NNT : 2017SACLS047

THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
L'UNIVERSITE PARIS-SUD
AU SEIN DU LABORATOIRE DES SIGNAUX ET SYSTEMES

ECOLE DOCTORALE N°580 STIC
Sciences et Technologies de l'Information et de la Communication

Réseaux, Information et Communications

Par

Mlle Maggie Mhanna

Privacy-Preserving Quantization Learning for Distributed Detection with Applications to Smart Meters

Thèse présentée et soutenue à Gif-sur-Yvette, le 13/01/2017 :

Après Avis des Rapporteurs :
Mme Inbar FIJALKOW (Professeur, ETIS, Université Cergy-Pontoise)
M. Matthieu BLOCH (Professeur Associé, Georgia Institute of Technology)

Composition du Jury :

Mme Marie-Laure BOUCHERET	Professeur, Université de Toulouse	Présidente
M. Matthieu BLOCH	Professeur Associé, Georgia Institute of Technology	Rapporteur
Mme Catuscia PALAMIDESSI	Directeur de Recherches, INRIA, Saclay and LIX	Examinateur
M. Pierre DUHAMEL	Professeur, Université Paris-Saclay	Directeur de thèse
M. Pablo PIANTANIDA	Professeur Associé, Université Paris-Saclay	Directeur de thèse

ABSTRACT

Abstract — *This thesis investigates source coding problems in which some secrecy should be ensured with respect to eavesdroppers.*

In the first part, we provide some new fundamental results on both detection and secrecy oriented source coding in the presence of side information at the receiving terminals. We provide several new results of optimality and single-letter characterization of the achievable rate-error-equivocation region, and propose practical algorithms to obtain solutions that are as close as possible to the optimal, which requires the design of optimal quantization in the presence of an eavesdropper. In the second part, we study the problem of secure estimation in a utility-privacy framework where the user is either looking to extract relevant aspects of complex data or hide them from a potential eavesdropper.

The objective is mainly centered on the development of a general framework that combines information theory with communication theory, aiming to provide a novel and powerful tool for security in Smart Grids. From a theoretical perspective, this research was able to quantify fundamental limits and thus the tradeoff between security and performance (estimation/detection).

Résumé — *Cette thèse porte sur quelques problèmes de codage de source dans lesquels on souhaite préserver la confidentialité vis à vis d'une écoute du canal.*

Dans la première partie, nous fournissons des nouveaux résultats fondamentaux sur le codage de source pour la détection (utilisateur légitime) et la confidentialité (vis à vis d'une écoute du canal) en présence d'informations secondaires aux terminaux de réception. Nous proposons plusieurs nouveaux résultats d'optimisation de la région de débit-erreur-équivocation réalisable, et proposons des algorithmes pratiques pour obtenir des solutions aussi proches que possible de l'optimal, ce qui nécessite la conception de quantificateurs en présence d'un eavesdropper. Dans la deuxième partie, nous étudions le problème de l'estimation sécurisée dans un cadre d'utilité-confidentialité où l'utilisateur recherche soit à extraire les aspects pertinents de données complexes ou bien à les cacher vis à vis d'un eavesdropper potentiel.

L'objectif est principalement axé sur l'élaboration d'un cadre général qui combine la théorie de l'information et la théorie de la communication, visant à fournir un nouvel outil pour la confidentialité dans les Smart Grids. D'un point de vue théorique, cette recherche a permis de quantifier les limites fondamentales et donc le compromis entre sécurité et performance (estimation / détection).

PUBLICATIONS

The material contained in this thesis appeared in the following publications.

Journal Papers

- ✓ *M. Mhanna, P. Piantanida and P. Duhamel*
  "Hypothesis testing with communication and security constraints",
  In preparation (see chapter 2).

- ✓ *M. Mhanna, P. Piantanida and P. Duhamel*
  "Quantization learning for distributed binary decision with privacy constraints",
  In preparation (see chapter 3).

- ✓ *M. Mhanna, P. Duhamel and P. Piantanida*
  "Quantization learning for a utility-privacy framework",
  In preparation (see chapter 4).

Conference Papers

- ✓ *Maggie Mhanna, Pablo Piantanida*
  "On secure distributed hypothesis testing",
  2015 IEEE International Symposium on Information Theory (ISIT),
  Hong Kong, Hong Kong SAR China, Jun 2015, (see chapter 2).

- ✓ *Maggie Mhanna, Pablo Piantanida, and Pierre Duhamel*
  "Quantization for distributed binary detection under secrecy constraints",
  2016 IEEE International Conference on Communications (ICC),
  Kuala Lumpur, Malaysia, Malaysia, May 2016, (see chapter 3).

- ✓ *Maggie Mhanna, Pierre Duhamel, and Pablo Piantanida*
  "Privacy-Preserving Quantization Learning for Distributed Binary Decision with Applications to Smart Meters",
  Submitted to 2017 IEEE International Conference on Communications (ICC), (see chapter 3).

- ✓ *Maggie Mhanna, Pablo Piantanida, and Pierre Duhamel*
  "Privacy-Preserving Quantization Learning with Applications to Smart Meters",
  2017 IEEE International Conference on Communications (ICC),
  Paris, France, May 2017, (see chapter 4).

CONTENTS

1 Introduction **1**

1.1 A Historical Review . . . 1
1.2 Elements of Practical Communication Systems . . . 3
1.3 Source Coding and Quantization . . . 5
1.3.1 Distributed Source Coding . . . 8
1.3.2 Detection-Oriented Source Coding . . . 9
1.3.3 Secrecy-Oriented Source Coding . . . 10
1.4 Secure Distributed Source Coding with Applications to Smart Meters . . . 13
1.4.1 Introduction . . . 13
1.4.2 Smart Grid Communication Infrastructure . . . 14
1.4.3 Smart Metering . . . 15
1.4.4 Automated control and incident detection (system surveillance) . . . 16
1.4.5 Information Security for the Smart Grid . . . 19

I Secure Distributed Binary Detection **21**

2 Hypothesis Testing with Communication and Security Constraints **23**

2.1 Introduction . . . 23
2.2 Problem Definition . . . 25
2.2.1 Notations . . . 25
2.2.2 System Model . . . 26
2.3 General Hypothesis Testing . . . 27
2.3.1 Rate-Error-Equivocation Region for the general HT . . . 28
2.3.2 Degraded Hypothesis with an Arbitrary Large Coding Rate . . . 28
2.4 Testing Against Independence . . . 29
2.4.1 Single-Letter Rate-Error-Equivocation Region . . . 30
2.4.2 Eve is the Statistician . . . 30
2.4.3 X follows an arbitrary distribution under Hypothesis H_1 . . . 31
2.4.4 Testing on Binary Sources . . . 32
2.4.5 Testing on Gaussian Sources . . . 33
2.4.6 Numerical Results . . . 33
2.5 Conclusion . . . 34

3 Quantization Learning for Distributed Binary Decision with Privacy Constraints 35
3.1 Introduction . . . 35
3.2 Problem Definition . . . 38
3.2.1 Notations . . . 38
3.2.2 System Model . . . 39
3.3 Scalar Quantization . . . 40
3.3.1 Quantizer Design . . . 41
3.4 Vector Quantization . . . 44
3.4.1 VQ Algorithm for Detection . . . 45
3.4.2 VQ Algorithm under Privacy Constraints . . . 47
3.4.3 Probabilities of Error . . . 50
3.5 Testing Against Independence with Memoryless Gaussian Sources . . . 51
3.5.1 Scalar Quantization of τ samples . . . 52
3.5.2 Vector Quantization . . . 54
3.6 Pratical Application: Smart Meter Fault Detection under Privacy Contraints . . . 60
3.6.1 Training Set . . . 62
3.6.2 Results . . . 64
3.7 Conclusion . . . 65

