Thèse n° 8094

# Feedback and Common Information: Bounds and Capacity for Gaussian Networks

Présentée le 21 mai 2021

Faculté informatique et communications
Laboratoire d’information dans les systèmes en réseaux
Programme doctoral en informatique et communications

pour l’obtention du grade de Docteur ès Sciences

par

**Erixhen SULA**

Acceptée sur proposition du jury

Prof. E. Telatar, président du jury
Prof. M. C. Gastpar, directeur de thèse
Prof. G. Kramer, rapporteur
Prof. M. Wigger, rapporteuse
Prof. B. Rimoldi, rapporteur

École
polytechnique
fédérale
de Lausanne

2021

Dedicated to my family.

*Love begins by taking care of the closest ones,*
*the ones at home.*
–Saint Mother Teresa

# Acknowledgements

On the journey of writing the thesis, I was wondering what could be the toughest chapter to write. It turns out that the toughest chapter is not reflecting any of the results, but is the Acknowledgements.

Firstly, my deepest gratitude goes to my supervisor, Prof. Michael Gastpar, who was actively involved in my graduation process, as a guiding expert he is. He taught me about various aspect of a research problem and to *think outside the box* whenever I encountered difficulty in progressing. One of his famous claims is that research is not only about the derivations of the toughest open problem, but as well as the way you deliver it to the audience in a simplistic way and I learned from his superb presentation skills. I am indebted to Prof. Michael Gastpar for helping with the process of the summer internship at Nokia Bell Labs, postdoc, but not only.

I had the opportunity to work with Dr. Junho Cho, during the summer internship at Nokia Bell Labs. I want to express my gratitude for his hospitality and friendship in my early days in the United States. His persistent advise to show up and present your own work has affected me ever since then.

It was honour to have on my thesis committee Prof. Gerhard Kramer, Prof. Bixio Rimoldi, Prof. Emre Telatar and Prof. Michèle Wigger. I really appreciate their effort to read the thesis and the comments on both the presentation and the thesis. In particular, Prof. Michèle Wigger had given invaluable comments about a conference paper on my early doctoral studies.

During the doctoral journey a special place holds our secretary, France Faille. Eventually, she is always present and ready to help. For me she was the driving force to learn French and encouraged me to attend French classes. In particular, I admire the way she welcomes students that join the lab. And thanks to Damir, my laptop was running properly.

An important acknowledgement goes to LINX group. Many thanks to my officemates Giel and Amedeo. It was a pleasure to share the office with you. My initial contact with the group was when I did a semester project with Jingge, who was always ready to answer whenever I knocked on his door. Thanks to Adriano, Jingge and Sung Hoon for the helpful conversations and the substantial impact they had on my master thesis. Many notable memories are shared during the conferences in Aachen, Colorado, Madrid and Sweden. In particular, most of them are attended with Su with many funny stories behind. I would like to thank also the fellow and current IPG labmates, Saeid, Ibrahim, Yanina, Eric, Jean, Clément, Elie, Mohamad, Nicolae, Yunus, Kirill, Pierre, Reka, Daria and all the other members.

A large part of the thesis is dedicated to my friends and cousins. I will start with my cousin Krenar. I want to thank him especially for providing me a place to stay initially and for the support I needed, I am forever indebted to him. I want to thank my good old friend Redi, for the hospitality, the time we spend in Lausanne and Zürich and flights we took together, that was a joy. During the time in the United States I want to thank in particular Erion, for the all the weekly adventures we had in Manhattan. Here in Lausanne I met great people, but with a few of them we were very close. On my arrival in Switzerland I used to spend most of the time with Veton, that culminates with celebrating birthdays on the same day. I want to thank Iurii and Vova for the barbecues and all the weekends we spend together. Special thanks to Valon, mostly during my final part of graduation, for going into each-others home without any sort of permission, for spending days and nights all over Lausanne. I would like also to thank my EPFL friends Araz and Besnik for all the funny stories. I am really thankful to Océane, for her patience to teach me french for over a year. Lastly, I want to mention my two cousins Armando and Eriglen that recently past away. We grew up together since primary school and their loss left a huge mark on me and also in my graduation journey.

Tani vjen pjesa më e vështirë, ku të humbin edhe fjalët, pra ëmbëlsira për në fund. Falenderimi më i përzemërt shkon për familjen time të shtrenjtë, mamin Filloreta, babin Petrit dhe motrën Arbora. Ata kanë qënë, janë dhe do te ngelen mbështetja dhe inspirimi im më i madh. Kjo tezë i dedikohet familjes dhe gjyshes time Ervehe që ka ndërruar jetë.

# Abstract

Network information theory studies the communication of information in a network and considers its fundamental limits. Motivating from the extensive presence of the networks in the daily life, the thesis studies the fundamental limits of particular networks including channel coding such as Gaussian multiple access channel with feedback and source coding such as lossy Gaussian Gray-Wyner network.

On one part, we establish the sum-Capacity of the Gaussian multiple-access channel with feedback. The converse bounds that are derived from the dependence-balance argument of Hekstra and Willems meet the achievable scheme introduced by Kramer. Even though the problem is not convex, the factorization of lower convex envelope method that is introduced by Geng and Nair, combined with a Gaussian property are invoked to compute the sum-Capacity. Additionally, we characterize the rate region of lossy Gaussian Gray-Wyner network for symmetric distortion. The problem is not convex, thus the method of factorization of lower convex envelope is used to show the Gaussian optimality of the auxiliaries. Both of the networks, are a long-standing open problem.

On the other part, we consider the common information that is introduced by Wyner and the natural relaxation of Wyner's common information. Wyner's common information is a measure that quantifies and assesses the commonality between two random variables. The operational significance of the newly introduced quantity is in Gray-Wyner network. Thus, computing the relaxed Wyner's common information is directly connected with computing the rate region in Gray-Wyner network. We derive a lower bound to Wyner's common information for any given source. The bound meets the exact Wyner's common information for sources that are expressed as sum of a common random variable and Gaussian noises. Moreover, we derive an upper bound on an extended variant of information bottleneck.

Finally, we use Wyner's common information and its relaxation as a tool to extract common information between datasets. Thus, we introduce a novel procedure to construct features from data, referred to as Common Information Components Analysis (CICA). We establish that in the case of Gaussian statistics, CICA precisely reduces to Canonical Correlation Analysis (CCA), where the relaxing parameter determines the number of CCA components that are extracted. In this sense, we establish a novel rigorous connection between information measures and CCA, and CICA is a strict generalization of the latter. Moreover, we show that CICA has several desirable features, including a natural extension to beyond just two data sets.

**Keywords:** Multiple access channel, feedback, Gray-Wyner network, Wyner's common information, common information component analysis, canonical correla-

tion analysis, Gaussian network, noise.

# Résumé

La théorie de l'information de réseau étudie les réseaux de communication de l'information et leurs limites fondamentales. Motivée par l'omniprésence des réseaux dans la vie quotidienne, la thèse étudie les limites fondamentales de réseaux particuliers, y compris le codage de canal tel que le canal Gaussien à accès multiple avec rétroaction, et le codage source tel que le réseau Gaussien Gray-Wyner avec perte.

D'une part, nous établissons la somme-capacité du canal Gaussien à accès multiple avec rétroaction. Les bornes inverses, qui sont dérivées de l'argument de l'équilibre de dépendance de Hekstra et Willems, répondent au schéma réalisable introduit par Kramer. La factorisation de la méthode d'enveloppe convexe inférieure introduite par Geng et Nair, combinée à une propriété Gaussienne, est utilisée pour calculer la somme-capacité lorsque le problème n'est pas convexe. De plus, nous caractérisons la région de proportion du réseau Gaussien de Gray-Wyner avec perte pour la distorsion symétrique. Le problème n'étant pas convexe, la méthode de factorisation de l'enveloppe convexe inférieure est utilisée pour montrer l'optimalité Gaussienne des auxiliaires. Les deux réseaux sont des problèmes ouverts de longue date.

D'autre part, nous considérons les informations communes introduites par Wyner et la relaxation naturelle des informations communes de Wyner. Les informations communes de Wyner sont une métrique qui quantifie et évalue la similitude entre deux variables aléatoires. L'importance opérationnelle de la quantité nouvellement introduite est dans le réseau de Gray-Wyner. Ainsi, le calcul des informations communes de Wyner relaxées est directement lié au calcul de la région de proportion dans le réseau de Gray-Wyner. Nous en déduisons une limite inférieure aux informations communes de Wyner pour une source donnée. La borne est exactement égale aux informations communes de Wyner pour des sources exprimées comme étant la somme d'une variable aléatoire commune et de bruits Gaussiens. Nous en déduisons également une borne supérieure sur une variante étendue de l'Information Bottleneck (IB).

Enfin, nous utilisons les informations communes de Wyner et leur assouplissement comme outil pour extraire des informations communes à plusieurs jeux de données. Ainsi, nous introduisons une nouvelle procédure pour construire des caractéristiques à partir de données, appelée analyse des composants d'information communs (CICA). Nous établissons que dans le cas de statistiques Gaussiens, l'ICCA se réduit précisément à l'analyse de corrélation canonique (CCA), où le paramètre relaxant détermine le nombre de composants CCA qui sont extraits. En ce sens, nous établissons un nouveau lien rigoureux entre les mesures d'information et le CCA, l'ICCA étant une généralisation stricte de ce dernier. De plus, nous montrons que l'ICCA a plusieurs caractéristiques souhaitables, notamment une extension naturelle au-delà de deux ensembles de données.

**Mots-clés:** Canal à accès multiple, retour d'information, réseau Gray-Wyner, informations communes de Wyner, analyse des composants d'information commune, analyse de corrélation canonique, réseau Gaussien, bruit.

# Contents

# 1 Introduction

Shannon [1] initially established the mathematical theory behind communication and introduced the notion of information theory. Point-to-point communication is studied and its fundamental limits are solved in the presence of noise [1]. Point-to-point communication is composed of a single sender and a single receiver where the information is sent from the sender to the receiver via the channel. Shannon's point-to-point communication is mainly composed of the following two fundamental problems.

- Channel Coding:
  Suppose that $X$ is input, $Y$ is output of the channel and $p(y|x)$ describes the channel. The probability of error is the probability that the encoded message at the sender side is different from the decoded message at the receiver side. To communicate the information reliably we wish to keep the probability of error as small as possible. The *channel capacity* $C$ is,

  $$C = \max_{p(x)} I(X;Y) \tag{1.1}$$

  that is the maximum communication rate in bits such that the probability of error is as small as possible. In order to find the capacity for a given channel with conditional probability $p(y|x)$, we need to optimize over the input probability $p(x)$.

- Lossy Source Coding:
  Suppose that source $X$ is compressed and sent through a noiseless channel and the receiver reconstructs it with some distortion $d(x, \hat{x})$, that is the distortion measure between the sent symbol $x$ and received symbol $\hat{x}$. The *rate-distortion function* $R(D)$ is

  $$R(D) = \min_{p(\hat{x}|x):\mathbb{E}[d(X,\hat{X})]\leq D} I(X;\hat{X}). \tag{1.2}$$

The scope of information theory is not limited to point-to-point communication. Network information theory studies the limits of information flow in networks consisting of multiple senders and multiple receivers. Interference, cooperation and/or feedback may be present in the network. The importance of networks lie in the presence in our daily life for instance, the cellphones that communicate form a telecommunication network, a group of computers that communicate form computer network, a collection of devices that communicate form a device network. Network information theory studies a simplistic version of aforementioned examples, that are far more complicated. In the thesis, we mainly focus on the following two networks that involve channel coding and lossy source coding as follows.

- Multiple Access Channel with Feedback

  Multiple access channel refers to multiple senders communicating information through a common channel to a single receiver. The receiver observes the sum of the messages of each user (and added independent noise in the additive channels). Moreover, the (perfect and causal) feedback is used during the transmission, which allows the senders to cooperate in communicating their messages to achieve higher rates than the absence of feedback. The presence of feedback in point-to-point communication does not help the capacity, however the feedback increases the capacity in the multiple access channel as it allows the senders to cooperate. The role of the receiver is to decode the messages send by each respective user.

- Gray-Wyner Network

  The network is composed of one sender and two receivers, that communicate messages through a common link and two private links. The common channel is provided to both receivers and each private channel is provided to the respective decoder. The sender communicates a pair of messages through the common and two private channels, where each receiver is interested in the respective message.

A subclass of the networks in information theory are the Gaussian networks, where in the channel coding, the channel is modelled as Gaussian and in the source coding the source is modelled as Gaussian. We focus on these networks and compute the capacity and rate-distortion function for the aforementioned networks.

## 1.1 Contributions

- **Gaussian Multiple Access Channel with Feedback**: In Chapter 3 the sum-Capacity is computed under a symmetric block power constraint by showing the optimality of Gaussian auxiliaries. We prove that the new outer bounds based on Hekstra-Willems dependence-balance argument meet the Fourier-Modulated Estimate Correction scheme. Our proof unifies all the previous partial proofs. The difficulty relies on the fact that the problem is not convex, thus the factorization of lower convex envelope method is used to compute the sum-Capacity and prove the Gaussian optimality.

- **(Relaxed) Wyner's Common Information**: In Chapter 4 we revive a natural relaxation of Wyner's common information, the so-called relaxed Wyner's common information that was initially proposed by Wyner, but not studied. More specifically, in the original definition of Wyner's common information, the conditional independence constraint is replaced by an upper bound on the conditional mutual information. We provide an alternative proof of (standard) Wyner's common information for Gaussian vector sources. Moreover, we solve the natural relaxation of Wyner's common information for Gaussian sources. The solution to natural relaxation of Wyner's common information for Gaussian sources is interpreted as a reverse water-filling procedure. Later on, the relaxed Wyner's common information is used in the common information component analysis algorithm to extract common information.

- **Gaussian Gray-Wyner Network**: In Chapter 5 we compute the rate region of Gaussian lossy Gray-Wyner network under symmetric mean-squared error distortion. We prove that it is optimal to select the auxiliary random variable to be jointly Gaussian with the source random variables.

- **Lower bound on (relaxed) Wyner's Common Information**: In Chapter 6 a lower bound to (relaxed) Wyner's common information is derived where the proof is fundamentally different from the method that were used to solve the Wyner's common information. We demonstrate that for a number of distributions the new lower bound is dominant to the existing bounds. When the distribution is written as the sum of a single arbitrary random variable and jointly Gaussians, then the new lower bound is tight.

- **Common Information Component Analysis**: In Chapter 7 we devise a novel algorithms, the so-called common information component analysis (CICA). The algorithm is an alternative way to extract common information from data. Extracting common information is not popular in feature extraction since two or more data are involved. The majority of the approaches deal with extraction of the essential information on a single data. The algorithms that perform common information extraction include canonical correlation analysis (CCA). The proposed algorithms is composed of two main steps. In the first one, we solve the relaxed Wyner's common information and the second one we project the common information from the earlier step back onto the original data. When the original data is generated from a jointly Gaussian source, our algorithms precisely extract the CCA components and a direct connection with CCA is established. A parameter that is exclusive to this algorithm allows to determines the number of the CCA components that are extracted. In the examples provided, the CICA outperforms CCA. Most importantly, CICA is dominant to other methods when extracting common information between three or more data, which is supported by example.

- **Upper Bound on Double Information Bottleneck**: In Chapter 8 we compute an upper bound on an extended variant of information bottleneck.

## 1.2 Notation

We use the following notation. Random variables are denoted by uppercase letters such as $X$ and their realizations by lowercase letters such as $x$. The alphabets in which they take their values will be denoted by calligraphic letters such as $\mathcal{X}$. The probability distribution function of random variable $X$ will be denoted by $p_X$

or $p(x)$ depending on the context. Let $\mathcal{P}$ be the set of all probability distribution, discrete or continuous depending on the context. Random column vectors are denoted by boldface uppercase letters and their realizations by boldface lowercase letters. Depending on the context we will denote the random column vector also as $X^n := (X_1, X_2, \dots, X_n)$. The $i$-th entry of the column vector $\boldsymbol{X}$ is denoted by $\boldsymbol{X}_i$ or $[\boldsymbol{X}]_i$. Calligraphic letters denote sets, e.g., $\mathcal{A}, \mathcal{B}, \mathcal{C}$. A subset $\mathcal{S}$ of the entries of the column vector $\boldsymbol{X}$ is denoted by $\boldsymbol{X}_{\mathcal{S}}$ or $[\boldsymbol{X}]_{\mathcal{S}}$. We denote matrices with uppercase letters, e.g., $A, B, C$. The $(i, j)$ element of matrix $A$ is denoted by $A_{ij}$ or $[A]_{ij}$. Let us denote with $I_n$ the identity matrix of dimension $n \times n$ and $0_n$ the zero matrix of dimension $n \times n$. For the cross-covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$, we use the shorthand notation $K_{\boldsymbol{XY}}$, and for the covariance matrix of a random vector $\boldsymbol{X}$ we use the shorthand notation $K_{\boldsymbol{X}} := K_{\boldsymbol{XX}}$. In slight abuse of notation, we will let $K_{(X,Y)}$ denote the covariance matrix of the stacked vector $(X, Y)^T$. Let $K^H$ be the Hermitian transpose matrix. We denote the Kullback-Leibler divergence with $D(.||.)$. The diagonal matrix is denoted by diag(.). We denote $\log^+(x) = \max(\log x, 0)$. The expression $X \sim \mathcal{N}(m, \sigma^2)$ denotes a Gaussian random variable with mean $m$ and variance $\sigma^2$. We denote the convergence in distribution (weak convergence) by $\overset{w}{\Rightarrow}$. We denote by $\breve{f}(x)$ the lower convex envelope of $f(x)$ with respect to $x$ and for random variables let $\breve{f}(X)$ (or $\breve{f}(p_X)$) be the lower convex envelope of $f(X)$ with respect to $p_X$. We denote by $h_b(x) := -x \log x - (1-x) \log 1 - x$ the binary entropy for $0 \leq x \leq 1$.

# Preliminaries 2

Network information theory is mainly composed of channel (coding) and source (coding) networks. We investigate simple networks when the channel is modelled as additive Gaussian noise and the source is modelled as Gaussian. In the additive Gaussian channel network, for a given probability density function of the channel, we seek the optimal probability density function of the input to attain the capacity of the channel. In the Gaussian source network, for a given probability density function of the source, we seek the optimal probability density function of the channel to attain the rate-distortion function.

## 2.1 Kac-Bernstein theorem

The following theorem is a property of the Gaussian distribution.

**Theorem 1** ([2, 3])**.** *If $X_1$ and $X_2$ are independent random variables, and if $\frac{1}{\sqrt{2}}(X_1 + X_2)$ and $\frac{1}{\sqrt{2}}(X_1 - X_2)$, then $X_1$, $X_2$ are normally distributed.*

Let us consider an application of Theorem 1 in a source coding problem.

## 2.2 Entropy maximization via Kac-Bernstein theorem

**Theorem 2.** *The unique maximizer of*

$$V(\sigma_X^2) = \max_{X:\mathbb{E}[X^2]\leq\sigma_X^2} h(X), \tag{2.1}$$

*is the Gaussian distribution with mean zero and variance $\sigma_X^2$.*

The proof of Theorem 2 is present in [4, Theorem 8.6.5]. Here we provide an alternative proof by using the Kac-Bernstein theorem.

*Proof.* Let $X$ be an optimizing random variable of $V(\sigma_X^2)$. Let $X_1$ and $X_2$ be two independent copies of the optimizing random variable $X$. Then,

$$\begin{aligned}
2V(\sigma_X^2) &= h(X_1) + h(X_2) && (2.2)\\
&= h(X_1, X_2) && (2.3)\\
&= h\left(\frac{X_1+X_2}{\sqrt{2}}, \frac{X_1-X_2}{\sqrt{2}}\right) && (2.4)\\
&= h\left(\frac{X_1+X_2}{\sqrt{2}}\right) + h\left(\frac{X_1-X_2}{\sqrt{2}}\right) - I\left(\frac{X_1+X_2}{\sqrt{2}}; \frac{X_1-X_2}{\sqrt{2}}\right) && (2.5)\\
&\leq 2V(\sigma_X^2) - I\left(\frac{X_1+X_2}{\sqrt{2}}; \frac{X_1-X_2}{\sqrt{2}}\right), && (2.6)
\end{aligned}$$

where (2.2) holds because $X_1$ and $X_2$ are optimizing random variables that attain $V(\sigma_X^2)$; (2.3) holds from the independence of $X_1$ and $X_2$; (2.4) holds because entropy is preserved for unitary transformations; (2.5) holds by applying the chain rule on the entropy and (2.6) holds from the definition of $V(\sigma_X^2)$. In order for (2.2)-(2.6) to hold, we need that $\frac{X_1+X_2}{\sqrt{2}}$ is independent of $\frac{X_1-X_2}{\sqrt{2}}$.

To sum up, by assumption $X_1$ and $X_2$ are independent and by combining (2.2)-(2.6) we proved that $\frac{X_1+X_2}{\sqrt{2}}$ is independent of $\frac{X_1-X_2}{\sqrt{2}}$. Thus, by Theorem 1, $X_1$ and $X_2$ must be Gaussian. □

## 2.3 Gaussian rate distortion via Kac-Bernstein theorem

Let us consider the rate distortion function. A *distortion function* is a mapping $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathcal{R}^+$ from the set of source alphabet-reproduction alphabet pairs into the set of nonnegative real numbers. The distortion $d(x, \hat{x})$ is a measure of the cost of representing the symbol $x$ by the symbol $\hat{x}$. The *square-error distortion*, $d(x, \hat{x}) = (x - \hat{x})^2$ is the most popular distortion measure used for continuous alphabets.

Let $X$ be a random variable and its reconstruction be $\hat{X}$. The problem is to find the optimal $\hat{X}$ that minimizes the distortion function, where probability density function $p(x)$ is given.

By the rate distortion Theorem [4, Theorem 10.2.1] extended to continuous alphabets with squared-error distortion, the optimization problem is defined as

$$V(D) := \inf_{p(\hat{x}|x):\mathbb{E}[(X-\hat{X})^2]\leq D} I(X; \hat{X}) \tag{2.7}$$

where $V(D)$ is the minimum achievable rate at distortion $D$. Let $X \sim \mathcal{N}(0, \sigma_x^2)$.

Let $(X_1, \hat{X}_1)$ and $(X_2, \hat{X}_2)$ be two identical and independent copies of $(X, \hat{X})$ and define

$$(X_{\theta_1}, \hat{X}_{\theta_1}) := \frac{1}{\sqrt{2}}(X_1 + X_2, \hat{X}_1 + \hat{X}_2), \quad (X_{\theta_2}, \hat{X}_{\theta_2}) := \frac{1}{\sqrt{2}}(X_1 - X_2, \hat{X}_1 - \hat{X}_2). \tag{2.8}$$

**Lemma 1.** *The statements listed below are true.*

1. *The following equality holds*

$$I(X_1, X_2; \hat{X}_1, \hat{X}_2) = I(X_{\theta_1}, X_{\theta_2}; \hat{X}_{\theta_1}, \hat{X}_{\theta_2}). \tag{2.9}$$

2. *The following inequality holds*

$$I(X_{\theta_1}, X_{\theta_2}; \hat{X}_{\theta_1}, \hat{X}_{\theta_2}) \geq I(X_{\theta_1}; \hat{X}_{\theta_1}) + I(X_{\theta_2}; \hat{X}_{\theta_2}), \tag{2.10}$$

*when $X_{\theta_1}$ and $X_{\theta_2}$ are independent.*

3. *The equality in* (2.10) *holds if and only if*

$$p(x_{\theta_1}, x_{\theta_2} | \hat{x}_{\theta_1}, \hat{x}_{\theta_2}) = p(x_{\theta_1} | \hat{x}_{\theta_1}) p(x_{\theta_2} | \hat{x}_{\theta_2}). \tag{2.11}$$

*Proof.* Item 1 follows from the fact that mutual information is invariant to linear transformation, i.e. $I(AX; B\hat{X}) = I(X; \hat{X})$ for linear transformation $A$ and $B$.

Item 2 is a consequence of

$$\begin{aligned} I(X_{\theta_1}, X_{\theta_2}; \hat{X}_{\theta_1}, \hat{X}_{\theta_2}) &= I(X_{\theta_1}; \hat{X}_{\theta_1}, \hat{X}_{\theta_2}) + I(X_{\theta_2}; \hat{X}_{\theta_1}, \hat{X}_{\theta_2} | X_{\theta_1}) &(2.12)\\ &= I(X_{\theta_1}; \hat{X}_{\theta_1}) + h(X_{\theta_2} | X_{\theta_1}) + I(X_{\theta_1}; \hat{X}_{\theta_2} | \hat{X}_{\theta_1}) \\ &\quad - h(X_{\theta_2} | X_{\theta_1}, \hat{X}_{\theta_1}, \hat{X}_{\theta_2}) &(2.13)\\ &\geq I(X_{\theta_1}; \hat{X}_{\theta_1}) + h(X_{\theta_2}) - h(X_{\theta_2} | \hat{X}_{\theta_2}) &(2.14)\\ &= I(X_{\theta_1}; \hat{X}_{\theta_1}) + I(X_{\theta_2}; \hat{X}_{\theta_2}) &(2.15) \end{aligned}$$

where (2.12) and (2.13) are application of the chain rule; (2.14) follows from conditioning reduces entropy, that is $h(X_{\theta_2} | X_{\theta_1}, \hat{X}_{\theta_1}, \hat{X}_{\theta_2}) \leq h(X_{\theta_2} | \hat{X}_{\theta_2})$, conditional mutual information is non-negative $I(X_{\theta_1}; \hat{X}_{\theta_2} | \hat{X}_{\theta_1}) \geq 0$ and $h(X_{\theta_2} | X_{\theta_1}) = h(X_{\theta_2})$ by independence of $X_{\theta_1}$ and $X_{\theta_2}$.

Item 3 follows from the equality conditions in (2.14), that are

$$\begin{aligned} I(X_{\theta_1}; \hat{X}_{\theta_2} | \hat{X}_{\theta_1}) &= 0, &(2.16)\\ I(X_{\theta_2}; X_{\theta_1}, \hat{X}_{\theta_1} | \hat{X}_{\theta_2}) &= 0. &(2.17) \end{aligned}$$

By adding Equation (2.16) and (2.17) we have that $I(X_{\theta_2}, \hat{X}_{\theta_2}; X_{\theta_1}, \hat{X}_{\theta_1}) = I(\hat{X}_{\theta_2}; \hat{X}_{\theta_1})$, that is interpreted as $p(x_{\theta_1}, x_{\theta_2} | \hat{x}_{\theta_1}, \hat{x}_{\theta_2}) = p(x_{\theta_1} | \hat{x}_{\theta_1}) p(x_{\theta_2} | \hat{x}_{\theta_2})$. □

The term $I(X; \hat{X})$ is a convex function of $p(\hat{x}|x)$ for a given distribution $p(x)$ and infimum always exits. The mutual information objective is not affected by the mean of the random variables, however the constraint is affected. In particular, for a zero mean $X$ the optimal $\hat{X}$ is mean zero as the term $\mathbb{E}[(X - \hat{X})^2]$ only decreases i.e. if $\hat{X} = \hat{X}_{\text{zm}} + m$, then $\mathbb{E}[(X - \hat{X}_{\text{zm}})^2] \leq D - m^2$. Now we present the main part.

**Lemma 2.** *Let $p^*(\hat{x}|x)$ attain $V(D)$ and let $(\hat{X}_1, \hat{X}_2) | ((X_1, X_2) = (x_1, x_2)) \sim p^*(\hat{x}_1 | x_1) p^*(\hat{x}_2 | x_2)$, then we have that $X_{\theta_1}$ and $X_{\theta_2}$ are conditionally independent given $(\hat{X}_1, \hat{X}_2)$ and attain $V(D)$.*

*Proof.* We have,

$$\begin{aligned} 2V(D) &= I(X_1; \hat{X}_1) + I(X_2; \hat{X}_2) &(2.18)\\ &= I(X_1, X_2; \hat{X}_1, \hat{X}_2) &(2.19)\\ &= I(X_{\theta_1}, X_{\theta_2}; \hat{X}_{\theta_1}, \hat{X}_{\theta_2}) &(2.20)\\ &\geq I(X_{\theta_1}; \hat{X}_{\theta_1}) + I(X_{\theta_2}; \hat{X}_{\theta_2}) &(2.21)\\ &\geq 2V(D), &(2.22) \end{aligned}$$

where (2.18) follows from the optimality of $p^*(\hat{x}_1|x_1)$ and $p^*(\hat{x}_2|x_2)$ satisfying $\mathbb{E}[(X_1-\hat{X}_1)^2] \leq D$ and $\mathbb{E}[(X_2-\hat{X}_2)^2] \leq D$; (2.19) follows from the assumption that $(\hat{X}_1, X_1)$ is independent of $(\hat{X}_2, X_2)$; (2.20) follows from Lemma 1, Item 1; (2.21) follows from Lemma 1, Item 2; (2.22) follows from definition of $V(D)$, and the constraint

$$\begin{aligned}\mathbb{E}[(X_{\theta_1} - \hat{X}_{\theta_1})^2] &= \frac{1}{2}\mathbb{E}[(X_1 - \hat{X}_1 + X_2 - \hat{X}_2)^2] && (2.23)\\ &= \frac{1}{2}\mathbb{E}[(X_1 - \hat{X}_1)^2 + (X_2 - \hat{X}_2)^2 + 2(X_1 - \hat{X}_1)(X_2 - \hat{X}_2)] && (2.24)\\ &\leq D + \mathbb{E}[(X_1 - \hat{X}_1)]\mathbb{E}[(X_2 - \hat{X}_2)] = D, && (2.25)\end{aligned}$$

where the third term in (2.24) is zero from the independence of $(\hat{X}_1, X_1)$ and $(\hat{X}_2, X_2)$.

The inequality starts with $2V(D)$ and ends with $2V(D)$ thus, we have equality in Equation (2.21) and by Lemma 1, Item 3, we have $p(x_{\theta_1}, x_{\theta_2}|\hat{x}_{\theta_1}, \hat{x}_{\theta_2}) = p(x_{\theta_1}|\hat{x}_{\theta_1})p(x_{\theta_2}|\hat{x}_{\theta_2})$, which implies that $p(x_{\theta_1}, x_{\theta_2}|\hat{x}_1, \hat{x}_2) = p(x_{\theta_1}|\hat{x}_1, \hat{x}_2)p(x_{\theta_2}|\hat{x}_1, \hat{x}_2)$, or $X_{\theta_1}$ and $X_{\theta_2}$ are conditionally independent given $(\hat{X}_1, \hat{X}_2)$. □

Let us denote $X|(\hat{X}_1 = \hat{x}_1)$ with $X_{\hat{x}_1}$. By assumption we have that $p(x_1, x_2|\hat{x}_1, \hat{x}_2) = p(x_1|\hat{x}_1)p(x_2|\hat{x}_2)$ or $X_{\hat{x}_1}$ is independent of $X_{\hat{x}_2}$ for any instances $(\hat{x}_1, \hat{x}_2)$. Let us define

$$X_{\theta_1}|((\hat{X}_1, \hat{X}_2) = (\hat{x}_1, \hat{x}_2)) := \frac{1}{\sqrt{2}}(X_{\hat{x}_1} + X_{\hat{x}_2}), \quad (2.26)$$

$$X_{\theta_2}|((\hat{X}_1, \hat{X}_2) = (\hat{x}_1, \hat{x}_2)) := \frac{1}{\sqrt{2}}(X_{\hat{x}_1} - X_{\hat{x}_2}). \quad (2.27)$$

Also, we showed that $p(x_{\theta_1}, x_{\theta_2}|\hat{x}_1, \hat{x}_2) = p(x_{\theta_1}|\hat{x}_1, \hat{x}_2)p(x_{\theta_2}|\hat{x}_1, \hat{x}_2)$ or $\frac{1}{\sqrt{2}}(X_{\hat{x}_1}+X_{\hat{x}_2})$ is independent of $\frac{1}{\sqrt{2}}(X_{\hat{x}_1} - X_{\hat{x}_2})$ for any instances $(\hat{x}_1, \hat{x}_2)$. Then, by Theorem 1, $X_{\hat{x}}$ is Gaussian for any instance $\hat{x}$. Since $X_{\hat{x}}$ are Gaussians for all instances $\hat{x}$,thus

$$\begin{aligned}h(X|\hat{X}) &= \mathbb{E}\left[\frac{1}{2}\log(2\pi e)\operatorname{Var}(X|\hat{X})\right] && (2.28)\\ &\leq \frac{1}{2}\log(2\pi e)\mathbb{E}[\operatorname{Var}(X|\hat{X})], && (2.29)\end{aligned}$$

where the inequality follows from the concavity of log function. Then, a valid lower bound is

$$\begin{aligned}I(X;\hat{X}) &= h(X) - h(X|\hat{X}) && (2.30)\\ &\geq \frac{1}{2}\log\frac{\sigma_x^2}{\mathbb{E}[\operatorname{Var}(X|\hat{X})]} && (2.31)\\ &\geq \frac{1}{2}\log\frac{\sigma_x^2}{D}, && (2.32)\end{aligned}$$

where the last inequality follows by using the law of total variance

$$\begin{aligned}\mathbb{E}[\operatorname{Var}(X|\hat{X})] &= \mathbb{E}[\operatorname{Var}(X - \hat{X}|\hat{X})] && (2.33)\\ &\leq \operatorname{Var}(X - \hat{X}) && (2.34)\\ &= \mathbb{E}[(X - \hat{X})^2] && (2.35)\\ &\leq D. && (2.36)\end{aligned}$$

The same result is obtained by a simpler proof in [4, Theorem 10.3.2], however the aim is to apply Theorem 1 in an information theoretical problem.

Let us analyse the application of Kac-Bernstein theorem. In the rate distortion problem we create two independent copies of $(X, \hat{X})$ that are, $(X_1, \hat{X}_1)$ and $(X_2, \hat{X}_2)$. The independence of $(X_1, \hat{X}_1)$ and $(X_2, \hat{X}_2)$ implies that $X_{\hat{x}_1}$ ($X_{\hat{x}_1}$ denotes $X|(\hat{X}_1 = \hat{x}_1)$) is independent of $X_{\hat{x}_1}$ for any instance $(\hat{x}_1, \hat{x}_2)$. By using Lemma 1 and 2, we show that $\frac{1}{\sqrt{2}}(X_{\hat{x}_1}+X_{\hat{x}_2})$ is independent of $\frac{1}{\sqrt{2}}(X_{\hat{x}_1}-X_{\hat{x}_2})$ for any instances $(\hat{x}_1, \hat{x}_2)$. Then, by Theorem 1, $X_{\hat{x}}$ is Gaussian for any instance $\hat{x}$.

## 2.4 Lower Convex Envelope

**Definition 1.** *The lower convex envelope $\breve{f}$ of the function $f : K \to \mathbb{R}$, where $K$ is a convex set is defined as pointwise supremum of all convex functions that lie under that function and is uniquely determined.*

Let $f : K \to \mathbb{R}$ be a lower semicontinuous function. The lower convex envelope $\breve{f}$ has the following properties

1. $\breve{f}(x)$ is a convex function of $x$ for $x \in K$,
2. $\breve{f}(x) \leq f(x)$ for all $x \in K$,
3. if $g$ is any other convex function such that $g(x) \leq f(x)$ for all $x \in K$, then $g(x) \leq \breve{f}(x)$ for all $x \in K$,
4. $\inf_{x\in K} f(x) = \inf_{x\in K} \breve{f}(x)$.

# Sum-Rate Capacity for Gaussian Multiple Access Channels with Feedback 3

## 3.1 Introduction

The feedback capacity of the *two-user* Gaussian multiple-access channel (GMAC)[1] was established by Ozarow [5]. The coding theorem was based on extending feedback strategies of Elias [6] and Schalkwijk and Kailath [7] (see [8]), while the converse followed from a cut-set argument. For more than two users the capacity region remains unknown. Thomas [9] proved that feedback can at most double the sum-rate capacity for any number of users. Iacobucci and Di Benedetto [10, 11] extended Ozarow's scheme to more than two users, but their strategies do not achieve capacity and perform worse than the no-feedback capacity for more than three users. Kramer [12] subsequently developed a method he called Fourier-Modulated Estimate Correction, or Fourier-MEC. For the symmetric GMAC and sufficiently large signal-to-noise ratio (SNR), the Fourier-MEC sum-rate meets the cut-set bound and is thus optimal. However, the problem remained open for low SNRs.

The coding schemes in [5, 6, 7, 10, 11, 12] start by mapping the message onto a point on the real line or complex plane, and they iteratively correct the receiver's estimate of this point by using linear minimum mean square error (LMMSE) estimation. There are many variants of the schemes. For example, one can convert complex-channel strategies to real-channel strategies [12], one can interpret the LMMSE step as posterior matching [13, 14], and one can use multi-dimensional Fourier transforms for Fourier-MEC, e.g., a Hadamard transform [12, 14].

New *outer* bounds on the capacity region were derived in [15] by applying the Hekstra-Willems *dependence-balance* argument [16]. This idea of dependence balance is to restrict the set of permissible input distributions to improve standard cut-set bounds. The intuitive argument given in [16] is that the amount of depen-

[1] The material of this chapter has appeared in
- E. Sula, M. Gastpar, and G. Kramer, "Sum-rate capacity for symmetric Gaussian multiple access channels with feedback," in *IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, USA, June 2018.
- E. Sula, M. Gastpar, and G. Kramer, "Sum-rate capacity for symmetric Gaussian multiple access channels with feedback," *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 2860−2871, 2020.

dence *consumed* cannot exceed the amount of dependence *produced*. Remarkably, the dependence-balance bounds of [15] match the Fourier-MEC sum-rates of [12] when evaluated with Gaussian signals. However, inferring the optimality of Gaussian signaling is not trivial. Investigations for *linear* feedback strategies appear in [17] where the maximum sum-rate is computed under a symmetric block power constraint.

Multiple-access channels (MACs) with feedback and non-Gaussian noise have also received attention. For example, outer bounds on the feedback capacity region of MACs with binary additive noise are derived in [18]. This class of MACs was also studied in [19, 20] to show that the Cover-Leung [21] achievable rate region can be improved. Other results for GMACs with imperfect and/or noisy feedback are presented in [22, 23, 24, 25].

### 3.1.1 Contribution

Our result unifies [12, 15, 17] and adds a missing piece of the puzzle. In particular, in [12] the problem remained open for low SNRs, in [15] the optimality of Gaussian signals was not proven, and in [17] the problem was solved for linear strategies only. Our starting point is the generalized dependence-balance outer bounds in [15], and we show that the best sum-rate obtained with these bounds is at most the sum-rate achieved in [12, Sec. V]. We thus have the following result.

**Theorem 3.** *The feedback sum-rate capacity of the $J$-user symmetric GMAC is*

$$C_{sum} = \frac{1}{2}\log_2(1 + PJ\alpha) \quad \textit{bits/channel use} \tag{3.1}$$

*where $P$ is the available power for each user and $\alpha$ is the unique solution satisfying $\alpha \in [1, J]$ and*

$$(1 + PJ\alpha)^{J-1} = (1 + P\alpha(J - \alpha))^J. \tag{3.2}$$

Achievability was established in [12, Section V] and the converse is given in Section 3.5. The converse combines the Lagrange-duality approach of [26] with a variant of the factorization of convex envelopes used in [27, 28] that was inspired by work on functional inequalities [29, 30]. The high level steps behind the proof are as follows:

- For the Lagrange dual of our capacity maximization problem, we establish the *existence* of a maximizing probability distribution (Appendix 3.7.1).
- For the Lagrange dual, we show that if a probability distribution is a maximizer, then so is the probability distribution of a random variable that is the sum of two independent random variables, each of which is distributed as the maximizing distribution (Lemma 3). This is the core of the argument, and it is established by the technique of factorization of convex envelopes.
- By induction, one can infer that the probability distribution of a random variable that is the sum of $2^\ell$ ($\ell \in \mathbb{Z}_+$) independent random variables, each distributed according to the maximizing distribution, must also be a maximizing distribution (Theorem 5).

- Finally, by a central limit theorem argument, a Gaussian distribution is a maximizer (Appendix 3.7.3).

Besides establishing the optimality of Gaussian signals, the non-convex Lagrangian dual problem is converted into a convex problem that is solved in Lemmas 4 – 7, which is an alternative proof to [17].

### 3.1.2 Organization

This chaper is organized as follows. In Section 3.2, we present the system model and in Section 3.3 we review existing capacity bounds. In Section 3.4, we give an upper bound on the sum-rate for general GMACs with feedback. In Section 3.5, we prove Theorem 3. Section 3.6 concludes the chapter and the appendices provide supporting results and proofs.

## 3.2 System Model

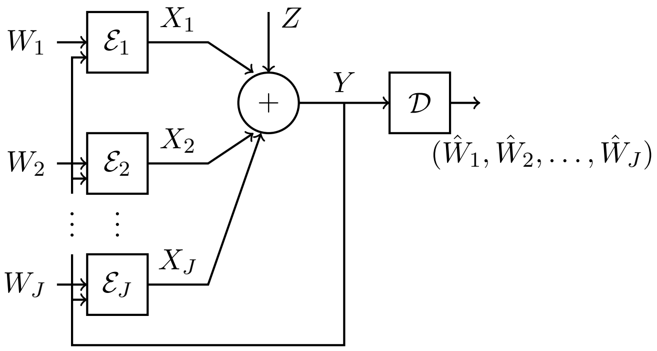


Figure 3.1 – The Multiple Access Channel with Feedback

Consider a GMAC with $J$ transmitters (called users) with channel input symbols $X_1, X_2, \ldots, X_J$, and a receiver with the channel output symbol $Y$. The received signal at time instant $i$ is

$$Y_i = Z_i + \sum_{j=1}^{J} g_j X_{j,i} \tag{3.3}$$

for $i = 1, 2, \ldots, n$, where $Z_1, Z_2, \ldots, Z_n$ is a string of independent and identically distributed (i.i.d.) zero-mean Gaussian noise variables with unit variance and $g_1, g_2, \ldots, g_J$ are channel gains. The $J$ channel inputs have the block power constraints

$$\sum_{i=1}^{n} \mathbb{E}\left[X_{j,i}^2\right] \leq nP_j, \quad j = 1, 2, \ldots, J. \tag{3.4}$$

The SNR of user $j$ is thus $P_j g_j^2$. If $P_1 g_1^2 = P_2 g_2^2 = \ldots = P_J g_J^2$, then the transmitters can be swapped without changing the capacity. For such models, we may as well set $P_j = P$ and $g_j = 1$ for all $j$, and we refer to this channel as the *symmetric* GMAC.

Let $W_j$ with $nR_j$ bits be the message of user $j$. The transmitted signal at time instant $i$ is

$$X_{j,i} = f_{j,i}(W_j, Y^{i-1}), \quad j = 1, 2, \ldots, J \tag{3.5}$$

where the $f_{j,i}(\cdot)$ are encoding functions to be optimized. The receiver puts out the estimates

$$\left(\hat{W}_1, \hat{W}_2, \ldots, \hat{W}_J\right) = g(Y^n) \tag{3.6}$$

where $g(\cdot)$ is a decoding function. The event that the receiver makes an error is

$$\mathcal{E} = \bigcup_{j=1}^{J} \left\{\hat{W}_j \neq W_j\right\}. \tag{3.7}$$

The rate-tuple $\boldsymbol{R} = (R_1, R_2, \ldots, R_J)$ is said to be *achievable* if, for any specified positive error probability $P_e$ and sufficiently large $n$, there are encoding functions and a decoder such that $\Pr[\mathcal{E}] \leq P_e$. The closure of the set of achievable $\boldsymbol{R}$ is called the *capacity region* $\mathcal{C}_{\text{MAC-FB}}$. We are interested in characterizing the *sum-rate capacity* $C_{\text{sum}}$, i.e., the maximum sum of the entries of any $\boldsymbol{R}$ in $\mathcal{C}_{\text{MAC-FB}}$.

## 3.3 Dependence Balance Bounds

### 3.3.1 Two-User Dependence Balance Bounds

Dependence balance bounds were introduced by Hekstra and Willems [16] for single output two-way channels. The tool generalizes to other models such as MACs with feedback. For example, for the two-user MAC with feedback, the achievable $(R_1, R_2)$ must satisfy

$$\begin{aligned} 0 &\leq R_1 \leq I(X_1; Y|X_2, T) \\ 0 &\leq R_2 \leq I(X_2; Y|X_1, T) \\ R_1 + R_2 &\leq I(X_1, X_2; Y|T) \end{aligned} \tag{3.8}$$

for some $p(t, x_1, x_2, y)$ for which

$$T - [X_1, X_2] - Y \text{ forms a Markov chain,} \tag{3.9}$$

$$I(X_1; X_2|T) \leq I(X_1; X_2|Y, T). \tag{3.10}$$

In [16, Section 7], the term $I(X_1; X_2|T)$ is interpreted as the amount of dependence *consumed*, and $I(X_1; X_2|Y, T)$ as the amount of dependence *produced* by communication. An interpretation of the inequality (3.10) is thus that the dependence consumed cannot exceed the dependence produced, i.e., communication is limited by *dependence balance.*

Other interpretations of this bound are described in [15]. Observe that (3.10) can be rewritten in the following two ways:

$$I(X_1; Y|T) + I(X_2; Y|T) \leq I(X_1, X_2; Y|T) \tag{3.11}$$

$$I(X_1, X_2; Y|T) \leq I(X_1; Y|X_2, T) + I(X_2; Y|X_1, T). \tag{3.12}$$

The bound (3.12) requires the set function $f : 2^{\{1,2\}} \to \mathbb{R}$ defined by

$$f(\{1\}) = I(X_1; Y|X_2, T) \tag{3.13}$$

$$f(\{2\}) = I(X_2; Y|X_1, T) \tag{3.14}$$

$$f(\{1, 2\}) = I(X_1, X_2; Y|T) \tag{3.15}$$

to be *submodular.* In other words, for valid choices of $p(t, x_1, x_2, y)$, the rate region defined by (3.8) is a *polymatroid.* We may thus interpret dependence balance as a submodularity (or polymatroid) constraint, i.e., communication is limited by submodularity.

