 \Big] dx \\
&= \frac{1}{2} \int \left[K''_{f'_X f'_X} {h'_X}^2 + \left(K''_{f_X f_X} - \frac{d}{dx} K''_{f_X f'_X} \right) {h_X}^2 \right] dx, \qquad (5.6)
\end{aligned}$$

where K'_{f_X} and $K'_{f'_X}$ are the first-order partial derivatives wrt f_X and f'_X, respectively, $K''_{f_X f'_X}$ is the second-order partial derivative wrt f_X and f'_X, $K''_{f_X f_X}$ is the second-order partial derivative wrt f_X, and $K''_{f'_X f'_X}$ is the second-order partial derivative wrt f'_X.*

Theorem 5.1 ([16]). *A necessary condition for the functional $U[f_X]$ in (5.1) to have an extremum (or, local optimum) for a given function f_{X^*} is the following:*

$$\delta U[f_{X^*}] = 0, \qquad (5.7)$$

for all admissible h_X. This implies

$$K'_{f_{X^*}} - \frac{d}{dx} K'_{f'_{X^*}} = 0, \qquad (5.8)$$

a result which is known as Euler's equation. When the functional in (5.1) includes multiple functions (e.g., $f_{X_1}, \ldots, f_{X_n}$) and multiple integrals wrt $x_1, \ldots, x_n$, then Eu-

*Throughout the section, the arguments of functionals or functions are omitted unless the arguments are ambiguous or confusing.

ler's equation in (5.8) is changed to

$$K'_{f_{X_i^*}} - \sum_{j=1}^{n} \frac{d}{dx_j} K'_{f'_{X_i^*}} = 0, \qquad i = 1, \ldots, n. \tag{5.9}$$

In particular, when the functional does not depend on the first-order derivative of the functions $f_{X_1}, \ldots, f_{X_n}$, the equation in (5.9) is simplified as

$$K'_{f_{X_i^*}} = 0, \qquad i = 1, \ldots, n. \tag{5.10}$$

Proof. Details of the proof of this theorem can be found e.g., in [16]. □

Theorem 5.2 ([16])**.** *A necessary condition for the functional $U[f_X]$ in (5.1) to have a minimum for a given f_{X^*} is the following:*

$$\delta^2 U[f_{X^*}] \geq 0, \tag{5.11}$$

for all admissible h_X. This implies

$$K''_{f'_{X^*} f'_{X^*}} \geq 0. \tag{5.12}$$

In particular, when the functional in (5.1) does not depend on the first-order derivative of the function f_X, the equation in (5.12) changes into

$$K''_{f_{X^*} f_{X^*}} \geq 0. \tag{5.13}$$

When the functional in (5.1) includes multiple functions (e.g., $f_{X_1}, \ldots, f_{X_n}$) and multiple integrals wrt $x_1, \ldots, x_n$, then the equation in (5.13) is changed into the

positive semi-definiteness of the following matrix:

$$\begin{bmatrix} K''_{f_{X_1} f_{X_1}} & \cdots & K''_{f_{X_1} f_{X_n}} \\ \vdots & \ddots & \vdots \\ K''_{f_{X_n} f_{X_1}} & \cdots & K''_{f_{X_n} f_{X_n}} \end{bmatrix}. \tag{5.14}$$

Proof. The inequality in (5.13) is easily derived from the inequality in (5.12) since $K''_{f'_X f'_X}$ and $K''_{f_X f'_X}$ are vanishing in (5.6) when the functional in (5.1) does not depend on the first-order derivative of the function f_X. Additional details of the proof can be found in [16]. □

Theorem 5.3 ([16])**.** *Given the functional*

$$U[f_X, f_Y] = \int_a^b K(x, f_X, f_Y, f'_X, f'_Y)dx, \tag{5.15}$$

assume that the admissible functions satisfy the following conditions:

$$f_X(a) = A_X, \quad f_X(b) = B_X, \quad f_Y(a) = A_Y, \quad f_Y(b) = B_Y,$$

$$k(x, f_X, f_Y) = 0, \tag{5.16}$$

$$L[f_X, f_Y] = \int_a^b \tilde{L}(x, f_X, f_Y, f'_X, f'_Y)dx = l, \tag{5.17}$$

where a, b, A_X, B_X, A_Y, B_Y, and l are constants, and $U[f_X, f_Y]$ is assumed to have an extremum for $f_X = f_{X^*}$ and $f_Y = f_{Y^*}$.

If f_{X^*} and f_{Y^*} are not extremals of $L[f_X, f_Y]$, or $k'_{f_{X^*}}$ and $k'_{f_{Y^*}}$ do not vanish simultaneously at any point in (5.16), there exists a constant λ or a function $\lambda(x)$ such that f_{X^*} and f_{Y^*} are extremals of the functional

$$\int_a^b \Big(K(x, f_X, f_Y, f'_X, f'_Y) + \lambda \tilde{L}(x, f_X, f_Y, f'_X, f'_Y) + \lambda(x) k(x, f_X, f_Y) \Big) dx. \tag{5.18}$$

Based on Theorem 5.3, the following corollary is derived.

Corollary 5.1. *Given the functional*

$$U[f_X, f_Y] = \int_a^b \int_a^b K(x, y, f_X, f_Y) dxdy, \tag{5.19}$$

assume that the admissible functions satisfy the following conditions:

$$\begin{aligned} &f_X(a, a) = A_X,\ f_X(b, b) = B_X,\ f_Y(a) = A_Y,\ f_Y(b) = B_Y,\ k(x, y, f_X, f_Y) = 0, \\ &L[f_X, f_Y] = \int_a^b \int_a^b \tilde{L}(x, y, f_X, f_Y) dxdy = l, \end{aligned} \tag{5.20}$$

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

Unless f_{X^*} and f_{Y^*} are extremals of $L[f_X, f_Y]$, or k'_{f_X} and k'_{f_Y} simultaneously vanish at any point of $k(x, y, f_X, f_Y)$, there exists a constant λ or a function $\lambda(y)$ such that $f_X = f_{X^*}$ and $f_Y = f_{Y^*}$ is an extremal of the functional

$$\int_a^b \left\{ \left(\int_a^b \left[K(x, y, f_X, f_Y) + \lambda \tilde{L}(x, y, f_X, f_Y) - \lambda(y) k(x, y, f_X) \right] dx \right) + \lambda(y) g(y, f_Y) \right\} dy \tag{5.21}$$

Proof. This corollary is a simple extension of Theorem 5.3 for multiple integrals. Therefore, the detailed proof is omitted. □

Based on Theorems 5.1, 5.2 and Corollary 5.1, we can derive the following corollary, which will be mainly used throughout this section.

Corollary 5.2. *Based on the functional defined in (5.21), the following necessary conditions are derived for the optimal solutions f_{X^*} and f_{Y^*}:*

$$K'_{f_{X^*}}(x, y, f_{X^*}, f_{Y^*}) - \lambda \tilde{L}'_{f_{X^*}}(x, y, f_{X^*}, f_{Y^*}) - \lambda(y) k'_{f_{X^*}}(x, y, f_{X^*}) = 0, \quad (5.22)$$

$$\int K'_{f_{Y^*}}(x, y, f_{X^*}, f_{Y^*}) - \lambda \tilde{L}'_{f_{Y^*}}(x, y, f_{X^*}, f_{Y^*}) dx + \lambda(y) g'_{f_{Y^*}}(y, f_{Y^*}) = 0, \quad (5.23)$$

and the matrix

$$\begin{bmatrix} G''_{f_{X^*}, f_{X^*}} & G''_{f_{X^*}, f_{Y^*}} \\ G''_{f_{Y^*}, f_{X^*}} & G''_{f_{Y^*}, f_{Y^*}} \end{bmatrix}, \quad (5.24)$$

where the functional G is defined as

$$\begin{aligned} G(x, y, f_{X^*}, f_{Y^*}) &= K(x, y, f_{X^*}, f_{Y^*}) - \lambda \tilde{L}(x, y, f_{X^*}, f_{Y^*}) - \lambda(y) k(x, y, f_{X^*}) \\ &\quad + \lambda(y) g(y, f_{Y^*}) q(x), \end{aligned}$$

and $q(x)$ is a function which satisfies $\int_a^b q(x)dx = 1$, is positive definite.

Proof. The equations in (5.22) and (5.23) are derived from the first-order variation condition in Theorem 5.1. Namely, the equations in (5.22) and (5.23) are Euler's equations for multiple integrals. The positive definiteness of the matrix in (5.24) is derived from the second-order variation condition in Theorem 5.2. Namely, this is the same as the one in (5.14). Since the proof is straightforward, the details of the proof are omitted here. □

5.3 MAX Entropy and MIN Fisher Information

This simple but significant result–given the second-order moment (or variance) of a random variable, a Gaussian density function maximizes the differential entropy

while it minimizes the Fisher information–is well-known. However, its complete rigorous proof can hardly be found. In this section, using calculus of variations, complete rigorous proofs will be provided.

Theorem 5.4 ([9])**.** *Given (the first-order) and the second-order moments of a random variable X, differential entropy of the random variable X is maximized when X is Gaussian, i.e.,*

$$h(X) \quad \leq \quad h(X_G), \tag{5.25}$$

where $h(\cdot)$ denotes differential entropy, and X_G is a Gaussian random variable whose (first-order) and second-order moments are identical to the one of X.

Proof. In [9], the proof relies on calculus of variations to find the first-order necessary condition, which confirms necessary optimal solutions. However, the first-order necessary condition shows neither whether the solutions are local minimal or local maximal nor whether the solutions are locally optimal or globally optimal. Therefore, an additional technique, the Kullback-Leibler divergence, was used to prove that the necessary solution globally maximizes the differential entropy. Unlike this proof, by confirming both the first-order and the second-order necessary conditions, we show that the optimal solution is a local maximal. Then, we prove that the local maximal is an actual global maximum achieving solution by showing that the local maximal is the only solution in the feasible set. Therefore, we can prove Theorem 5.4 purely based on calculus of variations. See Appendix C.1 for the details of the proof.

Remark 5.1. *Even though our proof is performed assuming constraints on the first-order and the second-order moments, the constraint of the first-order moment is not*

necessary. This will be shown in the proof of Theorem 5.5, which is the vector version of this theorem.

□

Similar to Theorem 5.4, given a correlation matrix (or a covariance matrix), a multi-variate Gaussian density function maximizes the differential entropy as shown by the following theorem.

Theorem 5.5 ([9], [43])**.** *Given (a mean vector $\boldsymbol{\mu}_X$) and a correlation matrix $\boldsymbol{\Omega}_X$, a Gaussian random vector maximizes the differential entropy, i.e.,*

$$h(\mathbf{X}) \quad \leq \quad h(\mathbf{X}_G), \tag{5.26}$$

where $h(\cdot)$ denotes differential entropy, $\mathbf{X}$ is an arbitrary but fixed random vector with the correlation matrix $\boldsymbol{\Omega}_X$, and $\mathbf{X}_G$ is a Gaussian random vector whose correlation matrix is identical to the one of $\mathbf{X}$.

Proof. See Appendix C.2.

Remark 5.2. *Our proof is different from the ones in [9], [43] in the sense that the proposed proof relies solely on variational calculus tools. Moreover, we show that the constraint related to the first-order moment is not necessary.*

Remark 5.3. *Depending on the existence of the constraint related to the mean vector, the mean of the optimal Gaussian density function is changed. However, the constraint on the mean vector is not necessarily required. Details of the proof are presented in Appendix C.2.*

□

If we only consider non-negative random variables, a Gaussian random variable is not the solution which maximizes the differential entropy. The following theorem shows that a half-normal random variable maximizes the differential entropy over the set of non-negative random variables.

Theorem 5.6. *Given an arbitrary but fixed non-negative random variable X and a half-normal random variable X_{HN}, whose second moments are identical to those of X, then the following relationship holds:*

$$h(X) \quad \leq \quad h(X_{HN}), \tag{5.27}$$

where $h(\cdot)$ denotes differential entropy.

Proof. See Appendix C.3. □

Similar to Theorems 5.4, 5.5, and 5.6, we can find a probability density function, which minimizes the Fisher information.

Theorem 5.7 (Cramér-Rao Inequality)**.** *Given (the first-order moment μ_X) and the second-order moment m_X^2, a Gaussian random variable X_G minimizes Fisher information, i.e.,*

$$J(X) \quad \geq \quad J(X_G), \tag{5.28}$$

where X is an arbitrary but fixed random variable with the first-order moment μ_X and the second-order moment m_X^2. Notation $J(\cdot)$ denotes the Fisher information, and it is defined as

$$J(X) \quad = \quad \int \left(\frac{\frac{d}{dx} f_X(x)}{f_X(x)} \right)^2 f_X(x) dx.$$

Proof. See Appendix C.4.

Remark 5.4. *Even though several versions of the proof of this theorem have been studied, this is the first rigorous proof of this theorem based on calculus of variations.*

□

Theorem 5.7 can be generalized for random vectors as shown in the following theorem.

Theorem 5.8 (Cramér-Rao Inequality (a vector version))**.** *Given an arbitrary but fixed random vector* $\mathbf{X}$ *and a Gaussian random vector* $\mathbf{X}_G$*, whose mean vectors and correlation matrices are identical, respectively,*

$$\mathbb{J}(\mathbf{X}) \succeq \mathbb{J}(\mathbf{X}_G), \tag{5.29}$$

where $\mathbb{J}(\cdot)$ *denotes Fisher information matrix, and it is defined as*

$$\mathbb{J}(\mathbf{X}) = \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{bmatrix}, \tag{5.30}$$

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

Proof. See Appendix C.5. □

Similar to Theorem 5.7, a half-normal and a chi density function minimize the Fisher information over the set of non-negative random variables as shown in the following two theorems.

Theorem 5.9. *Assume that the regularity condition for Fisher information is ignored. Given an arbitrary but fixed non-negative random variable X and a half-normal random variable X_{HN}, whose second order moments are identical to those of X, then the following inequality holds:*

$$J(X) \;\geq\; J(X_{HN}), \tag{5.31}$$

where $J(\cdot)$ denotes Fisher information. The regularity condition is the following relationship:

$$\int \frac{d}{dx} f(x) dx = 0. \tag{5.32}$$

Proof. See Appendix C.6. □

Theorem 5.10 ([2])**.** *Assume next that random variables, which satisfy the regularity condition in (5.32), are considered. Given an arbitrary but fixed non-negative random variable X and a chi-distributed random variable X_C, whose second-order moments are identical to those of X, then the following inequality holds:*

$$J(X) \;\geq\; J(X_C), \tag{5.33}$$

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

Proof. Unlike the proof in [2], by considering the first-order and the second-order moments instead of variance, we obtain the convex constraint sets. Since Fisher information is a strictly convex functional with respect to a probability density function, the variational problem is convex, and hence has an unique solution. The details of the proof are deferred to Appendix C.7. □

5.4 Worst Additive Noise Lemma

Worst additive noise lemma was introduced and exploited in several references [11], [43], [23], and it has been widely used in numerous applications. One of the main applications of the worst additive noise lemma pertains to the calculation of channel capacity under several different wireless communications scenarios such as the Gaussian MIMO broadcasting channel, Gaussian MIMO wire-tap channel, etc. In this section, the worst additive noise lemma for both random variables and random vectors will be proved solely based on calculus of variations.

Theorem 5.11. *Assume X is an arbitrary but fixed random variable and X_G is a Gaussian random variable, whose second-order moment is identical to the one of X, and it is denoted as m_X^2. Given a Gaussian random variable W_G, which is independent of both X and X_G, with the second-order moment m_W^2, then the following relationship holds:*

$$I(X + W_G; W_G) \geq I(X_G + W_G; W_G), \tag{5.34}$$

where $I(\cdot;\cdot)$ denotes mutual information.

Proof. The details of the proof are deferred to Appendix C.8. □

Similarly, Theorem 5.11 can be generalized to random vectors as shown in the following theorem.

Theorem 5.12. *Assume $\mathbf{X}$ is an arbitrary but fixed random vector and $\mathbf{X}_G$ is a Gaussian random vector, whose correlation matrix is identical to the one of $\mathbf{X}$, and it is denoted as $\mathbf{\Omega}_X$. Given a Gaussian random vector $\mathbf{W}_G$, which is independent of*

both $\mathbf{X}$ *and* $\mathbf{X}_G$*, with the correlation matrix* $\mathbf{\Omega}_W$*, then the following relation holds:*

$$I(\mathbf{X}+\mathbf{W}_G;\mathbf{W}_G) \geq I(\mathbf{X}_G+\mathbf{W}_G;\mathbf{W}_G). \tag{5.35}$$

Proof. Our novel proof is wholly based on calculus of variations arguments. The summary of our proof is the following. First, we construct a variational problem, which represents the inequality in (5.35) and required constraints. Second, using the first-order variation condition, we find 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 prove that the local minimum is also global. The details of the proof are presented to Appendix C.9. □

5.5 Entropy Power Inequality

Entropy power inequality (EPI) is a powerful result that found applicability in determining the capacity of scalar Gaussian broadcast channel [3], the capacity of Gaussian MIMO broadcast channel [32], [53], the secrecy capacity of Gaussian wire-tap channel [31], [41], etc., in conjunction with Fano's inequality and additional techniques such as the ones proposed in [53], [41]. In this section, we will prove several versions of EPI using calculus of variations techniques.

Theorem 5.13 (Entropy Power Inequality)**.** *For two independent random variables* X *and* W*, whose entropies and second-order moments are finite,*

$$h(a_X X + a_W W) \geq a_X^2 h(X) + a_W^2 h(W), \tag{5.36}$$

where $a_X^2 + a_W^2 = 1$*. The equality holds if and only if* X *and* W *are Gaussian random variables.*

Proof. See Appendix C.10. □

Theorem 5.14 (Entropy Power Inequality)**.** *For two independent random vectors* $\mathbf{X}$ *and* $\mathbf{W}$*, with finite entropies and correlation matrices, the following relation holds:*

$$h(a_X\mathbf{X} + a_W\mathbf{W}) \geq a_X^2 h(\mathbf{X}) + a_W^2 h(\mathbf{W}), \tag{5.37}$$

where $a_X^2 + a_W^2 = 1$*. The equality holds if and only if* $\mathbf{X}$ *and* $\mathbf{W}$ *are Gaussian random vectors and their covariance matrices* $\mathbf{\Sigma}_X$ *and* $\mathbf{\Sigma}_W$ *are identical.*

Proof. See Appendix C.11. □

5.6 Extremal Entropy Inequality

Extremal entropy inequality, 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, was proposed by Liu and Viswanath [32]. It 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 EPI. Therefore, new techniques ([53], [41]) were adopted in the proofs reported in [32], [41]. In this section, the extremal entropy inequality will be proved using calculus of variations.

Theorem 5.15. *Assume that* μ *is an arbitrary but fixed constant, where* $\mu \geq 1$*, and* r^2 *is a positive constant. A Gaussian random variable* W_G *with variance* σ_W^2 *is assumed to be independent of an arbitrary random variable* X*, with variance* $\sigma_X^2 \leq r^2$*. Then, there exists a Gaussian random variable* X_G^* *with variance* $\sigma_{X^*}^2$ *which satisfies the following inequality:*

$$h(X) - \mu h(X + W_G) \leq h(X_G^*) - \mu h(X_G^* + W_G), \tag{5.38}$$

where $\sigma_{X^*}^2 \leq r^2$.

Proof. See Appendix C.12. □

Theorem 5.15 can be generalized for random vectors as shown in the following two theorems.

Theorem 5.16. *Assume that* μ *is an arbitrary but fixed constant, where* $\mu \geq 1$*, and* $\mathbf{\Sigma}$ *is a positive semi-definite matrix. A Gaussian random vector* $\mathbf{W}_G$ *with positive definite covariance matrix* $\mathbf{\Sigma}_W$ *is assumed to be independent of an arbitrary random vector* $\mathbf{X}$ *whose covariance matrix* $\mathbf{\Sigma}_X$ *satisfies* $\mathbf{\Sigma}_X \preceq \mathbf{\Sigma}$*. Then, there exists a Gaussian random vector* $\mathbf{X}_G^*$ *with covariance matrix* $\mathbf{\Sigma}_{X^*}$ *which satisfies the following inequality:*

$$h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{W}_G) \leq h(\mathbf{X}_G^*) - \mu h(\mathbf{X}_G^* + \mathbf{W}_G), \tag{5.39}$$

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

Proof. See Appendix C.13. □

Remark 5.5. *As the extremal entropy inequality only shows the existence of necessary optimal solutions in [32] and [41], the current proof also shows the existence of necessary optimal solutions. In addition, the proposed proof only exploits calculus of variations tools. Namely, this proof does not adopt neither the channel enhancement technique and EPI in [32] nor the EPI and data processing inequality in [41].*

Theorem 5.17. *Assume that* μ *is an arbitrary but fixed constant, with* $\mu \geq 1$*, 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:*

$$h(\mathbf{X}+\mathbf{W}_G)-\mu h(\mathbf{X}+\mathbf{V}_G)\leq h(\mathbf{X}_G^*+\mathbf{W}_G)-\mu h(\mathbf{X}_G^*+\mathbf{V}_G), \tag{5.40}$$

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

Proof. See Appendix C.14. □

Remark 5.6. *The proposed proof does not borrow any techniques from [32]. Even though the proposed proof adopts the equality condition for the data processing inequality, a result which was also exploited in [41], the proposed proof is different from the one in [41] in the following sense. First, the proposed proof uses the equality condition of the data processing inequality only once while the proof in [41] used it twice. The proof in [32] 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 [41]; instead the current proof directly exploits a property of the conditional mutual information pertaining to a Markov chain.*

5.7 Applications

The importance of information theoretic inequalities such as EPI, extremal entropy inequality, etc., were already proved by several applications. For example, minimum Fisher information theorem (Cramér-Rao inequality) and maximum entropy theorem were used for developing min-max robust estimation techniques [49], [4], [48]. EPI was first adapted to prove a lower bound on the capacity of additive noise channels by Shannon [45]. Also, EPI was exploited for the scalar Gaussian

broadcast channel [3], the scalar quadratic Gaussian CEO problem [38], etc. The extremal entropy inequality can be used in the vector Gaussian broadcast channel [32], the distributed source coding with a single quadratic distortion constraint problem [32], and the Gaussian wire-tap channel [41], and so on. Even though these applications were traditionally addressed using the above mentioned information theoretic inequalities, we can directly approach these applications by means of variational calculus techniques.

5.8 Conclusions

In this section, we derived several fundamental information theoretic inequalities using a functional analysis framework. The main benefit for employing calculus of variations for proving information theoretic inequalities is the fact that the global optimal solution is obtained from the necessary conditions for optimality without additional calculations. The summary of our 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 entropy power inequality and the extremal entropy inequality were derived using calculus of variations. Finally, applications that could be addressed based on the proposed results were briefly mentioned.

6. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this dissertation, three major topics were studied. First, three relationships between information theory and statistical estimation theory–the equivalence between Stein's identity and De Bruijn's identity and two different extensions of De Bruijn's identity–were disclosed. Several applications based on the proposed relationships support the importance of the proposed results. Second, the Gaussian assumption was researched. This assumption was studied from two different perspectives: information theory and estimation theory. Based on these results, the min-max optimal approach was researched, and optimal training sequences for the channel and the frequency offset were proposed as an application of these results. Third, extremal entropy inequality was studied. By by-passing major techniques, the channel enhancement and KKT conditions, which were used in the previous proofs [32], this thesis presented a novel paradigm not only for the extremal entropy inequality but also for other applications such as the capacity of the vector Gaussian broadcast channel and the secrecy capacity of the Gaussian wire-tap channel, which were proved commonly established based on the channel enhancement technique. Finally, several fundamental information theoretic inequalities were proved using a functional analysis framework. The entropy maximizing theorem, Fisher information minimizing theorem, entropy power inequality, and extremal entropy inequality were established in the unified framework offered by calculus of variations. The major advantage for using calculus of variations for proving the information theoretic inequalities is the fact that the sufficient optimal solution is obtained from the necessary conditions for optimality without performing additional calculations.

Numerous possible future research directions could be considered. Many results

introduced in Sections 2 and 4 are mainly theoretical. There are many areas where the proposed results could be further adopted. Numerous applications of Stein's identity could be translated into the realm of De Bruijn identity. In particular, the proposed results may be useful for developing robust estimation and detection schemes in the presence of uncertainties in the distribution of observations. In addition, the novel approaches, proposed in Sections 4 and 5, can be adapted to many other open problems in information theory. For example, extending EPI or EEI to the case of positive-valued random variables or random variables whose values are restricted to an interval represent very challenging problems that could be successfully attacked within the proposed functional analysis framework by taking advantage of calculus of variations techniques. Similar extensions might be developed for the worst additive noise lemma.

REFERENCES

[1] A. R. Barron. Entropy and the central limit theorem. *The Annals. of Prob.*, 14:336–342, Jan. 1986.

[2] J. Bercher and C. Vignat. On minimum Fisher information distributions with restricted support and fixed variance. *Inform. Sci.*, 179:3832–3842, Nov. 2009.

[3] P. P. Bergmans. A simple converse for broadcast channels with additive white Gaussian noise. *IEEE Trans. Inf. Theory*, 20:279 – 280, Mar. 1974.

[4] O. Besson and P. Stoica. Training sequence selection for frequency offset estimation in frequency-selective channels. *Digital Signal Process.*, 13:106–127, Jan. 2003.

[5] N. M. Blachman. The convolution inequality for entropy power. *IEEE Trans. Inf. Theory*, 11:267 – 271, Apr 1965.

[6] L. Brown, A. DasGupta, L. R. Haff, and W. E. Strawderman. The heat equation and Stein's identity: Connections, applications. *Journ. of Stat. Planning and Infer.*, 136:2254–2278, July 2006.

[7] M. H. Costa. A new entropy power inequality. *IEEE Trans. Inform. Theory*, 31:751–760, Nov. 1985.

[8] M. H. M. Costa and T. M. Cover. On the similarity of the entropy power inequality and the Brunn-Minkowski inequality. *IEEE Trans. Inf. Theory*, 30:837–839, Nov. 1984.

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

[10] A. Dembo, T. M. Cover, and J. A. Thomas. Information theoretic inequalities. *IEEE Trans. Inf. Theory*, 37:1501–1518, Nov. 1991.

[11] S. N. Diggavi and T. M. Cover. The worst additive noise under a covariance constraint. *IEEE Trans. Inf. Theory*, 47:3072–3081, Nov. 2001.

[12] E. Ekrem and S. Ulukus. An alternative proof for the capacity region of the degraded Gaussian MIMO broadcast channel. *IEEE Trans. Inform. Theory*, pages 2427–2433, Apr. 2012.

[13] E. Ekrem and S. Ulukus. Secrecy capacity region of the Gaussian multi-receiver wiretap channel. *Proc. IEEE Int. Symp. Inform. Theory*, pages 2612–2616, Jun. 2009.

[14] E. Ekrem and S. Ulukus. The secrecy capacity region of the Gaussian MIMO multi-receiver wiretap channel. *IEEE Trans. Inf. Theory*, 57:2083 – 2114, Mar 2011.

[15] Y. C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Trans. Sig. Proc.*, 57:471–481, Feb. 2009.

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

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

[18] D. Guo, S. Shamai (Shitz), and S. Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Inform. Theory*, 51:1261–1282, Apr. 2005.

[19] D. Guo, S. Shamai (Shitz), and S. Verdú. Proof of entropy power inequalities via MMSE. *Proc. IEEE Int. Symp. Inform. Theory*, pages 1011–1015, Jul. 2006.

[20] D. Guo, S. Shamai (Shitz), and S. Verdú. Additive non-Gaussian noise channels: mutual information and conditional mean estimation. *Proc. IEEE Int. Symp. Inform. Theory*, pages 719–723, Sep. 2005.

[21] R. A. Horn and C. R. Johnson. *Matrix Analysis*. New York: Cambridge University Press, 1985.

[22] H. M. Hudson. A natural identity for exponential families with applications in multiparameter estimation. *Stat. and Prob. Letters*, 6:473–484, May 1978.

[23] S. Ihara. On the capacity of channels with additive non-Gaussian noise. *Inform. Contr.*, 37:34–39, Apr. 1978.

[24] O. Johnson. *Information Theory and the Central Limit Theorem*. London: Imperial College Press, 2004.

[25] O. Johnson. A conditional entropy power inequality for dependent variables. *IEEE Trans. Inf. Theory*, 50:1581 – 1583, Aug. 2004.

[26] S. K. Kattumannil. On Stein's identity and its applications. *Stat. and Prob. Letters*, 79:1444–1449, June 2009.

[27] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory (Vol 1)*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[28] A. Khisti and G. W. Wornell. Secure transmission with multiple antennas: The MIMOME channel. *IEEE Trans. Inf. Theory*, 56:5515 – 5532, Nov. 2010.

[29] A. Khisti, G. W. Wornell, A. Wiesel, and Y. Eldar. On the Gaussian MIMO wiretap channel. *Proc. IEEE Int. Symp. Information Theory*, pages 2471–2475, Jun. 2007.

[30] R. Liu, T. Liu, H. V. Poor, and S. Shamai (Shitz). A vector generalization of Costa's entropy-power inequality with applications. *IEEE Trans. Inf. Theory*, 56:1865 – 1879, Apr. 2010.

[31] T. Liu and S. Shamai (Shitz). A note on the secrecy capacity of the multiple-antenna wiretap channel. *IEEE Trans. Inf. Theory*, 55:2547 – 2553, Jun 2009.

[32] T. Liu and P. Viswanath. An extremal inequality motivated by multiterminal information-theoretic problems. *IEEE Trans. Inf. Theory*, 53:1839 – 1851, May 2007.

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

[34] J. H. Manton and Y. Hua. Rank reduction and James-Stein estimation. *IEEE Trans. Sig. Proc.*, 47:3121–3125, Nov. 1999.

[35] J. H. Manton, V. Krishnamurthy, and H. V. Poor. James-Stein state filtering algorithms. *IEEE Trans. Sig. Proc.*, 46:2431–2447, Sep. 1998.

[36] C. N. Morris. Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Stat.*, 11:515–529, Jun. 1983.

[37] F. Oggier and B. Hassibi. The secrecy capacity of the MIMO wiretap channel. *Proc. IEEE Int. Symp. Information Theory*, pages 524–528, Jul. 2008.

[38] Y. Oohama. The rate-distortion function for the quadratic gaussian ceo problem. *IEEE Trans. Inf. Theory*, 44:132 – 133, May 2011.

[39] D. P. Palomar and S. Verdú. Representation of mutual information via input estimates. *IEEE Trans. Inform. Theory*, 53:453–470, Feb. 2007.

[40] D. P. Palomar and S. Verdú. Gradient of mutual information in linear vector Gaussian channels. *IEEE Trans. Inform. Theory*, 52:141–154, Jan. 2006.

[41] S. Park, E. Serpedin, and K. Qaraqe. An alternative proof of an extremal inequality. IEEE Trans. Inf. Theory (submitted), arXiv:1201.6681.

[42] M. Payaro and D. P. Palomar. Hessian and concavity of mutual information, differential entropy, and entropy power in linear vector Gaussian channels. *IEEE Trans. Inform. Theory*, 55:3613–3628, Aug. 2009.

[43] O. Rioul. Information theoretic proofs of entropy power inequalities. *IEEE Trans. Inform. Theory*, 57:33–55, Jan. 2011.

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

[45] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623 – 656, Oct 1959.

[46] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inf.& Cont.*, 2:101–112, Jun. 1959.

[47] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *in Proc. Third Berkeley Symp. on Math. Statist. and Prob., (Univ. of Calif. Press)*, 1:197–206, 1956.

[48] P. Stoica and P. Babu. The Gaussian data assumption leads to the largest Cramér-Rao bound. *IEEE Signal Process. Mag.*, 28:132–133, May 2011.

[49] P. Stoica and O. Besson. Training sequence design for frequency offset and frequency-selective channel estimation. *IEEE Trans. Commun.*, 51:1910–1917, Nov. 2003.

[50] H. L. Van Trees. *Detection, Estimation, and Modulation Theory: Part I.* New York: Wiley, 2001.

[51] S. Verdú and D. Guo. A simple proof of the entropy-power inequality. *IEEE Trans. Inf. Theory*, 52:2165 – 2166, May 2006.

[52] C. Villani. A short proof of the concavity of entropy power. *IEEE Trans. Inform. Theory*, 46:1695–1696, Jul. 2000.

[53] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz). The capacity region of the Gaussian mutiple-input multiple-output broadcast channel. *IEEE Trans. Inf. Theory*, 52:3936 – 3964, Sep. 2006.

[54] A. D. Wyner. The wire-tap channel. *Bell Syst. Tech. J.*, 54:1355–1387, Oct. 1975.

APPENDIX A

STEIN'S IDENTITY AND DE BRUIJN'S IDENTITY

A.1 A Proof of Theorem 2.4

Since Theorem 2.5 is considered as a special case of Theorem 2.4, we only show the proof of Theorem 2.4 in this section.

Proof. [Theorem 2.4]

Prior to proving Theorem 2.4, we first introduce the following relationships in Lemma A.1, which are required for the proof.

Lemma A.1. *For random variables W, X and Y defined in equation (2.1) when Gaussian random variable W has zero mean and unit variance and random variable X has finite second-order moment, the following identities are satisfied:*

$$\textit{i)}\quad \frac{d}{da}\log f_Y(y;a)\bigg|_{y=u+\sqrt{a}w} = \frac{1}{2a^2}\left(\frac{\mathbb{E}_X\left[(y-X)^2 f_{Y|X}(y|X;a)\right]}{f_Y(y;a)} - a\right)\bigg|_{y=u+\sqrt{a}w},$$

$$\textit{ii)}\quad \frac{d}{da}\log f_Y(u+\sqrt{a}w;a) = \frac{1}{2a^2}\left(\frac{\mathbb{E}_X[(u-X)(y-X)f_{Y|X}(y|X;a)]}{f_Y(y;a)} - a\right)\bigg|_{y=u+\sqrt{a}w},$$

$$\textit{iii)}\quad \frac{d}{dy}\log f_Y(y;a)\bigg|_{y=u+\sqrt{a}w} = -\frac{\mathbb{E}_X\left[(y-X)f_{Y|X}(y|X;a)\right]}{a f_Y(y;a)}\bigg|_{y=u+\sqrt{a}w},$$

$$\textit{iv)}\quad \frac{w}{2\sqrt{a}}\frac{d}{dy}\log f_Y(y;a)\bigg|_{y=u+\sqrt{a}w} = \frac{d}{da}\log f_Y(u+\sqrt{a}w;a) - \left[\frac{d}{da}\log f_Y(y;a)\right]_{y=u+\sqrt{a}w},$$

where $f(y)|_{y=a}$ denotes $\lim_{y\to a} f(y)$. In some cases, to avoid confusion, $[f(y)]_{y=a}$ is used instead of $f(y)|_{y=a}$.