II Secure Lossy Estimation 69

4 Quantization Learning for A Utility-Privacy Framework 71
4.1 Introduction . . . 71
4.2 Lossy Source Coding with a Privacy Constraint . . . 73
4.2.1 Problem Definition . . . 73
4.2.2 Optimal Quantizer Design . . . 74
4.3 Lossy Source Coding of a Correlated Relevant Information with a Privacy Constraint . 76
4.3.1 Problem Definition . . . 76
4.3.2 Quantizer Design . . . 78
4.4 Application: Memoryless Gaussian Sources . . . 78
4.5 Pratical Application: Smart Meter Device Consumption Recovery under Privacy Constraints . . . 83
4.6 Conclusion . . . 86

5 Conclusion 89
5.1 General Comments . . . 89
5.2 Further Directions in the Detection Framework . . . 89
5.2.1 M-ary Hypothesis Testing . . . 89
5.2.2 Hypothesis Testing with Coded Side Information . . . 90

5.2.3 Decentralized Hypothesis Testing Network 91
5.3 Further Directions in the Estimation Framework 92

Appendices **95**

A Auxiliary Proofs of Chapter 2 **97**
A.1 Proof of Proposition 2.1 97
A.2 Proof of Proposition 2.2 104
A.2.1 Direct Part 105
A.2.2 Converse Part 105
A.3 Proof of Proposition 2.3 106
A.3.1 Proof of the Achievability Part 106
A.3.2 Proof of the Converse Part 107
A.4 Proof of Proposition 2.5 108
A.5 Proof of Proposition 2.6 110

III Synthèse de la Thèse, en Français **111**

6 Synthèse de la Thèse, en Français **113**
6.1 Introduction 113
6.2 Test d'Hypothèse avec Contraintes de Communication et de Sécurité 118
6.3 Apprentissage de Quantification pour une Décision Binaire Distribuée avec Contraintes de Confidentialité 123
6.4 Apprentissage de la Quantification pour un cadre de Utilité-Confidentialité 130

LIST OF FIGURES

1.1 Shannon's model of a memoryless communication channel. 3
1.2 Basic elements of a digital communication system. 4
1.3 Scalar quantization regions and representation points. 7
1.4 Secure communication system for lossy estimation of the source. 11
1.5 Smart metering communication. 15
1.6 Distributed HT against independence used for fault detection 17

2.1 HT with communication and privacy constraints. 24
2.2 Testing on binary sources. 32
2.3 Testing on Gaussian sources. 33
2.4 Error exponent as a function of the equivocation rate for different coding rates. 34
2.5 Error exponent as a function of the equivocation rate for different correlation values ρ_Y^2 with an arbitrary large coding rate. 34

3.1 Distributed detection under privacy constraints. 39
3.2 In (a), we show *Bhattacharyya distance* versus information leakage for different values of τ at rate $R = 1$. In (b), we consider different values of ρ_Z. 54
3.3 Type II vs type I probability of error. In (a), different combinations of rate-leakage scenarios are considered at $\tau = 1$. (b) shows the impact of the allowed information leakage on performance when $\tau = 10$. (c) shows the impact of the side information at Eve on performance-privacy trade-off. 55
3.4 Results for VQ optimal regions 58
3.5 Performance-privacy trade-offs for different VQ scenarios 59
3.6 An illustration of the studied model. 61
3.7 Row 1: optimal quantization regions for $n = 2$ (x and y axis are amplitudes 1 and 2 of the vector $\mathbf{x}$ respectively). Row 2: type II vs type I probabilities of error. In (d), we study the impact of the VQ dimension n when $R = 0.5$ and $\Delta \geq R$. In (e), we study the impact of the allowed level of information leakage on performance when $n = 4$ and $R = 0.5$. In all of these figures, $\tau = 5$. 65
3.8 Results for probability distribution fitting with lognormal. First row: illustration of the Kolmogorov-Smirnov statistic. Red line is the estimated lognormal CDF, blue line is ECDF, and the black arrow is the K-S statistic. Second row: empirical univariate distribution vs. fitted log-normal distribution of X and Y for $n = 1$. Third row: joint empirical distribution vs. fitted joint log-normal distribution of (X, Y) for $n = 1$. . . . 66

4.1 Lossy source coding with a privacy constraint. 73
4.2 Secure lossy coding of a correlated relevant information with privacy constraints. . . . 76
4.3 Results for VQ optimal regions 80

5.1 Distributed detection with coded side information under privacy constraints. 90
5.2 Decentralized detection under privacy constraints. 92
5.3 Decentralized estimation under privacy constraints. 93

A.1 Binary auxiliary RV. 109

6.1 Éléments de base d'un système de communication numérique. 114
6.2 Tests d'hypothèses avec contraintes de communication et de confidentialité. 120
6.3 Distributed detection under privacy constraints. 124
6.4 Une illustration du modèle étudié. 128
6.5 Codage de source à perte avec une contrainte de confidentialité. 130
6.6 Codage Source avec Perte d'une information Corrélée Pertinente avec une Contrainte de Confidentialité. 132

LIST OF TABLES

3.1 Optimal decision threshold(s) for different information leakages at $R = 1$, $R = 2$. The values of a_0 and a_M are excluded from the table. $\Delta^\star$ is the critical information leakage. 52
3.2 Optimal decision thresholds for different information leakages and correlations for $R = 1$. 54
3.3 Minimum β_n for type I error $\alpha_n \leq 0.01$ as n varies. 56
3.4 K-S statistic for different distribution laws for the univariate case. 63

4.1 The performance of the VQ for the memoryless Gaussian source when the decoder aims at recovering X^n, i.e., $SDR = 10\log_{10}(\sigma_X^2/D_X)$. 81
4.2 The performance of the VQ for the memoryless Gaussian source when the decoder aims at recovering Y^n, i.e., $SDR = 10\log_{10}(\sigma_Y^2/D_Y)$ when $\rho = 0.8$. 81
4.3 K-S statistic for different distribution laws when $n = 1$. 84
4.4 The performance of the VQ when the decoder aims at recovering X^n, i.e., $SDR = 10\log_{10}(\sigma_X^2/D_X)$. 85
4.5 The performance of the VQ when the decoder aims at recovering Y^n, i.e., $SDR = 10\log_{10}(\sigma_Y^2/D_Y)$. 85

ACKNOWLEDGEMENT

It's been a long road, but here I am at the end, but there are so many people to whom thanks I extend!

I would like to express my deepest appreciation to all those who provided me the possibility to complete this thesis. A special gratitude I give to my supervisors, Pierre Duhamel and Pablo Piantanida, whose contributions in stimulating suggestions and encouragement, helped me to coordinate my thesis. I would like to thank them for their availability, expertise and for providing the best conditions and environment for conducting my research. I could not have imagined having better advisors and mentors for my Ph.D study.

I also want to thank the other people with whom I had the opportunity to work with during the thesis, Prof. Leszek Szczecinski and Prof. Jean-Charles Grégoire for their hospitality during my stay at INRS, Canada, and for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives.

Many thanks go to "L2S" and its staff for their unconditioned help and support throughout the whole academic years.

I would also like to thank the Jury members for validating this work. Thanks to Prof. Inbar Fijalkow and Prof. Matthieu Bloch for rereading and commenting on my manuscript. I would also like to thank Prof. Catuscia Palamidessi and Prof. Marie-Laure Boucheret for accepting to be part of my Jury of examiners.

Last but not least, this journey would not have been possible without the support of my family, and friends. To mom, dad, Jules, Elyssa, and Ayssar, you have been so supreme! Thank you for encouraging me in all of my pursuits and inspiring me to follow my dreams.

CHAPTER 1

INTRODUCTION

1.1 A Historical Review

Today, we get in contact with communication in our daily lives in so many different ways. At home or at work, we get across many devices providing rapid communication from every corner of this world. Through the telephones at our hands, the televisions in our living rooms, or the computers connected to the internet at our offices or homes, we are able to communicate instantaneously with people from all parts of the globe and receive information about various developments and events that occur all around the world.