We remark that for two-user GMACs the dependence balance bound yields the same rate region as the standard cut-set bound. However, the dependence balance bound is more informative in the following sense. Consider jointly Gaussian $p(t, x_1, x_2, y)$. The optimal correlation coefficient $\rho^*$ in (3.17) below is the one that satisfies (3.10)-(3.12) with equality. However, dependence balance (or submodularity) limits $\rho$ to the range $[0, \rho^*]$, whereas the cut-set bound permits all $\rho$ in $[0, 1]$.

**Example 1** (Two-User Capacity with Feedback)**.** *Feedback enables the users to cooperate so as to increase rates. For $J = 2$ the capacity region is known to be [5]*

$$\mathcal{C}_{MAC\text{-}FB} = \bigcup_{0\le\rho\le1} \mathcal{R}(\rho) \tag{3.16}$$

*where $\mathcal{R}(\rho)$ is the set of rate pairs $(R_1, R_2)$ that satisfy*

$$\begin{aligned} 0 \le R_1 &\le \frac{1}{2}\log\left(1 + P_1(1-\rho^2)\right) \\ 0 \le R_2 &\le \frac{1}{2}\log\left(1 + P_2(1-\rho^2)\right) \\ R_1 + R_2 &\le \frac{1}{2}\log\left(1 + P_1 + P_2 + 2\rho\sqrt{P_1P_2}\right). \end{aligned} \tag{3.17}$$

*The parameter $\rho$ is the correlation coefficient of the two users. $\mathcal{C}_{MAC\text{-}FB}$ is here the same as a standard cut-set bound. However, we show that cut-set bounds are loose for $J > 2$, and that dependence balance bounds can characterize the fundamental limits of communication.*

### 3.3.2 Multi-User Dependence Balance Bounds

The two-user dependence balance concept was generalized to $J$ users in [20, Thm. 4] and more dependence balance bounds are derived in [15, Thm. 1]. The capacity region of the $J$-user MAC with feedback is a subset of the set of rate-tuples $(R_1, R_2, \ldots, R_J)$ satisfying

$$R_{\mathcal{S}} \le I(X_{\mathcal{S}}; Y | X_{\mathcal{S}^C}, T) \tag{3.18}$$

for all $\mathcal{S} \subseteq \mathcal{J}$, where $\mathcal{S}^C$ is the complement of $\mathcal{S}$, and where

$$T - [X_1, X_2, \ldots, X_J] - Y \text{ forms a Markov chain} \tag{3.19}$$

$$I(X_1, X_2, \ldots, X_J; Y|T) \le \frac{1}{M-1}\sum_{m=1}^{M} I(X_{\mathcal{S}_m^C}; Y | X_{\mathcal{S}_m}, T), \tag{3.20}$$

for any partition $\{\mathcal{S}_m\}_{m=1}^M$ of $\mathcal{J}$ into $M \ge 2$ subsets. One may again interpret (3.20) as a submodular constraint. For example, for the partition $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2\}$, …, $\mathcal{S}_J = \{J\}$ the dependence balance constraint (3.20) becomes

$$I(X_1, X_2, \ldots, X_J; Y|T) \le \frac{1}{J-1}\sum_{j=1}^{J} I(X_{\mathcal{J}\setminus\{j\}}; Y | X_j, T) \tag{3.21}$$

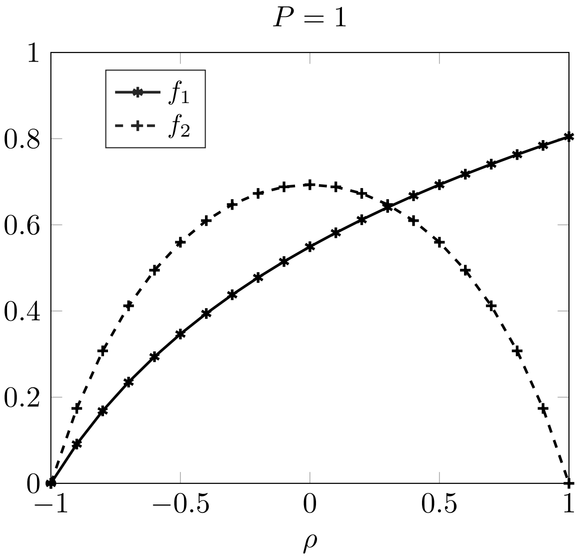


Figure 3.2 – Cut-set bounds for the sum-rate of a two-user symmetric GMAC with feedback.

where $\mathcal{J}\setminus\{j\}$ is the set $\{1, 2, \ldots, j-1, j+1, \ldots, J\}$. As usual, one can add the power constraints $\mathbb{E}[X_j^2] \le P_j$, $j \in \mathcal{J}$, to these bounds. Also, as for (3.11) for the two-user case, the bound (3.20) can be written as

$$\sum_{m=1}^{M} I(X_{\mathcal{S}_m}; Y|T) \le I(X_1, X_2, \ldots, X_J; Y|T). \tag{3.22}$$

### 3.3.3 Comparison to Cut-set Bounds

The cut-set bounds give the following result, see [5] and [4, Theorem 15.10.1].

**Proposition 1.** *For the two-user symmetric GMAC with feedback, we have*

$$C_{sum} \le \max_{0\le\rho\le1} \min\left\{ \underbrace{\frac{1}{2}\log(1+2P(1+\rho)}_{f_1(\rho)}, \underbrace{\log(1+P(1-\rho^2))}_{f_2(\rho)} \right\}. \tag{3.23}$$

The sum-rate on the right hand side (RHS) of (3.23) turns out to be achievable [5], and it is depicted in Figure 3.2. Similarly, again starting from [4, Theorem 15.10.1], we obtain the following result.

**Proposition 2.** *For the three-user symmetric GMAC with feedback, we have*

$$C_{sum} \le \max_{0\le\rho\le1} \min\left\{ \underbrace{\frac{1}{2}\log(1+3P(1+2\rho)}_{g_1(\rho)}, \right. \tag{3.24}$$

$$\left. \underbrace{\frac{3}{4}\log(1+2P(1-\rho)(1+2\rho))}_{g_2(\rho)}, \underbrace{\frac{3}{2}\log\left(1+\frac{P(1+2\rho)(1-\rho)}{1+\rho}\right)}_{g_3(\rho)} \right\}.$$

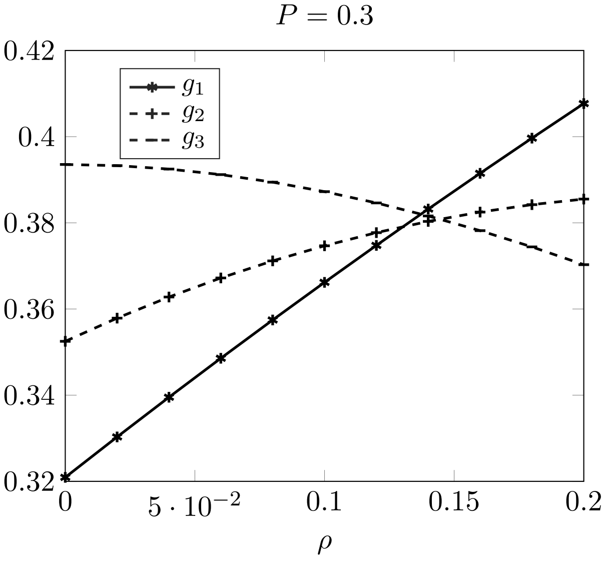


Figure 3.3 – Cut-set bounds for the sum-rate of a three-user symmetric GMAC with feedback and $P = 0.3$.

The sum-rate (3.25) is not generally achievable. Figure 3.3 illustrates the situation for the special case $P = 0.3$. The cut-set bound of Proposition 2 leads to an upper bound on the sum-rate given by the intersection point of the curves $g_2$ and $g_3$. However, we show that the capacity is given by the intersection of the curves $g_1$ and $g_2$; this intersection point is achieved by Fourier-MEC [12].

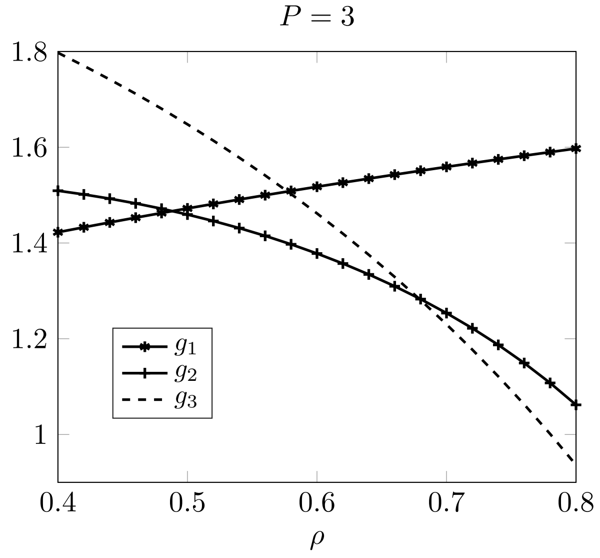


Figure 3.4 – Cut-set bounds for the sum-rate of a three-user symmetric GMAC with feedback and $P = 3$.

For the symmetric GMAC and large SNR, i.e., more than a certain threshold, the Fourier-MEC sum-rate meets the cut-set bound. For example, a sufficient condition for the cut-set bound to give the sum-rate capacity is that the SNR is greater than or equal to $2^{J+1}/J^2$ [12]. Observe that this threshold grows exponentially with the number of users. For $J = 3$ users, the threshold becomes $\frac{16}{9}$, thus we pick $P = 3$, that is strictly larger than the threshold. The cut-set bound of Proposition 2 is the intersection point of the curves $g_1$ and $g_2$ that is achieved by Fourier-MEC [12] illustrated in Figure 3.4.

## 3.4 General Converse Bound for $J$ Users

We derive the following upper bounds on the feedback sum-rate capacity of the general $J$-user GMAC.

**Theorem 4.** *For any $\lambda \geq 0$ and any partition $\{\mathcal{S}_m\}_{m=1}^{M}$ of $\mathcal{J}$ into $M \geq 2$ subsets, we have*

$$\begin{aligned} C_{\text{sum}} \leq \max_{p(x_1,x_2,\ldots,x_J)\in G_{\mathcal{G}}} & (1-\lambda)I(X_1,X_2,\ldots,X_J;Y) \\ & + \frac{\lambda}{M-1}\sum_{m=1}^{M} I(X_{\mathcal{S}_m^C};Y|X_{\mathcal{S}_m}) \end{aligned} \tag{3.25}$$

*where $G_{\mathcal{G}}$ is the set of zero-mean Gaussian distributions satisfying $\mathbb{E}[X_j^2] \leq P_j$ for $j = 1, 2, \ldots, J$.*

### 3.4.1 Proof of Theorem 4

Our converse bound starts from Section 3.3.2 and we use the shorthand $\boldsymbol{X} = (X_1, X_2, \ldots, X_J)$. We find it convenient to express our problem as a minimization, i.e., we seek to minimize $-I(\boldsymbol{X};Y|T)$. We consider only the sum-rate obtained by $\mathcal{S} = \mathcal{J}$ in (3.18) over the input distributions $p(t,\boldsymbol{x})$ that satisfy the dependence-balance constraint

$$I(\boldsymbol{X};Y|T) \leq \frac{1}{M-1}\sum_{m=1}^{M} I(X_{\mathcal{S}_m^C};Y|X_{\mathcal{S}_m},T) \tag{3.26}$$

for $(T,\boldsymbol{X})$ such that $T - \boldsymbol{X} - Y$ forms a Markov chain. We will treat the power constraints in two steps. First, for any fixed covariance matrix $K_{\boldsymbol{X}}$, we will optimize over all distributions satisfying $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T] = K_{\boldsymbol{X}}$. Then, we will optimize over all $K_{\boldsymbol{X}}$ whose diagonal entries are at most $P$.

After, we form the Lagrangian for our optimization problem as

$$s_\lambda(\boldsymbol{X}|T) :=(\lambda-1)I(\boldsymbol{X};Y|T) - \frac{\lambda}{M-1}\sum_{m=1}^{M} I(X_{\mathcal{S}_m^C};Y|X_{\mathcal{S}_m},T). \tag{3.27}$$

This can be rewritten as

$$s_\lambda(\boldsymbol{X}|T) =-\left(\frac{\lambda}{M-1}+1\right)I(\boldsymbol{X};Y|T) + \frac{\lambda}{M-1}\sum_{m=1}^{M} I(X_{\mathcal{S}_m};Y|T). \tag{3.28}$$

The lower convex envelope is defined as in [27], namely, as

$$\breve{s}_\lambda(\boldsymbol{X}) = \inf_{\substack{p(t|\boldsymbol{x}):\\ T-\boldsymbol{X}-Y}} \{s_\lambda(\boldsymbol{X}|T)\}$$

and we note that $\breve{s}_\lambda(\boldsymbol{X})$ is a convex function of $p(\boldsymbol{x})$ because $\breve{s}_\lambda(\boldsymbol{X})$ is the lower convex envelope of $s_\lambda(\boldsymbol{X})$, which is defined by discarding the random variable $T$ in (3.27). In addition, we define

$$\breve{s}_\lambda(\boldsymbol{X}|T) = \sum_t p(t)\breve{s}_\lambda(\boldsymbol{X}|T=t). \tag{3.29}$$

The dual function of our problem for $K_{\boldsymbol{X}} \succeq 0$ is

$$V_\lambda(K_{\boldsymbol{X}}) := \inf_{p(\boldsymbol{x}):\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]=K_{\boldsymbol{X}}} \left\{\breve{s}_\lambda(\boldsymbol{X})\right\}. \tag{3.30}$$

Alternatively, we have

$$V_\lambda(K_{\boldsymbol{X}}) = \inf_{\substack{p(t,\boldsymbol{x}):\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]=K_{\boldsymbol{X}}\\ T-\boldsymbol{X}-Y}} \left\{s_\lambda(\boldsymbol{X}|T)\right\}. \tag{3.31}$$

By the standard Lagrangian duality we bound the original optimization problem as follows

$$C_{\text{sum}} \le -\max_\lambda \inf_{\substack{p(t,\boldsymbol{x}):\ \mathbb{E}[X_j^2]\le P_j\\ T-\boldsymbol{X}-Y}} s_\lambda(\boldsymbol{X}|T) \tag{3.32}$$

$$= -\max_\lambda \inf_{\substack{K_{\boldsymbol{X}}\succeq 0:\\ [K_{\boldsymbol{X}}]_{jj}\le P_j}} \underbrace{\inf_{\substack{p(t,\boldsymbol{x}):\ \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]=K_{\boldsymbol{X}}\\ T-\boldsymbol{X}-Y}} s_\lambda(\boldsymbol{X}|T)}_{V_\lambda(K_{\boldsymbol{X}})}. \tag{3.33}$$

From now on we deal mainly with the dual function $V_\lambda(K_{\boldsymbol{X}})$. For $0 < \lambda \le 1$, the minimization problem (3.31) is a convex problem, and it follows from (3.27) and maximum entropy results that the optimizing distribution $p(t, x_1, x_2, \dots, x_J)$ is jointly Gaussian. The more difficult case is $\lambda > 1$. Our approach will be to establish that one of the distributions attaining the minimum in (3.31) is the Gaussian channel input, but we do not establish that this is the unique minimizer. This follows from a novel variant of the factorization of convex envelopes.

Consider two independent uses of the GMAC:

$$\begin{aligned} Y_1 &= G\boldsymbol{X}_1 + Z_1\\ Y_2 &= G\boldsymbol{X}_2 + Z_2 \end{aligned} \tag{3.34}$$

where $G = \begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent and identically distributed, and where $Z_1, Z_2 \sim \mathcal{N}(0,1)$ are independent. One key difference to [27] is that $G$ is *not* an invertible matrix. We define

$$\boldsymbol{X}_{\theta_1} = \frac{1}{\sqrt{2}}(\boldsymbol{X}_1 + \boldsymbol{X}_2), \quad \boldsymbol{X}_{\theta_2} = \frac{1}{\sqrt{2}}(\boldsymbol{X}_1 - \boldsymbol{X}_2), \tag{3.35}$$

$$Y_{\theta_1} = \frac{1}{\sqrt{2}}(Y_1 + Y_2), \quad Y_{\theta_2} = \frac{1}{\sqrt{2}}(Y_1 - Y_2). \tag{3.36}$$

We thus have

$$Y_{\theta_1} = G\boldsymbol{X}_{\theta_1} + \tilde{Z}_1, \quad Y_{\theta_2} = G\boldsymbol{X}_{\theta_2} + \tilde{Z}_2 \tag{3.37}$$

where $\tilde{Z}_1, \tilde{Z}_2 \sim \mathcal{N}(0,1)$ are independent. Moreover, we generalize the definition (3.28) to the two-letter extension as

$$\begin{aligned} s_\lambda(\boldsymbol{X}_1, \boldsymbol{X}_2|T) := &-\left(\frac{\lambda}{M-1}+1\right) I(\boldsymbol{X}_1, \boldsymbol{X}_2; Y_1, Y_2|T) \\ &+\frac{\lambda}{M-1}\sum_{m=1}^{M} I([\boldsymbol{X}_1]_{\mathcal{S}_m}, [\boldsymbol{X}_2]_{\mathcal{S}_m}; Y_1, Y_2|T). \end{aligned} \tag{3.38}$$

The following proposition establishes the existence of a minimizer in (3.30).

**Proposition 3.** *There is a pair of random variables $(T_*, \boldsymbol{X}_*)$ with $|\mathcal{T}_*| \le \frac{J(J+1)}{2}+1$ and $\mathbb{E}[\boldsymbol{X}_*\boldsymbol{X}_*^T] = K_{\boldsymbol{X}}$ such that*

$$V_\lambda(K_{\boldsymbol{X}}) = s_\lambda(\boldsymbol{X}_*|T_*). \tag{3.39}$$

*Proof*: Existence and the cardinality bound on $T_*$ are established in Appendix 3.7.1 by using a similar argument as in [27, Appendix 2A]. □

We can now establish the desired result.

**Lemma 3.** *Let $p_*(t,\boldsymbol{x})$ attain $V_\lambda(K_{\boldsymbol{X}})$ and let $(T_1, T_2, \boldsymbol{X}_1, \boldsymbol{X}_2) \sim p_*(t_1, \boldsymbol{x}_1)p_*(t_2, \boldsymbol{x}_2)$. Suppose $\boldsymbol{X}_t$ has conditional distribution $p_*(\boldsymbol{x}|T=t)$ and define*

$$\begin{aligned} \boldsymbol{X}_{\theta_1}|((T_1,T_2)=(t_1,t_2)) &:= \frac{1}{\sqrt{2}}(\boldsymbol{X}_{t_1}+\boldsymbol{X}_{t_2}), \\ Y_{\theta_1}|((T_1,T_2)=(t_1,t_2)) &:= \frac{1}{\sqrt{2}}(Y_{t_1}+Y_{t_2}), \\ \boldsymbol{X}_{\theta_2}|((T_1,T_2)=(t_1,t_2)) &:= \frac{1}{\sqrt{2}}(\boldsymbol{X}_{t_1}-\boldsymbol{X}_{t_2}), \\ Y_{\theta_2}|((T_1,T_2)=(t_1,t_2)) &:= \frac{1}{\sqrt{2}}(Y_{t_1}-Y_{t_2}). \end{aligned}$$

*Then, using $T=(T_1,T_2)$, we have*

1. *$(T, \boldsymbol{X}_{\theta_1})$ also attains $V_\lambda(K_{\boldsymbol{X}})$,*
2. *$(T, \boldsymbol{X}_{\theta_2})$ also attains $V_\lambda(K_{\boldsymbol{X}})$.*
3. *The joint distribution $(T, \boldsymbol{X}_{\theta_1}, \boldsymbol{X}_{\theta_2})$ must satisfy*
   - $I(Y_{\theta_1}; [\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m} \mid [\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}, T) = 0$
   - $I(Y_{\theta_2}; [\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m} \mid Y_{\theta_1}, [\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}, T) = 0$

   *for $m = 1, \ldots, M$.*

*Proof*: See Appendix 3.7.2. □

**Corollary 1.** *For every $\ell \in \mathbb{N}$ , let $n = 2^\ell$ and $(T^n, \boldsymbol{X}^n) \sim \prod_{i=1}^n p_*(t_i, \boldsymbol{x}_i)$. Then $(T^n, \tilde{\boldsymbol{X}}_n)$ achieves $V_\lambda(K_{\boldsymbol{X}})$ where $\tilde{\boldsymbol{X}}_n|(T^n = (t_1, t_2, \ldots, t_n)) := \frac{1}{\sqrt{n}}(\boldsymbol{X}_{t_1} + \boldsymbol{X}_{t_2} + \cdots + \boldsymbol{X}_{t_n})$. We choose $\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}, \ldots, \boldsymbol{X}_{t_n}$ to be independent random variables.*

*Proof*: The proof follows by induction using Lemma 3. □

**Theorem 5.** *There is a Gaussian distribution (i.e., $T$ can be chosen to be a constant) that achieves $V_\lambda(K_{\boldsymbol{X}})$.*

*Proof*: See Appendix 3.7.3. □

Note that our approach does not establish the uniqueness of the minimizing distribution. Using Theorem 5 in equation (3.33) completes the proof of Theorem 4.

## 3.5 Feedback sum-rate capacity for symmetric GMACs

The proof of Theorem 3 is a special case of the proof of Theorem 4 with the partition $\mathcal{S}_1 = \{1\}, \ldots, \mathcal{S}_J = \{J\}$ where the dependence balance constraint is given in (3.21) or (3.22). We tackle the resulting (non-convex) optimization problem with Lagrange duality.

Consider the covariance matrix

$$K_{\boldsymbol{X}} = \begin{pmatrix} Q_1 & \rho_{12}\sqrt{Q_1 Q_2} & \ldots & \rho_{1J}\sqrt{Q_1 Q_J} \\ \rho_{21}\sqrt{Q_2 Q_1} & Q_2 & \ldots & \rho_{2J}\sqrt{Q_2 Q_J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{J1}\sqrt{Q_J Q_1} & \rho_{J2}\sqrt{Q_J Q_2} & \ldots & Q_J \end{pmatrix}. \tag{3.40}$$

**Lemma 4.** *We have the bound*

$$-C_{\text{sum}} = \min_{\substack{p(t,\boldsymbol{x}):\mathbb{E}[X_j^2]\le P_j \\ T-\boldsymbol{X}-Y \\ \textit{subject to } (3.20)}} -I(\boldsymbol{X};Y|T) \ge \max_{\lambda} \min_{K_{\boldsymbol{X}}\succeq 0: Q_j \le P_j} q(\lambda, K_{\boldsymbol{X}}) \tag{3.41}$$

*where*

$$\begin{aligned} q(\lambda, K_{\boldsymbol{X}}) = {} & \frac{(\lambda-1)}{2}\log\left(1 + \sum_{j,k=1}^{J} [K_{\boldsymbol{X}}]_{jk}\right) \\ & - \frac{\lambda}{2(J-1)} \sum_{j=1}^{J} \log\left(1 + \sum_{\ell,k=1}^{J} [K_{\boldsymbol{X}}]_{\ell k} - \frac{\left(\sum_{k=1}^{J} [K_{\boldsymbol{X}}]_{jk}\right)^2}{Q_j}\right). \end{aligned} \tag{3.42}$$

*Proof*: See Appendix 3.7.4. □

We have shown that the optimal input distributions are Gaussian. At this point the problem is similar to the one in [17], and we can use Lemmas 4 and 5 from [17] to complete the optimization. However, the converse in [17] relies on a specific covariance matrix form with only two variables, and this does not necessarily work for asymmetric power constraints. Therefore, we provide a different analysis that applies to asymmetric power constraints.

Consider the covariance matrix

$$M_{\boldsymbol{X}} = \begin{pmatrix} P_1 & \rho_{12}\sqrt{P_1 P_2} & \ldots & \rho_{1J}\sqrt{P_1 P_J} \\ \rho_{21}\sqrt{P_2 P_1} & P_2 & \ldots & \rho_{2J}\sqrt{P_2 P_J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{J1}\sqrt{P_J P_1} & \rho_{J2}\sqrt{P_J P_2} & \ldots & P_J \end{pmatrix}. \tag{3.43}$$

**Lemma 5.** *For every $\lambda \geq 0$, we have*

$$\max_{\lambda} \min_{K_{\boldsymbol{X}} \succeq 0 : Q_j \leq P_j} q(\lambda, K_{\boldsymbol{X}}) \geq \max_{\lambda} \min_{\rho_{12},\ldots,\rho_{J(J-1)} : M_{\boldsymbol{X}} \succeq 0} q(\lambda, M_{\boldsymbol{X}}). \tag{3.44}$$

*Proof*: See Appendix 3.7.5. □

Note that Lemma 5 holds with equality, while it suffices to have it as an inequality with the specified direction.

We now set all power constraints to be the same, i.e., $P_1 = P_2 = \cdots = P_J = P$.

**Lemma 6.** *For every $\lambda \geq 0$, we have*

$$\max_{\lambda} \min_{\rho_{12},\ldots,\rho_{J(J-1)} : M_{\boldsymbol{X}} \succeq 0} q(\lambda, M_{\boldsymbol{X}}) \geq \max_{\lambda} \min_{\beta \in [0,J]} \tag{3.45}$$
$$\left\{ \frac{(\lambda - 1)}{2} \log(1 + JP\beta) - \frac{J\lambda}{2(J-1)} \log\left(1 + P\beta\left(J - \beta\right)\right) \right\}.$$

*Proof*: See Appendix 3.7.6. □

**Remark 1.** *The proof of Lemma 6 is different from [17] (see also [12]).*

We define the function

$$\ell(\beta, J, P) = \frac{1}{2} \log\left(1 + JP\beta\right) - \frac{J}{2(J-1)} \log\left(1 + P\beta\left(J - \beta\right)\right). \tag{3.46}$$

**Lemma 7.** *We have*

$$\max_{\lambda} \min_{\beta \in [0,J]} \left\{ -\frac{1}{2} \log(1 + JP\beta) + \lambda \ell(\beta, J, P) \right\} = -\frac{1}{2} \log\left(1 + JP\beta^*\right). \tag{3.47}$$

*where $\beta^*$ is the unique solution to $\beta^* \in [1, J]$ and* $(1+JP\beta^*)^{J-1} = \left(1 + P\beta^*\left(J - \beta^*\right)\right)^J$.

*Proof*: We have

$$\max_{\lambda} \min_{\beta \in [0,J]} \left\{ -\frac{1}{2} \log(1 + JP\beta) + \lambda \ell(\beta, J, P) \right\} = \min_{\beta \in [1,J] : \ell(\beta,J,P) \leq 0} \left\{ -\frac{1}{2} \log(1 + JP\beta) \right\} \tag{3.48}$$

$$= -\frac{1}{2} \log\left(1 + JP\beta^*\right) \tag{3.49}$$

where $\beta^*$ is the unique solution satisfying $\beta^* \in [1, J]$ and $(1 + PJ\beta^*)^{J-1} = (1 + P\beta^*(J - \beta^*))^J$, (3.48) follows from strong duality as the problem is convex from Lemma 8 in Appendix 3.7.7 and satisfies Slater's condition. Slater's condition holds because there exists a $\beta$ such that $\ell(\beta, J, P) < 0$, e.g., $\beta = 1$ so the primal problem is strictly feasible. Equation (3.49) follows from the Karush-Kuhn-Tucker (KKT) conditions for a convex problem which satisfy Slater's condition for the optimal $\beta^*$ and $\lambda^*$. We start by showing that $\lambda^* \neq 0$. Suppose that $\lambda^* = 0$, then from the KKT conditions we have

$$\frac{\partial}{\partial \beta} \left\{ -\frac{1}{2} \log(1 + JP\beta) + \lambda \ell(\beta, J, P) \right\} \Bigg|_{\lambda=0} = -\frac{JP}{2(1 + JP\beta)} = 0 \tag{3.50}$$

which implies that $P = 0$. This is impossible, so by contradiction we have $\lambda^* \neq 0$. Now the complementary slackness condition $\lambda^* \cdot \ell(\beta^*, J, P) = 0$ gives $\ell(\beta^*, J, P) = 0$, which is equivalent to

$$(1 + JP\beta^*)^{J-1} = (1 + P\beta^* (J - \beta^*))^J . \tag{3.51}$$

This equation has a unique solution for $\beta^* \in [1, J]$, see [12, Lemma 1], [17, Appendix A]. □

By combining Lemmas 4-7 we obtain

$$C_{\text{sum}} \leq \frac{1}{2} \log (1 + PJ\beta^*) \tag{3.52}$$

where $\beta^*$ is the unique solution satisfying $\beta^* \in [1, J]$ and (3.51).

## 3.6 Conclusions

We derived a new converse bound that combines the Lagrange duality approach of [15] with a novel variation of the factorization of convex envelopes [27]. The new converse bound meets the achievable sum-rate of the Fourier MEC scheme, thus establishing the sum-rate capacity for the $J$-user symmetric GMAC with feedback.

It remains to see whether, as in [5], Fourier MEC combined with successive interference cancellation can achieve all rate points in the capacity region of the symmetric GMAC with feedback. For asymmetric transmit power constraints, however, it is known that Fourier-MEC can be improved by using modulation frequencies other than the uniformly-spaced frequencies $(j-1)/J$ for $j \in \mathcal{J}$. A few more variations of MEC strategies are described in [12, Sec. VIII].

**Example 2.** *Fourier MEC does not meet the dependence-balance bound under asymmetric power constraints. For example, consider three users with the power constraints $P_1 \leq 1, P_2 \leq 4$ and $P_3 \leq 9$, for which Fourier-MEC achieves the sum-rate $R_{sum} = 1.6215$ bits/use, see [12, Sec. III]. However, the dependence balance bound permits a larger sum-rate, since the choice $(\rho_{12}, \rho_{13}, \rho_{23}) = (0.5, 0.44, 0.58)$ satisfies the dependence balance constraints and permits $R_{sum} = 1.6427$.*

## 3.7 Appendix

### 3.7.1 Proof of Proposition 3

The difference compared to [27] is Proposition 8, where instead of the continuity we prove only lower semi-continuity via a different technique.

**Proposition 4** ([31, Lemma 1]). *Suppose that $Y_n$ and $Y$ have continuous densities $f_n(y)$ and $f(y)$ with respect to the Lebesgue measure on $\mathbb{R}$. If $Y_n \overset{w}{\Rightarrow} Y$ and*

$$\sup_n |f_n(y)| \leq M(y) < \infty, \ \forall y \in \mathbb{R} \tag{3.53}$$

*and $f_n$* is equicontinuous, i.e., $\forall\, y, \epsilon > 0$, $\exists \delta(y, \epsilon), n(y, \epsilon)$ *such that $\|y - y_1\| < \delta(y, \epsilon)$ implies that $|f_n(y) - f_n(y_1)| < \epsilon \ \forall n \geq n(y, \epsilon)$, then for any compact subset $C$ of $\mathbb{R}$*

*we have*

$$\sup_{y \in C} |f_n(y) - f(y)| \to 0 \text{ as } n \to \infty. \tag{3.54}$$

*If $\{f_n\}$ is uniformly equicontinuous, i.e., $\delta(y,\epsilon)$, $n(y,\epsilon)$ do not depend on $y$, and $f(y_n) \to 0$ whenever $\|y_n\| \to \infty$ then*

$$\sup_{y \in \mathbb{R}} |f_n(y) - f(y)| = \|f_n(y) - f(y)\|_\infty \to 0 \text{ as } n \to \infty. \tag{3.55}$$

**Proposition 5** ([27, Proposition 16])**.** *Let $\{\boldsymbol{X}_n\}$ be a sequence of random variables satisfying $Y_n = G\boldsymbol{X}_n + \boldsymbol{Z}$ where $\boldsymbol{Z} \sim \mathcal{N}(0, I)$ is independent of $\{\boldsymbol{X}_n\}$ and $f_n(y)$ represents the density of $Y_n$. Then the collection of functions $\{f_n(y)\}$ is uniformly bounded and uniformly equicontinuous.*

**Definition 2.** *A collection of random variables $\boldsymbol{X}_n$ on $\mathbb{R}^N$ is said to be* tight *if for every $\epsilon > 0$ there is a compact set $C_\epsilon \subset \mathbb{R}^N$ such that $P(\boldsymbol{X}_n \notin C_\epsilon) \leq \epsilon$, $\forall n$.*

**Proposition 6** ([27, Proposition 17])**.** *Consider a sequence of random variables $\{\boldsymbol{X}_n\}$ for which $\mathbb{E}[\boldsymbol{X}_n \boldsymbol{X}_n^T] = K$, $\forall n$. Then the sequence is tight.*

**Theorem 6** (Prokhorov)**.** *If $\{\boldsymbol{X}_n\}$ is a tight sequence of random variables in $\mathbb{R}^N$ then there exists a subsequence $\{\boldsymbol{X}_{n_i}\}$ and a limiting probability distribution $\boldsymbol{X}_*$ such that $\boldsymbol{X}_{n_i} \overset{w}{\Rightarrow} \boldsymbol{X}_*$.*

**Proposition 7** ([27, Proposition 18])**.** *Let $\boldsymbol{X}_n \overset{w}{\Rightarrow} \boldsymbol{X}_*$ and let $Z \sim \mathcal{N}(0,1)$ be pairwise independent of $\{\boldsymbol{X}_n\}$, $\boldsymbol{X}_*$. Let $Y_n = G\boldsymbol{X}_n + Z$, $Y_* = G\boldsymbol{X}_* + Z$. Further let $\mathbb{E}[\boldsymbol{X}_n \boldsymbol{X}_n^T] = K$, $\mathbb{E}[\boldsymbol{X}_* \boldsymbol{X}_*^T] = K$. Let $f_n(y)$ denote the density of $Y_n$ and $f_*(y)$ denote the density of $Y_*$. Then we have*

1. *$Y_n \overset{w}{\Rightarrow} Y_*$,*
2. *$f_n(y) \to f_*(y)$ for all $y$,*
3. *$h(Y_n) \to h(Y_*)$.*

**Proposition 8** (Lower Semi-continuity)**.** *Let $\boldsymbol{X}_n \overset{w}{\Rightarrow} \boldsymbol{X}_*$ and $Y_n = G\boldsymbol{X}_n + Z$, $Y_* = G\boldsymbol{X}_* + Z$, where $Z \sim \mathcal{N}(0,1)$ is pairwise independent of $\{\boldsymbol{X}_n\}$, $\boldsymbol{X}_*$. Let $s_\lambda(\boldsymbol{X}_n) = (\lambda - 1)h(Y_n) + \left(\frac{\lambda}{J-1} + 1\right) h(Z) - \frac{\lambda}{J-1} \sum_{j=1}^{J} h(Y_n | X_{jn})$ and $s_\lambda(\boldsymbol{X}_*)$ similarly. Then*

1. *$(Y_n, X_{1n}) \overset{w}{\Rightarrow} (Y_*, X_{1*})$,*
2. *$\liminf_{n\to\infty} s_\lambda(\boldsymbol{X}_n) \geq s_\lambda(\boldsymbol{X}_*)$.*

*Proof*: The first part follows from pointwise convergence of characteristic functions (which is equivalent to weak convergence by Levy's continuity theorem) since $\Phi_{(\boldsymbol{X}_n, Z)}(\boldsymbol{u}, v) = \mathbb{E}[e^{i\boldsymbol{u}^T \boldsymbol{X}_n + ivZ}] = \mathbb{E}[e^{i\boldsymbol{u}^T \boldsymbol{X}_n}]\mathbb{E}[e^{ivZ}] = \Phi_{\boldsymbol{X}_n}(\boldsymbol{u})\Phi_Z(v)$. By letting $n \to \infty$ we have $\Phi_{\boldsymbol{X}_*}(\boldsymbol{u})\Phi_Z(v) = \mathbb{E}[e^{i\boldsymbol{u}^T \boldsymbol{X}_*}]\mathbb{E}[e^{ivZ}] = \mathbb{E}[e^{i\boldsymbol{u}^T \boldsymbol{X}_* + ivZ}] = \Phi_{(\boldsymbol{X}_*, Z)}(\boldsymbol{u}, v)$. To relate $(Y_n, X_{1n})$ with $(\boldsymbol{X}_n, Z)$ we use the linear transformation $(Y_n, X_{1n})^T = A(\boldsymbol{X}_n, Z)^T$ for a deterministic matrix $A$. By using the previous steps and the linear dependence we obtain $\lim_{n\to\infty} \Phi_{(Y_n, X_{1n})}(\boldsymbol{t}) = \lim_{n\to\infty} \Phi_{(\boldsymbol{X}_n, Z)}(A\boldsymbol{t}) = \Phi_{(X_*, Z)}(A\boldsymbol{t}) = \Phi_{(Y_*, X_{1*})}(\boldsymbol{t})$.

For the second part, we fix $\delta > 0$ and define $N_\delta \sim \mathcal{N}(0,\delta)$ pairwise independent of $\{\boldsymbol{X}_n\}$, $\boldsymbol{X}_*$. By the third claim of proposition 7, we obtain

$$\begin{aligned}&(\lambda-1)h(Y_n)+\left(\frac{\lambda}{J-1}+1\right)h(Z)-\frac{\lambda}{J-1}\sum_{j=1}^{J}h(Y_n|X_{jn}+N_\delta)\\ &\to(\lambda-1)h(Y_*)+\left(\frac{\lambda}{J-1}+1\right)h(Z)-\frac{\lambda}{J-1}\sum_{j=1}^{J}h(Y_*|X_{j*}+N_\delta)\end{aligned} \tag{3.56}$$

as $n \to \infty$. From the Markov chain $(X_{1n}+N_\delta)-X_{1n}-Y_n$ and the data processing inequality we have $h(Y_n|X_{1n}) \le h(Y_n|X_{1n}+N_\delta)$. By using the aforementioned inequality, we have

$$\liminf_{n\to\infty} s_\lambda(\boldsymbol{X}_n) \ge (\lambda-1)h(Y_*)+\left(\frac{\lambda}{J-1}+1\right)h(Z)-\frac{\lambda}{J-1}\sum_{j=1}^{J}h(Y_*|X_{j*}+N_\delta). \tag{3.57}$$

Since the RHS of (3.57) is continuous in $\delta$ for $\delta > 0$, we take $\delta \downarrow 0$. We have $\liminf_{n\to\infty} s_\lambda(\boldsymbol{X}_n) \ge s_\lambda(\boldsymbol{X}_*)$, which is the definition of lower semi-continuity. This proves the second claim. □

**Theorem 7** ([32, Theorem 1])**.** *Let $\{Y_i \in \mathcal{C}\}$ be a sequence of continuous random variables with densities $\{f_i\}$, and $Y_*$ be a continuous random variable with density $f_*$ such that $f_i \to f_*$ pointwise. Let $\|y\| = \sqrt{y^\dagger y}$ denote the Euclidean norm of $y \in \mathcal{C}$. If the conditions*

$$\max\{\sup_y f_i(y), \sup_y f_*(y)\} \le F \tag{3.58}$$

$$\max\{\int \|y\|^\kappa f_i(y)dy, \int \|y\|^\kappa f_*(y)dy\} \le L \tag{3.59}$$

*hold for some $\kappa > 1$ and for all $i$ then $h(Y_i) \to h(Y_*)$.*

**Remark 2.** *We have* $\liminf_i h(Y_i) \ge h(Y_*)$ *due to the upper bound on the densities and* $\limsup_i h(Y_i) \le h(Y_*)$ *due to the moment constraints.*

*Proof of Proposition 3*: The proof is similar to [27, Proposition 7]. Define

$$v_\lambda(\hat{K}) = \inf_{p(\boldsymbol{x}):E[\boldsymbol{X}\boldsymbol{X}^T]=\hat{K}} s_\lambda(\boldsymbol{X}). \tag{3.60}$$

Let $\boldsymbol{X}_n$ be a sequence of random variables such that $\mathbb{E}[\boldsymbol{X}_n\boldsymbol{X}_n^T] = \hat{K}$ and $s_\lambda(\boldsymbol{X}_n) \downarrow v_\lambda(\hat{K})$. By the covariance constraint (Proposition 6) the sequence of random variables $\boldsymbol{X}_n$ forms a tight sequence and by Theorem 6 there exists $X^*_{\hat{K}}$ and a convergent subsequence such that $\boldsymbol{X}_{n_i} \overset{w}{\Rightarrow} \boldsymbol{X}^*_{\hat{K}}$. From Proposition 7 and 8 we have $s_\lambda(\boldsymbol{X}^*_{\hat{K}}) = v_\lambda(\hat{K})$. For $\lambda \ge 0$, we have

$$\begin{aligned} v_\lambda(\hat{K}) = s_\lambda(\boldsymbol{X}^*_{\hat{K}}) &\geq -\left(\frac{\lambda}{J-1}+1\right) I(\boldsymbol{X}^*_{\hat{K}};Y) \\ &\geq \left(\frac{-\lambda-J+1}{2(J-1)}\right)\log(1+G\hat{K}G^T) = C_\lambda. \end{aligned} \tag{3.61}$$

Recall that $V_\lambda(K_{\boldsymbol{X}})$ is defined using a convex combination:

$$V_\lambda(K_{\boldsymbol{X}}) = \inf_{\substack{(T,\boldsymbol{X}):\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]=K_{\boldsymbol{X}} \\ T-\boldsymbol{X}-Y}} s_\lambda(\boldsymbol{X}|T). \tag{3.62}$$

To obtain the best convex combination subject to the covariance constraint it suffices to consider the family of maximizers $\boldsymbol{X}^*_{\hat{K}}$ for $\hat{K} \succeq 0$. Thus, we have

$$V_\lambda(K_{\boldsymbol{X}}) = \inf_{\substack{\alpha_i,\hat{K}_i:\alpha_i\geq 0,\sum_i \alpha_i=1 \\ \sum_i \alpha_i\hat{K}_i=K_{\boldsymbol{X}}}} \sum\nolimits_i \alpha_i v_\lambda(\hat{K}_i). \tag{3.63}$$

It takes $\frac{J(J+1)}{2}$ constraints to preserve the covariance matrix and one constraint to preserve $\sum_i \alpha_i v_\lambda(\hat{K}_i)$. Hence, by using the Bunt-Caratheodory theorem, we can consider convex combinations of at most $m := \frac{J(J+1)}{2}+1$ points, i.e., we have

$$V_\lambda(K_{\boldsymbol{X}}) = \inf_{\substack{\alpha_i,\hat{K}_i:\alpha_i\geq 0,\sum_{i=1}^m \alpha_i=1 \\ \sum_{i=1}^m \alpha_i\hat{K}_i=K_{\boldsymbol{X}}}} \sum_{i=1}^m \alpha_i v_\lambda(\hat{K}_i). \tag{3.64}$$

Consider any sequence of convex combinations $(\{\alpha_i^n\},\{K_i^n\})$ that approaches the supremum as $n\to\infty$. Using compactness of the $m-$dimensional simplex, we can assume w.l.o.g. that $\alpha_i^n \overset{n\to\infty}{\to} \alpha_i^*$, $i=1,\ldots,m$. If any $\alpha_i^*=0$, since $\alpha_i^n K_i^n = K_{\boldsymbol{X}}$ and $v_\lambda(K_i^n) \geq C_\lambda$ it is easy to see that $\alpha_i^n v_\lambda(K_i^n) \overset{n\to\infty}{\to} 0$. Thus we can assume that $\min_{i=1,\ldots,m}\alpha_i^* = \alpha^* > 0$. We can find a convergent subsequence for each $i$, $1\leq i\leq m$, so that $K_i^{n_k} \overset{k\to\infty}{\to} K_i^*$. We thus have

$$V_\lambda(K_{\boldsymbol{X}}) = \sum_{i=1}^m \alpha_i^* v_\lambda(\hat{K}_i^*). \tag{3.65}$$

In other words, we can find a pair of random variables $(T_*,\boldsymbol{X}_*)$ with $|\mathcal{T}| \leq \frac{J(J+1)}{2}+1$ such that $V_\lambda(K_{\boldsymbol{X}}) = s_\lambda(\boldsymbol{X}_*|T_*)$. □

### 3.7.2 Proof of Lemma 3

Now we state and prove four key propositions needed for the proof of the lemma.