Proof. Since $f_{Y|X}(y|x;a)$ is normally distributed with mean x and variance a, the

following relationships hold:

$$f_{Y|X}(y|x;a) = \frac{1}{\sqrt{2\pi a}} \exp\left(-\frac{(y-x)^2}{2a}\right), \tag{A.1}$$

$$\frac{d}{dy} f_{Y|X}(y|x;a) = -\frac{1}{a}(y-x) f_{Y|X}(y|x;a), \tag{A.2}$$

$$\frac{d}{da} f_{Y|X}(y|x;a) = \left(-\frac{1}{2a} + \frac{1}{2a^2}(y-x)^2\right) f_{Y|X}(y|x;a), \tag{A.3}$$

$$\begin{aligned}\frac{d}{da} f_{Y|X}(u+\sqrt{a}w|x;a) &= f_{Y|X}(u+\sqrt{a}w|x;a) \\ &\quad \times \left(-\frac{1}{2a} + \frac{1}{2a^2}(u+\sqrt{a}w-x)(u-x)\right).\end{aligned} \tag{A.4}$$

Equation (A.4) is true since

$$\begin{aligned}&\frac{d}{da} f_{Y|X}(u+\sqrt{a}w|x;a) \\ =&\frac{d}{da}\left[\frac{1}{\sqrt{2\pi a}} \exp\left(-\frac{1}{2a}(u+\sqrt{a}w-x)^2\right)\right] \\ =&-\frac{1}{2a}\left(\frac{1}{\sqrt{2\pi a}} \exp\left(-\frac{1}{2a}(u+\sqrt{a}w-x)^2\right)\right) \\ &+\left(\frac{1}{\sqrt{2\pi a}} \exp\left(-\frac{1}{2a}(u+\sqrt{a}w-x)^2\right)\right) \\ &\quad \times\left(-\frac{2(u+\sqrt{a}w-x)(\frac{w}{2\sqrt{a}})a - (u+\sqrt{a}w-x)^2}{2a^2}\right) \\ =&-\frac{1}{2a} f_{Y|X}(u+\sqrt{a}w|x;a) \\ &+f_{Y|X}(u+\sqrt{a}w|x;a)\left(-\frac{(u+\sqrt{a}w-x)(u-x)}{2a^2}\right).\end{aligned}$$

Based on equation (A.3), i) is proved by following these calculations:

$$\begin{aligned}\frac{d}{da}\log f_Y(y;a)\bigg|_{y=u+\sqrt{a}w} &= \frac{\mathbb{E}_X\left[\frac{d}{da} f_{Y|X}(y|X;a)\right]}{f_Y(y;a)}\bigg|_{y=u+\sqrt{a}w} \\ &= \frac{1}{2a^2}\left(\frac{\mathbb{E}_X\left[(y-X)^2 f_{Y|X}(y|X;a)\right]}{f_Y(y;a)} - a\right)\bigg|_{y=u+\sqrt{a}w}\end{aligned} \tag{A.5}$$

Second, equation ii) is proved by the following calculations:

$$\begin{aligned}
&\frac{d}{da}\log f_Y(u+\sqrt{a}w;a)\\
=&\frac{\mathbb{E}_X\left[\frac{d}{da}f_{Y|X}(u+\sqrt{a}w|X;a)\right]}{f_Y(u+\sqrt{a}w;a)}\\
=&\frac{\mathbb{E}_X\left[-\frac{1}{2a}f_{Y|X}(u+\sqrt{a}w|X;a)\right]}{f_Y(u+\sqrt{a}w;a)}\\
&+\frac{\mathbb{E}_X\left[\frac{1}{2a^2}(u+\sqrt{a}w-X)(u-X)f_{Y|X}(u+\sqrt{a}w|X;a)\right]}{f_Y(u+\sqrt{a}w;a)}\\
=&\frac{-af_Y(u+\sqrt{a}w;a)}{2a^2f_Y(u+\sqrt{a}w;a)}\\
&+\frac{\mathbb{E}_X\left[(u+\sqrt{a}w-X)(u-X)f_{Y|X}(u+\sqrt{a}w|X;a)\right]}{2a^2f_Y(u+\sqrt{a}w;a)}
\end{aligned}\tag{A.6}$$

$$\begin{aligned}
=&\frac{1}{2a^2}\left(\frac{\mathbb{E}_X[(u+\sqrt{a}w-X)(u-X)f_{Y|X}(u+\sqrt{a}w|X;a)]}{f_Y(u+\sqrt{a}w;a)}-a\right)\\
=&\frac{1}{2a^2}\left(\frac{\mathbb{E}_X[(y-X)(u-X)f_{Y|X}(y|X;a)]}{f_Y(y;a)}-a\right)\Bigg|_{y=u+\sqrt{a}w}.
\end{aligned}\tag{A.7}$$

The equality in (A.6) is due to equation (A.4).

Third, equation iii) is proved based on equation (A.2) as follows:

$$\begin{aligned}
\frac{d}{dy}\log f_Y(y;a)\Big|_{y=u+\sqrt{a}w} &= \frac{\mathbb{E}_X\left[\frac{d}{dy}f_{Y|X}(y|X;a)\right]}{f_Y(y;a)}\Bigg|_{y=u+\sqrt{a}w}\\
&= \frac{-\mathbb{E}_X\left[(y-X)f_{Y|X}(y|X;a)\right]}{af_Y(y;a)}\Bigg|_{y=u+\sqrt{a}w}.
\end{aligned}\tag{A.8}$$

The equality in (A.8) is due to equation (A.2).

Equation iv) is trivial since equation (A.8) multiplied by $w/2\sqrt{a}$ is equal to equation (A.7) minus equation (A.5), and the proof is completed. □

Like the proof of Theorem 2.3 in [6], the equivalence is proved by showing that each identity is derived from the other one, using Lemma A.1.

First, in the generalized Stein's identity, all necessary functions are defined as follows:

$$r(y;a) = -\frac{d}{dy}\log f_Y(y;a), \quad k(y) = 1, \quad t(y;a) = -\frac{\frac{d}{dy}f_Y(y;a)}{f_Y(y;a)}, \quad \text{and} \quad \nu = 0. \tag{A.9}$$

Then, De Bruijn's identity is derived from the generalized Stein's identity as follows.

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_Y\left[\frac{d}{dY}r(Y;a)\right] \\
=&\frac{1}{2}\mathbb{E}_Y\left[r(Y;a)t(Y;a)\right] \quad \text{(generalized Stein's identity)} \qquad (A.10)\\
=&-\frac{1}{2}\int_{-\infty}^{\infty}\frac{d}{dy}\mathbb{E}_X\left[f_{Y|X}(y|X;a)\right]r(y;a)dy \\
=&-\mathbb{E}_X\left[\int_{-\infty}^{\infty}\frac{(y-X)}{2a}f_{Y|X}(y|X;a)\frac{d}{dy}\log f_Y(y;a)dy\right] \\
=&-\int_{-\infty}^{\infty}f_X(u)\underbrace{\int_{-\infty}^{\infty}\frac{(y-u)}{2a}f_{Y|X}(y|u;a)\frac{d}{dy}\log f_Y(y;a)dy}_{(A)}du. \qquad (A.11)
\end{aligned}$$

The interchangeability among integrals and derivatives are due to the dominated convergence theorem and Fubini's theorem.

Changing the variable as $y = u + \sqrt{a}w$, equation (A) is expressed as

$$
\begin{aligned}
&\int_{-\infty}^{\infty} \frac{(y-u)}{2a} f_{Y|X}(y|u;a) \frac{d}{dy} \log f_Y(y;a) dy \\
&= \int_{-\infty}^{\infty} \frac{\sqrt{a}w}{2a} f_{Y|X}(u+\sqrt{a}w|u;a) \left[\frac{d}{dy} \log f_Y(y;a) \right]_{y=u+\sqrt{a}w} \sqrt{a} dw \\
&= \int_{-\infty}^{\infty} f_{Y|X}(u+\sqrt{a}w|u;a) \left(\frac{d}{da} \log f_Y(u+\sqrt{a}w;a) \right. \\
&\qquad\qquad \left. - \left[\frac{d}{da} \log f_Y(y;a) \right]_{y=u+\sqrt{a}w} \right) \sqrt{a} dw \qquad \text{(A.12)} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) \frac{d}{da} \log f_Y(u+\sqrt{a}w;a) dw \\
&\quad - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) \left[\frac{d}{da} \log f_Y(y;a) \right]_{y=u+\sqrt{a}w} dw \\
&= \frac{d}{da} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) \log f_Y(u+\sqrt{a}w;a) dw \\
&\quad - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) \left[\frac{d}{da} \log f_Y(y;a) \right]_{y=u+\sqrt{a}w} dw. \qquad \text{(A.13)}
\end{aligned}
$$

The equality in equation (A.12) is due to Lemma A.1, iv).

Re-defining the variable $w = (y-u)/\sqrt{a}$, equation (A.11) is expressed as

$$
\begin{aligned}
&-\int_{-\infty}^{\infty} f_X(u)\left(\int_{-\infty}^{\infty}\frac{(y-u)}{2a} f_{Y|X}(y|u;a)\frac{d}{dy}\log f_Y(y;a)dy\right)du \\
&=\int_{-\infty}^{\infty} f_X(u)\Bigg(\int_{-\infty}^{\infty} f_{Y|X}(y|u;a)\frac{d}{da}\log f_Y(y;a)dy \\
&\qquad\qquad -\frac{d}{da}\int_{-\infty}^{\infty} f_{Y|X}(y|u;a)\log f_Y(y;a)dy\Bigg)du \qquad \text{(A.14)} \\
&=\int_{-\infty}^{\infty} f_Y(y;a)\frac{d}{da}\log f_Y(y;a)dy \\
&\qquad\qquad -\frac{d}{da}\int_{-\infty}^{\infty} f_Y(y;a)\log f_Y(y;a)dy \qquad \text{(A.15)} \\
&=\int_{-\infty}^{\infty}\frac{d}{da} f_Y(y;a)dy - \frac{d}{da}\int_{-\infty}^{\infty} f_Y(y;a)\log f_Y(y;a)dy \\
&=\frac{d}{da}\int_{-\infty}^{\infty} f_Y(y;a)dy - \frac{d}{da}\int_{-\infty}^{\infty} f_Y(y;a)\log f_Y(y;a)dy \\
&=-\frac{d}{da}\int_{-\infty}^{\infty} f_Y(y;a)\log f_Y(y;a)dy \\
&=\frac{d}{da}h(Y).
\end{aligned}
$$

The equality in (A.14) is due to the change of variable, and the equality in (A.15) is because of the independence of $f_X(u)$ with respect to a.

Since the left-hand side of equation (A.10) is equal to $J(Y)/2$, we obtain De Bruijn's identity:

$$\frac{1}{2}J(Y) \;=\; \frac{d}{da}h(Y),$$

from the generalized Stein's identity.

Second, the generalized Stein's identity is derived from De Bruijn's identity. We

define the function

$$g(y;a) \;=\; \int_0^y r(u;a)du + q(a), \tag{A.16}$$

where $q(a) = -\log f_Y(y;a)|_{y=0}$. Here, $q(a)$ is always real-valued due to the following:

$$\begin{aligned} f_Y(y;a)\Big|_{y=0} &= \lim_{y\to 0} \mathbb{E}_X[f_{Y|X}(y|X;a)] \\ &= \mathbb{E}_X\left[\lim_{y\to 0} \frac{1}{\sqrt{2\pi a}} \exp\left(-\frac{1}{2a}(y-X)^2\right)\right] \\ &= \mathbb{E}_X\left[\frac{1}{\sqrt{2\pi a}} \exp\left(-\frac{1}{2a}X^2\right)\right] \\ &\leq \frac{1}{\sqrt{2\pi a}}. \end{aligned} \tag{A.17}$$

The last inequality is due to $\exp(-\frac{1}{2a}X^2) \leq 1$. In addition, equation (A.17) is always greater than zero unless $f_X(x)$ is identical to zero or a is infinite. However, neither case holds. Therefore, $q(a)$ is always mapping to a real-valued number.

Then, the expectation of $g(y;a)$ is expressed as

$$
\begin{aligned}
&\mathbb{E}_Y\left[g(Y;a)\right] \\
&= \int_{-\infty}^{\infty} f_Y(y;a)\left(\int_0^y r(u;a)du + q(a)\right)dy \\
&= \int_0^{\infty}\int_0^y f_Y(y;a)r(u;a)dudy \\
&\qquad + \int_{-\infty}^{0}\int_0^y f_Y(y;a)r(u;a)dudy + q(a) \\
&= \int_0^{\infty}\int_0^y f_Y(y;a)r(u;a)dudy \\
&\qquad - \int_{-\infty}^{0}\int_y^0 f_Y(y;a)r(u;a)dudy + q(a) \\
&= \int_0^{\infty}\left(\int_u^{\infty} f_Y(y;a)dy\right)r(u;a)du \\
&\qquad - \int_{-\infty}^{0}\left(\int_{-\infty}^{u} f_Y(y;a)dy\right)r(u;a)du + q(a) \\
&= \mathbb{E}_X\left[\int_0^{\infty}\left(\int_u^{\infty} f_{Y|X}(y|X;a)dy\right)r(u;a)du\right] \\
&\quad -\mathbb{E}_X\left[\int_{-\infty}^{0}\left(\int_{-\infty}^{u} f_{Y|X}(y|X;a)dy\right)r(u;a)du\right] + q(a) \\
&= \mathbb{E}_X\left[\int_0^{\infty}\left(1-\Phi\left(\frac{u-X}{\sqrt{a}}\right)\right)r(u;a)du\right] \\
&\quad -\mathbb{E}_X\left[\int_{-\infty}^{0}\Phi\left(\frac{u-X}{\sqrt{a}}\right)r(u;a)du\right] + q(a), \qquad \text{(A.18)}
\end{aligned}
$$

where $\Phi(\cdot)$ denotes the standard normal cumulative density function.

We differentiate both sides of equation (A.18) with respect to parameter a as

follows.

$$
\begin{aligned}
\frac{d}{da}\mathbb{E}_Y\left[g(Y;a)\right] &= \frac{d}{da}\mathbb{E}_X\left[\int_0^\infty \left(1-\Phi\left(\frac{u-X}{\sqrt{a}}\right)\right) r(u;a)du\right] \\
&\quad -\mathbb{E}_X\left[\int_{-\infty}^0 \Phi\left(\frac{u-X}{\sqrt{a}}\right) r(u;a)du\right] + \frac{d}{da}q(a) \\
&= -\mathbb{E}_X\left[\int_0^\infty \left(\frac{d}{da}\Phi\left(\frac{u-X}{\sqrt{a}}\right)\right) r(u;a)du\right] \\
&\quad +\mathbb{E}_X\left[\int_0^\infty \left(1-\Phi\left(\frac{u-X}{\sqrt{a}}\right)\right) \frac{d}{da}r(u;a)du\right] \\
&\quad -\mathbb{E}_X\left[\int_{-\infty}^0 \left(\frac{d}{da}\Phi\left(\frac{u-X}{\sqrt{a}}\right)\right) r(u;a)du\right] \\
&\quad -\mathbb{E}_X\left[\int_{-\infty}^0 \Phi\left(\frac{u-X}{\sqrt{a}}\right) \frac{d}{da}r(u;a)du\right] + \frac{d}{da}q(a) \\
&= -\mathbb{E}_X\left[\int_{-\infty}^\infty \frac{d}{da}\Phi\left(\frac{u-X}{\sqrt{a}}\right) r(u;a)du\right] \\
&\quad \underbrace{+\mathbb{E}_X\left[\int_0^\infty \left(1-\Phi\left(\frac{u-X}{\sqrt{a}}\right)\right) \frac{d}{da}r(u;a)du\right]}_{(B)} \\
&\quad \underbrace{-\mathbb{E}_X\left[\int_{-\infty}^0 \Phi\left(\frac{u-X}{\sqrt{a}}\right) \frac{d}{da}r(u;a)du\right]}_{(C)} + \frac{d}{da}q(a). \quad \text{(A.19)}
\end{aligned}
$$

Equations (B) and (C) are further processed as

$$
\begin{aligned}
&\mathbb{E}_X\left[\int_0^\infty \left(1-\Phi\left(\frac{u-X}{\sqrt{a}}\right)\right)\frac{d}{da}r(u;a)du\right] \\
&\quad -\mathbb{E}_X\left[\int_{-\infty}^0 \Phi\left(\frac{u-X}{\sqrt{a}}\right)\frac{d}{da}r(u;a)du\right] \\
=\ &\mathbb{E}_X\left[\int_0^\infty\int_u^\infty f_{Y|X}(y|X;a)dy\frac{d}{da}r(u;a)du\right] \\
&\quad -\mathbb{E}_X\left[\int_{-\infty}^0\int_{-\infty}^u f_{Y|X}(y|X;a)dy\frac{d}{da}r(u;a)du\right] \\
=\ &\mathbb{E}_X\left[\int_0^\infty\int_0^y \frac{d}{da}r(u;a)du f_{Y|X}(y|X;a)dy\right] \\
&\quad -\mathbb{E}_X\left[\int_{-\infty}^0\int_y^0 \frac{d}{da}r(u;a)du f_{Y|X}(y|X;a)dy\right] \\
=\ &\mathbb{E}_X\left[\int_0^\infty\int_0^y \frac{d}{da}r(u;a)du f_{Y|X}(y|X;a)dy\right] \\
&\quad +\mathbb{E}_X\left[\int_{-\infty}^0\int_0^y \frac{d}{da}r(u;a)du f_{Y|X}(y|X;a)dy\right] \\
=\ &\mathbb{E}_X\left[\int_{-\infty}^\infty\int_0^y \frac{d}{da}r(u;a)du f_{Y|X}(y|X;a)dy\right]. \qquad \text{(A.20)}
\end{aligned}
$$

The interchangeability among integrals is due to Fubini's theorem and dominated convergence theorem.

Due to equation (A.16),

$$
\frac{d}{da}g(y;a) \;=\; \frac{d}{da}\int_0^y r(u;a)du + \frac{d}{da}q(a),
$$

equation (A.20) is further simplified as follows:

$$\begin{aligned}
& \mathbb{E}_X \left[\int_{-\infty}^{\infty} \int_0^y \frac{d}{da} r(u;a) du f_{Y|X}(y|X;a) dy \right] \\
= & \int_{-\infty}^{\infty} \left(\frac{d}{da} \int_0^y r(u;a) du \right) f_Y(y;a) dy \\
= & \int_{-\infty}^{\infty} f_Y(y;a) \frac{d}{da} g(y;a) dy - \frac{d}{da} q(a) \\
= & -\int_{-\infty}^{\infty} f_Y(y;a) \frac{d}{da} \log f_Y(y;a) dy - \frac{d}{da} q(a) \qquad \text{(A.21)} \\
= & -\frac{d}{da} q(a).
\end{aligned}$$

The equality in (A.21) holds because $g(y;a) = -\log f_Y(y;a)$.

Therefore, the last three terms in equation (A.19) vanish, and equation (A.19) is expressed as

$$\begin{aligned}
& -\mathbb{E}_X \left[\int_{-\infty}^{\infty} \frac{d}{da} \Phi \left(\frac{u - X}{\sqrt{a}} \right) r(u;a) du \right] \\
= & \mathbb{E}_X \left[\int_{-\infty}^{\infty} \frac{(u - X)}{2a\sqrt{a}} \left[\frac{d}{dy} \Phi (y) \right]_{y = \frac{u-X}{\sqrt{a}}} r(u;a) du \right] \\
= & \mathbb{E}_X \left[\int_{-\infty}^{\infty} \frac{(u - X)}{2a\sqrt{a}} \phi \left(\frac{u - X}{\sqrt{a}} \right) r(u;a) du \right] \\
= & \frac{1}{2} \int_{-\infty}^{\infty} \mathbb{E}_X \left[\frac{(u - X)}{a} \frac{1}{\sqrt{2\pi a}} \exp \left(-\frac{(u - X)^2}{2a} \right) \right] r(u;a) du \\
= & -\frac{1}{2} \int_{-\infty}^{\infty} \mathbb{E}_X \left[\frac{d}{dy} f_{Y|X}(y|X;a) \right] r(u;a) du \\
= & -\frac{1}{2} \int_{-\infty}^{\infty} \frac{\frac{d}{du} f_Y(u;a)}{f_Y(u;a)} r(u;a) f_Y(u;a) du \\
= & \frac{1}{2} \mathbb{E}_Y \left[t(Y;a) r(Y;a) \right],
\end{aligned}$$

where $\phi(\cdot)$ denotes the standard normal probability density function, and $t(y;a) = -(\frac{d}{dy} f_Y(y;a))/f_Y(y;a)$.

Since

$$\begin{aligned}\frac{d}{da}h(Y) &= \frac{d}{da}\mathbb{E}_Y\left[g(Y;a)\right] \\ &= \frac{1}{2}\mathbb{E}_Y\left[t(Y;a)r(Y;a)\right],\end{aligned}$$

and

$$\frac{1}{2}J(Y) = \frac{1}{2}\mathbb{E}_Y\left[\frac{d}{dY}r(Y;a)\right],$$

from De Bruijn's identity, we derive the generalized Stein's identity:

$$\begin{aligned}\frac{d}{da}h(Y) &= \frac{1}{2}J(Y) \\ \iff \mathbb{E}_Y\left[t(Y;a)r(Y;a)\right] &= \mathbb{E}_Y\left[\frac{d}{dY}r(Y;a)\right],\end{aligned}$$

where $\iff$ denotes equivalence between before and after the notation. □

A.2 A Proof of Theorem 2.6

Based on equation (2.16), Theorem 2.6 is proved next using integration by parts and the dominated convergence theorem.

Proof. [Theorem 2.6]

$$\begin{aligned}\frac{d}{da}h(Y) &= -\int_{-\infty}^{\infty}\left(1+\log f_Y(y;a)\right)\frac{d}{da}f_Y(y;a)dy \\ &= -\int_{-\infty}^{\infty}\frac{d}{da}f_Y(y;a)dy - \int_{-\infty}^{\infty}\log f_Y(y;a)\frac{d}{da}f_Y(y;a)dy \qquad \text{(A.22)} \\ &= -\int_{-\infty}^{\infty}\log f_Y(y;a)\frac{d}{da}\mathbb{E}_X\left[f_{Y|X}(y|X;a)\right]dy \\ &= -\int_{-\infty}^{\infty}\log f_Y(y;a)\mathbb{E}_X\left[\frac{d}{da}f_{Y|X}(y|X;a)\right]dy. \qquad \text{(A.23)}\end{aligned}$$

The interchangeability between integral and derivative is due to assumptions (2.17a) and (2.17b).

Using equation (2.16), equation (A.23) is expressed as

$$
\begin{aligned}
&-\int_{-\infty}^{\infty} \log f_Y(y;a)\mathbb{E}_X\left[\frac{d}{da}f_{Y|X}(y|X;a)\right]dy \\
=&\frac{1}{2a}\int_{-\infty}^{\infty} \log f_Y(y;a)\mathbb{E}_X\left[\frac{d}{dy}\left((y-X)f_{Y|X}(y|X;a)\right)\right]dy \\
=&\frac{1}{2a}\int_{-\infty}^{\infty} \log f_Y(y;a)\frac{d}{dy}\mathbb{E}_X\left[(y-X)f_{Y|X}(y|X;a)\right]dy \qquad \text{(A.24)} \\
=&\frac{1}{2a}\log f_Y(y;a)\mathbb{E}_X\left[(y-X)f_{Y|X}(y|X;a)\right]\Big|_{y=-\infty}^{\infty} \\
&-\frac{1}{2a}\int_{-\infty}^{\infty}\frac{d}{dy}\log f_Y(y;a)\mathbb{E}_X\left[(y-X)f_{Y|X}(y|X;a)\right]dy \qquad \text{(A.25)} \\
=&-\frac{1}{2a}\int_{-\infty}^{\infty}\frac{d}{dy}\log f_Y(y;a)\mathbb{E}_X\left[(y-X)f_{Y|X}(y|X;a)\right]dy \qquad \text{(A.26)} \\
=&-\frac{1}{2a}\int_{-\infty}^{\infty}\frac{d}{dy}f_Y(y;a)\mathbb{E}_X\left[(y-X)\frac{f_{Y|X}(y|X;a)}{f_Y(y;a)}\right]dy, \qquad \text{(A.27)}
\end{aligned}
$$

where $f(y)|_{y=a_1}^{a_2}$ denotes $\lim_{y\to a_2} f(y) - \lim_{y\to a_1} f(y)$.

The first term in equation (A.25) vanishes due to the following relationship:

$$
\begin{aligned}
&\log f_Y(y;a)\mathbb{E}_X\left[(y-X)f_{Y|X}(y|X;a)\right]\Big|_{y=-\infty}^{\infty} \\
=\;& yf_Y(y;a)\log f_Y(y;a)\Big|_{y=-\infty}^{\infty} \\
&-\mathbb{E}_X\left[Xf_{Y|X}(y|X;a)\right]\log f_Y(y;a)\Big|_{y=-\infty}^{\infty}. \qquad \text{(A.28)}
\end{aligned}
$$

The first term in (A.28) is expressed as

$$
\begin{aligned}
&yf_Y(y;a)\log f_Y(y;a)\Big|_{y=-\infty}^{\infty} \\
=\;& 2y\sqrt{f_Y(y;a)}\sqrt{f_Y(y;a)}\log\sqrt{f_Y(y;a)}\Big|_{y=-\infty}^{\infty}. \qquad \text{(A.29)}
\end{aligned}
$$

Due to assumptions (2.17d), $y\sqrt{f_Y(y;a)}$ converges to zero as y goes to $\pm\infty$. Since $x \log x$ becomes zero as x goes to zero and $f_Y(y;a)$ converges to zero as y goes to $\pm\infty$, $\sqrt{f_Y(y;a)} \log \sqrt{f_Y(y;a)}$ in (A.29) also becomes zero as y approaches $\pm\infty$.

Similarly, the second term in (A.28) is re-written as

$$\begin{aligned}
&\mathbb{E}_X \left[X f_{Y|X}(y|X;a) \right] \log f_Y(y;a) \Big|_{y=-\infty}^{\infty} \\
&= \underbrace{\frac{\mathbb{E}_X \left[X f_{Y|X}(y|X;a) \right]}{\sqrt{f_Y(y;a)}}}_{(a_1)} 2 \underbrace{\sqrt{f_Y(y;a)} \log \sqrt{f_Y(y;a)}}_{(a_2)} \Bigg|_{y=-\infty}^{\infty} . \qquad \text{(A.30)}
\end{aligned}$$

Since factor (a_2) tends to zero as y approaches $\pm\infty$, and factor (a_1) is bounded due to assumption (2.17d), the right-hand side of equation (A.30) approaches zero as y goes to $\pm\infty$. Therefore, the first term in equation (A.25) is zero, and the equality in (A.26) is verified.

Again, using integration by parts, equation (A.27) is expressed as

$$\begin{aligned}
&-\frac{1}{2a} \int_{-\infty}^{\infty} \frac{d}{dy} f_Y(y;a) \mathbb{E}_X \left[(y - X) \frac{f_{Y|X}(y|X;a)}{f_Y(y;a)} \right] dy \\
&= -\frac{1}{2a} f_Y(y;a) \mathbb{E}_X \left[(y - X) \frac{f_{Y|X}(y|X;a)}{f_Y(y;a)} \right] \Bigg|_{y=-\infty}^{\infty} \\
&\quad + \frac{1}{2a} \int_{-\infty}^{\infty} f_Y(y;a) \frac{d}{dy} \mathbb{E}_X \left[(y - X) \frac{f_{Y|X}(y|X;a)}{f_Y(y;a)} \right] dy \qquad \text{(A.31)} \\
&= \frac{1}{2a} \int_{-\infty}^{\infty} f_Y(y;a) \frac{d}{dy} \mathbb{E}_X \left[(y - X) \frac{f_{Y|X}(y|X;a)}{f_Y(y;a)} \right] dy \qquad \text{(A.32)} \\
&= \frac{1}{2a} \int_{-\infty}^{\infty} f_Y(y;a) \frac{d}{dy} \left(y - \mathbb{E}_X \left[X \frac{f_{Y|X}(y|X;a)}{f_Y(y;a)} \right] \right) dy \\
&= \frac{1}{2a} \left\{ 1 - \mathbb{E}_Y \left[\frac{d}{dY} \mathbb{E}_{X|Y} \left[X|Y \right] \right] \right\} . \qquad \text{(A.33)}
\end{aligned}$$

The equality in (A.32) is verified by the following procedure: the first part of

equation (A.31) is re-written as

$$\begin{aligned} & -\frac{1}{2a} f_Y(y;a)\mathbb{E}_X\left[(y-X)\frac{f_{Y|X}(y|X;a)}{f_Y(y;a)}\right]\Bigg|_{y=-\infty}^{\infty} \\ = & -\frac{1}{2a}\left(y f_Y(y;a) - \mathbb{E}_X\left[X f_{Y|X}(y|X;a)\right]\right)\Bigg|_{y=-\infty}^{\infty} \\ = & \ 0. \end{aligned} \tag{A.34}$$

Due to assumptions (2.17c) and (2.17d), both terms $y f_Y(y;a)$ and $\mathbb{E}_X[X f_{Y|X}(y|X;a)]$ become zero as y goes to $\pm\infty$, and equation (A.34) is zero.

Therefore,

$$\frac{d}{da}h(Y) = \frac{1}{2a}\left\{1 - \mathbb{E}_Y\left[\frac{d}{dY}\mathbb{E}_{X|Y}\left[X|Y\right]\right]\right\},$$

and the proof is completed. □

A.3 A Proof of Theorem 2.7

Proof. [Theorem 2.7]

From equation (A.22), we know

$$\begin{aligned} & \frac{d}{da}h(Y) \\ = & -\int_{-\infty}^{\infty}\frac{d}{da}f_Y(y;a)dy - \int_{-\infty}^{\infty}\log f_Y(y;a)\frac{d}{da}f_Y(y;a)dy \\ = & -\int_{-\infty}^{\infty}\log f_Y(y;a)\frac{d}{da}f_Y(y;a)dy. \end{aligned}$$

Therefore, the second derivative of differential entropy is expressed as

$$\begin{aligned}\frac{d^2}{da^2}h(Y) &= -\int_{-\infty}^{\infty}\frac{d}{da}\log f_Y(y;a)\frac{d}{da}f_Y(y;a)dy - \int_{-\infty}^{\infty}\log f_Y(y;a)\frac{d^2}{da^2}f_Y(y;a)dy, \\ &= -J_a(Y) - \int_{-\infty}^{\infty}\log f_Y(y;a)\frac{d^2}{da^2}f_Y(y;a)dy. \end{aligned} \tag{A.35}$$

The last equality is due to the definition of Fisher information with respect to parameter a in (2.7).