The communication growth has been unbelievable during the past two decades. Smart hand-held communications devices are now at the disposal of almost every individual and have became an essential part of modern human life. Today, as a result of communication and computing technologies advancements, the term "smart" has been popular recently in different contexts. Smart cities, smart-grid, etc. have become main innovation agendas of research organizations, technology vendors, and governments.

Progress in telecommunications over the past couple of decades has been revolutionary. Communication technology as we see it today became important with telegraphy, then telephony, then video, then computer communication, and the superb mixture of all of these inexpensive, small portable devices at our disposal today [1].

The beginnings of what we know now as modern digital communication stem from the work of Nyquist (1924) [2]. Nyquist developed a model in which a transmitted signal has the general form of

$$s(t) = \sum_n a_n g(t - nT) \tag{1.1}$$

where $g(t)$ represents a basic pulse shape and $\{a_n\}$ is a binary data sequence of $\{\pm 1\}$ transmitted at a rate of $1/T$ bits/s. Nyquist investigated the problem of determining the maximum signaling rate and the optimum pulse shape that can be used over a telegraph channel of a given bandwidth without causing inter-symbol interference. His studies led him to conclude that the maximum pulse rate is 2 times the channel bandwidth. This rate is now called the Nyquist rate.

In the light of Nyquist's work, Hartley (1928) [3] investigated the amount of information that can be transmitted reliably over a band-limited channel when multiple amplitude levels are allowed. Hartley argued that the maximum number of distinguishable pulse levels that can be transmitted and received reliably over a communications channel is limited by the dynamic range of the signal amplitude and

the precision with which the receiver can distinguish amplitude levels.

Another significant advance in the development of communications was the work of Wiener (1942) [4, 5], who considered the problem of estimating a desired signal waveform $s(t)$ in the presence of additive noise $n(t)$, based on observation of the received signal $r(t) = s(t) + n(t)$. This problem arises in signal demodulation. Wiener determined the linear filter whose output is the best mean-square approximation to the desired signal $s(t)$.

Hartley's and Nyquist's results on the maximum transmission rate of digital information were precursors to the work of Claude Shannon; often called the father of the Digital Age. In the beginning of his paper [6] and later his book [7], Shannon acknowledged the work done before him, by such pioneers as Nyquist and Hartley at Bell Labs. Though their influence was profound, it was Shannon who revolutionized communication, and defined a new field of communication research that we now know as Information Theory. One of those key concepts was his definition of the limit for channel capacity.

Information theory is one of the few scientific fields fortunate enough to have an identifiable beginning - Claude Shannon's 1948 paper. The story of the evolution of how it progressed from a single theoretical paper to a broad field that has redefined our world is a fascinating one. For perhaps the first 25 years of its existence, information theory served as a rich source of academic research problems [8, 9, 10, 11] and as a tempting suggestion that its approaches can make communication systems more efficient and more reliable. Aside from small experiments and a few particular military systems, the theory was regarded as a beautiful theory and had little interaction with practice. However, by the mid 1970's, communication systems started implementing information theoretic ideas extensively.

Shannon formulated the basic problem of reliable transmission of information in statistical terms, using probabilistic models for informations sources and communications channels. Based on such a statistical formulation, he adopted a logarithmic measure for the average information content of the source, later called entropy (more specifically, Shannon entropy). The entropy H of a discrete random variable X with possible values $\{x_1, x_2, \dots\}$ and probability mass function $p(X)$ is defined as

$$\mathrm{H}(X) = \mathrm{E}[-\log_b(\mathrm{p}(X))] = -\sum_i \mathrm{p}(x_i)\log_b \mathrm{p}(x_i). \tag{1.2}$$

Here E is the expected value operator, b is the base of the logarithm used. Common values of b are 2. When $b = 2$, the unit of entropy is commonly referred to as bits. When the distribution is continuous rather than discrete, the sum is replaced with an integral.

Shannon also defined the notion of channel capacity and provided a mathematical framework by which one can compute it. The key result states that the capacity of the channel, as defined above, is given by the maximum of the mutual information between the input and output of the channel, where the maximization is with respect to the input distribution. Let U and V be the random variables representing the input and output of the channel, respectively. Let $p_{V|U}(v|u)$ be the conditional distribution function of V given U, which is an inherent fixed property of the memoryless communications channel. Then the choice of the marginal distribution $p_U(u)$ completely determines the joint distribution

$p_{U,V}(u,v)$ which, in turn, induces a mutual information $I(U;V)$. The channel capacity is defined as

$$C = \sup_{p_U} I(U;V) \tag{1.3}$$

The importance of the channel capacity is as follows: If the information rate R from the source is less than the channel capacity C ($R < C$) then it is theoretically possible to achieve reliable error-free transmission through the channel using some appropriate coding. However, if $R > C$, reliable transmission is no longer possible regardless of the amount of signal processing performed at the transmitter and receiver [12].

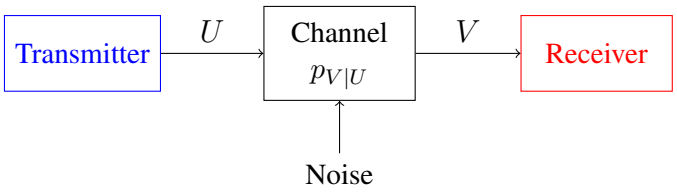


Figure 1.1: Shannon's model of a memoryless communication channel.

Following Shannon's work, came the classic publication of Hamming (1950) on error detection correction to diminish the channel noise impact [13]. Hamming's work inspired many researchers in the years that followed, and a variety of new and powerful codes were discovered, many of which are used today in the implementation of modern communication systems [14, 15].

1.2 Elements of Practical Communication Systems

Information Theory that we know today was not only the work of Claude Shannon but the result of many significant contributions made by different individuals, from a variety of backgrounds, who took Shannon's ideas and expanded upon them. The diversity and directions of their objectives and interests formed the shape of the Information Theory of today.

The increase in demand for transmission during the last three to four decades, couples with the development of more sophisticated integrated circuits, has led to the development of very efficient and more reliable digital communication systems. In the course of these developments, Shannon's original results and the generalization of his results on maximum transmission limits over a channel and on bounds on the performance achieved have served as benchmarks for any communication system design. The theoretical limits derived by Shannon and other researchers that contributed to the development of information theory serve as an ultimate goal in the continuing efforts to design and develop more efficient digital communication systems.

In the most fundamental sense, communication involves implicitly the transmission from one point to another through a succession of processes. Figure 1.2 illustrates the functional diagram and the basic elements of a digital communication system [1].

The design of a communication system was based on two concepts. The first is to view all communication sources as being representable by binary sequences. The second is to design communication

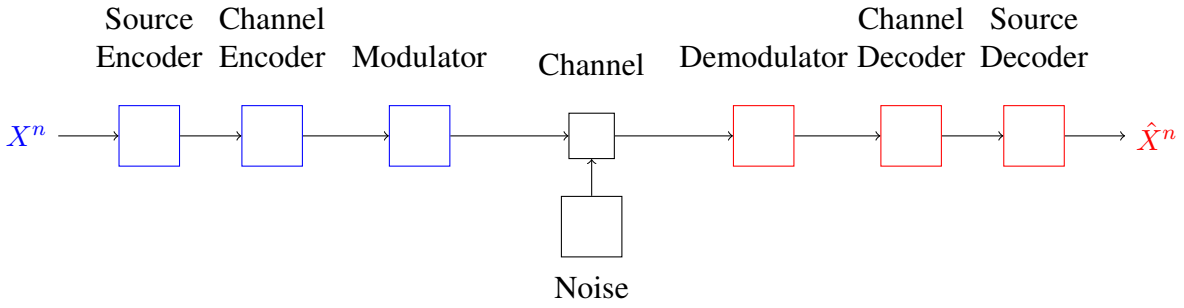


Figure 1.2: Basic elements of a digital communication system.

systems that first convert the source output into a binary sequence and then convert that binary sequence into a form suitable for transmission over particular physical media such as cable, twisted wire pair, optical fiber, or electromagnetic radiation through space.