**Proposition 9.** $I(\boldsymbol{X}_1,\boldsymbol{X}_2;Y_1,Y_2) = I(\boldsymbol{X}_{\theta_1},\boldsymbol{X}_{\theta_2};Y_{\theta_1},Y_{\theta_2})$.

*Proof*: The function $f(x,y) = \big((x+y)/\sqrt{2},(x-y)/\sqrt{2}\big)$ is bijective. □

**Proposition 10.** *The chain* $Y_{\theta_1}-\boldsymbol{X}_{\theta_1}-\boldsymbol{X}_{\theta_2}-Y_{\theta_2}$ *is Markov and we have*

$$I(\boldsymbol{X}_{\theta_1},\boldsymbol{X}_{\theta_2};Y_{\theta_1},Y_{\theta_2}) = I(\boldsymbol{X}_{\theta_1};Y_{\theta_1}) + I(\boldsymbol{X}_{\theta_2};Y_{\theta_2}|Y_{\theta_1}). \tag{3.66}$$

*Proof*: The Markovity follows by (3.37). We further compute

$$\begin{aligned} I(\boldsymbol{X}_{\theta_1}, \boldsymbol{X}_{\theta_2}; Y_{\theta_1}, Y_{\theta_2}) &= I(\boldsymbol{X}_{\theta_1}, \boldsymbol{X}_{\theta_2}; Y_{\theta_1}) + I(\boldsymbol{X}_{\theta_1}, \boldsymbol{X}_{\theta_2}; Y_{\theta_2}|Y_{\theta_1}) \\ &= I(\boldsymbol{X}_{\theta_1}; Y_{\theta_1}) + I(\boldsymbol{X}_{\theta_2}; Y_{\theta_1}|\boldsymbol{X}_{\theta_1}) \\ &\quad + I(\boldsymbol{X}_{\theta_2}; Y_{\theta_2}|Y_{\theta_1}) + I(\boldsymbol{X}_{\theta_1}; Y_{\theta_2}|Y_{\theta_1}, \boldsymbol{X}_{\theta_2}) \\ &= I(\boldsymbol{X}_{\theta_1}; Y_{\theta_1}) + I(\boldsymbol{X}_{\theta_2}; Y_{\theta_2}|Y_{\theta_1}) \end{aligned} \tag{3.67}$$

where the last step follows from the Markov chain. □

**Proposition 11.** *For any $\lambda \geq 0$ we have*

$$s_\lambda(\boldsymbol{X}_{\theta_1}, \boldsymbol{X}_{\theta_2}|T) \geq s_\lambda(\boldsymbol{X}_{\theta_1}|T) + s_\lambda(\boldsymbol{X}_{\theta_2}|Y_{\theta_1}, T) \tag{3.68}$$

*with equality if and only if we have*

- $I(Y_{\theta_1}; [\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}|[\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}, T) = 0$
- $I(Y_{\theta_2}; [\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}|Y_{\theta_1}, [\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}, T) = 0$

*for $m = 1, \ldots, M$.*

*Proof*: We compute

$$s_\lambda(\boldsymbol{X}_{\theta_1}, \boldsymbol{X}_{\theta_2}|T) - s_\lambda(\boldsymbol{X}_{\theta_1}|T) - s_\lambda(\boldsymbol{X}_{\theta_2}|Y_{\theta_1}, T) \tag{3.69}$$

$$\begin{aligned} &= \left(\frac{\lambda}{M-1} + 1\right)(-I(\boldsymbol{X}_{\theta_1}, \boldsymbol{X}_{\theta_2}; Y_{\theta_1}, Y_{\theta_2}|T) + I(\boldsymbol{X}_{\theta_1}; Y_{\theta_1}|T) \\ &+I(\boldsymbol{X}_{\theta_2}; Y_{\theta_2}|Y_{\theta_1}, T)) + \frac{\lambda}{M-1}\left(\sum_{m=1}^{M} I([\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}, [\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}; Y_{\theta_1}, Y_{\theta_2}|T)\right. \\ &-I([\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}; Y_{\theta_1}|T) - I([\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}; Y_{\theta_2}|Y_{\theta_1}, T)) \end{aligned} \tag{3.70}$$

$$\begin{aligned} &= \frac{\lambda}{M-1}\sum_{m=1}^{M}(I([\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}, [\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}; Y_{\theta_1}|T) - I([\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}; Y_{\theta_2}|Y_{\theta_1}, T) \\ &+I([\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}, [\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}; Y_{\theta_2}|Y_{\theta_1}, T) - I([\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}; Y_{\theta_1}|T)) \end{aligned} \tag{3.71}$$

$$\begin{aligned} &= \frac{\lambda}{M-1}\sum_{m=1}^{M}(I([\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}; Y_{\theta_1}|[\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}, T) \\ &+I([\boldsymbol{X}_{\theta_1}]_{\mathcal{S}_m}; Y_{\theta_2}|Y_{\theta_1}, [\boldsymbol{X}_{\theta_2}]_{\mathcal{S}_m}, T)) \geq 0 \end{aligned} \tag{3.72}$$

where (3.71) follows from Proposition 10. The last step follows from the nonnegativity of mutual information. □

*Proof of Lemma 3*: We have

$$\begin{aligned} 2V_\lambda(K_{\boldsymbol{X}}) &= s_\lambda(\boldsymbol{X}_1|T_1) + s_\lambda(\boldsymbol{X}_2|T_2) && (3.73) \\ &= s_\lambda(\boldsymbol{X}_1, \boldsymbol{X}_2|T_1, T_2) && (3.74) \\ &= s_\lambda(\boldsymbol{X}_{\theta_1}, \boldsymbol{X}_{\theta_2}|T_1, T_2) && (3.75) \\ &\geq s_\lambda(\boldsymbol{X}_{\theta_1}|T_1, T_2) + s_\lambda(\boldsymbol{X}_{\theta_2}|Y_{\theta_1}, T_1, T_2) && (3.76) \\ &\geq \breve{s}_\lambda(\boldsymbol{X}_{\theta_1}) + \breve{s}_\lambda(\boldsymbol{X}_{\theta_2}|Y_{\theta_1}) && (3.77) \\ &\geq \breve{s}_\lambda(\boldsymbol{X}_{\theta_1}) + \breve{s}_\lambda(\boldsymbol{X}_{\theta_2}) && (3.78) \\ &\geq 2V_\lambda(K_{\boldsymbol{X}}), && (3.79) \end{aligned}$$

where (3.73) is valid for the distribution $p_*(t,\boldsymbol{x})$ that attains $V_\lambda(K_{\boldsymbol{X}})$; (3.74) follows since $(T_1,\boldsymbol{X}_1)$ and $(T_2,\boldsymbol{X}_2)$ are independent by assumption; (3.75) can be proved in the same way as Proposition 9; (3.76) follows by Proposition 11; (3.77) follows from

$$\begin{aligned} s_\lambda(\boldsymbol{X}_{\theta_2}|T,Y_{\theta_1}) &= \sum_{y_{\theta_1}} p(y_{\theta_1}) s_\lambda(\boldsymbol{X}_{\theta_2}|T,Y_{\theta_1}=y_{\theta_1}) &(3.80)\\ &\geq \sum_{y_{\theta_1}} p(y_{\theta_1}) \breve{s}_\lambda(\boldsymbol{X}_{\theta_2}|Y_{\theta_1}=y_{\theta_1}) &(3.81)\\ &= \breve{s}_\lambda(\boldsymbol{X}_{\theta_2}|Y_{\theta_1}) &(3.82)\end{aligned}$$

where (3.81) follows because $\breve{s}_\lambda(\boldsymbol{X}_{\theta_2}|Y_{\theta_1}=y_{\theta_1})$ is the lower convex envelope of $s_\lambda(\boldsymbol{X}_{\theta_2}|Y_{\theta_1}=y_{\theta_1})$ and the chain $T-\boldsymbol{X}_{\theta_2}-Y_{\theta_2}$ conditioned on $Y_{\theta_1}=y_{\theta_1}$ is Markov (Markovity is implied by (3.37) where $T=(T_1,T_2)$) and (3.82) is the definition of $\breve{s}_\lambda(.|.)$; step (3.78) follows since $\breve{s}_\lambda(\boldsymbol{X}_{\theta_1})$ is convex in $p(\boldsymbol{x}_{\theta_1})$ and by Jensen's inequality $\breve{s}_\lambda(\boldsymbol{X}_{\theta_2}|Y_{\theta_1}) \geq \breve{s}_\lambda(\boldsymbol{X}_{\theta_2})$; (3.79) follows from definition of $V_\lambda(K_{\boldsymbol{X}})$ and by checking the constraint

$$\begin{aligned}\mathbb{E}[\boldsymbol{X}_{\theta_1}\boldsymbol{X}_{\theta_1}^T] &= \sum_{t_1,t_2} p_*(t_1)p_*(t_2)\frac{1}{2}\left(\mathbb{E}[\boldsymbol{X}_{t_1}\boldsymbol{X}_{t_1}^T]+\mathbb{E}[\boldsymbol{X}_{t_2}\boldsymbol{X}_{t_2}^T]\right)\\ &= \sum_t p_*(t)\mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_t^T] = K_{\boldsymbol{X}}.\end{aligned}$$

We now see that all inequalities (3.73) – (3.79) are equalities, and Equation (3.76) combined with Proposition 11 proves the third claim. The first two claims follow from Equation (3.79), see the definition of $V_\lambda(K_{\boldsymbol{X}})$. □

### 3.7.3 Proof of Theorem 5

The proof is the same as in [27, Appendix IV] and we include it for completeness. Define the set of typical sequences as

$$\mathcal{T}^{(n)}(T) := \left\{t^n : \left||\{i : t_i = t\}| - np_*(t)\right| \leq n\omega_n p_*(t), \forall t \in [1:m]\right\},$$

where $\omega_n$ is any sequence such that $\omega_n \to 0$ as $n\to\infty$ and $\omega_n\sqrt{n}\to\infty$ as $n\to\infty$. For any sequence $t^n \in \mathcal{T}^{(n)}(T)$, let $A_n(t) = |\{i : t_i = t\}|$. Thus, the mean of $A_n(t)$ is $np_*(t)$ and the variance is $np_*(t)(1-p_*(t))$. For instance, one may use $\omega_n = \frac{\log n}{\sqrt{n}}$ . By Chebyshev's inequality, we have

$$P(\left||\{i : t_i = t\}| - np_*(t)\right| > n\omega_n p_*(t)) \leq \frac{1-p_*(t)}{p_*(t)\omega_n^2 n}.$$

Hence $P(t^n \notin \mathcal{T}^{(n)}(T)) \to 0$ as $n\to\infty$. Consider a sequence of random variables $\hat{\boldsymbol{X}}_n := \tilde{\boldsymbol{X}}_n|(T^n = t^n)$.

**Proposition 12.** $\hat{\boldsymbol{X}}_n \overset{w}{\Rightarrow} \mathcal{N}(0, \sum_{t=1}^m p_*(t)K_t)$.

*Proof*: We know that $A_n(t) \in np_*(t)(1\pm\omega_n)$, $\forall t$. Consider a $\boldsymbol{c}$ with real entries and $\|\boldsymbol{c}\| = 1$. Let $\hat{\boldsymbol{X}}_{n,i}^{\boldsymbol{c}} := \frac{1}{\sqrt{n}}\boldsymbol{c}^T\boldsymbol{X}_{t_i}$ such that $\hat{\boldsymbol{X}}_{n,i}^{\boldsymbol{c}}$ and $\hat{\boldsymbol{X}}_{n,j}^{\boldsymbol{c}}$ are independent random variables for $i \neq j$ (recall that the $\boldsymbol{X}_{t_i}$ have zero mean). Observe that $\sum_{i=1}^{n} \hat{\boldsymbol{X}}_{n,i}^{\boldsymbol{c}} = \boldsymbol{c}^T \hat{\boldsymbol{X}}_n$. Note that

$$\sum_{i=1}^{n} \mathbb{E}[(\hat{\boldsymbol{X}}_{n,i}^{\boldsymbol{c}})^2] = \frac{1}{n} \sum_t A_n(t) \boldsymbol{c}^T K_t \boldsymbol{c} \to \boldsymbol{c}^T \left( \sum_t p_*(t) K_t \right) \boldsymbol{c},$$

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}[(\hat{\boldsymbol{X}}_{n,i}^{\boldsymbol{c}})^2; |\hat{\boldsymbol{X}}_{n,i}^{\boldsymbol{c}}| > \epsilon_1] &= \frac{1}{n} \sum_t A_n(t) \mathbb{E}[\boldsymbol{c}^T \boldsymbol{X}_t \boldsymbol{X}_t^T \boldsymbol{c}; \boldsymbol{c}^T \boldsymbol{X}_t \boldsymbol{X}_t^T \boldsymbol{c} > n\epsilon_1^2] \\ &\leq \sum_t p_*(t)(1+\omega_n) \mathbb{E}[\boldsymbol{c}^T \boldsymbol{X}_t \boldsymbol{X}_t^T \boldsymbol{c}; \boldsymbol{c}^T \boldsymbol{X}_t \boldsymbol{X}_t^T \boldsymbol{c} > n\epsilon_1^2] \to 0.\end{aligned}$$

For the last step, we used that the $K_t$'s are bounded, and hence $\boldsymbol{c}^T \boldsymbol{X}_t$ has a bounded second moment. The Lindeberg-Feller Central Limit Theorem gives $\sum_{i=1}^{n} \hat{\boldsymbol{X}}_{n,i}^{\boldsymbol{c}} \overset{w}{\Rightarrow} \mathcal{N}(0, \boldsymbol{c}^T \sum_t p_*(t) K_t \boldsymbol{c})$. Hence, using the Cramér-Wold theorem we obtain $\hat{\boldsymbol{X}}_n \overset{w}{\Rightarrow} \mathcal{N}(0, \sum_t p_*(t) K_t)$. □

**Proposition 13.** *Given any $\delta > 0$, there exists $N_0$ such that $\forall n > N_0$ we have for all $t^n \in \mathcal{T}^{(n)}(T)$*

$$s_\lambda(\tilde{\boldsymbol{X}}_n | T^n = t^n) - s_\lambda(\boldsymbol{X}^*) \leq \delta$$

*where $\boldsymbol{X}^* \sim \mathcal{N}(0, \sum_t p_*(t) K_t)$.*

*Proof*: Suppose the claim is not true. Then we have a subsequence $t^{n_k} \in \mathcal{T}^{n_k}(T)$ and random variable $\tilde{\boldsymbol{X}}_{n_k} | T^{n_k} = t^{n_k}$ such that

$$s_\lambda(\tilde{\boldsymbol{X}}_{n_k} | t^{n_k}) > s_\lambda(\boldsymbol{X}^*) + \delta, \forall k.$$

However from Proposition 12 we know that $\tilde{\boldsymbol{X}}_{n_k} | t^{n_k} \overset{w}{\Rightarrow} \boldsymbol{X}^*$ and from Proposition 7 we have $s_\lambda(\tilde{\boldsymbol{X}}_{n_k} | t^{n_k}) \to s_\lambda(\boldsymbol{X}^*)$, a contradiction. □

*Proof of Theorem 5*: We know from Corollary 1 that for every $\ell \in \mathbb{N}$ and $n = 2^\ell$, the pair $(T^n, \tilde{\boldsymbol{X}}_n)$ achieves $V_\lambda(K_{\boldsymbol{X}})$. Hence we have

$$\begin{aligned} V_\lambda(K_{\boldsymbol{X}}) &= \sum_{t^n} p_*(t^n) s_\lambda(\tilde{\boldsymbol{X}}_n | t^n) \\ &= \sum_{t^n \in \mathcal{T}^{(n)}(T)} p_*(t^n) s_\lambda(\tilde{\boldsymbol{X}}_n | t^n) + \sum_{t^n \notin \mathcal{T}^{(n)}(T)} p_*(t^n) s_\lambda(\tilde{\boldsymbol{X}}_n | t^n). \end{aligned}$$

For a given $t^n$, let $\hat{\boldsymbol{X}} := \tilde{\boldsymbol{X}}_n | T^n = t^n$ so that $\mathbb{E}[\hat{\boldsymbol{X}} \hat{\boldsymbol{X}}^T] \preceq \sum_{t=1}^{m} K_t$. Thus, we have $s_\lambda(\hat{\boldsymbol{X}}) \leq C_\lambda$ for some fixed constant that is independent of $t^n$. Using Proposition 13 we can upper bound $V_\lambda(K_{\boldsymbol{X}})$ for large $n$ by

$$\begin{aligned} V_\lambda(K_{\boldsymbol{X}}) &= \sum_{t^n \in \mathcal{T}^{(n)}(T)} p_*(t^n) s_\lambda(\tilde{\boldsymbol{X}}_n | t^n) + \sum_{t^n \notin \mathcal{T}^{(n)}(T)} p_*(t^n) s_\lambda(\tilde{\boldsymbol{X}}_n | t^n) \\ &\leq \sum_{t^n \in \mathcal{T}^{(n)}(T)} p_*(t^n)(s_\lambda(\boldsymbol{X}^*) + \delta) + C_\lambda \sum_{t^n \notin \mathcal{T}^{(n)}(T)} p_*(t^n) \\ &= P(t^n \in \mathcal{T}^{(n)}(T))(s_\lambda(\boldsymbol{X}^*) + \delta) + C_\lambda P(t^n \notin \mathcal{T}^{(n)}(T)). \end{aligned}$$

Here $\boldsymbol{X}^* \sim \mathcal{N}(0, \sum_t p_*(t) K_t)$. Since $P(t^n \in \mathcal{T}^{(n)}(T)) \to 1$ as $n \to \infty$ we have $V_\lambda(K_{\boldsymbol{X}}) \leq s_\lambda(\boldsymbol{X}^*) + \delta$. However, $\delta > 0$ is arbitrary, and hence $V_\lambda(K_{\boldsymbol{X}}) \leq s_\lambda(\boldsymbol{X}^*)$. The other direction $V_\lambda(K_{\boldsymbol{X}}) \geq s_\lambda(\boldsymbol{X}^*)$ follows from the definition of $V_\lambda(K_{\boldsymbol{X}})$ and $\sum_t p_*(t) K_t = K_{\boldsymbol{X}}$. □

### 3.7.4 Proof of Lemma 4

For the GMAC with channel gains $g_j = 1$ for all $j$ we have

$$K_{\boldsymbol{X}Y} = \begin{pmatrix} K_{\boldsymbol{X}} & \operatorname{Cov}(\boldsymbol{X}, Y) \\ \operatorname{Cov}(\boldsymbol{X}, Y)^T & K_Y \end{pmatrix} = \begin{pmatrix} K_{\boldsymbol{X}} & K_{\boldsymbol{X}}\mathbf{1} \\ (K_{\boldsymbol{X}}\mathbf{1})^T & 1 + \mathbf{1}^T K_{\boldsymbol{X}} \mathbf{1} \end{pmatrix} \tag{3.83}$$

where $K_Y = 1 + \mathbf{1}^T K_{\boldsymbol{X}} \mathbf{1}$, $\operatorname{Cov}(\boldsymbol{X}, Y) = K_{\boldsymbol{X}}\mathbf{1}$ and $\mathbf{1}$ is the column vector of all ones. By standard Lagrangian duality, we have

$$-C_{\text{sum}} = \min_{\substack{p(t,\boldsymbol{x}):\ \mathbb{E}[X_j^2] \leq P_j \\ T - \boldsymbol{X} - Y \\ \text{subject to (3.20)}}} -I(\boldsymbol{X}; Y | T) \geq \max_\lambda \min_{p(\boldsymbol{x}) \in G_{\mathcal{G}}} s_\lambda(\boldsymbol{X}) \tag{3.84}$$

$$= \max_\lambda \min_{K_{\boldsymbol{X}} \succeq 0: Q_j \leq P_j} \frac{(\lambda - 1)}{2} \log \det K_Y - \frac{\lambda}{2(J-1)} \log \frac{\prod_{j=1}^J \det K_{X_j Y}}{\prod_{j=1}^J \det K_{X_j}} \tag{3.85}$$

$$= \max_\lambda \min_{K_{\boldsymbol{X}} \succeq 0: Q_j \leq P_j} \left\{ \frac{(\lambda - 1)}{2} \log \left( 1 + \sum_{j,k=1}^J [K_{\boldsymbol{X}}]_{jk} \right) \right.$$
$$\left. - \frac{\lambda}{2(J-1)} \sum_{j=1}^J \log \left( 1 + \sum_{\ell,k=1}^J [K_{\boldsymbol{X}}]_{\ell k} - \frac{\left( \sum_{k=1}^J [K_{\boldsymbol{X}}]_{jk} \right)^2}{Q_j} \right) \right\} \tag{3.86}$$

where (3.84) follows from Theorem 4, and step (3.85) follows by inserting $\mathcal{S}_1 = \{1\}, \ldots, \mathcal{S}_J = \{J\}$ in (3.27) to obtain

$$s_\lambda(\boldsymbol{X}) = (\lambda - 1) I(\boldsymbol{X}; Y) - \frac{\lambda}{J-1} \sum_{j=1}^J I(X_j; Y). \tag{3.87}$$

By using Gaussian inputs in (3.87), we obtain

$$I(\boldsymbol{X}; Y) = \frac{1}{2} \log \det K_Y, \quad I(X_j; Y) = \frac{1}{2} \log \frac{\det K_{X_j Y}}{\det K_{X_j}} \tag{3.88}$$

and combining the two equalities we have

$$\min_{p(\boldsymbol{x}) \in G_{\mathcal{G}}} s_\lambda(\boldsymbol{X}) = \min_{K_{\boldsymbol{X}} \succeq 0: Q_j \leq P_j} \frac{(\lambda - 1)}{2} \log \det K_Y - \frac{\lambda}{2(J-1)} \log \frac{\prod_{j=1}^J \det K_{X_j Y}}{\prod_{j=1}^J \det K_{X_j}}. \tag{3.89}$$

The RHS of (3.89) represents the function $q(\cdot)$. Step (3.86) above follows by computing the determinants and simplifying. □

### 3.7.5 Proof of Lemma 5

The function $q(\cdot)$ in (3.89) can be rewritten as

$$2q(\lambda, K_{\boldsymbol{X}}) = -\log\det K_Y - \frac{\lambda}{J-1}\log\frac{\prod_{j=1}^{J}\det K_{X_jY}}{(\det K_Y)^{J-1}\prod_{j=1}^{J}\det K_{X_j}}. \tag{3.90}$$

We define $K'_{\boldsymbol{X}}$ to be the same as $K_{\boldsymbol{X}}$ except that the (1,1) entry of $K'_{\boldsymbol{X}}$ is $P_1$ rather than $Q_1$. Then from (3.83) we have

$$K'_{X_1Y} = K_{X_1Y} \circ \begin{pmatrix} D & F \\ F & E \end{pmatrix}, \tag{3.91}$$

$$K'_{X_jY} = K_{X_jY} \circ \begin{pmatrix} 1 & 1 \\ 1 & E \end{pmatrix}, \quad j \neq 1 \tag{3.92}$$

where '$\circ$' denotes Hadamard multiplication, and where

$$D = \frac{P_1}{Q_1}, \quad E = \frac{K_Y + P_1 - Q_1}{K_Y}, \quad F = \frac{\mathrm{Cov}\,(X_1, Y) + P_1 - Q_1}{\mathrm{Cov}\,(X_1, Y)}. \tag{3.93}$$

Observe that $D > 1$ and $E > 1$. Oppenheim's inequality ($\det K'_{X_1Y} \geq DE\det K_{X_1Y}$) [33, p. 480] thus gives

$$2q(\lambda, K'_{\boldsymbol{X}}) = -\log EK_Y - \frac{\lambda}{J-1}\log\frac{\left(\prod_{j=2}^{J}\det K'_{X_jY}\right)\det K'_{X_1Y}}{E^{J-1}DK_Y^{J-1}\prod_{j=1}^{J}K_{X_j}} \tag{3.94}$$

$$\leq -\log EK_Y - \frac{\lambda}{J-1}\log E\frac{\prod_{j=1}^{J}\det K_{X_jY}}{K_Y^{J-1}\prod_{j=1}^{J}K_{X_j}} \tag{3.95}$$

$$= -(1 + \frac{\lambda}{J-1})\log E + 2q(\lambda, K_{\boldsymbol{X}}) \leq 2q(\lambda, K_{\boldsymbol{X}}). \tag{3.96}$$

The minimum is attained for $K'_{\boldsymbol{X}}$, however $K'_{\boldsymbol{X}}$ is not in a standard covariance matrix form. We show that it can be rewritten as a standard covariance matrix with the following correlation coefficients

$$K'_{\boldsymbol{X}} = \begin{pmatrix} P_1 & \rho_{12}\sqrt{Q_1Q_2} & \cdots & \rho_{1J}\sqrt{Q_1Q_J} \\ \rho_{21}\sqrt{Q_2Q_1} & Q_2 & \cdots & \rho_{2J}\sqrt{Q_2Q_J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{J1}\sqrt{Q_JQ_1} & \rho_{J2}\sqrt{Q_JQ_2} & \cdots & Q_J \end{pmatrix} \tag{3.97}$$

$$= \begin{pmatrix} P_1 & \rho_{12}\sqrt{\frac{Q_1}{P_1}}\sqrt{P_1Q_2} & \cdots & \rho_{1J}\sqrt{\frac{Q_1}{P_1}}\sqrt{P_1Q_J} \\ \rho_{21}\sqrt{\frac{Q_1}{P_1}}\sqrt{Q_2P_1} & Q_2 & \cdots & \rho_{2J}\sqrt{Q_2Q_J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{J1}\sqrt{\frac{Q_1}{P_1}}\sqrt{Q_JP_1} & \rho_{J2}\sqrt{Q_JQ_2} & \cdots & Q_J \end{pmatrix}.$$

With the same approach we attain the minimum for $Q_2 = P_2$, $Q_3 = P_3$, …, $Q_J = P_J$. Thus, we have the desired lower bound for the original problem.

### 3.7.6 Proof of Lemma 6

Consider the arithmetic mean

$$\rho = \frac{1}{J(J-1)} \left( \sum_{\substack{j,k=1 \\ j \neq k}}^{J} \rho_{jk} \right). \tag{3.98}$$

For the inequality in Lemma 6 to hold we need to show that

$$\prod_{j=1}^{J} \left( 1 + \sum_{\ell,k=1}^{J} [M_{\boldsymbol{X}}]_{\ell k} - \frac{\left(\sum_{k=1}^{J} [M_{\boldsymbol{X}}]_{jk}\right)^2}{P} \right) \leq (1 + P\beta\,(J-\beta))^J. \tag{3.99}$$

We prove the desired result as follows:

$$\begin{aligned}
&\prod_{j=1}^{J} \left( 1 + \sum_{\ell,k=1}^{J} [M_{\boldsymbol{X}}]_{\ell k} - \frac{\left(\sum_{k=1}^{J} [M_{\boldsymbol{X}}]_{jk}\right)^2}{P} \right) \\
&\leq \left( 1 + \sum_{\ell,k=1}^{J} [M_{\boldsymbol{X}}]_{\ell k} - \sum_{j=1}^{J} \frac{\left(\sum_{k=1}^{J} [M_{\boldsymbol{X}}]_{jk}\right)^2}{JP} \right)^J && (3.100) \\
&\leq \left( 1 + \sum_{\ell,k=1}^{J} [M_{\boldsymbol{X}}]_{\ell k} - \left( \sum_{j,k=1}^{J} \frac{[M_{\boldsymbol{X}}]_{jk}}{J\sqrt{P}} \right)^2 \right)^J && (3.101) \\
&= (1 + P\beta\,(J-\beta))^J && (3.102)
\end{aligned}$$

where (3.100) follows from the arithmetic-geometric mean (AM-GM) which is valid for non-negative real numbers, and (3.101) follows by the Cauchy-Schwarz inequality

$$\underbrace{(1^2 + 1^2 + \cdots + 1^2)}_{J\text{-ones}} \left( \sum_{j=1}^{J} \left( \sum_{k=1}^{J} \frac{[M_{\boldsymbol{X}}]_{jk}}{\sqrt{P}} \right)^2 \right) \geq \left( \sum_{j,k=1}^{J} \frac{[M_{\boldsymbol{X}}]_{jk}}{\sqrt{P}} \right)^2. \tag{3.103}$$

For equality in both (3.100) and (3.101) a sufficient and necessary condition is $\rho_{12} = \rho_{13} = \cdots = \rho_{(J-1)J} = \rho$. We define $\beta = 1 + (J-1)\rho$. To obtain the expression in (3.102) we use the identity $\sum_{\ell,k=1}^{J} [M_{\boldsymbol{X}}]_{\ell k} = PJ\beta$. Consider the matrix $M_{\boldsymbol{X}}$ in (3.43) where the users have the same power and the same correlation coefficient for each pair of users. Then the eigenvalues of $M_{\boldsymbol{X}}$ are $P(1-\rho)$ with multiplicity one and $P(1+(J-1)\rho)$ with multiplicity $J-1$. Since $M_{\boldsymbol{X}} \succeq 0$, we find that $-1/(J-1) \leq \rho \leq 1$ and $0 \leq \beta \leq J$.

### 3.7.7 Convexity

**Lemma 8.** *The problem* $\min\limits_{\beta\in[0,J]:\ell(\beta,J,P)\leq 0}\left\{-\frac{1}{2}\log(1+JP\beta)\right\}$ *is convex, where*

$$\ell(\beta,J,P)=\frac{1}{2}\log\left(1+JP\beta\right)-\frac{J}{2(J-1)}\log\left(1+P\beta\left(J-\beta\right)\right). \tag{3.104}$$

*Proof*: For a fixed $P$ and $J$, the term $-\log(1+JP\beta)$ is convex in $\beta$. We now show that $\ell(\beta,J,P)$ is convex in $\beta$ for a fixed $P$ and $J$. The second derivative of $\ell(\beta,J,P)$ with respect to $\beta$ can be written as

$$\begin{aligned}\frac{\partial^2\ell(\beta,J,P)}{\partial\beta^2}&=\frac{JP}{2(J-1)(1+JP\beta)^2(1+P\beta(J-\beta))^2}\cdot\\&\quad\big(J^2P^3\beta^2(\beta-1)^2+JP(P\beta^2-1)^2+4JP^2\beta^3\\&\quad+2P\beta^2+2J2P^2\beta+2JP\beta+2\big)\geq 0,\end{aligned} \tag{3.105}$$

since RHS of (3.105) has only non-negative terms. □

# 4 On Wyner's Common Information for Gaussians

## 4.1 Introduction

Wyner's Common Information [34] is a measure of dependence between two random variables. Its operational significance lies in network information theory problems (including a canonical information-theoretic model of the problem of coded caching) as well as in distributed simulation of shared randomness. Specifically, for a pair of random variables, Wyner's common information can be described by the search for the most compact third variable that makes the pair conditionally independent. Compactness is measured in terms of the mutual information between the pair and the third variable. The value of Wyner's common information is the minimum of this mutual information. The main difficulty of Wyner's common information is finding the optimal choice for the third variable. Indeed, explicit solutions are known only for a handful of special cases, including the binary symmetric double source and the case of jointly Gaussian random variables.

In the same paper [34, Section 4.2], Wyner also proposes a natural relaxation[1] of his common information, obtained by replacing conditional independence with an upper bound on the conditional mutual information. This relaxation is again directly related to network information theory problems, including the Gray-Wyner source coding network [35]. In the present chapter, we study this relaxation in the special case of jointly Gaussian random variables.

### 4.1.1 Related Work and Contribution

The development of Wyner's common information started with the consideration of a particular network source coding problem, now referred to as the Gray-Wyner network [35]. From this consideration, Wyner extracted the compact form of the common information in [34], initially restricting attention to the case of discrete ran-

[1] The material of this chapter has appeared in
- M. Gastpar and E. Sula, "Relaxed Wyner's common information," in *Proceedings of the 2019 IEEE Information Theory Workshop*, Visby, Sweden, August 2019.
- E. Sula and M. Gastpar, "On Wyner's common information in the Gaussian case," *CoRR*, vol. abs/1912.07083, 2019. [Online]. Available: http://arxiv.org/abs/1912.07083.

dom variables. Extensions to continuous random variables are considered in [36, 37], with a closed-form solution for the Gaussian case. Our work provides an alternative and fundamentally different proof of this same formula (along with a generalization). In the same line of work Wyner's common information is computed in additive Gaussian channels [38]. A local characterization of Wyner's common information is provided in [39], by optimizing over weakly dependent random variables. In [40] Witsenhausen managed to give closed-form formulas for a class of distributions he refers to as "L-shaped." The concept of Wyner's common information has also been extended using other information measures [41]. Other related works include [42, 23]. Wyner's common information has many applications, including to communication networks [34], to caching [43, Section III.C] and to source coding [44].

Other variants of Wyner's common information include [45, 46]. In [45], the conditional independence constraint is replaced by the conditional maximal correlation constraint, whereas in [46], the mutual information objective is replaced by the entropy. The relaxation of Wyner's common information studied in this paper is different from the above variants in the sense that it can be expressed using only mutual information.

The main difficulty in dealing with Wyner's common information is the fact that it is not a convex optimization problem. Specifically, while the objective is convex, the constraint set is not a convex set : taking convex combinations does not respect the constraint of conditional independence. The main contributions of our work concern explicit solutions to this non-convex optimization problem in the special case when the underlying random variables are jointly Gaussian. Our contributions include the following:

1. We establish an alternative and fundamentally different proof of the well-known formula for (standard) Wyner's common information in the Gaussian case, both for scalars and for vectors. Our proof leverages the technique of factorization of convex envelopes [27].

2. In doing so, we establish a more general formula for the Gaussian case of a natural relaxation of Wyner's common information. This relaxation was proposed by Wyner. In it, the constraint of conditional independence is replaced by an upper bound on the conditional mutual information. The quantity is of independent interest, for example establishing a rigorous connection between Canonical Correlation Analysis and Wyner's Common Information [47].

## 4.2 Preliminaries

### 4.2.1 Wyner's Common Information

Wyner's common information is defined for two random variables $X$ and $Y$ of arbitrary fixed joint distribution $p(x, y)$.

**Definition 3.** *For random variables $X$ and $Y$ with joint distribution $p(x, y)$, Wyner's common information is defined as*

$$C(X;Y) = \inf_{p(w|x,y)} I(X,Y;W) \text{ such that } I(X;Y|W) = 0. \tag{4.1}$$

Wyner's common information satisfies a number of interesting properties. We state some of them below in Lemmas 9 and 10 for a generalized definition given in Definition 4.

We note that explicit formulas for Wyner's common information are known only for a small number of special cases. The case of the doubly symmetric binary source is solved completely in [34] and can be written as

$$C(X;Y) = 1 + h_b(a_0) - 2h_b\left(\frac{1-\sqrt{1-2a_0}}{2}\right), \tag{4.2}$$

where $a_0$ denotes the probability that the two sources are unequal (assuming without loss of generality $a_0 \leq \frac{1}{2}$). In this case, the optimizing $W$ in Equation (4.1) can be chosen to be binary. Further special cases of discrete-alphabet sources appear in [40].

Moreover, when $X$ and $Y$ are jointly Gaussian with correlation coefficient $\rho$, then $C(X;Y) = \frac{1}{2}\log\frac{1+|\rho|}{1-|\rho|}$. Note that for this example, $I(X;Y) = \frac{1}{2}\log\frac{1}{1-\rho^2}$. This case was solved in [36, 37] using a parameterization of conditionally independent distributions. We note that an alternative proof follows from our arguments below.

### 4.2.2 A Natural Relaxation of Wyner's Common Information

Wyner, in [34, Section 4.2], defines an auxiliary quantity $\Gamma(\delta_1, \delta_2)$. Starting from this definition, it is natural to introduce the following quantity:

**Definition 4.** *For jointly continuous random variables $X$ and $Y$ with joint distribution $p(x,y)$, we define*

$$C_\gamma(X;Y) = \inf_{p(w|x,y)} I(X,Y;W) \text{ such that } I(X;Y|W) \leq \gamma. \tag{4.3}$$

With respect to [34, Section 4.2], we have that $C_\gamma(X;Y) = H(X,Y) - \Gamma(0,\gamma)$. Comparing Definitions 3 and 4, we see that in $C_\gamma(X;Y)$, the constraint of conditional independence is relaxed into an upper bound on the conditional mutual information. Specifically, for $\gamma = 0$, we have $C_0(X;Y) = C(X;Y)$, the regular Wyner's common information. In this sense, it is tempting to refer to $C_\gamma(X;Y)$ as *relaxed* Wyner's common information. The following lemma summarizes some basic properties.

**Lemma 9.** *$C_\gamma(X;Y)$ satisfies the following basic properties:*

1. *$C_\gamma(X;Y) \geq \max\{I(X;Y) - \gamma, 0\}$.*
2. *Data processing inequality: If $X-Y-Z$ form a Markov chain, then $C_\gamma(X;Z) \leq \min\{C_\gamma(X;Y), C_\gamma(Y;Z)\}$.*
3. *$C_\gamma(X;Y)$ is a convex and continuous function of $\gamma$ for $\gamma \geq 0$.*
4. *If $Z - X - Y$ forms a Markov chain, then $C_\gamma((X,Z);Y) = C_\gamma(X;Y)$.*
5. *The cardinality of $\mathcal{W}$ may be restricted to $|\mathcal{W}| \leq |\mathcal{X}||\mathcal{Y}| + 1$.*
6. *If $f(\cdot)$ and $g(\cdot)$ are one-to-one functions, then $C_\gamma(f(X);g(Y)) = C_\gamma(X;Y)$.*
7. *For discrete $X$, we have $C_\gamma(X;X) = \max\{H(X) - \gamma, 0\}$.*

Proofs are given in Appendix 4.6.1.

A further property of $C_\gamma(X;Y)$ is a tensorization result for independent pairs, which we will use below to solve the case of the Gaussian vector source.

**Lemma 10** (Tensorization)**.** *Let $\{(X_i,Y_i)\}_{i=1}^n$ be $n$ independent pairs of random variables. Then*

$$C_\gamma(X^n;Y^n) = \min_{\{\gamma_i\}_{i=1}^n : \sum_{i=1}^n \gamma_i = \gamma} \sum_{i=1}^n C_{\gamma_i}(X_i;Y_i). \tag{4.4}$$

The proof is given in Appendix 4.6.2. The lemma has an intuitive interpretation in $\mathbb{R}^2$ plane. If we express $C_\gamma(X^n;Y^n)$ as a region in $\mathbb{R}^2$, which is determined by $(\gamma, C_\gamma(X^n;Y^n))$, then the computation of $(\gamma, C_\gamma(X^n;Y^n))$ is simply the Minkowski sum of the individual regions which are determined by $(\gamma_i, C_{\gamma_i}(X_i;Y_i))$.

## 4.3 The Scalar Gaussian Case

One of the main technical contributions of this work is a closed-form formula for $C_\gamma(X;Y)$ in the case where $X$ and $Y$ are jointly Gaussian.

**Theorem 8.** *When $X$ and $Y$ are jointly Gaussian with correlation coefficient $\rho$, then*

$$C_\gamma(X;Y) = \frac{1}{2}\log^+\left(\frac{1+|\rho|}{1-|\rho|}\cdot\frac{1-\sqrt{1-e^{-2\gamma}}}{1+\sqrt{1-e^{-2\gamma}}}\right). \tag{4.5}$$

The proof is given below in Section 4.3.2.

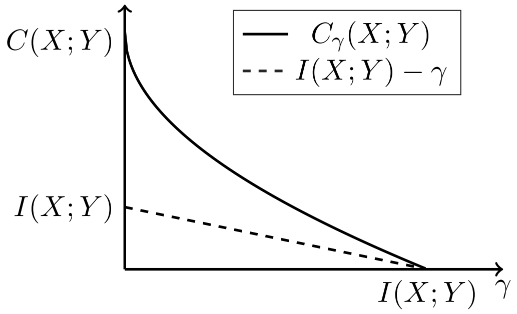


Figure 4.1 – $C_\gamma(X;Y)$ for jointly Gaussian $X$ and $Y$ for the case $\rho = 1/2$, thus, we have $C(X;Y) = \log\sqrt{3}$ and $I(X;Y) = \log(2/\sqrt{3})$. The dashed line is the lower bound from Lemma 9, Item 1).

Furthermore, in Figure 4.2 we exploit the relaxed Wyner's common information from Theorem 8, by comparing the curves for different values of $\gamma$, versus $\rho$, that is the correlation coefficient of the Gaussian random variable $X$ and $Y$.

### 4.3.1 Preliminary Results for the Proof of Theorem 8

The following results are used as intermediate tools in the proof of the main results.

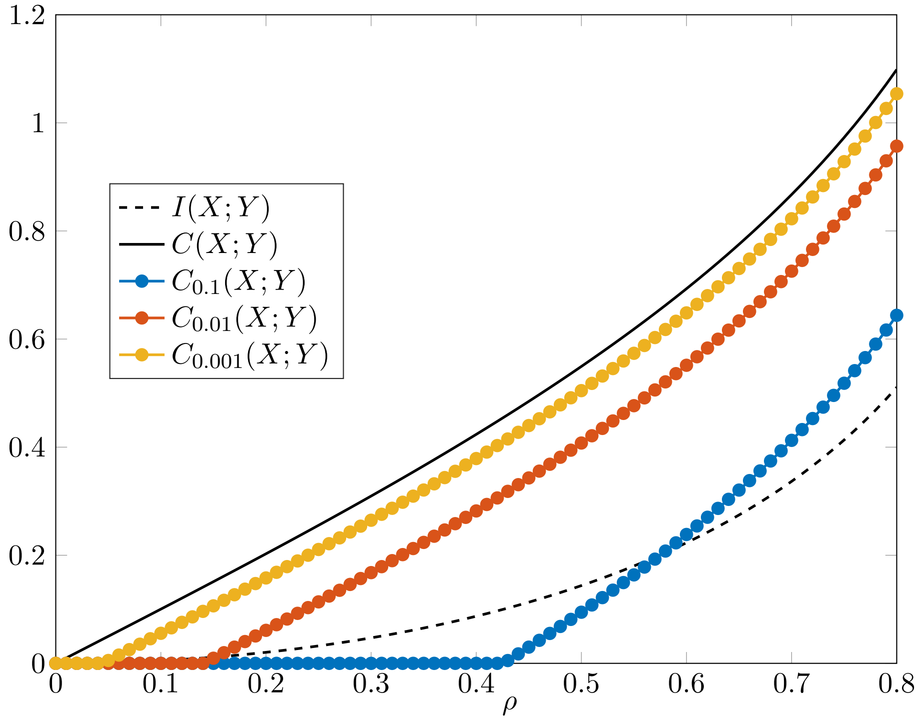


Figure 4.2 – Comparison of mutual information $I(X;Y)$, Wyner's common information $C(X;Y)$ and relaxed Wyner's common information $C_\gamma(X;Y)$ for different values of $\gamma$.

**Theorem 9.** *For $K \succeq 0$, $0 < \lambda < 1$, there exists a $0 \preceq K' \preceq K$ and $(X', Y') \sim \mathcal{N}(0, K')$ such that $(X,Y) \sim p_{X,Y}$ with covariance matrix $K$ the following inequality holds:*

$$\inf_W h(Y|W) + h(X|W) - (1+\lambda)h(X,Y|W) \geq h(Y') + h(X') - (1+\lambda)h(X',Y'). \tag{4.6}$$

*Proof.* The theorem is a consequence of [48, Theorem 2], for a specific choice of $p = \frac{1}{\lambda} + 1$. An extended proof is given in Appendix 4.6.3. □

To leverage Theorem 9, we need to understand the covariance matrix $K'$. In [48], the right hand side in Equation (4.6) is further lower bounded as $h(Y) + h(X) - (1+\lambda)h(X,Y)$, where $(X,Y) \sim \mathcal{N}(0,K)$ (correlation coefficient of matrix $K$ is $\rho$ and the diagonal entries are unit), which holds for $\lambda < \rho$. This choice establishes the hypercontractivity bound $(1+\rho)I(W;X,Y) \geq I(W;X) + I(W;Y)$ (for jointly Gaussian $X, Y$ and any $W$). Unfortunately, for the problem of Wyner's common information, this leads to a loose lower bound, which can be seen as follows:

$$C_{\gamma=0}(X;Y) = \inf_{p(w|x,y):I(X;Y|W)=0} I(X,Y;W) \tag{4.7}$$

$$= \inf_{p(w|x,y):I(X,Y;W)+I(X;Y)-I(W;X)-I(W;Y)=0} I(X,Y;W) \tag{4.8}$$

$$\geq \inf_{p(w|x,y):I(X;Y)-\rho I(X,Y;W)\leq 0} I(X,Y;W) \tag{4.9}$$

$$= \inf_{p(w|x,y):I(X,Y;W)\geq \frac{I(X;Y)}{\rho}} I(X,Y;W) \tag{4.10}$$

$$= \frac{I(X;Y)}{\rho} = \frac{1}{2}\frac{\log \frac{1}{1-\rho^2}}{\rho}, \tag{4.11}$$

where (4.9) follows from $(1+\rho)I(W;X,Y) \geq I(W;X) + I(W;Y)$.