From equation (2.16), we derive an additional relationship between the second order differentials with respect to y and a:

$$\begin{aligned}\frac{d^2}{da^2}f_{Y|X}(y|x;a) &= \frac{d}{da}\left(-\frac{1}{2a}\frac{d}{dy}\left((y-x)f_{Y|X}(y|x;a)\right)\right) \\ &= \frac{1}{2a^2}\frac{d}{dy}\left((y-x)f_{Y|X}(y|x;a)\right) \\ &\quad +\frac{1}{4a^2}\frac{d}{dy}\left((y-x)\left(\frac{d}{dy}\left((y-x)f_{Y|X}(y|x;a)\right)\right)\right).\end{aligned}$$

Since

$$\begin{aligned}&\frac{d^2}{dy^2}\left((y-x)^2 f_{Y|X}(y|x;a)\right) \\ =&\frac{d^2}{dy^2}\left[(y-x)\left((y-x)f_{Y|X}(y|x;a)\right)\right] \\ =&\frac{d}{dy}((y-x)f_{Y|X}(y|x;a)) + \frac{d}{dy}\left((y-x)\frac{d}{dy}((y-x)f_{Y|X}(y|x;a))\right),\end{aligned}$$

we obtain the following relationship:

$$\begin{aligned}\frac{d^2}{da^2}f_{Y|X}(y|x;a) &= \frac{1}{4a^2}\frac{d^2}{dy^2}\left((y-x)^2 f_{Y|X}(y|x;a)\right) \\ &\quad +\frac{1}{4a^2}\frac{d}{dy}\left((y-x)f_{Y|X}(y|x;a)\right).\end{aligned} \tag{A.36}$$

Taking the expected value of both sides of (A.36),

$$\frac{d^2}{da^2} f_Y(y;a) = \frac{1}{4a^2} \left\{ \frac{d^2}{dy^2} \mathbb{E}_X \left[(y-X)^2 f_{Y|X}(y|X;a) \right] + \frac{d}{dy} \mathbb{E}_X \left[(y-X) f_{Y|X}(y|X;a) \right] \right\}. \tag{A.37}$$

After substituting $(d^2 f_Y(y;a)/da^2)$, from equation (A.37), into equation (A.35), the second term of (A.35) takes the expression:

$$\begin{aligned} & -\int_{-\infty}^{\infty} \log f_Y(y;a) \frac{d^2}{da^2} f_Y(y;a) dy \\ = & \underbrace{-\frac{1}{4a^2} \int_{-\infty}^{\infty} \log f_Y(y;a) \frac{d^2}{dy^2} \mathbb{E}_X \left[(y-X)^2 f_{Y|X}(y|X;a) \right] dy}_{(D)} \\ & \underbrace{-\frac{1}{4a^2} \int_{-\infty}^{\infty} \log f_Y(y;a) \frac{d}{dy} \mathbb{E}_X [(y-X) f_{Y|X}(y|X;a)] \, dy}_{(E)} . \end{aligned}$$

Term (E) is exactly of the same form as (A.24), and therefore,

$$\begin{aligned} & -\frac{1}{4a^2} \int_{-\infty}^{\infty} \log f_Y(y;a) \frac{d}{dy} \mathbb{E}_X \left[(y-X) f_{Y|X}(y|X;a) \right] dy \\ = & -\frac{1}{4a^2} \mathbb{E}_Y \left[\frac{d}{dY} \mathbb{E}_{X|Y} \left[Y - X | Y \right] \right] \\ = & -\frac{1}{2a} \frac{d}{da} h(Y). \end{aligned} \tag{A.38}$$

Term (D) is further simplified by the following procedures:

$$\begin{aligned}
&-\frac{1}{4a^2}\int_{-\infty}^{\infty}\log f_Y(y;a)\frac{d^2}{dy^2}\mathbb{E}_X[(y-X)^2 f_{Y|X}(y|X;a)]\,dy \\
=&-\frac{1}{4a^2}\log f_Y(y;a)\frac{d}{dy}\mathbb{E}_X[(y-X)^2 f_{Y|X}(y|X;a)]\Big|_{y=-\infty}^{\infty} \\
&+\frac{1}{4a^2}\int_{-\infty}^{\infty}\frac{d}{dy}\log f_Y(y;a)\frac{d}{dy}\mathbb{E}_X[(y-X)^2 f_{Y|X}(y|X;a)]\,dy.
\end{aligned} \tag{A.39}$$

The first part of (A.39) is expressed as

$$
\begin{aligned}
&-\frac{1}{4a^2}\log f_Y(y;a)\frac{d}{dy}\mathbb{E}_X\big[(y-X)^2 f_{Y|X}(y|X;a)\big]\Big|_{y=-\infty}^{\infty}\\
=&-\frac{1}{4a^2}\log f_Y(y;a)\Bigg(\mathbb{E}_X[2(y-X)f_{Y|X}(y|X;a)]\\
&\quad+\mathbb{E}_X\bigg[(y^2-2Xy+X^2)\frac{d}{dy}f_{Y|X}(y|X;a)\bigg]\Bigg)\Bigg|_{y=-\infty}^{\infty}\\
=&-\frac{1}{4a^2}\log f_Y(y;a)\Bigg(2yf_Y(y;a)-2\mathbb{E}_X[Xf_{Y|X}(y|X;a)]\\
&\quad+y^2\frac{d}{dy}f_Y(y;a)-2y\mathbb{E}_X\bigg[X\frac{d}{dy}f_{Y|X}(y|X;a)\bigg]\\
&\quad+\mathbb{E}_X\bigg[X^2\frac{d}{dy}f_{Y|X}(y|X;a)\bigg]\Bigg)\Bigg|_{y=-\infty}^{\infty}\\
=&-\frac{1}{2a^2}\underbrace{\sqrt{f_Y(y;a)}\log\sqrt{f_Y(y;a)}}_{(b_1)}\Bigg(2\underbrace{y\sqrt{f_Y(y;a)}}_{(b_2)}+\mathbb{E}_X\Bigg[\underbrace{X^2\frac{\frac{d}{dy}f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}}_{(b_3)}\Bigg]\Bigg)\\
&-\frac{1}{a^2}\underbrace{\sqrt[4]{f_Y(y;a)}\log\sqrt[4]{f_Y(y;a)}}_{(b_1)}\times\Bigg(\underbrace{y^2\sqrt[4]{f_Y(y;a)}}_{(b_2)}\underbrace{\mathbb{E}_X\Bigg[\frac{\frac{d}{dy}f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}\Bigg]}_{(b_3)}\\
&\quad-2\underbrace{y\sqrt[4]{f_Y(y;a)}}_{(b_2)}\mathbb{E}_X\Bigg[\underbrace{X\frac{\frac{d}{dy}f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}}_{(b_3)}\Bigg]\Bigg)\\
&+\frac{1}{a^2}\underbrace{\sqrt{f_Y(y;a)}\log\sqrt{f_Y(y;a)}}_{(b_1)}\underbrace{\frac{\mathbb{E}_X[Xf_{Y|X}(y|X;a)]}{\sqrt{f_Y(y;a)}}}_{(b_4)}\Bigg|_{y=-\infty}^{\infty}.
\end{aligned}
$$

Since $x\log x$ becomes zero as x approaches zero and $f_Y(y;a)$ converges to zero as y goes to $\pm\infty$, factor (b_1) is zero as $y\to\pm\infty$. Due to assumptions (2.19c) and (2.19d), term (b_2) becomes zero as $y\to\pm\infty$ and term (b_3) is bounded. Also, factor (b_4) must be bounded due to assumption (2.19e). Therefore, as $y\to\pm\infty$, the first part of

equation (A.39) vanishes.

Then, equation (A.39) is further processed using integration by parts as follows:

$$\begin{aligned}
&\frac{1}{4a^2}\int_{-\infty}^{\infty}\frac{d}{dy}\log f_Y(y;a)\frac{d}{dy}\mathbb{E}_X[(y-X)^2 f_{Y|X}(y|X;a)]\,dy \\
=&\frac{1}{4a^2}\frac{d}{dy}\log f_Y(y;a)\mathbb{E}_X[(y-X)^2 f_{Y|X}(y|X;a)]\,\Big|_{y=-\infty}^{\infty} \\
&-\frac{1}{4a^2}\int_{-\infty}^{\infty}\frac{d^2}{dy^2}\log f_Y(y;a)\mathbb{E}_X[(y-X)^2 f_{Y|X}(y|X;a)]\,dy. \qquad \text{(A.40)}
\end{aligned}$$

Again, the first part of equation (A.40) is re-written as

$$\begin{aligned}
&\frac{1}{4a^2}\frac{d}{dy}\log f_Y(y;a)\mathbb{E}_X[(y-X)^2 f_{Y|X}(y|X;a)]\,\Big|_{y=-\infty}^{\infty} \\
=&\frac{1}{4a^2}\mathbb{E}_X\left[\frac{\frac{d}{dy}f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}\right]\mathbb{E}_X\left[(y-X)^2\frac{f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}\right]\Bigg|_{y=-\infty}^{\infty} \\
=&\frac{1}{4a^2}\underbrace{\mathbb{E}_X\left[\frac{\frac{d}{dy}f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}\right]}_{(c_1)}\underbrace{y^2\sqrt{f_Y(y;a)}}_{(c_2)} \\
&-2\frac{1}{4a^2}\underbrace{\mathbb{E}_X\left[\frac{\frac{d}{dy}f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}\right]}_{(c_1)}\underbrace{y\sqrt[4]{f_Y(y;a)}}_{(c_2)}\underbrace{\mathbb{E}_X\left[X\frac{f_{Y|X}(y|X;a)}{(f_Y(y;a))^{3/4}}\right]}_{(c_3)} \\
&+\frac{1}{4a^2}\underbrace{\mathbb{E}_X\left[\frac{\frac{d}{dy}f_{Y|X}(y|X;a)}{\sqrt{f_Y(y;a)}}\right]}_{(c_1)}\underbrace{\sqrt[4]{f_Y(y;a)}}_{(c_2)}\underbrace{\mathbb{E}_X\left[X^2\frac{f_{Y|X}(y|X;a)}{(f_Y(y;a))^{3/4}}\right]}_{(c_3)}\Bigg|_{y=-\infty}^{\infty}. \qquad \text{(A.41)}
\end{aligned}$$

Factors (c_1) and (c_3) are bounded due to assumptions (2.19c) and (2.19e), and, by assumption (2.19d), factor (c_2) approaches zero as $y \to \pm\infty$. Then, equation (A.40)

is expressed as

$$\begin{aligned}
&\frac{1}{4a^2}\int_{-\infty}^{\infty}\frac{d}{dy}\log f_Y(y;a)\frac{d}{dy}\mathbb{E}_X\big[(y-X)^2 f_{Y|X}(y|X;a)\big]\,dy \\
=&-\frac{1}{4a^2}\int_{-\infty}^{\infty}\frac{d^2}{dy^2}\log f_Y(y;a)\mathbb{E}_X\big[(y-X)^2 f_{Y|X}(y|X;a)\big]\,dy. \qquad \text{(A.42)}
\end{aligned}$$

Using equations (A.38) and (A.42), equation (A.35) is expressed as

$$\begin{aligned}
\frac{d^2}{da^2}h(Y) &= -J_a(Y) - \int_{-\infty}^{\infty}\log f_Y(y;a)\frac{d^2}{da^2}f_Y(y;a)dy \\
&= -J_a(Y) - \frac{1}{2a}\frac{d}{da}h(Y) - \frac{1}{4a^2}\mathbb{E}_Y\left[\frac{d}{dY}S_Y(Y)\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]\right] \\
&= -J_a(Y) - \frac{1}{4a^2}\mathbb{E}_Y\left[\frac{d}{dY}\mathbb{E}_{X|Y}\left[(Y-X)|Y\right]\right] \\
&\quad -\frac{1}{4a^2}\mathbb{E}_Y\left[\frac{d}{dY}S_Y(Y)\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]\right],
\end{aligned}$$

and the proof is completed. □

A.4 A proof of Lemma 2.1

Proof. [Lemma 2.1]

Before we prove this lemma, we first introduce two lemmas which are necessary to prove Lemma 2.1.

Lemma A.2. *Given the channel $Y = X + \sqrt{a}W$ in (2.1), the following identity holds:*

$$\frac{d}{da}J(Y) \;=\; -\mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)^2\right], \qquad \text{(A.43)}$$

where X is an arbitrary but fixed random variable with a finite second-order moment, and W is a Gaussian random variable with zero mean and unit variance.

Proof. In Theorems 2.4, 2.5, we showed the equivalence among De Bruijn, generalized Stein, and heat equation identities for specific conditions. Therefore, using one of the identities, this lemma can be proved. In this proof, Theorem 2.3 (the heat equation identity) will be used with $g(y) = S_Y(y)^2$. Unlike the definition of $g(y)$ in Theorem 2.3, $g(y)$ is dependent on the parameter a. Therefore, we use the notation $g(y;a)$ instead of $g(y)$. Since $J(Y) = \mathbb{E}[S_Y(Y)^2]$, the right-hand side of (A.43) is expressed as

$$\begin{aligned}\frac{d}{da}J(Y) &= \frac{d}{da}\mathbb{E}_Y\left[S_Y(Y)^2\right] \\ &= \int_{-\infty}^{\infty}\frac{d}{da}f_Y(y;a)g(y;a)dy + \mathbb{E}_Y\left[\frac{d}{da}g(Y;a)\right]. \end{aligned} \tag{A.44}$$

By the heat equation identity, the first term in equation (A.44) is expressed as

$$\int_{-\infty}^{\infty}\frac{d}{da}f_Y(y;a)g(y;a)dy \;=\; \frac{1}{2}\mathbb{E}_Y\left[\frac{d^2}{dY^2}g(Y;a)\right].$$

Using integration by parts, the second term in equation (A.44) is expressed as

$$\mathbb{E}_Y\left[\frac{d}{da}g(Y;a)\right]\frac{1}{2}\mathbb{E}_Y\left[\frac{d^2}{dY^2}g(Y;a)\right] - \mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)^2\right] + 2\mathbb{E}_Y\left[S_Y(Y)^2\frac{d}{dY}S_Y(Y)\right].$$

Therefore, equation (A.44) takes the form:

$$\begin{aligned}&\int_{-\infty}^{\infty}\frac{d}{da}f_Y(y;a)g(y;a)dy + \mathbb{E}_Y\left[\frac{d}{da}g(Y;a)\right] \\ =\;& -\mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)^2\right] + \underbrace{\mathbb{E}_Y\left[\frac{d^2}{dY^2}g(Y;a)\right] + 2\mathbb{E}_Y\left[S_Y(Y)^2\frac{d}{dY}S_Y(Y)\right]}_{(F)}.\end{aligned}$$

Performing an integration by parts, the term (F) is shown to be equal to zero, and

the proof is completed.

Remark A.1. *A vector version of this lemma was reported in [42]. The reasons why we introduce both this lemma and its proof are not only to present alternative proofs, but also to explain the usefulness of our novel results. For example, Lemma A.2 was proved based on the heat equation identity, which is a novel approach to prove this lemma. At the same time, this lemma can also be alternatively proved using Theorem 2.7 or Corollary 2.4.*

□

Lemma A.3 (Fisher Information Inequality)**.** *Consider the channel* $Y = X + \sqrt{a}W$ *in (2.1), where the random variable* X *is assumed to have an arbitrary distribution but a fixed second-order moment and* W *is normally distributed with zero mean and unit variance. Then, the following inequality is always satisfied:*

$$\frac{1}{J(Y)} \geq \frac{1}{J(X)} + \frac{1}{J(\sqrt{a}W)},$$

where the equality holds if and only if X *is normally distributed.*

Proof. Using Lemma A.2 (equivalently, Theorem 2.7 or Corollary 2.4 can be used),

$$\begin{aligned} -\frac{d}{da}J(Y) &= \mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)^2\right] \\ &\geq \mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)\right]^2 \\ &= J(Y)^2. \end{aligned} \tag{A.45}$$

Equation (A.45) is expressed as

$$-\frac{d}{da}J(Y) \geq J(Y)^2,$$

and it is equivalent to

$$-\frac{\frac{d}{da}J(Y)}{J(Y)^2} \geq 1 \quad \Longleftrightarrow \quad \frac{d}{da}\left(\frac{1}{J(Y)}\right) \geq 1. \tag{A.46}$$

Since inequality (A.46) is satisfied for any a,

$$\begin{aligned} & \int_0^a \frac{d}{dt}\left(\frac{1}{J(Y)}\right) dt \geq \int_0^a 1 dt, \\ \Longleftrightarrow \quad & \frac{1}{J(Y)} - \frac{1}{J(X)} \geq a, \\ \Longleftrightarrow \quad & \frac{1}{J(Y)} \geq \frac{1}{J(X)} + \frac{1}{J(\sqrt{a}W)}. \end{aligned} \tag{A.47}$$

Since W is normally distributed with unit variance, $a = 1/J(\sqrt{a}W)$, and the last equivalence holds. The last equation in (A.47) denotes the Fisher information inequality, and the proof is completed.

Remark A.2. *This proof uses neither the convolutional inequality, the data processing inequality, nor the EPI, unlike previous proofs. The proof only relies on De Bruijn's identity, Stein's identity, or the heat equation identity. Namely, Theorem 2.1, 2.2, 2.3, or 2.7 is the only adopted result, and Theorems 2.4, 2.5 ensure Theorem 2.1, 2.2, 2.3, or 2.7 can be equivalently adopted to the proof. Even though Lemma A.2 was used in this proof, Lemma A.2 itself was also proved using one of the above identities. Therefore, this proof only uses our results.*

□

Now, based on Lemma A.3, the proof of Lemma 2.1 is straightforward. From

Lemma A.3,

$$\begin{aligned} & \frac{1}{J(Y)} \geq \frac{1}{J(X)} + \frac{1}{J(\sqrt{a}W)} \\ \iff \quad & J(Y) \leq \frac{J(X)J(\sqrt{a}W)}{J(X) + J(\sqrt{a}W)}. \end{aligned} \tag{A.48}$$

Since X and W are independent, and W is normally distributed,

$$\begin{aligned} \mathbb{E}_X\left[J(Y|X)\right] &= \int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{\infty} \left(\frac{d}{dx} \log f_{Y|X}(y|x;a) \right)^2 f_{Y|X}(y|x;a) dy dx \\ &= \int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{\infty} \frac{1}{a^2} (y-x)^2 f_{Y|X}(y|x;a) dy dx \\ &= \frac{1}{a} \\ &= J(\sqrt{a}W). \end{aligned} \tag{A.49}$$

The equality in (A.49) is due to $\mathbb{E}_{Y|X}[(Y-X)^2|X=x] = a$.

For a Gaussian random variable W,

$$J(Y) \;=\; \frac{1}{a} - \frac{1}{a^2} Var(X|Y), \tag{A.50}$$

where $Var(X|Y)$ stands for $\mathbb{E}_{X,Y}[(X - \mathbb{E}_{X|Y}[X|Y])^2]$ ([18], [43]).

Substituting $Var(X|Y)$ and $\mathbb{E}_X[J(Y|X)]$ for $J(Y)$ and $J(\sqrt{a}W)$, respectively,

equation (A.48) is expressed as

$$\begin{aligned}
& J(Y) \leq \frac{J(X)J(\sqrt{a}W)}{J(X)+J(\sqrt{a}W)}, \\
\iff\ & \frac{1}{a}-\frac{1}{a^2}Var(X|Y) \leq \frac{J(X)J(\sqrt{a}W)}{J(X)+J(\sqrt{a}W)}, \\
\iff\ & Var(X|Y) \geq \frac{1}{J(X)+J(\sqrt{a}W)}, \\
\iff\ & Var(X|Y) \geq \frac{1}{J(X)+\mathbb{E}_X\left[J(Y|X)\right]}.
\end{aligned}$$

Since $Var(X|Y)$ is equal to the minimum mean square error,

$$\begin{aligned}
MSE(\hat{X}) &\geq MMSE(\hat{X}) \\
&= Var(X|Y) \\
&\geq \frac{1}{J(X)+\mathbb{E}_X\left[J(Y|X)\right]},
\end{aligned}$$

where $\hat{X}$ denotes a Bayesian estimator, and the obtained inequality is the Bayesian Cramér-Rao lower bound (BCRLB). □

A.5 A Proof of Lemma 2.2

Proof. [Lemma 2.2]

When a is zero, the right-hand side of (2.22) is zero due to the following relations:

$$\begin{aligned}
N(X|Y) &= \frac{1}{2\pi e}\exp(2h(X|Y)) \\
&= \frac{1}{2\pi e}\exp(2(h(X)+h(Y|X)-h(Y))) \\
&= \frac{1}{2\pi e}\exp(2(h(X)+h(\sqrt{a}W)-h(Y))) \\
&= \frac{N(X)N(\sqrt{a}W)}{N(Y)} \\
&= \frac{aN(X)N(W)}{N(X+\sqrt{a}W)}.
\end{aligned}$$

Therefore, when a goes to zero,

$$\begin{aligned} \lim_{a\to 0} N(X|Y) &= \lim_{a\to 0} \frac{aN(X)N(W)}{N(X+\sqrt{a}W)} \\ &= 0. \end{aligned} \tag{A.51}$$

The equality is due to the fact that $\lim_{a\to 0} N(X+\sqrt{a}W) = N(X)$. Since the left-hand side of (2.22) is always greater than or equal to zero, the inequality in (2.22) is satisfied when a is zero.

Without loss of generality, from now on, we assume that $a > 0$.

Since $h(X|Y) = h(X) + h(Y|X) - h(Y)$, by Theorem 2.1 (De Bruijn's identity),

$$\begin{aligned} \frac{d}{da}N(X|Y) &= \frac{d}{da}\left(\frac{1}{2\pi e}\exp\left(2h(X|Y)\right)\right) \\ &= 2N(X|Y)\left\{\frac{d}{da}h(X) + \frac{d}{da}h(Y|X) - \frac{d}{da}h(Y)\right\} \\ &= 2N(X|Y)\left\{\frac{1}{2a} - \frac{1}{2}J(Y)\right\} \qquad \text{(A.52)} \\ &= N(X|Y)\frac{1}{a^2}Var(X|Y). \qquad \text{(A.53)} \end{aligned}$$

Since $h(X)$ is independent of a and $h(Y|X) = h(\sqrt{a}W)$, $(d/da)h(X)$ is zero, and $(d/da)h(Y|X) = 1/2a$. Therefore, the equality in (A.52) is satisfied. The equality in (A.53) is due to equation (A.50).

Based on equation (A.50),

$$\begin{aligned} \frac{d}{da}Var(X|Y) &= \frac{d}{da}\left[a - a^2 J(Y)\right] \\ &= \frac{d}{da}\left[a - a^2\left(2\frac{d}{da}h(Y)\right)\right]. \end{aligned} \tag{A.54}$$

The equality in (A.54) is due to Theorem 2.1.

Using Corollary 2.4 and equation (A.50), equation (A.54) is further processed as

$$\begin{aligned}
\frac{d}{da}\left[a - a^2\left(2\frac{d}{da}h(Y)\right)\right] &= 1 - 2a\left(2\frac{d}{da}h(Y)\right) + a^2\left(-2\frac{d^2}{da^2}h(Y)\right) \\
&= 1 - 2aJ(Y) + a^2\mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)^2\right] \qquad \text{(A.55)} \\
&\geq 1 - 2aJ(Y) + a^2J(Y)^2 \qquad \text{(A.56)} \\
&= (1 - aJ(Y))^2 \\
&= \frac{1}{a^2}Var(X|Y)^2.
\end{aligned}$$

The equality in (A.55) is due to Theorem 2.1 and Corollary 2.4, and the inequality in (A.56) holds because

$$\begin{aligned}
\mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)^2\right] &\geq \left(\mathbb{E}_Y\left[\frac{d}{dY}S_Y(Y)\right]\right)^2 \\
&= J(Y)^2.
\end{aligned}$$

Therefore,

$$\frac{d}{da}Var(X|Y) \geq \frac{1}{a^2}Var(X|Y)^2. \qquad \text{(A.57)}$$

Using equations (A.53) and (A.57), we obtain the following inequality:

$$\frac{d}{da}\log N(X|Y) \leq \frac{d}{da}\log Var(X|Y).$$

Since $N(X_G|Y_G) = Var(X_G|Y_G)$, where X_G and Y_G denote Gaussian random variables whose variances are equal to X and Y, respectively, the following inequality

also holds:

$$\frac{d}{da}\left(\log N(X_G|Y_G) - \log N(X|Y)\right) \geq \frac{d}{da}\left(\log Var(X_G|Y_G) - \log Var(X|Y)\right). \quad \text{(A.58)}$$

By performing an integration, from 0 to a, of both sides in (A.58), equation (A.58) is expressed as

$$\begin{aligned}
& \int_0^a \frac{d}{dt}\left(\log N_t(X_G|Y_G) - \log N_t(X|Y)\right) dt \\
& \quad \geq \int_0^a \frac{d}{dt}\left(\log Var_t(X_G|Y_G) - \log Var_t(X|Y)\right) dt \\
\Leftrightarrow \quad & \log N_t(X_G|Y_G) - \log N_t(X|Y)\Big|_{t=0}^{a} \\
& \quad \geq \log Var_t(X_G|Y_G) - \log Var_t(X|Y)\Big|_{t=0}^{a} \\
\Leftrightarrow \quad & \log N_a(X_G|Y_G) - \log N_a(X|Y) \\
& \quad - \lim_{t\to 0}\left(\log N_t(X_G|Y_G) - \log N_t(X|Y)\right) \\
& \quad \geq \log Var_a(X_G|Y_G) - \log Var_a(X|Y) \\
& \quad - \lim_{t\to 0}\left(\log Var_t(X|Y) - \log Var_t(X_G|Y_G)\right) \qquad \text{(A.59)} \\
\Leftrightarrow \quad & \log N_a(X|Y) \leq \log Var_a(X|Y), \qquad \text{(A.60)}
\end{aligned}$$

where $\Leftrightarrow$ stands for equivalence between before and after the notation, subscript t or a denotes that a function depends on a parameter t or a, respectively (the subscript is only used when there may be a confusion between an actual parameter variable and a dummy variable).

The equivalence in (A.60) is due to the following: $N_a(X_G|Y_G) = Var_a(X_G|Y_G)$,

$$\begin{aligned}
& \lim_{t\to 0} \left(\log N_t(X_G|Y_G) - \log N_t(X|Y)\right) \\
= & \lim_{t\to 0} \log \frac{N_t(X_G|Y_G)}{N_t(X|Y)} \\
= & \lim_{t\to 0} \log \left(\frac{N(X_G)N_t(Y_G|X_G)}{N_t(Y_G)} \Big/ \frac{N(X)N_t(Y|X)}{N_t(Y)} \right) \\
= & \lim_{t\to 0} \log \left(\frac{N(X_G)N(\sqrt{t}W)}{N(X_G+\sqrt{t}W)} \Big/ \frac{N(X)N(\sqrt{t}W)}{N(X+\sqrt{t}W)} \right) \\
= & \lim_{t\to 0} \log \left(\frac{N(X_G)N(X+\sqrt{t}W)}{N(X)N(X_G+\sqrt{t}W)} \right) \\
= & \log \left(\frac{N(X_G)N(X)}{N(X)N(X_G)} \right) \\
= & \ 0, \qquad \text{(A.61)}
\end{aligned}$$

and

$$\begin{aligned}
& \lim_{t\to 0} \left(\log Var_t(X_G|Y_G) - \log Var_t(X|Y)\right) \\
= & \lim_{t\to 0} \Big(\log\Big(t - t^2 J(X_G + \sqrt{t}W)\Big) - \log\Big(t - t^2 J(X + \sqrt{t}W)\Big)\Big) \qquad \text{(A.62)} \\
= & \lim_{t\to 0} \Big(\log\Big(1 - tJ(X_G + \sqrt{t}W)\Big) - \log\Big(1 - tJ(X + \sqrt{t}W)\Big)\Big) \\
= & \log(1) - \log(1) \\
= & \ 0,
\end{aligned}$$

where W is a Gaussian random variable. The equality in (A.62) is due to equation (A.50).

Since $\log x$ is an increasing function with respect to x, equation (A.60) is equiv-

alent to

$$N(X|Y) \;\leq\; Var(X|Y),$$

and the proof is completed. □

A.6 A Proof of Lemma 2.3

Proof. [Lemma 2.3]

When $a = 0$, both sides of the inequality in (2.25) are zero, and the inequality in (2.25) is satisfied. Therefore, without loss of generality, we assume that $a > 0$.

$$\begin{aligned}
\frac{d}{da}\log N(X|Y) &= \frac{1}{N(X|Y)}\frac{d}{da}N(X|Y) \\
&= \frac{1}{a^2}Var(X|Y) && \text{(A.63)} \\
&\geq \frac{1}{a^2}\frac{1}{J(X)+J(\sqrt{a}W)} && \text{(A.64)} \\
&= \frac{d}{da}\log\left(\frac{1}{J(X)+J(\sqrt{a}W)}\right),
\end{aligned}$$

where W is a Gaussian random variable with zero mean and unit variance. The equality in (A.63) is due to equation (A.53), the inequality in (A.64) is because of BCRLB.

Since $N(X_G|Y_G)$ is equal to $1/(J(X_G)+J(\sqrt{a}W))$, where X_G and Y_G are Gaussian random variables whose variances are equal to X and Y, respectively, the following inequality is satisfied:

$$\begin{aligned}
&\frac{d}{da}(\log N(X_G|Y_G) - \log N(X|Y))) \\
\leq &\frac{d}{da}\left(\log\frac{1}{J(X_G)+J(\sqrt{a}W)} - \log\frac{1}{J(X)+J(\sqrt{a}W)}\right). && \text{(A.65)}
\end{aligned}$$

By integrating both sides in (A.65), equation (A.65) is equivalent to the following:

$$
\begin{aligned}
&\int_0^a \frac{d}{dt}\left(\log N_t(X_G|Y_G) - \log N_t(X|Y)\right) dt \\
&\leq \int_0^a \frac{d}{dt}\left(\log\frac{1}{J(X_G)+J(\sqrt{t}W)} - \log\frac{1}{J(X)+J(\sqrt{t}W)}\right) dt \\
\Leftrightarrow\quad & \log N_a(X_G|Y_G) - \log N_a(X|Y) - \lim_{t\to 0}\left(\log N_t(X_G|Y_G) - \log N_t(X|Y)\right) \\
&\leq \log\frac{1}{J(X_G)+J(\sqrt{a}W)} - \log\frac{1}{J(X)+J(\sqrt{a}W)} \\
&\quad - \lim_{t\to 0}\left(\log\frac{1}{J(X_G)+J(\sqrt{t}W)} - \log\frac{1}{J(X)+J(\sqrt{t}W)}\right) \\
\Leftrightarrow\quad & \log N(X|Y) \geq \log\frac{1}{J(X)+J(\sqrt{a}W)},
\end{aligned}
\tag{A.66}
$$

where $\Leftrightarrow$ denotes the equivalence between before and after the notation, and subscript a or t of a function means dependency of the function with respect to a or t, respectively. The equivalence in (A.66) is due to the following: $N(X_G|Y_G)$ is equal to $1/(J(X_G)+J(\sqrt{a}W))$, and

$$
\begin{aligned}
& \lim_{t\to 0}\left(\log\frac{1}{J(X_G)+J(\sqrt{t}W)} - \log\frac{1}{J(X)+J(\sqrt{t}W)}\right) \\
=\ & \lim_{t\to 0}\left(\log\frac{t}{tJ(X_G)+J(W)} - \log\frac{t}{tJ(X)+J(W)}\right) \\
=\ & \lim_{t\to 0}\log\frac{tJ(X)+J(W)}{tJ(X_G)+J(W)} \\
=\ & \log\frac{J(W)}{J(W)} \\
=\ & 0,
\end{aligned}
\tag{A.67}
$$

and

$$
\lim_{t\to 0}\left(\log N_t(X_G|Y_G) - \log N_t(X|Y)\right) = 0
$$

due to equation (A.61).

Since $\log x$ is a increasing function with respect to x, the inequality in (A.66) is equivalent to

$$N(X|Y) \geq \frac{1}{J(X) + J(\sqrt{a}W)}. \tag{A.68}$$

Since we have already proved that $N(X|Y)$ is a lower bound for any Bayesian estimator in Lemma 2.2, the inequality in (A.68) means that the lower bound $N(X|Y)$, the left-hand side of (A.68), is tighter than BCRLB, the right-hand side of (A.68). □

A.7 A Proof of Lemma 2.4 (Costa's EPI)

Proof. [Lemma 2.4]

The proof will be conducted in two different ways.

1. Instead of proving equation (2.26), we are going to prove the inequality in (2.27).

   Using De Bruijn's identity,

$$\begin{aligned} \frac{d^2}{da^2}N(Y) &= 2\frac{d}{da}N(Y)\frac{d}{da}h(Y) + 2N(Y)\frac{d^2}{da^2}h(Y), \\ &= N(Y)\left(J(Y)^2 + 2\frac{d^2}{da^2}h(Y)\right), \end{aligned}$$

   where $Y = X + \sqrt{a}W$. Since $N(Y) \geq 0$, proving the inequality in (2.27) is equivalent to proving the following inequality:

$$J(Y)^2 + 2\frac{d^2}{da^2}h(Y) \leq 0. \tag{A.69}$$

Using Theorem 2.7, the inequality in (A.69) is expressed as

$$J(Y)^2 - 2J_a(Y) - \frac{1}{2a^2}\mathbb{E}_Y\left[\frac{d}{dY}\mathbb{E}_{X|Y}\left[Y-X|Y\right]\right] - \frac{1}{2a^2}\mathbb{E}_Y\left[\frac{d}{dY}S_Y(Y)\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]\right] \leq 0. \tag{A.70}$$

By Corollary 2.4, equation (A.70) is equivalent to

$$\begin{aligned} &J(Y)^2 - 2J_a(Y) - \frac{1}{2a^2}\mathbb{E}_Y\left[\frac{d}{dY}\mathbb{E}_{X|Y}\left[Y-X|Y\right]\right] \\ &\quad - \frac{1}{2a^2}\mathbb{E}_Y\left[\frac{d}{dY}S_Y(Y)\mathbb{E}_{X|Y}\left[(Y-X)^2|Y\right]\right] \\ &= J(Y)^2 - \mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)^2\right] \\ &= -\mathbb{E}_Y\left[\left(J(Y) + \frac{d}{dY}S_Y(Y)\right)^2\right] \\ &\leq 0. \end{aligned} \tag{A.71}$$

Since $J(Y) = -\mathbb{E}[(d/dY)S_Y(Y)]$ and $\mathbb{E}[S_Y(Y)] = 0$, the equality holds in (A.71). Therefore,

$$\begin{aligned} \frac{d^2}{da^2}N(Y) &= -\mathbb{E}_Y\left[\left(J(Y) + \frac{d}{dY}S_Y(Y)\right)^2\right], \\ &\leq 0, \end{aligned}$$

and the proof is completed.

Remark A.3. *This proof mostly follows the proof in [52]. However, by using Theorem 2.7 to prove Costa's EPI, we show that Costa's EPI can be proved by De Bruijn-like identity without using the Fisher information inequality.*

2. In the second proof, the inequality (2.27) is proved by a slightly different

method.

First, define a function $l(a)$ as follows:

$$l(a) = -\frac{J(X)}{1+aJ(X)} + J(Y), \tag{A.72}$$

where $Y = X + \sqrt{a}W$, X is an arbitrary but fixed random variable, W is a Gaussian random variable, and X and W are independent of each other.

For arbitrary non-negative real-valued a, $l(a) \leq 0$, and it is proved by the following procedure; using Lemma A.2 (Theorem 2.7 or Corollary 2.4 can be used instead of Lemma A.2),

$$\begin{aligned} -\frac{d}{da}J(Y) &= \mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)^2\right] \\ &\geq \mathbb{E}_Y\left[\left(\frac{d}{dY}S_Y(Y)\right)\right]^2 \\ &= J(Y)^2. \end{aligned} \tag{A.73}$$

Equation (A.73) is equivalent to the following inequalities:

$$\begin{aligned} & -\frac{\frac{d}{da}J(Y)}{J(Y)^2} \geq 1 \\ \iff & \frac{d}{da}\left(\frac{1}{J(Y)}\right) \geq 1. \end{aligned} \tag{A.74}$$

Since inequality (A.74) is satisfied for arbitrary non-negative real-valued a,

$$\begin{aligned} & \int_0^a \frac{d}{dt}\left(\frac{1}{J(Y)}\right) dt \geq \int_0^a 1 dt \\ \iff & \frac{1}{J(Y)} - \frac{1}{J(X)} \geq a \\ \iff & J(Y) \leq \frac{J(X)}{1 + aJ(X)}, \end{aligned} \tag{A.75}$$

and therefore, equation (A.72) is always non-positive.

Since $J(Y)$ converges to $J(X)$ as a approaches zero, $l(0) = 0$, and the following inequality holds for an arbitrary but fixed random variable X and arbitrary small non-negative real-valued ϵ:

$$l(\epsilon) - l(0) = -\frac{J(X)}{1 + \epsilon J(X)} + J(X + \sqrt{\epsilon}W) \tag{A.76}$$

$$\leq 0. \tag{A.77}$$

Therefore,

$$\left.\frac{d}{d\epsilon} l(\epsilon)\right|_{\epsilon=0} \leq 0, \tag{A.78}$$

for an arbitrary but fixed random variable X.