The source, e.g., speech waveform, image waveform, or text file, may be either an analog signal, such as audio or video signal, or a digital signal, such as the output of a teletype machine, that is discrete in time and has a finite number of output characters.

The messages produced by the source are converted into a sequence of binary digits. Ideally, we would like to represent the source output (message) by a few binary digits as possible. In other words, we seek an efficient representation of the source output that results in little or no redundancy. The process of efficiently converting the output of either an analog or digital source into a sequence of binary digits is called *source coding* or *data compression*.

The sequence of binary digits from the source encoder, which we call the *information sequence*, is passed to the *channel encoder*. The purpose of the channel encoder is to introduce, in a controlled manner, some redundancy in the binary information sequence that can be used at the receiver to overcome the effect of noise and interference encountered during the transmission of the signal through the channel. Thus, the added redundancy serves to increase the reliability of the received data and improves the fidelity of the received signal.

The binary sequence at the output of the channel encoder is passed to the digital modulator, which serves as the interface to the communication channel. Since nearly all communication channels encountered in practice are capable of transmitting electric signals (waveforms), the primary purpose of the digital modulator is to map the binary information sequence into signal waveforms.

The communication channel is a physical medium that is used to send the signal from the transmitter to the receiver. In wireless transmission, the channel maybe the atmosphere (free space). On the other hand, telephone channels usually employ a variety of physical media, including wire lines, optical fiber cables, and wireless (microwave radio). Whatever the physical medium used for transmission of the information, the essential feature is that the transmitted signal is corrupted in a random manner by a variety of possible mechanisms, such as additive thermal noise generated by electronic devices, man-made noise, example, auto Mobile ignition noise, and atmospheric noise, electrical lightning discharges during thunderstorms, etc.

At the receiving end of a digital communication system, the digital demodulator processes the

channel corrupted transmitted waveform and reduces the waveforms to a sequence of numbers that represent estimates of the transmitted data symbols. the sequence of numbers is passed to the channel decoder, which attempts to reconstruct the original information sequence from knowledge of the code used by the channel encoder and the redundancy contained in the received data.

As a final step, the source decoder accepts the output sequence from the channel decoder and, from knowledge of the source encoding method used, attempts to reconstruct the original signal from the source. Due to channel decoding errors and possible distortion introduced by the source encoder and, perhaps, the source decoder, the signal of the output of the source decoder is an approximation to the original source output. The difference or some function of the difference between the original signal and the reconstructed signal is a measure of the distortion introduced by the digital communication system.

1.3 Source Coding and Quantization

In this thesis, our main focus will be on the *source coding* component. Optimal communication systems for source transmission could be constructed by separately designing source codes for the source and error correcting codes for the channel. Hence a work such as this one that focuses only on the source or signal coding aspects, does not inherently mean a loss of generality. An effective overall system can always be constructed by cascading a good signal coding system with a good error control system. In fact, most practical communication systems for source transmission today are based on the separation.

The theory of source coding was first formulated by Shannon where a minimum rate of lossless data compression was established. This rate is the same as the entropy rate H of the source that was defined earlier. The exact value of this rate depends on the information source, more specifically, the statistical nature of the source. It is possible to compress the source, in a lossless manner, with compression rate close to H.

Shannon also developed the theory of lossy data compression. This is better known as rate-distortion theory [16]. In lossy data compression, the decompressed data does not have to be exactly the same as the original data. Instead, some amount of distortion, D, is tolerated. Shannon showed that, for a given source (with all its statistical properties known) and a given distortion measure, there is a function, $R(D)$, called the rate-distortion function. The theory says that if D is the tolerable amount of distortion, then $R(D)$ is the best possible compression rate.

When the compression is lossless (i.e., no distortion or $D = 0$), the best possible compression rate is $R(0) = H$ (for a finite alphabet source). In other words, the best possible lossless compression rate is the entropy rate. In this sense, rate-distortion theory is a generalization of lossless data compression theory, where we went from no distortion ($D = 0$) to some distortion ($D > 0$).

Lossless data compression theory and rate-distortion theory collectively form the source coding theory. Source coding theory sets fundamental limits on the performance of all data compression algorithms. The theory, in itself, does not specify exactly how to design and implement these algorithms. It does, however, provide some hints and guidelines on how to achieve optimal performance.

Quantization; a form of lossy compression method; maps amplitude values into a discrete range, so that the quantized signal takes on only a discrete, usually finite, set of values. Therefore, quantization results in loss of information by introducing distortion into the quantized signal which cannot be eliminated. Increasing the number of discrete outputs of a quantizer, typically reduces the distortion, but cannot eliminate it. The fundamental trade-off in this choice is the resulting signal quality versus the amount of data needed to represent each input.

The set of inputs of a quantizer can be scalars or vectors. If they are scalars, we call the quantizers scalar quantizers. If they are vectors, we call them vector quantizers. Both scalar and vector quantizers play an important role in data compression and have therefore been studied extensively. While scalar quantization is used primarily for analog-to-digital conversion, vector quantization is used with sophisticated digital signal processing, where in most cases the input signal already has some form of digital representation and the desired output is a compressed version of the digital signal. Vector quantization is usually, but not exclusively, used for the purpose of data compression.

A vector can be used to describe almost any type of pattern, such as a segment of a speech waveform or an image, simply by forming a vector of samples from the waveform or image. Vector quantization can be viewed as a form of pattern recognition where an input pattern is approximated by one of a predetermined set of standard patterns, or in other language, the input pattern is matched with one of a stored set of templates or codewords. Vector quantization is far more than a formal generalization of scalar quantization. In the last few years it has become an important technique in speech recognition as well as in speech and image compression, and its importance and application are growing.

A vector quantizer Q of dimension n and size M is a mapping from a vector (or a "point") in n-dimensional Euclidean space, $\mathbb{R}^n$, into a finite set $\mathcal{C}$ containing M output or reproduction points, called code vectors or codewords. Thus,

$$Q : \mathbb{R}^n \mapsto \mathcal{C} \tag{1.4}$$

where $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M\}$ and $\mathbf{c}_j \in \mathbb{R}^n$ for each $j \in \mathcal{J} \equiv \{1, 2, \ldots, M\}$. The set $\mathcal{C}$ is called codebook or the code and has size M, meaning it has M distinct elements, each a vector in $\mathbb{R}^n$. The resolution, code rate, or, simply, rate of a vector quantizer is $R = (\log_2 M)/n$, which measures the number of bits per vector component used to represent the input vector and gives an indication of the accuracy or precision that is achievable with a vector quantizer if the codebook is well-designed.

Associated with every M point vector quantizer is a partition of $\mathbb{R}^n$ into M regions or cells, $\mathcal{R}_j$ for $j \in \mathcal{J}$. The jth cell is defined by

$$\mathcal{R}_j = \{\mathbf{x} \in \mathbb{R}^n : Q(\mathbf{x}) = \mathbf{c}_j\} , \tag{1.5}$$

From the definition of the cells, it follows that

$$\bigcup_j \mathcal{R}_j = \mathbb{R}^n \quad \text{and} \quad \mathcal{R}_j \bigcap \mathcal{R}_{j'} = \emptyset \quad \text{for} \quad j \neq j' \tag{1.6}$$

In fact, every quantizer can be viewed as a combined effect of two successive operations, an encoder f, and a decoder, g. The encoder is a mapping $f : \mathbb{R}^n \mapsto \mathcal{J}$ where $\mathcal{J} = \{1, 2, \ldots, M\}$, and the decoder is the mapping $g : \mathcal{J} \mapsto \mathcal{C}$. Thus if $Q(\mathbf{x}) = \mathbf{c}_j$ then $f(\mathbf{x}) = j$ and $g(j) = \mathbf{c}_j$. With these definitions we have $Q(\mathbf{x}) = g(f(\mathbf{x}))$.