We now show that by a different lower bound on the right hand side in Equation (4.6), we can indeed get a tight lower bound for the problem of Wyner's common information as well as its relaxation $C_\gamma(X;Y)$. Specifically, we have the following lower bound:

**Lemma 11.** *For $(X', Y') \sim \mathcal{N}(0, K')$, the following inequality holds*

$$\min_{K':0\preceq K'\preceq \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}} h(X') + h(Y') - (1+\lambda)h(X',Y')$$

$$\geq \frac{1}{2}\log\frac{1}{1-\lambda^2} - \frac{\lambda}{2}\log(2\pi e)^2\frac{(1-\rho)^2(1+\lambda)}{1-\lambda}, \tag{4.12}$$

*where $\lambda \leq \rho$.*

*Proof.* The proof is given in Appendix 4.6.6. □

### 4.3.2 Proof of Theorem 8

The proof of the converse for Theorem 8 involves two main steps. In this section, we prove that one optimal distribution is jointly Gaussian via a variant of the factorization of convex envelope. Then, we tackle the resulting optimization problem with Lagrange duality. Let us start form the lower bound first.

**Lemma 12.** *When $X$ and $Y$ are jointly Gaussian with correlation coefficient $\rho$ and unit variance, then $C_\gamma(X;Y) \geq \frac{1}{2}\log^+\left(\frac{1+|\rho|}{1-|\rho|}\cdot\frac{1-\sqrt{1-e^{-2\gamma}}}{1+\sqrt{1-e^{-2\gamma}}}\right)$.*

*Proof.* The lower bound is derived in the following lines

$$C_\gamma(X;Y) = \inf_{W:I(X;Y|W)\leq\gamma} I(X,Y;W) \tag{4.13}$$

$$\geq \inf_W (1+\mu)I(X,Y;W) - \mu I(X;W) - \mu I(Y;W) + \mu I(X;Y) - \mu\gamma \tag{4.14}$$

$$= h(X,Y) - \mu\gamma + \mu \inf_W h(X|W) + h(Y|W) - (1+\frac{1}{\mu})h(X,Y|W) \tag{4.15}$$

$$\geq h(X,Y) - \mu\gamma + \mu \min_{K':0\preceq K'\preceq \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}} h(X') + h(Y') - (1+\frac{1}{\mu})h(X',Y') \tag{4.16}$$

$$\geq \frac{1}{2}\log(2\pi e)^2(1-\rho^2) - \mu\gamma + \frac{\mu}{2}\log\frac{\mu^2}{\mu^2-1} - \frac{1}{2}\log(2\pi e)^2\frac{(1-\rho)^2(\mu+1)}{\mu-1} \tag{4.17}$$

$$\geq \log^+\left(\frac{1+|\rho|}{1-|\rho|}\cdot\frac{1-\sqrt{1-e^{-2\gamma}}}{1+\sqrt{1-e^{-2\gamma}}}\right) \tag{4.18}$$

where (4.14), is a bound for all $\mu \geq 0$; (4.16) follows from Theorem 9 where $(X', Y') \sim \mathcal{N}(0, K')$, $\mu := \frac{1}{\lambda}$ and for the assumption $0 < \lambda < 1$ to be satisfied we need $\mu > 1$; (4.17) follows from Lemma 11 for $\mu \geq \frac{1}{\rho}$ and (4.18) follows by maximizing the function

$$g(\mu) := \frac{1}{2} \log (2\pi e)^2 (1-\rho^2) - \mu\gamma + \frac{\mu}{2} \log \frac{\mu^2}{\mu^2 - 1} - \frac{1}{2} \log (2\pi e)^2 \frac{(1-\rho)^2(\mu+1)}{\mu - 1}, \tag{4.19}$$

for $\mu \geq \frac{1}{\rho}$. Now we need to choose the tightest bound where $\mu \geq \frac{1}{\rho}$, which is $\max_{\mu \geq \frac{1}{\rho}} g(\mu)$ and function $g$ is concave in $\mu$,

$$\frac{\partial^2 g}{\partial \mu^2} = -\frac{1}{\mu(\mu^2 - 1)} < 0. \tag{4.20}$$

By studying the monotonicity we obtain

$$\frac{\partial g}{\partial \mu} = -\frac{1}{2} \log \frac{\mu^2 - 1}{\mu^2} - \gamma, \tag{4.21}$$

since the function is concave the maximum has to be when the derivative vanishes which leads to the optimal solution $\mu_* = \frac{1}{\sqrt{1-e^{-2\gamma}}}$, where $\mu_* \geq \frac{1}{\rho}$. Substituting for the optimal $\mu_*$ we obtain

$$C_\gamma(X;Y) \geq g\left(\frac{1}{\sqrt{1-e^{-2\gamma}}}\right) = \frac{1}{2} \log^+ \left(\frac{1+\rho}{1-\rho} \cdot \frac{1-\sqrt{1-e^{-2\gamma}}}{1+\sqrt{1-e^{-2\gamma}}}\right). \tag{4.22}$$

□

Now let us move the attention to the upper bound. Let us assume (without loss of generality) that $X$ and $Y$ have unit variance and are non-negatively correlated with correlation coefficient $\rho \geq 0$. Since they are jointly Gaussian, we can express them as

$$X = \sigma W + \sqrt{1-\sigma^2} N_X \tag{4.23}$$

$$Y = \sigma W + \sqrt{1-\sigma^2} N_Y, \tag{4.24}$$

where $W, N_X, N_Y$ are jointly Gaussian, and where $W \sim \mathcal{N}(0,1)$ is independent of $(N_X, N_Y)$. Letting the covariance of the vector $(N_X, N_Y)$ be

$$K_{(N_X, N_Y)} = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \tag{4.25}$$

for some $0 \leq \alpha \leq \rho$, we find that we need to choose $\sigma^2 = \frac{\rho - \alpha}{1-\alpha}$. Specifically, let us select $\alpha = \sqrt{1-e^{-2\gamma}}$, for some $0 \leq \gamma \leq \frac{1}{2} \log \frac{1}{1-\rho^2}$. For this choice, we find $I(X;Y|W) = \gamma$ and

$$I(X,Y;W) = \frac{1}{2} \log \frac{(1+\rho)(1-\alpha)}{(1-\rho)(1+\alpha)}. \tag{4.26}$$

## 4.4 The Vector Gaussian Case

In this section, we consider the case where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian random vectors. The key observation is that in this case, there exist invertible matrices $A$ and $B$ such that $A\boldsymbol{X}$ and $B\boldsymbol{Y}$ are vectors of independent pairs, exactly like in Lemma 10. Therefore, we can use Lemma 10 to give an explicit formula for the relaxed Wyner's common information between *arbitrarily correlated* jointly Gaussian random vectors, as stated in the following theorem.

**Theorem 10.** *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be jointly Gaussian random vectors of length $n$ and covariance matrix $K_{(\boldsymbol{X},\boldsymbol{Y})}$. Then,*

$$C_\gamma(\boldsymbol{X};\boldsymbol{Y}) = \min_{\gamma_i:\sum_{i=1}^n \gamma_i=\gamma} \sum_{i=1}^n C_{\gamma_i}(X_i;Y_i), \tag{4.27}$$

*where*

$$C_{\gamma_i}(X_i;Y_i) = \frac{1}{2}\log^+ \frac{(1+\rho_i)(1-\sqrt{1-e^{-2\gamma_i}})}{(1-\rho_i)(1+\sqrt{1-e^{-2\gamma_i}})} \tag{4.28}$$

*and $\rho_i$ (for $i = 1,\ldots,n$) are the singular values of $K_{\boldsymbol{X}}^{-1/2}K_{\boldsymbol{X}\boldsymbol{Y}}K_{\boldsymbol{Y}}^{-1/2}$, where $K_{\boldsymbol{X}}^{-1/2}$ and $K_{\boldsymbol{Y}}^{-1/2}$ are defined to mean that only the positive eigenvalues are inverted.*

**Remark 3.** *Note that we do not assume that $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ are of full rank. Moreover, note that the case where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are of unequallength is included: Simply invoke Lemma 9, Item 4), to append the shorter vector with independent Gaussians so as to end up with two vectors of the same length.*

*Proof.* Note that the mean is irrelevant for the problem at hand, so we assume it to be zero without loss of generality. The first step of the proof is to apply the same transform used, e.g., in [44]. Namely, we form $\hat{\boldsymbol{X}} = K_{\boldsymbol{X}}^{-1/2}\boldsymbol{X}$ and $\hat{\boldsymbol{Y}} = K_{\boldsymbol{Y}}^{-1/2}\boldsymbol{Y}$, where $K_{\boldsymbol{X}}^{-1/2}$ and $K_{\boldsymbol{Y}}^{-1/2}$ are defined to mean that only the positive eigenvalues are inverted. Let us denote the rank of $K_{\boldsymbol{X}}$ by $r_X$ and the rank of $K_{\boldsymbol{Y}}$ by $r_Y$. Then, we have

$$K_{\hat{\boldsymbol{X}}} = \begin{pmatrix} I_{r_X} & 0 \\ 0 & 0_{n-r_X} \end{pmatrix} \tag{4.29}$$

and

$$K_{\hat{\boldsymbol{Y}}} = \begin{pmatrix} I_{r_Y} & 0 \\ 0 & 0_{n-r_Y} \end{pmatrix} \tag{4.30}$$

Moreover, we have $K_{\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}} = K_{\boldsymbol{X}}^{-1/2}K_{\boldsymbol{X}\boldsymbol{Y}}K_{\boldsymbol{Y}}^{-1/2}$. Let us denote the singular value decomposition of this matrix by $K_{\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}} = R_{\boldsymbol{X}}\Lambda R_{\boldsymbol{Y}}$. Define $\tilde{\boldsymbol{X}} = R_{\boldsymbol{X}}^T\hat{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}} = R_{\boldsymbol{Y}}\hat{\boldsymbol{Y}}$, which implies that $K_{\tilde{\boldsymbol{X}}} = K_{\hat{\boldsymbol{X}}}$, $K_{\tilde{\boldsymbol{Y}}} = K_{\hat{\boldsymbol{Y}}}$, and $K_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}}} = \Lambda$. The second step of the proof is to observe that the mappings from $\boldsymbol{X}$ to $\tilde{\boldsymbol{X}}$ and from $\boldsymbol{Y}$ to $\tilde{\boldsymbol{Y}}$, respectively, are linear one-to-one and mutual information is preserved under such transformation. Hence, we have $C_\gamma(\boldsymbol{X};\boldsymbol{Y}) = C_\gamma(\tilde{\boldsymbol{X}};\tilde{\boldsymbol{Y}})$. The third, and key, step of the proof is now to observe that $\{(X_i,Y_i)\}_{i=1}^n$ are $n$ independent pairs of random variables. Hence, we can apply Lemma 10. The final step is to apply Theorem 8 separately to each of the independent pairs, thus establishing the claimed formula. □

In the remainder of this section, we explore the structure of the allocation problem in Theorem 10, that is, the problem of optimally choosing the values of $\gamma_i$. As we will show, the answer is of the water-filling type. That is, there is a "water level" $\gamma^*$. Then, all $\gamma_i$ whose corresponding correlation coefficient $\rho_i$ is large enough will be set equal to $\gamma^*$. The remaining $\gamma_i$, corresponding to those $i$ with low correlation coefficient $\rho_i$, will be set to their respective maximal values (all of which are smaller than $\gamma^*$). To establish this result, we prefer to change notation as follows. We define $\alpha_i = \sqrt{1-e^{-2\gamma_i}}$. With this, we can express the allocation problem in Theorem 10 as

$$C_\gamma(\boldsymbol{X};\boldsymbol{Y}) = \min_{\alpha_1,\alpha_2,\cdots,\alpha_n} \sum_{i=1}^n \frac{1}{2}\log^+ \frac{(1+\rho_i)(1-\alpha_i)}{(1-\rho_i)(1+\alpha_i)} \text{ such that } \sum_{i=1}^n \frac{1}{2}\log\frac{1}{1-\alpha_i^2} \le \gamma. \tag{4.31}$$

Moreover, defining

$$C(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}, \quad I(\rho) = \frac{1}{2}\log\frac{1}{1-\rho^2}, \tag{4.32}$$

we can rewrite Equation (4.31) as

$$C_\gamma(\boldsymbol{X};\boldsymbol{Y}) = \min_{\alpha_1,\alpha_2,\cdots,\alpha_n} \sum_{i=1}^n (C(\rho_i)-C(\alpha_i))^+ \text{ such that } \sum_{i=1}^n I(\alpha_i) \le \gamma. \tag{4.33}$$

**Theorem 11.** *The solution to the allocation problem of Theorem 10 can be expressed as*

$$C_\gamma(\boldsymbol{X};\boldsymbol{Y}) = \sum_{i=1}^n (C(\rho_i)-\beta^*)^+, \tag{4.34}$$

*where $\beta^*$ is selected such that*

$$\sum_{i=1}^n \min\{f(\beta^*), I(\rho_i)\} = \gamma, \tag{4.35}$$

*where*

$$f(\beta^*) = \frac{1}{2}\log\frac{(\exp(2\beta^*)+1)^2}{4\exp(2\beta^*)}. \tag{4.36}$$

*Proof of Theorem 11.* Note that (4.33) can be rewritten as

$$C_\gamma(\boldsymbol{X};\boldsymbol{Y}) = \min_{\gamma_1,\gamma_2,\cdots,\gamma_n} \sum_{i=1}^n \left(C(\rho_i)-C(I^{-1}(\gamma_i))\right)^+ \text{ such that } \sum_{i=1}^n \gamma_i \le \gamma, \tag{4.37}$$

and thus, for notational compactness, let us define

$$g(x) = C(I^{-1}(x)) = \frac{1}{2}\log\frac{1+\sqrt{1-e^{-2x}}}{1-\sqrt{1-e^{-2x}}}, \tag{4.38}$$

which is a strictly concave, strictly increasing function. We also define its inverse,

$$f(x) = g^{-1}(x) = I(C^{-1}(x)) = \frac{1}{2}\log\frac{1}{1-\left(\frac{\exp(2x)-1}{\exp(2x)+1}\right)^2} = \frac{1}{2}\log\frac{(\exp(2x)+1)^2}{4\exp(2x)}, \tag{4.39}$$

which is a strictly convex, strictly increasing function.

Without loss of generality, suppose that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_n$. The objective function is composed of $n$ terms which can be active or not, meaning that they can be either positive or zero. Since the function $C(\rho)$ is increasing in $\rho$, we have that $C(\rho_1) \geq C(\rho_2) \geq \cdots \geq C(\rho_n)$. To summarize the intuition of the proof, note that the $n$-th term, *i.e.,* $(C(\rho_n) - g(\gamma_n))^+$, will be inactive first. Therefore, by increasing $\gamma$ then the terms will become inactive in a decreasing fashion until we are left with only the first term active and the rest inactive.

Let us start with the case when all the terms are active, which implies that $\sum_{i=1}^n (C(\rho_i) - g(\gamma_i))^+ = \sum_{i=1}^n (C(\rho_i) - g(\gamma_i))$ Then, by the concavity of $g(\gamma_i)$, we have

$$\sum_{i=1}^n g(\gamma_i) \leq ng(\frac{\gamma}{n}), \tag{4.40}$$

thus an optimal choice is $\gamma^* = \frac{\gamma}{n}$, for all $i$. Hence, in our notation, in this case $\beta^* = g(\frac{\gamma}{n})$. Clearly, all the terms are active in the interval $0 \leq \gamma \leq nI(\rho_n)$, with the reasoning that if the $n$-th terms is active then the rest of the terms is active too. Next, consider the case when the $n$-th term is inactive and the rest is active. Therefore, $\sum_{i=1}^n (C(\rho_i) - g(\gamma_i))^+ = \sum_{i=1}^{n-1} (C(\rho_i) - g(\gamma_i))$ and by the concavity of $g(\gamma_i)$, we have

$$\sum_{i=1}^{n-1} g(\gamma_i) \leq (n-1)g\left(\frac{\gamma}{n-1}\right), \tag{4.41}$$

thus an optimal choice is $\gamma^* = \frac{\gamma-\gamma_n}{n-1}$, for all $i \in \{1, 2, \cdots, n-1\}$. The optimal choice for $\gamma_n$ is $\gamma_n = I(\rho_n)$, which makes the $n$-th term exactly zero. This scenario will happen in the interval, $nI(\rho_n) < \gamma \leq I(\rho_n) + (n-1)I(\rho_{n-1})$. Instead, the corresponding $\beta^*$ in our notation is $\beta^* = g\left(\frac{\gamma - I(\rho_n)}{n-1}\right)$. In general, let us consider the case when $k$-th term is active and $k+1$-th is inactive. By a similar argument as above, the optimal choice is $\gamma^* = \frac{\gamma - \sum_{i=k+1}^n \gamma_i}{n-k}$ for $i \in \{1, 2, \cdots, k\}$ and $\gamma_i = I(\rho_i)$ for $i \in \{k+1, \cdots, n\}$. This scenario will happen in the interval $(k+1)I(\rho_{k+1}) + \sum_{i=k+2}^n I(\rho_i) < \gamma \leq kI(\rho_k) + \sum_{i=k+1}^n I(\rho_i)$. Importantly, observe that the optimal $\gamma_i$ can be rewritten as $\gamma_i = \min\{I(\rho_i), \gamma^*\}$, therefore the solution to the allocation problem can be expressed as

$$C_\gamma(\boldsymbol{X}; \boldsymbol{Y}) = \sum_{i=1}^n (C(\rho_i) - g(\gamma^*))^+, \tag{4.42}$$

where $\gamma^*$ is selected such that

$$\sum_{i=1}^n \min\{\gamma^*, I(\rho_i)\} = \gamma. \tag{4.43}$$

The solution to the allocation problem can be rewritten as

$$C_\gamma(\boldsymbol{X};\boldsymbol{Y}) = \sum_{i=1}^{n} \left(C(\rho_i) - \beta^*\right)^+ , \tag{4.44}$$

where $\beta^*$ is selected such that

$$\sum_{i=1}^{n} \min\left\{f(\beta^*), I(\rho_i)\right\} = \gamma. \tag{4.45}$$

□

Theorem 11 shows that the allocation problem has a natural reverse water-filling interpretation which can be visualized in two dual ways. First, we could consider the space of the $\gamma_i$ parameters, which leads to Figure 4.3: None of the $\gamma_i$ should be selected larger than the corresponding $I(\rho_i)$, and those $\gamma_i$ that are strictly smaller than their maximum value should all be equal. This graphically identifies the optimal value $\gamma^*$, and thus, the resulting solution to our optimization problem. Alternatively, we could consider directly the space of the individual contributions to the objective, denoted by $C(\rho_i)$ in Equation (4.37), which leads to Figure 4.4.

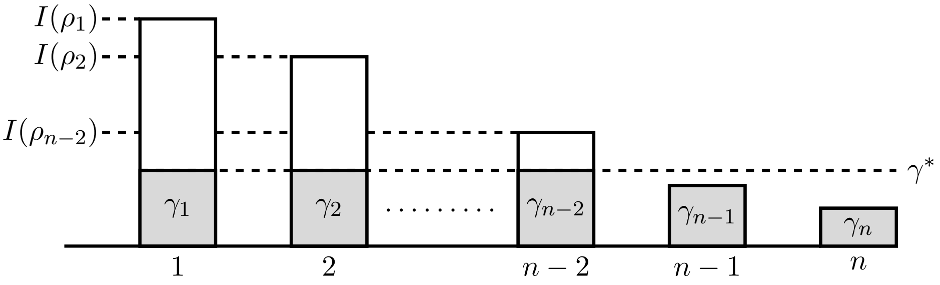


Figure 4.3 – Example of reverse water-filling. The (whole) bars represent the $\gamma_i$-s which make $C_{\gamma_i}(X_i;Y_i) = 0$, and the shaded area of the bars is the proper allocation $\gamma_i$ to minimize the original problem. In this example, $\gamma = \sum_{i=1}^{n} \gamma_i$ is chosen such that $C_{\gamma_{n-1}}(X_{n-1};Y_{n-1}) = C_{\gamma_n}(X_n;Y_n) = 0$.

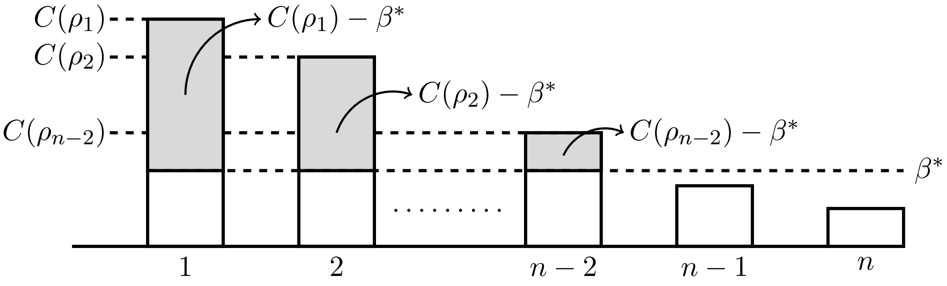


Figure 4.4 – Example of reverse water-filling. The (whole) bars represent the (standard) Wyner's common information of each individual pair, respectively. The shaded area of the bars is the respective contribution to $C_\gamma(\boldsymbol{X};\boldsymbol{Y})$. In this example, $\gamma$ is chosen such that $(C(\rho_{n-1}) - \beta^*)^+ = (C(\rho_n) - \beta^*)^+ = 0$.

## 4.5 Conclusion

We studied a natural relaxation of Wyner's common information, whereby the constraint of conditional independence is replaced by an upper bound on the conditional mutual information. This leads to a novel and different optimization problem. We established a number of properties of this novel quantity, including a chain rule type formula for the case of independent pairs of random variables. For the case of jointly Gaussian sources, both scalar and vector, we presented a closed-form expression for the relaxed Wyner's common information.

## 4.6 Appendix

### 4.6.1 Proof of Lemma 9

For Item 1), the inequality follows from the fact that mutual information is non-negative. If $\gamma \geq I(X;Y)$, we may select $W$ to be a constant, thus we have equality to zero. If $\gamma < I(X;Y)$, then the lower bound proved in the next item establishes that we cannot have equality to zero. Also, observe that the Lagrangian for the relaxed Wyner's common information problem of Equation (4.3) is $L(\lambda, p(w|x,y)) = I(X,Y;W) + \lambda(I(X;Y|W) - \gamma)$. From Lagrange duality, we thus have the lower bound $C_\gamma(X;Y) \geq \inf_{p(w|x,y)} L(\lambda, p(w|x,y))$, for all positive $\lambda$. Setting $\lambda = 1$, we have $\inf_{p(w|x,y)}(I(X,Y;W) + I(X;Y|W) - \gamma) = \inf_{p(w|x,y)}(I(X;Y) + I(X;W|Y) + I(Y;W|X) - \gamma) = I(X;Y) - \gamma$. For Item 2), observe that for fixed $p(x,y,z)$, we can write

$$C_\gamma(X;Y) = \inf_{p(x,y,z)p(w|x,y):I(X;Y|W)\leq\gamma} I(X,Y;W) \tag{4.46}$$

$$\geq \inf_{p(x,y,z)p(w|x,y):I(X;Y|W)\leq\gamma} I(X,Z;W), \tag{4.47}$$

due to the Markov chain $(X,Z) - (X,Y) - W$. Moreover, note that since we consider only joint distributions of the form $p(x,y)p(z|y)p(w|x,y)$, we also have the Markov chain $(X,W) - Y - Z$, which implies the Markov chain $X - (W,Y) - Z$. The latter implies $I(X;Y|W) \geq I(X;Z|W)$. Hence,

$$C_\gamma(X;Y) \geq \inf_{p(x,y,z)p(w|x,y):I(X;Z|W)\leq\gamma} I(X,Z;W) \geq C_\gamma(X;Z). \tag{4.48}$$

By the same token, $C_\gamma(Y;Z) \geq C_\gamma(X;Z)$, which completes the proof. Item 3) follows directly from [34, Corollary 4.5]. For Item 4), on the one hand, we have

$$C_\gamma((X,Z);Y) = \inf_{p(w|x,y,z):I(X,Z;Y|W)\leq\gamma} I(X,Z,Y;W) \tag{4.49}$$

$$\leq \inf_{p(w|x,y):I(X;Y|W)+I(Z;Y|W,X)\leq\gamma} I(X,Y;W) + I(Z;W|X,Y) \tag{4.50}$$

$$= C_\gamma(X;Y) \tag{4.51}$$

where in Equation (4.50) we add the constraint that conditioned on $(X,Y)$, $W$ is selected to be *independent* of $Z$, which cannot reduce the value of the infimum. That is, for such a choice of $W$, we have the Markov chain $Z - (X,Y) - W$, thus $I(Z;W|X,Y) = 0$. Furthermore, observe that the factorization $p(x,y,z,w) =$

$p(x,y)p(z|x)p(w|x,y)$ also implies the factorization $p(x,y,z,w) = p(x,w)p(z|x)p(y|w,x)$. Hence, we also have the Markov chain $Z-(W,X)-Y$, thus $I(Z;Y|W,X) = 0$, which thus establishes the last step. Conversely, observe that

$$\begin{aligned} C_\gamma((X,Z);Y) &= \inf_{p(w|x,y,z):I(X,Z;Y|W)\leq\gamma} I(X,Y,Z;W) && (4.52)\\ &\geq \inf_{p(w|x,y):I(X;Y|W)\leq\gamma} I(X,Y;W) + \inf_{p(w|x,y,z):I(X,Z;Y|W)\leq\gamma} I(Z;W|X,Y) && (4.53)\\ &\geq C_\gamma(X;Y) && (4.54) \end{aligned}$$

where (4.53) follows from the fact that the infimum of the sum is lower bounded by the sum of the infimums and the fact that relaxing constraints cannot increase the value of the infimum, and (4.54) follows from non-negativity of the second term.

Item 5) is a standard cardinality bound, following from the arguments in [49]. For the context at hand, see also Theorem 1 in [43, p.6396]. Item 6) follows because all involved mutual information terms are invariant to one-to-one transforms. For Item 7), note that we can express $C_\gamma(X;X) = H(X) - \max_{p(w|x):H(X|W)\leq\gamma} H(X|W)$, which directly gives the result.

### 4.6.2 Proof of Lemma 10

The achievability part, that is, the inequality

$$C_\gamma(X^n;Y^n) \leq \min_{\{\gamma_i\}_{i=1}^n:\sum_{i=1}^n \gamma_i=\gamma} \sum_{i=1}^n C_{\gamma_i}(X_i;Y_i), \qquad (4.55)$$

merely corresponds to a particular choice of $W$ in the definition given in Equation (4.3). Specifically, let $W = (W_1, W_2, \ldots, W_n)$, and choose $\{(X_i, Y_i, W_i)\}_{i=1}^n$ to be $n$ independent triples of random vectors. The converse is more subtle. We prove the case $n=2$ first, followed by induction. For $n=2$, we have

$$\begin{aligned} &\inf_{p(w|x_1,x_2,y_1,y_2):I(X_1,X_2;Y_1,Y_2|W)\leq\gamma} I(X_1,X_2,Y_1,Y_2;W)\\ &\geq \inf_{p(w|x_1,x_2,y_1,y_2):I(X_1;Y_1|W)+I(X_2;Y_2|W,X_1)\leq\gamma} I(X_1,Y_1;W) + I(X_2,Y_2;W,X_1) && (4.56)\\ &= \min_{\gamma_1+\gamma_2=\gamma} \left\{ \inf_{\substack{p(w|x_1,x_2,y_1,y_2):I(X_1;Y_1|W)\leq\gamma_1,\\ I(X_2;Y_2|W,X_1)\leq\gamma_2}} I(X_1,Y_1;W) + I(X_2,Y_2;W,X_1) \right\} && (4.57)\\ &\geq \min_{\gamma_1+\gamma_2=\gamma} \left\{ \inf_{p(w|x_1,x_2,y_1,y_2):I(X_1;Y_1|W)\leq\gamma_1,I(X_2;Y_2|W,X_1)\leq\gamma_2} I(X_1,Y_1;W) \right.\\ &\quad \left. + \inf_{p(\tilde{w}|x_1,x_2,y_1,y_2):I(X_1;Y_1|\tilde{W})\leq\gamma_1,I(X_2;Y_2|\tilde{W},X_1)\leq\gamma_2} I(X_2,Y_2;\tilde{W},X_1) \right\} && (4.58)\\ &\geq \min_{\gamma_1+\gamma_2=\gamma} \left\{ \inf_{p(w|x_1,y_1):I(X_1;Y_1|W)\leq\gamma_1} I(X_1,Y_1;W) \right.\\ &\quad \left. + \inf_{p(\tilde{w}|x_1,x_2,y_2):I(X_2;Y_2|\tilde{W},X_1)\leq\gamma_2} I(X_2,Y_2;\tilde{W},X_1) \right\} && (4.59) \end{aligned}$$

$$\geq \min_{\gamma_1+\gamma_2=\gamma} \left\{ \inf_{p(w_1|x_1,y_1):I(X_1;Y_1|W_1)\leq\gamma_1} I(X_1,Y_1;W_1) \right.$$
$$\left. + \inf_{p(\tilde{w},\tilde{x}_1|x_2,y_2):I(X_2;Y_2|\tilde{W},\tilde{X}_1)\leq\gamma_2} I(X_2,Y_2;\tilde{W},\tilde{X}_1) \right\} \tag{4.60}$$

where (4.56) follows from

$$I(X_1,X_2,Y_1,Y_2;W) = I(X_1,Y_1;W) + I(X_2,Y_2;W|X_1,Y_1) + I(X_1,Y_1;X_2,Y_2) \tag{4.61}$$
$$= I(X_1,Y_1;W) + I(X_2,Y_2;W,X_1,Y_1) \tag{4.62}$$
$$\geq I(X_1,Y_1;W) + I(X_2,Y_2;W,X_1) \tag{4.63}$$

and the constraint is relaxed as follows

$$\gamma \geq I(X_1,X_2;Y_1,Y_2|W) = I(X_1;Y_1,Y_2|W) + I(X_2;Y_1,Y_2|W,X_1) \tag{4.64}$$
$$\geq I(X_1;Y_1|W) + I(X_2;Y_2|W,X_1), \tag{4.65}$$

(4.57) follows from splitting the minimization, (4.58) follows from minimizing each subproblem individually which would result in a lower bound to the original problem, (4.59) follows from reducing the number of constraints resulting into a lower bound and (4.60) follows from introducing $\tilde{X}_1$ as a random variable to be optimized, whereas preceding $X_1$ had a given distribution. In other words, the preceding minimization is taken over $p(\tilde{w}|x_2,y_2,x_1)p(x_1|x_2,y_2)$ where $p(x_1|x_2,y_2)$ has a given distribution, whereas now the minimization is taken over $p(\tilde{w}|x_2,y_2,\tilde{x}_1)p(\tilde{x}_1|x_2,y_2)$, where we also optimize over $p(\tilde{x}_1|x_2,y_2)$. Lastly, denoting $W_2 = (\tilde{W},\tilde{X}_1)$, this can be expressed as

$$\inf_{p(w|x_1,x_2,y_1,y_2):I(X_1,X_2;Y_1,Y_2|W)\leq\gamma} I(X_1,X_2,Y_1,Y_2;W)$$
$$\geq \min_{\gamma_1+\gamma_2=\gamma} \left\{ \inf_{p(w_1|x_1,y_1):I(X_1;Y_1|W_1)\leq\gamma_1} I(X_1,Y_1;W_1) + \inf_{p(w_2|x_2,y_2):I(X_2;Y_2|W_2)\leq\gamma_2} I(X_2,Y_2;W_2) \right\} \tag{4.66}$$

After proving it for $n=2$, we will use the standard induction. In other words, we will assume that the converse holds for $n-1$ i.e.

$$C_{\bar{\gamma}}(X^{n-1};Y^{n-1}) \geq \min_{\gamma_i:\sum_{i=1}^{n-1}\gamma_i=\bar{\gamma}} \sum_{i=1}^{n-1} C_{\gamma_i}(X_i;Y_i), \tag{4.67}$$

after we prove it for $n$ as follows,

$$\inf_{p(w|x^n,y^n):I(X^n;Y^n|W)\leq\gamma} I(X^n,Y^n;W)$$
$$\geq \inf_{\substack{p(w|x^n,y^n):\\ I(X^{n-1};Y^{n-1}|W)+I(X_n;Y_n|W,X^{n-1})\leq\gamma}} I(X^{n-1},Y^{n-1};W) + I(X_n,Y_n;W,X^{n-1}) \tag{4.68}$$

$$= \min_{\bar{\gamma}+\gamma_n=\gamma} \left\{ \inf_{\substack{p(w|x^n,y^n):\\ I(X^{n-1};Y^{n-1}|W)\leq\sum_{i=1}^{n-1}\gamma_i,\\ I(X_n;Y_n|W,X^{n-1})\leq\gamma_n}} I(X^{n-1},Y^{n-1};W) + I(X_n,Y_n;W,X^{n-1}) \right\} \tag{4.69}$$

$$\begin{aligned}
&\geq \min_{\bar{\gamma}+\gamma_n=\gamma} \Bigg\{ \inf_{p(w|x^n,y^n):\substack{I(X^{n-1};Y^{n-1}|W)\leq\sum_{i=1}^{n-1}\gamma_i, \\ I(X_n;Y_n|W,X^{n-1})\leq\gamma_n}} I(X^{n-1},Y^{n-1};W) \\
&\quad + \inf_{p(\tilde{w}|x^n,y^n):\substack{I(X^{n-1};Y^{n-1}|\tilde{W})\leq\sum_{i=1}^{n-1}\gamma_i, \\ I(X_n;Y_n|\tilde{W},X^{n-1})\leq\gamma_n}} I(X_n,Y_n;\tilde{W},X^{n-1}) \Bigg\} && (4.70) \\
&\geq \min_{\bar{\gamma}+\gamma_n=\gamma} \Bigg\{ \inf_{p(w|x^{n-1},y^{n-1}):I(X^{n-1};Y^{n-1}|W)\leq\sum_{i=1}^{n-1}\gamma_i} I(X^{n-1},Y^{n-1};W) \\
&\quad + \inf_{p(\tilde{w}|x^n,y_n):I(X_n;Y_n|\tilde{W},X^{n-1})\leq\gamma_n} I(X_n,Y_n;\tilde{W},X^{n-1}) \Bigg\} && (4.71) \\
&\geq \min_{\bar{\gamma}+\gamma_n=\gamma} \Bigg\{ \inf_{p(w|x^{n-1},y^{n-1}):I(X^{n-1};Y^{n-1}|W)\leq\sum_{i=1}^{n-1}\gamma_i} I(X^{n-1},Y^{n-1};W) \\
&\quad + \inf_{p(\tilde{w},\tilde{x}^{n-1}|x_n,y_n):I(X_n;Y_n|\tilde{W},\tilde{X}^{n-1})\leq\gamma_n} I(X_n,Y_n;\tilde{W},\tilde{X}^{n-1}) \Bigg\} && (4.72) \\
&= \min_{\bar{\gamma}+\gamma_n=\gamma} \left\{ C_{\bar{\gamma}} I(X^{n-1};Y^{n-1}) + \inf_{p(w_n|x_n,y_n):I(X_n;Y_n|W_n)\leq\gamma_n} I(X_n,Y_n;W_n) \right\} && (4.73) \\
&\geq \min_{\bar{\gamma}+\gamma_n=\gamma} \left\{ C_{\gamma_n}(X_n;Y_n) + \min_{\gamma_i:\sum_{i=1}^{n-1}\gamma_i=\bar{\gamma}} \sum_{i=1}^{n-1} C_{\gamma_i}(X_i;Y_i) \right\} && (4.74) \\
&= \min_{\gamma_i:\sum_{i=1}^{n}\gamma_i=\gamma} \sum_{i=1}^{n} C_{\gamma_i}(X_i;Y_i), && (4.75)
\end{aligned}$$

where (4.68) follows from

$$\begin{aligned}
I(X^n,Y^n;W) &= I(X^{n-1},Y^{n-1};W) + I(X_n,Y_n;W|X^{n-1},Y^{n-1}) \\
&\quad + I(X_n,Y_n;X^{n-1},Y^{n-1}) && (4.76) \\
&= I(X^{n-1},Y^{n-1};W) + I(X_n,Y_n;W,X^{n-1},Y^{n-1}) && (4.77) \\
&\geq I(X^{n-1},Y^{n-1};W) + I(X_n,Y_n;W,X^{n-1}) && (4.78)
\end{aligned}$$

and the constraint is relaxed as follows

$$\begin{aligned}
\gamma \geq I(X^n;Y^n|W) &= I(X^{n-1};Y^n|W) + I(X_n;Y^n|W,X^{n-1}) && (4.79) \\
&\geq I(X^{n-1};Y^{n-1}|W) + I(X_n;Y_n|W,X^{n-1}). && (4.80)
\end{aligned}$$

Equation (4.69) follows from the same argument as (4.57), (4.70) follows from the same argument as (4.58), (4.71) follows follows from the same argument as (4.59), (4.72) follows from a similar argument as (4.60), (4.73) follows from denoting $W_n = (\tilde{W}, \tilde{X}^{n-1})$, and (4.74) follows from the induction hypothesis (4.67).

### 4.6.3 Proof of Theorem 9

The techniques to establish the optimality of Gaussian distributions is used in [27] and is known as factorization of lower convex envelope. Let us define the following object

$$V(K) := \inf_{(X,Y):K_{(X,Y)}=K} \inf_W h(Y|W) + h(X|W) - (1+\lambda)h(X,Y|W), \tag{4.81}$$

where $\lambda$ is a real number, $0 < \lambda < 1$ and $K$ is an arbitrary covariance matrix. Let $\ell_\lambda(X,Y) = h(Y)+h(X)-(1+\lambda)h(X,Y)$, and $\breve{\ell}_\lambda(X,Y) = \inf_W h(Y|W)+h(X|W)-(1+\lambda)h(X,Y|W)$, where $\breve{\ell}_\lambda(X,Y)$ is the lower convex envelope of $\ell_\lambda(X,Y)$.

First, in Section 4.6.4, we prove that the infimum is attained, then, in Section 4.6.5, we prove that a Gaussian $W$ attains the infimum in Equation (4.81). Together, these two arguments establish Theorem 9.

### 4.6.4 The infimum in Equation (4.81) is attained

**Proposition 14** (Proposition 17 in [27]). *Consider a sequence of random variables $\{X_n, Y_n\}$ such that $K_{(X_n,Y_n)} \preceq K$ for all $n$, then the sequence is tight.*

**Theorem 12** (Prokhorov). *If $\{X_n, Y_n\}$ is a tight sequence then there exists a subsequence $\{X_{n_i}, Y_{n_i}\}$ and a limiting probability distribution $\{X_*, Y_*\}$ such that $\{X_{n_i}, Y_{n_i}\} \overset{w}{\Rightarrow} \{X_*, Y_*\}$ converges weakly in distribution.*

Note that $\ell_\lambda(X,Y) = h(Y) + h(X) - (1+\lambda)h(X,Y)$ can be written as $(1+\lambda)I(X;Y) - \lambda[h(X)+h(Y)]$. Thus, it is enough to show that this expression is lower semi-continuous. We will show by utilizing the following theorem.

**Theorem 13** ([50]). *If $p_{X_n,Y_n} \overset{w}{\Rightarrow} p_{X,Y}$ and $q_{X_n,Y_n} \overset{w}{\Rightarrow} q_{X,Y}$, then $D(p_{X,Y}||q_{X,Y}) \leq \liminf_{n\to\infty} D(p_{X_n,Y_n}||q_{X_n,Y_n})$.*

Observe that $I(X;Y) = D(p_{X,Y}||q_{X,Y})$, where $q_{X,Y} = p_X p_Y$. For the theorem to hold we need to check the assumptions. First, from Theorem 12, we have $p_{X_n,Y_n} \overset{w}{\Rightarrow} p_{X,Y}$. Second, since the marginal distributions converge weakly if the joint distribution converges weakly, we also have $q_{X_n,Y_n} \overset{w}{\Rightarrow} q_{X,Y}$. Therefore,

$$I(X;Y) \leq \liminf_{n\to\infty} I(X_n;Y_n). \tag{4.82}$$

To preserve the covariance matrix $K_{(X,Y)}$, there are three degrees of freedom plus one degree of freedom coming from minimizing the objective, thus $|\mathcal{W}| \leq 4$ is enough to attain the minimum.

Let us introduce $\delta > 0$ and define $N_\delta \sim \mathcal{N}(0,\delta)$, being independent of $\{X_n\}$, $X$, $\{Y_n\}$ and $Y$. From the entropy power inequality, we have

$$h(X_n + N_\delta) \geq h(X_n) \tag{4.83}$$

$$h(Y_n + N_\delta) \geq h(Y_n), \tag{4.84}$$

and moreover, for Gaussian perturbations, we have

$$\liminf_{n\to\infty} h(X_n + N_\delta) = h(X + N_\delta). \tag{4.85}$$

This results in

$$\begin{aligned}
\liminf_{n\to\infty} \ell_\lambda(X_n,Y_n) &= \liminf_{n\to\infty} (1+\lambda)I(X_n;Y_n) - \lambda[h(X_n) + h(Y_n)] && (4.86)\\
&\geq \liminf_{n\to\infty} (1+\lambda)I(X_n;Y_n) - \lambda[h(X_n+N_\delta) + h(Y_n+N_\delta)] && (4.87)\\
&\geq (1+\lambda)I(X;Y) - \lambda[h(X+N_\delta) + h(Y+N_\delta)], && (4.88)
\end{aligned}$$

where (4.87) follows from (4.83) and (4.88) follows from (4.82), (4.85). Letting $\delta \to 0$, we obtain the weak semicontinuity of our object $\liminf_{n\to\infty} \ell_\lambda(X_n, Y_n) \geq \ell_\lambda(X, Y)$.