Since the inequality in (A.78) holds for an arbitrary random variable X, we define X as $\tilde{X} + \sqrt{a}\tilde{W}$, where $\tilde{X}$ is an arbitrary but fixed random variable, $\tilde{W}$ is a Gaussian random variable whose variance is identical to the variance of W, and $\tilde{X}$, $\tilde{W}$, and W are independent of one another. Then, the inequality

in (A.78) is equivalent to the following inequalities:

$$
\begin{aligned}
0 &\geq \left(\frac{J(\tilde{X}+\sqrt{a}\tilde{W})}{1+\epsilon J(\tilde{X}+\sqrt{a}\tilde{W})}\right)^2\Bigg|_{\epsilon=0} + \frac{d}{d\epsilon}J(\tilde{X}+\sqrt{a}\tilde{W}+\sqrt{\epsilon}W)\Bigg|_{\epsilon=0} \\
\Leftrightarrow 0 &\geq \left(\frac{J(\tilde{X}+\sqrt{a}\tilde{W})}{1+\epsilon J(\tilde{X}+\sqrt{a}\tilde{W})}\right)^2\Bigg|_{\epsilon=0} + \frac{d}{d\epsilon}J(\tilde{X}+\sqrt{a+\epsilon}\tilde{W})\Bigg|_{\epsilon=0} \qquad \text{(A.79)} \\
\Leftrightarrow 0 &\geq \left(\frac{J(\tilde{X}+\sqrt{a}\tilde{W})}{1+\epsilon J(\tilde{X}+\sqrt{a}\tilde{W})}\right)^2\Bigg|_{\epsilon=0} + \frac{d}{da}J(\tilde{X}+\sqrt{a+\epsilon}\tilde{W})\Bigg|_{\epsilon=0} \qquad \text{(A.80)} \\
\Leftrightarrow 0 &\geq J(\tilde{X}+\sqrt{a}\tilde{W})^2 + \frac{d}{da}J(\tilde{X}+\sqrt{a}\tilde{W}), \qquad \text{(A.81)}
\end{aligned}
$$

where $\Leftrightarrow$ denotes the equivalence between before and after the notation. The equivalence in (A.79) is due to the fact that $J(\tilde{X}+\sqrt{a}\tilde{W}+\sqrt{\epsilon}W) = J(\tilde{X}+\sqrt{a+\epsilon}\tilde{W})$ for independent Gaussian random variables W and $\tilde{W}$ whose variances are identical to each other. The inequality in (A.80) holds due to the following procedure: first, the Fisher information $J(\tilde{X}+\sqrt{a+\epsilon}\tilde{W})$ is expressed as

$$
\begin{aligned}
& J(\tilde{X}+\sqrt{a+\epsilon}\tilde{W}) \\
=& \int_{-\infty}^{\infty} \frac{d}{dy} f_Y(y;a,\epsilon) \frac{d}{dy} \log f_Y(y;a,\epsilon) dy \\
=& \int_{-\infty}^{\infty} \frac{d}{dy} \mathbb{E}_{\tilde{X}}\Big[f_{Y|\tilde{X}}(y|\tilde{X};a,\epsilon)\Big] \frac{d}{dy} \log \mathbb{E}_{\tilde{X}}\Big[f_{Y|\tilde{X}}(y|\tilde{X};a,\epsilon)\Big] dy \\
=& \int_{-\infty}^{\infty} \frac{d}{dy} \mathbb{E}_{\tilde{X}}\left[\frac{1}{\sqrt{2\pi(a+\epsilon)}} \exp\left(-\frac{1}{2(a+\epsilon)}(y-\tilde{X})^2\right)\right] \\
& \times \frac{d}{dy} \log \mathbb{E}_{\tilde{X}}\left[\frac{1}{\sqrt{2\pi(a+\epsilon)}} \exp\left(-\frac{1}{2(a+\epsilon)}(y-\tilde{X})^2\right)\right] dy, \qquad \text{(A.82)}
\end{aligned}
$$

where $Y = \tilde{X}+\sqrt{a+\epsilon}\tilde{W}$. Since $f_{Y|\tilde{X}}(y|\tilde{x};a,\epsilon)$ is a Gaussian density function with mean $\tilde{x}$ and variance $a+\epsilon$, the equality in (A.82) holds. In equation

(A.82), a and ϵ are symmetrically included in the equation, and therefore,

$$\frac{d}{d\epsilon}J(\tilde{X}+\sqrt{a+\epsilon}\tilde{W})=\frac{d}{da}J(\tilde{X}+\sqrt{a+\epsilon}\tilde{W}).$$

Since random variable $\tilde{X}$ is arbitrary and a is an arbitrary non-negative real-valued number in equation (A.81), the proof is completed.

□

A.8 Derivation of Equation (2.16)

Given the channel model (2.1), random variables X and W are independent of each other, a is a deterministic parameter, and random variable Y is the summation of X and $\sqrt{a}W$. Therefore, between the two probability density functions $f_{Y|X}(y|x;a)$ and $f_W(w)$, there exists a relationship that can be established as follows.

$$\begin{aligned} f_{Y|X}(y|x;a) &= \frac{1}{\sqrt{a}}f_W(w)\Big|_{w=\frac{y-x}{\sqrt{a}}} \\ &= \frac{1}{\sqrt{a}}f_W\left(\frac{y-x}{\sqrt{a}}\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{d}{dy}f_{Y|X}(y|x;a) &= \frac{1}{\sqrt{a}}\left(\frac{d}{dy}f_W\left(\frac{y-x}{\sqrt{a}}\right)\right) \\ &= \frac{1}{\sqrt{a}}\left(\frac{1}{\sqrt{a}}\frac{d}{dw}f_W\left(w\right)\right)\Big|_{w=\frac{y-x}{\sqrt{a}}}, \end{aligned}$$

and

$$\begin{aligned}
&\frac{d}{da}f_{Y|X}(y|x;a)\\
=\;&\frac{d}{da}\left(\frac{1}{\sqrt{a}}f_W\left(\frac{y-x}{\sqrt{a}}\right)\right)\\
=\;&-\frac{1}{2a\sqrt{a}}f_W\left(\frac{y-x}{\sqrt{a}}\right)+\frac{1}{\sqrt{a}}\frac{d}{da}f_W\left(\frac{y-x}{\sqrt{a}}\right)\\
=\;&-\frac{1}{2a\sqrt{a}}f_W\left(\frac{y-x}{\sqrt{a}}\right)+\frac{1}{\sqrt{a}}\left(-\frac{1}{2a\sqrt{a}}(y-x)\frac{d}{dw}f_W(w)\Big|_{w=\frac{y-x}{\sqrt{a}}}\right). \quad \text{(A.83)}
\end{aligned}$$

Equation (A.83) is further processed as

$$\begin{aligned}
&-\frac{1}{2a\sqrt{a}}f_W\left(\frac{y-x}{\sqrt{a}}\right)+\frac{1}{\sqrt{a}}\left(-\frac{1}{2a\sqrt{a}}(y-x)\frac{d}{dw}f_W(w)\Big|_{w=\frac{y-x}{\sqrt{a}}}\right)\\
=&-\frac{1}{2a}\left[\frac{1}{\sqrt{a}}f_W\left(\frac{y-x}{\sqrt{a}}\right)+\frac{y-x}{\sqrt{a}}\left(\frac{1}{\sqrt{a}}\frac{d}{dw}f_W(w)\Big|_{w=\frac{y-x}{\sqrt{a}}}\right)\right]\\
=&-\frac{1}{2a}\left[\left(\frac{d}{dy}(y-x)\right)f_{Y|X}(y|x;a)+(y-x)\frac{d}{dy}f_{Y|X}(y|x;a)\right]\\
=&-\frac{1}{2a}\frac{d}{dy}\left[(y-x)f_{Y|X}(y|x;a)\right],
\end{aligned}$$

and therefore,

$$\frac{d}{da}f_{Y|X}(y|x;a) \;=\; -\frac{1}{2a}\frac{d}{dy}\left[(y-x)f_{Y|X}(y|x;a)\right].$$

A.9 Explanation of Assumptions (2.17) in Corollaries 2.2, 2.3

1. Corollary 2.2

   Given the channel $Y = X + \sqrt{a}W$ in (2.1), W is assumed to be exponentially distributed with unit parameter, i.e., its pdf $f_W(w)$ is defined as $\exp(-w)U(w)$,

where $U(\cdot)$ denotes a unit step function. Since random variables X and W are independent of each other, conditional density function $f_{Y|X}(y|x;a)$ is expressed as

$$f_{Y|X}(y|x;a) = \frac{1}{\sqrt{a}} \exp\left(\frac{y-x}{\sqrt{a}}\right) U(y-x), \tag{A.84}$$

and its derivatives with respect to y and a are respectively denoted as

$$\frac{d}{dy} f_{Y|X}(y|x;a) = -\frac{1}{\sqrt{a}} f_{Y|X}(y|x;a) + \frac{1}{\sqrt{a}} \exp\left(\frac{y-x}{\sqrt{a}}\right) \delta(y-x), \tag{A.85}$$

$$\frac{d}{da} f_{Y|X}(y|x;a) = -\frac{1}{2a} f_{Y|X}(y|x;a) + \frac{(y-x)}{2a\sqrt{a}} f_{Y|X}(y|x;a), \tag{A.86}$$

where $\delta(\cdot)$ is a Dirac delta function.

The absolute values of equations (A.85), (A.86) are bounded as

$$\begin{aligned} \left|\frac{d}{dy} f_{Y|X}(y|x;a)\right| &= \left|-\frac{1}{\sqrt{a}} f_{Y|X}(y|x;a) + \frac{1}{\sqrt{a}} \exp\left(\frac{y-x}{\sqrt{a}}\right) \delta(y-x)\right| \\ &\leq \left|\frac{1}{\sqrt{a}} f_{Y|X}(y|x;a)\right| + \left|\frac{1}{\sqrt{a}} \exp\left(\frac{y-x}{\sqrt{a}}\right) \delta(y-x)\right| \\ &\leq \frac{1}{a} + \frac{1}{\sqrt{a}} \exp\left(\frac{y-x}{\sqrt{a}}\right) \delta(y-x), \end{aligned} \tag{A.87}$$

and

$$\begin{aligned} \left|\frac{d}{da} f_{Y|X}(y|x;a)\right| &= \left|-\frac{1}{2a} f_{Y|X}(y|x;a) + \frac{(y-x)}{2a\sqrt{a}} f_{Y|X}(y|x;a)\right| \\ &\leq \left|\frac{1}{2a} f_{Y|X}(y|x;a)\right| + \left|\frac{(y-x)}{2a\sqrt{a}} f_{Y|X}(y|x;a)\right| \quad \text{(A.88)} \\ &\leq \frac{1}{2a\sqrt{a}} + E, \quad \text{(A.89)} \end{aligned}$$

where $E = \max_y[(y-x) f_{Y|X}(y|x;a)]$. Since $f_{Y|X}(y|x;a)$ is exponentially de-

creasing as y approaches ∞, the real valued E always exists. In addition, $\max_y f(Y|X)(y|x;a) = 1/\sqrt{a}$, and therefore, the inequalities in (A.87) and (A.89) are satisfied.

The right-hand side of (A.87) and (A.89) are now integrable as follows:

$$\begin{aligned}
&\mathbb{E}_X\left[\frac{1}{a} + \frac{1}{\sqrt{a}}\exp\left(\frac{y-X}{\sqrt{a}}\right)\delta(y-X)\right] = \frac{1}{a} + f_X(y), \\
&\mathbb{E}_X\left[\frac{1}{2a\sqrt{a}} + E\right] = \frac{1}{2a\sqrt{a}} + E.
\end{aligned} \tag{A.90}$$

If a function $f_X(x)$ is bounded, by dominated convergence theorem, assumption (2.17a) is verified.

Second, assumption (2.17b) is verified as follows.

$$\left|\frac{d}{da}\left(f_Y(y;a)\log f_Y(y;a)\right)\right| \tag{A.91}$$

$$\begin{aligned}
&\leq \left|\log f_Y(y;a)\frac{d}{da}f_Y(y;a)\right| + \left|\frac{d}{da}f_Y(y;a)\right| \\
&= \left|\log f_Y(y;a)\mathbb{E}_X\left[-\frac{1}{2a}f_{Y|X}(y|X;a)+\frac{(y-X)}{2a\sqrt{a}}f_{Y|X}(y|X;a)\right]\right| + \left|\frac{d}{da}f_Y(y;a)\right| \\
&= \left|\sqrt{f_Y(y;a)}\log f_Y(y;a)\left(-\frac{1}{2a}\sqrt{f_Y(y;a)}\right.\right. \\
&\qquad \left.\left.+\frac{y}{2a\sqrt{a}}\sqrt{f_Y(y;a)}-\frac{\mathbb{E}_X\left[Xf_{Y|X}(y|X;a)\right]}{2a\sqrt{a}\sqrt{f_Y(y;a)}}\right)\right| + \left|\frac{d}{da}f_Y(y;a)\right| \\
&= \underbrace{\left|2\sqrt{f_Y(y;a)}\log\sqrt{f_Y(y;a)}\right|}_{(d_1)} \\
&\qquad \times \underbrace{\left|-\frac{1}{2a}\sqrt{f_Y(y;a)}+\frac{y}{2a\sqrt{a}}\sqrt{f_Y(y;a)}-\frac{\mathbb{E}_X[Xf_{Y|X}(y|X;a)]}{2a\sqrt{a}\sqrt{f_Y(y;a)}}\right|}_{(d_2)} + \underbrace{\left|\frac{d}{da}f_Y(y;a)\right|}_{(d_3)}
\end{aligned} \tag{A.92}$$

$$\leq K\left|2\sqrt{f_Y(y;a)}\log\sqrt{f_Y(y;a)}\right| + \left|\frac{d}{da}f_Y(y;a)\right|.$$

The term (d_3) is bounded by an integrable function due to equation (A.88), factor (d_2) is bounded by a constant K due to assumptions (2.17c) and (2.17d),

which will be proved later, and factor (d_1) is bounded, and it is integrable:

$$
\begin{aligned}
&\int_0^\infty \left|\sqrt{f_Y(y;a)}\log\sqrt{f_Y(y;a)}\right| dy \\
=&\frac{1}{2}\int_0^\infty \left|\sqrt{f_Y(y;a)}\log f_Y(y;a)\right| dy \\
=&\frac{1}{2}\int_0^\infty \left(\mathbb{E}_X\left[\frac{1}{\sqrt{a}}\exp\left(-\frac{1}{\sqrt{a}}(y-X)\right)U(y-X)\right]\right)^{\frac{1}{2}} \\
&\times\log\mathbb{E}_X\left[\frac{1}{\sqrt{a}}\exp\left(-\frac{1}{\sqrt{a}}(y-X)\right)U(y-X)\right] dy \\
=&\frac{1}{2}\int_0^\infty \frac{1}{\sqrt[4]{a}}\exp\left(-\frac{1}{2\sqrt{a}}y\right)\left(\mathbb{E}_X\left[\exp\left(\frac{1}{\sqrt{a}}X\right)U(y-X)\right]\right)^{\frac{1}{2}} \\
&\times\left|\log\left(\frac{1}{\sqrt{a}}\exp\left(-\frac{1}{\sqrt{a}}y\right)\mathbb{E}_X\left[\exp\left(\frac{1}{\sqrt{a}}X\right)U(y-X)\right]\right)\right| dy \\
\leq&\frac{1}{2}\int_0^\infty \frac{1}{\sqrt[4]{a}}\exp\left(-\frac{1}{2\sqrt{a}}y\right)\left(\mathbb{E}_X\left[\exp\left(\frac{1}{\sqrt{a}}X\right)\right]\right)^{\frac{1}{2}} \\
&\times\left|\log\left(\frac{1}{\sqrt{a}}\exp\left(-\frac{1}{\sqrt{a}}y\right)\mathbb{E}_X\left[\exp\left(\frac{1}{\sqrt{a}}X\right)\right]\right)\right| dy \\
\leq&\frac{1}{2}\int_0^\infty \frac{1}{\sqrt[4]{a}}\exp\left(-\frac{1}{2\sqrt{a}}y\right)\left(M_X\left(\frac{1}{\sqrt{a}}\right)\right)^{\frac{1}{2}} \\
&\times\left|\log\left(\frac{1}{\sqrt{a}}\exp\left(-\frac{1}{\sqrt{a}}y\right)M_X\left(\frac{1}{\sqrt{a}}\right)\right)\right| dy, \qquad \text{(A.93)}
\end{aligned}
$$

where $M_X(\cdot)$ denotes the moment generating function of X. If the moment generating function of X exists, then equation (A.93) is bounded and integrable, and so does the term (d_1). Therefore, term (d_1) is integrable with respect to y, and assumption (2.17b) is verified by dominated convergence theorem.

Similarly, assumption (2.17c) is verified as follows.

$$\begin{aligned} \left|f_{Y|X}(y|x;a)\right| &= \left|\frac{1}{\sqrt{a}}\exp\left(\frac{y-x}{\sqrt{a}}\right)U(y-x)\right| \\ &\leq \frac{1}{\sqrt{a}}, \qquad \text{(A.94)} \\ \left|x f_{Y|X}(y|x;a)\right| &= \left|x\frac{1}{\sqrt{a}}\exp\left(\frac{y-x}{\sqrt{a}}\right)U(y-x)\right| \\ &\leq \frac{1}{\sqrt{a}}x, \qquad \text{(A.95)} \end{aligned}$$

and the right hand-side terms of (A.94) and (A.95) are integrable as

$$\begin{aligned} \mathbb{E}_X\left[\frac{1}{\sqrt{a}}\right] &= \frac{1}{\sqrt{a}}, \\ \mathbb{E}_X\left[\frac{1}{\sqrt{a}}X\right] &= \frac{1}{\sqrt{a}}\mathbb{E}_X[X], \qquad \text{(A.96)} \end{aligned}$$

and if $E_X[X]$ exists, assumption (2.17c) is satisfied.

Since $f_{Y|X}(y|x;a)$ is exponentially decreasing, $\lim_{y\to\infty} y^2 f_Y(y;a)$ is zero. In addition,

$$\begin{aligned} & \lim_{y\to 0} y^2 f_Y(y;a) \\ = & \lim_{y\to 0} \mathbb{E}_X\left[y^2 f_{Y|X}(y|X;a)\right] \\ = & \lim_{y\to 0} \mathbb{E}_X\left[y^2\frac{1}{\sqrt{a}}\exp\left(\frac{y-x}{\sqrt{a}}\right)U(y-x)\right] \\ = & \mathbb{E}_X\left[0\times\frac{1}{\sqrt{a}}\exp\left(\frac{-x}{\sqrt{a}}\right)U(-x)\right] \\ = & 0. \qquad \text{(A.97)} \end{aligned}$$

Assumption (2.17d) is expressed as

$$\begin{aligned}
\frac{\mathbb{E}_X\left[X f_{Y|X}(y|X;a)\right]}{\sqrt{f_Y(y;a)}} &= \frac{\mathbb{E}_X\left[X f_{Y|X}(y|X;a)\right]}{f_Y(y;a)}\sqrt{f_Y(y;a)} \\
&= \frac{\int_0^\infty x f_X(x)\frac{1}{\sqrt{a}}\exp\left(\frac{y-x}{\sqrt{a}}\right)U(y-x)dx}{\int_0^\infty f_X(x)\frac{1}{\sqrt{a}}\exp\left(\frac{y-x}{\sqrt{a}}\right)U(y-x)dx}\sqrt{f_Y(y;a)} \quad \text{(A.98)} \\
&\leq \frac{y\int_0^y f_X(x)\frac{1}{\sqrt{a}}\exp\left(\frac{y-x}{\sqrt{a}}\right)dx}{\int_0^y f_X(x)\frac{1}{\sqrt{a}}\exp\left(\frac{y-x}{\sqrt{a}}\right)dx}\sqrt{f_Y(y;a)} \quad \text{(A.99)} \\
&= y\sqrt{f_Y(y;a)}.
\end{aligned}$$

The inequality in (A.99) is due to the fact that, in (A.98), the term inside integral is non-negative, x is increasing, and integration is performed from 0 to y.

Therefore, the assumptions in (2.17) require the following conditions: 1) existence of $\mathbb{E}_X[X]$, 2) existence of $M_X(\cdot)$, 3) bounded pdf $f_X(x)$, and these are further simplified into the existence of the moment generating function of X and bounded pdf $f_X(x)$.

2. Corollary 2.3

   Given the channel $Y = X + \sqrt{a}W$ in (2.1), W is assumed to be a gamma random variable, and its pdf is expressed as

$$f_W(w) = \frac{1}{\Gamma(\alpha)}w^{\alpha-1}\exp(-w)U(w),$$

   where $\Gamma(\cdot)$ is a gamma function, $U(\cdot)$ denotes a unit step function, and $\alpha \geq 2$. Since random variables X and W are independent of each other, the conditional

density function $f_{Y|X}(y|x;a)$ is expressed as

$$f_{Y|X}(y|x;a) = \frac{1}{\sqrt{a}\Gamma(\alpha)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha-1}\exp\left(-\frac{y-x}{\sqrt{a}}\right)U(y-x), \quad \text{(A.100)}$$

and its derivatives are denoted as

$$\begin{aligned}
&\frac{d}{dy}f_{Y|X}(y|x;a) \\
=&-\frac{1}{\sqrt{a}}f_{Y|X}(y|x;a) + \frac{1}{a\Gamma(\alpha-1)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha-2}\exp\left(-\frac{y-x}{\sqrt{a}}\right)U(y-x), \quad \text{(A.101)} \\
&\frac{d}{da}f_{Y|X}(y|x;a) \\
=&-\frac{\alpha}{2a}f_{Y|X}(y|x;a) + \frac{\alpha}{2a}\left(\frac{1}{\sqrt{a}\Gamma(\alpha+1)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha}\exp\left(-\frac{y-x}{\sqrt{a}}\right)U(y-x)\right). \quad \text{(A.102)}
\end{aligned}$$

The absolute values of equations (A.101), (A.102) are bounded as

$$\begin{aligned}
&\left|\frac{d}{dy}f_{Y|X}(y|x;a)\right| \\
=&\left| -\frac{1}{\sqrt{a}}f_{Y|X}(y|x;a) + \frac{1}{a\Gamma(\alpha-1)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha-2}\exp\left(-\frac{y-x}{\sqrt{a}}\right)U(y-x)\right| \\
\leq&\left|\frac{1}{\sqrt{a}}f_{Y|X}(y|x;a)\right| + \left|\frac{1}{a\Gamma(\alpha-1)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha-2}\exp\left(-\frac{y-x}{\sqrt{a}}\right)U(y-x)\right| \\
=&\left|\frac{1}{\sqrt{a}}f_{Y|X}(y|x;a)\right| + \left|\frac{1}{\sqrt{a}}f_{Y_{\alpha-1}|X}(y|x;a)\right| \\
=&\frac{1}{\sqrt{a}}f_{Y|X}(y|x;a) + \frac{1}{\sqrt{a}}f_{Y_{\alpha-1}|X}(y|x;a), \quad \text{(A.103)}
\end{aligned}$$

where

$$f_{Y_{\alpha-1}|X}(y|x;a) = \frac{1}{\sqrt{a}\Gamma(\alpha-1)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha-2}\exp\left(-\frac{y-x}{\sqrt{a}}\right)U(y-x), \quad \text{(A.104)}$$

i.e., this is a gamma density function with two parameters defined as $\alpha-1$ and

1, and

$$\begin{aligned}
&\left|\frac{d}{da}f_{Y|X}(y|x;a)\right| \\
=&\left|-\frac{\alpha}{2a}f_{Y|X}(y|x;a)+\frac{\alpha}{2a}\left(\frac{1}{\sqrt{a}\Gamma(\alpha+1)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha}\exp\left(-\frac{y-x}{\sqrt{a}}\right)\right)\right| \\
\leq&\left|\frac{\alpha}{2a}f_{Y|X}(y|x;a)\right|+\left|\frac{\alpha}{2a}\left(\frac{1}{\sqrt{a}\Gamma(\alpha+1)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha}\exp\left(-\frac{y-x}{\sqrt{a}}\right)\right)\right| \\
=&\left|\frac{\alpha}{2a}f_{Y|X}(y|x;a)\right|+\left|\frac{\alpha}{2a}f_{Y_{\alpha+1}|X}(y|x;a)\right| \\
=&\frac{\alpha}{2a}f_{Y|X}(y|x;a)+\frac{\alpha}{2a}f_{Y_{\alpha+1}|X}(y|x;a), \qquad \text{(A.105)}
\end{aligned}$$

where

$$f_{Y_{\alpha+1}|X}(y|x;a) \;=\frac{1}{\sqrt{a}\Gamma(\alpha+1)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha}\exp\left(-\frac{y-x}{\sqrt{a}}\right)U(y-x), \qquad \text{(A.106)}$$

i.e., this is a gamma density function with two parameters defined as $\alpha+1$ and 1.

Since $f_{Y_{\alpha-1}|X}(y|x;a)$, $f_{Y|X}(y|x;a)$, and $f_{Y_{\alpha+1}|X}(y|x;a)$ are all integrable, the right-hand side of (A.103) and (A.105) are integrable as

$$\begin{aligned}
&\mathbb{E}_X\left[\frac{1}{\sqrt{a}}f_{Y|X}(y|X;a)+\frac{1}{\sqrt{a}}f_{Y_{\alpha-1}|X}(y|X;a)\right] \\
=&\frac{1}{\sqrt{a}}f_Y(y;a)+\frac{1}{\sqrt{a}}f_{Y_{\alpha-1}}(y;a), \qquad \text{(A.107)} \\
&\mathbb{E}_X\left[\frac{\alpha}{2a}f_{Y|X}(y|X;a)+\frac{\alpha}{2a}f_{Y_{\alpha+1}|X}(y|X;a)\right] \\
=&\frac{\alpha}{2a}f_Y(y;a)+\frac{\alpha}{2a}f_{Y_{\alpha+1}}(y;a), \qquad \text{(A.108)}
\end{aligned}$$

where $f_{Y_{\alpha-1}}(y;a)=\mathbb{E}_X[f_{Y_{\alpha-1}|X}(y|X;a)]$, and $f_{Y_{\alpha+1}}(y;a)=\mathbb{E}_X[f_{Y_{\alpha+1}|X}(y|X;a)]$. Therefore, assumption (2.17a) is verified by dominated convergence theorem.

Second, assumption (2.17b) is verified as follows.

$$\left|\frac{d}{da}\left(f_Y(y;a)\log f_Y(y;x)\right)\right| \tag{A.109}$$

$$\begin{aligned}
&\leq \left|\log f_Y(y;x)\frac{d}{da}f_Y(y;a)\right| + \left|\frac{d}{da}f_Y(y;a)\right| \\
&= \left|\log f_Y(y;x)\mathbb{E}_X\left[-\frac{1}{2a}f_{Y|X}(y|X;a)+\frac{(y-X)}{2a\sqrt{a}}f_{Y|X}(y|X;a)\right]\right| \\
&\quad + \left|\frac{d}{da}f_Y(y;a)\right| \\
&= \left|2\sqrt{f_Y(y;x)}\log\sqrt{f_Y(y;x)}\left(-\frac{1}{2a}\sqrt{f_Y(y;x)}\right.\right. \\
&\qquad \left.\left.+\frac{y}{2a\sqrt{a}}\sqrt{f_Y(y;x)}-\frac{\mathbb{E}_X\left[Xf_{Y|X}(y|X;a)\right]}{2a\sqrt{a}\sqrt{f_Y(y;x)}}\right)\right| + \left|\frac{d}{da}f_Y(y;a)\right| \\
&= \underbrace{\left|2\sqrt{f_Y(y;x)}\log\sqrt{f_Y(y;x)}\right|}_{(e_1)} \\
&\qquad \times \underbrace{\left|-\frac{1}{2a}\sqrt{f_Y(y;x)}+\frac{y}{2a\sqrt{a}}\sqrt{f_Y(y;x)}-\frac{\mathbb{E}_X\left[Xf_{Y|X}(y|X;a)\right]}{2a\sqrt{a}\sqrt{f_Y(y;x)}}\right|}_{(e_2)} \\
&\quad + \underbrace{\left|\frac{d}{da}f_Y(y;a)\right|}_{(e_3)}.
\end{aligned} \tag{A.110}$$

The factors (e_1), (e_2), and (e_3) can be verified using exactly the same reasons as the factors (d_1), (d_2), and (d_3), in (A.92), respectively. Therefore, like equation (A.93), the existence of moment generating function of X is required.

Assumption (2.17c) is confirmed by the following procedures.

Since $f_{Y|X}(y|x;a)$ is exponentially decreasing, $\lim_{y\to\infty} y^2 f_Y(y;a)$ is zero. By the same procedure as equation (A.97), $y^2 f_Y(y;a)$ becomes zero as y approaches zero.

In addition,

$$\left|f_{Y|X}(y|x;a)\right| \leq f_{Y|X}(y|x;a)\Big|_{y=x+\sqrt{a}(\alpha-1)}, \tag{A.111}$$

$$\left|xf_{Y|X}(y|x;a)\right| \leq xf_{Y|X}(y|x;a)\Big|_{y=x+\sqrt{a}(\alpha-1)}. \tag{A.112}$$

The inequalities above are due to the fact that the function $f_{Y|X}(y|x;a)$ is always nonnegative, and it is maximized at $y = x + \sqrt{a}(\alpha - 1)$. Therefore, the right-hand sides of (A.111) and (A.112) are integrable as

$$\begin{aligned} & \mathbb{E}_X\left[\frac{1}{\sqrt{a}\Gamma(\alpha)}(\alpha-1)^{\alpha-1}\exp(-(\alpha-1))\right] \\ = \; & \frac{1}{\sqrt{a}\Gamma(\alpha)}(\alpha-1)^{\alpha-1}\exp(-(\alpha-1)), \\ & \mathbb{E}_X\left[X\frac{1}{\sqrt{a}\Gamma(\alpha)}(\alpha-1)^{\alpha-1}\exp(-(\alpha-1))\right] \\ = \; & \frac{1}{\sqrt{a}\Gamma(\alpha)}(\alpha-1)^{\alpha-1}\exp(-(\alpha-1))\mathbb{E}_X[X], \end{aligned} \tag{A.113}$$

and, if $\mathbb{E}_X[X]$ exits, by dominated convergence theorem, assumption (2.17c) is verified.

Finally, assumption (2.17d) is expressed as

$$\begin{aligned} \frac{\mathbb{E}_X\left[Xf_{Y|X}(y|X;a)\right]}{\sqrt{f_Y(y;a)}} &= \frac{\mathbb{E}_X\left[Xf_{Y|X}(y|X;a)\right]}{f_Y(y;a)}\sqrt{f_Y(y;a)} \\ &= \frac{\int_0^\infty x f_X(x)\frac{1}{\sqrt{a}\Gamma(\alpha)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha-1}\exp\left(-\frac{y-x}{\sqrt{a}}\right)U(y-x)dx}{\int_0^\infty f_X(x)\frac{1}{\sqrt{a}\Gamma(\alpha)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha-1}\exp\left(-\frac{y-x}{\sqrt{a}}\right)U(y-x)dx}\sqrt{f_Y(y;a)} \qquad \text{(A.114)} \\ &\leq \frac{y\int_0^y f_X(x)\frac{1}{\sqrt{a}\Gamma(\alpha)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha-1}\exp\left(-\frac{y-x}{\sqrt{a}}\right)dx}{\int_0^y f_X(x)\frac{1}{\sqrt{a}\Gamma(\alpha)}\left(\frac{y-x}{\sqrt{a}}\right)^{\alpha-1}\exp\left(-\frac{y-x}{\sqrt{a}}\right)dx}\sqrt{f_Y(y;a)} \qquad \text{(A.115)} \\ &= y\sqrt{f_Y(y;a)}. \end{aligned}$$

The inequality in (A.115) is due to the fact that, in (A.114), the term inside integral is non-negative, x is increasing, and the integration with respect to x is performed from 0 to y.

Therefore, in this case, the assumptions in (2.17) require the existence of the mean and moment generating function of X, and these are further simplified to the existence of the moment generating function of X.

APPENDIX B

EXTREMAL ENTROPY INEQUALITY

B.1 Proof of Lemma 4.7

Proving $\mathbf{\Sigma}_{X^*} \preceq \mathbf{\Sigma}_X$ is equivalent to proving the following:

$$\mathbf{\Sigma}_{X^*} \preceq \mathbf{\Sigma}_X \tag{B.1}$$

$$\iff \mathbf{\Sigma}_{\tilde{W}} \preceq (\mu - 1)\,\mathbf{\Sigma}_X \tag{B.2}$$

$$\iff \left((\mathbf{\Sigma}_X + \mathbf{\Sigma}_W)^{-1} + \mathbf{L}\right)^{-1} - \mathbf{\Sigma}_X \preceq (\mu - 1)\,\mathbf{\Sigma}_X \tag{B.3}$$

$$\iff (\mathbf{\Sigma}_X + \mathbf{\Sigma}_W)^{-1} + \mathbf{L} \succeq \mu^{-1}\mathbf{\Sigma}_X^{-1} \tag{B.4}$$

Since there always exists a non-singular matrix which simultaneously diagonalizes two positive semi-definite matrices [21], there exists a non-singular matrix $\mathbf{Q}$ which simultaneously diagonalize both $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_W$ as follows:

$$\mathbf{Q}^T\mathbf{\Sigma}_X\mathbf{Q} = \mathbf{I}, \tag{B.5}$$

$$\mathbf{Q}^T\mathbf{\Sigma}_W\mathbf{Q} = \mathbf{D}_W, \tag{B.6}$$

where $\mathbf{I}$ is an identity matrix, and $\mathbf{D}_W$ is a diagonal matrix. Since $\mathbf{Q}$ is a non-singular matrix, the inverse of $\mathbf{Q}$ always exists, and $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_W$ are expressed as

$$\mathbf{\Sigma}_X = \mathbf{Q}^{-T}\mathbf{Q}^{-1}, \tag{B.7}$$

$$\mathbf{\Sigma}_W = \mathbf{Q}^{-T}\mathbf{D}_W\mathbf{Q}^{-1}. \tag{B.8}$$

If we define $\mathbf{D}_L$ as a diagonal matrix whose i^{th} diagonal element is represented as d_{L_i}, and which it is defined as

$$d_{L_i} = \begin{cases} 0 & \text{if } d_{W_i} \leq \mu - 1 \\ \frac{d_{W_i} - (\mu - 1)}{\mu\left(1 + d_{W_i}\right)} & \text{if } d_{W_i} > \mu - 1 \end{cases} \tag{B.9}$$

where d_{W_i} denotes the i^{th} diagonal element of $\mathbf{D}_W$, and define $\mathbf{L}$ as

$$\mathbf{L} \;=\; \mathbf{Q}\mathbf{D}_L\mathbf{Q}^T, \tag{B.10}$$

the equation (B.4) is equivalent to

$$\left(\mathbf{\Sigma}_X + \mathbf{\Sigma}_W\right)^{-1} + \mathbf{L} \succeq \mu^{-1}\mathbf{\Sigma}_X^{-1} \tag{B.11}$$

$$\iff \left(\mathbf{Q}^{-T}\mathbf{Q}^{-1} + \mathbf{Q}^{-T}\mathbf{D}_W\mathbf{Q}^{-1}\right)^{-1} + \mathbf{Q}\mathbf{D}_L\mathbf{Q}^T \succeq \mu^{-1}\mathbf{Q}\mathbf{Q}^T \tag{B.12}$$

$$\iff \left(\mathbf{I} + \mathbf{D}_W\right)^{-1} + \mathbf{D}_L \succeq \mu^{-1}\mathbf{I}. \tag{B.13}$$

The equation (B.13) always holds since $\mathbf{D}_L$ is defined as in (B.9) and (B.10) to satisfy (B.13). Therefore, the inequality (B.1) is also satisfied.