In the context of a waveform communication system, the encoder transmits the index j of the selected level, $\mathbf{c}_j$, chosen to represent an input sample, and not the value $\mathbf{c}_j$ itself. Thus, if the rate R is an integer, one could assign to each index j a unique binary R-tuple. This binary R-tuple can be transmitted or stored and then the decoder reconstructs the corresponding reproduction value. Note that the decoding can be implemented by a table look-up procedure, where the table or codebook contains the output set which can be stored with extremely high precision without affecting the transmission rate R.

When $n = 1$, vector quantizers become scalar quantizers. In that case, each quantization region is an interval; and each region $\mathcal{R}_j$ is then represented by a representation point $a_j \in \mathbb{R}$. Thus an M-level quantizer is specified by $M + 1$ interval endpoints, $b_0 = -\infty, b_1, \ldots, b_{M-1}, b_M = +\infty$ and M representation points, $a_1, \ldots, a_M$. A quantization region can be thus written as $\mathcal{R}_j =]b_{j-1}, b_j]$ as shown in Figure 1.3.

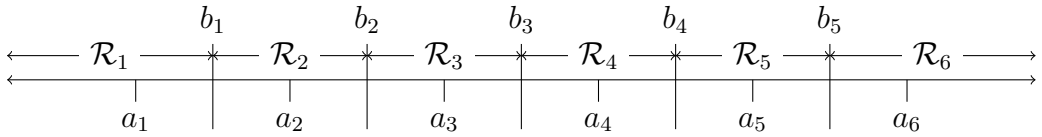


Figure 1.3: Scalar quantization regions and representation points.

Scalar quantizers have many advantages including the facts that they are the simplest and the cheapest. Furthermore, they are usually memoryless and thus avoid transmission delays caused by complex signal processing at the node. It is known that vector quantizers may require lower rates than their scalar counterparts even for independent source outputs, yet they are generally more complex and consume more processing time.

For more on quantization and its information-theoretic foundation, rate-distortion theory, see the original papers by Kolmogorov [17] and Shannon [18], book by Gersho and Gray [19], a collection of papers edited by Abut [20], surveys by Gray [21] and Kieffer [22], and a comprehensive review by Gray and Neuhoff [23].

The purpose of quantization is to provide a limited precision description of a previously unknown input vector. It is only because the input is not known in advance that it is necessary to quantize. Thus the input must be modeled as a random variable, having some specific statistical character, usually specified by its probability density function. Consequently, the error introduced in quantizing this input will also be random. To conveniently assess the performance of a particular quantizer, we need a single number that indicates the overall quality degradation or distortion incurred over the lifetime of its use with a particular statistically specified input. In addition, some kind of overall measure of performance, usually based on statistical averaging, is required that takes into account the input pdf as well as the

specific quantizer characteristic. The most common measure of the distortion between two vectors is the squared error defined by

$$d(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \tag{1.7}$$

For an overall measure of performance that considers the lifetime performance of the quantizer we can use either a worst-case value or a statistical average of some suitable measure of the distortion. For a bounded input, the worst-case absolute error, often more simply called the maximum error, is the maximum of $d(\mathbf{x}, \hat{\mathbf{x}})$ taken over all possible inputs. For an unbounded input with a finite quantizer the worst-case error is infinite and hence is not a meaningful performance measure. The statistical average of the distortion is usually a more informative and meaningful performance measure in general it can be written as

$$D = \frac{1}{n}\,\mathbb{E}[d(X^n, Q(X^n))] = \frac{1}{n}\int_{-\infty}^{+\infty} d(\mathbf{x}, Q(\mathbf{x}))p_{X^n}(\mathbf{x})d\mathbf{x} \tag{1.8}$$

Where $p_{X^n}(\mathbf{x})$ is the pdf of X^n. The average distortion D is called the mean squared error (MSE).

The mean square error (MSE) is the most common distortion measure used in information theory which forms the basis for the larger part of lossy compression algorithms in use nowadays [23]. The MSE is a norm on the distortion sustained in estimating the source from its encoded values. In algorithmics and in signal processing, Lloyd-Max is an algorithm that allows the construction of the optimal quantizer using the MSE distortion [24, 25]. The Lloyd algorithm for quantizer design works by iterating between two partial solutions

- Given a set of representation codewords $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M\}$, how should the quantization regions $\{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M\}$ be chosen?
- Given a set of quantization regions $\{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M\}$, how should the representation codewords $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M\}$ be chosen?

1.3.1 Distributed Source Coding

The problem of source coding becomes significantly more interesting and challenging in a network context. Several new scenarios arise:

- Different information of the source information may be available to seperate encoding terminals that cannot cooperate.
- Decoders may have access to additional side information about the source information.

Distributed source coding (DSC) [26], which refers to the separate encoding and joint decoding of multiple correlated data sources, has received considerable attention from the signal processing [27], communications [28, 29, 30] and information theory [31, 32] communities.

In [31], Slepian and Wolf (SW) proved that, by exploiting the correlation between the sources only at the decoder (but not at the encoder), a total rate equal to the joint entropy suffices to achieve lossless compression. Supposing X and Y are two statistically dependent discrete random processes, taking values in finite alphabets, which are encoded by two separate encoders, but are decoded by a joint decoder. Then the SW result says that even if the encoders are independent, to achieve reliable transmission of both X and Y , the rates of transmission must be such that $R_X \geq H(X|Y)$, $R_Y \geq H(Y|X)$, and $R_X + R_Y \geq H(X,Y)$.

Instead of lossless compression in Slepian–Wolf theorem, Wyner and Ziv [32, 33] established the information theoretic bounds for lossy compression when the decoder has access to an additional information source, correlated with the input, termed the side information. In [32], they derived the rate-distortion function for source coding with side information, i.e., Wyner-Ziv (WZ) coding; the solution employs an auxiliary random variable, which forms a Markov chain with the source and the side information. This constraint typically imposes a rate loss in coding rate, when compared to the predictive coding case, in which the encoder also has access to the side information [34]. The Wyner-Ziv rate is given by

$$R_{X|Y}^{WZ}(D) = \min_{\substack{p(u|x): \\ U-X-Y \\ E[d(U,X)]\leq D}} I(X;U|Y) \tag{1.9}$$

Source coding problems with decoder side information are a special case of distributed source coding problems. The role and potential benefit of Side Information (S.I.) in lossless and lossy data compression is a central theme in information theory. In ways that are well understood for various source coding systems, S.I. can be a valuable resource, resulting in significant performance boosts relative to the case where it is absent.

1.3.2 Detection-Oriented Source Coding

The design of the source encoder/decoder depends on the objective of the communication system. Lloyd, for instance, aimed at designing systems for the particular application of recovering the source. In many other applications, the goal is to detect a source rather than estimate it. For example a radar operator must decide if what he sees on the radar screen indicates the presence of a plane (the signal) or the presence of parasites (the noise). This type of applications was the original framework of *Signal Detection Theory* (see the founding work of Green & Swets, 1966) [35].

Detection, decision making, and hypothesis testing (HT) may sometimes refer to the same thing. The meaning has been extended in the communication field to detect which one, among a set of mutually exclusive alternatives, is correct. These mutually exclusive alternatives are usually referred to as hypotheses. The primary problem that we consider in this thesis is the binary hypothesis testing problem in which we assume that there are two possible hypotheses, the null hypothesis H_0 and the alternative hypothesis H_1.