### 4.6.5 A Gaussian auxiliary $W$ attains the infimum in Equation (4.81)

This proof is an extended version of the arguments in [48], that is instead of (4.81) for $0 < \lambda < 1$, we consider

$$V(K) := \inf_{(X,Y,Z):K_{(X,Y,Z)}=K} \inf_W h(X|W) + h(Y|W) + h(Z|W) - (1+\lambda)h(X,Y,Z|W), \tag{4.89}$$

for $1 < \lambda < 2$. We start by creating two identical and independent copies of the minimizer $(W, X, Y, Z)$, which are $(W_1, X_1, Y_1, Z_1)$ and $(W_2, X_2, Y_2, Z_2)$. In addition, let us denote with $X_w$ the random variable $X|(W = w)$ and define

$$X_{\theta_1}|((W_1, W_2) = (w_1, w_2) := \frac{X_{w_1} + X_{w_2}}{\sqrt{2}},\ X_{\theta_2}|((W_1, W_2) = (w_1, w_2) := \frac{X_{w_1} - X_{w_2}}{\sqrt{2}}, \tag{4.90}$$

$$Y_{\theta_1}|((W_1, W_2) = (w_1, w_2) := \frac{Y_{w_1} + Y_{w_2}}{\sqrt{2}},\ Y_{\theta_2}|((W_1, W_2) = (w_1, w_2) := \frac{Y_{w_1} - Y_{w_2}}{\sqrt{2}}, \tag{4.91}$$

$$Z_{\theta_1}|((W_1, W_2) = (w_1, w_2) := \frac{Z_{w_1} + Z_{w_2}}{\sqrt{2}},\ Z_{\theta_2}|((W_1, W_2) = (w_1, w_2) := \frac{Z_{w_1} - Z_{w_2}}{\sqrt{2}}. \tag{4.92}$$

Thus, we have

$$\begin{aligned}
2V(K) &= h(X_1, X_2|W_1, W_2) + h(Y_1, Y_2|W_1, W_2) + h(Z_1, Z_2|W_1, W_2) \\
&\qquad - (1+\lambda)h(X_1, X_2, Y_1, Y_2, Z_1, Z_2|W_1, W_2) && (4.93) \\
&= h(X_{\theta_1}, X_{\theta_2}|W_1, W_2) + h(Y_{\theta_1}, Y_{\theta_2}|W_1, W_2) + h(Z_{\theta_1}, Z_{\theta_2}|W_1, W_2) \\
&\qquad - (1+\lambda)h(X_{\theta_1}, X_{\theta_2}, Y_{\theta_1}, Y_{\theta_2}, Z_{\theta_1}, Z_{\theta_2}|W_1, W_2) && (4.94) \\
&= h(X_{\theta_1}|W_1, W_2) + h(Y_{\theta_1}|W_1, W_2) + h(Z_{\theta_1}|W_1, W_2) \\
&\qquad - (1+\lambda)h(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}|W_1, W_2) + h(X_{\theta_2}|X_{\theta_1}, W_1, W_2) \\
&\qquad + h(Y_{\theta_2}|Y_{\theta_1}, W_1, W_2) + h(Z_{\theta_2}|Z_{\theta_1}, W_1, W_2) \\
&\qquad - (1+\lambda)h(X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2) && (4.95) \\
&= h(X_{\theta_1}|W_1, W_2) + h(Y_{\theta_1}|W_1, W_2) + h(Z_{\theta_1}|W_1, W_2) \\
&\qquad - (1+\lambda)h(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}|W_1, W_2) + h(X_{\theta_2}|X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2) \\
&\qquad + h(Y_{\theta_2}|X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2) + h(Z_{\theta_2}|X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2) \\
&\qquad - (1+\lambda)h(X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2) \\
&\qquad + I(X_{\theta_2}; Y_{\theta_1}, Z_{\theta_1}|X_{\theta_1}, W_1, W_2) + I(Y_{\theta_2}; X_{\theta_1}, Z_{\theta_1}|Y_{\theta_1}, W_1, W_2) \\
&\qquad + I(Z_{\theta_2}; X_{\theta_1}, Y_{\theta_1}|Z_{\theta_1}, W_1, W_2) && (4.96) \\
&\geq 2V(K) + I(X_{\theta_2}; Y_{\theta_1}, Z_{\theta_1}|X_{\theta_1}, W_1, W_2) + I(Y_{\theta_2}; X_{\theta_1}, Z_{\theta_1}|Y_{\theta_1}, W_1, W_2) \\
&\qquad + I(Z_{\theta_2}; X_{\theta_1}, Y_{\theta_1}|Z_{\theta_1}, W_1, W_2), && (4.97)
\end{aligned}$$

where (4.94) follows from entropy preservation under bijective transformation and (4.97) follows from definition of $V(K)$ such that $K_{(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1})} \preceq K$. The above set of inequalities would imply that

$$\begin{aligned}
I(X_{\theta_2}; Y_{\theta_1}, Z_{\theta_1} | X_{\theta_1}, W_1, W_2) &= 0, \\
I(Y_{\theta_2}; X_{\theta_1}, Z_{\theta_1} | Y_{\theta_1}, W_1, W_2) &= 0, \\
I(Z_{\theta_2}; X_{\theta_1}, Y_{\theta_1} | Z_{\theta_1}, W_1, W_2) &= 0. \qquad (4.98)
\end{aligned}$$

By switching the roles of indexes, we get

$$\begin{aligned}
I(X_{\theta_1}; Y_{\theta_2}, Z_{\theta_2} | X_{\theta_2}, W_1, W_2) &= 0, \\
I(Y_{\theta_1}; X_{\theta_2}, Z_{\theta_2} | Y_{\theta_2}, W_1, W_2) &= 0, \\
I(Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2} | Z_{\theta_2}, W_1, W_2) &= 0. \qquad (4.99)
\end{aligned}$$

By factorizing in another way we have

$$\begin{aligned}
2V(K) &= h(X_1, X_2 | W_1, W_2) + h(Y_1, Y_2 | W_1, W_2) + h(Z_1, Z_2 | W_1, W_2) \\
&\quad - (1+\lambda) h(X_1, X_2, Y_1, Y_2, Z_1, Z_2 | W_1, W_2) \qquad (4.100) \\
&= h(X_{\theta_1}, X_{\theta_2} | W_1, W_2) + h(Y_{\theta_1}, Y_{\theta_2} | W_1, W_2) + h(Z_{\theta_1}, Z_{\theta_2} | W_1, W_2) \\
&\quad - (1+\lambda) h(X_{\theta_1}, X_{\theta_2}, Y_{\theta_1}, Y_{\theta_2}, Z_{\theta_1}, Z_{\theta_2} | W_1, W_2) \qquad (4.101) \\
&= h(X_{\theta_1} | W_1, W_2) + h(Y_{\theta_1} | W_1, W_2) + h(Z_{\theta_1} | W_1, W_2) \\
&\quad - (1+\lambda) h(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1} | W_1, W_2) + h(X_{\theta_2} | W_1, W_2) \\
&\quad + h(Y_{\theta_2} | W_1, W_2) + h(Z_{\theta_2} | W_1, W_2) - (1+\lambda) h(X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2} | W_1, W_2) \\
&\quad - I(X_{\theta_1}; X_{\theta_2} | W_1, W_2) - h(Y_{\theta_1}; Y_{\theta_2} | W_1, W_2) - h(Z_{\theta_1}; Z_{\theta_2} | W_1, W_2) \\
&\quad + (1+\lambda) I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2} | W_1, W_2) \qquad (4.102) \\
&\geq 2V(K) - I(X_{\theta_1}; X_{\theta_2} | W_1, W_2) - I(Y_{\theta_1}; Y_{\theta_2} | W_1, W_2) \\
&\quad - I(Z_{\theta_1}; Z_{\theta_2} | W_1, W_2) + (1+\lambda) I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2} | W_1, W_2). \\
&\qquad (4.103)
\end{aligned}$$

By combining the above inequalities, it implies that

$$\begin{aligned}
(1+\lambda) I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2} | W_1, W_2) &\leq I(X_{\theta_1}; X_{\theta_2} | W_1, W_2) \\
&\quad + I(Y_{\theta_1}; Y_{\theta_2} | W_1, W_2) + I(Z_{\theta_1}; Z_{\theta_2} | W_1, W_2). \qquad (4.104)
\end{aligned}$$

By considering another factorization, we have

$$\begin{aligned}
2V(K) &= h(X_1, X_2 | W_1, W_2) + h(Y_1, Y_2 | W_1, W_2) + h(Z_1, Z_2 | W_1, W_2) \\
&\quad - (1+\lambda) h(X_1, X_2, Y_1, Y_2, Z_1, Z_2 | W_1, W_2) \qquad (4.105) \\
&= h(X_{\theta_1}, X_{\theta_2} | W_1, W_2) + h(Y_{\theta_1}, Y_{\theta_2} | W_1, W_2) + h(Z_{\theta_1}, Z_{\theta_2} | W_1, W_2) \\
&\quad - (1+\lambda) h(X_{\theta_1}, X_{\theta_2}, Y_{\theta_1}, Y_{\theta_2}, Z_{\theta_1}, Z_{\theta_2} | W_1, W_2) \qquad (4.106) \\
&= h(X_{\theta_1} | W_1, W_2) + h(Y_{\theta_1} | W_1, W_2) + h(Z_{\theta_1} | W_1, W_2) \\
&\quad - (1+\lambda) h(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1} | W_1, W_2) + h(X_{\theta_2} | X_{\theta_1}, W_1, W_2) \\
&\quad + h(Y_{\theta_2} | Y_{\theta_1}, W_1, W_2) + h(Z_{\theta_2} | Z_{\theta_1}, W_1, W_2) \\
&\quad - (1+\lambda) h(X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2} | X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2) \qquad (4.107) \\
&= h(X_{\theta_1} | W_1, W_2) + h(Y_{\theta_1} | W_1, W_2) + h(Z_{\theta_1} | W_1, W_2) \\
&\quad - (1+\lambda) h(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1} | W_1, W_2) + h(X_{\theta_2} | X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2) \\
&\quad + h(Y_{\theta_2} | X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2) + h(Z_{\theta_2} | X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2)
\end{aligned}$$

$$
\begin{aligned}
&\quad - (1+\lambda)h(X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}, W_1, W_2)\\
&\quad + I(X_{\theta_2}; Y_{\theta_1}, Z_{\theta_1}|X_{\theta_1}, W_1, W_2) + I(Y_{\theta_2}; X_{\theta_1}, Z_{\theta_1}|Y_{\theta_1}, W_1, W_2)\\
&\quad + I(Z_{\theta_2}; X_{\theta_1}, Y_{\theta_1}|Z_{\theta_1}, W_1, W_2) && (4.108)\\
&\geq V(K) + h(X_{\theta_1}|W_1, W_2) + h(Y_{\theta_1}|W_1, W_2) + h(Z_{\theta_1}|W_1, W_2)\\
&\quad - (1+\lambda)h(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}|W_1, W_2) && (4.109)\\
&= V(K) + h(X_{\theta_1}|X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}, W_1, W_2) + I(X_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2)\\
&\quad + h(Z_{\theta_1}|X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}, W_1, W_2) + I(Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2)\\
&\quad + h(Y_{\theta_1}|X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}, W_1, W_2) + I(Y_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2)\\
&\quad - (1+\lambda)I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2)\\
&\quad - (1+\lambda)h(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}|X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}, W_1, W_2) && (4.110)\\
&\geq 2V(K) + I(X_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2) + I(Y_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2)\\
&\quad + I(Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2)\\
&\quad - (1+\lambda)I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2) && (4.111)
\end{aligned}
$$

where the set of inequalities implies that

$$
\begin{aligned}
(1+\lambda)I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2) &\geq I(X_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2)\\
+ I(Y_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2) &+ I(Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2). \quad (4.112)
\end{aligned}
$$

By making use of (4.98) we can simplify the previous inequality as follows

$$
\begin{aligned}
(1+\lambda)I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2) &\geq I(X_{\theta_1}; X_{\theta_2}|W_1, W_2)\\
&+ I(Y_{\theta_1}; Y_{\theta_2}|W_1, W_2) + I(Z_{\theta_1}; Z_{\theta_2}|W_1, W_2). \quad (4.113)
\end{aligned}
$$

By combining (4.104) and (4.113) we have

$$
\begin{aligned}
(1+\lambda)I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}|W_1, W_2) &= I(X_{\theta_1}; X_{\theta_2}|W_1, W_2)\\
&+ I(Y_{\theta_1}; Y_{\theta_2}|W_1, W_2) + I(Z_{\theta_1}; Z_{\theta_2}|W_1, W_2). \quad (4.114)
\end{aligned}
$$

Now resume in Equation (4.109) and try another factorization thus, we have

$$
\begin{aligned}
2V(K) &\geq V(K) + h(X_{\theta_1}|W_1, W_2) + h(Y_{\theta_1}|W_1, W_2) + h(Z_{\theta_1}|W_1, W_2)\\
&\quad - (1+\lambda)h(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}|W_1, W_2) && (4.115)\\
&= V(K) + h(X_{\theta_1}|X_{\theta_2}, W_1, W_2) + h(Y_{\theta_1}|X_{\theta_2}, W_1, W_2) + h(Z_{\theta_1}|X_{\theta_2}, W_1, W_2)\\
&\quad - (1+\lambda)h(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}|X_{\theta_2}, W_1, W_2) + I(X_{\theta_1}; X_{\theta_2}|W_1, W_2)\\
&\quad + I(Y_{\theta_1}; X_{\theta_2}|W_1, W_2) + I(Z_{\theta_1}; X_{\theta_2}|W_1, W_2)\\
&\quad - (1+\lambda)I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}|W_1, W_2) && (4.116)\\
&\geq 2V(K) + I(X_{\theta_1}; X_{\theta_2}|W_1, W_2) + I(Y_{\theta_1}; X_{\theta_2}|W_1, W_2)\\
&\quad + I(Z_{\theta_1}; X_{\theta_2}|W_1, W_2) - (1+\lambda)I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}|W_1, W_2) && (4.117)
\end{aligned}
$$

and the set of inequalities implies that

$$
\begin{aligned}
(1+\lambda)I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}|W_1, W_2) &\geq I(X_{\theta_1}; X_{\theta_2}|W_1, W_2)\\
&+ I(Y_{\theta_1}; X_{\theta_2}|W_1, W_2) + I(Z_{\theta_1}; X_{\theta_2}|W_1, W_2). \quad (4.118)
\end{aligned}
$$

By making use of (4.98) the above inequality simplifies into

$$\lambda I(X_{\theta_1};X_{\theta_2}|W_1,W_2) \geq I(Y_{\theta_1};X_{\theta_2}|W_1,W_2) + I(Z_{\theta_1};X_{\theta_2}|W_1,W_2). \tag{4.119}$$

Once again, we resume in Equation (4.109) thus, we have

$$\begin{aligned}
2V(K) \geq\ & V(K) + h(X_{\theta_1}|W_1,W_2) + h(Y_{\theta_1}|W_1,W_2) + h(Z_{\theta_1}|W_1,W_2) \\
& - (1+\lambda)h(X_{\theta_1},Y_{\theta_1},Z_{\theta_1}|W_1,W_2) && (4.120)\\
=\ & V(K) + h(X_{\theta_1}|Y_{\theta_2},W_1,W_2) + h(Y_{\theta_1}|Y_{\theta_2},W_1,W_2) + h(Z_{\theta_1}|Y_{\theta_2},W_1,W_2) \\
& - (1+\lambda)h(X_{\theta_1},Y_{\theta_1},Z_{\theta_1}|Y_{\theta_2},W_1,W_2) + I(X_{\theta_1};Y_{\theta_2}|W_1,W_2) \\
& + I(Y_{\theta_1};Y_{\theta_2}|W_1,W_2) + I(Z_{\theta_1};Y_{\theta_2}|W_1,W_2) \\
& - (1+\lambda)I(X_{\theta_1},Y_{\theta_1},Z_{\theta_1};Y_{\theta_2}|W_1,W_2) && (4.121)\\
\geq\ & 2V(K) + I(X_{\theta_1};Y_{\theta_2}|W_1,W_2) + I(Y_{\theta_1};Y_{\theta_2}|W_1,W_2) \\
& + I(Z_{\theta_1};Y_{\theta_2}|W_1,W_2) - (1+\lambda)I(X_{\theta_1},Y_{\theta_1},Z_{\theta_1};Y_{\theta_2}|W_1,W_2) && (4.122)
\end{aligned}$$

where the set of inequalities implies that

$$\begin{aligned}
(1+\lambda)I(X_{\theta_1},Y_{\theta_1},Z_{\theta_1};Y_{\theta_2}|W_1,W_2) \geq\ & I(X_{\theta_1};Y_{\theta_2}|W_1,W_2) \\
& + I(Y_{\theta_1};Y_{\theta_2}|W_1,W_2) + I(Z_{\theta_1};Y_{\theta_2}|W_1,W_2). && (4.123)
\end{aligned}$$

By making use of (4.98) the above inequality simplifies into

$$\lambda I(Y_{\theta_1};Y_{\theta_2}|W_1,W_2) \geq I(X_{\theta_1};Y_{\theta_2}|W_1,W_2) + I(Z_{\theta_1};Y_{\theta_2}|W_1,W_2). \tag{4.124}$$

For the last time we resume in Equation (4.109) and try another factorization thus, we have

$$\begin{aligned}
2V(K) \geq\ & V(K) + h(X_{\theta_1}|W_1,W_2) + h(Y_{\theta_1}|W_1,W_2) + h(Z_{\theta_1}|W_1,W_2) \\
& - (1+\lambda)h(X_{\theta_1},Y_{\theta_1},Z_{\theta_1}|W_1,W_2) && (4.125)\\
=\ & V(K) + h(X_{\theta_1}|Z_{\theta_2},W_1,W_2) + h(Y_{\theta_1}|Z_{\theta_2},W_1,W_2) + h(Z_{\theta_1}|Z_{\theta_2},W_1,W_2) \\
& - (1+\lambda)h(X_{\theta_1},Y_{\theta_1},Z_{\theta_1}|Z_{\theta_2},W_1,W_2) + I(X_{\theta_1};Z_{\theta_2}|W_1,W_2) \\
& + I(Y_{\theta_1};Z_{\theta_2}|W_1,W_2) + I(Z_{\theta_1};Z_{\theta_2}|W_1,W_2) \\
& - (1+\lambda)I(X_{\theta_1},Y_{\theta_1},Z_{\theta_1};Z_{\theta_2}|W_1,W_2) && (4.126)\\
\geq\ & 2V(K) + I(X_{\theta_1};Z_{\theta_2}|W_1,W_2) + I(Y_{\theta_1};Z_{\theta_2}|W_1,W_2) \\
& + I(Z_{\theta_1};Z_{\theta_2}|W_1,W_2) - (1+\lambda)I(X_{\theta_1},Y_{\theta_1},Z_{\theta_1};Z_{\theta_2}|W_1,W_2) && (4.127)
\end{aligned}$$

where the set of inequalities implies that

$$\begin{aligned}
(1+\lambda)I(X_{\theta_1},Y_{\theta_1},Z_{\theta_1};Z_{\theta_2}|W_1,W_2) \geq\ & I(X_{\theta_1};Z_{\theta_2}|W_1,W_2) \\
& + I(Y_{\theta_1};Z_{\theta_2}|W_1,W_2) + I(Z_{\theta_1};Z_{\theta_2}|W_1,W_2). && (4.128)
\end{aligned}$$

By making use of (4.98) the above inequality simplifies into

$$\lambda I(Z_{\theta_1};Z_{\theta_2}|W_1,W_2) \geq I(X_{\theta_1};Z_{\theta_2}|W_1,W_2) + I(Y_{\theta_1};Z_{\theta_2}|W_1,W_2). \tag{4.129}$$

By switching the role of the index $\theta_1$ and $\theta_2$ for (4.119), (4.124) and (4.129) we get

$$\begin{aligned}
\lambda I(X_{\theta_1};X_{\theta_2}|W_1,W_2) &\geq I(X_{\theta_1};Y_{\theta_2}|W_1,W_2) + I(X_{\theta_1};Z_{\theta_2}|W_1,W_2), && (4.130)\\
\lambda I(Y_{\theta_1};Y_{\theta_2}|W_1,W_2) &\geq I(Y_{\theta_1};X_{\theta_2}|W_1,W_2) + I(Y_{\theta_1};Z_{\theta_2}|W_1,W_2), && (4.131)\\
\lambda I(Z_{\theta_1};Z_{\theta_2}|W_1,W_2) &\geq I(Z_{\theta_1};X_{\theta_2}|W_1,W_2) + I(Z_{\theta_1};Y_{\theta_2}|W_1,W_2). && (4.132)
\end{aligned}$$

By using the non-negativity of conditional mutual information we have

$$\begin{aligned}
&I(Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Z_{\theta_2}|X_{\theta_1}, Y_{\theta_2}) \\
&\quad = I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Z_{\theta_2}|Y_{\theta_2}) - I(X_{\theta_1}; X_{\theta_2}, Z_{\theta_2}|Y_{\theta_2}) && (4.133)\\
&\quad = I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Z_{\theta_2}, Y_{\theta_2}) - I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; Y_{\theta_2})\\
&\qquad - I(X_{\theta_1}; X_{\theta_2}, Z_{\theta_2}, Y_{\theta_2}) + I(X_{\theta_1}; Y_{\theta_2}) && (4.134)\\
&\quad = I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Z_{\theta_2}, Y_{\theta_2}) - I(Y_{\theta_1}; Y_{\theta_2})\\
&\qquad - I(X_{\theta_1}; X_{\theta_2}) + I(X_{\theta_1}; Y_{\theta_2}) && (4.135)\\
&\quad = \frac{1}{1+\lambda}\left[I(X_{\theta_1}; X_{\theta_2}) + I(Y_{\theta_1}; Y_{\theta_2}) + I(Z_{\theta_1}; Z_{\theta_2})\right] - I(Y_{\theta_1}; Y_{\theta_2})\\
&\qquad - I(X_{\theta_1}; X_{\theta_2}) + I(X_{\theta_1}; Y_{\theta_2}) \geq 0 && (4.136)
\end{aligned}$$

where (4.135) follows from (4.98) and (4.136) follows from (4.114). By rewriting the last inequality we get

$$-\lambda I(X_{\theta_1}; X_{\theta_2}) - \lambda I(Y_{\theta_1}; Y_{\theta_2}) + I(Z_{\theta_1}; Z_{\theta_2}) + (1+\lambda) I(X_{\theta_1}; Y_{\theta_2}) \geq 0. \qquad (4.137)$$

By using the non-negativity of conditional mutual information we have

$$\begin{aligned}
&I(X_{\theta_1}, Y_{\theta_1}; X_{\theta_2}, Z_{\theta_2}|Z_{\theta_1}, Y_{\theta_2}) \\
&\quad = I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Z_{\theta_2}|Y_{\theta_2}) - I(Z_{\theta_1}; X_{\theta_2}, Z_{\theta_2}|Y_{\theta_2}) && (4.138)\\
&\quad = I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Z_{\theta_2}, Y_{\theta_2}) - I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; Y_{\theta_2})\\
&\qquad - I(Z_{\theta_1}; X_{\theta_2}, Z_{\theta_2}, Y_{\theta_2}) + I(Z_{\theta_1}; Y_{\theta_2}) && (4.139)\\
&\quad = I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Z_{\theta_2}, Y_{\theta_2}) - I(Y_{\theta_1}; Y_{\theta_2})\\
&\qquad - I(Z_{\theta_1}; Z_{\theta_2}) + I(Z_{\theta_1}; Y_{\theta_2}) && (4.140)\\
&\quad = \frac{1}{1+\lambda}\left[I(X_{\theta_1}; X_{\theta_2}) + I(Y_{\theta_1}; Y_{\theta_2}) + I(Z_{\theta_1}; Z_{\theta_2})\right] - I(Y_{\theta_1}; Y_{\theta_2})\\
&\qquad - I(Z_{\theta_1}; Z_{\theta_2}) + I(Z_{\theta_1}; Y_{\theta_2}) \geq 0 && (4.141)
\end{aligned}$$

where (4.140) follows from (4.98) and (4.141) follows from (4.114). By rewriting the last inequality we get

$$-\lambda I(Z_{\theta_1}; Z_{\theta_2}) - \lambda I(Y_{\theta_1}; Y_{\theta_2}) + I(X_{\theta_1}; X_{\theta_2}) + (1+\lambda) I(Z_{\theta_1}; Y_{\theta_2}) \geq 0. \qquad (4.142)$$

By adding (4.137) and (4.142) and using (4.124) for $1 < \lambda < 2$, we obtain

$$\lambda I(Y_{\theta_1}; Y_{\theta_2}) \geq I(X_{\theta_1}; X_{\theta_2}) + I(Z_{\theta_1}; Z_{\theta_2}). \qquad (4.143)$$

In a similar fashion we obtain

$$\begin{aligned}
\lambda I(X_{\theta_1}; X_{\theta_2}) &\geq I(Y_{\theta_1}; Y_{\theta_2}) + I(Z_{\theta_1}; Z_{\theta_2}), && (4.144)\\
\lambda I(Z_{\theta_1}; Z_{\theta_2}) &\geq I(X_{\theta_1}; X_{\theta_2}) + I(Y_{\theta_1}; Y_{\theta_2}). && (4.145)
\end{aligned}$$

By adding up (4.143), (4.144) and (4.145) we obtain

$$(2-\lambda)\left[I(X_{\theta_1}; X_{\theta_2}) + I(Y_{\theta_1}; Y_{\theta_2}) + I(Z_{\theta_1}; Z_{\theta_2})\right] \leq 0, \qquad (4.146)$$

thus, $I(X_{\theta_1}; X_{\theta_2}) = I(Y_{\theta_1}; Y_{\theta_2}) = I(Z_{\theta_1}; Z_{\theta_2}) = 0$, which implies that

$$I(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1}; X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2}) = 0. \qquad (4.147)$$

The following statements are true.

- The pair $(X_1, Y_1, Z_1)$ and $(X_2, Y_2, Z_2)$ are conditionally independent given $W_1 = w_1, W_2 = w_2$ from assumption, i.e. $(X_{w_1}, Y_{w_1}, Z_{w_1})$ and $(X_{w_2}, Y_{w_2}, Z_{w_2})$ are independent.
- The pair $(X_{\theta_1}, Y_{\theta_1}, Z_{\theta_1})$ and $(X_{\theta_2}, Y_{\theta_2}, Z_{\theta_2})$ are conditionally independent given $W_1 = w_1, W_2 = w_2$, in other words we have $\left(\frac{X_{w_1}+X_{w_2}}{\sqrt{2}}, \frac{Y_{w_1}+Y_{w_2}}{\sqrt{2}}, \frac{Z_{w_1}+Z_{w_2}}{\sqrt{2}}\right)$ and $\left(\frac{X_{w_1}-X_{w_2}}{\sqrt{2}}, \frac{Y_{w_1}-Y_{w_2}}{\sqrt{2}}, \frac{Z_{w_1}-Z_{w_2}}{\sqrt{2}}\right)$ are independent. This follows from (4.147).

By applying Theorem 1 on the above listed statements, we can infer that $(X, Y, Z)|\{W = w\} \sim \mathcal{N}(0, K_w)$, where $K_w$ might depend on the realization of $W = w$. We will now argue that this is not the case. To make a brief summary we have shown the existence part thus, by choosing $W$ to be the trivial random variable a single Gaussian (i.e. not a Gaussian mixture) is one of the possible minimizers. Let us suppose that there are two Gaussian minimizers $\mathcal{N}(0, K_{w_1})$ and $\mathcal{N}(0, K_{w_2})$, where $K_{w_1} \neq K_{w_2}$. Consider the random variable $(W, X, Y, Z)$ where, $(X, Y, Z)|\{W = w_1\} \sim \mathcal{N}(0, K_{w_1})$ and $(X, Y, Z)|\{W = w_2\} \sim \mathcal{N}(0, K_{w_2})$. Therefore the triple $(W, X, Y, Z)$ also attains $V(K)$ and satisfies the covariance constraint. At the same time, we showed that the sum and the difference are also minimizers and they must be independent of each other, which happens only when $K_{w_1} = K_{w_2}$. In other words, $K_w$ does not depend on the realization $W = w$, and the $(X, Y, Z)$ is a single Gaussian (i.e. not a Gaussian mixture). We established that $(X, Y, Z)$ is a unique Gaussian minimizer. The marginal of $W$ does not affect the defined problem and without loss of optimality we can assume it to be Gaussian (or by an optimal transform on $W$). Thus,

$$V(K) = h(X|W) + h(Y|W) + h(Z|W) - (1+\lambda)h(X, Y, Z|W). \tag{4.148}$$

Furthermore, there exists a decomposition $(X, Y, Z) = W + (X', Y', Z') \sim \mathcal{N}(0, K)$, where $W$ is independent of $(X', Y', Z')$ and $W \sim \mathcal{N}(0, K - K')$ and $(X', Y', Z') \sim \mathcal{N}(0, K')$. Then,

$$V(K) = h(X') + h(Y') + h(Z') - (1+\lambda)h(X', Y', Z'), \tag{4.149}$$

thus establishing Theorem 9, because

$$V(K) \leq \inf_W h(X|W) + h(Y|W) + h(Z|W) - (1+\lambda)h(X, Y, Z|W), \tag{4.150}$$

by the definition of $V(K)$ in (4.81).

### 4.6.6 Proof of Lemma 11

Let us parametrize $K'$ as $K' = \begin{pmatrix} \sigma_X^2 & q\sigma_X\sigma_Y \\ q\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \succeq 0$. By substituting we obtain

$$\min_{K': 0 \preceq K' \preceq \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}} h(X') + h(Y') - (1+\lambda)h(X', Y')$$

$$= \min_{(\sigma_X, \sigma_Y, q) \in \mathcal{A}_\rho} \frac{1}{2}\log (2\pi e)^2\sigma_X^2\sigma_Y^2 - \frac{1+\lambda}{2}\log (2\pi e)^2\sigma_X^2\sigma_Y^2(1-q^2) \tag{4.151}$$

where the set $\mathcal{A}_\rho$ is defined as

$$\mathcal{A}_\rho := \left\{(\sigma_X, \sigma_Y, q) : \begin{pmatrix} \sigma_X^2 - 1 & q\sigma_X\sigma_Y - \rho \\ q\sigma_X\sigma_Y - \rho & \sigma_Y^2 - 1 \end{pmatrix} \preceq 0\right\}. \tag{4.152}$$

Matrices of dimension two by two are negative definite (semi-definite) if and only if the trace is negative (non-positive) and determinant is positive (non-negative). Thus, the set $\mathcal{A}_\rho$ is

$$\mathcal{A}_\rho = \left\{(\sigma_X, \sigma_Y, q) : \substack{\sigma_X^2+\sigma_Y^2\leq 2,\\ (1-q^2)\sigma_X^2\sigma_Y^2+2\rho q\sigma_X\sigma_Y+1-\rho^2-(\sigma_X^2+\sigma_Y^2)\geq 0}\right\}. \tag{4.153}$$

Let us define

$$\mathcal{B}_\rho := \left\{(\sigma_X, \sigma_Y, q) : \substack{\sigma_X\sigma_Y\leq 1,\\ (1-q^2)\sigma_X^2\sigma_Y^2+2\rho q\sigma_X\sigma_Y+1-\rho^2-2\sigma_X\sigma_Y)\geq 0}\right\}, \tag{4.154}$$

and the inequality $\sigma_X^2 + \sigma_Y^2 \geq 2\sigma_X\sigma_Y$, implies that $\mathcal{A}_\rho \subseteq \mathcal{B}_\rho$. By reparametrizing $\sigma^2 = \sigma_X\sigma_Y$, the set $\mathcal{B}_\rho$ becomes

$$\mathcal{D}_\rho := \left\{(\sigma^2, q) : \substack{\sigma^2\leq 1,\\ (\sigma^2(1-q)-1+\rho)(\sigma^2(1+q)-1-\rho)\geq 0}\right\}. \tag{4.155}$$

The second inequality in the definition of the set $\mathcal{D}_\rho$ has roots $\sigma^2 = \frac{1+\rho}{1+q}$ and $\sigma^2 = \frac{1-\rho}{1-q}$ when meet with equality. Thus, we can rewrite the set $\mathcal{D}_\rho$ as

$$\mathcal{D}_\rho = \left\{(\sigma^2, q) : \substack{\rho\geq q, \quad \sigma^2(1-q)\leq 1-\rho\\ \rho<q, \quad \sigma^2(1+q)\leq 1+\rho}\right\}. \tag{4.156}$$

Thus, we have

$$\min_{(\sigma_X,\sigma_Y,q)\in\mathcal{A}_\rho} \frac{1}{2}\log{(2\pi e)^2\sigma_X^2\sigma_Y^2} - \frac{1+\lambda}{2}\log{(2\pi e)^2\sigma_X^2\sigma_Y^2(1-q^2)} \geq \min_{(\sigma^2,q)\in\mathcal{D}_\rho} f(\lambda, \sigma^2, q) \tag{4.157}$$

where,

$$f(\lambda, \sigma^2, q) = \frac{1}{2}\log{(2\pi e)^2\sigma^4} - \frac{1+\lambda}{2}\log{(2\pi e)^2\sigma^4(1-q^2)}. \tag{4.158}$$

For now let us assume $\rho$ is positive and start from the case $\rho \geq q$. Then, by weak duality we have

$$\min_{(\sigma^2,q)\in\mathcal{D}_\rho} f(\lambda, \sigma^2, q) \geq \min_{\sigma^2,q} f(\lambda, \sigma^2, q) + \mu(\sigma^2(1-q) - 1 + \rho)), \tag{4.159}$$

for any $\mu \geq 0$. By applying Karush-Kuhn-Tucker (KKT) conditions on the right hand side of (4.159) we get

$$\frac{\partial}{\partial\sigma^2} = -\frac{\lambda}{\sigma^2} + \mu(1-q) = 0, \tag{4.160}$$

$$\frac{\partial}{\partial q} = \frac{(1+\lambda)q}{1-q^2} - \mu\sigma^2 = 0, \tag{4.161}$$

$$\mu(\sigma^2(1-q) - 1 + \rho)) = 0, \tag{4.162}$$

where (4.160), (4.161) is known as stationary condition and (4.162) is known as complementary slackness condition. By using (4.160) have

$$\mu = \frac{\lambda}{\sigma^2(1-q)}. \tag{4.163}$$

By using (4.161) we have

$$\mu = \frac{(1+\lambda)q}{\sigma^2(1-q^2)}. \tag{4.164}$$

By equating (4.163) and (4.164) we deduce that $q_* = \lambda$. Since $\lambda > 0$, then $\mu \neq 0$ and by using (4.162) we have $\sigma_*^2 = \frac{1-\rho}{1-\lambda}$. In addition, $\mu_* = \frac{\lambda}{1-\rho}$.

Since the KKT conditions are satisfied by $q_*, \sigma_*^2$ and $\mu_*$ then strong duality holds. Thus, we have

$$\begin{aligned} \min_{(\sigma^2,q)\in\mathcal{D}_\rho} f(\lambda,\sigma^2,q) &= \max_{\mu\geq 0}\min_{\sigma^2,q} f(\lambda,\sigma^2,q) + \mu(\sigma^2(1-q)-1+\rho)) && (4.165)\\ &= f(\lambda, \frac{1-\rho}{1-\lambda}, \lambda) && (4.166)\\ &= \frac{1}{2}\log\frac{1}{1-\lambda^2} - \frac{\lambda}{2}\log{(2\pi e)^2}\frac{(1-\rho)^2(1+\lambda)}{1-\lambda}. && (4.167) \end{aligned}$$

By combining (4.151), (4.157), (4.165) and (4.167) we get the desired lower bound.

For the case $\rho < q$, let us optimize over $\sigma^2$ for any fixed $q$. The function $f$ is decreasing in $\sigma^2$. Also, the function $f$ is convex in $\sigma^2$. Since the object is continuous in $\sigma^2$ and the constraint is linear for a given value of $q$, then the optimal choice is $\sigma^2 = \frac{1+\rho}{1+q}$. Thus,

$$\min_{(\sigma^2,q)\in\mathcal{D}_\rho} f(\lambda,\sigma^2,q) \geq \min_{q\in[\rho,1]} f(\lambda, \frac{1+\rho}{1+q}, q). \tag{4.168}$$

The function on the right hand side can be written as

$$f(\lambda, \frac{1+\rho}{1+q}, q) = \frac{1}{2}\log{(2\pi e)^2}\frac{(1+\rho)^2}{(1+q)^2} - \frac{1+\lambda}{2}\log{(2\pi e)^2}\frac{(1+\rho)^2(1-q)}{(1+q)}. \tag{4.169}$$

The function is convex and increasing in $q$ for $q \in [\rho, 1]$,

$$\begin{aligned} \frac{\partial f}{\partial q} &= \frac{q+\lambda}{1-q^2} > 0, && (4.170)\\ \frac{\partial^2 f}{\partial q^2} &= \frac{1+q^2+2\lambda q}{(1-q^2)^2} > 0, && (4.171) \end{aligned}$$

thus, the optimal value of $q^* = \rho$ and $\sigma^2 = 1$. To conclude we show that $f(\lambda, \frac{1-\rho}{1-\lambda}, \lambda) \leq f(\lambda, 1, \rho)$ for $\lambda \leq \rho$. To show this we define

$$\begin{aligned} h(\lambda) &:= f(\lambda, \frac{1-\rho}{1-\lambda}, \lambda) - f(\lambda, 1, \rho) && (4.172)\\ &= \frac{1}{2}\log\frac{1-\rho^2}{1-\lambda^2} - \frac{\lambda}{2}\log\frac{(1+\lambda)(1-\rho)}{(1-\lambda)(1+\rho)}, && (4.173) \end{aligned}$$

and function $h$ is increasing in $\lambda$,

$$\frac{\partial h}{\partial \lambda} = -\frac{1}{2} \log \frac{(1+\lambda)(1-\rho)}{(1-\lambda)(1+\rho)} \geq 0, \quad \text{for } \lambda \leq \rho \tag{4.174}$$

and it is concave in $\lambda$,

$$\frac{\partial^2 h}{\partial \lambda^2} = -\frac{1}{1-\lambda^2} < 0, \tag{4.175}$$

thus, $h(\lambda) \leq h(\rho) = 0$. Then, $f(\lambda, 1, \rho) \geq f(\lambda, \frac{1-\rho}{1-\lambda}, \lambda)$. The argument goes through also for the case when $\rho$ is negative, which completes the proof.

# 5 The Gaussian Lossy Gray-Wyner Network

## 5.1 Introduction

Source coding for network scenarios has a long history, starting with the work of Slepian and Wolf [51] concerning the distributed compression of correlated sources in a lossless reconstruction setting. In this work, we study a source coding network introduced by Gray and Wyner [35]. In this network, there is a single encoder. It encodes a pair of sources, $(X, Y)$, into three messages, namely, a common message and two private messages. There are two decoders, both receiving the common message, and each receiver has access to the respective private message. For this problem, both in the setting of lossless and of lossy reconstruction, Gray and Wyner fully characterized the optimal rate(-distortion) regions in [35], up to the optimization over a single auxiliary random variable (which represents the common message). An alternative operational interpretation of the Gray-Wyner network as a model for a caching system has been proposed in [43, Section III.C].

For Gaussian sources the Gray-Wyner network [35] problem remained unsolved. A closed form solution is given in [35] by assuming that the auxiliaries are Gaussian. Partial progress was made in [52, 37], when the sum of the common rate and the private rates is exactly equal to the joint rate distortion function. For this corner case, it is known that Wyner's common information is the smallest rate needed on the common channel. In the present chapter, we solve[1] the Gray-Wyner network [35] for Gaussian sources, encompassing all previous partial results.

### 5.1.1 Contribution

The contributions of the present chapter includes, for the Gaussian lossy Gray-Wyner network under symmetric mean-squared error distortion, we prove that it is optimal to select the auxiliary random variable to be jointly Gaussian with the source random variables and we compute closed-form solutions of the common rate versus the sum of private rates. That is reflected in Theorem 15.

[1] The material of this chapter has appeared in
- E. Sula and M. Gastpar, "The Gaussian lossy Gray-Wyner network," in *54th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, 2020.

## 5.2 System Model

Let us assume that the probability of the pair $(X, Y)$ is given and $X \in \mathcal{X}$, $Y \in \mathcal{Y}$. Let $S_c, S_x$ and $S_y$ be messages represented by $nR_c, nR_x$ and $nR_y$ bits. Let

$$(S_c, S_x, S_y) = f_{\mathcal{E}}(X^n, Y^n), \tag{5.1}$$

where $f_{\mathcal{E}}(.)$ is the encoding function, $X^n \in \mathcal{X}^n$ and $Y^n \in \mathcal{Y}^n$. Let

$$\hat{X}^n = f_{\mathcal{D}_x}(S_c, S_x), \quad \hat{Y}^n = f_{\mathcal{D}_y}(S_c, S_y), \tag{5.2}$$

where $f_{\mathcal{D}_x}(.), f_{\mathcal{D}_y}(.)$ are the decoding function, $\hat{X}^n \in \hat{\mathcal{X}}^n$ and $\hat{Y}^n \in \hat{\mathcal{Y}}^n$. The system has a distortion $(\Delta_x, \Delta_y)$, where

$$\Delta_x = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} d_x(X_k, \hat{X}_k)\right], \quad \Delta_y = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} d_y(Y_k, \hat{Y}_k)\right]. \tag{5.3}$$

Let the event $\mathcal{E}_r$ be

$$\mathcal{E}_r = \{D_x < \Delta_x\} \cup \{D_y < \Delta_y\}. \tag{5.4}$$

A rate triple $(R_0, R_1, R_2)$ is said to be $(D_x, D_y)$-achievable if, for any specified positive error probability $P_e$ and sufficiently large $n$, there are encoding and decoding functions such that $\Pr[\mathcal{E}_r] \leq P_e$. The closure of the set of achievable $(R_0, R_1, R_2)$ is called $\mathcal{R}(D_x, D_y)$.

## 5.3 The lossless Gray-Wyner Network

Gray and Wyner in [35] introduced a particular network source coding problem referred to as the Gray-Wyner network.

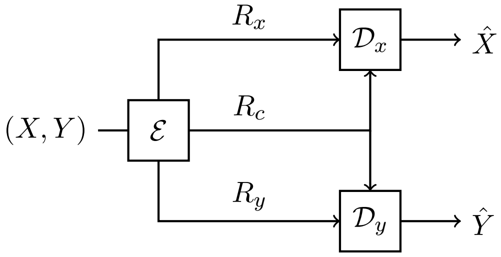


Figure 5.1 – The Gray-Wyner Network

The Gray-Wyner network [35] is composed of one joint sender and two receivers. The purpose of this network is to convey the joint source $(X, Y)$ (where source $X$ and $Y$ are correlated) to the two receivers, such that each receiver gets only one of the source, either $X$ or $Y$. In other words, receiver or decoder $\mathcal{D}_x$ wants to obtain source $X$, and receiver or decoder $\mathcal{D}_y$ wants to obtain source $Y$. The network is consisting of three links or channels as described in the figure. The central link, of rate $R_c$, is provided to both receivers. In addition, each receiver also has access to only one private link. From now on we denote the rates of the private links by $R_x$

and $R_y$, respectively. The main result of [35, Theorem 4], says that the rate region of lossless Gray-Wyner network ($\Delta_x = \Delta_y = 0$ in Section 5.2) is given by the closure of the union of the regions

$$\mathcal{R} = \{(R_c, R_x, R_y) : R_c \geq I(X, Y; W), R_x \geq H(X|W), R_y \geq H(Y|W)\}, \tag{5.5}$$

where the union is over all probability distributions $p(w, x, y)$ with marginals $p(x, y)$.

## 5.4 The Gaussian lossy Gray-Wyner Network

As in the original work of Gray and Wyner [35], one may instead ask for *lossy* reconstructions of the original sources $X$ and $Y$ with respect to fidelity criteria.

**Theorem 14** (Theorem 6, Equation (40), in [35]). *The rate region of the lossy Gray-Wyner network is given by the closure of the union of the regions*

$$\mathcal{R}(D_x, D_y) = \{(R_c, R_x, R_y) : R_c \geq I(X, Y; W), R_x \geq R_{X|W}(D_x), R_y \geq R_{Y|W}(D_y)\}, \tag{5.6}$$

*where $R_{X|W}$ and $R_{Y|W}$ are conditional rate-distortion function, for a given probability density function of $(X, Y)$.*

This motivates the following definition that is directly linked with the quantity $T(\boldsymbol{\lambda})$ in [35]), that is

$$T(\boldsymbol{\lambda}) = \min_{(R_c, R_x, R_y) \in \mathcal{R}(D_x, D_y)} R_c + \lambda_x R_x + \lambda_y R_y. \tag{5.7}$$

**Definition 5** (Gray-Wyner rate-distortion function). *For random variables $X$ and $Y$ with joint distribution $p(x, y)$, the Gray-Wyner rate-distortion function is defined as*

$$\mathrm{R}_{\boldsymbol{D}, \boldsymbol{\alpha}}(X, Y) = \inf I(X, Y; W) \tag{5.8}$$

*such that $I(X; \hat{X}|W) \leq \alpha_x$ and $I(Y; \hat{Y}|W) \leq \alpha_y$, where the minimum is over all probability distributions $p(\hat{x}, \hat{y}, w, x, y)$ with marginals $p(x, y)$ and satisfying*

$$\mathbb{E}[d_x(X, \hat{X})] \leq D_x \text{ and } \mathbb{E}[d_y(Y, \hat{Y})] \leq D_y, \tag{5.9}$$

*where $d_x(\cdot, \cdot)$ and $d_y(\cdot, \cdot)$ are arbitrary single-letter distortion measures (as in, e.g., [35, Eqn. (30) ff.]).*

Let us consider a special case of Definition 5 for which we can derive a closed-form solution. For a fixed probability distribution $p(x, y)$, we define

$$\mathrm{R}_{D, \alpha}(X, Y) = \inf I(X, Y; W) \tag{5.10}$$

such that $I(X; \hat{X}|W) + I(Y; \hat{Y}|W) \leq \alpha$, where the minimum is over all probability distributions $p(\hat{x}, \hat{y}, w, x, y)$ with marginals $p(x, y)$ and satisfying

$$\mathbb{E}[d_x(X, \hat{X})] \leq D \text{ and } \mathbb{E}[d_y(Y, \hat{Y})] \leq D, \tag{5.11}$$

where $D_x = D_y = D$.

**Theorem 15.** *Let $X$ and $Y$ be jointly Gaussian with mean zero, equal variance $\sigma^2$, and with correlation coefficient $\rho$. Let $d_x(\cdot,\cdot)$ and $d_y(\cdot,\cdot)$ be the mean-squared error distortion measure. Then,*

$$\mathrm{R}_{D,\alpha}(X,Y) = \begin{cases} \frac{1}{2}\log^+ \frac{1+\rho}{2\frac{D}{\sigma^2}e^{\alpha}+\rho-1}, & \text{if } \sigma^2(1-\rho) \leq De^{\alpha} \leq \sigma^2 \\ \frac{1}{2}\log^+ \frac{1-\rho^2}{\frac{D^2}{\sigma^4}e^{2\alpha}}, & \text{if } De^{\alpha} \leq \sigma^2(1-\rho), \end{cases} \tag{5.12}$$

*that is defined in (5.10).*

The proof of this theorem is given in Section 5.4.1.

**Remark 4.** *Assuming that auxiliaries are jointly Gaussian with the sources, the same formula was derived in [53, Theorem 4.3] via a different reasoning.*

Figure 5.2 will illustrate the piecewise function of (5.12) in terms of $De^{\alpha}$, for the specific choice of $\rho = 0.5$ and $\sigma^2 = 1$.