We know that $\mathbf{\Sigma}_{X'}$ is $\mathbf{\Sigma}_X - \mathbf{\Sigma}_{X^*}$. Since $\mathbf{\Sigma}_{X^*} = (\mu - 1)^{-1}\mathbf{\Sigma}_{\tilde{W}}$, $\mathbf{\Sigma}_{X'}$ is expressed as $\mathbf{\Sigma}_X - (\mu - 1)^{-1}\mathbf{\Sigma}_{\tilde{W}}$, and

$$\mathbf{\Sigma}_{X'}\mathbf{L} \;=\; \left(\mathbf{\Sigma}_X - (\mu - 1)^{-1}\,\mathbf{\Sigma}_{\tilde{W}}\right)\mathbf{L}, \tag{B.14}$$

and the equation (B.14) is re-written as

$$
\begin{aligned}
& \boldsymbol{\Sigma}_{X'}\mathbf{L} \\
= \; & \left(\boldsymbol{\Sigma}_X - (\mu-1)^{-1}\boldsymbol{\Sigma}_{\tilde{W}}\right)\mathbf{L}, && \text{(B.15)} \\
= \; & \Bigg\{\mathbf{Q}^{-T}\mathbf{Q}^{-1} - (\mu-1)^{-1} \\
& \qquad \times\left(\left(\left(\mathbf{Q}^{-T}\mathbf{Q}^{-1}+\mathbf{Q}^{-T}\mathbf{D}_W\mathbf{Q}^{-1}\right)^{-1}+\mathbf{Q}\mathbf{D}_L\mathbf{Q}^T\right)^{-1}-\mathbf{Q}^{-T}\mathbf{Q}^{-1}\right)\Bigg\} \\
& \times\mathbf{Q}\mathbf{D}_L\mathbf{Q}^T \\
= \; & (\mu-1)^{-1}\mathbf{Q}^{-T}\left(\mu\mathbf{I}-\left(\left(\mathbf{I}+\mathbf{D}_W\right)^{-1}+\mathbf{D}_L\right)^{-1}\right)\mathbf{D}_L\mathbf{Q}^T && \text{(B.16)} \\
= \; & \mathbf{0}. && \text{(B.17)}
\end{aligned}
$$

The equality (B.16) is due to the equations (B.7), (B.8), and (B.10), and the equality (B.17) is due to (B.9). Therefore, by defining $\boldsymbol{\Sigma}_{\tilde{W}} = ((\boldsymbol{\Sigma}_X+\boldsymbol{\Sigma}_W)^{-1}+\mathbf{L})^{-1}-\boldsymbol{\Sigma}_X$, we can make $\boldsymbol{\Sigma}_{\tilde{W}}$ satisfy

$$
\boldsymbol{\Sigma}_{\tilde{W}} \preceq (\mu-1)\boldsymbol{\Sigma}_X, \quad \boldsymbol{\Sigma}_{X'}\mathbf{L}=\mathbf{0}, \tag{B.18}
$$

and the proof is completed.

Remark B.1. *Since the optimization problem in [32] is generally nonconvex, the existence of optimal solution must be proved [32], [53], and this step is very complicated. However, in our proof, Lemmas 4.7 and 4.8 serve as a substitute for this step since we by-pass KKT-condition related parts using the data processing inequality. This makes the proposed proof much simpler.*

B.2 Proof of Lemma 4.8

Proving $\mathbf{\Sigma}_{\tilde{W}} \preceq \mu^{-1}\mathbf{\Sigma}_V$ is equivalent to proving the following:

$$\mathbf{\Sigma}_{\tilde{W}} \preceq \mu^{-1}\mathbf{\Sigma}_V \tag{B.19}$$

$$\iff \mathbf{\Sigma}_W^{-1} + \mathbf{K} \succeq \mu\mathbf{\Sigma}_V^{-1} \tag{B.20}$$

Since there always exists a non-singular matrix which simultaneously diagonalizes two positive semi-definite matrices [21], there exists a non-singular matrix $\mathbf{Q}$ which simultaneously diagonalize both $\mathbf{\Sigma}_W^{-1}$ and $\mathbf{\Sigma}_W^{-1}$ as follows:

$$\mathbf{Q}^T\mathbf{\Sigma}_W\mathbf{Q} = \mathbf{D}_W \tag{B.21}$$

$$\mathbf{Q}^T\mathbf{\Sigma}_V\mathbf{Q} = \mathbf{I}, \tag{B.22}$$

where $\mathbf{I}$ is an identity matrix, and $\mathbf{D}_W$ is a diagonal matrix. Since $\mathbf{Q}$ is a non-singular matrix, the inverse of $\mathbf{Q}$ always exists, and $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_V$ are expressed as

$$\mathbf{\Sigma}_W = \mathbf{Q}^{-T}\mathbf{D}_W\mathbf{Q}^{-1}, \tag{B.23}$$

$$\mathbf{\Sigma}_V = \mathbf{Q}^{-T}\mathbf{Q}^{-1}. \tag{B.24}$$

If we define $\mathbf{D}_K$ as a diagonal matrix whose i^{th} diagonal element is represented as d_{K_i}, and which it is defined as

$$d_{K_i} = \begin{cases} 0 & \text{if } d_{W_i} \leq \mu^{-1} \\ \mu - \frac{1}{d_{W_i}} & \text{if } d_{W_i} > \mu^{-1} \end{cases} \tag{B.25}$$

where d_{W_i} denotes the i^{th} diagonal element of $\mathbf{D}_W$, and define $\mathbf{K}$ as

$$\mathbf{K} \;=\; \mathbf{Q}\mathbf{D}_K\mathbf{Q}^T, \tag{B.26}$$

then the equation (B.20) is equivalent to

$$\boldsymbol{\Sigma}_W^{-1} + \mathbf{K} \succeq \mu\boldsymbol{\Sigma}_V^{-1} \tag{B.27}$$

$$\iff \left(\mathbf{Q}^{-T}\mathbf{D}_W\mathbf{Q}^{-1}\right)^{-1} + \mathbf{Q}\mathbf{D}_K\mathbf{Q}^{-1} \succeq \mu\left(\mathbf{Q}^{-T}\mathbf{Q}^{-1}\right)^{-1} \tag{B.28}$$

$$\iff \mathbf{D}_W^{-1} + \mathbf{D}_K \succeq \mu\mathbf{I}. \tag{B.29}$$

The equation (B.29) always holds since $\mathbf{D}_K$ is defined in (B.25). Therefore, the inequality (B.19) is also satisfied.

We know that $\boldsymbol{\Sigma}_{X^*}$ is $(\mu-1)^{-1}(\boldsymbol{\Sigma}_V - \mu\boldsymbol{\Sigma}_{\tilde{W}})$. Therefore,

$$\boldsymbol{\Sigma}_{X^*}\mathbf{K} \;=\; (\mu-1)^{-1}\left(\boldsymbol{\Sigma}_V - \mu\boldsymbol{\Sigma}_{\tilde{W}}\right)\mathbf{K}, \tag{B.30}$$

and the equation (B.30) is re-written as

$$\boldsymbol{\Sigma}_{X^*}\mathbf{K} \;=\; (\mu-1)^{-1}\left(\boldsymbol{\Sigma}_V - \mu\boldsymbol{\Sigma}_{\tilde{W}}\right)\mathbf{K} \tag{B.31}$$

$$= (\mu-1)^{-1}\left(\mathbf{Q}^{-T}\mathbf{Q}^{-1} - \mu\left(\left(\mathbf{Q}^{-T}\mathbf{D}_W\mathbf{Q}^{-1}\right)^{-1} + \mathbf{Q}\mathbf{D}_K\mathbf{Q}^T\right)^{-1}\right) \times\mathbf{Q}\mathbf{D}_K\mathbf{Q}^T \tag{B.32}$$

$$= (\mu-1)^{-1}\mathbf{Q}^{-1}\left(\mathbf{I} - \mu\left(\mathbf{D}_W^{-1} + \mathbf{D}_K\right)^{-1}\right)\mathbf{D}_K\mathbf{Q}^T \tag{B.33}$$

$$= (\mu-1)^{-1}\mu\mathbf{Q}^{-T}\left(\mu^{-1}\mathbf{I} - \left(\mathbf{D}_W^{-1} + \mathbf{D}_K\right)^{-1}\right)\mathbf{D}_K\mathbf{Q}^T \tag{B.34}$$

$$= \mathbf{0}. \tag{B.35}$$

The equality (B.32) is due to the equations (B.23), (B.24), and (B.26), and the

equality (B.35) is due to (B.25). Therefore, by defining $\mathbf{\Sigma}_{\tilde{W}} = (\mathbf{\Sigma}_W^{-1} + \mathbf{K})^{-1}$, we can make $\mathbf{\Sigma}_{\tilde{W}}$ satisfy

$$\mathbf{\Sigma}_{\tilde{W}} \preceq \mu^{-1}\mathbf{\Sigma}_V, \quad \mathbf{\Sigma}_{X^*}\mathbf{K} = \mathbf{0}, \tag{B.36}$$

and the proof is completed.

Remark B.2. *In Lemmas 4.7 and 4.8, we specify the structure of positive semi-definite matrices* $\mathbf{L}$ *and* $\mathbf{K}$*, and this gives more details on the structure of the covariance matrix of the optimal solution.*

APPENDIX C

INFORMATION THEORETIC INEQUALITIES

C.1 Proof of Theorem 5.4

Proof. To prove the inequality in (5.25), we first construct a functional problem as follows.

$$\min_{f_X} \quad \int f_X(x) \log f_X(x) dx, \tag{C.1}$$

$$\text{s. t.} \quad \int f_X(x) dx = 1, \tag{C.2}$$

$$\int x f_X(x) dx = \mu_X,$$

$$\int x^2 f_X(x) dx = m_X^2, \tag{C.3}$$

where μ_X is the first-order moment of X, and m_X represents the second-order moment of X.

Using Theorem 5.3, the functional problem in (C.1) is expressed as

$$\min_{f_X} \quad U[f_X], \tag{C.4}$$

where $U[f_X] = \int K(x, f_X) dx$, $K(x, f_X) = f_X(x) \left(\log f_X(x) + \alpha_0 + \alpha_1 x + \alpha_2 x^2\right)$, α_0, α_1, and α_2 are Lagrange multipliers.

The optimal density function f_{X^*} must satisfy the first-order variation condition as follows:

$$K'_{f_X} - \frac{d}{dx} K'_{f'_X}\Big|_{f_X = f_{X^*}} = 1 + \log f_{X^*}(x) + \alpha_0 + \alpha_1 x + \alpha_2 x^2 = 0. \tag{C.5}$$

Considering the constraints in (C.2)-(C.3) and the equation in (C.5), it follows that

$$\begin{aligned} f_{X^*}(x) &= \frac{1}{\sqrt{2\pi\frac{1}{2\alpha_2}}} \exp\left\{-\frac{1}{2\frac{1}{2\alpha_2}}\left(x+\frac{\alpha_1}{2\alpha_2}\right)^2\right\} \sqrt{2\pi\frac{1}{2\alpha_2}} \exp\left\{-\alpha_0 - 1 + \frac{\alpha_1^2}{4\alpha_2}\right\} \\ &= \frac{1}{\sqrt{2\pi(m_X^2-\mu_X^2)}} \exp\left\{-\frac{1}{2(m_X^2-\mu_X^2)}(x-\mu_X)^2\right\}, \end{aligned} \tag{C.6}$$

where

$$\begin{aligned} \alpha_0 &= -1 + \frac{\mu_X^2}{2(m_X^2-\mu_X^2)} + \frac{1}{2}\log 2\pi(m_X^2-\mu_X^2), \\ \alpha_1 &= -\frac{\mu_X}{m_X^2-\mu_X^2}, \\ \alpha_2 &= \frac{1}{2(m_X^2-\mu_X^2)}. \end{aligned} \tag{C.7}$$

Since the second-order variation of $U[f_X]$ is expressed as

$$K''_{f_X f_X}\Big|_{f_X=f_{X^*}} = \frac{1}{f_{X^*}(x)}, \tag{C.8}$$

and it is positive, the optimal solution f_{X^*} minimizes the variational problem in (C.1).

These first-order and second-order conditions are not sufficient but necessary for the optimal solution. However, as shown in (C.5) and (C.6), there exists only one solution, the Gaussian density function, in the feasible set. Therefore, the Gaussian density function is also sufficient in this case.

Therefore, a negative differential entropy $-h(X)$ is minimized (or, equivalently $h(X)$ is maximized) when $f_X(x)$ is Gaussian, and the proof is completed. □

C.2 Proof of Theorem 5.5

Proof. We first construct a functional problem, which represents the inequality in (5.26) and required constraints, as follows:

$$\min_{f_X} \quad \int f_X(\mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x}, \tag{C.9}$$

$$\text{s. t.} \quad \int f_X(\mathbf{x}) d\mathbf{x} = 1, \tag{C.10}$$

$$\int \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) d\mathbf{x} = \mathbf{\Omega}_X. \tag{C.11}$$

Using Theorem 5.3, the functional problem in (C.9) is expressed as

$$\min_{f_X} \quad U[f_X], \tag{C.12}$$

where $U[f_X] = \int K(\mathbf{x}, f_X) d\mathbf{x} = \int f_X(\mathbf{x}) \left(\log f_X(\mathbf{x}) + \alpha + \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{ij} x_i x_j \right) d\mathbf{x}$, and α and λ_{ij} are Lagrange multipliers.

Based on Theorem 5.1 or Corollary 5.2, by checking the first-order variation condition, we can find the optimal solution $f_{X^*}(\mathbf{x})$ as follows.

$$K'_{f_X} - \frac{d}{dx} K'_{f'_X} \Big|_{f_X = f_{X^*}} = 1 + \log f_{X^*}(\mathbf{x}) + \alpha + \mathbf{x}^T \mathbf{\Lambda} \mathbf{x} = 0, \tag{C.13}$$

$$\tag{C.14}$$

Considering the constraints in (C.10) and (C.11),

$$\begin{aligned} f_{X^*}(\mathbf{x}) &= \exp\left\{-\mathbf{x}^T \mathbf{\Lambda} \mathbf{x} - \alpha - 1\right\} \\ &= (2\pi)^{-\frac{n}{2}} \left| \frac{1}{2} \mathbf{\Lambda}^{-1} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{x}^T \left(\frac{1}{2} \mathbf{\Lambda}^{-1} \right)^{-1} \mathbf{x} \right\} (2\pi)^{\frac{n}{2}} \left| \frac{1}{2} \mathbf{\Lambda}^{-1} \right|^{\frac{1}{2}} \exp\left\{-1 - \alpha\right\} \\ &= (2\pi)^{-\frac{n}{2}} \left| \mathbf{\Omega}_X \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{x}^T \mathbf{\Omega}_X^{-1} \mathbf{x} \right\}, \end{aligned} \tag{C.15}$$

where

$$\begin{aligned} \alpha &= -1 + \frac{1}{2}\log\left(2\pi\right)^n \left|\mathbf{\Omega}_X\right|, \\ \mathbf{\Lambda} &= \frac{1}{2}\mathbf{\Omega}_X^{-1}. \end{aligned} \tag{C.16}$$

Here, two remarks are 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 [32], we can equivalently re-write the functional problem in (C.9) and its constraints in (C.11) as

$$\min_{f_{\bar{X}}} \quad \int f_{\bar{X}}(\mathbf{x}) \log f_{\bar{X}}(\mathbf{x}) d\mathbf{x}, \tag{C.17}$$

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

$$\int \mathbf{x}\mathbf{x}^T f_{\bar{X}}(\mathbf{x}) d\mathbf{x} = \mathbf{\Omega}_{\bar{X}}, \tag{C.18}$$

where $\bar{\mathbf{X}}$ is a random vector with correlation matrix $\mathbf{\Omega}_{\bar{X}}$, and $\mathbf{\Omega}_{\bar{X}}$ is a positive definite matrix. Therefore, without loss of generality, we assume the correlation matrix $\mathbf{\Omega}_X$ is invertible. Second, if an additional constraint, related to the mean vector of $\mathbf{X}$, $\boldsymbol{\mu}_X$, is given, the optimal solution is a multi-variate Gaussian density function, whose mean is $\boldsymbol{\mu}_X$, instead of the multi-variate Gaussian density function, which has zero mean, in (C.15) (cf. Appendix C.1).

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 (C.9).

Therefore, a differential entropy $-h(\mathbf{X})$ is minimized (or, equivalently $h(\mathbf{X})$ is maximized) when $\mathbf{X}$ is a multi-variate Gaussian random vector with zero mean and a covariance matrix $\mathbf{\Sigma}_X$. Even though Theorems 5.1, 5.2 are necessary conditions for the minimum, in this case, a multi-variate Gaussian density function is an actual solution since there is only one solution, a multi-variate Gaussian density function, in the feasible set. □

C.3 Proof of Theorem 5.6

Proof. We first construct a functional problem, which represents the inequality in (5.27) and required constraints, as follows:

$$\min_{f_X} \quad \int_0^\infty f_X(x) \log f_X(x) dx, \tag{C.19}$$

$$\text{s. t.} \quad \int_0^\infty f_X(x) dx = 1,$$

$$\int_0^\infty x^2 f_X(x) dx = m_X^2. \tag{C.20}$$

Using Theorem 5.3, the functional problem in (C.19) is expressed as

$$\min_{f_X} \quad U[f_X], \tag{C.21}$$

where $U[f_X] = \int K(x, f_X) dx$, $K(x, f_X) = f_X(x) \left(\log f_X(x) + \alpha_0 + \alpha_1 x^2\right)$, and α_0 and α_1 are Lagrange multipliers.*

Based on Theorem 5.1 or Corollary 5.2, the first-order variation condition of $U[f_X]$

*For the simplicity of notations, the range of integration will not be explicitly expressed in the rest of this proof. Throughout the section, the range of integration will not be explicitly denoted unless the range is ambiguous.

is considered as follows.

$$K'_{f_X} - \frac{d}{dx} K'_{f'_X}\Big|_{f_X = f_{X^*}} = 1 + \log f_{X^*}(x) + \alpha_0 + \alpha_1 x^2 = 0. \tag{C.22}$$

Considering the constraints in (C.20) and the equation in (C.22),

$$\begin{aligned} f_{X^*}(x) &= \frac{1}{\sqrt{\pi \frac{1}{4\alpha_1}}} \exp\left\{-\frac{1}{2\frac{1}{2\alpha_1}} x^2\right\} \sqrt{\pi \frac{1}{4\alpha_1}} \exp\{-\alpha_0 - 1\} \\ &= \frac{1}{\sqrt{\frac{\pi m_X^2}{2}}} \exp\left\{-\frac{1}{2m_X^2} x^2\right\}, \qquad x \geq 0, \end{aligned} \tag{C.23}$$

where

$$\begin{aligned} \alpha_0 &= -1 + \frac{1}{2} \log \frac{\pi m_X^2}{2}, \\ \alpha_1 &= \frac{1}{2m_X^2}. \end{aligned}$$

Since

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

and the second-order variation $\delta^2 U[f_{X^*}] > 0$, the optimal solution f_{X^*} is a minimal solution for the variational problem in (C.19).

These first-order and second-order conditions are not sufficient but necessary for the optimal. However, as shown in (C.22) and (C.23), there exists only one solution, a half-normal density function, in the feasible set. Therefore, a half-normal density function is also sufficient in this problem.

Therefore, given the second-order moment, the negative differential entropy $-h(X)$ is minimized (or, equivalently $h(X)$ is maximized) over the set of non-negative ran-

dom variables when $f_X(x)$ is a half-normal density function.

Remark C.1. *Since a half-normal random variable has a fixed mean, if we add a constraint of the mean such as $\mathbb{E}_X[X] = \mu_X$ in (C.20), the inequality in (5.27) is not true except $\mu_X = \sqrt{2m_X^2/\pi}$, where μ_X and m_X^2 are the first-order moment and the second-order moment of X, respectively.*

□

C.4 Proof of Theorem 5.7

Proof. We first construct a functional problem, which represents the inequality in (5.28) and required constraints, as follows:

$$\min_{f_X} \quad \int \frac{f_X'(x)^2}{f_X(x)} dx, \tag{C.24}$$

$$\text{s. t.} \quad \int f_X(x)dx = 1,$$

$$\int x f_X(x)dx = \mu_X,$$

$$\int x^2 f_X(x)dx = m_X^2. \tag{C.25}$$

Using Theorem 5.3, the functional problem in (C.24) is expressed as

$$\min_{f_X} \quad U[f_X], \tag{C.26}$$

where $U[f_X] = \int K(x, f_X, f_X')dx$, $K(x, f_X, f_X') = (f_X'(x)^2/f_X(x)) + \alpha_0 f_X(x) + \alpha_1 x f_X(x) + \alpha_2 x^2 f_X(x)$, and α_0, α_1, and α_2 are the Lagrange multipliers.

Based on Theorem 5.1 or Corollary 5.2, the first-order variation is investigated

as follows.

$$K'_{f_X} - \frac{d}{dx}K'_{f'_X}\Big|_{f_X=f_{X^*}} = \left(\frac{f'_{X^*}(x)}{f_{X^*}(x)}\right)^2 - 2\frac{{f_{X^*}}''(x)}{f_{X^*}(x)} + \alpha_0 + \alpha_1 x + \alpha_2 x^2 = 0, \quad \text{(C.27)}$$

Unlike Theorem 5.4, we cannot directly calculate $f_{X^*}(x)$ from the equation in (C.27). Fortunately, when $f_{X^*}(x)$ is a Gaussian density function, $(f'_{X^*}(x)/f_{X^*}(x))^2 - 2(f''_{X^*}(x)/f_{X^*}(x))$ in (C.27) is expressed as a quadratic function, which is similar to the quadratic parts in (C.27).

Due to the constraints in (C.25), a Gaussian density function $f_{X^*}(x)$ is defined as

$$f_{X^*}(x) = \frac{1}{\sqrt{2\pi(m_X^2 - \mu_X^2)}} \exp\left\{-\frac{1}{2\left(m_X^2 - \mu_X^2\right)}\left(x - \mu_X\right)^2\right\}. \quad \text{(C.28)}$$

By substituting $f_{X^*}(x)$ in (C.28) for the equation in (C.27),

$$\begin{aligned}
& \left(-\frac{1}{m_X^2 - \mu_X^2}\left(x - \mu_X\right)\right)^2 - 2\left\{\left(-\frac{1}{m_X^2 - \mu_X^2}\left(x - \mu_X\right)\right)^2 - \frac{1}{m_X^2 - \mu_X^2}\right\} \\
& +\alpha_0 + \alpha_1 x + \alpha_2 x^2 \\
= \; & -\frac{1}{\left(m_X^2 - \mu_X^2\right)^2}x^2 + \frac{2\mu_X}{\left(m_X^2 - \mu_X^2\right)^2}x + \left(-\frac{\mu_X^2}{\left(m_X^2 - \mu_X^2\right)^2} + \frac{2}{m_X^2 - \mu_X^2}\right) \\
& +\alpha_0 + \alpha_1 x + \alpha_2 x^2 \\
= \; & 0. \qquad \text{(C.29)}
\end{aligned}$$

Since the equations in (C.29) must be satisfied for any x,

$$\begin{aligned}
\alpha_0 &= \frac{\mu_X^2}{\left(m_X^2 - \mu_X^2\right)^2} - \frac{2}{m_X^2 - \mu_X^2}, \\
\alpha_1 &= -\frac{2\mu_X}{\left(m_X^2 - \mu_X^2\right)^2}, \\
\alpha_2 &= \frac{1}{\left(m_X^2 - \mu_X^2\right)^2}. \qquad \text{(C.30)}
\end{aligned}$$

Since

$$K''_{f'_{X^*}f'_{X^*}} = 2\frac{1}{f_{X^*}(x)} > 0 \tag{C.31}$$

and the second-order variation $\delta^2 U[f_{X^*}]$ is positive, the optimal solution f_{X^*} minimizes the variational problem in (C.24).

Therefore, Fisher information $J(X)$ is minimized when $f_X(x)$ is Gaussian. Even though Theorems 5.1, 5.2 are necessary conditions for the minimum, in this case, a Gaussian density function is sufficiently optimal due to the following fact: the objective function is strictly convex and the constraint sets are convex. Therefore, the proof is completed.

Remark C.2. *Even though this result is well-known in the literature (e.g., [2], [43]), this is the first rigorous proof based on calculus of variations.*

Remark C.3. *The constraint related to the first-order moment in (C.25), is not required in this case. Without the constraint, the optimal solution is a Gaussian density function, which has zero mean.*

□

C.5 Proof of Theorem 5.8

Proof. We first construct a functional problem, which represents the inequality in (5.29) and the required constraints as follows:

$$\min_{f_X} \quad \int \boldsymbol{\xi}^T \nabla f_X(\mathbf{x}) \nabla f_X(\mathbf{x})^T \boldsymbol{\xi} \frac{1}{f_X(\mathbf{x})} d\mathbf{x}, \tag{C.32}$$

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

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

$$\int \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) d\mathbf{x} = \boldsymbol{\Omega}_X, \tag{C.33}$$

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

Using Theorem 5.3, the functional problem in (C.32) is expressed as

$$\min_{f_X} \quad U[f_X], \tag{C.34}$$

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

Based on Theorem 5.1 or 5.2, 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.

$$K'_{f_X} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i} K'_{f'_{X_i}} \Big|_{f_X = f_{X^*}} = 0, \tag{C.35}$$

where

$$
\begin{aligned}
K'_{f_X} &= -\frac{\boldsymbol{\xi}^T \nabla f_X(\mathbf{x}) \nabla f_X(\mathbf{x})^T \boldsymbol{\xi}}{f_X(\mathbf{x})^2} + \alpha + \boldsymbol{\zeta}^T \mathbf{x} + \mathbf{x}^T \boldsymbol{\Lambda} \mathbf{x}, \\
\frac{\partial}{\partial x_i} K'_{f'_{X_i}} &= \frac{\partial}{\partial x_i} \left(\frac{2 \sum_{j=1}^{n} \frac{\partial}{\partial x_j} f_X(\mathbf{x}) \xi_i \xi_j}{f_X(\mathbf{x})} \right) \\
&= \frac{2 \sum_{j=1}^{n} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f_X(\mathbf{x}) \xi_i \xi_j}{f_X(\mathbf{x})} - \frac{2 \sum_{j=1}^{n} \frac{\partial}{\partial x_j} f_X(\mathbf{x}) \xi_i \xi_j \frac{\partial}{\partial x_i} f_X(\mathbf{x})}{f_X(\mathbf{x})^2}. \qquad \text{(C.36)}
\end{aligned}
$$

Therefore, the left-hand side of the equation in (C.35) is expressed as

$$
\begin{aligned}
& K'_{f_X} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i} K'_{f'_{X_i}} \\
= & \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial}{\partial x_i} f_X(\mathbf{x}) \frac{\partial}{\partial x_j} f_X(\mathbf{x}) \xi_i \xi_j}{f_X(\mathbf{x})^2} - \frac{2 \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} f_X(\mathbf{x}) \xi_i \xi_j}{f_X(\mathbf{x})} + \alpha + \sum_{i=1}^{n} \zeta_i x_i \\
& + \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{ij} x_i x_j \qquad \text{(C.37)} \\
= & 0. \qquad \text{(C.38)}
\end{aligned}
$$

Unlike Theorem 5.5, we cannot directly calculate $f_{X^*}(\mathbf{x})$ from the equation in (C.35). Fortunately, the first two parts in equation (C.37) are expressed as a quadratic function when $f_{X^*}(\mathbf{x})$ is a multi-variate Gaussian density function, and therefore, the multi-variate Gaussian density function satisfies the equality in (C.38). When $f_{X^*}(\mathbf{x})$ is a multi-variate Gaussian density function:

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

where $\mathbf{\Sigma}_X = \mathbf{\Omega}_X - \boldsymbol{\mu}_X \boldsymbol{\mu}_X^T$,

$$\mathbf{\Sigma}_X^{-1} = \begin{bmatrix} \sigma^2_{X_{11}} & \cdots & \sigma^2_{X_{1n}} \\ \vdots & \ddots & \vdots \\ \sigma^2_{X_{n1}} & \cdots & \sigma^2_{X_{nn}} \end{bmatrix}, \tag{C.39}$$

its partial derivative is expressed as

$$\begin{aligned} \frac{\partial}{\partial x_i} f_{X^*}(\mathbf{x}) &= -\frac{1}{2}\left(\sum_{l=1}^{n} \sigma^2_{X_{il}} \left(x_l - \mu_{X_l}\right) + \sum_{m=1}^{n} \sigma^2_{X_{mi}} \left(x_m - \mu_{X_m}\right)\right) f_{X^*}(\mathbf{x}) \\ \frac{\partial}{\partial x_j}\frac{\partial}{\partial x_i} f_{X^*}(\mathbf{x}) &= -\frac{1}{2}\left(\sigma^2_{X_{ij}} + \sigma^2_{X_{ji}}\right) f_{X^*}(\mathbf{x}) \\ &\quad + \frac{1}{4}\left(\sum_{l=1}^{n} \sigma^2_{X_{il}} \left(x_l - \mu_{X_l}\right) + \sum_{m=1}^{n} \sigma^2_{X_{mi}} \left(x_m - \mu_{X_m}\right)\right) \\ &\quad \times \left(\sum_{l=1}^{n} \sigma^2_{X_{jl}} \left(x_l - \mu_{X_l}\right) + \sum_{m=1}^{n} \sigma^2_{X_{mj}} \left(x_m - \mu_{X_m}\right)\right) f_{X^*}(\mathbf{x}). \end{aligned} \tag{C.40}$$

Without loss of generality, the covariance matrix $\mathbf{\Sigma}_X$ is assumed to be invertible due to the same reason mentioned in Appendix C.2.

By substituting the equations in (C.40) into the equations (C.37), it turns out

that

$$
\begin{aligned}
& 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} \left[\left(x_l - \mu_{X_l} \right) \left(x_m - \mu_{X_m} \right) \left(\frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i \xi_j \left(\sigma^2_{X_{il}} + \sigma^2_{X_{li}} \right) \left(\sigma^2_{X_{jm}} + \sigma^2_{X_{mj}} \right) \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} \left[\left(x_l - \mu_{X_l} \right) \left(x_m - \mu_{X_m} \right) \boldsymbol{\xi}^T \boldsymbol{\Sigma}_{X_{lm}} \boldsymbol{\xi} \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 \\
= \; & \left(\mathbf{x} - \boldsymbol{\mu}_X \right)^T \boldsymbol{\Omega} \left(\mathbf{x} - \boldsymbol{\mu}_X \right) + \boldsymbol{\xi}^T \boldsymbol{\Psi} \boldsymbol{\xi} + \alpha + \boldsymbol{\zeta}^T \mathbf{x} + \mathbf{x}^T \boldsymbol{\Lambda} \mathbf{x} \\
= \; & \left(\mathbf{x}^T \boldsymbol{\Omega} \mathbf{x} + \mathbf{x}^T \boldsymbol{\Lambda} \mathbf{x} \right) + \left(\boldsymbol{\zeta}^T \mathbf{x} - 2 \boldsymbol{\mu}_X^T \boldsymbol{\Omega} \mathbf{x} \right) + \boldsymbol{\mu}_X^T \boldsymbol{\Omega} \boldsymbol{\mu}_X + \boldsymbol{\xi}^T \boldsymbol{\Psi} \boldsymbol{\xi} + \alpha \\
= \; & 0, \qquad \text{(C.41)}
\end{aligned}
$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{X_{lm}} &= \begin{bmatrix} \Sigma^{lm}_{X_{11}} & \cdots & \Sigma^{lm}_{X_{1n}} \\ \vdots & \ddots & \vdots \\ \Sigma^{lm}_{X_{n1}} & \cdots & \Sigma^{lm}_{X_{nn}} \end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1n} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{nn} \end{bmatrix}, \\
\boldsymbol{\Psi} &= \begin{bmatrix} \psi_{11} & \cdots & \psi_{1n} \\ \vdots & \ddots & \vdots \\ \psi_{n1} & \cdots & \psi_{nn} \end{bmatrix}, \quad \boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{n1} & \cdots & \omega_{nn} \end{bmatrix}, \\
\Sigma^{lm}_{X_{ij}} &= \frac{1}{4}\left(\sigma^2_{X_{il}} + \sigma^2_{X_{li}}\right)\left(\sigma^2_{X_{jm}} + \sigma^2_{X_{mj}}\right) \\
&= \sigma^2_{X_{li}}\sigma^2_{X_{jm}}, \qquad i = 1,\ldots,n, \quad j = 1,\ldots,n, \\
&\qquad\qquad\qquad\qquad\quad l = 1,\ldots,n, \quad m = 1,\ldots,n, \\
\psi_{ij} &= 2\sigma^2_{X_{ij}}, \qquad i = 1,\ldots,n, \quad j = 1,\ldots,n, \\
\omega_{lm} &= \boldsymbol{\xi}^T\boldsymbol{\Sigma}_{X_{lm}}\boldsymbol{\xi}, \qquad l = 1,\ldots,n, \quad m = 1,\ldots,n. \qquad \text{(C.42)}
\end{aligned}$$

Therefore, the Lagrange multipliers α and λ_{ij} are defined as

$$\begin{aligned}
\alpha &= -\boldsymbol{\mu}_X^T\boldsymbol{\Omega}\boldsymbol{\mu}_X - \boldsymbol{\xi}^T\boldsymbol{\Psi}\boldsymbol{\xi}, \\
\boldsymbol{\zeta} &= 2\boldsymbol{\Omega}\boldsymbol{\mu}_X, \\
\boldsymbol{\Lambda} &= -\boldsymbol{\Omega}. \qquad \text{(C.43)}
\end{aligned}$$

Since the second-order variation condition is positive

$$K''_{f'_X f'_X} = 2\frac{1}{f_{X^*}(\mathbf{x})} > 0, \qquad \text{(C.44)}$$

the optimal solution $f_{X^*}(\mathbf{x})$ minimizes the variational problem in (C.32). Therefore, the Fisher information matrix $\mathbb{J}(\mathbf{X})$ is minimized when $f_{X^*}(\mathbf{x})$ is a multi-variate

Gaussian, i.e., $\mathbb{J}(\mathbf{X}) \succeq \mathbb{J}(\mathbf{X}_G)$. Even though Theorems 5.1, 5.2 are necessary conditions for the minimum, in this case, the multi-variate Gaussian density function is sufficiently minimum since the objective function is strictly convex and its constraint sets are convex. □

C.6 Proof of Theorem 5.9

Proof. We first construct a functional problem, which represents the inequality in (5.31) and required constraints, as follows:

$$\min_{f_X} \quad \int_0^\infty \frac{f'_X(x)^2}{f_X(x)} dx, \tag{C.45}$$

$$\text{s. t.} \quad \int_0^\infty f_X(x)dx = 1,$$

$$\int_0^\infty x^2 f_X(x)dx = m_X^2. \tag{C.46}$$

Using Theorem 5.3, the functional problem in (C.45) is expressed as

$$\min_{f_X} \quad U[f_X], \tag{C.47}$$

where $U[f_X] = \int K(x, f_X, f'_X)dx$, $K(x, f_X, f'_X) = (f'_X(x)^2/f_X(x)) + f_X(x)\,(\alpha_0 + \alpha_1 x^2)$, and α_0 and α_1 are the Lagrange multipliers.