A HT can be erroneous in two different ways. Either the detector decides H_1 when H_0 is the true

hypothesis or it decides H_0 when H_1 is the correct hypothesis. The probabilities of these two errors, respectively denoted by α and β, together give the performance of a detector (or a test) [36].

The conventional HT problem; when no communication is required; is to decide between two alternative distributions from the observed data which is available at the statistician, i.e., $H_0 : p_X$ versus $H_1 : p_{\bar{X}}$. The optimal error exponent was characterized by Stein [37, 38]:

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n = -D(p_X||p_{\bar{X}})$$

with β_n being the Type II probability of error, subject to the Type I probability of error $\alpha_n \leq \epsilon$.

Distributed hypothesis testing (HT) has first been investigated by Berger [39] and Ahlswede & Csiszar in [40]. This was a first step to combine two seemingly different problems which have been studied separately in the fields of standard statistics inference and information theory. This scenario consists of a decoder (detector) that is required to perform a binary decision based remotely collected data sent from the encoder.

In such problems, the goal is hence to define an appropriate distortion measure allowing to design a quantizer adapted to hypothesis testing. The MSE criterion is appropriate when the aim is to reconstruct the source. However, it can be unreasonable when other applications are concerned. A quantizer designed to minimize mean squared error may recklessly lose information necessary for good detection performance. In detection problems, the main degradation measure should be the two types of error. However, minimizing the error probabilities is difficult and inconvenient to carry out on many occasions. Consequently, several suboptimum performance measures such as the measures of dissimilarity that are easier to manipulate are studied.

1.3.3 Secrecy-Oriented Source Coding

Nowadays, a huge amounts of information flow in the network. With this huge amount of information, the main task for network designers is to ensure that the data can be transmitted reliably and also securely across the network. The latter requirement is becoming increasingly acute, especially when sensitive information is involved. Let us imagine a network in which information flows from one node to another through a number of intermediate nodes. The system design generally makes use of these intermediate nodes to help the transmission. However, these nodes might be public devices or terminals which we cannot fully trust with access to significant amounts of our information. This scenario leads to a natural trade-off between cooperation and secrecy in the system and motivates the study of secure communication and compression.

At the beginning of the information theory era, the majority of studies focused only on problems of reliable communication. Recently, extensive research is concentrating on secure communication, i.e., when the goal is to design a communication system that is both reliable and secure. Conventional techniques for achieving confidentiality and communication networks are based on cryptographic encryption where security was only taken into consideration in the application layer of the OSI model. In encryption, the transmitter uses a key to encrypt source information, i.e., plaintext, converted into

ciphertext. The intended receiver extract original plaintext from a ciphertext by a corresponding key. If an eavesdropping has access to the ciphertext, but it does not know the corresponding decryption key, then it cannot obtain the source information. The chief failings of this notion of security are the assumptions placed on the attacker. As a practical matter, the attacker is usually assumed to have a limited time or limited computed computational resources to that it cannot test all possible keys to extract the source information.

Shannon [41] introduced the information theoretic notion of security considered at the physical layer, where secrecy at the eavesdropper is measured through the conditional uncertainty of the source given the message taking into account the different characteristics possessed by the eavesdropper and the legitimate receiver. A system is information-theoretically secure if its security derives purely from information theory [42]. That is, it cannot be broken even when the adversary has unlimited computing power. The adversary simply does not have enough information to break the encryption. Hence, information-theoretic security makes no computational assumptions on the attacker, and is accepted as the strictest form of security [43].

Claude Shannon defined the notion of perfect secrecy [41]. If a secret message J is encrypted to form a cryptogram C using a secret key K, then perfect secrecy is achieved if

$$H(J|C) = H(J) \tag{1.10}$$

That is, if the cipher-text provides no information about the message. Note that the entropies in 1.10 are calculated assuming K is chosen according to some random key distribution. We note here, that the quantity $H(J|C)$ is typically called the message equivocation, and $H(K|C)$ is called the key equivocation.

For the physical-layer security design, we not only wish to obtain arbitrarily low probability of decoding error for Bob, as in traditional channel coding, but also wish to provide some level of security against Eve as seen in Figure 1.4.

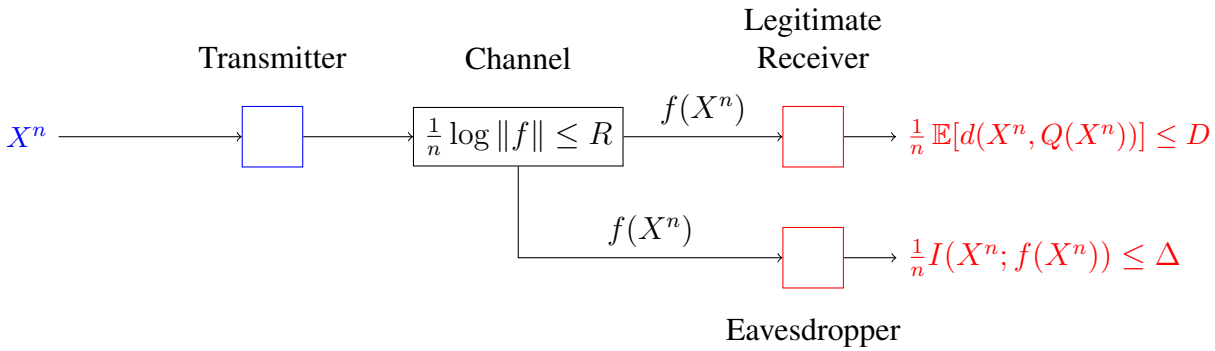


Figure 1.4: Secure communication system for lossy estimation of the source.

Information theoretic study of the physical layer security was pioneered by Wyner [44] for the case of the Wiretap channel. Wyner has shown that perfect security is attainable as long as the channel of the eavesdropper is a degraded version of the legitimate user's one. In his setting, a legitimate receiver

obtains the message over the main channel, while the eavesdropper obtains a degraded version of the message, through an additional channel called the wire-tap channel. For the general degraded wiretap channel, the secrecy capacity [45] can be bounded as $C_s \geq C_m - C_w$, where C_m is the channel capacity of the main channel, and C_w is the channel capacity of the wiretap channel. Note, for some channels, $C_s = C_m - C_w$, but in the general degraded wiretap channel, the secrecy capacity is at least equal to this difference.

Csiszar & Korner [46] studied the case of broadcast channels instead. Several extensions to other multi-user channels [42] including secure coding for multiple access channels, etc. were then considered. In all these settings, the eavesdropper is assumed to be obtaining different information than that of the legitimate receiver via a wiretap or a broadcast channel. On the contrary, in this thesis, we assume that the eavesdropper can obtain the same information obtained by the legitimate receiver.

While practical codes exist that obtain varying levels of information-theoretic security, most designs suffer from one or more of several drawbacks. If an eavesdropper has a better channel than a legitimate receiver, say the channel capacity of the wiretap channel C_w exceeds that of the main channel C_m, then the scheme will likely fail. For some channels, $C_w > C_m$ implies that the secrecy capacity of the system is zero. For example, consider what occurs when an eavesdropper has a noise-free channel. These types of scenarios necessitate the coupling of physical-layer security schemes with additional protection. Of course, cryptography can fill that role nicely.

When a system exposes its secret information to an unauthorised entity, the information exposed is called *Information leakage*. The information leakage can be measured using different ways depending on the designer's point of view. One of the most common metrics for measuring information leakage is the mutual information of (i.e. the amount of information shared between) the system's secret data and the data accessible by an eavesdropper.