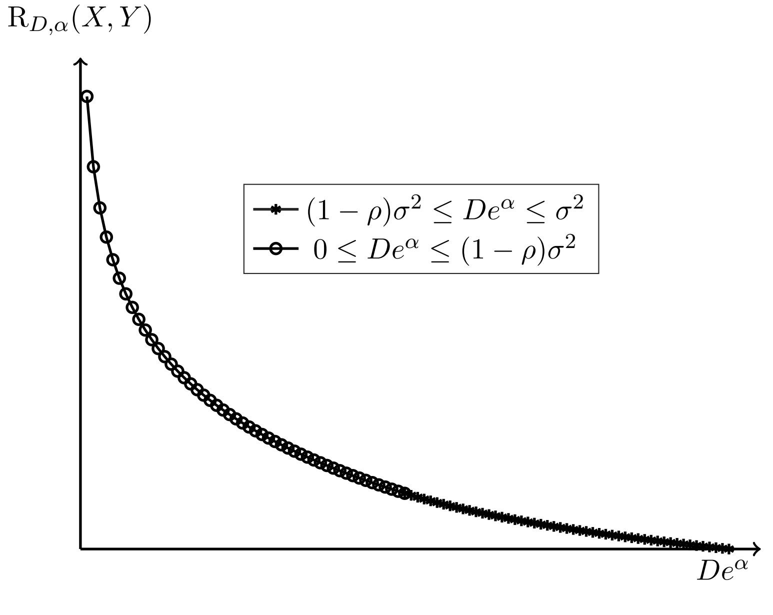


Figure 5.2 – Piecewise function, $\mathrm{R}_{D,\alpha}(X,Y)$ versus $De^{\alpha}$.

### 5.4.1 Proof of Theorem 15

Let $K$ be the covariance matrix with unit entries in the main diagonal and $\rho$ entries in the off-diagonal. First, we consider the lower bounds for definition 5. We observe that for mean-squared error, a scheme attaining distortion $D$ for sources of variance $\sigma^2$ is a scheme attaining distortion $D/\sigma^2$ on unit-variance sources, and vice versa. Therefore, for ease of notation, in the sequel, we assume that the sources are of unit variance. Then, we can bound:

$$
\begin{aligned}
&\mathrm{R}_{D,\alpha}(X,Y) && (5.13)\\
&= \inf_{\substack{W,\hat{X},\hat{Y}: I(X;\hat{X}|W)+I(Y;\hat{Y}|W)\le\alpha \\ \mathbb{E}[(X-\hat{X})^2]\le D \\ \mathbb{E}[(Y-\hat{Y})^2]\le D}} I(X,Y;W) && (5.14)\\
&\ge \inf_{\substack{W,\hat{X},\hat{Y}: \mathbb{E}[(X-\hat{X})^2]\le D \\ \mathbb{E}[(Y-\hat{Y})^2]\le D}} I(X,Y;W) + \nu I(X;\hat{X}|W)\\
&\qquad + \nu(I(Y;\hat{Y}|W) - \alpha) && (5.15)\\
&= \inf_{\substack{W,\hat{X},\hat{Y}: \mathbb{E}[(X-\hat{X})^2]\le D \\ \mathbb{E}[(Y-\hat{Y})^2]\le D}} h(X,Y) - \nu\alpha + \nu(h(X|W) + h(Y|W))\\
&\qquad - h(X,Y|W) - \nu(h(X|W,\hat{X}) + h(Y|W,\hat{Y})) && (5.16)\\
&\ge h(X,Y) - \nu\alpha + \nu \inf_{W} h(X|W) + h(Y|W) - \frac{1}{\nu} h(X,Y|W)\\
&\qquad + \inf_{\substack{W,\hat{X},\hat{Y}: \mathbb{E}[(X-\hat{X})^2]\le D \\ \mathbb{E}[(Y-\hat{Y})^2]\le D}} -\nu(h(X|W,\hat{X}) + h(Y|W,\hat{Y})) && (5.17)\\
&\ge h(X,Y) - \nu\alpha + \nu\cdot \min_{0\preceq K'\preceq \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}} h(X') + h(Y') - \frac{1}{\nu} h(X',Y')\\
&\qquad + \nu\cdot\left( \min_{\substack{(W,\hat{X},\hat{Y})\in\mathcal{P}_G: \\ \mathbb{E}[(X-\hat{X})^2]\le D}} -h(X|W,\hat{X}) + \min_{\substack{(W,\hat{X},\hat{Y})\in\mathcal{P}_G: \\ \mathbb{E}[(Y-\hat{Y})^2]\le D}} -h(Y|W,\hat{Y}) \right) && (5.18)\\
&= h(X,Y) - \nu\alpha - \nu\log(2\pi e D)\\
&\qquad + \nu\cdot \min_{0\preceq K'\preceq \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}} h(X') + h(Y') - \frac{1}{\nu} h(X',Y') && (5.19)\\
&= \frac{1}{2}\log(2\pi e)^2(1-\rho^2) - \nu\alpha - \nu\log(2\pi e D)\\
&\qquad + \frac{\nu}{2}\log\frac{\nu^2}{2\nu-1} - \frac{1-\nu}{2}\log(2\pi e)^2\frac{(1-\rho)^2}{2\nu-1} && (5.20)\\
&= \begin{cases} \frac{1}{2}\log^+ \frac{1+\rho}{2De^\alpha+\rho-1}, & \text{if } 1-\rho \le De^\alpha \le 1\\ \frac{1}{2}\log^+ \frac{1-\rho^2}{D^2e^{2\alpha}}, & \text{if } De^\alpha \le 1-\rho. \end{cases} && (5.21)
\end{aligned}
$$

where (5.15) follows from weak duality for $\nu \ge 0$; (5.17) follows from bounding the infimum of the sum with the sum of the infima of its summands, and the fact that relaxing the constraints cannot increase the value of the infimum; (5.18) follows from Theorem 9 where $\nu := \frac{1}{1+\lambda}$ and for the constraint $0 \le \lambda < 1$ (indeed we can also include zero) to be satisfied we need $\frac{1}{2} < \nu \le 1$ and [9, Lemma 1] on each of the terms; (5.19) follows by observing

$$
\begin{aligned}
h(X|W,\hat{X}) &= h(X-\hat{X}|W,\hat{X}) && (5.22)\\
&\le h(X-\hat{X}) && (5.23)
\end{aligned}
$$

$$\leq \frac{1}{2}\log(2\pi e D), \tag{5.24}$$

where the last step is due to the fact that $\mathbb{E}[(X-\hat{X})^2] \leq D$; (5.20) follows from Lemma 11 for $\nu \geq \frac{1}{1+\rho}$; and (5.21) follows from maximizing

$$\ell(\nu) := \frac{1}{2}\log{(2\pi e)^2(1-\rho^2)} - \nu\alpha - \nu\log{(2\pi e D)} \tag{5.25}$$

$$+\frac{\nu}{2}\log\frac{\nu^2}{2\nu-1} - \frac{1-\nu}{2}\log{(2\pi e)^2\frac{(1-\rho)^2}{2\nu-1}}, \tag{5.26}$$

for $1 \geq \nu \geq \frac{1}{1+\rho}$. Now we need to choose the tightest bound $\max_{1\geq\nu\geq\frac{1}{1+\rho}} \ell(\nu)$. Note that the function $\ell$ is concave since

$$\frac{\partial^2 \ell}{\partial \nu^2} = -\frac{1}{\nu(2\nu-1)} < 0. \tag{5.27}$$

Since it also satisfies monotonicity

$$\frac{\partial \ell}{\partial \nu} = \log\frac{\nu(1-\rho)}{(2\nu-1)De^{\alpha}}, \tag{5.28}$$

its maximal value occurs when the derivative vanishes, that is, when $\nu_* = \frac{De^{\alpha}}{2De^{\alpha}-1+\rho}$. Substituting for the optimal $\nu_*$ we get

$$\mathrm{R}_{D,\alpha}(X,Y) \geq \ell\left(\frac{De^{\alpha}}{2De^{\alpha}-1+\rho}\right) = \frac{1}{2}\log^{+}\frac{1+\rho}{2De^{\alpha}-1+\rho}, \tag{5.29}$$

for $1 \geq \nu_* \geq \frac{1}{1+\rho}$, which means the expression is valid for $1-\rho \leq De^{\alpha} \leq 1$.

The other case is $De^{\alpha} \leq 1-\rho$. In this case note that $\nu(1-\rho) \geq \nu De^{\alpha} \geq (2\nu-1)De^{\alpha}$ for $\nu \leq 1$. This implies $\frac{\nu(1-\rho)}{(2\nu-1)De^{\alpha}} \geq 1$, thus we have $\frac{\partial \ell}{\partial \nu} \geq 0$. Since the function is concave and increasing the maximum is attained at $\nu_* = 1$, thus

$$\mathrm{R}_{D,\alpha}(X,Y) \geq \ell\,(1) = \frac{1}{2}\log^{+}\frac{1-\rho^2}{D^2e^{2\alpha}}, \tag{5.30}$$

where the expression is valid for $De^{\alpha} \leq 1-\rho$. As stated at the beginning of the proof, this is the correct formula assuming unit-variance sources. For sources of variance $\sigma^2$, it suffices to replace $D$ with $D/\sigma^2$, which leads to the expression given in the theorem statement.

The upper bound follows by plugging in jointly Gaussian random variables, that was derived in in [53, Theorem 4.3].

## 5.5 Conclusion

For the Gaussian lossy Gray-Wyner network under symmetric mean-squared error distortion, the rate region of the common rate versus the sum of the private rates, is fully characterized.

# 6 Lower Bound on (relaxed) Wyner's Common Information

## 6.1 Introduction

Extracting and assessing common features amongst multiple variables is a natural task occurring in many different problem settings. Wyner's common information in Definition 3 provides one answer to this, which was originally defined for finite alphabets. For a pair of random variables, it seeks to find the most compact third variable that makes the pair conditionally independent. Compactness is measured in terms of the mutual information between the pair and the third variable. In [34], Wyner also identifies two operational interpretations. The first concerns a source coding network often referred to as the Gray-Wyner network. For this scenario, Wyner's common information characterizes the smallest common rate required to enable two decoders to recover $X$ and $Y$, respectively, in a lossless fashion. The second operational interpretation pertains to the distributed simulation of common randomness. Here, Wyner's common information characterizes the smallest number of random bits that need to be shared between the processors. In subsequent work, Wyner's common information was extended to continuous random variables and was computed for a pair of Gaussian random variables [36, 37] and for a pair of additive "Gaussian channel" distributions [38]. Other related works include [42, 23]. Wyner's common information has many applications, including to communication networks [34], to caching [43, Section III.C], to source coding [44], and to feature extraction [54].

A natural extension of Wyner's common information is given in Definition 4. Different from Wyner's common information the constraint of conditional independence is relaxed into an upper bound on the conditional mutual information.

In this chapter, we derive a new lower bound [1] on relaxed Wyner's common information for continuous random variables. The proof is based on a method known as factorization of convex envelopes, which was originally introduced in [27]. The proof strategy is fundamentally different from the techniques that were used to

[1] The material of this chapter has appeared in
- E. Sula and M. Gastpar, "Lower bound on Wyner's common information," *CoRR*, vol. abs/2102.08157, 2021. [Online]. Available: https://arxiv.org/abs/2102.08157.

solve Wyner's original common information problem. Specifically, for the latter, the generic approach is to first characterize the class of variables that enable conditional independence, and then inside this class to find the optimal variable. By contrast, we lower bound the relaxed Wyner's common information problem by a convex problem, which we can then solve via optimizing.

We illustrate the promise of the new lower bound by considering Gaussian mixture distributions. We also establish that the new lower bound is tight for a simple case of the so-called "Gaussian channels" distribution. Here, $X$ and $Y$ can be written as the sum of a single arbitrary random variable and jointly Gaussian noises. We note that for this special case, Wyner's common information was previously found, using different methods, in [38].

### 6.1.1 Main Result

Here we present our lower bound on relaxed Wyner's common information. The relaxed Wyner's common information $C_\gamma(X;Y)$ is given in Definition 4. The bound is given in terms of the entropy of the pair, entropy and relaxed Wyner's common information for Gaussian random variables. The theorem says:

**Theorem 16.** *Let $(X,Y)$ have probability density functions $p_{(X,Y)}$ that satisfy the covariance matrix $K_{(X,Y)}$. Let, $(X_g,Y_g)\sim\mathcal{N}(0,K_{(X,Y)})$, then*

$$C_\gamma(X;Y)\geq\max\{C_\gamma(X_g;Y_g)+h(X,Y)-h(X_g,Y_g),0\},\tag{6.1}$$

*where*

$$C_\gamma(X_g;Y_g)=\frac{1}{2}\log^+\left(\frac{1+|\rho|}{1-|\rho|}\cdot\frac{1-\sqrt{1-e^{-2\gamma}}}{1+\sqrt{1-e^{-2\gamma}}}\right),\tag{6.2}$$

*and $\rho$ is the correlation coefficient between $X$ and $Y$.*

The proof is given in Section 6.3. The statement of the theorem is in accordance with max-entropy statement where the probability density functions have given covariances. Interestingly, once we plug in Gaussian random variables and additive "Gaussian channel" distributions, then the bound is attained with equality.

**Remark 5.** *For $\gamma=0$, in [34], it is showed that $C(X;Y)\geq I(X;Y)$. In Section 6.2 we show that our lower bound from Theorem 16 can be tighter.*

**Remark 6.** *For $\gamma=0$, we have*

$$C(X;Y)\geq\max\{C(X_g;Y_g)+h(X,Y)-h(X_g,Y_g),0\},\tag{6.3}$$

*and the same bound is derived in [55, Theorem 6.4.3] and [45, Theorem 1].*

**Remark 7.** *The bound of Theorem 16 can be expressed equivalently as*

$$C_\gamma(X;Y)\geq C_\gamma(X_g;Y_g)-D\left(p_{(X,Y)}\,\|p_{(X_g,Y_g)}\right).\tag{6.4}$$

**Remark 8.** *The bound of Theorem 16 can be negative (if not for the correction). If we choose $X$ and $Y$ to be independent, then $X_g$ and $Y_g$ will be independent as well. Thus, the bound in (6.4) becomes*

$$C_\gamma(X;Y) \geq -D\left(p_X \| p_{X_g}\right) - D\left(p_Y \| p_{Y_g}\right), \tag{6.5}$$

*that is a negative bound from the positivity of the Kullback-Leibler divergence.*

In the later section, we provide pairs of random variable and compute our lower bounds on Wyner's common information to verify the usefulness of the derived bound.

## 6.2 Additive "Gaussian Channel" Distributions

In this section, we consider the distributions that are described as follows. Let $(\hat{X}, \hat{Y})$ be a Gaussian distribution with mean zero and covariance matrix

$$K_{(\hat{X},\hat{Y})} = \begin{pmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{pmatrix}. \tag{6.6}$$

Then, we consider the two-dimensional source given by

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \hat{X} \\ \hat{Y} \end{pmatrix} + \begin{pmatrix} A \\ B \end{pmatrix}. \tag{6.7}$$

Let $(A, B)$ be arbitrary random variables with mean zero and covariance

$$K_{(A,B)} = \begin{pmatrix} \sigma_A^2 & r\sigma_A\sigma_B \\ r\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix}, \tag{6.8}$$

where $\sigma_A = \sigma_B$ and $(A, B)$ is independent of the pair $(\hat{X}, \hat{Y})$. For this particular distribution, we evaluate our lower bound in (6.1) and also provide an upper bound.

### 6.2.1 Lower Bound

We have that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ and

$$\mathbb{E}[X^2] = \mathbb{E}[\hat{X}^2] + \mathbb{E}[A^2] = 1 + \sigma_A^2, \tag{6.9}$$

$$\mathbb{E}[XY] = \mathbb{E}[\hat{X}\hat{Y}] + \mathbb{E}[AB] = \hat{\rho} + r\sigma_A^2. \tag{6.10}$$

By symmetry $\mathbb{E}[Y^2] = \mathbb{E}[X^2]$ and

$$\rho = \frac{\mathbb{E}[XY]}{\sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}} = \frac{\hat{\rho} + r\sigma_A^2}{1 + \sigma_A^2}. \tag{6.11}$$

Let $\rho \geq 0$ so the formula given in Theorem 16 evaluates to

$$\begin{aligned}
C_\gamma(X;Y) &\geq C_\gamma(X_g;Y_g) + h(X,Y) - h(X_g,Y_g) && (6.12)\\
&= \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\cdot\frac{1-\sqrt{1-e^{-2\gamma}}}{1+\sqrt{1-e^{-2\gamma}}}\right) + h(X,Y)\\
&\quad - \frac{1}{2}\log(2\pi e)^2\left((1+\sigma_A^2)^2 - (\hat{\rho}+r\sigma_A^2)^2\right) && (6.13)\\
&= h(X,Y) - \log\left(2\pi e(1-\hat{\rho}+(1-r)\sigma_A^2\right) + \frac{1}{2}\log\left(\frac{1-\sqrt{1-e^{-2\gamma}}}{1+\sqrt{1-e^{-2\gamma}}}\right). && (6.14)
\end{aligned}$$

where (6.13) follows from substituting for $K_{(X,Y)}$ and (6.14) follows from substituting for $\rho$ computed in (6.11).

### 6.2.2 Upper Bound

Next we give an upper bound on Wyner's common information for the example of this section. To accomplish this, rewrite the pair $(\hat{X},\hat{Y})$ as

$$\begin{aligned}
\hat{X} &= \sqrt{\alpha}V + Z_x,\\
\hat{Y} &= \sqrt{\alpha}V + Z_y, && (6.15)
\end{aligned}$$

where $0 \leq \alpha \leq \hat{\rho}$ where $V$ is independent of $(Z_x,Z_y)$, $V \sim \mathcal{N}(0,1)$ and $(Z_x,Z_y) \sim \mathcal{N}\left(0,\begin{pmatrix}1-\alpha & \hat{\rho}-\alpha\\ \hat{\rho}-\alpha & 1-\alpha\end{pmatrix}\right)$. Then, we select $W$ to be $W = (\sqrt{\alpha}V + A, \sqrt{\alpha}V + B)$ for $0 \leq \alpha \leq \hat{\rho}$. By combining (6.7) and (6.15) we can rewrite the pair $(X,Y)$ as

$$\begin{aligned}
X =& \sqrt{\alpha}V + A + Z_x,\\
Y =& \sqrt{\alpha}V + B + Z_y. && (6.16)
\end{aligned}$$

Let us compute the constraint, so we have

$$\begin{aligned}
I(X;Y|W) &= I(\sqrt{\alpha}V + A + Z_x; \sqrt{\alpha}V + B + Z_y|W) && (6.17)\\
&= I(Z_x;Z_y|W) && (6.18)\\
&= I(Z_x;Z_y) && (6.19)\\
&= \frac{1}{2}\log\frac{1}{1-\left(\frac{\hat{\rho}-\alpha}{1-\alpha}\right)^2} && (6.20)\\
&= \gamma, && (6.21)
\end{aligned}$$

where (6.18) follows by subtracting the parts that are in the conditioning by recalling that $W = (\sqrt{\alpha}V+A, \sqrt{\alpha}V+B)$, (6.19) follows from independence of $W$ and $(Z_x,Z_y)$, (6.20) follows from computation of the mutual information with Gaussians and (6.21) for

$$\alpha = \frac{\hat{\rho}-\sqrt{1-e^{-2\gamma}}}{1-\sqrt{1-e^{-2\gamma}}}. \quad (6.22)$$

Thus, the upper bound is

$$
\begin{aligned}
C_\gamma(X;Y) &\leq I(X,Y;W) && (6.23)\\
&= h(X,Y) - h(\sqrt{\alpha}V + A + Z_x, \sqrt{\alpha}V + B + Z_y|W) && (6.24)\\
&= h(X,Y) - h(Z_x, Z_y|W) && (6.25)\\
&= h(X,Y) - h(Z_x, Z_y) && (6.26)\\
&= h(X,Y) - \frac{1}{2}\log(2\pi e)^2\left((1-\alpha)^2 - (\hat{\rho}-\alpha)^2\right) && (6.27)\\
&= h(X,Y) - \frac{1}{2}\log(2\pi e)^2(1-\hat{\rho})^2\left(\frac{1-\sqrt{1-e^{-2\gamma}}}{1+\sqrt{1-e^{-2\gamma}}}\right). && (6.28)
\end{aligned}
$$

where (6.23) follows from the definition of $C_\gamma(X;Y)$ where $W$ satisfies $I(X;Y|W) = \gamma$, (6.24) follows by rewriting the mutual information, (6.25) follows from subtracting the parts that are in the conditioning, (6.26) follows from independence of $W$ and $(Z_x, Z_y)$, (6.27) follows from the Gaussian pair $(Z_x, Z_y)$ and (6.28) follows from substituting for $\alpha$ given in (6.22).

### 6.2.3 Example 1

Let us choose $(A, B)$ doubly symmetric binary distribution where $p_{(A,B)}(A = B = \sigma_A) = p_{(A,B)}(A = B = -\sigma_A) = \frac{1+r}{4}$ and $p_{(A,B)}(A = -B = \sigma_A) = p_{(A,B)}(A = -B = -\sigma_A) = \frac{1-r}{4}$. Note that for these choices, the covariance matrix of $A$ and $B$ is given by Equation (6.8). If we select $A = B$ or $r = 1$, this model is precisely the model studied in Example 1. A numerical evaluation is shown in Figure 6.1.

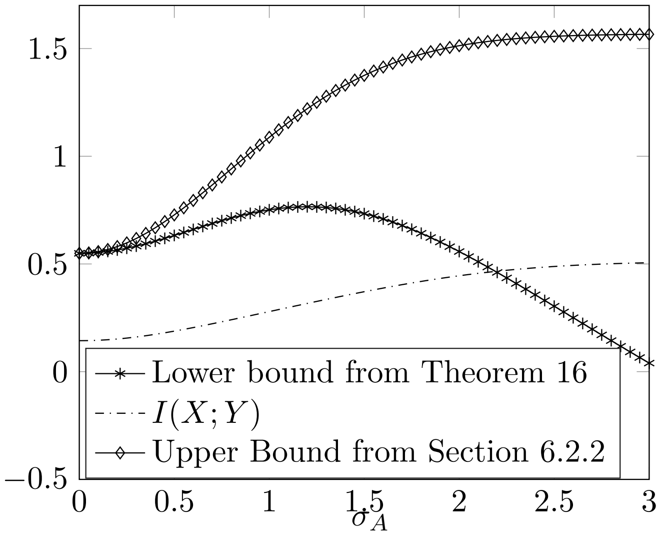


Figure 6.1 – The $*$-line is the lower bound on $C(X;Y)$ from Theorem 16 and the $\diamond$-line is the upper bound on $C(X;Y)$ from Section 6.2.2. The dashed line is the mutual information $I(X;Y)$. In this setup we plot the bounds on $C(X;Y)$ in nats versus $\sigma_A$ for $\hat{\rho} = 0.5$ and $r = 0.9$.

### 6.2.4 Example 2

**Lemma 13.** *For the additive "Gaussian channel" distributions described in (6.7) and $A = B$, we have*

$$C_\gamma(X;Y) = h(X,Y) - \frac{1}{2}\log(2\pi e)^2(1-\hat{\rho})^2\left(\frac{1-\sqrt{1-e^{-2\gamma}}}{1+\sqrt{1-e^{-2\gamma}}}\right). \tag{6.29}$$

The proof follows from the fact that the lower bound (6.14) and upper bound (6.28) coincide when $A = B$, which means $r = 1$. For $\gamma = 0$, the same result is derived by a different approach in [38]. In Figure 6.2, we illustrate Lemma 13, for $A$ binary $\pm\sigma_A$ with uniform probability.

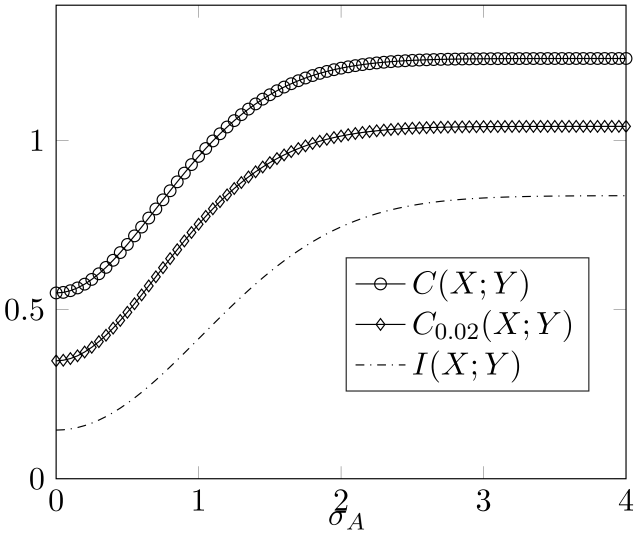


Figure 6.2 – The o-line is the Wyner's common information $C(X;Y)$ and the ⋄-line is the relaxed Wyner's common information $C_{0.02}(X;Y)$ for the specified Gaussian mixture distribution. The dashed line is the mutual information $I(X;Y)$. In this setup we plot the three curves in nats versus $\sigma_A$ for $\hat{\rho} = 0.5$.

## 6.3 Proof of lower bound on relaxed Wyner's common information

Note that the mean of the random variables does not affect the Wyner's common information and its relaxed variant thus, we assume mean zero for both $X$ and $Y$. Also, the relaxed Wyner's common information is invariant to scaling of $X$ and $Y$. Thus, without loss of generality we assume $X$ and $Y$ to be mean zero, unit variance and correlation coefficient $\rho$, so we proceed as follows

$$\begin{aligned}
&C_\gamma(X;Y)\\
&= \inf_{W:I(X;Y|W)\le\gamma} I(X,Y;W) &(6.30)\\
&\ge \inf_W (1+\mu)I(X,Y;W) - \mu I(X;W) - \mu I(Y;W) + \mu I(X;Y) - \mu\gamma &(6.31)\\
&= \mu \inf_W h(X|W) + h(Y|W) - (1+\frac{1}{\mu})h(X,Y|W) + h(X,Y) - \mu\gamma &(6.32)
\end{aligned}$$

$$\geq \mu \min_{K':0 \preceq K' \preceq \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}} h(X') + h(Y') - (1 + \frac{1}{\mu}) h(X', Y') + h(X, Y) - \mu\gamma \tag{6.33}$$

$$\geq h(X, Y) + \frac{\mu}{2} \log \frac{\mu^2}{\mu^2 - 1} - \frac{1}{2} \log (2\pi e)^2 \frac{(1-\rho)^2(\mu+1)}{\mu - 1} - \mu\gamma \tag{6.34}$$

$$\geq h(X, Y) - h(X_g, Y_g) + C(X_g; Y_g) \tag{6.35}$$

where (6.31) follows from weak duality and the bound is valid for all $\mu \geq 0$; (6.32) follows from simplification; (6.33) follows from Theorem 9 under the assumption that $\mu > 1$ where $(X', Y') \sim \mathcal{N}(0, K')$; (6.34) follows from Lemma 11 under the assumption $\mu \geq \frac{1}{\rho}$ and (6.35) follows by maximizing the function

$$g(\mu) = h(X, Y) - \mu\gamma + \frac{\mu}{2} \log \frac{\mu^2}{\mu^2 - 1} - \frac{1}{2} \log (2\pi e)^2 \frac{(1-\rho)^2(\mu+1)}{\mu - 1}, \tag{6.36}$$

for $\mu \geq \frac{1}{\rho}$. Now we need to solve $\max_{\mu \geq \frac{1}{\rho}} g(\mu)$. The function $g$ is concave in $\mu$,

$$\frac{\partial^2 g}{\partial \mu^2} = -\frac{1}{\mu(\mu^2 - 1)} < 0, \tag{6.37}$$

and by studying the monotonicity we obtain

$$\frac{\partial g}{\partial \mu} = -\frac{1}{2} \log \frac{\mu^2 - 1}{\mu^2} - \gamma. \tag{6.38}$$

Since the function is concave, the maximum is attained when the first derivative vanishes. That leads to the optimal solution $\mu_* = \frac{1}{\sqrt{1 - e^{-2\gamma}}}$, where $\mu_*$ has to satisfy $\mu_* \geq \frac{1}{\rho}$. Substituting for the optimal solution we get

$$C_\gamma(X; Y) \geq g\left(\frac{1}{\sqrt{1 - e^{-2\gamma}}}\right) \tag{6.39}$$

$$= h(X, Y) - h(X_g, Y_g) + C_\gamma(X_g; Y_g). \tag{6.40}$$

## 6.4 Vector Wyner's common information

It is well-known that for $n$ independent pairs of random variables, we have

$$C(X^n; Y^n) = \sum_{i=1}^{n} C(X_i; Y_i). \tag{6.41}$$

For the proof see [56, Lemma 2] by letting $\gamma = 0$.

By making use of Theorem 16 and (6.41) we can lower bound the Wyner's common information for $n$ independent pairs of random variables as

$$C(X^n; Y^n) \geq \sum_{i=1}^{n} C(X_{g_i}; Y_{g_i}) + h(X_i, Y_i) - h(X_{g_i}, Y_{g_i}). \tag{6.42}$$

An interesting problem is finding a bound for arbitrary $(X^n, Y^n)$, for any dependencies between $X^n$ and $Y^n$. This is not studied here and is left for future investigation.

# Common Information Components Analysis 7

## 7.1 Introduction

Understanding relations between two (or more) sets of variates is key to many tasks in data analysis and beyond. To approach this problem, it is natural to reduce each of the sets of variates separately in such a way that the reduced descriptions fully capture the *commonality* between the two sets, while suppressing aspects that are individual to each of the sets. This permits to understand the relation between the two sets without obfuscation. A popular framework to accomplish this task follows the classical viewpoint of *dimensionality reduction* and is referred to as Canonical Correlation Analysis (CCA) [57]. CCA seeks the best *linear* extraction, i.e., we consider linear projections of the original variates. Via the so-called Kernel trick, this can be extended to cover arbitrary (fixed) function classes. Motivated from CCA, we introduce a novel method of *feature extraction* namely common information components analysis (CICA)[1]

Wyner's common information is a well-known and established measure of the dependence of two random variables. Intuitively, it seeks to extract a third random variable such that the two random variables are conditionally independent given the third, but at the same time, that third variable is as compact as possible. Compactness is measured in terms of the mutual information that the third random variable retains about the original two. The resulting optimization problem is not a convex problem (because the constraint set is not a convex set), and therefore, not surprisingly, closed-form solutions are rare. A natural generalization of Wyner's common information is obtained by replacing the constraint of conditional independence by a limit on the conditional mutual information. If the limit is set equal to zero, we return precisely to the case of conditional independence. Exactly like mutual information, Wyner's common information and its generalization are endowed

[1] The material of this chapter has appeared in
- M. Gastpar and E. Sula, "Common information components analysis," in *Proceedings of the 2020 Information Theory and Applications (ITA) Workshop*, San Diego, USA, February 2020.
- E. Sula and M. Gastpar, "Common information components analysis," *Entropy Special Issue on The Role of Signal Processing and Information Theory in Modern Machine Learning*, vol. 23, no. 2, 2021.

with a clear operational meaning. They characterize the fundamental limits of data compression (in the Shannon sense) for a certain network situation.

### 7.1.1 Related Work

Connections between CCA and Wyner's common information have been explored in the past. It is well known that for Gaussian vectors, (standard, non-relaxed) Wyner's common information is attained by all of the CCA components together, see [44]. This has been further interpreted, see e.g. [58]. Needless to say, having all of the CCA components together essentially amounts to a one-to-one transform of the original data into a new basis. It does not yet capture the idea of feature extraction or dimensionality reduction. To put our work into context, it is only the *relaxation* of Wyner's common information [59, 56] that permits to conceptualize the sequential, one-by-one recovery of the CCA components, and thus, the spirit of dimensionality reduction.

CCA also appears in a number of other problems related to information measures and probabilistic models. For example, in the so-called Gaussian information bottleneck problem, the optimizing solution can be expressed in terms of the CCA components [60], and an interpretation of CCA as a (Gaussian) probabilistic model was presented in [61].

Generalizations of CCA have appeared before in the literature. The most prominent is built around maximal correlation. Here, one seeks arbitrary remappings of the original data in such a way as to maximize their correlation coefficient. This perspective culminates in the well-known *alternating conditional expectation* (ACE) algorithm [62].

Feature extraction and dimensionality reduction have a vast literature attached to them and it is beyond the scope of this chapter to provide a comprehensive overview. In a part of that literature, information measures play a key role. Prominent examples are independent components analysis (ICA) [63] and the information bottleneck [64, 65], amongst others. More recently, feature extraction alternations via information theory are presented in [66, 67]. In [66] the estimation of Rényi's quadratic entropy is studied, whereas in [67] standard information theoretic measures such as Kullback-Leibler divergence are used for fault diagnosis. Other slightly related feature extraction methods that perform dimensionality reduction on a single dataset include [68, 69, 70, 71, 72, 73, 74]. More concretely, in [68] a sparse Support Vector Machine (SVM) approach is used for feature extraction. In [69] feature extraction is performed via regression by using curvilinearity instead of linearity. In [70] compressed sensing is used to extract features when the data has a sparse representation. In [71], an invariant mapping method is invoked to map the high dimensional data to low dimensional data that is based on a neighbourhood relation. In [72] feature extraction is performed for a partial learning of the geometry of the manifold. In [73] distance correlation measure (a measure with similar properties as the regular Pearson correlation coefficient) is proposed as a new feature extraction method. In [74] kernel principal component analysis is used to perform feature extraction and allow for the extraction of non-linearities. In [75] feature extraction is done by a robust regression based approach and in [76] a linear regression approach is used to extract features.

### 7.1.2 Contributions

The contributions of our work are the following:

- We introduce a novel suit of algorithms, referred to as CICA. These algorithms are characterized by a two-step procedure. In the first step, a relaxation of Wyner's common information is extracted. The second step can be interpreted as a form of projection of the common information back onto the original data so as to obtain the respective features. A free parameter $\gamma$ is introduced to control the complexity of the extracted features.

- We establish that for the special case where the original data are jointly Gaussian, our algorithms precisely extract the CCA components. In this case, the parameter $\gamma$ determines how many of the CCA components are extracted. In this sense, we establish a new rigorous connection between information measures and CCA.

- We present initial results on how to extend CICA to more than two variates.

- Via a number of paradigmatic examples, we illustrate that for *discrete data,* CICA gives intuitively pleasing results while other methods, including CCA, do not. This is most pronounced in a simple example with three sources described in Section 7.7.1.

### 7.1.3 A Simple Example with Synthetic Data

To set the stage and in guise of an informal problem statement, let us consider a simple example involving synthetic data. Specifically, we consider two-dimensional data, that is, the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are of length 2. The goal is to extract, separately from each of the two, a one-dimensional description in such a way as to extract the commonality between $\boldsymbol{X}$ and $\boldsymbol{Y}$ while suppressing their individual features. For simplicity, in the present artificial example, we will assume that the entries of the vectors only take value in a small finite set, namely, $\{0, 1, 2, 3\}$. To illustrate the point, we consider the following special statistical model:

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} U \oplus X_2 \\ X_2 \end{pmatrix}, \tag{7.1}$$

and

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} U \oplus Y_2 \\ Y_2 \end{pmatrix}, \tag{7.2}$$

where $U, X_2$, and $Y_2$ are mutually independent uniform random variables over the set $\{0, 1, 2, 3\}$ and $\oplus$ denotes addition modulo 4.

The reason for this special statistical structure is so that it is obvious what should be extracted, namely, $\boldsymbol{X}$ should be reduced to $U$, and $\boldsymbol{Y}$ should also be reduced to $U$. This reduces both $\boldsymbol{X}$ and $\boldsymbol{Y}$ to "one-dimensional" descriptions, and these one-dimensional descriptions capture precisely the dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$. In this simple example, all the commonality between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is captured by $U$. More formally, conditioned on $U$, the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent.

The interesting observation of this example is that any *pair* of components of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *independent* of each other, such as, for example $X_1$ and $Y_1$. Therefore, the joint covariance matrix of the merged vector $(\boldsymbol{X}, \boldsymbol{Y})$ is a scaled identity matrix. This implies that any method that only uses the covariance matrix as input, including CCA, cannot find any commonalities between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in this example.

By contrast, the algorithmic procedure discussed in the present chapter will correctly extract the desired answer. In Figure 7.1, we show numerical simulation outcomes for a couple of approaches. Specifically, in *(a),* we can see that in this particular example, CCA fails to extract the common features. This, of course, was done on purpose: For the synthetic data at hand, the global covariance matrix is merely a scaled identity matrix, and since CCA's only input is the covariance matrix, it does not actually do anything in this example. In *(b),* we show the performance of the approximate gradient-descent based implementation of the CICA algorithm proposed in this chapter, as detailed in Section 7.6. In this simple example, this precisely coincides with the ideal theoretical performance of CICA as in Generic Procedure 1, but in general, the gradient-descent based implementation is not guaranteed to find the ideal solution.

At this point, we should stress that for such a simple example, many other approaches would also lead to the same, correct answer. One of them is maximal correlation. In that perspective, one seeks to separately reduce $\boldsymbol{X}$ and $\boldsymbol{Y}$ by applying possibly non-linear functions $f(\cdot)$ and $g(\cdot)$ in such a way as to maximize the correlation between $f(\boldsymbol{X})$ and $g(\boldsymbol{Y})$. Clearly, for the simple example at hand, selecting $f(\boldsymbol{X}) = X_1 \oplus X_2$ and $g(\boldsymbol{Y}) = Y_1 \oplus Y_2$ leads to correlation one, and is thus a maximizer.

Finally, the present example is also too simplistic to express the finer information-theoretic structure of the problem. One step up is the example presented in Section 7.5 below, where the commonality between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is not merely an equality (the component $U$ above), but rather a probabilistic dependency.

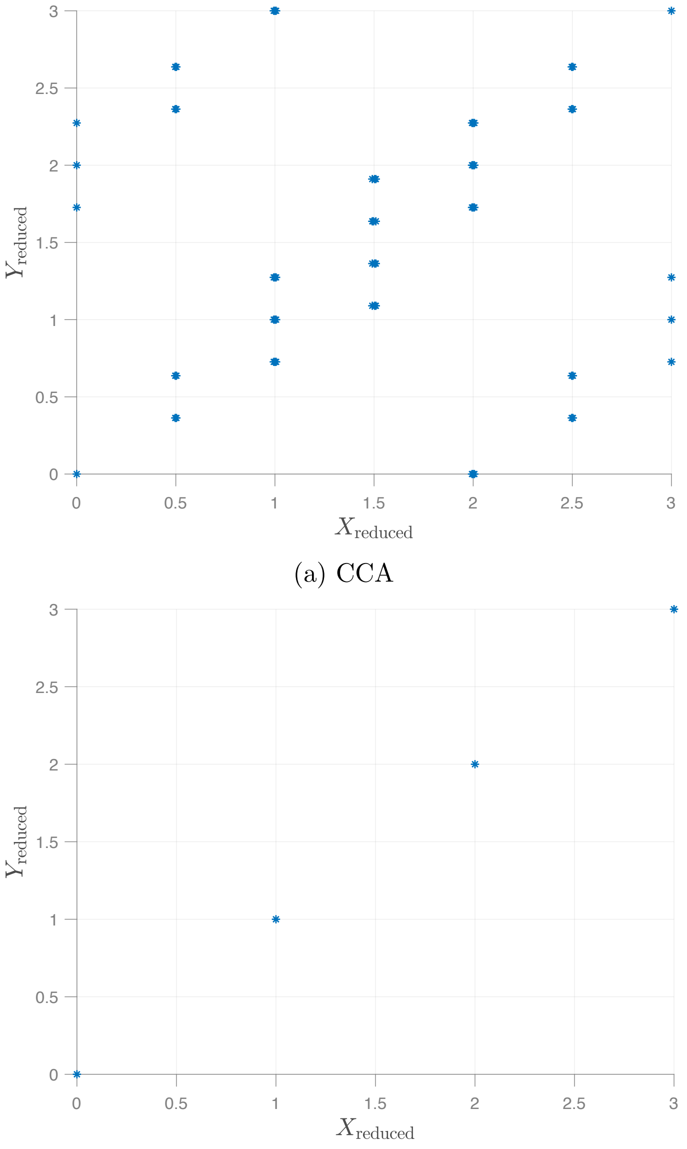
## 7.2 Wyner's Common Information and Its Relaxation

The main framework and underpinning of the proposed algorithm is Wyner's common information given in Definition 3 and its extension given in Definition 4.

It is important to observe that the Wyner's common information is not a convex problem. First, we observe that $I(X, Y; W)$ is indeed a convex function of $p(w|x, y)$, which is a well-known fact, see e.g. [4, Theorem 2.7.4]. The issue is with the constraint set. The set of distributions $p(w|x, y)$ for which $I(X; Y|W) \leq \gamma$ is not a convex set. To provide some intuition for the structure of this set, let us consider $I(X; Y|W)$ as a function of $p(w|x, y)$, and examine its (non-)convexity. The relation between the two is described by the epigraph

$$\text{epigraph}\{I(X;Y|W)\} = \{(p(w|x,y), \gamma) : p(w|x,y) \in \mathcal{P}, \gamma \geq I(X;Y|W)\}. \quad (7.3)$$

The function $I(X; Y|W)$ is convex in $p(w|x, y)$ if and only if its epigraph is a convex set which would imply that the set of distributions $p(w|x, y)$ for which $I(X; Y|W) \leq \gamma$ is also convex. Now we present an example that $I(X; Y|W)$ is not a convex function of $p(w|x, y)$.



(a) CCA

(b) CICA

Figure 7.1 – The situation for the synthetic data as described in example described in Section 7.1.3. Figure *7.1a* shows the scatterplot for two one-dimensional features extracted by CCA. Apparently, the approach is not able to extract the commonality between the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ in this synthetic example. Figure *7.1b* shows the performance of the heuristic algorithm of CICA described in Section 7.6, which in this simple example ends up matching the ideal theoretical performance of CICA as in Generic Procedure 1 for $n = 10^5$ data samples.

**Example 3.** *Let the distributions* $p(x, y), p_1(w|x, y), p_2(w|x, y)$ *be*

$$p(x, y) = \begin{bmatrix} \frac{2}{5} & \frac{1}{10} \\ \frac{1}{10} & \frac{2}{5} \end{bmatrix}, p_1(w|x, y) = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{3}{4} & \frac{3}{4} & \frac{3}{4} & \frac{3}{4} \end{bmatrix}, p_2(w|x, y) = \begin{bmatrix} \frac{1}{2} & \frac{3}{4} & \frac{3}{4} & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}, \tag{7.4}$$

*respectively. For this example, one can evaluate numerically that under* $p_1(w|x,y)$*, we have* $I_{p_1}(X;Y|W) < 0.279$ *and under* $p_2(w|x,y)$*, we have* $I_{p_2}(X;Y|W) < 0.262$*. By the same token, one can show that under* $(p_1(w|x,y) + p_2(w|x,y))/2$*, we have* $I_{(p_1+p_2)/2}(X;Y|W) > 0.274$*. Hence, we conclude that for this example,*

$$I_{(p_1+p_2)/2}(X;Y|W) > \frac{1}{2}\left(I_{p_1}(X;Y|W) + I_{p_2}(X;Y|W)\right), \tag{7.5}$$

*which proves that* $I(X;Y|W)$ *cannot be convex.*

## 7.3 The Algorithm

The main technical result of this chapter is to establish that the outcome of a specific procedure induced by the relaxed Wyner's common information is tantamount to CCA whenever the original underlying distribution is Gaussian. In preparation for this, in this section, we present the proposed algorithm. In doing so, we will assume that the distribution of the data is $p(\boldsymbol{x},\boldsymbol{y})$. In many applications involving CCA, the data distribution may not be known, but rather, a number of samples of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are provided, based on which CCA would then estimate the covariance matrix. A similar perspective can be taken on our procedure, but is left for future work. A short discussion can be found in Section 7.8 below.

### 7.3.1 High-level Description

The proposed algorithm takes as input the distribution $p(\boldsymbol{x},\boldsymbol{y})$ of the data, as well as a level $\gamma$. The level $\gamma$ is a non-negative real number and may be thought of as a resolution level or a measure of coarseness: If $\gamma = 0$, then the full commonality (or common information) between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is extracted in the sense that conditioned on the common information, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent. Conversely, if $\gamma$ is large, then only the most important part of the commonality is extracted. Fixing the level $\gamma$, the idea of the proposed algorithm is to evaluate the relaxed Wyner's Common Information of Equation (4.3) between the information sources (data sets) at the chosen level $\gamma$. This evaluation will come with an associated conditional distribution $p_\gamma(w|x,y)$, namely, the conditional distribution attaining the minimum in the optimization problem of Equation (4.3). The second half of the proposed algorithm consists in leveraging the minimizing $p_\gamma(w|x,y)$ in such a way as to separately reduce $\boldsymbol{X}$ and $\boldsymbol{Y}$ to those features that best express the commonality. This may be thought of as a type of projection of the minimizing random variable $W$ back onto $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. For the case of Gaussian statistics, this can be made precise.