Based on Theorem 5.1 or 5.2, the first-order and the second-order variation conditions of $U[f_X]$ will be considered as follows. First, the optimal solution $f_{X^*}(x)$ must satisfy the following first-order variation condition:

$$K'_{f_X} - \frac{d}{dx} K'_{f'_X}\Big|_{f_X = f_{X^*}} = \left(\frac{f'_{X^*}(x)}{f_{X^*}(x)}\right)^2 - 2\frac{f''_{X^*}(x)}{f_{X^*}(x)} + \alpha_0 + \alpha_1 x^2 = 0. \tag{C.48}$$

When $f_{X^*}(x)$ is a half-normal density function, $(f'_{X^*}(x)/f_{X^*}(x))^2 - 2(f''_{X^*}(x)/f_{X^*}(x))$

in (C.48) is expressed as a quadratic function, and therefore the equation in (C.48) can be satisfied.

Considering the constraints in (C.46) and $f_{X^*}(x) = (1/\sqrt{\pi m_X^2/2})\exp(-x^2/(2m_X^2))$, where $x > 0$,

$$\begin{aligned} & \left(-\frac{1}{m_X^2}x\right)^2 - 2\left\{\left(-\frac{1}{m_X^2}x\right)^2 - \frac{1}{m_X^2}\right\} + \alpha_0 + \alpha_1 x^2 \\ = \; & -\frac{1}{m_X^4}x^2 + \frac{2}{m_X^2} + \alpha_0 + \alpha_1 x^2 \\ = \; & 0. \end{aligned} \tag{C.49}$$

Since the equation in (C.49) is satisfied for any x,

$$\begin{aligned} \alpha_0 &= -\frac{2}{m_X^2}, \\ \alpha_1 &= \frac{1}{m_X^4}. \end{aligned} \tag{C.50}$$

Now, the second-order variation condition is considered as follows. Since

$$K''_{f'_X f'_X}\Big|_{f_X = f_{X^*}} = 2\frac{1}{f_{X^*}(x)} > 0, \tag{C.51}$$

the second-order variation of $\delta^2 U[f_{X^*}] > 0$, and therefore f_{X^*} minimizes the variational problem in (5.33). Therefore, the Fisher information $J(X)$ is minimized when $f_X(x)$ is half normal. Even though Theorems 5.1, 5.2 are necessary conditions for the minimum, in this case, a half normal density function is sufficiently optimal due to the strict convexity of the objective function and the convexity of the constraint set in (C.45) and (C.46). Therefore, the proof is completed. □

C.7 Proof of Theorem 5.10

Proof. We first construct a functional problem, which represents the inequality in (5.33) and required constraints, as follows:

$$\min_{f_X} \quad \int \frac{f'_X(x)^2}{f_X(x)} dx, \tag{C.52}$$

$$\text{s. t.} \quad \int f_X(x)dx = 1,$$

$$\int x^2 f_X(x)dx = m_X^2. \tag{C.53}$$

Using Theorem 5.3, the functional problem in (C.52) is expressed as

$$\min_{f_X} \quad U[f_X], \tag{C.54}$$

where $U[f_X] = \int K(x, f_X, f'_X)dx$, $K(x, f_X, f'_X) = (f'_X(x)^2/f_X(x)) + f_X(x)\,(\alpha_0 + \alpha_1 x^2)$, and α_0 and α_1 are the Lagrange multipliers.

Based on Theorem 5.1 or Corollary 5.2, by confirming the first-order variation condition, the optimal solution $f_{X^*}(x)$ can be found as follows:

$$K'_{f_X} - \frac{d}{dx} K'_{f'_X}\Big|_{f_X = f_{X^*}} = \left(\frac{f'_{X^*}(x)}{f_{X^*}(x)}\right)^2 - 2\frac{f''_{X^*}(x)}{f_{X^*}(x)} + \alpha_0 + \alpha_1 x^2 = 0. \tag{C.55}$$

Unfortunately, we cannot directly calculate $f_{X^*}(x)$ from the equation in (C.55). Instead, we try to search density functions which satisfy the equation in (C.55). The first two parts, $(f'_{X^*}(x)/f_{X^*}(x))^2 - 2(f''_{X^*}(x)/f_{X^*}(x))$, in equation (C.55) are expressed as a quadratic function when $f_{X^*}(x)$ is a chi density function with 3 degrees of freedom. Therefore, the chi density function satisfies the equation in (C.55).

Considering the constraints in (C.53) and defining $f_{X^*}(x)$ as

$$f_{X^*}(x) = \sqrt{\frac{2}{\pi}}\frac{x^2}{a^3}\exp\left(-\frac{x^2}{2a^2}\right),$$

where $a = \sqrt{m_X^2/3}$, the equation in (C.55) is expressed as

$$\begin{aligned} & \left(\frac{2}{x} - \frac{x}{a^2}\right)^2 - 2\left(\frac{x^2}{a^4} + \frac{2}{x^2} - \frac{5}{a^2}\right) + \alpha_0 + \alpha_1 x^2 \\ = \; & -\frac{1}{a^4}x^2 + \frac{6}{a^2} + \alpha_0 + \alpha_1 x^2 \\ = \; & 0. \end{aligned} \tag{C.56}$$

Since the equation in (C.56) must be satisfied for any x,

$$\begin{aligned} \alpha_0 &= -\frac{6}{a^2} = -\frac{18}{m_X^2}, \\ \alpha_1 &= \frac{1}{a^4} = \left(\frac{3}{m_X^2}\right)^2. \end{aligned} \tag{C.57}$$

Now, using the second-order variation condition, we will confirm that the optimal solution f_{X^*} actually minimizes the variational problem in (C.52) as shown in the following equation:

$$K''_{f'_X f'_X}\Big|_{f_X = f_{X^*}} = 2\frac{1}{f_{X^*}(x)} > 0. \tag{C.58}$$

Therefore, the Fisher information $J(X)$ is minimized when $f_X(x)$ is a chi density function with 3 degrees of freedom and the second-order moment m_X^2. Even though Theorems 5.1, 5.2 are necessary conditions for the minimum, in this case, the chi density function is sufficiently minimum since the variational problem in (C.52) is strictly convex and the constraint set in (C.53) is convex. Therefore, the proof is

completed.

Remark C.4. *Both a half normal density function and a chi-density function satisfy Euler's equation. Therefore, these two functions are the optimal solutions which minimize Fisher information for non-negative random variables. However, a half normal density function does not obey the regularity condition for Fisher information while a chi density function satisfies the regularity condition.*

□

C.8 Proof of Theorem 5.11

Proof. To prove the inequality in (5.34), the following functional problem is constructed:

$$\min_{f_X} \quad \int\int f_X(x) f_{Y|X}(y|x) \left[-\log \left(\int f_X(x) f_{Y|X}(y|x) dx \right) + \log f_X(x) \right] dxdy \quad \text{(C.59)}$$

$$\text{s. t.} \quad \int f_X(x) dx = 1,$$
$$\int x^2 f_X(x) dx = m_X^2. \quad \text{(C.60)}$$

After substituting the random variable Y for $X + W_G$, its density function $f_Y(y)$ is expressed as

$$\begin{aligned} f_Y(y) &= \int f_X(x) f_{Y|X}(y|x) dx \\ &= \int f_X(x) f_W(y-x) dx. \end{aligned} \quad \text{(C.61)}$$

Then, the problem in (C.59) and its constraints in (C.60) are expressed as

$$\begin{aligned}
\min_{f_X, f_Y} \quad & \int\int f_X(x) f_W(y-x) \left[-\log f_Y(y) + \log f_X(x) \right] dxdy \qquad \text{(C.62)} \\
\text{s. t.} \quad & \int\int f_X(x) f_W(y-x) dxdy = 1, \\
& \int\int x^2 f_X(x) f_W(y-x) dxdy = m_X^2, \\
& \int y^2 f_Y(x) dy = m_Y^2, \\
& f_Y(y) = \int f_X(x) f_W(y-x) dx. \qquad \text{(C.63)}
\end{aligned}$$

Using Lagrange multipliers, the functional problem in (C.62) is denoted as

$$\begin{aligned}
\min_{f_X, f_Y} \quad & \int \Bigg(\int f_X(x) f_W(y-x) \left[-\log f_Y(y) + \log f_X(x) + \alpha_0 + \alpha_1 x^2 - \lambda(y) \right] dx \\
& + f_Y(y) \left[\alpha_2 y^2 + \lambda(y) \right] \Bigg) dy. \qquad \text{(C.64)}
\end{aligned}$$

Define a functional U as

$$U[f_X, f_Y] \;=\; \int \left(\int K(x, y, f_X, f_Y) dx \right) + \tilde{K}(y, f_Y) dy, \qquad \text{(C.65)}$$

where* $K(x, y, f_X, f_Y) = f_X(x) f_W(y-x)[-\log f_Y(y) + \log f_X(x) + \alpha_0 + \alpha_1 x^2 - \lambda(y)]$, and $\tilde{K}(y, f_Y) = f_Y(y) \left[\alpha_2 y^2 + \lambda(y) \right]$.

Now, we have to find f_{X^*} and f_{Y^*} which satisfy the first-order variation condition,

*The equation in (C.65) is denoted as $\int (\int K dx) + \tilde{K} dy$ for the simplicity of notation.

$\delta U = 0$.

$$
\begin{aligned}
& K'_{f_X}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} \\
= \;& f_W(y-x)\left(-\log f_{Y^*}(y) + \log f_{X^*}(x) + \alpha_0 + \alpha_1 x^2 + 1 - \lambda(y)\right) \\
= \;& 0 \qquad \text{(C.66)} \\
& \int K'_{f_Y} dx + \tilde{K}'_{f_Y}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} \\
= \;& -\int f_{X^*}(x) f_W(y-x) dx \frac{1}{f_{Y^*}(y)} + \alpha_2 y^2 + \lambda(y) \\
= \;& 0. \qquad \text{(C.67)}
\end{aligned}
$$

Since the equations in (C.66) and (C.67) are satisfied for any x and y,

$$
\begin{aligned}
& -\log f_{Y^*}(y) + c_Y - \lambda(y) = 0, \\
& \log f_{X^*}(x) + \alpha_0 + \alpha_1 x^2 + 1 - c_{Y^*} = 0, \\
& \lambda(y) = 1 - \alpha_2 y^2, \qquad \text{(C.68)}
\end{aligned}
$$

where c_Y is a constant.

Therefore,

$$
\begin{aligned}
f_{X^*}(x) &= \exp\left(-\alpha_0 - \alpha_1 x^2 - 1 + c_Y\right), \\
f_{Y^*}(y) &= \exp\left(c_Y - 1 + \alpha_2 y^2\right),
\end{aligned}
$$

and $f_{X^*}(x)$ and $f_{Y^*}(x)$ are re-written as

$$\begin{aligned} f_X(x) &= \exp\left(-\alpha_0 - \alpha_1 x^2 - 1 + c_Y\right) \\ &= \frac{1}{\sqrt{2\pi\frac{1}{2\alpha_1}}} \exp\left\{-\frac{1}{2\frac{1}{2\alpha_1}}x^2\right\} \sqrt{2\pi\frac{1}{2\alpha_1}} \exp\left\{-\alpha_0 - 1 + c_Y\right\}, \qquad \text{(C.69)} \\ f_Y(y) &= \exp\left(c_Y - 1 + \alpha_2 y^2\right) \\ &= \frac{1}{\sqrt{2\pi\left(-\frac{1}{2\alpha_2}\right)}} \exp\left\{-\frac{1}{2\left(-\frac{1}{2\alpha_2}\right)}y^2\right\} \sqrt{2\pi\left(-\frac{1}{2\alpha_2}\right)} \exp\left\{c_Y - 1\right\}. \qquad \text{(C.70)} \end{aligned}$$

Considering the constraints in (C.63), the Lagrange multipliers in (C.69) and (C.70) are expressed as

$$\begin{aligned} \alpha_0 &= -1 + c_Y + \frac{1}{2}\log 2\pi m_X^2 \\ &= \frac{1}{2}\log\frac{m_X^2}{m_Y^2}, \\ \alpha_1 &= \frac{1}{2m_X^2}, \\ \alpha_2 &= -\frac{1}{2m_Y^2}, \\ c_Y &= 1 - \frac{1}{2}\log 2\pi m_Y^2. \end{aligned} \qquad \text{(C.71)}$$

Therefore, Gaussian density functions f_{X^*} and f_{Y^*} satisfy the first-order variation condition, $\delta U = 0$.

Now, the second-order variation condition must be considered, and, for the minimum, it requires the positive definiteness of the matrix,

$$\left[\begin{array}{cc} K''_{f_X f_X} & K''_{f_X f_Y} \\ K''_{f_Y f_X} & K''_{f_Y f_Y} \end{array}\right]\Bigg|_{f_X = f_{X^*}, f_Y = f_{Y^*}}. \qquad \text{(C.72)}$$

The elements of the matrix in (C.72) are calculated as

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

and the matrix in (C.72) is positive definite. Therefore, $\delta^2 U > 0$, the optimal solutions f_{X^*} and f_{Y^*} minimize the variational problem in (C.62). Even though the optimal solutions are necessarily optimal, there are only Gaussian density functions f_{X^*} and f_{Y^*} in the feasible set, i.e., Gaussian density functions f_{X^*} and f_{Y^*} are the only ones which satisfy the equations in (C.66) and (C.67). Therefore, these optimal solutions are actually sufficient.

In conclusion, given the second-order moment, a Gaussian random variable X_G minimizes the mutual information $I(X + W_G; W_G)$, and the proof is completed. □

C.9 Proof of Theorem 5.12

Proof. To prove the inequality in (5.35), we first construct a functional problem as follows:

$$\begin{aligned}
\min_{f_X} \quad & -\int\int f_X(\mathbf{x}) f_{Y|X}(\mathbf{y}|\mathbf{x}) \log \left(\int f_X(\mathbf{x}) f_{Y|X}(\mathbf{y}|\mathbf{x}) d\mathbf{x} \right) d\mathbf{x} d\mathbf{y} \quad \text{(C.74)} \\
& + \int\int f_X(\mathbf{x}) f_{Y|X}(\mathbf{y}|\mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x} d\mathbf{y} \\
\text{s. t.} \quad & \int f_X(\mathbf{x}) d\mathbf{x} = 1, \\
& \int \mathbf{x} f_X(\mathbf{x}) d\mathbf{x} = \boldsymbol{\mu}_X, \\
& \int \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) d\mathbf{x} = \boldsymbol{\Omega}_X. \quad \text{(C.75)}
\end{aligned}$$

By substituting the random vector $\mathbf{Y}$ for $\mathbf{X} + \mathbf{W}_G$, where $\mathbf{X}$ and $\mathbf{W}_G$ are independent of each other, in (5.35), its density function $f_Y(\mathbf{y})$ and conditional density function $f_{Y|X}(\mathbf{y}|\mathbf{x})$ are expressed as

$$\begin{aligned}
f_Y(\mathbf{y}) &= \int f_X(\mathbf{x}) f_{Y|X}(\mathbf{y}|\mathbf{x}) d\mathbf{x}, \quad \text{(C.76)} \\
f_{Y|X}(\mathbf{y}|\mathbf{x}) &= f_W(\mathbf{y} - \mathbf{x}), \quad \text{(C.77)}
\end{aligned}$$

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 $f_{Y|X}(\mathbf{y}|\mathbf{x})$, and appropriately changing the constrains in (C.75), the variational

problem in (C.74) can be expressed as

$$\min_{f_X, f_Y} \quad \int\int f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) \left[-\log f_Y(\mathbf{y}) + \log f_X(\mathbf{x})\right] d\mathbf{x} d\mathbf{y} \tag{C.78}$$

$$\begin{aligned}
\text{s. t.} \quad & \int\int f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x} d\mathbf{y} = 1, \\
& \int\int \mathbf{x} f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x} d\mathbf{y} = \boldsymbol{\mu}_X, \\
& \int\int \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x} d\mathbf{y} = \boldsymbol{\Omega}_X, \\
& \int f_Y(\mathbf{y}) d\mathbf{y} = 1, \\
& \int \mathbf{y} f_Y(\mathbf{y}) d\mathbf{y} = \boldsymbol{\mu}_Y, \\
& \int \mathbf{y}\mathbf{y}^T f_Y(\mathbf{x}) d\mathbf{y} = \boldsymbol{\Omega}_Y, \\
& f_Y(\mathbf{y}) = \int f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x}.
\end{aligned} \tag{C.79}$$

The functional problem in (C.78) is changed into the following equivalent problem:

$$\begin{aligned}
\min_{f_X, f_Y} \quad & \int \left(\int f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) \left[-\log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) + \alpha_0 + \sum_{i=1}^{n} \zeta_i x_i \right.\right. \\
& \left.\left. + \sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} x_i x_j - \lambda(\mathbf{y}) \right] d\mathbf{x} \right) \\
& + f_Y(\mathbf{y}) \left[\alpha_1 + \sum_{i=1}^{n} \eta_i y_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \theta_{ij} y_i y_j + \lambda(\mathbf{y}) \right] d\mathbf{y},
\end{aligned} \tag{C.80}$$

where $\mathbf{x}^T = [x_1, \ldots, x_n]$, $\mathbf{y}^T = [y_1, \ldots, y_n]$, and α_0, α_1, ζ_i, γ_{ij}, η_i, θ_{ij}, and $\lambda(\mathbf{y})$ are Lagrange multipliers.

Let's define 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

$$\begin{aligned} K(\mathbf{x},\mathbf{y},f_X,f_Y) &= f_X(\mathbf{x})f_W(\mathbf{y}-\mathbf{x})[-\log f_Y(\mathbf{y})+\log f_X(\mathbf{x})+\alpha_0+\sum_{i=1}^{n}\zeta_i x_i \\ &\quad +\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}x_i x_j-\lambda(\mathbf{y})], \\ \tilde{K}(\mathbf{y},f_Y) &= f_Y(\mathbf{y})\left[\alpha_1+\sum_{i=1}^{n}\eta_i y_i+\sum_{i=1}^{n}\sum_{j=1}^{n}\theta_{ij}y_i y_j+\lambda(\mathbf{y})\right]. \end{aligned} \tag{C.81}$$

Based on the first-order variation condition, we can find the optimal solution, f_{X^*} and f_{Y^*}, as follows.

$$\begin{aligned} & K'_{f_X}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} \\ =\ & f_W(\mathbf{y}-\mathbf{x})\Bigg(-\log f_{Y^*}(\mathbf{y})+\log f_{X^*}(\mathbf{x})+\alpha_0+\sum_{i=1}^{n}\zeta_i x_i \\ & \qquad +\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}x_i x_j+1-\lambda(\mathbf{y})\Bigg) \\ =\ & f_W(\mathbf{y}-\mathbf{x})\left(-\log f_{Y^*}(\mathbf{y})+\log f_{X^*}(\mathbf{x})+\alpha_0+\boldsymbol{\zeta}\mathbf{x}^T+\mathbf{x}^T\boldsymbol{\Gamma}\mathbf{x}+1-\lambda(\mathbf{y})\right) \\ =\ & 0 \end{aligned} \tag{C.82}$$

$$\begin{aligned} & \int K'_{f_Y}d\mathbf{x}+\tilde{K}'_{f_Y}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} \\ =\ & -\int f_{X^*}(\mathbf{x})f_W(\mathbf{y}-\mathbf{x})d\mathbf{x}\frac{1}{f_{Y^*}(\mathbf{y})}+\alpha_1+\sum_{i=1}^{n}\eta_i y_i+\sum_{i=1}^{n}\sum_{j=1}^{n}\theta_{ij}y_i y_j+\lambda(\mathbf{y}) \\ =\ & -\int f_{X^*}(\mathbf{x})f_W(\mathbf{y}-\mathbf{x})d\mathbf{x}\frac{1}{f_{Y^*}(\mathbf{y})}+\alpha_1+\boldsymbol{\eta}^T\mathbf{y}+\mathbf{y}^T\boldsymbol{\Theta}\mathbf{y}+\lambda(\mathbf{y}) \\ =\ & 0, \end{aligned} \tag{C.83}$$

where

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nn} \end{bmatrix}, \quad \mathbf{\Theta} = \begin{bmatrix} \theta_{11} & \cdots & \theta_{1n} \\ \vdots & \ddots & \vdots \\ \theta_{n1} & \cdots & \theta_{nn} \end{bmatrix}, \tag{C.84}$$

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

Since the equalities in (C.82) and (C.83) must be satisfied for any $\mathbf{x}$ and $\mathbf{y}$,

$$\begin{aligned} 0 &= -\log f_{Y^*}(\mathbf{y}) - \lambda(\mathbf{y}), \\ 0 &= \log f_{X^*}(\mathbf{x}) + \alpha_0 + \boldsymbol{\zeta}\mathbf{x}^T + \mathbf{x}^T\mathbf{\Gamma}\mathbf{x} + 1, \\ \lambda(\mathbf{y}) &= 1 - \alpha_1 - \boldsymbol{\eta}^T\mathbf{y} - \mathbf{y}^T\mathbf{\Theta}\mathbf{y}, \end{aligned} \tag{C.85}$$

and

$$\begin{aligned} f_{X^*}(\mathbf{x}) &= \exp\left(-\alpha_0 - \boldsymbol{\zeta}^T\mathbf{x} - \mathbf{x}^T\mathbf{\Gamma}\mathbf{x} - 1\right), \\ f_{Y^*}(\mathbf{y}) &= \exp\left(-1 + \alpha_1 + \boldsymbol{\eta}^T\mathbf{y} + \mathbf{y}^T\mathbf{\Theta}\mathbf{y}\right). \end{aligned} \tag{C.86}$$

Considering the constraints in (C.79), $f_{X^*}(\mathbf{x})$ and $f_{Y^*}(\mathbf{x})$ in (C.86) are expressed as

$$\begin{aligned} f_{X^*}(\mathbf{x}) &= (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}_X|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_X\right)^T \mathbf{\Sigma}_X^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_X\right)\right\}, \\ &= \exp\left\{-\frac{1}{2}\log\left(2\pi\right)^n |\mathbf{\Sigma}_X| - \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_X^{-1}\mathbf{x} + \boldsymbol{\mu}_X^T\mathbf{\Sigma}_X^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_X^T\mathbf{\Sigma}_X^{-1}\boldsymbol{\mu}_X\right\} \\ &= \exp\left(-\alpha_0 - \boldsymbol{\zeta}^T\mathbf{x} - \mathbf{x}^T\mathbf{\Gamma}\mathbf{x} - 1\right), \\ f_{Y^*}(\mathbf{y}) &= (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}_Y|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \boldsymbol{\mu}_Y\right)^T \mathbf{\Sigma}_Y^{-1}\left(\mathbf{y} - \boldsymbol{\mu}_Y\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\log\left(2\pi\right)^n |\mathbf{\Sigma}_Y| - \frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}_Y^{-1}\mathbf{y} + \boldsymbol{\mu}_Y^T\mathbf{\Sigma}_Y^{-1}\mathbf{y} - \frac{1}{2}\boldsymbol{\mu}_Y^T\mathbf{\Sigma}_Y^{-1}\boldsymbol{\mu}_Y\right\} \\ &= \exp\left(-1 + \alpha_1 + \boldsymbol{\eta}^T\mathbf{y} + \mathbf{y}^T\mathbf{\Theta}\mathbf{y}\right), \end{aligned} \tag{C.87}$$

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 a covariance matrix of $\mathbf{W}_G$. Based on the equations in (C.87),

$$\begin{aligned}
\alpha_0 &= -1 + \frac{1}{2}\log(2\pi)^n |\boldsymbol{\Sigma}_X| + \frac{1}{2}\boldsymbol{\mu}_X^T \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\mu}_X, \\
\alpha_1 &= 1 - \frac{1}{2}\log(2\pi)^n |\boldsymbol{\Sigma}_Y| - \frac{1}{2}\boldsymbol{\mu}_Y^T \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{\mu}_Y, \\
\boldsymbol{\Gamma} &= \frac{1}{2}\boldsymbol{\Sigma}_X^{-1}, \\
\boldsymbol{\zeta} &= -\boldsymbol{\mu}_X^T \boldsymbol{\Sigma}_X^{-1}, \\
\boldsymbol{\Theta} &= \frac{1}{2}\boldsymbol{\Sigma}_Y^{-1}, \\
\boldsymbol{\eta} &= -\boldsymbol{\mu}_Y^T \boldsymbol{\Sigma}_Y^{-1}.
\end{aligned} \tag{C.88}$$

Therefore, the optimal solutions f_{X^*} and f_{Y^*} are multi-variate Gaussian density functions (without loss of generality, we assume that the covariance matrix $\boldsymbol{\Sigma}_X$ is invertible due to the reason mentioned in Appendix C.2).

Now, by confirming the second-order variation condition, we will show that the optimal solutions f_{X^*} and f_{Y^*} minimize the variational functional in (C.78). Based on Theorem 5.2, we will show that the following matrix is positive definite:

$$\begin{bmatrix} K''_{f_X f_X} & K''_{f_X f_Y} \\ K''_{f_Y f_X} & K''_{f_Y f_Y} \end{bmatrix} \succ \mathbf{0}. \tag{C.89}$$

Since the elements of the matrix in (C.89) are defined as

$$\begin{aligned}
K''_{f_X f_X}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} &= \frac{f_W(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})},\\
K''_{f_Y f_Y}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} &= \frac{f_{X^*}(\mathbf{x})f_W(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})^2},\\
K''_{f_X f_Y}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} &= -\frac{f_W(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})},\\
K''_{f_Y f_X}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} &= -\frac{f_W(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})},
\end{aligned} \tag{C.90}$$

the matrix is a positive definite matrix, and therefore $\delta^2 U > 0$. Therefore, the optimal solutions f_{X^*} and f_{Y^*} actually minimize the variational functional in (C.78). Even though these optimal solutions are necessarily optimal, there exists only one solution, which is a multi-variate Gaussian density function, which satisfies Euler's equation in (C.82) and (C.83). Therefore, f_{X^*} and f_{Y^*} are also sufficient in this case.

Remark C.5. *The constraints related to the mean vectors in (C.79) are unnecessary. Without these constraints, the optimal solutions are still multi-variate Gaussian density functions but the mean vectors are changed into zero.*

□

C.10 Proof of Theorem 5.13

Proof. To prove the entropy power inequality, we slightly change the inequality in (5.36) into the following relationship:

$$h(\tilde{X}+\tilde{W}) \geq a_X^2 h(\tilde{X}) + a_W^2 h(\tilde{W}) - \log a_X - \log a_W, \tag{C.91}$$

where $\tilde{X} = a_X X$ and $\tilde{X} = a_W W$. Since a_X and a_W are constants, they do not affect the optimization, and we can ignore these two terms.

Based on the inequality in (C.91) and required constraints, construct the following functional problem (for the simplicity of the notation, we simply denote $\tilde{X}$ and $\tilde{W}$ as X and W):

$$\min_{f_X, f_W, f_Y} \int\int f_X(x) f_W(y-x) \left(-\log f_Y(y) + a_X^2 \log f_X(x) + a_W^2 \log f_W(y-x) \right) dxdy \tag{C.92}$$

$$\begin{aligned}
\text{s.t.} \quad & \int\int f_X(x) f_W(y-x)\, dxdy = 1, \\
& \int\int y^2 f_X(x) f_W(y-x)\, dxdy = m_{Y^*}^2, \\
& \int\int x^2 f_X(x) f_W(y-x) dxdy = m_{X^*}^2, \\
& \int\int (y-x)^2 f_X(x) f_W(y-x) dxdy = m_{W^*}^2, \\
& -\int\int f_X(x) f_W(y-x) \log f_X(x) dxdy = p_X, \\
& -\int\int f_X(x) f_W(y-x) \log f_W(y-x) dxdy = p_W, \\
& f_Y(y) = \int f_X(x) f_W(y-x)\, dx,
\end{aligned} \tag{C.93}$$

where $m_{X^*}^2$, $m_{W^*}^2$, and $m_{Y^*}^2$ denote the second-order moments of the optimal solutions of X, W, and Y, respectively. The constraints related to the second-order moments mean that all random variables have finite second-order moments. Also, the constraints related to p_X and p_W mean that random variables X and W have finite entropies, respectively, where p_X and p_W are constants. Without loss of generality, the zero mean condition is assumed for all random variables (in the case of non-zero mean, all constraints related to the second-order moments are changed into constraints related to the covariance matrices).

Using Lagrange multipliers, the problem in (C.92) and the constraints in (C.93)

are reformulated as the following equivalent problem:

$$\min_{f_X, f_W, f_Y} \quad \int \left(\int K(x, y, f_X, f_W, f_Y) dx \right) + \tilde{K}(y, f_Y) dy, \tag{C.94}$$

where

$$\begin{aligned} K(x, y, f_X, f_W, f_Y) &= f_X(x) f_W(y - x) \Big(- \log f_Y(y) + (a_X^2 - \lambda_X) \log f_X(x) \\ &\quad + (a_W^2 - \lambda_W) \log f_W(y - x) + \alpha_0 + \alpha_1 y^2 + \alpha_2 x^2 \\ &\quad + \alpha_3 (y - x)^2 - \lambda(y) \Big), \\ \tilde{K}(y, f_Y) &= \lambda(y) f_Y(y). \end{aligned} \tag{C.95}$$

The first-order partial derivative is expressed as

$$\begin{aligned} & K'_{f_X} \Big|_{f_X = f_{X^*}, f_W = f_{W^*}, f_Y = f_{Y^*}} \\ = \; & f_{W^*}(y - x) \Big(- \log f_{Y^*}(y) + (a_X^2 - \lambda_X) \log f_{X^*}(x) + (a_W^2 - \lambda_W) \log f_{W^*}(y - x) \\ & + \alpha_0 + \alpha_1 y^2 + \alpha_2 x^2 + \alpha_3 (y - x)^2 - \lambda(y) + a_X^2 - \lambda_X \Big), \\ & K'_{f_W} \Big|_{f_X = f_{X^*}, f_W = f_{W^*}, f_Y = f_{Y^*}} \\ = \; & f_{X^*}(x) \Big(- \log f_{Y^*}(y) + (a_X^2 - \lambda_X) \log f_{X^*}(x) + (a_W^2 - \lambda_W) \log f_{W^*}(y - x) \\ & + \alpha_0 + \alpha_1 y^2 + \alpha_2 x^2 + \alpha_3 (y - x)^2 - \lambda(y) + a_W^2 - \lambda_W \Big), \\ & \left(\int K dx + \tilde{K} \right)'_{f_Y} \Big|_{f_X = f_{X^*}, f_W = f_{W^*}, f_Y = f_{Y^*}} \\ = \; & - \int f_{X^*}(x) f_{W^*}(y - x) \, dx \frac{1}{f_{Y^*}(y)} + \lambda(y). \end{aligned} \tag{C.96}$$

Due to the first-order variation condition, $\delta U[f_{X^*}, f_{W^*}, f_{Y^*}] = 0$, the optimal

solutions f_{X^*}, f_{W^*}, and f_{Y^*}, must satisfy the following relationships:

$$\begin{aligned}
-\log f_{Y^*}(y) + \alpha_1 y^2 - \lambda(y) + c_Y &= 0, \\
(a_X^2 - \lambda_X)\log f_{X^*}(x) + \alpha_2 x^2 + c_X &= 0, \\
(a_W^2 - \lambda_W)\log f_{W^*}(y-x) + \alpha_3 (y-x)^2 + \alpha_0 + a_W^2 - \lambda_W - c_X - c_Y &= 0, \\
-1 + \lambda(y) &= 0, \\
a_W^2 - \lambda_W - a_X^2 + \lambda_X &= 0, \qquad \text{(C.97)}
\end{aligned}$$

and therefore,

$$\begin{aligned}
f_{Y^*}(y) &= \exp\left\{\alpha_1 y^2 - \lambda(y) + c_Y\right\}, \\
f_{X^*}(x) &= \exp\left\{\frac{1}{a_X^2 - \lambda_X}\left(-\alpha_2 x^2 - c_X\right)\right\}, \\
f_{W^*}(y-x) &= \exp\left\{\frac{1}{a_W^2 - \lambda_W}\left(-\alpha_3 (y-x)^2 - \alpha_0 - a_W^2 + \lambda_W + c_X + c_Y\right)\right\}, \\
\lambda(y) &= 1. \qquad \text{(C.98)}
\end{aligned}$$

Considering the constraints in (C.93), the equations in (C.98) are expressed as

$$
\begin{aligned}
f_{Y^*}(y) &= \frac{1}{\sqrt{2\pi\left(-\frac{1}{2\alpha_1}\right)}} \exp\left\{-\frac{1}{2\left(-\frac{1}{2\alpha_1}\right)} y^2\right\} \sqrt{2\pi\left(-\frac{1}{2\alpha_1}\right)} \exp\left\{-\lambda(y) + c_Y\right\} \\
&= \frac{1}{\sqrt{2\pi m_{Y^*}^2}} \exp\left\{-\frac{1}{2m_{Y^*}^2} y^2\right\}, \\
f_{X^*}(x) &= \frac{1}{\sqrt{2\pi\left(\frac{a_X^2-\lambda_X}{2\alpha_2}\right)}} \exp\left\{-\frac{1}{2\left(\frac{a_X^2-\lambda_X}{2\alpha_2}\right)} x^2\right\} \\
&\quad \times \sqrt{2\pi\left(\frac{a_X^2-\lambda_X}{2\alpha_2}\right)} \exp\left\{-\frac{c_X}{a_X^2-\lambda_X}\right\} \\
&= \frac{1}{\sqrt{2\pi m_{X^*}^2}} \exp\left\{-\frac{1}{2m_{X^*}^2} x^2\right\}, \\
f_{W^*}(y-x) &= \frac{1}{\sqrt{2\pi\left(\frac{a_W^2-\lambda_W}{2\alpha_3}\right)}} \exp\left\{-\frac{1}{2\left(\frac{a_W^2-\lambda_W}{2\alpha_3}\right)} (y-x)^2\right\} \\
&\quad \times \sqrt{2\pi\left(\frac{a_W^2-\lambda_W}{2\alpha_3}\right)} \exp\left\{\frac{-\alpha_0 - a_W^2 + \lambda_W + c_X + c_Y}{a_W^2-\lambda_W}\right\} \\
&= \frac{1}{\sqrt{2\pi m_{W^*}^2}} \exp\left\{-\frac{1}{2m_{W^*}^2} (y-x)^2\right\},
\end{aligned}
\tag{C.99}
$$

where

$$
\begin{aligned}
\alpha_0 &= -(a_W^2 - \lambda_W) + c_X + c_Y + \frac{a_W^2 - \lambda_W}{2} \log 2\pi m_{W^*}^2 \\
\alpha_1 &= -\frac{1}{2m_{Y^*}^2}, \\
\alpha_2 &= \frac{a_X^2 - \lambda_X}{2m_{X^*}^2}, \qquad \text{(C.100)} \\
\alpha_3 &= \frac{a_W^2 - \lambda_W}{2m_{W^*}^2}, \qquad \text{(C.101)} \\
c_X &= \frac{a_X^2 - \lambda_X}{2} \log 2\pi m_{X^*}^2 \\
c_Y &= 1 - \frac{1}{2} \log 2\pi m_{Y^*}^2, \\
a_X^2 - \lambda_X &= a_W^2 - \lambda_W \geq 1, \qquad \text{(C.102)} \\
m_{X^*}^2 &= \frac{1}{2\pi e} \exp\{2p_X\}, \\
m_{W^*}^2 &= \frac{1}{2\pi e} \exp\{2p_W\}, \\
m_{Y^*}^2 &= m_{X^*}^2 + m_{W^*}^2 \\
&= \frac{1}{2\pi e} \exp\{2p_X\} + \frac{1}{2\pi e} \exp\{2p_W\}.
\end{aligned}
$$

The inequality in (C.102) is due to the second-order variation condition, which will be justified next.