In this work, we consider a secure lossy source coding problem under an information leakage rate constraint. The secure source coding problem is essentially a source coding problem with an additional secrecy constraint. For a given source sequence X^n, and the output of the encoder/quantizer $f(X^n)$, that is available to the eavesdropper, the information leakage rate is defined as a normalized mutual information $\frac{1}{n} I(X^n; f(X^n))$. The solution to a secure lossy source coding problem is the optimal tradeoff between transmission rate, incurred distortion at the decoder, and information leakage rate at the eavesdropper in the form of a rate-distortion-leakage region.

This thesis (in Chapters 2 and 3) provides some new fundamental results on both detection and secrecy oriented source coding in the presence of side informations at the receiving terminals.

More precisely, we investigate the problem of secure multiterminal HT with side information at both the detector and the eavesdropper. This scenario consists of three nodes:

- A main encoder (referred to as Alice), that observes a local source,
- A legitimate receiver (referred to as Bob), that wishes to estimate the joint distribution of Alice's source and the directly available side information from a compressed version received through a (public) rate-limited channel,

- An eavesdropper (referred to as Eve), that perfectly observes the information bits sent by Alice to Bob, and has also access to a correlated source which can be used as side information.

In Chapter 2, we study the fundamental limits of the problem, i.e. the trade-off between the maximum achievable error exponent at the detector, (i.e., the minimum type II error probability for a fixed type I probability of error), the coding rate at the encoder and the leakage rate at the eavesdropper. Whereas in Chapter 3, we propose practical algorithms to obtain solutions that are as close as possible to the optimal, which requires the design of optimal quantization (both scalar and vector) in the presence of an eavesdropper.

1.4 Secure Distributed Source Coding with Applications to Smart Meters

1.4.1 Introduction

Governments and power companies across the world have recognized that the traditional grid, which has not significantly changed in 100 years, was not designed to meet the increased demands of a restructured electricity marketplace, the energy needs of a digital society, or the increased use and variability of renewable power production and must be replaced and upgraded by more efficient, flexible and intelligent energy-distribution networks.

This radical change is also stimulated by the pressing need to de-carbonise electricity supply, to replace ageing assets and to make effective use of rapidly developing information and communication technologies (ICTs). These aims all converge in the Smart Grid. The Smart Grid uses advanced information and communication to control this new energy system reliably and efficiently.

In its current state, the grid consists of four major components: generation produces electric energy in different manners, e.g., by burning fossil fuels, inducing nuclear reaction, harnessing water (hydroelectric dams), wind, solar, and tidal forces; transmission moves electricity via a very high voltage infrastructure; distribution steps down current and spreads out for consumption; and consumption, i.e., industrial, commercial, and residential, uses the electric energy in a multitude of ways.

The part of the power system supplying energy has good communication links to ensure its effective operation, to enable market transactions, to maintain the security of the system, and to facilitate the integrated operation of the generators and the transmission circuits. This part of the power system has some automatic control systems though these may be limited to local, discrete functions to ensure predictable behaviour by the generators and the transmission network during major disturbances.

The distribution system, feeding load, is very extensive but is almost entirely passive with little communication and only limited local controls. Other than for the very largest loads (for example, in a steelworks or in aluminium smelters), there is no real-time monitoring of either the voltage being offered to a load or the current being drawn by it. There is very little interaction between the loads and the power system other than the supply of load energy whenever it is demanded. The present revolution

in communication systems, particularly stimulated by the internet, offers the possibility of much greater monitoring and control throughout the power system and hence more effective, flexible and lower cost operation. The Smart Grid is an opportunity to use new ICTs (Information and Communication Technologies) to revolutionise the electrical power system. However, due to the huge size of the power system and the scale of investment that has been made in it over the years, any significant change will be expensive and requires careful justification.

The consensus among climate scientists is clear that man-made greenhouse gases are leading to dangerous climate change. Hence ways of using energy more effectively and generating electricity without the production of CO2 must be found. The effective management of loads and reduction of losses and wasted energy needs accurate information while the use of large amounts of renewable generation requires the integration of the load in the operation of the power system in order to help balance supply and demand. Smart meters are an important element of the Smart Grid as they can provide information about the loads and hence the power flows throughout the network. Once all the parts of the power system are monitored, its state becomes observable and many possibilities for control emerge.

The Smart Grid vision is to give much greater visibility to lower voltage networks and to enable the participation of customers in the operation of the power system, particularly through Smart Meters and Smart Homes. The Smart Grid will support improved energy efficiency and allow a much greater utilisation of renewables.

1.4.2 Smart Grid Communication Infrastructure

The operation of smart grid will feature bi-directional digital communication, bi-directional power flow, and consumer empowerment with enhanced situation awareness. As such, adaptive signal processing, distributed detection and estimation, statistical signal processing, signal representation and data compression, machine learning, optimization methods, efficient computational algorithms, etc., will all prove to be important tools to make possible some of the important features envisioned for the smart grid – demand response, distribution automation, self-healing, improved security, etc.

The communication infrastructure of a power system typically consists of SCADA systems with dedicated communication channels to and from the System Control Centre and a Wide Area Network (WAN). The SCADA systems connect all the major power system operational facilities, that is, the central generating stations, the transmission grid substations and the primary distribution substations to the System Control Centre.

An essential development of the Smart Grid is to extend communication throughout the distribution system and to establish two-way communications with customers through Neighbourhood Area Networks (NANs) covering the areas served by distribution substations. Customers' premises will have Home Area Networks (HANs). The interface of the Home and Neighbourhood Area Networks will be through a smart meter or smart interfacing device.

Smart meters may be used in various ways and thus lead to different requirements for the metering

communication system. Automated meter reading (AMR) requires only occasional transmission of recorded energy data (perhaps once a month) while advanced metering infrastructure (AMI) requires frequent bi-directional communication (perhaps every 30 minutes). The use of smart meters to support Smart Grid operation of the distribution network has not yet been implemented widely but is likely to place severe demands on the communication system.

1.4.3 Smart Metering

The connection of a large amount of intermittent renewable generation alters the pattern of the output of central generation and the power flows in both transmission and distribution circuits. One solution to this increase in variability is to add large-scale energy storage devices to the power system. This is often not practical at present due to technical limitations and cost. Therefore, flexibility in the demand side is seen as another way to enable the integration of a large amount of renewable energy.

Effective implementation of the Demand-Side Integration needs an advanced ICT (Information and Communication Technology) infrastructure and good knowledge of system loads. However, the electro-mechanical meters that are presently installed in domestic premises have little or no communication ability and do not transmit information of the load in real time. Smart metering refers to systems that measure, collect, analyse, and manage energy use using advanced ICT. The concept includes two-way communication networks between smart meters and various actors in the energy supply system. The smart meter's objective is to provide real-time or near-real-time information exchange and advanced control capabilities.

Electronic smart meters not only can measure instantaneous power and the amount of energy consumed over time but also other parameters such as power factor, reactive power, voltage and frequency, with high accuracy. Data can be measured and stored at specific intervals. Moreover, electronic meters are not sensitive to external magnets or orientation of the meter itself, so they are more tamperproof and more reliable.

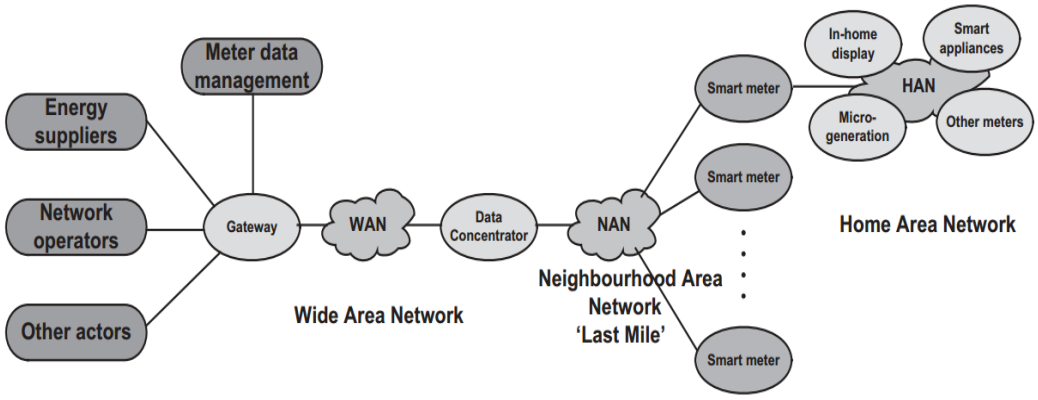


Figure 1.5: Smart metering communication.