### 7.3.2 Main Steps of the Algorithm

The algorithm proposed here starts from the joint distribution of the data, $p(\boldsymbol{x},\boldsymbol{y})$. Estimates of this distribution can be obtained from data samples $X^n$ and $Y^n$ via standard techniques. The main steps of the procedure can then be described as follows:

**Generic Procedure 1** (CICA). *1. Select a real number* $\gamma$*, where* $0 \leq \gamma \leq I(\boldsymbol{X};\boldsymbol{Y})$*. This is the compression level: A low value of* $\gamma$ *represents low compression,* *and thus, many components are retained. A high value of $\gamma$ represents high compression, and thus, only a small number of components are retained.*

2. *Solve the relaxed Wyner's common information problem,*

$$\min_{p(w|\boldsymbol{x},\boldsymbol{y})} I(\boldsymbol{X},\boldsymbol{Y};W) \text{ such that } I(\boldsymbol{X};\boldsymbol{Y}|W) \leq \gamma, \tag{7.6}$$

*leading to an associated conditional distribution $p_\gamma(w|\boldsymbol{x},\boldsymbol{y})$.*

3. *Using the conditional distribution $p_\gamma(w|\boldsymbol{x},\boldsymbol{y})$ found in Step 2), the dimension-reduced data sets can now be found via one of the following three variants:*

   *a) Version 1: MAP (maximum a posteriori):*

$$u(\boldsymbol{x}) = \arg\max_w p_\gamma(w|\boldsymbol{x}), \tag{7.7}$$

$$v(\boldsymbol{y}) = \arg\max_w p_\gamma(w|\boldsymbol{y}). \tag{7.8}$$

   *b) Version 2: Conditional Expectation:*

$$u(\boldsymbol{x}) = \mathbb{E}[W|\boldsymbol{X} = \boldsymbol{x}], \tag{7.9}$$

$$v(\boldsymbol{y}) = \mathbb{E}[W|\boldsymbol{Y} = \boldsymbol{y}]. \tag{7.10}$$

   *c) Version 3: Marginal Integration:*

$$u(\boldsymbol{x}) = \int_{\boldsymbol{y}} p(\boldsymbol{y})\mathbb{E}[W|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}]d\boldsymbol{y}, \tag{7.11}$$

$$v(\boldsymbol{y}) = \int_{\boldsymbol{x}} p(\boldsymbol{x})\mathbb{E}[W|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}]d\boldsymbol{x}. \tag{7.12}$$

The present chaper focuses on the three versions given here because for these three versions, we can establish Theorem 17, showing that in the case of Gaussian statistics, all three versions lead exactly to CCA. Second, we note that for concrete examples, it is often evident which of the versions is preferable. For example, in Section 7.5, we consider a binary example where the associated $W$ in Step 2 of our algorithm is also binary. In this case, Version 1 will reduce the original binary vector $\boldsymbol{X}$ to a binary scalar, which is perhaps the most desirable outcome. By contrast, Versions 2 and 3 require an explicit embedding of the binary example in the reals, and will reduce the original binary vector $\boldsymbol{X}$ to a real-valued scalar, which might not be as insightful.

## 7.4 For Gaussian, CICA is CCA

In this section, we consider the special case where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian random vectors. Since the mean has no bearing on either CCA or Wyner's common information, we will assume it to be zero in the sequel, without loss of generality. One key ingredient for this argument is a well-known change of basis, see for example [44], which we will now introduce in detail. Note that the mean will not change any mutual information term, thus we assume it to be zero without loss of generality. We first need to introduce notation for CCA. To this end, let us express the covariance matrices, as usual, in terms of their eigendecompositions as

$$K_{\boldsymbol{X}} = Q_x \begin{pmatrix} \Lambda_{r_X} & 0 \\ 0 & 0_{n-r_X} \end{pmatrix} Q_x^T \tag{7.13}$$

and

$$K_{\boldsymbol{Y}} = Q_y \begin{pmatrix} \Lambda_{r_Y} & 0 \\ 0 & 0_{n-r_Y} \end{pmatrix} Q_y^T, \tag{7.14}$$

where $r_X$ and $r_Y$ denote the rank of $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$, respectively. Starting from this, we define the matrices

$$K_{\boldsymbol{X}}^{-1/2} = Q_x \begin{pmatrix} \Lambda_{r_X}^{-1/2} & 0 \\ 0 & 0_{n-r_X} \end{pmatrix} Q_x^T \tag{7.15}$$

and

$$K_{\boldsymbol{Y}}^{-1/2} = Q_y \begin{pmatrix} \Lambda_{r_Y}^{-1/2} & 0 \\ 0 & 0_{n-r_Y} \end{pmatrix} Q_y^T, \tag{7.16}$$

where for a diagonal matrix $\Lambda$ with strictly positive entries, $\Lambda_{r_Y}^{-1/2}$ denotes the diagonal matrix whose diagonal entries are the reciprocals of the square roots of the entries of the matrix $\Lambda$. Using these matrices, we apply the following linear transformation

$$\hat{\boldsymbol{X}} = K_{\boldsymbol{X}}^{-1/2}\boldsymbol{X} \tag{7.17}$$

$$\hat{\boldsymbol{Y}} = K_{\boldsymbol{Y}}^{-1/2}\boldsymbol{Y}. \tag{7.18}$$

In the new coordinates, the covariance matrices of $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{Y}}$, respectively, can be shown to be

$$K_{\hat{\boldsymbol{X}}} = \begin{pmatrix} I_{r_X} & 0 \\ 0 & 0_{n-r_X} \end{pmatrix} \tag{7.19}$$

and

$$K_{\hat{\boldsymbol{Y}}} = \begin{pmatrix} I_{r_Y} & 0 \\ 0 & 0_{n-r_Y} \end{pmatrix}. \tag{7.20}$$

Moreover, we have

$$K_{\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}} = K_{\boldsymbol{X}}^{-1/2} K_{\boldsymbol{X}\boldsymbol{Y}} K_{\boldsymbol{Y}}^{-1/2}. \tag{7.21}$$

Let us denote the singular value decomposition of this matrix by

$$K_{\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}} = U\Sigma V^H. \tag{7.22}$$

where $\Sigma$ contains, on its diagonal, the ordered singular values of this matrix, denoted by $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_n$. Also, let us define

$$\tilde{\boldsymbol{X}} = U^H \hat{\boldsymbol{X}} \tag{7.23}$$

$$\tilde{\boldsymbol{Y}} = V^H \hat{\boldsymbol{Y}}, \tag{7.24}$$

which implies that $K_{\tilde{\boldsymbol{X}}} = K_{\hat{\boldsymbol{X}}}$, $K_{\tilde{\boldsymbol{Y}}} = K_{\hat{\boldsymbol{Y}}}$, and $K_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}}} = \Sigma$.

Next, we will leverage this change of basis to establish Wyner's common information and its relaxation for the Gaussian vector case, and then to prove the connection between Generic Procedure 1 and CCA.

### 7.4.1 Wyner's Common Information and Its Relaxation in the Gaussian case

For the case where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian random vectors, a full and explicit solution to the optimization problem of Wyner's common information defined in Equation 4.3 is found in [56]. To give some high-level intuition, the proof starts by mapping from $\boldsymbol{X}$ to $\tilde{\boldsymbol{X}}$ and from $\boldsymbol{Y}$ to $\tilde{\boldsymbol{Y}}$, as in Equations (7.23)-(7.24). This preserves all mutual information expressions as well as joint Gaussianity. Moreover, due to the structure of the covariance matrices of the vectors $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$, we have that $\{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^n$ are $n$ independent pairs of Gaussian random variables. Thus, by the tensorization property (see Lemma 10), the vector problem can be reduced to $n$ parallel scalar problems. The solution of the scalar problem is the main technical contribution of [56], and we refer to that paper for the detailed proof.

### 7.4.2 CICA in the Gaussian case and the exact connection with CCA

In this section, we consider the proposed CICA algorithm in the special case where the data distribution is $p(\boldsymbol{x}, \boldsymbol{y})$, a (multivariate) Gaussian distribution. We establish that in this case, the classic CCA is a solution to all versions of the proposed CICA algorithm. In this sense, CICA is a strict generalization of CCA. CCA is briefly reviewed in Appendix 7.9.1. Leveraging the matrices $U$ and $V$ defined via the singular value decomposition in Equation (7.22), CCA performs the dimensonality reduction

$$u(\boldsymbol{x}) = U_k^H \hat{\boldsymbol{x}} = U_k^H K_{\boldsymbol{X}}^{-1/2} \boldsymbol{x} \tag{7.25}$$

$$v(\boldsymbol{y}) = V_k^H \hat{\boldsymbol{y}} = V_k^H K_{\boldsymbol{Y}}^{-1/2} \boldsymbol{y}, \tag{7.26}$$

where the matrix $U_k$ contains the first $k$ columns of $U$ (that is, the $k$ left singular vectors corresponding to the largest singular values), and the matrix $V_k$ the respective right singular vectors. We refer to these as the "top $k$ CCA components."

**Theorem 17.** *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be jointly Gaussian random vectors. Then:*

1. *The top $k$ CCA components are a solution to* **all three** *versions of Generic Procedure 1.*

2. *The parameter $\gamma$ controls the number $k$ as follows:*

$$k(\gamma) = \begin{cases} n, & \text{if } 0 \le \gamma < ng(\rho_n), \\ n-1, & \text{if } ng(\rho_n) \le \gamma < (n-1)g(\rho_{n-1}) + g(\rho_n), \\ n-2, & \text{if } (n-1)g(\rho_{n-1}) + g(\rho_n) \le \gamma \\ & \quad < (n-2)g(\rho_{n-2}) + g(\rho_{n-1}) + g(\rho_n), \\ \vdots, & \vdots, \\ \ell & \text{if } (\ell+1)g(\rho_{\ell+1}) + \sum_{i=\ell+2}^n g(\rho_i) \le \gamma \\ & \quad < \ell g(\rho_\ell) + \sum_{i=\ell+1}^n g(\rho_i), \\ \vdots, & \vdots, \\ 1, & \text{if } 2g(\rho_2) + \sum_{i=2}^n g(\rho_i) \le \gamma < \sum_{i=1}^n g(\rho_i), \\ 0, & \text{if } \sum_{i=1}^n g(\rho_i) \le \gamma, \end{cases} \tag{7.27}$$

*where $g(\rho) = \frac{1}{2} \log \frac{1}{1-\rho^2}$.*

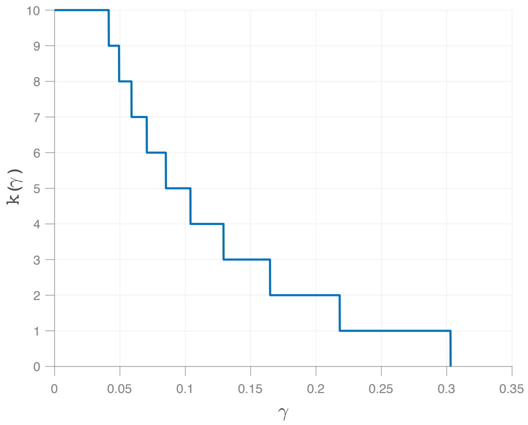


Figure 7.2 – Illustration of the function $k(\gamma)$ from Theorem 17 for the concrete case where $\boldsymbol{X}$ and $\boldsymbol{Y}$ have $n = 10$ components each and the correlation coefficients are $\rho_m = 1/(m+1)$.

**Remark 9.** *Note that $k(\gamma)$ is a decreasing, integer-valued function. An illustration for a special case is given in Figure 7.2.*

*Proof.* The main contribution of the theorem is the first item, *i.e.,*the connection between CCA and Generic Procedure 1 in the case where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian. The proof follows along the steps of the CICA procedure: We first show that in Step 2, when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian, then the minimizing $W$ may be taken jointly Gaussian with $\boldsymbol{X}$ and $\boldsymbol{Y}$. Then, we establish that in Step 3, with the $W$ from Step 2, we indeed obtain that the dimension-reduced representations $u(\boldsymbol{x})$ and $v(\boldsymbol{y})$ turn into the top $k$ CCA components. In detail:

*Step 2 of Generic Procedure 1:* The technical heavy lifting for this step in the case where $p(\boldsymbol{x}, \boldsymbol{y})$ is a multivariate Gaussian distribution is presented in [56]. We shall briefly summarize it here. In the case of Gaussian vectors, the solution to the optimization problem in Equation (4.3) is most easily described in two steps. First, we apply the change of basis indicated in Equations (7.17)-(7.18). This is a one-to-one transform, leaving all information expressions in Equation (4.3) unchanged. In the new basis, we have $n$ independent pairs. By the tensorization property (see Lemma 10), when $\boldsymbol{X}$ and $\boldsymbol{Y}$ consist of independent pairs, the solution to the optimization problem in Equation (4.3) can be reduced to $n$ separate scalar optimizations. The remaining crux then is solving the scalar Gaussian version of the optimization problem in Equation (4.3). This is done in [56, Theorem 3] via an argument of factorization of convex envelope. The full solution to the optimization problem is given in Equation (4.27)-(4.28). The remaining allocation problem over the non-negative numbers $\gamma_i$ can be shown to lead to a water-filling solution, given in [56, Theorem 8]. More explicitly, to understand this solution, start by setting $\gamma = I(\boldsymbol{X}; \boldsymbol{Y})$. Then, the corresponding $C_\gamma(\boldsymbol{X}; \boldsymbol{Y}) = 0$ and the optimizing distribution $p_\gamma(w|\boldsymbol{x}, \boldsymbol{y})$ trivializes. Now, as we lower $\gamma$, the various terms in the sum in Equation (4.27) start to become non-zero, starting with the term with the largest correlation coefficient $\rho_1$. Hence, an optimizing distribution $p_\gamma(w|\boldsymbol{x}, \boldsymbol{y})$ can be expressed as $\boldsymbol{W}_\gamma = U_k^H K_{\boldsymbol{X}}^{-1/2} \boldsymbol{X} + V_k^H K_{\boldsymbol{Y}}^{-1/2} \boldsymbol{Y} + \boldsymbol{Z}$, where the matrices $U_k$ and $V_k$ are precisely the top $k$ CCA components (see Equations (7.25)-(7.26) and the following discussion), and $\boldsymbol{Z}$ is additive Gaussian noise with mean zero, independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

*Step 3 of Generic Procedure 1:*For the algorithm, we need the corresponding conditional marginals, $p_\gamma(w|\boldsymbol{x})$ and $p_\gamma(w|\boldsymbol{y})$. By symmetry, it suffices to prove one formula. Changing basis as in Equations (7.17)-(7.18), we can write

$$
\begin{aligned}
\mathbb{E}[W|\boldsymbol{X}] &= \mathbb{E}[U_k^H \hat{\boldsymbol{X}} + V_k^H \hat{\boldsymbol{Y}} + \boldsymbol{Z}|\hat{\boldsymbol{X}}] && (7.28)\\
&= U_k^H \hat{\boldsymbol{X}} + V_k^H \mathbb{E}[\hat{\boldsymbol{Y}}|\hat{\boldsymbol{X}}] && (7.29)\\
&= U_k^H \hat{\boldsymbol{X}} + V_k^H \left( \mathbb{E}[\hat{\boldsymbol{Y}}\hat{\boldsymbol{X}}^H] \left( \mathbb{E}[\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^H] \right)^{-1} \hat{\boldsymbol{X}} \right) && (7.30)\\
&= U_k^H \hat{\boldsymbol{X}} + V_k^H K_{\hat{\boldsymbol{Y}}\hat{\boldsymbol{X}}} \hat{\boldsymbol{X}} && (7.31)\\
&= U_k^H \hat{\boldsymbol{X}} + \left( K_{\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}} V_k \right)^H \hat{\boldsymbol{X}}. && (7.32)
\end{aligned}
$$

The first summand contains exactly the top $k$ CCA components extracted from $\boldsymbol{X}$, which is the claimed result. The second summand requires further scrutiny. To proceed, we observe that for CCA, the projection vectors obey the relationship (see Equation (7.87))

$$
\boldsymbol{u} = \alpha K_{\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}} \boldsymbol{v}, \qquad (7.33)
$$

for some real-valued constant $\alpha$. Thus, combining the top $k$ CCA components, we can write

$$
U_k = D K_{\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}} V_k, \qquad (7.34)
$$

where $D$ is a diagonal matrix. Hence,

$$
\begin{aligned}
\mathbb{E}[W|\boldsymbol{X}] &= U_k^H \hat{\boldsymbol{X}} + D^{-1} U_k^H \hat{\boldsymbol{X}} && (7.35)\\
&= \tilde{D} U_k^H \hat{\boldsymbol{X}}, && (7.36)
\end{aligned}
$$

where $\tilde{D}$ is the diagonal matrix

$$
\tilde{D} = I + D^{-1}. \qquad (7.37)
$$

This is precisely the top $k$ CCA components (note that the solution to the CCA problem (7.82) is only specified up to a scaling). This establishes the theorem for the case of Version 2) of the proposed algorithm. Clearly, it also establishes that $p_\gamma(w|\boldsymbol{x})$ is a Gaussian distribution with mean given by (7.36), thus establishing the theorem for Version 1) of the proposed algorithm. The proof for Version 3) follows along similar lines and is thus omitted. □

## 7.5 A Binary Example

In this section we carry through a theoretical study of a somewhat more general case of the example discussed in Section 7.1.3 that is believed to be within the reach of practical data. In order to do a theoretical study we need to constrain the data into binary for the reason that computing the Wyner's common information for doubly binary symmetric source is already known.

Let us illustrate the proposed algorithm via a simple example. Consider the vector $(U, X_2, V, Y_2)$ of binary random variables. Suppose that $(U, V)$ are a doubly symmetric binary source. This means that $U$ is uniform and $V$ is the result of passing $U$ through a binary symmetric ("bit-flipping") channel with flip probability denoted by $a_0$ to match the notation in [34, Sec. 3]. Without loss of generality, we may assume $a_0 \leq \frac{1}{2}$. Meanwhile, $X_2$ and $Y_2$ are independent binary uniform random variables, also independent of the pair $(U, V)$. We will then form the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ as

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} U \oplus X_2 \\ X_2 \end{pmatrix}, \tag{7.38}$$

and

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} V \oplus Y_2 \\ Y_2 \end{pmatrix}, \tag{7.39}$$

where $\oplus$ denotes the modulo-reduced addition, as usual. How do various techniques perform for this example?

- Let us first analyze the behavior and outcome of CCA in this particular example. The key observation is that any pair amongst the four entries in these two vectors, $X_1, X_2, Y_1$, and $Y_2$, are (pairwise) independent binary uniform random variables. Hence, the overall covariance matrix of the merged random vector $(\boldsymbol{X}^T, \boldsymbol{Y}^T)^T$ is merely a scaled identity matrix. This, in turn, implies that CCA as described in Equations (7.25) and (7.26) merely boils down to the identity mapping. Concretely, this means that for CCA, in this example, the best one-dimensional projections are *ex aequo* any pair of one coordinate of the vector $\boldsymbol{X}$ with one coordinate of the vector $\boldsymbol{Y}$. As we have already explain above, any such pair is merely a pair of independent (and identically distributed) random variables, so CCA does not extract any dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$ at all. Of course, this is the main point of the present example.

- How does CICA perform in this example? We selected this example because it represents one of the only cases for which a closed-form solution to the optimization problem in Equation (7.6) is known, at least in the case $\gamma = 0$. To see this, let us first observe that in our example, we have

$$p(u, v, x_2, y_2) = p(u, v)p(x_2)p(y_2). \tag{7.40}$$

  Next, we observe that

$$\begin{aligned} C_\gamma(\boldsymbol{X}; \boldsymbol{Y}) &= C_\gamma(U, X_2; V, Y_2) && (7.41)\\ &= C_\gamma(U; V, Y_2) && (7.42)\\ &= C_\gamma(U; V) && (7.43) \end{aligned}$$

  where (7.42) follows from Lemma 9, Item 4, and the Markov chain $X_2 - U - (V, Y_2)$ that is satisfied from (7.40). The last equation (7.43) follows from

Lemma 9, Item 4, and the Markov chain $Y_2 - V - U$ that is satisfied from (7.40). That is, in this simple example, solving the optimization problem of Equation (7.6) is tantamount to solving the optimization problem in Equation (7.43). For the latter, the solution is well known, see [34, Sec. 3]. Specifically, we can express the conditional distribution $p_\gamma(w|\boldsymbol{x}, \boldsymbol{y})$ that solves the optimization problem of Equation (7.6) and is required for Step 3 of Generic Procedure 1 as follows:

$$p_{\gamma=0}(w|\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 1-\nu, & \text{if } w = 0, x_1 \oplus x_2 = 0, y_1 \oplus y_2 = 0, \\ \nu, & \text{if } w = 1, x_1 \oplus x_2 = 0, y_1 \oplus y_2 = 0, \\ \nu, & \text{if } w = 0, x_1 \oplus x_2 = 1, y_1 \oplus y_2 = 1, \\ 1-\nu, & \text{if } w = 1, x_1 \oplus x_2 = 1, y_1 \oplus y_2 = 1, \\ \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{7.44}$$

where

$$\nu = \frac{1}{2} - \frac{\sqrt{1-2a_0}}{2(1-a_0)}. \tag{7.45}$$

Let us now apply Version 1 (the MAP version) of Generic Procedure 1. To this end, we also need to calculate $p_\gamma(w|\boldsymbol{x})$ and $p_\gamma(w|\boldsymbol{y})$. Again, for $\gamma = 0$, these can be expressed in closed form as follows:

$$p_{\gamma=0}(w|\boldsymbol{x}) = \begin{cases} 1-a_1, & \text{if } w = 0, x_1 \oplus x_2 = 0, \\ a_1, & \text{if } w = 1, x_1 \oplus x_2 = 0, \\ a_1, & \text{if } w = 0, x_1 \oplus x_2 = 1, \\ 1-a_1, & \text{if } w = 1, x_1 \oplus x_2 = 1, \end{cases} \tag{7.46}$$

where

$$a_1 = \frac{1}{2}\left(1 - \sqrt{1-2a_0}\right). \tag{7.47}$$

The formula for $p_\gamma(w|\boldsymbol{y})$ follows by symmetry and shall be omitted. The final step is to follow Equations (7.7)-(7.8) and find $\arg\max_w p_{\gamma=0}(w|\boldsymbol{x})$ for each $\boldsymbol{x}$ as well as $\arg\max_w p_{\gamma=0}(w|\boldsymbol{y})$ for each $\boldsymbol{y}$. For the example at hand, these can be compactly expressed as

$$u(\boldsymbol{x}) = \arg\max_w p_\gamma(w|\boldsymbol{x}) = x_1 \oplus x_2 = u, \tag{7.48}$$

$$v(\boldsymbol{y}) = \arg\max_w p_\gamma(w|\boldsymbol{y}) = y_1 \oplus y_2 = v, \tag{7.49}$$

from the fact that $0 \leq a_0 \leq \frac{1}{2}$ that implies $0 \leq a_1 \leq \frac{1}{2}$. Hence, we find that for CICA as described in Generic Procedure 1, an optimal solution is to reduce $\boldsymbol{X}$ to $U$ and $\boldsymbol{Y}$ to $V$. This captures all the dependence between the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, which appears to be the most desirable outcome.

As a final note, we point out that it is conceptually straightforward to evaluate Versions 2 and 3 (conditional expectation) of Generic Procedure 1 in this example, but this would require embedding the considered binary alphabets into the real numbers. This makes it a less satisfying option for the simple example at hand.

## 7.6 A Gradient Descent Based Implementation

As we discussed above, in our problem, the objective is indeed a convex function of the optimization variables (but the constraint set is not convex). Clearly, this gives hope that gradient-based techniques may lead to interesting solutions. In this section, we examine a first tentative implementation and check it against ground truth for some simple examples.

For convex problems, gradient descent will guarantee convergence to the optimal solution and for non-convex problems it will guarantee only local convergence. Gradient descent runs in iterative steps, where each step does a local linear approximation and the step size depends on a learning parameter that is $\alpha$ for our problem. In our work we want to minimize the objective $I(W;X,Y)$ when the constraint $I(X;Y|W)$ is held below a $\gamma-$level.

Instead we apply a variant of gradient descent where we minimize the weighted sum of objective $I(W;X,Y)$ and the constraint $I(X;Y|W)$, which is $I(W;X,Y)+\lambda I(X;Y|W)$. The parameter $\lambda$ will permit some control on the constraint, thus sweeping all its possible values. We present the algorithm where $C(p(w|x,y))$ will be a function of $p(w|x,y)$ that will represent $I(W;X,Y)$ and $J(p(w|x,y))$ will be a function of $p(w|x,y)$ that will represent $I(X;Y|W)$.

**Algorithm 1:** Approximate CICA Algorithm via Gradient Descent

```
Set $\alpha, \lambda$, $error$ ;
$\beta = \lambda \cdot \alpha$ ;
Initialise $p(w|x,y)$ $randomly$ ;
Initialise $C_{new} \leftarrow 1, C_{old} \leftarrow 0$ ;
while $|C_{new} - C_{old}| > error$ do
    $C_{old} \leftarrow C_{new}$ ;
    $p(w|x,y) \leftarrow p(w|x,y) + \alpha \frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} + \beta \frac{\partial J(p(w|x,y))}{\partial p(w|x,y)}$ ;        // update step
    $C_{new} \leftarrow C(p(w|x,y))$ ;
Output $C_\gamma \leftarrow C_{new}$, $\gamma \leftarrow J(p(w|x,y))$ ;
Function $C(p(w|x,y)) \leftarrow \sum_{x,y,w} p(w|x,y)p(x,y) \log \frac{p(w|x,y)}{\sum_{x',y'} p(x',y')p(w|x',y')}$ ;
    // $I(W;X,Y)$
Function $J(p(w|x,y)) \leftarrow$
    $\sum_{x,y,w} p(w|x,y)p(x,y) \log \frac{p(w|x,y)p(x,y) \sum_{x',y'} p(x',y')p(w|x',y')}{\sum_{x''} p(w|x'',y)p(x'',y) \sum_{y''} p(w|x,y'')p(y'',x)}$ ;
    // $I(X;Y|W)$
```

The exact computation of the stated update step is presented in the following lemma.

**Lemma 14** (Computation of the update step). *Let $p(x,y)$ be a fixed distribution, then the updating steps for the gradient descent are*

$$\frac{\partial C(p(w|x,y)}{\partial p(w|x,y)} = p(x,y) \log \frac{p(w|x,y)}{\sum_{x',y'} p(x',y')p(w|x',y')}, \tag{7.50}$$

$$\frac{\partial J(p(w|x,y)}{\partial p(w|x,y)} = p(x,y) \log \frac{p(w|x,y) \sum_{x',y'} p(x',y')p(w|x',y')}{\sum_{x''} p(w|x'',y)p(x''|y) \sum_{y''} p(w|x,y'')p(y''|x)}. \tag{7.51}$$

*Proof.* Let the function $C$ be as defined above

$$C(p(w|x,y)) = \sum_{x,y,w} p(w|x,y)p(x,y) \log \frac{p(w|xy)}{\sum_{x',y'} p(w|x',y')p(x',y')}, \tag{7.52}$$

and in terms of information theoretic terms the function is $C(p(w|x,y)) = I(W;X,Y)$. In addition, $C(p(w|x,y))$ is a convex function of $p(w|x,y)$, shown in [4, Theorem 2.7.4]. Taking the first derivative we get

$$\begin{aligned}\frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} &= p(x,y) \log \frac{p(w|x,y)}{\sum_{x',y'} p(w|x',y')p(x',y')} + p(w|x,y)p(x,y)\frac{1}{p(w|x,y)} \\ &\quad - \sum_{x'',y''} p(w|x'',y'')p(x'',y'')\frac{p(x,y)}{\sum_{x',y'} p(w|x',y')p(x',y')} \qquad (7.53) \\ &= p(x,y) \log \frac{p(w|x,y)}{\sum_{x',y'} p(w|x',y')p(x',y')}. \qquad (7.54)\end{aligned}$$

On the other hand, the term $I(X;Y|W)$ can be expressed as

$$\begin{aligned} I(X;Y|W) &= I(W;X,Y) - I(W;X) - I(W;Y) + I(X;Y) \qquad (7.55) \\ &= C(p(w|x,y)) - C(p(w|x)) - C(p(w|y)) + I(X;Y). \qquad (7.56)\end{aligned}$$

Taking the derivative with respect to $p(w|x,y)$ becomes easier once $I(X;Y|W)$ is written in terms of function $C$ and we already know the derivative of $C$ from (7.54). Thus, the derivative would be

$$\begin{aligned}\frac{\partial J(p(w|x,y)}{\partial p(w|x,y)} &= \frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} - \frac{\partial C(p(w|x))}{\partial p(w|x,y)} - \frac{\partial C(p(w|y))}{\partial p(w|x,y)} \qquad (7.57) \\ &= \frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} - \frac{\partial C(p(w|x))}{\partial p(w|x)}\frac{\partial p(w|x)}{\partial p(w|x,y)} - \frac{\partial C(p(w|y))}{\partial p(w|y)}\frac{\partial p(w|y)}{\partial p(w|x,y)} \\ & \qquad (7.58) \\ &= p(x,y) \log \frac{p(w|x,y)}{\sum_{x',y'} p(w|x',y')p(x',y')} \\ &\quad - p(x) \log \frac{p(w|x)}{\sum_{x''} p(w|x'')p(x'')} p(y|x) - p(y) \log \frac{p(w|y)}{\sum_{y''} p(w|y'')p(y'')} p(x|y) \\ & \qquad (7.59) \\ &= p(x,y) \log \frac{p(w|x,y) \sum_{x',y'} p(x',y')p(w|x',y')}{\sum_{x''} p(w|x'',y)p(x''|y) \sum_{y''} p(w|x,y'')p(y''|x)}. \qquad (7.60)\end{aligned}$$

where (7.58) is an application of the chain rule and the rest is straightforward computation. □

**Remark 10.** *In practice, it is useful and computationally cheaper to replace the derivative formulas in Lemma 14 by their standard approximations. That is, the updating step in line 7 of Algorithm 1 is replaced by*

$$\frac{\partial C(p(w|x,y))}{\partial p(w|x,y)} \approx \frac{C(p(w|x,y) + \Delta) - C(p(w|x,y))}{\Delta}, \tag{7.61}$$

$$\frac{\partial J(p(w|x,y))}{\partial p(w|x,y)} \approx \frac{J(p(w|x,y) + \Delta) - J(p(w|x,y))}{\Delta}, \tag{7.62}$$

*for a judicious choice of $\Delta$. This is the version that was used to for Figure 7.1b, with $\Delta = 10^{-3}$. We point out that in the general case, the error introduced by this approximation is not bounded.*

## 7.7 Extension to More Than Two Sources

It is unclear how one would extend CCA to more than two databases. By contrast, for CICA, this extension is conceptually straightforward. For Wyner's common information, in Equation (4.1), it suffices to replace the objective in the minimization by $I(X_1, X_2, \ldots, X_M; W)$ and to keep the constraint of conditional independence. To obtain an interesting algorithm, we now need to relax the constraint of conditional independence. The most natural way is via the conditional version of Watanabe's total correlation [77], leading to the following definition:

**Definition 6** (Relaxed Wyner's Common Information for $M$ variables)**.** *For a fixed probability distribution $p(x_1, x_2, \ldots, x_M)$, we define*

$$C_\gamma(X_1; X_2; \ldots; X_M) = \inf I(X_1, X_2, \ldots, X_M; W) \tag{7.63}$$

*such that $\sum_{i=1}^{M} H(X_i|W) - H(X_1, X_2, \ldots, X_M|W) \leq \gamma$, where the infimum is over all probability distributions $p(w, x_1, x_2, \ldots, x_M)$ with marginal $p(x_1, x_2, \ldots, x_M)$.*

Not surprisingly, explicit closed-form solution are difficult to find. One simple case appears below as part of the example presented in Section 7.7.1, see Lemma 16. By analogy to Lemma 9, we can again state basic properties.

**Lemma 15.** *$C_\gamma(X_1; X_2; \ldots; X_M)$ satisfies the following basic properties:*

1. *$C_\gamma(X_1; X_2; \ldots; X_M) \geq \frac{1}{M-1} \max\{\sum_{i=1}^{M} H(X_i) - H(X_1, X_2, \ldots, X_M) - \gamma, 0\}$.*
2. *$C_\gamma(X_1; X_2; \ldots; X_M)$ is a convex and continuous function of $\gamma$ for $\gamma \geq 0$.*
3. *If $Z - X_1 - (X_2, \ldots, X_M)$ forms a Markov chain, then*
$$C_\gamma((X_1, Z); X_2; \ldots; X_M) = C_\gamma(X_1; X_2; \ldots; X_M). \tag{7.64}$$
4. *The cardinality of $\mathcal{W}$ may be restricted to $|\mathcal{W}| \leq \prod_{i=1}^{M} |\mathcal{X}_i| + 1$.*
5. *If $f_i(\cdot)$ are one-to-one functions, then*
$$C_\gamma(f_1(X_1); f_2(X_2); \ldots; f_M(X_M)) = C_\gamma(X_1; X_2; \ldots; X_M). \tag{7.65}$$
6. *For discrete $X$, we have $C_\gamma(X; X; \ldots; X) = \max\{H(X) - \frac{\gamma}{M-1}, 0\}$.*

Proofs for these basic properties can be found in Appendix 7.9.2.

Leveraging Definition 6, it is conceptually straightforward to extend CICA (that is, Generic Procedure 1) to the case of $M$ databases as follows. For completeness, we include an explicit statement of the resulting procedure.

**Generic Procedure 2** (CICA with multiple sources). *1. Select a real number $\gamma$, where $0 \leq \gamma \leq \sum_{i=1}^{M} H(\boldsymbol{X}_i) - H(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)$. This is the compression level: A low value of $\gamma$ represents low compression, and thus, many components are retained. A high value of $\gamma$ represents high compression, and thus, only a small number of components are retained.*

*2. Solve the relaxed Wyner's common information problem,*

$$\min_{p(w|\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_M)} I(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M; W) \tag{7.66}$$

$$\text{such that } \sum_{i=1}^{M} H(\boldsymbol{X}_i|W) - H(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M|W) \leq \gamma, \tag{7.67}$$

*leading to an associated conditional distribution $p_\gamma(w|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M)$.*

*3. Using the conditional distribution $p_\gamma(w|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M)$ found in Step 2), the dimension-reduced data sets can now be found via one of the following three variants:*

*a) Version 1: MAP (maximum a posteriori):*

$$u_i(\boldsymbol{x}_i) = \arg\max_w p_\gamma(w|\boldsymbol{x}_i), \tag{7.68}$$

*for $i = 1, 2, \ldots, M$.*

*b) Version 2: Conditional Expectation:*

$$u_i(\boldsymbol{x}_i) = \mathbb{E}[W|\boldsymbol{X}_i = \boldsymbol{x}_i] \tag{7.69}$$

*for $i = 1, 2, \ldots, M$.*

*c) Version 3: Marginal Integration:*

$$\begin{aligned} u_i(\boldsymbol{x}_i) = \int_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{i-1},\boldsymbol{x}_{i+1},\ldots,\boldsymbol{x}_M} & p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_M) \\ & \mathbb{E}[W|\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_M = \boldsymbol{x}_M] d\boldsymbol{x}_1 \cdots d\boldsymbol{x}_{i-1} d\boldsymbol{x}_{i+1} \cdots d\boldsymbol{x}_M \end{aligned} \tag{7.70}$$

*for $i = 1, 2, \ldots, M$.*

Clearly, Generic Procedure 2 closely mirrors Generic Procedure 1. The key difference is that there is no direct analog of Theorem 17. This is no surprise since it is unclear how CCA would be extended to beyond the case of two sources. Nonetheless, it would be very interesting to explore what Generic Procedure 2 boils down to in the special case when all vectors are jointly Gaussian. At the current time, this is unknown. In fact, the explicit solution to the optimization problem in Definition 6 is presently an open problem.

Instead, we illustrate the promise of Generic Procedure 2 via a simple binary example in the next section. The example mirrors some of the basic properties of the example tackled in Section 7.5.

### 7.7.1 A Binary Example With Three Sources

In this section, we develop an example with three sources that borrows some of the ideas from the example discussed in Section 7.5. In a sense, the present example is even more illustrative because in it, *any* two distinct components of the original vectors $\boldsymbol{X}_1, \boldsymbol{X}_2$, and $\boldsymbol{X}_3$, are (pairwise) independent. Therefore, any method based on pairwise measures, including CCA and maximal correlation, would not identify any commonality at all. Specifically, we consider the following simple statistical model:

$$\boldsymbol{X}_1 = \begin{pmatrix} U \\ Z_1 \end{pmatrix}, \boldsymbol{X}_2 = \begin{pmatrix} V \\ Z_2 \end{pmatrix}, \boldsymbol{X}_3 = \begin{pmatrix} U \oplus V \\ Z_3 \end{pmatrix}, \tag{7.71}$$

where $U, V, Z_1, Z_2, Z_3$ are independent uniform binary variables and $\oplus$ denotes modulo-2 addition. We observe that amongst these three vectors, any pair is independent. This implies, for example, that any correlation-based technique (including maximal correlation) will not identify any relevant features, since correlation is a pairwise measure. By contrast, we can show that one output of Algorithm 2 is indeed to select $W = (U, V)$, for $\gamma = 0$. Thus, the algorithm would reduce each of the three vectors to their first component, which is the intuitively pleasing answer in this case. By going through the steps of the Generic Procedure 2, for $\gamma = 0$, where the the joint distribution satisfies

$$p(u, v, u \oplus v, z_1, z_2, z_3) = p(u, v, u \oplus v)p(z_1)p(z_2)p(z_3) \tag{7.72}$$

we have that

$$\begin{aligned} C(\boldsymbol{X}_1; \boldsymbol{X}_2; \boldsymbol{X}_3) &= C(U, Z_1; V, Z_2; U \oplus V, Z_3) && (7.73) \\ &= C(U; V, Z_2; U \oplus V, Z_3) && (7.74) \\ &= C(U; V; U \oplus V, Z_3) && (7.75) \\ &= C(U; V; U \oplus V) && (7.76) \end{aligned}$$

where we use Lemma 15, Item 3, together with the Markov chain $Z_1 - U - (Z_2, V, Z_3, U \oplus V)$ that follows from (7.72) to prove step (7.74). Similarly, the Markov chain $Z_2 - V - (U, Z_3, U \oplus V)$ proves step (7.75) by making use of Lemma 15, Item 3. A similar argument is used for the last step (7.76). Managing to compute $C(U; V; U \oplus V)$ is equivalent to computing $C(\boldsymbol{X}_1; \boldsymbol{X}_2; \boldsymbol{X}_3)$ and we demonstrate how to compute it in the next part.

**Lemma 16.** *Let $U, V$ be independent uniform binary variables and $\oplus$ denotes modulo-2 addition. Then, the optimal solution to*

$$C_{\gamma=0}(U; V; U \oplus V) = \inf_{W : H(U|W) + H(V|W) + H(U \oplus V|W) - H(U,V,U \oplus V|W) = 0} I(W; U, V, U \oplus V) \tag{7.77}$$

*is $W = (U, V)$ where the expression evaluates to two.*

The proof is given in Appendix 7.9.3. If we apply Version 1 of Step 3 of Generic Procedure 2, we obtain

$$\arg\max_{w} p_{\gamma=0}(w|\boldsymbol{x}_1) = \{(u, 0), (u, 1)\}, \tag{7.78}$$

that is, in this case, the maximizer is not unique. However, as we observe that the set of maximizers is a deterministic function of $u$ alone, it is natural to reduce as follows:

$$u_1(\boldsymbol{x}_1) = u. \tag{7.79}$$

By the same token, we can reduce

$$u_2(\boldsymbol{x}_2) = v, \tag{7.80}$$

$$u_3(\boldsymbol{x}_3) = u \oplus v. \tag{7.81}$$

In this example, it is clear that this indeed extracts all of the dependency there is between our three sources, and thus, is the correct answer.

As pointed out above, in this simple example, any pair of the random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2$, and $\boldsymbol{X}_3$ are (pairwise) independent, which implies that the classic tools based on pairwise measures (CCA, maximal correlation) cannot identify any commonality between $\boldsymbol{X}_1, \boldsymbol{X}_2$, and $\boldsymbol{X}_3$.

## 7.8 Conclusion

We introduce a novel two-step procedure that we refer to as CICA. The first step consists in an information minimization problem related to Wyner's common information, while the second can be thought of as a type of back-projection. We prove that in the special case of Gaussian statistics, this two-step procedure precisely extracts the CCA components. A free parameter $\gamma$ in CICA permits to select the number of CCA components that are being extracted. In this sense, the chapter establishes a novel rigorous connection between CCA and information measures. A number of simple examples are presented. It is also shown how to extend the novel algorithm to more than two sources.

Future work includes a more in-depth study and consideration to assess the practical promise of this novel algorithm. This will also require to move beyond the current setting where it was assumed that the probability distribution of the data at hand was provided directly. Instead, this distribution has to be estimated from data, and one needs to understand what limitations this additional constraint will end up imposing.

## 7.9 Appendix

### 7.9.1 A Brief Review of Canonical Correlation Analysis (CCA)

A brief review of CCA [57] is presented. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be zero-mean real-valued random vectors with covariance matrices $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$, respectively. Moreover, let $K_{\boldsymbol{XY}} = \mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^H]$. We first apply the change of basis as in (7.17)-(7.18). CCA seeks to find vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ such as to maximize the correlation between $\boldsymbol{u}^H\hat{\boldsymbol{X}}$ and $\boldsymbol{v}^H\hat{\boldsymbol{Y}}$, that is,

$$\max_{\boldsymbol{u},\boldsymbol{v}} \frac{\mathbb{E}[\boldsymbol{u}^H\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}^H\boldsymbol{v}]}{\sqrt{\mathbb{E}[|\boldsymbol{u}^H\hat{\boldsymbol{X}}|^2]}\sqrt{\mathbb{E}[|\boldsymbol{v}^H\hat{\boldsymbol{Y}}|^2]}}, \tag{7.82}$$

which can be rewritten as

$$\max_{\boldsymbol{u},\boldsymbol{v}} \frac{\boldsymbol{u}^H K_{\hat{X}\hat{Y}} \boldsymbol{v}}{\|\boldsymbol{u}\| \, \|\boldsymbol{v}\|}, \tag{7.83}$$

where

$$K_{\hat{X}\hat{Y}} = K_{\boldsymbol{X}}^{-1/2} K_{\boldsymbol{XY}} K_{\boldsymbol{Y}}^{-1/2}. \tag{7.84}$$

Note that this expression is *invariant* to arbitrary (separate) scaling of $\boldsymbol{u}$ and $\boldsymbol{v}$. To obtain a unique solution, we could choose to impose that both vectors be unit vectors,

$$\max_{\boldsymbol{u},\boldsymbol{v}:\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=1} \boldsymbol{u}^H K_{\hat{X}\hat{Y}} \boldsymbol{v}. \tag{7.85}$$

From Cauchy-Schwarz, for a fixed $\boldsymbol{u}$, the maximizing (unit-norm) $\boldsymbol{v}$ is given by

$$\boldsymbol{v} = \frac{K_{\hat{X}\hat{Y}}^H \boldsymbol{u}}{\left\| K_{\hat{X}\hat{Y}}^H \boldsymbol{u} \right\|}, \tag{7.86}$$

or equivalently, for a fixed $\boldsymbol{v}$, the maximizing (unit-norm) $\boldsymbol{u}$ is given by

$$\boldsymbol{u} = \frac{K_{\hat{X}\hat{Y}} \boldsymbol{v}}{\left\| K_{\hat{X}\hat{Y}} \boldsymbol{v} \right\|}. \tag{7.87}$$

Plugging in the latter, we obtain

$$\max_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \frac{\boldsymbol{v}^H K_{\hat{X}\hat{Y}}^H K_{\hat{X}\hat{Y}} \boldsymbol{v}}{\left\| K_{\hat{X}\hat{Y}} \boldsymbol{v} \right\|}, \tag{7.88}$$

or, dividing through,

$$\max_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \left\| K_{\hat{X}\hat{Y}} \boldsymbol{v} \right\|. \tag{7.89}$$

The solution to this problem is well known: $\boldsymbol{v}$ is the right singular vector corresponding to the largest singular vector of the matrix $K_{\hat{X}\hat{Y}} = K_{\boldsymbol{X}}^{-1/2} K_{\boldsymbol{XY}} K_{\boldsymbol{Y}}^{-1/2}$. Evidently, $\boldsymbol{u}$ is the corresponding left singular vector. Restarting again from Equation (7.82), but restricting to vectors that are orthogonal to the optimal choices of the first round leads to the second CCA components, and so on.