Consider now the conditions for the second variation of the functional problem:

$$
\begin{aligned}
K''_{f_X f_X}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= \frac{(a_X^2-\lambda_X)f_{W^*}(y-x)}{f_{X^*}(x)},\\
K''_{f_W f_W}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= \frac{(a_W^2-\lambda_W)f_{X^*}(x)}{f_{W^*}(y-x)},\\
\left(\int K dx+\tilde{K}\right)''_{f_Y f_Y}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= \frac{f_{X^*}(x)f_{W^*}(y-x)}{f_{Y^*}(y)^2},\\
K''_{f_X f_W}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= a_W^2-\lambda_W,\\
K''_{f_W f_X}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= a_X^2-\lambda_X,\\
K''_{f_X f_Y}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= -\frac{f_{W^*}(y-x)}{f_{Y^*}(y)},\\
K''_{f_Y f_X}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= -\frac{f_{W^*}(y-x)}{f_{Y^*}(y)},\\
K''_{f_W f_Y}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= -\frac{f_{X^*}(x)}{f_{Y^*}(y)},\\
K''_{f_Y f_W}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= -\frac{f_{X^*}(x)}{f_{Y^*}(y)}.
\end{aligned}
\tag{C.103}
$$

To satisfy $\delta^2 J \geq 0$, the following condition must hold:

$$
\begin{aligned}
&\begin{bmatrix} h_X & h_W & h_Y \end{bmatrix}
\begin{bmatrix} K''_{f_X f_X} & K''_{f_X f_W} & K''_{f_X f_Y} \\ K''_{f_W f_X} & K''_{f_W f_W} & K''_{f_W f_Y} \\ K''_{f_Y f_X} & K''_{f_Y f_W} & K''_{f_Y f_Y} \end{bmatrix}
\begin{bmatrix} h_X \\ h_W \\ h_Y \end{bmatrix}\Bigg|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}}\\
=\;& K''_{f_X f_X}h_X^2+K''_{f_W f_W}h_W^2+K''_{f_Y f_Y}h_Y^2+(K''_{f_X f_W}+K''_{f_W f_X})h_X h_W\\
&+(K''_{f_W f_Y}+K''_{f_Y f_W})h_W h_Y+(K''_{f_X f_Y}+K''_{f_Y f_X})h_Y h_X\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}}\\
\geq\;& 0.
\end{aligned}
\tag{C.104}
$$

Using the defined quantities in (C.103), the equation in (C.104) is expressed as

follows:

$$
\begin{aligned}
& K''_{f_{X^*}f_{X^*}}h_X^2 + K''_{f_{W^*}f_{W^*}}h_W^2 + K''_{f_{Y^*}f_{Y^*}}h_Y^2 + (K''_{f_{X^*}f_{W^*}} + K''_{f_{W^*}f_{X^*}})h_X h_W \\
& +(K''_{f_{W^*}f_{Y^*}} + K''_{f_{Y^*}f_{W^*}})h_W h_Y + (K''_{f_{X^*}f_{Y^*}} + K''_{f_{Y^*}f_{X^*}})h_Y h_X \\
= \;& \frac{(a_X^2 - \lambda_X) f_{W^*}(y-x)}{f_{X^*}(x)} h_X(x)^2 + \frac{(a_W^2 - \lambda_W) f_{X^*}(x)}{f_{W^*}(y-x)} h_W(y-x)^2 \\
& + \frac{f_{X^*}(x) f_{W^*}(y-x)}{f_{Y^*}(y)^2} h_Y(y)^2 + 2(a_W^2 - \lambda_W) h_X(x) h_W(y-x) \\
& - 2\frac{f_{X^*}(x)}{f_{Y^*}(y)} h_W(y-x) h_Y(y) - 2\frac{f_{W^*}(y-x)}{f_{Y^*}(y)} h_X(x) h_Y(y) \\
= \;& \frac{f_{W^*}(y-x)}{f_{X^*}(x)} \Bigg((a_W^2 - \lambda_W) h_X(x)^2 + (a_W^2 - \lambda_W) \frac{f_{X^*}(x)^2}{f_{W^*}(y-x)^2} h_W(y-x)^2 \\
& \qquad + \frac{f_{X^*}(x)^2}{f_{Y^*}(y)^2} h_Y(y)^2 + 2(a_W^2 - \lambda_W) \frac{f_{X^*}(x)}{f_{W^*}(y-x)} h_X(x) h_W(y-x) \\
& \qquad - 2\frac{f_{X^*}(x)^2}{f_{W^*}(y-x) f_{Y^*}(y)} h_W(y-x) h_Y(y) - 2\frac{f_{X^*}(x)}{f_{Y^*}(y)} h_X(x) h_Y(y) \Bigg) \\
= \;& \frac{f_{W^*}(y-x)}{f_{X^*}(x)} \left(h_X(x) + \frac{f_{X^*}(x)}{f_{W^*}(y-x)} h_W(y-x) - \frac{f_{X^*}(x)}{f_{Y^*}(y)} h_Y(y) \right)^2 \\
\geq \;& 0, \qquad \text{(C.105)}
\end{aligned}
$$

where $a_W^2 - \lambda_W = a_X^2 - \lambda_X \geq 1$.

Therefore, the optimal solutions, f_{X^*}, f_{W^*}, and f_{Y^*}, minimize the variational problem in (C.92). Even though f_{X^*}, f_{W^*}, and f_{Y^*} are necessarily optimal, they are sufficiently optimal since only Gaussian density functions are in the feasible constraints set. □

C.11 Proof of Theorem 5.14

Proof. Similar to the proof shown in Appendix C.10, we first construct the following functional problem:

$$
\begin{aligned}
\min_{f_X, f_W, f_Y} \; & \int\int f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) \left(-\log f_Y(\mathbf{y}) + a_X^2 \log f_X(\mathbf{x}) + a_W^2 \log f_W(\mathbf{y}-\mathbf{x}) \right) d\mathbf{x} d\mathbf{y} \\
\text{s.t.} \quad & \int\int f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x})\, d\mathbf{x} d\mathbf{y} = 1, \\
& \int\int \mathbf{y}\mathbf{y}^T f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x})\, d\mathbf{x} d\mathbf{y} = \mathbf{\Omega}_{X^*} + \mathbf{\Omega}_{W^*}, \\
& \int\int \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x})\, d\mathbf{x} d\mathbf{y} = \mathbf{\Omega}_{X^*}, \\
& \int\int (\mathbf{y}-\mathbf{x})(\mathbf{y}-\mathbf{x})^T f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x})\, d\mathbf{x} d\mathbf{y} = \mathbf{\Omega}_{W^*}, \\
& -\int\int f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x} d\mathbf{y} = p_X, \\
& -\int\int f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) \log f_W(\mathbf{y}-\mathbf{x})\, d\mathbf{x} d\mathbf{y} = p_W, \\
& f_Y(\mathbf{y}) = \int f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x})\, d\mathbf{x},
\end{aligned} \tag{C.106}
$$

where p_X and p_W are constants, and the constraints related to these constants mean the entropies of $\mathbf{X}$ and $\mathbf{W}$ are finite. The matrices $\mathbf{\Omega}_{X^*}$ and $\mathbf{\Omega}_{W^*}$ denote the correlation matrices of the optimal random vectors $\mathbf{X}^*$ and $\mathbf{W}^*$, respectively. The constraints related to these correlation matrices mean that the correlation matrices of random vectors $\mathbf{X}$ and $\mathbf{W}$ exist. Without loss of generality, the mean vectors of $\mathbf{X}$ and $\mathbf{W}$ are assumed to be zero (If $\mathbf{X}$ and $\mathbf{W}$ have non-zero mean vectors, the constraints related to the correlation matrices are changed into the ones related to the covariance matrices.).

Using Lagrange multipliers, the problem in (C.106) is changed into the following

optimization problem:

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

where

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}, f_X, f_W, f_Y) &= f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) \Big(- \log f_Y(\mathbf{y}) + (a_X^2 - \lambda_X) \log f_X(\mathbf{x}) \\ &\quad + (a_W^2 - \lambda_W) \log f_W(\mathbf{y} - \mathbf{x}) + \alpha + \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} y_i y_j \\ &\quad + \sum_{i=1}^{n} \sum_{j=1}^{n} \phi_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \theta_{ij} (y_i - x_i)(y_j - x_j) - \lambda(\mathbf{y}) \Big), \\ \tilde{K}(\mathbf{y}, f_Y) &= \lambda(\mathbf{y}) f_Y(\mathbf{y}). \end{aligned} \tag{C.107}$$

Then,

$$\begin{aligned} K'_{f_X} &= f_W(\mathbf{y} - \mathbf{x}) \Big(- \log f_Y(\mathbf{y}) + (a_X^2 - \lambda_X) \log f_X(\mathbf{x}) + (a_W^2 - \lambda_W) \log f_W(\mathbf{y} - \mathbf{x}) \\ &\quad + \alpha + \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} y_i y_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \phi_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \theta_{ij} (y_i - x_i)(y_j - x_j) \\ &\quad - \lambda(\mathbf{y}) + a_X^2 - \lambda_X \Big), \\ K'_{f_W} &= f_X(\mathbf{x}) \Big(- \log f_Y(\mathbf{y}) + (a_X^2 - \lambda_X) \log f_X(\mathbf{x}) + (a_W^2 - \lambda_W) \log f_W(\mathbf{y} - \mathbf{x}) + \alpha \\ &\quad + \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} y_i y_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \phi_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \theta_{ij} (y_i - x_i)(y_j - x_j) \\ &\quad - \lambda(\mathbf{y}) + a_W^2 - \lambda_W \Big), \\ \left(\int K d\mathbf{x} + \tilde{K} \right)'_{f_Y} &= - \int f_X(\mathbf{x}) f_W(\mathbf{y} - \mathbf{x}) \, d\mathbf{x} \frac{1}{f_Y(\mathbf{y})} + \lambda(\mathbf{y}). \end{aligned} \tag{C.108}$$

To satisfy $\delta U[f_{X^*}, f_{W^*}, f_{Y^*}] = 0$,

$$
\begin{aligned}
&-\log f_{Y^*}(\mathbf{y}) + \sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} y_i y_j - \lambda(\mathbf{y}) + c_Y = 0, \\
&(a_X^2 - \lambda_X) \log f_{X^*}(\mathbf{x}) + \sum_{i=1}^{n}\sum_{j=1}^{n} \phi_{ij} x_i x_j + c_X = 0, \\
&(a_W^2 - \lambda_W) \log f_{W^*}(\mathbf{y}-\mathbf{x}) + \sum_{i=1}^{n}\sum_{j=1}^{n} \theta_{ij} (y_i - x_i)(y_j - x_j) + \alpha \\
&\qquad\qquad + a_W^2 - \lambda_W - c_X - c_Y = 0, \\
&-1 + \lambda(\mathbf{y}) = 0, \\
&a_W^2 - \lambda_W - a_X^2 + \lambda_X = 0.
\end{aligned} \tag{C.109}
$$

Since the equations in (C.109) must be satisfied for any $\mathbf{x}$ and $\mathbf{y}$, the optimal solution f_{X^*}, f_{W^*}, and f_{W^*} are expressed as

$$
\begin{aligned}
f_{Y^*}(\mathbf{y}) &= \exp\left\{ \sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} y_i y_j - \lambda(\mathbf{y}) + c_Y \right\} \\
&= \exp\left\{ \mathbf{y}^T \mathbf{\Gamma} \mathbf{y} - 1 + c_Y \right\}, \\
f_{X^*}(\mathbf{x}) &= \exp\left\{ \frac{1}{a_X^2 - \lambda_X} \left(- \sum_{i=1}^{n}\sum_{j=1}^{n} \phi_{ij} x_i x_j - c_X \right) \right\} \\
&= \exp\left\{ -\frac{1}{a_X^2 - \lambda_X} \left(\mathbf{x}^T \mathbf{\Phi} \mathbf{x} + c_X \right) \right\}, \\
f_{W^*}(\mathbf{y}-\mathbf{x}) &= \exp\left\{ \frac{1}{a_W^2 - \lambda_W} \left(- \sum_{i=1}^{n}\sum_{j=1}^{n} \theta_{ij} (y_i - x_i)(y_j - x_j) \right.\right. \\
&\qquad\qquad \left.\left. - \alpha - a_W^2 + \lambda_W + c_X + c_Y \right) \right\} \\
&= \exp\left\{ -\frac{1}{a_W^2 - \lambda_W} \left((\mathbf{y}-\mathbf{x})^T \mathbf{\Theta} (\mathbf{y}-\mathbf{x}) + \alpha + a_W^2 - \lambda_W - c_X - c_Y \right) \right\} \\
\lambda(\mathbf{y}) &= 1.
\end{aligned} \tag{C.110}
$$

Considering the constraints in (C.106), the equations in (C.110) are further processed as

$$
\begin{aligned}
f_{Y^*}(\mathbf{y}) &= \frac{1}{(2\pi)^{\frac{n}{2}}\left|-\frac{1}{2}\mathbf{\Gamma}^{-1}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\left(-\frac{1}{2}\mathbf{\Gamma}^{-1}\right)^{-1}\mathbf{y}\right\} \\
&\quad \times(2\pi)^{\frac{n}{2}}\left|-\frac{1}{2}\mathbf{\Gamma}^{-1}\right|^{\frac{1}{2}}\exp\left\{-\lambda(\mathbf{y})+c_Y\right\} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\mathbf{\Omega}_{X^*}+\mathbf{\Omega}_{W^*}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\left(\mathbf{\Omega}_{X^*}+\mathbf{\Omega}_{W^*}\right)^{-1}\mathbf{y}\right\}, \\
f_{X^*}(\mathbf{x}) &= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\frac{a_X^2-\lambda_X}{2}\mathbf{\Phi}^{-1}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathbf{x}^T\left(\frac{a_X^2-\lambda_X}{2}\mathbf{\Phi}^{-1}\right)^{-1}\mathbf{x}\right\} \\
&\quad \times(2\pi)^{\frac{n}{2}}\left|\frac{a_X^2-\lambda_X}{2}\mathbf{\Phi}^{-1}\right|^{\frac{1}{2}}\exp\left\{-\frac{c_X}{a_X^2-\lambda_X}\right\} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\mathbf{\Omega}_{X^*}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{\Omega}_{X^*}^{-1}\mathbf{x}\right\}, \\
f_{W^*}(\mathbf{y}-\mathbf{x}) &= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\frac{a_W^2-\lambda_W}{2}\mathbf{\Theta}^{-1}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{x})^T\left(\frac{a_W^2-\lambda_W}{2}\mathbf{\Theta}^{-1}\right)^{-1}(\mathbf{y}-\mathbf{x})\right\} \\
&\quad \times(2\pi)^{\frac{n}{2}}\left|\frac{a_W^2-\lambda_W}{2}\mathbf{\Theta}^{-1}\right|^{\frac{1}{2}}\exp\left\{\frac{-\alpha-a_W^2+\lambda_W+c_X+c_Y}{a_W^2-\lambda_W}\right\} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\mathbf{\Omega}_{W^*}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{x})^T\mathbf{\Omega}_{W^*}^{-1}(\mathbf{y}-\mathbf{x})\right\}, \qquad \text{(C.111)}
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha &= -(a_W^2 - \lambda_W) + c_X + c_Y + \frac{a_W^2 - \lambda_W}{2} \log\left((2\pi)^n \left|\mathbf{\Omega}_{W^*}\right|\right) \\
\mathbf{\Gamma} &= -\frac{1}{2}\left(\mathbf{\Omega}_{X^*} + \mathbf{\Omega}_{W^*}\right)^{-1}, \\
\mathbf{\Phi} &= \frac{a_X^2 - \lambda_X}{2}\mathbf{\Omega}_{X^*}^{-1}, \qquad \text{(C.112)} \\
\mathbf{\Theta} &= \frac{a_W^2 - \lambda_W}{2}\mathbf{\Omega}_{W^*}^{-1}, \qquad \text{(C.113)} \\
c_X &= \frac{a_X^2 - \lambda_X}{2} \log\left((2\pi)^n \left|\mathbf{\Omega}_{X^*}\right|\right) \\
c_Y &= 1 - \frac{1}{2} \log\left((2\pi)^n \left|\mathbf{\Omega}_{X^*} + \mathbf{\Omega}_{W^*}\right|\right), \\
a_W^2 - \lambda_W &= a_X^2 - \lambda_X \geq 1, \qquad \text{(C.114)} \\
\left|\mathbf{\Omega}_{X^*}\right| &= \left(\frac{1}{2\pi e} \exp\left\{\frac{2}{n} p_X\right\}\right)^n, \qquad \text{(C.115)} \\
\left|\mathbf{\Omega}_{W^*}\right| &= \left(\frac{1}{2\pi e} \exp\left\{\frac{2}{n} p_W\right\}\right)^n. \qquad \text{(C.116)}
\end{aligned}
$$

Without loss of generality, the matrices $\mathbf{\Omega}_{X^*}$ and $\mathbf{\Omega}_{W^*}$ are assumed to be invertible due to the same reasons mentioned in Appendix C.2. The relationships in (C.114) are obtained based on the second-order variation condition, which will be shown later in this proof.

Therefore, we can always find the Lagrange multipliers.

Now, consider the conditions for the second-order variation condition:

$$
\begin{aligned}
K''_{f_X f_X}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= \frac{(a_X^2-\lambda_X)f_{W^*}(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})},\\
K''_{f_W f_W}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= \frac{(a_W^2-\lambda_W)f_{X^*}(\mathbf{x})}{f_{W^*}(\mathbf{y}-\mathbf{x})},\\
\left(\int K d\mathbf{x}+\tilde{K}\right)''_{f_Y f_Y}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= \frac{f_{X^*}(\mathbf{x})f_{W^*}(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})^2},\\
K''_{f_X f_W}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= a_W^2-\lambda_W,\\
K''_{f_W f_X}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= a_X^2-\lambda_X,\\
K''_{f_X f_Y}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= -\frac{f_{W^*}(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})},\\
K''_{f_Y f_X}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= -\frac{f_{W^*}(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})},\\
K''_{f_W f_Y}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= -\frac{f_{X^*}(\mathbf{x})}{f_{Y^*}(\mathbf{y})},\\
K''_{f_Y f_W}\Big|_{f_X=f_{X^*},f_W=f_{W^*},f_Y=f_{Y^*}} &= -\frac{f_{X^*}(\mathbf{x})}{f_{Y^*}(\mathbf{y})}.
\end{aligned}
\tag{C.117}
$$

To satisfy $\delta^2 U[f_{X^*}, f_{W^*}, f_{Y^*}] \geq 0$, the following must hold:

$$
\begin{aligned}
&\begin{bmatrix} h_X & h_W & h_Y \end{bmatrix}
\begin{bmatrix}
K''_{f_{X^*}f_{X^*}} & K''_{f_{X^*}f_{W^*}} & K''_{f_{X^*}f_{Y^*}}\\
K''_{f_{W^*}f_{X^*}} & K''_{f_{W^*}f_{W^*}} & K''_{f_{W^*}f_{Y^*}}\\
K''_{f_{Y^*}f_{X^*}} & K''_{f_{Y^*}f_{W^*}} & K''_{f_{Y^*}f_{Y^*}}
\end{bmatrix}
\begin{bmatrix} h_X \\ h_W \\ h_Y \end{bmatrix}\\
= \;& K''_{f_{X^*}f_{X^*}}h_X^2 + K''_{f_{W^*}f_{W^*}}h_W^2 + K''_{f_{Y^*}f_{Y^*}}h_Y^2 + (K''_{f_{X^*}f_{W^*}} + K''_{f_{W^*}f_{X^*}})h_X h_W\\
& + (K''_{f_{W^*}f_{Y^*}} + K''_{f_{Y^*}f_{W^*}})h_W h_Y + (K''_{f_{X^*}f_{Y^*}} + K''_{f_{Y^*}f_{X^*}})h_Y h_X \qquad \text{(C.118)}\\
\geq \;& 0.
\end{aligned}
$$

Using the defined quantities in (C.117), the equation in (C.118) is expressed as

follows:

$$
\begin{aligned}
& K''_{f_{X^*}f_{X^*}}h_X^2 + K''_{f_{W^*}f_{W^*}}h_W^2 + K''_{f_{Y^*}f_{Y^*}}h_Y^2 + (K''_{f_{X^*}f_{W^*}} + K''_{f_{W^*}f_{X^*}})h_X h_W \\
& +(K''_{f_{W^*}f_{Y^*}} + K''_{f_{Y^*}f_{W^*}})h_W h_Y + (K''_{f_{X^*}f_{Y^*}} + K''_{f_{Y^*}f_{X^*}})h_Y h_X \\
= \; & \frac{(a_X^2 - \lambda_X) f_{W^*}(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})} h_X(\mathbf{x})^2 + \frac{(a_W^2 - \lambda_W) f_{X^*}(\mathbf{x})}{f_{W^*(\mathbf{y}-\mathbf{x})}} h_W(\mathbf{y}-\mathbf{x})^2 \\
& + \frac{f_{X^*}(\mathbf{x}) f_{W^*}(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})^2} h_Y(\mathbf{y})^2 + 2\frac{(a_W^2 - \lambda_W)}{a_W} h_X(\mathbf{x}) h_W(\mathbf{y}-\mathbf{x}) \\
& - 2\frac{f_{X^*}(\mathbf{x})}{f_{Y^*}(\mathbf{y})} h_W(\mathbf{y}-\mathbf{x}) h_Y(\mathbf{y}) - 2\frac{f_{W^*}(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})} h_X(\mathbf{x}) h_Y(\mathbf{y}) \\
= \; & \frac{f_{W^*}(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})} \Bigg((a_W^2 - \lambda_W) h_X(\mathbf{x})^2 + (a_W^2 - \lambda_W) \frac{f_{X^*}(\mathbf{x})^2}{f_{W^*}(\mathbf{y}-\mathbf{x})^2} h_W(\mathbf{y}-\mathbf{x})^2 \\
& \qquad + \frac{f_{X^*}(\mathbf{x})^2}{f_{Y^*}(\mathbf{y})^2} h_Y(\mathbf{y})^2 + 2(a_W^2 - \lambda_W) \frac{f_{X^*}(\mathbf{x})}{f_{W^*}(\mathbf{y}-\mathbf{x})} h_X(\mathbf{x}) h_W(\mathbf{y}-\mathbf{x}) \\
& \qquad - 2\frac{f_{X^*}(\mathbf{x})^2}{f_{W^*}(\mathbf{y}-\mathbf{x}) f_{Y^*}(\mathbf{y})} h_W(\mathbf{y}-\mathbf{x}) h_Y(\mathbf{y}) - 2\frac{f_{X^*}(\mathbf{x})}{f_{Y^*}(\mathbf{y})} h_X(\mathbf{x}) h_Y(\mathbf{y}) \Bigg) \\
\geq \; & \frac{f_{W^*}(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})} \left(h_X(\mathbf{x}) + \frac{f_{X^*}(\mathbf{x})}{f_{W^*}(\mathbf{y}-\mathbf{x})} h_W(\mathbf{y}-\mathbf{x}) - \frac{f_{X^*}(\mathbf{x})}{f_{Y^*}(\mathbf{y})} h_Y(\mathbf{y}) \right)^2 \\
\geq \; & 0, \qquad \text{(C.119)}
\end{aligned}
$$

where $a_W^2 - \lambda_W = a_X^2 - \lambda_X \geq 1$.

Therefore, the optimal solutions, f_{X^*}, f_{W^*}, and f_{Y^*}, minimize the variational problem in (C.106). Even though f_{X^*}, f_{W^*}, and f_{Y^*} are necessarily minimum solutions, multi-variate Gaussian density functions are the only ones in the feasible set. However, unlike Theorem 5.13, the correlation matrices are not explicitly defined as shown in (C.115) and (C.116), and there are more than one Gaussian density functions which satisfy the first-order and the second-order variation conditions. Therefore, we need an additional step to determine the correlation matrices $\mathbf{\Omega}_{X^*}$ and $\mathbf{\Omega}_{W^*}$ as follows.

Based on the first-order and the second-order variation conditions, we know the

optimal solutions of the functional problem in (C.106) are multi-variate Gaussian density functions f_{X^*} and f_{W^*} whose correlation matrices are $\mathbf{\Omega}_{X^*}$ and $\mathbf{\Omega}_{W^*}$, respectively. Therefore, the inequality in (5.37) is expressed as

$$\begin{aligned} & h(a_X\mathbf{X} + a_W\mathbf{W}) - a_X^2 h(\mathbf{X}) - a_W^2 h(\mathbf{W}) \\ \geq\ & h(a_X\mathbf{X}^* + a_W\mathbf{W}^*) - a_X^2 h(\mathbf{X}^*) - a_W^2 h(\mathbf{W}^*) \\ =\ & \frac{1}{2}\log(2\pi e)^n \left|a_X^2\mathbf{\Omega}_{X^*} + a_W^2\mathbf{\Omega}_{W^*}\right| - \frac{a_X^2}{2}\log(2\pi e)^n \left|\mathbf{\Omega}_{X^*}\right| - \frac{a_W^2}{2}\log(2\pi e)^n \left|\mathbf{\Omega}_{W^*}\right| \\ \geq\ & 0. \end{aligned} \tag{C.120}$$

Since $\log|\cdot|$ is a concave function and $a_X^2 + a_W^2 = 1$, the inequality in (C.120) is proved using Jensen's inequality. Therefore,

$$h(a_X\mathbf{X} + a_W\mathbf{W}) \geq a_X^2 h(\mathbf{X}) + a_W^2 h(\mathbf{W}), \tag{C.121}$$

and the proof is completed.

Remark C.6. *In (C.120), equality holds if and only if* $\mathbf{\Omega}_{X^*} = \mathbf{\Omega}_{W^*}$. *Since the optimal multi-variate Gaussian density functions have zero mean vectors, in this case, the correlation matrices are equal to the covariance matrices. Therefore, the equality condition requires identical covariance matrices. However, the equality condition is not required in the proof of EPI.*

□

C.12 Proof of Theorem 5.15

Proof. Now, construct the following variational problem, which represents the inequality in (5.38) and required constraints:

$$\min_{f_X, f_Y} \int\int f_X(x) f_W(y-x) \left(-\mu \log f_Y(y) + \log f_X(x) + \mu(\mu-1) \log f_W(y-x)\right) dxdy \tag{C.122}$$

$$\begin{aligned}
\text{s.t.} \quad & \int\int f_X(x) f_W(y-x) dxdy = 1, \\
& \int\int (y-\mu_Y)^2 f_X(x) f_W(y-x) dxdy = \sigma_{Y^*}^2, \\
& \int\int (y-\mu_Y)^2 f_X(x) f_W(y-x) dxdy = \int\int (x-\mu_X)^2 f_X(x) f_W(y-x) dxdy \\
& \qquad\qquad + \int\int (y-x-\mu_W)^2 f_X(x) f_W(y-x) dxdy, \\
& \int\int (x-\mu_X)^2 f_X(x) f_W(y-x) dxdy \leq r^2, \\
& -\int\int f_X(x) f_W(y-x) \log f_X(x) dxdy = p, \\
& f_Y(y) = \int\int f_X(x) f_W(y-x) dxdy,
\end{aligned} \tag{C.123}$$

where p and r are constants, and $\sigma_{Y^*}^2$ stands for the variance of the optimal solution Y.

Using Lagrange multipliers, the functional problem in (C.122) is expressed as

$$\min_{f_X, f_Y} \quad \int \left(\int K(x, y, f_X, f_Y) dx \right) + \tilde{K}(y, f_Y) dy, \tag{C.124}$$

where

$$
\begin{aligned}
K(x,y,f_X,f_Y) = & f_X(x) f_W(y-x)\Big(-\mu \log f_Y(y) + \log f_X(x) + \mu(\mu-1)\log f_W(y-x) \\
& +\alpha_0 + \beta_1 (y-\mu_Y)^2 + \beta_2 (y-\mu_Y)^2 - \beta_2 (x-\mu_X)^2 \\
& -\beta_2 (y-x-\mu_W)^2 + \beta_3 (x-\mu_X)^2 - \gamma_1 \log f_X(x) - \lambda(y)\Big), \\
\tilde{K}(y,f_Y) &= \lambda(y) f_Y(y). \qquad \text{(C.125)}
\end{aligned}
$$

Due to the first-order variation condition,

$$
\begin{aligned}
& K'_{f_X}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} \\
= \; & f_W(y-x)\Big(-\mu \log f_{Y^*}(y) + \log f_{X^*}(x) + \mu(\mu-1)\log f_W(y-x) + \alpha_0 \\
& +\beta_1 (y-\mu_Y)^2 + \beta_2 (y-\mu_Y)^2 - \beta_2 (x-\mu_X)^2 - \beta_2 (y-x-\mu_W)^2 \\
& +\beta_3 (x-\mu_X)^2 - \gamma_1 \log f_{X^*}(x) - \lambda(y) + 1 - \gamma_1\Big) \\
= \; & 0, \qquad \text{(C.126)} \\
& \int K'_{f_Y} dx + \tilde{K}'_{f_Y}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} \\
= \; & -\mu \frac{\int f_{X^*}(x) f_W(y-x) dx}{f_{Y^*}(y)} + \lambda(y) \\
= \; & 0. \qquad \text{(C.127)}
\end{aligned}
$$

Since the equations in (C.126) and (C.127) must be satisfied for any x and y,

$$
\begin{aligned}
\lambda(y) &= \mu, \\
f_{Y^*}(y) &= \exp\left\{\frac{1}{\mu}\left(\left(\beta_1+\beta_2\right)\left(y-\mu_{Y^*}\right)^2+c_Y\right)\right\} \\
&= \frac{1}{\sqrt{2\pi\left(-\frac{\mu}{2(\beta_1+\beta_2)}\right)}}\exp\left\{-\frac{1}{2\left(-\frac{\mu}{2(\beta_1+\beta_2)}\right)}\left(y-\mu_{Y^*}\right)^2\right\} \\
&\quad\times\sqrt{2\pi\left(-\frac{\mu}{2\left(\beta_1+\beta_2\right)}\right)}\exp\left\{\frac{c_Y}{\mu}\right\} \\
f_W(y-x) &= \exp\left\{\frac{\beta_2}{\mu\left(\mu-1\right)}\left(y-x-\mu_W\right)^2-\frac{c_W}{\mu\left(\mu-1\right)}\right\} \\
&= \frac{1}{\sqrt{2\pi\left(-\frac{\mu(\mu-1)}{2(\beta_2)}\right)}}\exp\left\{-\frac{1}{2\left(-\frac{\mu(\mu-1)}{2(\beta_2)}\right)}\left(y-x-\mu_W\right)^2\right\} \\
&\quad\times\sqrt{2\pi\left(-\frac{\mu\left(\mu-1\right)}{2\left(\beta_2\right)}\right)}\exp\left\{-\frac{c_W}{\mu\left(\mu-1\right)}\right\}, \\
f_{X^*}(x) &= \exp\left\{\frac{1}{1-\gamma_1}\left(\left(\beta_2-\beta_3\right)\left(x-\mu_{X^*}\right)^2-\alpha_0+\mu-1+\gamma_1+c_W+c_Y\right)\right\} \\
&= \frac{1}{\sqrt{2\pi\left(-\frac{1-\gamma_1}{2(\beta_2-\beta_3)}\right)}}\exp\left\{-\frac{1}{2\left(-\frac{1-\gamma_1}{2(\beta_2-\beta_3)}\right)}\left(x-\mu_{X^*}\right)^2\right\} \\
&\quad\times\sqrt{2\pi\left(-\frac{1-\gamma_1}{2\left(\beta_2-\beta_3\right)}\right)}\exp\left\{\frac{-\alpha_0+\mu-1+\gamma_1+c_W+c_Y}{1-\gamma_1}\right\}. \qquad \text{(C.128)}
\end{aligned}
$$

Considering the constraints in (C.123), the equations in (C.128) are further pro-

cessed as follows:

$$
\begin{aligned}
f_{Y^*}(y) &= \frac{1}{\sqrt{2\pi\left(-\frac{\mu}{2(\beta_1+\beta_2)}\right)}}\exp\left\{-\frac{1}{2\left(-\frac{\mu}{2(\beta_1+\beta_2)}\right)}\left(y-\mu_{Y^*}\right)^2\right\} \\
&\quad \times\sqrt{2\pi\left(-\frac{\mu}{2\left(\beta_1+\beta_2\right)}\right)}\exp\left\{\frac{c_Y}{\mu}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma_{Y^*}^2}}\exp\left\{-\frac{1}{2\sigma_{Y^*}^2}\left(y-\mu_{Y^*}\right)^2\right\}, \\
f_W(y-x) &= \frac{1}{\sqrt{2\pi\left(-\frac{\mu(\mu-1)}{2(\beta_2)}\right)}}\exp\left\{-\frac{1}{2\left(-\frac{\mu(\mu-1)}{2(\beta_2)}\right)}\left(y-x-\mu_W\right)^2\right\} \\
&\quad \times\sqrt{2\pi\left(-\frac{\mu\left(\mu-1\right)}{2\left(\beta_2\right)}\right)}\exp\left\{-\frac{c_W}{\mu\left(\mu-1\right)}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma_W^2}}\exp\left\{-\frac{1}{2\sigma_W^2}\left(y-x-\mu_W\right)^2\right\}, \\
f_{X^*}(x) &= \frac{1}{\sqrt{2\pi\left(-\frac{1-\gamma_1}{2(\beta_2-\beta_3)}\right)}}\exp\left\{-\frac{1}{2\left(-\frac{1-\gamma_1}{2(\beta_2-\beta_3)}\right)}\left(x-\mu_{X^*}\right)^2\right\} \\
&\quad \times\sqrt{2\pi\left(-\frac{1-\gamma_1}{2\left(\beta_2-\beta_3\right)}\right)}\exp\left\{\frac{-\alpha_0+\mu-1+\gamma_1+c_W+c_Y}{1-\gamma_1}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma_{X^*}^2}}\exp\left\{-\frac{1}{2\sigma_{X^*}^2}\left(x-\mu_{X^*}\right)^2\right\}, \qquad \text{(C.129)}
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_0 &= \mu - (1-\gamma_1) + c_W + c_Y + \frac{1-\gamma_1}{2}\log\left(2\pi\sigma_{X^*}^2\right) \\
&= \frac{\mu(\mu-1)}{2}\log\left(2\pi m_W^2\right) - \frac{\mu}{2}\log\left(2\pi m_Y^2\right) + \frac{\mu}{2}\log\left(2\pi m_X^2\right), \\
\beta_1 &= -\beta_2 - \frac{\mu}{2\sigma_{Y^*}^2} \\
&= \frac{\mu(\mu-1)}{2\sigma_W^2} - \frac{\mu}{2\sigma_{Y^*}^2} \\
\beta_2 &= -\frac{\mu(\mu-1)}{2\sigma_W^2}, \\
\beta_3 &= \beta_2 + \frac{(1-\gamma_1)}{2\sigma_{X^*}^2} \\
&= -\frac{\mu(\mu-1)}{2\sigma_W^2} + \frac{(1-\gamma_1)}{2\sigma_{X^*}^2} \qquad \text{(C.130)} \\
&\geq 0, \\
c_W &= \frac{\mu(\mu-1)}{2}\log\left(2\pi\sigma_W^2\right), \\
c_Y &= -\frac{\mu}{2}\log\left(2\pi\sigma_{Y^*}^2\right), \\
\sigma_{X^*}^2 &= \frac{1}{2\pi e}\exp\{2p\} \leq r^2, \qquad \text{(C.131)} \\
\sigma_{Y^*}^2 &= \sigma_{X^*}^2 + \sigma_W^2, \\
\gamma_1 &\leq 1-\mu. \qquad \text{(C.132)}
\end{aligned}
$$

The constant p must be chosen to satisfy the inequality in (C.131) due to Theorem 5.4. The inequality in (C.132) is due to the second-order variation condition, which will be presented later in this proof. Therefore, by appropriately choosing p, the Lagrange multipliers always exist, and therefore, the necessary optimal solutions, which are Gaussian, exist.

To make the second variation positive, we need the positive-definiteness of the

following matrix:

$$\left[\begin{array}{cc} K''_{f_X f_X} & K''_{f_X f_Y} \\ K''_{f_Y f_X} & K''_{f_Y f_Y} \end{array} \right] \Bigg|_{f_X = f_{X^*}, f_Y = f_{Y^*}} \tag{C.133}$$

and it requires the following:

$$\begin{aligned} & \left[\begin{array}{cc} h_X & h_Y \end{array} \right] \left[\begin{array}{cc} K''_{f_X f_X} & K''_{f_X f_Y} \\ K''_{f_Y f_X} & K''_{f_Y f_Y} \end{array} \right] \left[\begin{array}{c} h_X \\ h_Y \end{array} \right] \Bigg|_{f_X = f_{X^*}, f_Y = f_{Y^*}} \\ = \; & K''_{f_X f_X} h_X^2 + K''_{f_Y f_Y} h_Y^2 + (K''_{f_X f_Y} + K''_{f_Y f_X}) h_Y h_X \Big|_{f_X = f_{X^*}, f_Y = f_{Y^*}} \\ > \; & 0, \end{aligned} \tag{C.134}$$

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

$$\begin{aligned} K''_{f_X f_X} &= \frac{(1-\gamma_1) f_W(y-x)}{f_X(x)}, \\ K''_{f_X f_Y} &= -\frac{\mu f_W(y-x)}{f_Y(y)}, \\ K''_{f_Y f_X} &= -\frac{\mu f_W(y-x)}{f_Y(y)}, \\ K''_{f_Y f_Y} &= \frac{\mu f_X(x) f_W(y-x)}{f_Y(y)^2}, \end{aligned} \tag{C.135}$$

the equation in (C.134) requires the following:

$$\begin{aligned} & \frac{(1-\gamma_1) f_W(y-x)}{f_{X^*}(x)} h_X(x)^2 - 2\frac{\mu f_W(y-x)}{f_{Y^*}(y)} h_X(x) h_Y(y) + \frac{\mu f_{X^*}(x) f_W(y-x)}{f_{Y^*}(y)^2} h_Y(y)^2 \\ \geq \; & \frac{\mu f_W(y-x)}{f_{X^*}(x)} \left(h_X(x) - \frac{f_{X^*}(x)}{f_{Y^*}(y)} h_Y(y) \right)^2, \end{aligned} \tag{C.136}$$

where $\gamma_1 \leq 1 - \mu$. Similar to the complementary slackness in KKT conditions, when

$\beta_3 = 0$ in (C.130), $\sigma_{X^*}^2 = (1-\gamma_1)\,\mu^{-1}(\mu-1)^{-1}\sigma_{W^*}^2$, and it requires $(1-\gamma_1)\,\mu^{-1}(\mu-1)^{-1}\sigma_{W^*}^2 < r^2$ (If $\gamma_1 = 1-\mu$, then $\sigma_{X^*}^2 = (\mu-1)^{-1}\sigma_{W^*}^2$). Otherwise, $\sigma_{X^*}^2 = r^2 \leq (1-\gamma_1)\,\mu^{-1}(\mu-1)^{-1}\sigma_{W^*}^2$.

In conclusion, the Gaussian density function, whose variance is $\sigma_{X^*}^2$, minimizes the variational problem in (C.122), and the proof is completed.

Remark C.7. *Unlike other theorems shown in this section, Theorem 5.15 only requires to find necessarily optimal solutions, which is the same as Theorem in [32].*

□

C.13 Proof of Theorem 5.16

Proof. We first construct the following variational problem (without loss of generality, we assume the mean vectors of $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Y}$ are zeros. (cf. Appendix C.12)):

$$\min_{f_X, f_Y} \iint f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) \left(-\mu \log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) + \mu\,(\mu-1) \log f_W(\mathbf{y}-\mathbf{x})\right) d\mathbf{x}d\mathbf{y} \tag{C.137}$$

$$\begin{aligned}
\text{s.t.} \quad & \iint f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x}d\mathbf{y} = 1, \\
& \iint \mathbf{y}\mathbf{y}^T f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x}d\mathbf{y} = \iint \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x}d\mathbf{y}, \\
& \qquad\qquad + \iint (\mathbf{y}-\mathbf{x})\,(\mathbf{y}-\mathbf{x})^T f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x}d\mathbf{y}, \\
& \iint \mathbf{x}\mathbf{x}^T f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x}d\mathbf{y} \preceq \mathbf{\Sigma}, \\
& \iint \mathbf{y}\mathbf{y}^T f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x}d\mathbf{y} = \mathbf{\Sigma}_{Y^*}, \\
& -\iint f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x}d\mathbf{y} = p_X, \\
& f_Y(\mathbf{y}) = \iint f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x}d\mathbf{y},
\end{aligned} \tag{C.138}$$

where p_X is a constant, and $\mathbf{\Sigma}_{Y^*}$ is the covariance matrix of the optimal solution of $\mathbf{Y}$. Without loss of generality, the matrix $\mathbf{\Sigma}$ is assumed to be a positive definite matrix due to the same reason mentioned in [32].

This problem is more appropriately changed as follows:

$$\min_{f_X, f_Y} \iint f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) \left(-\mu \log f_Y(\mathbf{y}) + \log f_X(\mathbf{x}) + \mu(\mu-1) \log f_W(\mathbf{y}-\mathbf{x})\right) d\mathbf{x} d\mathbf{y} \tag{C.139}$$

$$\begin{aligned}
\text{s.t.} \quad & \iint f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x} d\mathbf{y} = 1, \\
& \iint \left(y_i y_j - x_i x_j - (y-x)_i (y-x)_j\right) f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x} d\mathbf{y} = 0, \\
& \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\int \int x_i x_j \xi_i \xi_j f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x} d\mathbf{y} \right) \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij}^2 \xi_i \xi_j, \\
& \iint y_i y_j f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x} d\mathbf{y} = \sigma_{Y_{ij}^*}^2, \\
& - \iint f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) \log f_X(\mathbf{x}) d\mathbf{x} d\mathbf{y} = p_X, \\
& f_Y(\mathbf{y}) = \iint f_X(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x}) d\mathbf{x} d\mathbf{y},
\end{aligned} \tag{C.140}$$

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

Using Lagrange multipliers, the functional problem in (C.139) and the constraints in (C.140) are expressed as

$$\min_{f_X, f_Y} \quad \int \left(\int K(\mathbf{x}, \mathbf{y}, f_X, f_Y) d\mathbf{x} \right) + \tilde{K}(\mathbf{y}, f_Y) d\mathbf{y}, \tag{C.141}$$

where

$$K(\mathbf{x},\mathbf{y},f_X,f_Y)=f_X(\mathbf{x})f_W(\mathbf{y}-\mathbf{x})\Big(-\mu\log f_Y(\mathbf{y})+\log f_X(\mathbf{x})+\mu\left(\mu-1\right)\log f_W(\mathbf{y}-\mathbf{x})$$
$$+\alpha_0+\sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\gamma_{ij}y_iy_j-\gamma_{ij}x_ix_j-\gamma_{ij}\left(y-x\right)_i\left(y-x\right)_j+\theta x_ix_j\xi_i\xi_j$$
$$+\phi_{ij}y_iy_j\Big)-\alpha_1\log f_X(\mathbf{x})-\lambda(\mathbf{y})\Big),$$
$$\tilde{K}(\mathbf{y},f_Y)=\lambda(\mathbf{y})f_Y(\mathbf{y}). \tag{C.142}$$

Then, the first-order variation condition is checked as follows.

$$\begin{aligned}
& K'_{f_X}\Big|_{f_X=f_{X^*},f_Y=f_{Y^*}} \\
= \;& f_W(\mathbf{y}-\mathbf{x})\Big(-\mu\log f_{Y^*}(\mathbf{y})+(1-\alpha_1)\log f_{X^*}(\mathbf{x}) \\
& +\mu\left(\mu-1\right)\log f_W(\mathbf{y}-\mathbf{x})+\alpha_0+\sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\gamma_{ij}y_iy_j-\gamma_{ij}x_ix_j \\
& -\gamma_{ij}\left(y-x\right)_i\left(y-x\right)_j+\theta x_ix_j\xi_i\xi_j+\phi_{ij}y_iy_j+\Big)-\lambda(\mathbf{y})+1-\alpha_1\Big) \\
= \;& 0. \qquad \text{(C.143)}
\end{aligned}$$

$$\begin{aligned}
& K'_{f_X}\Big|_{f_Y=f_{X^*},f_Y=f_{Y^*}} \\
= \;& -\frac{\mu\int f_X(\mathbf{x})f_W(\mathbf{y}-\mathbf{x})d\mathbf{x}}{f_Y(\mathbf{y})}+\lambda(\mathbf{y}) \\
= \;& 0. \qquad \text{(C.144)}
\end{aligned}$$

Since the equalities in (C.143) and (C.144) must be satisfied for any $\mathbf{x}$ and $\mathbf{y}$,

$$
\begin{aligned}
\lambda(\mathbf{y}) &= \mu, \\
f_{Y^*}(\mathbf{y}) &= \exp\left\{\frac{1}{\mu}\left(\mathbf{y}^T\left(\mathbf{\Gamma}+\mathbf{\Phi}\right)\mathbf{y}+c_Y\right)\right\} \\
&= (2\pi)^{-\frac{n}{2}}\left|-\frac{\mu}{2}\left(\mathbf{\Gamma}+\mathbf{\Phi}\right)^{-1}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{y}^T\left(-\frac{\mu}{2}\left(\mathbf{\Gamma}+\mathbf{\Phi}\right)^{-1}\right)^{-1}\mathbf{y}\right\} \\
&\quad\times(2\pi)^{\frac{n}{2}}\left|-\frac{\mu}{2}\left(\mathbf{\Gamma}+\mathbf{\Phi}\right)^{-1}\right|^{\frac{1}{2}}\exp\left\{\frac{c_Y}{\mu}\right\} \\
f_W(\mathbf{y}-\mathbf{x}) &= \exp\left\{\frac{1}{\mu\left(\mu-1\right)}\left(\left(\mathbf{y}-\mathbf{x}\right)^T\mathbf{\Gamma}\left(\mathbf{y}-\mathbf{x}\right)-c_W\right)\right\} \\
&= (2\pi)^{-\frac{n}{2}}\left|-\frac{\mu\left(\mu-1\right)}{2}\mathbf{\Gamma}^{-1}\right|^{-\frac{1}{2}} \\
&\quad\times\exp\left\{-\frac{1}{2}\left(\mathbf{y}-\mathbf{x}\right)^T\left(-\frac{\mu\left(\mu-1\right)}{2}\mathbf{\Gamma}^{-1}\right)^{-1}\left(\mathbf{y}-\mathbf{x}\right)\right\} \\
&\quad\times(2\pi)^{\frac{n}{2}}\left|-\frac{\mu\left(\mu-1\right)}{2}\mathbf{\Gamma}^{-1}\right|^{\frac{1}{2}}\exp\left\{-\frac{c_W}{\mu\left(\mu-1\right)}\right\}, \\
f_{X^*}(\mathbf{x}) &= \exp\left\{\frac{1}{1-\alpha_1}\left(\mathbf{x}^T\left(\mathbf{\Gamma}-\theta\mathbf{\Xi}\right)\mathbf{x}-\alpha_0+\mu-1+\alpha_1+c_W+c_Y\right)\right\} \\
&= (2\pi)^{-\frac{n}{2}}\left|-\frac{1-\alpha_1}{2}\left(\mathbf{\Gamma}-\theta\mathbf{\Xi}\right)^{-1}\right|^{-\frac{1}{2}} \\
&\quad\times\exp\left\{-\frac{1}{2}\mathbf{x}^T\left(-\frac{1-\alpha_1}{2}\left(\mathbf{\Gamma}-\theta\mathbf{\Xi}\right)^{-1}\right)^{-1}\mathbf{x}\right\} \\
&\quad\times(2\pi)^{\frac{n}{2}}\left|-\frac{1-\alpha_1}{2}\left(\mathbf{\Gamma}-\theta\mathbf{\Xi}\right)^{-1}\right|^{\frac{1}{2}} \\
&\quad\times\exp\left\{\frac{-\alpha_0+\mu-1+\alpha_1+c_W+c_Y}{1-\alpha_1}\right\}, \qquad \text{(C.145)}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{\Phi} &= \begin{bmatrix} \phi_{11} & \cdots & \phi_{1n} \\ \vdots & \ddots & \vdots \\ \phi_{n1} & \cdots & \phi_{nn} \end{bmatrix}, \quad \mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nn} \end{bmatrix}, \quad \mathbf{\Xi} = \begin{bmatrix} \xi_1\xi_1 & \cdots & \xi_1\xi_n \\ \vdots & \ddots & \vdots \\ \xi_n\xi_1 & \cdots & \xi_n\xi_n \end{bmatrix}, \\
\mathbf{x} &= [x_1, \cdots, x_n]^T, \\
\mathbf{y} &= [y_1, \cdots, y_n]^T, \\
\theta &\geq 0. \qquad \text{(C.146)}
\end{aligned}
$$

Considering the constraints in (C.140), the equations in (C.145) are further processed

as follows.

$$
\begin{aligned}
f_{Y^*}(y) &= (2\pi)^{-\frac{n}{2}} \left| -\frac{\mu}{2} (\mathbf{\Gamma} + \mathbf{\Phi})^{-1} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{y}^T \left(-\frac{\mu}{2} (\mathbf{\Gamma} + \mathbf{\Phi})^{-1} \right)^{-1} \mathbf{y} \right\} \\
&\quad \times (2\pi)^{\frac{n}{2}} \left| -\frac{\mu}{2} (\mathbf{\Gamma} + \mathbf{\Phi})^{-1} \right|^{\frac{1}{2}} \exp\left\{ \frac{c_Y}{\mu} \right\} \\
&= (2\pi)^{-\frac{n}{2}} \left| \mathbf{\Sigma}_{Y^*} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{y}^T \mathbf{\Sigma}_{Y^*}^{-1} \mathbf{y} \right\}, \\
f_W(y - x) &= (2\pi)^{-\frac{n}{2}} \left| -\frac{\mu(\mu - 1)}{2} \mathbf{\Gamma}^{-1} \right|^{-\frac{1}{2}} \\
&\quad \times \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{x})^T \left(-\frac{\mu(\mu - 1)}{2} \mathbf{\Gamma}^{-1} \right)^{-1} (\mathbf{y} - \mathbf{x}) \right\} \\
&\quad \times (2\pi)^{\frac{n}{2}} \left| -\frac{\mu(\mu - 1)}{2} \mathbf{\Gamma}^{-1} \right|^{\frac{1}{2}} \exp\left\{ -\frac{c_W}{\mu(\mu - 1)} \right\} \\
&= (2\pi)^{-\frac{n}{2}} \left| \mathbf{\Sigma}_W \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{x})^T \mathbf{\Sigma}_W^{-1} (\mathbf{y} - \mathbf{x}) \right\}, \\
f_{X^*}(x) &= (2\pi)^{-\frac{n}{2}} \left| -\frac{1 - \alpha_1}{2} (\mathbf{\Gamma} - \theta \mathbf{\Xi})^{-1} \right|^{-\frac{1}{2}} \\
&\quad \times \exp\left\{ -\frac{1}{2} \mathbf{x}^T \left(-\frac{1 - \alpha_1}{2} (\mathbf{\Gamma} - \theta \mathbf{\Xi})^{-1} \right)^{-1} \mathbf{x} \right\} \\
&\quad \times (2\pi)^{\frac{n}{2}} \left| -\frac{1 - \alpha_1}{2} (\mathbf{\Gamma} - \theta \mathbf{\Xi})^{-1} \right|^{\frac{1}{2}} \\
&\quad \times \exp\left\{ \frac{-\alpha_0 + \mu - 1 + \alpha_1 + c_W + c_Y}{1 - \alpha_1} \right\} \\
&= (2\pi)^{-\frac{n}{2}} \left| \mathbf{\Sigma}_{X^*} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathbf{x}^T \mathbf{\Sigma}_{X^*}^{-1} \mathbf{x} \right\}, \qquad \text{(C.147)}
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_0 &= \mu - (1-\alpha_1) + c_W + c_Y + \frac{1-\alpha_1}{2}\log(2\pi)^n |\mathbf{\Sigma}_{X^*}| \\
&= \mu - (1-\alpha_1) + \frac{\mu(\mu-1)}{2}\log(2\pi)^n |\mathbf{\Sigma}_W| \\
&\quad -\frac{\mu}{2}\log(2\pi)^n |\mathbf{\Sigma}_{Y^*}| + \frac{1-\alpha_1}{2}\log(2\pi)^n |\mathbf{\Sigma}_{X^*}|, \\
\mathbf{\Gamma} &= -\frac{\mu(\mu-1)}{2}\mathbf{\Sigma}_W^{-1}, \\
\mathbf{\Phi} &= -\mathbf{\Gamma} - \frac{\mu}{2}\mathbf{\Sigma}_{Y^*}^{-1} \\
&= \frac{\mu(\mu-1)}{2}\mathbf{\Sigma}_W^{-1} - \frac{\mu}{2}(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_W)^{-1}, \\
\mathbf{\Sigma}_{X^*} &= -\frac{1-\alpha_1}{2}(\mathbf{\Gamma} - \theta\mathbf{\Xi})^{-1} \\
&= \frac{1-\alpha_1}{2}\left(\frac{\mu(\mu-1)}{2}\mathbf{\Sigma}_W^{-1} + \theta\mathbf{\Xi}\right)^{-1} && \text{(C.148)} \\
&\succeq \mathbf{0}, && \text{(C.149)} \\
\theta &\geq 0, \\
\alpha_1 &\leq 1-\mu, && \text{(C.150)} \\
c_W &= \frac{\mu(\mu-1)}{2}\log(2\pi)^n |\mathbf{\Sigma}_W|, \\
c_Y &= -\frac{\mu}{2}\log(2\pi)^n |\mathbf{\Sigma}_{Y^*}|, \\
|\mathbf{\Sigma}_{X^*}| &= \left(\frac{1}{2\pi e}\exp\left\{\frac{2}{n}p_X\right\}\right)^n \leq |\mathbf{\Sigma}|. && \text{(C.151)}
\end{aligned}
$$

The inequality in (C.149) is always satisfied since the matrix $\mathbf{\Xi}$ is non-zero positive semi-definite and θ is non-negative. The inequality in (C.151) will be proved later in this proof. The constant p_X must be chosen to satisfy the inequality in (C.151). Then, the Lagrange multipliers always exist, and necessary optimal solutions exist.

Interestingly, similar to the complementary slackness in KKT conditions, when $\theta = 0$ in (C.148), $\mathbf{\Sigma}_{X^*} = (1-\alpha_1)\mu^{-1}(\mu-1)^{-1}\mathbf{\Sigma}_W$, and it requires $(1-\alpha_1)\mu^{-1}(\mu-1)^{-1}\mathbf{\Sigma}_W \preceq \mathbf{\Sigma}$. When θ is non-zero, the equation in (C.148) is positive semi-definite,

and it means $\mathbf{\Sigma}_{X^*} = (1-\alpha_1)\,\mu^{-1}\,(\mu-1)^{-1}\,\mathbf{\Sigma}_{\tilde{W}}$, where $\mathbf{\Sigma}_{\tilde{W}} = \mathbf{\Sigma}_W - \mathbf{\Sigma}_{\hat{W}}$, where $\mathbf{\Sigma}_{\hat{W}}$ and $\mathbf{\Sigma}_{\tilde{W}}$ are positive semi-definite matrices. When $1-\alpha_1 = \mu$, then $\mathbf{\Sigma}_{X^*} = (\mu-1)^{-1}\,\mathbf{\Sigma}_{\tilde{W}}$, which is exactly the same as the one in [32] and [41].

To make the second variation positive, we need the positive-definiteness of the following matrix:

$$\begin{bmatrix} K''_{f_{X^*}f_{X^*}} & K''_{f_{X^*}f_{Y^*}} \\ K''_{f_{Y^*}f_{X^*}} & K''_{f_{Y^*}f_{Y^*}} \end{bmatrix}, \tag{C.152}$$

and it requires the following condition to hold:

$$\begin{aligned} & \begin{bmatrix} h_X & h_Y \end{bmatrix} \begin{bmatrix} K''_{f_{X^*}f_{X^*}} & K''_{f_{X^*}f_{Y^*}} \\ K''_{f_{Y^*}f_{X^*}} & K''_{f_{Y^*}f_{Y^*}} \end{bmatrix} \begin{bmatrix} h_X \\ h_Y \end{bmatrix} \\ = \;& K''_{f_{X^*}f_{X^*}} h_X^2 + K''_{f_{Y^*}f_{Y^*}} h_Y^2 + (K''_{f_{X^*}f_{Y^*}} + K''_{f_{Y^*}f_{X^*}}) h_Y h_X \\ \geq \;& 0, \end{aligned} \tag{C.153}$$

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

$$\begin{aligned} K''_{f_{X^*}f_{X^*}} &= \frac{(1-\alpha_1) f_W(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})}, \\ K''_{f_{X^*}f_{Y^*}} &= -\frac{\mu f_W(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})}, \\ K''_{f_{Y^*}f_{X^*}} &= -\frac{\mu f_W(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})}, \\ K''_{f_{Y^*}f_{Y^*}} &= \frac{\mu f_{X^*}(\mathbf{x}) f_W(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})^2}, \end{aligned} \tag{C.154}$$

the equation in (C.153) requires

$$\frac{(1-\alpha_1)f_W(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})}h_X(\mathbf{x})^2 - 2\frac{\mu f_W(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})}h_X(\mathbf{x})h_Y(\mathbf{y}) + \frac{\mu f_{X^*}(\mathbf{x})f_W(\mathbf{y}-\mathbf{x})}{f_{Y^*}(\mathbf{y})^2}h_Y(\mathbf{y})^2$$
$$\geq \frac{\mu f_W(\mathbf{y}-\mathbf{x})}{f_{X^*}(\mathbf{x})}\left(h_X(\mathbf{x}) - \frac{f_{X^*}(\mathbf{x})}{f_{Y^*}(\mathbf{y})}h_Y(\mathbf{y})\right)^2, \tag{C.155}$$

where $\alpha_1 \geq 1-\mu$.

Therefore, the optimal solutions f_{X^*} and f_{Y^*} minimize the functional problem in (C.139), and the proof is completed. □

C.14 Proof of Theorem 5.17

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 (5.40) is written as follows:

$$\begin{aligned} & \mu h(\mathbf{X}+\mathbf{V}_G) - h(\mathbf{X}+\mathbf{W}_G) \\ \geq\ & \mu h(\mathbf{X}+\mathbf{V}_G) - h(\mathbf{X}+\tilde{\mathbf{W}}_G) - h(\mathbf{W}_G) + h(\tilde{\mathbf{W}}_G) \\ =\ & \mu h(\mathbf{X}+\tilde{\mathbf{W}}_G+\hat{\mathbf{V}}_G) - h(\mathbf{X}+\tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G+\hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G). \end{aligned} \tag{C.156}$$

Since the equation will be minimized over $f_X(\mathbf{x})$, the last two terms in (C.156) 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 (5.40) is equivalently expressed as the following variational problem:

$$
\begin{aligned}
\min_{f_{\hat{X}}, f_Y} \quad & \mu h(\mathbf{Y}) - h(\hat{\mathbf{X}}) - \mu(\mu-1) h(\hat{\mathbf{V}}_G) \\
\text{s. t.} \quad & \int\int f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y}-\mathbf{x}) d\mathbf{x} d\mathbf{y} - 1 = 0, \\
& \int\int f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y}-\mathbf{x}) \mathbf{x}\mathbf{x}^T d\mathbf{x} d\mathbf{y} - \mathbf{\Sigma}_{\hat{X}} \preceq \mathbf{0}, \\
& \int\int f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y}-\mathbf{x}) \mathbf{y}\mathbf{y}^T d\mathbf{x} d\mathbf{y} - \mathbf{\Sigma}_{Y^*} = \mathbf{0}, \\
& \int\int f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y}-\mathbf{x}) \left(\mathbf{y}\mathbf{y}^T - \mathbf{x}\mathbf{x}^T - (\mathbf{y}-\mathbf{x})(\mathbf{y}-\mathbf{x})^T\right) d\mathbf{x} d\mathbf{y} = \mathbf{0}, \\
& -\int\int f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y}-\mathbf{x}) \log f_{\hat{X}}(\mathbf{x}) d\mathbf{x} d\mathbf{y} = p_{\hat{X}}, \qquad \text{(C.157)} \\
& f_Y(\mathbf{y}) = \int f_{\hat{X}}(\mathbf{x}) f_{\hat{V}}(\mathbf{y}-\mathbf{x}) d\mathbf{x},
\end{aligned}
$$

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

The variational problem in (C.157) is exactly the same as the one in (C.139). Therefore, using the same method in the proof of Theorem 5.16, we obtain the following inequality (see the details of the proof in Appendix C.13).

$$
\begin{aligned}
& \mu h(\mathbf{X} + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X} + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G) \\
\geq \; & \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G). \qquad \text{(C.158)}
\end{aligned}
$$

By appropriately choosing $\mathbf{X}^*_G$ and $\tilde{\mathbf{W}}_G$, the right-hand side of the equation in (C.158) is expressed as

$$
\begin{aligned}
& \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G) - h(\tilde{\mathbf{W}}_G + \hat{\mathbf{W}}_G) + h(\tilde{\mathbf{W}}_G) \\
= \; & \mu h(\mathbf{X}^*_G + \tilde{\mathbf{W}}_G + \hat{\mathbf{V}}_G) - h(\mathbf{X}^*_G + \mathbf{W}_G). \qquad \text{(C.159)}
\end{aligned}
$$

The equality in (C.159) is due to the equality condition of data processing inequality in [41]. For the completeness of the proof, we introduce a technique, which is slightly different from the one in [41].

To satisfy the equality in the equation (C.159), the equality condition in the following lemma must be satisfied.

Lemma C.1 (Data Processing Inequality [9]). *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:*

$$I(\mathbf{Y}_1; \mathbf{Y}_3) \leq I(\mathbf{Y}_1; \mathbf{Y}_2). \tag{C.160}$$

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

In Lemma C.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

$$
\begin{aligned}
I(\mathbf{Y}_1;\mathbf{Y}_2|\mathbf{Y}_3) &= h(\mathbf{Y}_1|\mathbf{Y}_3) - h(\mathbf{Y}_1|\mathbf{Y}_2,\mathbf{Y}_3) \\
&= \frac{1}{2}\log(2\pi e)^n \left|\mathbf{\Sigma}_{Y_1|Y_3}\right| - \frac{1}{2}\log(2\pi e)^n \left|\mathbf{\Sigma}_{Y_1|Y_2}\right| \\
&= \frac{1}{2}\log(2\pi e)^n \left|\mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_1}\mathbf{\Sigma}_{Y_3}^{-1}\mathbf{\Sigma}_{Y_1}\right| - \frac{1}{2}\log(2\pi e)^n \left|\mathbf{\Sigma}_{Y_1} - \mathbf{\Sigma}_{Y_1}\mathbf{\Sigma}_{Y_2}^{-1}\mathbf{\Sigma}_{Y_1}\right| \\
&= \frac{1}{2}\log(2\pi e)^n \left|\mathbf{\Sigma}_{X^*} - \mathbf{\Sigma}_{X^*}\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}} + \mathbf{\Sigma}_{\hat{W}}\right)^{-1}\mathbf{\Sigma}_{X^*}\right| \\
&\qquad\qquad - \frac{1}{2}\log(2\pi e)^n \left|\mathbf{\Sigma}_{X^*} - \mathbf{\Sigma}_{X^*}\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)^{-1}\mathbf{\Sigma}_{X^*}\right| \\
&= \frac{1}{2}\log(2\pi e)^n \left|\mathbf{\Sigma}_{X^*}\right| \left|I - \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}} + \mathbf{\Sigma}_{\hat{W}}\right)^{-1}\mathbf{\Sigma}_{X^*}\right| \\
&\qquad\qquad - \frac{1}{2}\log(2\pi e)^n \left|\mathbf{\Sigma}_{X^*}\right| \left|I - \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)^{-1}\mathbf{\Sigma}_{X^*}\right| \\
&= \frac{1}{2}\log(2\pi e)^n \left|I - \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}} + \mathbf{\Sigma}_{\hat{W}}\right)^{-1}\mathbf{\Sigma}_{X^*}\right| \\
&\qquad\qquad - \frac{1}{2}\log(2\pi e)^n \left|I - \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)^{-1}\mathbf{\Sigma}_{X^*}\right| \\
&= \frac{1}{2}\log(2\pi e)^n \left|I - \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{W}\right)^{-1}\mathbf{\Sigma}_{X^*}\right| \\
&\qquad\qquad - \frac{1}{2}\log(2\pi e)^n \left|I - \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)^{-1}\mathbf{\Sigma}_{X^*}\right| \\
&= 0. \qquad\qquad \text{(C.161)}
\end{aligned}
$$

If $\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{W}\right)^{-1}\mathbf{\Sigma}_{X^*} = \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)^{-1}\mathbf{\Sigma}_{X^*}$, the equality in (C.161) is satisfied, the equality condition in Lemma C.1 holds, and therefore, the equality in (C.159) is proved. The validity of $\left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{W}\right)^{-1}\mathbf{\Sigma}_{X^*} = \left(\mathbf{\Sigma}_{X^*} + \mathbf{\Sigma}_{\tilde{W}}\right)^{-1}\mathbf{\Sigma}_{X^*}$ is proved by Lemma 8 in [41].

Therefore, $I(\mathbf{Y}_1;\mathbf{Y}_2|\mathbf{Y}_3)=0$, and, from the equations in (C.156), (C.158), and

(C.159), 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.

□