### 7.9.2 Proof of Lemma 15

For item 1) we proceed as follows

$$C_\gamma(X_1; X_2; \ldots; X_M) = \inf_{\substack{W: H(X_1|W)+H(X_2|W)+\cdots+H(X_M|W) \\ -H(X_1,X_2,\ldots,X_M|W)\leq\gamma}} I(W; X_1, X_2, \ldots, X_M) \tag{7.90}$$

$$\geq \inf_W L(\lambda, p(w|x_1, x_2, \ldots, x_M)) \tag{7.91}$$

where we used weak duality for $\lambda \geq 0$ and $L(\lambda, p(w|x_1, x_2, \ldots, x_M)$ is

$$L(\lambda, p(w|x_1, x_2, \ldots, x_M) := I(W; X_1, X_2, \ldots, X_M) + \lambda[H(X_1|W) + H(X_2|W) + \ldots \\ + H(X_M|W) - H(X_1, X_2, \ldots, X_M|W) - \gamma]. \quad (7.92)$$

By setting $\lambda = \frac{1}{M-1}$, we obtain

$$\begin{aligned} & L(\frac{1}{M-1}, p(w|x_1, x_2, \ldots, x_M)) & (7.93) \\ & = \frac{M}{M-1} I(W; X_1, X_2, \ldots, X_M) \\ & - \frac{1}{M-1}[I(W; X_1) + I(W; X_2) + \cdots + I(W; X_M)] \\ & + \frac{1}{M-1}[H(X_1) + H(X_2) + \cdots + H(X_M) - H(X_1, X_2, \ldots, X_M) - \gamma] & (7.94) \\ & = \frac{1}{M-1}[I(W; X_2, \ldots, X_M|X_1) + \cdots + \frac{1}{M-1}[I(W; X_1, \ldots, X_{M-1}|X_M) \\ & + \frac{1}{M-1}[H(X_1) + H(X_2) + \cdots + H(X_M) - H(X_1, X_2, \ldots, X_M) - \gamma], & (7.95) \end{aligned}$$

where the infimum of $L(\frac{1}{M-1}, p(w|x_1, x_2, \ldots, x_M))$ in (7.95) is attained for the trivial random variable $W$, thus $C_\gamma(X_1; X_2; \ldots; X_M) \geq \frac{1}{M-1}[H(X_1) + H(X_2) + \cdots + H(X_M) - H(X_1, X_2, \ldots, X_M) - \gamma]$. Item 2) follows from a similar argument as in [34, Corollary 4.5]. For item 3) we start by showing both sides of the inequality that will result in equality. One side of the inequality is shown below

$$\begin{aligned} & C_\gamma(X_1, Z; X_2; \ldots; X_M) & (7.96) \\ & = \inf_{W: H(X_1, Z|W) + H(X_2|W) + \cdots + H(X_M|W) - H(X_1, Z, X_2, \ldots, X_M|W) \leq \gamma} I(W; X_1, Z, X_2, \ldots, X_M) & (7.97) \\ & = \inf_{\substack{W: H(X_1|W) + H(X_2|W) + \cdots + H(X_M|W) \\ -H(X_1, X_2, \ldots, X_M|W) + I(Z; X_2, \ldots, X_M|X_1, W) \leq \gamma}} I(W; X_1, X_2, \ldots, X_M) + I(W; Z|X_1, X_2, \ldots, X_M) \\ & & (7.98) \\ & \leq C_\gamma(X_1; X_2; \ldots; X_M) & (7.99) \end{aligned}$$

where the last inequality follows by restricting the possible set of $W$, such that $W$ and $Z$ are conditionally independent given $(X_1, X_2, \ldots, X_M)$,

$$I(Z; W|X_1, X_2, \ldots, X_M) = 0. \quad (7.100)$$

From the statement of the lemma we have $Z - X_1 - (X_2, \ldots, X_M)$,

$$I(Z; X_2, \ldots, X_M|X_1) = 0. \quad (7.101)$$

By adding (7.100) and (7.101) we get $I(Z; W, X_2, \ldots, X_M|X_1) = 0$. That implies $I(Z; X_2, \ldots, X_M|X_1, W) = 0$, which appears in the constraint of (7.98). For the other part of the inequality we proceed as follows

$$C_\gamma(X_1, Z; X_2; \dots; X_M) \tag{7.102}$$

$$= \inf_{\substack{W: H(X_1|W)+H(X_2|W)+\dots+H(X_M|W) \\ -H(X_1,X_2,\dots,X_M|W)+I(Z;X_2,\dots,X_M|X_1,W)\le\gamma}} I(W; X_1, X_2, \dots, X_M) + I(W; Z|X_1, X_2, \dots, X_M) \tag{7.103}$$

$$\ge C_\gamma(X_1; X_2; \dots; X_M), \tag{7.104}$$

where the last part follows by relaxing the constraint set as $I(Z; X_2, \dots, X_M|X_1, W) \ge 0$ and by further bounding the terms in the objective, $I(W; Z|X_1, X_2, \dots, X_M) \ge 0$.

Item 4) is a standard cardinality bound, following from a similar argument in [49]. Item 5) follows because all involved mutual information terms are invariant to one-to-one transforms. For item 6) we apply the definition of relaxed Wyner's common information for $M$ variables and we have

$$C_\gamma(X; X; \dots; X) = \inf_{W:(M-1)H(X|W)\le\gamma} I(X; W) \tag{7.105}$$

$$= H(X) - \sup_{W:(M-1)H(X|W)\le\gamma} H(X|W) \tag{7.106}$$

$$= H(X) - \frac{\gamma}{M-1}. \tag{7.107}$$

### 7.9.3 Proof of Lemma 16

An upper bound to the problem is to pick $W = (U, V)$, thus

$$C(U; V; U \oplus V) \le H(U, V, U \oplus V) = 2. \tag{7.108}$$

Another equivalent way of writing the problem is by splitting the constraint into two constraints, as we already know that the constraint cannot be smaller than zero, so it has to be exactly zero and it can be written in the following way

$$C(U; V; U \oplus V) = \inf_{\substack{W: I(U\oplus V; U,V|W)=0 \\ I(U;V|W)=0}} I(W; U, V, U \oplus V). \tag{7.109}$$

By using weak duality for $\lambda \ge 0$, a lower bound to the problem would be the following

$$C(U; V; U \oplus V) \ge \inf_{W:U-W-V} I(W; U, V, U \oplus V) + \lambda[H(U, V|W) + H(U \oplus V|W) - H(U, V, U \oplus V|W)]. \tag{7.110}$$

By further using the constraint $U - W - V$ the above expression can be written as

$$C(U; V; U \oplus V) \ge \inf_{W:U-W-V} I(W; U, V, U \oplus V) + \lambda[H(U|W) + H(V|W) + H(U \oplus V|W) - H(U, V, U \oplus V|W)] \tag{7.111}$$

$$= H(U, V, U \oplus V) + \inf_{W:U-W-V} \lambda[H(U|W) + H(V|W)] + \lambda H(U \oplus V|W) - (1 + \lambda)H(U, V, U \oplus V|W) \tag{7.112}$$

$$\begin{aligned}
&\geq H(U,V,U\oplus V) + \inf_{\tilde{U},\tilde{V}} \inf_{\substack{W:\tilde{U}-W-\tilde{V}\\ \tilde{U}\oplus\tilde{V}-(\tilde{U},\tilde{V})-W}} \lambda[H(\tilde{U}|W)+H(\tilde{V}|W)]\\
&\qquad + \lambda H(\tilde{U}\oplus\tilde{V}|W) - (1+\lambda)H(\tilde{U},\tilde{V},\tilde{U}\oplus\tilde{V}|W) \qquad (7.113)\\
&= 2 + \inf_{\tilde{U},\tilde{V}} \inf_{\substack{W:\tilde{U}-W-\tilde{V}\\ \tilde{U}\oplus\tilde{V}-(\tilde{U},\tilde{V})-W}} \underbrace{\lambda H(\tilde{U}\oplus\tilde{V}|W) - H(\tilde{U}|W) - H(\tilde{V}|W)}_{r(\tilde{U},\tilde{V}|W)}\\
&\qquad\qquad (7.114)\\
&= 2 + \inf_{\tilde{U},\tilde{V}} \check{r}(\tilde{U},\tilde{V}) \qquad (7.115)
\end{aligned}$$

where (7.113) is a consequence of allowing a minimization (if minimum exists) over binary random variables $\tilde{U},\tilde{V}$ and the rest of equalities is straightforward manipulation. The last equation is in terms of the lower convex envelope with respect to the distribution $p_{\tilde{U}}p_{\tilde{V}}$. The aim is to search for the tightest bound over $\lambda$ by studying the lower convex envelope with respect to $p_{\tilde{U}}p_{\tilde{V}}$, which is fact for binary and independent $\tilde{U},\tilde{V}$ can be simplified into

$$\lambda H(\tilde{U}\oplus\tilde{V}) - H(\tilde{U}) - H(\tilde{V}) = \lambda h_b(\alpha\beta + (1-\alpha)(1-\beta)) - h_b(\alpha) - h_b(\beta) \qquad (7.116)$$

where $0 \leq \alpha,\beta \leq 1$. Thus, $\inf_{\tilde{U},\tilde{V}} \check{r}(\tilde{U},\tilde{V}) = \inf_{\alpha,\beta} \check{r}(\alpha,\beta)$. Note that (7.116) is a continuous function of $\alpha,\beta$ so a first order and a second order differentiation will be enough to compute the lower convex envelope. As a result for $\lambda \geq 2$, the lower convex envelope of the right hand side of (7.116) is just zero, thus completing the proof.

# Upper Bound on Double Information Bottleneck 8

## 8.1 Introduction

Information bottleneck [65] is an information theoretic measure closely related to rate distortion theory and is defined for a pair of random variables labelled as input and output. More precisely, we seek to compress the input such that it preserves the maximum information about the output. The information bottleneck is defined in [65] as

$$\sup_{\substack{T: I(T;X)\leq\alpha \\ T-X-Y}} I(T;Y) \tag{8.1}$$

Recent studies have shown the usefulness and importance, however the information bottleneck computation for a given pair of random variables is known only in special cases. For a pair of Gaussian random variables, the information bottleneck is computed in [60]. This chapter considers an extended variant of information bottleneck that is defined below.

**Definition 7.** *Double information bottleneck for a pair of random variables $(X, Y)$, is defined as*

$$I_D(\alpha) = \sup_{\substack{(U,V): I(U;X)+I(V;Y)\leq\alpha \\ U-X-Y-V}} I(U;V) \tag{8.2}$$

We study only the case when the pair $(X, Y)$ is Gaussian. While the optimal auxiliary random variable for the standard information bottleneck turns out to be Gaussian, for the double information bottleneck it is not generally the case.

## 8.2 Main Result

The main contribution of this chapter is the following theorem.

**Theorem 18.** *For a pair of Gaussian random variables $(X, Y)$ where $X$ and $Y$ have unit variance, correlation coefficient $\rho$ and $|\rho| \geq 1 - e^{-\alpha}$, we have*

$$I_D(\alpha) \leq g_1(\alpha, \rho), \tag{8.3}$$

*where*

$$g_1(\alpha, \rho) = \alpha - \frac{1}{2} \log \frac{1 + |\rho|}{2e^{-\alpha} - 1 + |\rho|}. \tag{8.4}$$

*Proof.* Go to Appendix 8.4.1 □

## 8.3 Result Comparison

### 8.3.1 Lower bound evaluated with Gaussians

**Proposition 15.** *For a pair of Gaussian random variables $(X, Y)$ where $X$ and $Y$ have unit variance, correlation coefficient $\rho$, we have*

$$I_D(\alpha) \geq g_2(\alpha, \rho), \tag{8.5}$$

*where*

$$g_2(\alpha, \rho) = \frac{1}{2} \log \frac{1}{1 - \rho^2(1 - e^{-\alpha})^2}. \tag{8.6}$$

*Proof.* A trivial lower bound is by setting $(X, Y, U, V)$ to be jointly Gaussian. For the Markov chain $U - X - Y - V$ to hold we need

$$\rho_{uv} = \rho_{ux}\rho_{vy}\rho. \tag{8.7}$$

In addition, we need to satisfy the constraint

$$I(U; X) + I(V; Y) = \frac{1}{2} \log \frac{1}{1 - \rho_{ux}^2} + \frac{1}{2} \log \frac{1}{1 - \rho_{vy}^2} = \alpha. \tag{8.8}$$

Then, by setting $\rho_{ux} = \rho_{vy}$, we have

$$\rho_{ux} = \rho_{vy} = \sqrt{1 - e^{-\alpha}}, \tag{8.9}$$

and $\rho_{uv} = \rho(1 - e^{-\alpha})$ by using (8.7) and (8.9). Thus, the lower bound is

$$\begin{aligned} I(U; V) &= \frac{1}{2} \log \frac{1}{1 - \rho_{uv}^2} && (8.10) \\ &= \frac{1}{2} \log \frac{1}{1 - \rho^2(1 - e^{-\alpha})^2}. && (8.11) \end{aligned}$$

□

### 8.3.2 (Strong) data processing inequality upper bound

**Proposition 16.** *For a pair of Gaussian random variables $(X, Y)$ where $X$ and $Y$ have unit variance, correlation coefficient $\rho$, we have*

$$I_D(\alpha) \leq \min\left\{g_3(\alpha, \rho), I(X;Y)\right\}, \tag{8.12}$$

*where*

$$g_3(\alpha, \rho) = \rho^2 \frac{\alpha}{2}. \tag{8.13}$$

*Proof.* We bound the objective $I(U;V) \leq I(X;Y)$ by data processing inequality (DPI), thus

$$I_D(\alpha) \leq I(X;Y), \tag{8.14}$$

is a trivial upper bound.

For a pair of Gaussian random variables $(X, Y)$, where $X$ and $Y$ have a correlation coefficient $\rho$ and $U$ satisfies the Markov chain $U - X - Y$, the strong data processing inequality in [78] implies that

$$\rho^2 I(U;X) \geq I(U;Y). \tag{8.15}$$

For the Markov chain $U - X - Y - V$, the constraint of $I_D$ is relaxed as follows

$$\begin{aligned} \rho^2\alpha &\geq \rho^2 I(U;X) + \rho^2 I(V;Y) && (8.16) \\ &\geq I(U;Y) + I(X;V) && (8.17) \\ &\geq 2I(U;V) && (8.18) \end{aligned}$$

where (8.16) follows from the original constraint, (8.17) follows from strong data processing inequality in (8.15) and (8.18) follows from data processing inequality. Thus, $I(U;V) \leq \rho^2 \frac{\alpha}{2}$ and

$$I_D(\alpha) \leq \rho^2 \frac{\alpha}{2}. \tag{8.19}$$

By combining (8.14) and (8.19) we have

$$I_D(\alpha) \leq \min\left\{\rho^2 \frac{\alpha}{2}, I(X;Y)\right\}. \tag{8.20}$$

□

### 8.3.3 Comparison with Theorem 18

Let us compare the upper bound from Theorem 18 with the upper bound derived from (strong) data processing inequality.

**Proposition 17.** *For a pair of Gaussian random variables $(X, Y)$ where $X$ and $Y$ have unit variance, correlation coefficient $\rho$ and $|\rho| \geq 1 - e^{-\alpha}$, we have*

1. $I(X;Y) \geq g_1(\alpha, \rho)$,

*2.* $g_3(\alpha,\rho) \leq g_1(\alpha,\rho)$ *for* $|\rho| \leq 1-e^{-1}$,

*where* $g_1$ *is given in Equation (8.4) and* $g_3$ *is given in Equation (8.13).*

*Proof.* Go to Appendix 8.4.2. □

Figure 8.1 makes a comparison with the previous results in the literature and shows that the our derived upper bound is better than existing bounds in a certain regime of $\alpha$ and $\rho$. Proposition 17, Item 2 implies that for $\alpha \leq 1$, the bound from Theorem 18 is worse than the existing results, thus we choose $\alpha = 2.6$ and we plot $\rho$ versus $I_D(\alpha = 2.6)$. A more natural plot would be, fixing $\rho$ and plotting $\alpha$ versus $I_D(\alpha)$, however for visual purposes prefer fixing $\alpha$ and plotting $\rho$ versus $I_D(\alpha)$. One of the limits of the bound in Theorem 18 is that the bound is valid only for $1 \geq \rho \geq 1-e^{-\alpha}$ or in other words for $0 \leq \alpha \leq \log\frac{1}{1-\rho}$. Thus, for a fixed $\alpha$ the range of $\rho$ is limited and for a fixed $\rho$ the range of $\alpha$ is limited. Another limit of the bound in Theorem 18 comes from Proposition 17, Item 2 that for $\alpha \leq 1$ the bound is strictly worse than existing results. In Figure 8.1 we show the dominance of the bound for a relatively large value of $\alpha$ only for the validity region $\rho \geq 1-e^{-\alpha}$.

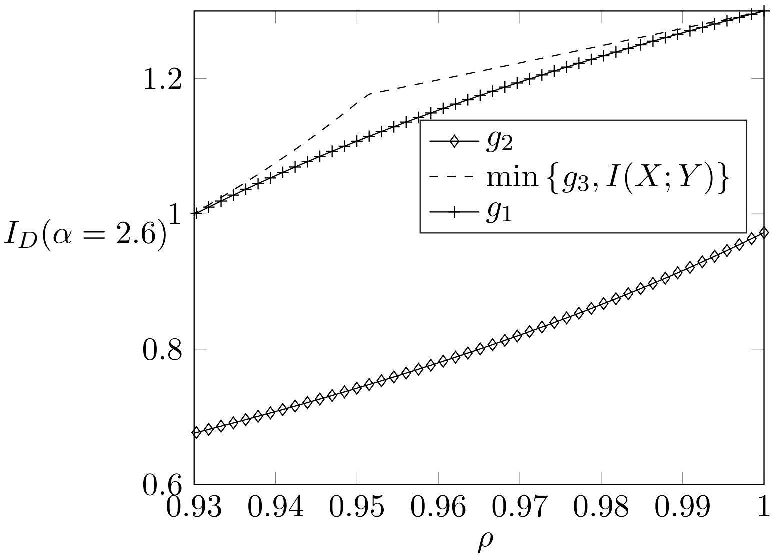


Figure 8.1 – The ◇-line is a lower bound evaluated with jointly Gaussian auxiliaries, the dashed line is the upper bound from (strong) data processing inequality, i.e. $\min\left\{\rho^2\frac{\alpha}{2}, I(X;Y)\right\}$, the +-line is the upper bound from Theorem 18. We plot the bounds on $I_D(\alpha = 2.6)$ in nats versus $\rho$.

## 8.4 Appendix

### 8.4.1 Proof of Theorem 18

Let us consider the case when $\rho \geq 0$, thus we have

$$-I_D(\alpha) = \inf_{\substack{(U,V):I(U;X)+I(V;Y)\leq\alpha \\ U-X-Y-V}} -I(U;V) \tag{8.21}$$

$$
\begin{aligned}
&\geq \inf_{(U,V):U-X-Y-V} -I(U;V)+\lambda(I(U;X)+I(V;Y)-\alpha) && (8.22)\\
&= \inf_{(U,V):U-X-Y-V} I(U,X;Y|V)+I(X;V|U)-I(X;Y)\\
&\quad +\lambda(I(U;X)+I(V;Y)-\alpha) && (8.23)\\
&= \inf_{(U,V):U-X-Y-V} -h(X,Y|U,V)+(1-\lambda)h(X|U)+(1-\lambda)h(Y|V)\\
&\quad +\lambda h(X)+\lambda h(Y)-I(X;Y)-\lambda\alpha && (8.24)\\
&\geq \inf_{(U,V):U-X-Y-V} -h(X,Y|U,V)+(1-\lambda)h(X|U,V)+(1-\lambda)h(Y|U,V)\\
&\quad +\lambda h(X)+\lambda h(Y)-I(X;Y)-\lambda\alpha && (8.25)\\
&\geq \lambda h(X)+\lambda h(Y)-I(X;Y)-\lambda\alpha\\
&\quad +(1-\lambda)\inf_{(U,V)} h(X|U,V)+h(Y|U,V)-\frac{1}{1-\lambda}h(X,Y|U,V) && (8.26)\\
&\geq \lambda h(X)+\lambda h(Y)-I(X;Y)-\lambda\alpha\\
&\quad +(1-\lambda)\cdot \inf_{K':0\preceq K'\preceq\begin{pmatrix}1&\rho\\ \rho&1\end{pmatrix}} h(X')+h(Y')-\frac{1}{1-\lambda}h(X',Y') && (8.27)\\
&\geq -\frac{1}{2}\log\frac{1}{1-\rho^2}-\lambda\alpha+\frac{1-\lambda}{2}\log\frac{(1-\lambda)^2}{1-2\lambda}-\frac{\lambda}{2}\log\frac{(1-\rho)^2}{1-2\lambda} && (8.28)
\end{aligned}
$$

where (8.22) follows from weak duality, (8.23) follows from the fact that the Markov chain $U-X-Y-V$ is written in terms of mutual information as

$$I(U,X;V,Y)=I(X;Y). \tag{8.29}$$

Now rewrite $I(U,X;V,Y)$ in order to dissociate the term $I(U;V)$, thus we have

$$I(U,X;V,Y)=I(U,X;Y|V)+I(X;V|U)+I(U;V). \tag{8.30}$$

By combining (8.29) and (8.30) we get

$$I(U;V)=I(X;Y)-I(U,X;Y|V)-I(X;V|U). \tag{8.31}$$

Equation (8.24) follows from rearranging the terms, (8.25) follows from conditioning reduces entropy i.e. $h(X|U)\geq h(X|U,V)$, (8.26) follows from relaxing the constraint. For $\lambda\geq 1$, from the max-entropy argument, the optimal distribution is jointly Gaussian. The remaining case is $0<\lambda<1$. Equation (8.27) follows from Theorem 9 for $0<\lambda<\frac{1}{2}$, where $(X',Y')\sim\mathcal{N}(0,K')$ and (8.28) follows from Lemma 11 for $\lambda\leq\frac{\rho}{1+\rho}$. Then, we find the tightest bound in (8.28) i.e. we maximize (8.28) for $0<\lambda\leq\frac{\rho}{1+\rho}$. Let us define

$$f(\lambda):=-\frac{1}{2}\log\frac{1}{1-\rho^2}-\lambda\alpha+\frac{1-\lambda}{2}\log\frac{(1-\lambda)^2}{1-2\lambda}-\frac{\lambda}{2}\log\frac{(1-\rho)^2}{1-2\lambda}, \tag{8.32}$$

thus we need to solve

$$\max_{\lambda:0<\lambda\leq\frac{\rho}{1+\rho}} f(\lambda). \tag{8.33}$$

The function $f$ is concave for $0 < \lambda \leq \frac{\rho}{1+\rho}$ because

$$\frac{\partial^2 f}{\partial \lambda^2} = \frac{1}{(1-\lambda)(1-2\lambda)} \leq 0. \tag{8.34}$$

By examining the monotonicity

$$\frac{\partial f}{\partial \lambda} = -\log \frac{(1-\rho)(1-\lambda)}{1-2\lambda} - \alpha \tag{8.35}$$

where the function is concave for $0 < \lambda \leq \frac{\rho}{1+\rho}$, thus the maximum value is when the first order derivative vanishes,

$$\lambda^* = \frac{e^{-\alpha} - 1 + \rho}{2e^{-\alpha} - 1 + \rho}. \tag{8.36}$$

Thus, we have

$$I_D(\alpha) \leq -f(\lambda^*) = \alpha - \frac{1}{2} \log \frac{1+\rho}{2e^{-\alpha} - 1 + \rho}. \tag{8.37}$$

The validity condition is $0 < \lambda^* \leq \frac{\rho}{1+\rho}$, that is

$$\rho \geq 1 - e^{-\alpha}. \tag{8.38}$$

The result for negative $\rho$ follows similarly.

### 8.4.2 Proof of Proposition 17

Let us assume that $\rho \geq 0$. In order to prove Proposition 17, Item 1 we need to show that

$$\frac{1}{2} \log \frac{1}{1-\rho^2} \geq \alpha - \frac{1}{2} \log \frac{1+\rho}{2e^{-\alpha} - 1 + \rho}, \tag{8.39}$$

holds for $\rho \geq 1 - e^{-\alpha}$. The above inequality is simplified as follows

$$(1-\rho)(2e^{-\alpha} - 1 + \rho) \leq e^{-2\alpha}, \tag{8.40}$$

that is further simplified as follows

$$(e^{-\alpha} - 1 + \rho)^2 \geq 0. \tag{8.41}$$

Equation (8.41) holds for any $\rho$ and $\alpha$.

In order to prove Proposition 17, Item 2 we define

$$h(\alpha) := \rho^2 \frac{\alpha}{2} - \alpha + \frac{1}{2} \log \frac{1+\rho}{2e^{-\alpha} - 1 + \rho}. \tag{8.42}$$

The function $h$ is a convex function of $\alpha$ because

$$\frac{\partial^2 h}{\partial \alpha^2} = \frac{(1-\rho)e^{-\alpha}}{2e^{-\alpha} - 1 + \rho} > 0. \tag{8.43}$$

The possible value of $\alpha$ are $0 \leq \alpha \leq \log \frac{1}{1-\rho}$. Observe that $h(0) = 0$. Note that $h\left(\log \frac{1}{1-\rho}\right) < 0$ combined with the convexity of function $h$ and $h(0) = 0$ is sufficient to establish that $h(\alpha) \leq 0$. Now we need to show that $h\left(\log \frac{1}{1-\rho}\right) < 0$ for $\rho \leq 1 - e^{-1}$. Let us compute the function $h$ at $\alpha = \log \frac{1}{1-\rho}$, that is

$$h\left(\log \frac{1}{1-\rho}\right) = \frac{1}{2}\log(1+\rho) + \frac{1-\rho^2}{2}\log(1-\rho). \tag{8.44}$$

To show that $h\left(\log \frac{1}{1-\rho}\right) < 0$, we need to prove that

$$(1+\rho)^{\frac{1}{1+\rho}} \leq \left(\frac{1}{1-\rho}\right)^{1-\rho}, \tag{8.45}$$

that is obtained from (8.44). Let us define, $x_1 = 1+\rho$ and $x_2 = \frac{1}{1-\rho}$ where $x_1 < x_2$. In addition, the constraint $\rho \leq 1-e^{-1}$, implies that $x_2 < e$. In order to prove (8.45), we need to show that $g(x) = x^{\frac{1}{x}}$ is an increasing function for $0 < x < e$. We have,

$$\frac{\partial g}{\partial x} = \frac{1 - \log x}{x^2} x^{\frac{1}{x}} > 0, \tag{8.46}$$

for $0 < x < e$, thus completing the proof.

# 9 Conclusion

In this thesis, we solve two long-standing open problems such as Gaussian multiple access channel with feedback and Gaussian lossy Gray-Wyner network. For the Gaussian multiple access channel with feedback we compute the sum-Capacity under a symmetric block power constraint by proving the optimality of Gaussian auxiliaries. For Gaussian lossy Gray-Wyner network we compute the rate region of common rate versus sum of the private rates under symmetric mean-squared error distortion by proving the jointly Gaussian optimality of the auxiliary random variables. Further, we study a relaxed variant of Wyner's common information that has an operational meaning on Gray-Wyner network. We compute the relaxed Wyner's common information for Gaussian random vectors and work out lower bounds for any given distribution of the source, that are tight in certain cases. We build on relaxed Wyner's common information to devise a novel algorithm, the so-called common information component analysis. The proposed algorithm is able to extract common features when two or more data are involved. The examples indicate that common information component analysis is dominant to other methods when extracting common information between two or more data. Future research directions include the followings.

- Asymmetric Gaussian multiple access channel with feedback: We managed to compute the sum-Capacity when symmetric block power constraint is imposed on each user. How to deal with asymmetric block power constraint case? In the asymmetric case, we have derived a converse bound to the sum-Capacity, that does not meet the achievable scheme of Kramer [12]. We believe that the achievable scheme can be improved.

- Asymmetric Gaussian lossy Gray-Wyner network: We managed to compute the rate region of central rate versus sum of the private rates under symmetric distortion. A natural extension is to consider the asymmetric distortion case. Moreover, instead of the central rate versus sum of the private rates it would be interesting to consider the weighted sum of rates, that is to fully characterize the rate region of Gaussian lossy Gray-Wyner network.

- Common information component analysis of multiple data: Common information component analysis is build upon relaxed Wyner's common information, which is computationally expensive. So far we have explored the algorithm for two synthetic sets of data to demonstrate its potential. An interesting direction is to seek an efficient implementation of common information component analysis for two data sets or more.

# Bibliography

[1] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Tehnical Journal*, vol. 27, pp. 379–423, 1948.

[2] M. Kac, “On a characterization of the normal distribution,” *American Journal of Mathematics*, vol. 61, pp. 726–728, 1939.

[3] S. N. Bernstein, “On a property which characterizes a Gaussian distribution,” *Proceedings of the Leningrad Polytechnic Institute*, vol. 217, no. 3, pp. 21–22, 1941.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2005.

[5] L. Ozarow, “The capacity of the white Gaussian multiple access channel with feedback,” *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 623 – 629, July 1984.

[6] P. Elias, “Channel capacity without coding,” MIT Research Laboratory of Electronics, Cambridge, MA, Quarterly Progress Report, October 1956.

[7] J. Schalkwijk and T. Kailath, “A coding scheme for additive noise channels with feedback–I: No bandwidth constraint,” *IEEE Trans. Inf. Theory*, vol. 12, pp. 172 – 182, April 1966.

[8] R. G. Gallager and B. Nakiboglu, “Variations on a theme by Schalkwijk and Kailath,” *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 6–17, January 2010.

[9] J. Thomas, “Feedback can at most double Gaussian multiple access channel capacity,” *IEEE Trans. Inf. Theory*, vol. 33, no. 5, pp. 711–716, September 1987.

[10] M. S. Iacobucci and M. G. D. Benedetto, “A feedback code for the multiple access channel (MAC): a case study,” in *IEEE Glob. Comm. Conf., Comm. Theory Mini Conf.*, Phoenix, AZ, Nov. 1997, pp. 128–132.

[11] ——, “Design and performance of a code for the multiple access channel with feedback,” in *IEEE Int. Symp. Inform. Theory*, Cambridge, MA, Aug. 1998, p. 125.

[12] G. Kramer, “Feedback strategies for white Gaussian interference networks,” *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1423–1438, June 2002.

[13] O. Shayevitz and M. Feder, “Optimal feedback communication via posterior matching,” *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1221, Mar. 2011.

[14] L. V. Truong, “Posterior matching scheme for Gaussian multiple access channel with feedback,” in *IEEE Inf. Theory Workshop*, Hobart, TAS, Australia, November 2014, pp. 476–480.

[15] G. Kramer and M. Gastpar, “Dependence balance and the Gaussian multiaccess channel with feedback,” in *IEEE Inf. Theory Workshop*, Punta del Este, Uruguay, March 2006, pp. 198–202.

[16] A. Hekstra and F. Willems, “Dependence balance bounds for single-output two-way channels,” *IEEE Trans. Inf. Theory*, vol. 35, pp. 44 – 53, January 1989.

[17] E. Ardestanizadeh, M. A. Wigger, Y.-H. Kim, and T. Javidi, “Linear-feedback sum-capacity for Gaussian multiple access channels,” *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 224–236, January 2012.

[18] R. Tandon and S. Ulukus, “Outer bounds for multiple-access channels with feedback using dependence balance,” *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4494–4507, October 2009.

[19] G. Kramer, *Directed Information for Channels with Feedback.* Konstanz: Hartung-Gorre Verlag, 1998, ETH Series in Information Processing, Vol. 11.

[20] ——, “Capacity results for the discrete memoryless network,” *IEEE Trans. Inf. Theory*, vol. 49, pp. 4–21, January 2003.

[21] T. M. Cover and C. S. K. Leung, “An achievable rate region for the multiple access channel with feedback,” *IEEE Trans. Inf. Theory*, vol. 27, no. 3, pp. 292–298, May 1981.

[22] M. Gastpar and G. Kramer, “On cooperation via noisy feedback,” in *Int. Zurich Seminar*, Zurich, Switzerland, Feb. 2006, pp. 146–149.

[23] A. Lapidoth and M. Wigger, “Conditional and relevant common information,” in *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Eilat, Israel, November 2016.

[24] R. Tandon and S. Ulukus, “On the capacity region of the Gaussian multiple access channel with noisy feedback,” in *IEEE Int. Conf. Commun. (ICC)*, Dresden, Germany, June 2009.

[25] ——, “Dependence balance based outer bounds for Gaussian networks with cooperation and feedback,” *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4063–4086, June 2011.

[26] M. Gastpar and G. Kramer, “On noisy feedback for interference channels,” in *Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Nov. 2006, pp. 216–220.

[27] Y. Geng and C. Nair, “The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages,” *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2087 – 2104, April 2014.

[28] T. Courtade, “Strengthening the entropy power inequality,” in *IEEE Int. Symp. Inf. Theory*, Barcelona, Spain, July 2016.

[29] E. Lieb, “Gaussian kernels have only Gaussian maximizers,” *Inventiones mathematicae*, pp. 179–208, 1990.

[30] E. Carlen, “Superadditivity of Fisher’s information and logarithmic Sobolev inequalities,” *J. Functional Analysis*, vol. 101, no. 1, pp. 194–211, Oct. 1991.

[31] D. D. Boos, “On a converse to Scheffe’s theorem,” *The Annals of Statistics*, vol. 13, no. 1, pp. 423–427, 1985.

[32] M. Godavarti and A. Hero, “Convergence of differential entropies,” in *IEEE Int. Symp. Inf. Theory*, vol. 50, no. 1, January 2004, pp. 171–176.

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

[34] A. Wyner, “The common information of two dependent random variables,” *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163–179, March 1975.

[35] R.M.Gray and A.D.Wyner, “Source coding for a simple network,” *Bell Syst. Tech. J*, vol. 53, no. 9, pp. 1681–1721, 1974.

[36] G. Xu, W. Liu, and B. Chen, “Wyner’s common information for continuous random variables - a lossy source coding interpretation,” in *Annual Conference on Information Sciences and Systems*, Baltimore, MD, USA, March 2011.

[37] ——, “A lossy source coding interpretation of Wyner’s common information,” *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 754 – 768, February 2016.

[38] P. Yang and B. Chen, “Wyner’s common information in Gaussian channels,” in *IEEE International Symposium on Information Theory*, Honolulu, HI, USA, 2014, pp. 3112–3116.

[39] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, “A local characterization for Wyner common information,” in *IEEE International Symposium on Information Theory*, Los Angeles, California, USA, June 2020.

[40] H. S. Witsenhausen, “Values and bounds for the common information of two discrete random variables,” *SIAM J. Appl. Math*, vol. 31, no. 2, pp. 313–333, September 1976.

[41] L. Yu and V. Y. F. Tan, “Wyner’s common information under Rényi divergence measures,” *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3616–3632, 2018.

[42] G. O. Veld and M. Gastpar, "Total correlation of Gaussian vector sources on the Gray-Wyner network," in *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, USA, September 2016.

[43] C.-Y. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching: Sequential coding for computing," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6393 – 6406, August 2016.

[44] S. Satpathy and P. Cuff, "Gaussian secure source coding and Wyner's common information," in *2015 IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, China, June 2015, pp. 116–120.

[45] L. Yu, H. Li, and C. W. Chen, "Generalized common informations: Measuring commonness by the conditional maximal correlation," *CoRR*, vol. abs/1610.09289, 2016. [Online]. Available: https://arxiv.org/abs/1610.09289

[46] G. R. Kumar, C. T. Li, and A. E. Gamal, "Exact common information," in *IEEE International Symposium on Information Theory (ISIT)*, Honolulu, HI, USA, August 2014.

[47] M. Gastpar and E. Sula, "Common information components analysis," in *Proceedings of the 2020 Information Theory and Applications (ITA) Workshop*, San Diego, USA, February 2020.

[48] C. Nair, "An extremal inequality related to hypercontractivity of Gaussian random variables," in *Proceedings of the Information Theory and Applications Workshop (ITA)*, San Diego, CA, USA, February 2014, pp. 1–7.

[49] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Transactions on Information Theory*, vol. 21, no. 6, pp. 629–637, 1975.

[50] E. Posner, "Random coding strategies for minimum entropy," *IEEE Transactions on Information Theory*, vol. 21, no. 4, pp. 388 – 391, 1975.

[51] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[52] K. B. Viswanatha, E. Akyol, and K. Rose, "The lossy common information of correlated sources," *IEEE Transactions on Information Theory*, vol. 60, no. 6, pp. 3238 – 3253, June 2014.

[53] G. J. Op 't Veld and M. C. Gastpar, "A Gaussian source coding perspective on caching and total correlation," Ph.D. dissertation, EPFL, 2017.

[54] E. Sula and M. Gastpar, "Common information components analysis," *Entropy Special Issue on The Role of Signal Processing and Information Theory in Modern Machine Learning*, vol. 23, no. 2, 2021.

[55] J. Liu, "Information theory from a functional viewpoint," Ph.D. dissertation, Princeton University, 2018.

[56] E. Sula and M. Gastpar, "On Wyner's common information in the Gaussian case," *CoRR*, vol. abs/1912.07083, 2019. [Online]. Available: http://arxiv.org/abs/1912.07083

[57] H. Hotelling, "Relations between two sets of variants," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, December 1936.

[58] S.-L. Huang, G. W. Wornell, and L. Zheng, "Gaussian universal features, canonical correlations, and common information," in *2018 IEEE Information Theory Workshop (ITW)*, November 2018.

[59] M. Gastpar and E. Sula, "Relaxed Wyner's common information," in *Proceedings of the 2019 IEEE Information Theory Workshop*, Visby, Sweden, August 2019.

[60] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *Journal of Machine Learning Research*, vol. 6, pp. 165–188, 2005.

[61] F. Bach and M. Jordan, "A probabilistic interpretation of canonical correlation analysis," University of California, Berkeley, Department of Statistics, Tech. Rep. 688, April 2005.

[62] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *J. Am. Stat. Assoc.*, vol. 80, no. 391, pp. 580–598, September 1985.

[63] P. Comon, "Independent component analysis," in *Internat. Signal Processing Workshop on High-Order Statistics*, Chamrousse, France, July 1991, pp. 111–120.

[64] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 5, pp. 493–501, September 1975.

[65] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *The 37th annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, U.S.A., Sep. 1999, pp. 368–377.

[66] B. H. Xiao-Tong Yuan and B.-G. Hu, "Robust feature extraction via information theoretic learning," in *26th Annual International Conference on Machine Learning*, June 2009.

[67] H. Wang and P. Chen, "A feature extraction method based on information theory for fault diagnosis of reciprocating machinery," *Sensors*, vol. 9, no. 4, pp. 2415–2436, 2009.

[68] J. Bi, K. P. Bennett, M. Embrechts, C. M. Breneman, and M. Song, "Dimensionality reduction via sparse support vector machines," *Journal of Machine Learning Research*, vol. 3, pp. 1229–1243, 2003.

[69] V. Laparra, J. Malo, and G. Camps-Valls, “Dimensionality reduction via regression in hyperspectral imagery,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1026 – 1036, 2015.

[70] J. Gao, Q. Shi, and Caetano, “Dimensionality reduction via compressive sensing,” *Pattern Recognition Letters*, vol. 33, no. 9, pp. 1163–1170, 2012.

[71] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, USA, June 2006.

[72] Z. Zhang and H. Zha, “Principal manifolds and nonlinear dimensionality reduction via tangent space alignment,” *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.

[73] P. Vepakomma, “Supervised dimensionality reduction via distance correlation maximization,” *Electronic Journal of Statistics*, vol. 12, no. 1, pp. 960–984, 2018.

[74] J. Wu, J. Wang, and L. Liu, “Feature extraction via KPCA for classification of gait patterns,” *Human Movement Science*, vol. 26, no. 3, pp. 393–411, 2007.

[75] Z. Lai, D. Mo, W. K. Wong, Y. Xu, D. Miao, and D. Zhang, “Robust discriminant regression for feature extraction,” *IEEE Transactions on Cybernetics*, vol. 48, no. 8, pp. 2472 – 2484, August 2018.

[76] H. Wang, Y. Zhang, N. R. Waytowich, D. J. Krusienski, G. Zhou, J. Jin, X. Wang, and A. Cichocki, “Discriminative feature extraction via multivariate linear regression for SSVEP-based BCI,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 5, pp. 532 – 541, May 2016.

[77] S. Watanabe, “Information theoretical analysis of multivariate correlation,” *IBM Journal of Research and Development*, vol. 4, no. 1, pp. 66–82, Jan 1960.

[78] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, “On hypercontractivity and a data processing inequality,” in *IEEE International Symposium on Information Theory*, Honolulu, HI, USA, July 2014.

# Curriculum Vitae

## Personal Information

| | |
|---|---|
| Address | Chemin des Paleyres 10, 1006 Lausanne, Switzerland |
| E-mail | erixhensula@yahoo.com |
| Phone | +41774397399 |
| Nationality | Albanian |

## Education

| | |
|---|---|
| 2009 – 2013 | B.Sc in Electrical and Electronics Engineering, Orta Doğu Teknik Üniversitesi, Turkey. |
| 2013 – 2016 | M.Sc in Communication Systems, École Polytechnique Fédérale de Lausanne, Switzerland. |
| 2016 – 2021 | Ph.D. in Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Switzerland.<br>*Supervisor*: Prof. Michael Gastpar |

## Internship Experience

| | |
|---|---|
| Feb 2015 – Jul 2015 | *Topic*: Powerline communication on long-distance high-voltage lines. ABB Schweiz AG company, Baden Switzerland.<br>*Supervisor*: Dr. Dacfey Dzung |
| Jul 2018 – Sep 2018 | *Topic*: Optimal Supply Power Allocation Capacity for Optical Fibers using Multi Layer Neural Networks. Nokia BELL Labs, New Jersey USA. *Supervisor*: Dr. Junho Cho. |

## Teaching Experience

- Signals and Systems II, spring 2018, spring 2019 and spring 2020.
- Information theory and coding, autumn 2018.
- Modern digital communications: a hands-on approach, autumn 2017 and autumn 2019.
- Probabilities and statistics, spring 2017.

## Awards

| | |
|---|---|
| 2010 | Dr. Bülent Kerim Altay Success Award, Orta Doğu Teknik Üniversitesi, Turkey. (award for maximum grades) |
| 2016 | EDIC Computer and Communication Sciences Fellowship, École Polytechnique Fédérale de Lausanne, Switzerland. |
| 2018 | TPC-Chairs-Choice-Sessions, IEEE International Symposium on Information Theory, Vail, Colorado. |
| 2020 | Early Postdoc.Mobility Fellowship, Swiss National Science Foundation. |

## Publication

### Conferences

- M. Gastpar and E. Sula, “Common information components analysis,” in *Proceedings of the 2020 Information Theory and Applications (ITA) Workshop*, San Diego, USA, February 2020.

- E. Sula and M. Gastpar, “The Gaussian lossy Gray-Wyner network,” in *54th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, 2020.

- J. Cho, G. Raybon, E. Burrows, J.-C. Antona, N. K. Fontaine, R. Ryf, H. Chen, S. Chandrasekhar, E. Sula, S. Olsson, S. Grubb, and P. J. Winzer, “Optimizing gain shaping filters with neural networks for maximum cable capacity under electrical power constraints,” in *European Conference on Optical Communications (ECOC)*, Brussels, Belgium, December 2020.

- J. Cho, S. Chandrasekhar, E. Sula, S. Olsson, E. Burrows, G. Raybon, R. Ryf, N. Fontaine, J.-C. Antona, S. Grubb, P. Winzer, and A. Chraplyvy, “Maximizing fiber cable capacity under a supply power constraint using deep neural networks,” in *Optical Fiber Communications Conference and Exhibition (OFC)*, San Diego, CA, USA, March 2020.

- M. Gastpar and E. Sula, “Relaxed Wyner’s common information,” in *Proceedings of the 2019 IEEE Information Theory Workshop*, Visby, Sweden, August 2019.

- E. Sula, M. Gastpar, and G. Kramer, “Sum-rate capacity for symmetric Gaussian multiple access channels with feedback,” in *IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, USA, June 2018.

- E. Sula, J. Zhu, A. Pastore, S. H. Lim, and M. Gastpar, Compute-forward multiple access (CFMA) with nested LDPC codes, in *IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, June 2017.

## Journals

- E. Sula and M. Gastpar, “Common information components analysis,” *Entropy Special Issue on The Role of Signal Processing and Information Theory in Modern Machine Learning*, vol. 23, no. 2, 2021.
- J. Cho, S. Chandrasekhar, E. Sula, S. Olsson, E. Burrows, G. Raybon, R. Ryf, N. Fontaine, J.-C. Antona, S. Grubb, P. Winzer, and A. Chraplyvy, “Supply-power-constrained cable capacity maximization using multi-layer neural networks,” *Journal of Lightwave Technology*, vol. 38, no. 14, pp. 3652 – 3662, 2020.
- E. Sula, M. Gastpar, and G. Kramer, “Sum-rate capacity for symmetric Gaussian multiple access channels with feedback,” *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 2860 – 2871, 2020.
- E. Sula, J. Zhu, A. Pastore, S. H. Lim, and M. Gastpar, “Compute-forward multiple access (CFMA): practical implementations,” *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1133 – 1147, 2019.

## Preprints

- E. Sula and M. Gastpar, “On Wyner’s common information in the Gaussian case,” *CoRR*, vol. abs/1912.07083, 2019. [Online]. Available: `http://arxiv.org/abs/1912.07083`
- E. Sula and M. Gastpar, “Lower bound on Wyner’s common information,” *CoRR*, vol. abs/2102.08157, 2021. [Online]. Available: `https://arxiv.org/abs/2102.08157`