A typical communications architecture for smart metering is shown in Figure 1.5. It has three com-

munications interfaces: Wide Area Network (WAN), Neighbourhood Area Network (NAN) and Home Area Network (HAN) [47].

- Home-area network: A Home-Area Network (HAN) is an integrated system of smart meter, in-home display, microgeneration, smart appliances, smart sockets, HVAC (Heating, Ventilation, Air Conditioning) facilities and plug-in hybrid/electric vehicles. A HAN uses wired or wireless communications and networking protocols to ensure the interoperability of networked appliances and the interface to a smart meter. It also includes security mechanisms to protect consumer data and the metering system.

- Neighbourhood area network: The primary function of the Neighbourhood Area Network (NAN) is to transfer consumption readings from smart meters. The NAN should also facilitate diagnostic messages, firmware upgrades and real-time or near real-time messages for the power system support.

- Data concentrator: The data concentrator acts as a relay between the smart meters and the gateway. It manages the meters by automatically detecting them, creates and optimises repeating chains (if required to establish reliable communication), coordinates the bi-directional delivery of data, and monitors the conditions of the meters.

- Meter data management system: The core of a meter data management system is a database. It typically provides services such as data acquisition, validation, adjustment, storage and calculation (for example, data aggregation), in order to provide refined information for customer service and system operation purposes such as billing, demand forecasting and demand response. A major issue in the design and implementation of a meter data management system is how to make it open and flexible enough to integrate to existing business/enterprise applications and deliver better services and more value to customers while ensuring data security. Besides the common database functionalities, a meter data management system for smart metering also provides functions such as remote meter connection/disconnection, power status verification, supply restoration verification and on-demand reading of remote smart meters.

1.4.4 Automated control and incident detection (system surveillance)

This thesis focuses mainly on distributed binary detection when side information is directly available at the decoder with possible applications in smart grids. Applications may arise in the context of testing against independence; i.e. when the two remote sources are independent under the alternative hypothesis. In that case distributed detection can be used to notify a fault at the level of the smart meter, an incident at the level of the TSO, or a voltage/frequency fluctuation etc.

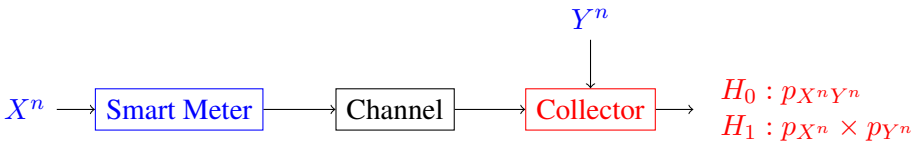


Figure 1.6: Distributed HT against independence used for fault detection

Fault Detection

The integrity of smart meter data is a paramount concern for customers (as well as utilities). A lack of confidence may lead to a public and political backlash, like the one against Pacific Gas and Electricity (PG&E) in Bakersfield, California in 2009. Thousands of customers complained that their smart meters were overcharging them. PG&E attributed the higher bills to a rate increase and spikes in usage due to summer weather [48]. However, in at least some cases, it appears that billing discrepancies were due to improper installation or malfunctioning equipment [49]. The consequences for PG&E are ongoing lawsuits, political pressure for a moratorium on deployment and increased scrutiny of their smart meters [50], [51]. Some PG&E customers now verify the integrity of their billing independently using consumer-device wireless power meters [52].

Smart meters have created a growing array of serious problems and as any other devices, meters could face failure problems causing them to communicate false and unreliable data to the provider/collector. To elaborate, the source in this case could be the readings of a smart meter of a certain consumption unit while the decoder could be the collector or the provider, having access to side information representing some useful correlated data (history, etc.). Distributed detection can be then adopted to test the joint distribution of these two data. Any absence of correlation could indicate a smart meter deficiency and hence a fault detection at the level of the smart meter could be investigated.

In the control of electric power systems, especially in the wide area backup protection of electric power systems, the prerequisite of protection device's accurate, fast and reliable performance is its corresponding fault type and fault location can be discriminated quickly and defined exactly. In [53], the author uses discriminant analysis theory to detect faults in smart meters.

Incident Detection

Experts agreed that the Transmission System Operators TSOs and the Distribution System Operators DSOs need to perform monitoring actions to detect possible incidents affecting the power grid as a whole and also in each member state MS. In European-wide incidents, many experts consider that TSOs should be the organisations in charge of monitoring and triggering alarms. Incidents may take place at the consumer level (smart metering at the customer) [54], or at the DSO/TSO level. Distributed hypothesis testing against independence can also be used to detect such incidents.

Detection of Non-technical Losses

Technical losses are naturally occurring losses (caused by actions internal to the power system) and consist mainly of power dissipation in electrical system components such as transmission lines, power transformers, measurement systems, etc. Technical losses are possible to compute and control, provided the power system in question consists of known quantities of loads.

Non-technical losses (NTL), on the other hand, are caused by actions external to the power system, or are caused by loads and conditions that the technical losses computation failed to take into account. NTL are more difficult to measure because these losses are often unaccounted for by the system operators and thus have no recorded information [55]. The most probable causes of NTL are:

- Electricity theft
- Non-payment by customers
- Errors in technical losses computation
- Errors in accounting and record keeping that distort technical information
- Component breakdowns that drastically increase losses

In the same way mentioned in the previous sections, monitoring the distribution of the data collected could help us detect such losses. A more precise Hypothesis testing could later be performed in order to define the type of this technical losses (Component breakdowns for example will have much more impact on the distribution than electricity theft and so on).

Reducing Voltage and Frequency Fluctuations

Voltage fluctuations can be described as repetitive or random variations of the voltage envelope due to sudden changes in the real and reactive power drawn by a load. To allow equipment connected to the power system to operate correctly it is important for both the utility and their customers to ensure that the operating voltage of the system remains within the boundaries set by the appropriate standards. Reducing the effects of voltage fluctuations could be done using different strategies.

Frequency fluctuations should also be taken into consideration. Any inequality between production and consumption results in an instantaneous change in frequency from nominal, frequency should be always monitored and controlled. Traditionally, frequency regulation is provided by varying the power output of generators which have restricted ramp rates. New energy storage technologies can also be useful by rapidly changing their outputs and providing frequency regulation with very fast response to frequency fluctuations.

In HT, the remote source could be the measured voltage or frequency at different time-slots, and the side information could be the desired values or some historical data. HT could be used to measure the correlation between these random variables, lower correlation means higher fluctuations and action should be taken while higher correlation could mean more stability of the electic network.

1.4.5 Information Security for the Smart Grid

With millions of customers becoming part of the Smart Grid, the information and communication infrastructure will use different communication technologies and network architectures that may become vulnerable to theft of data or malicious cyber attacks. Ensuring information security in the Smart Grid is a much more complex task than in conventional power systems because the systems are so extensive and integrated with other networks. Potentially sensitive personal data is transmitted and, in order to control costs, public ICT infrastructure such as the Internet will be used. Moreover, the size of the physical system widens the range of possible attacks and constrains the set of feasible countermeasures.

The Smart Grid requires reliable and secure delivery of information in real time. . Delays in the delivery of information accurately and safely are less tolerable in the Smart Grid than for much commercial data transmission as the information is required for real-time or near real-time monitoring and control. A lot of monitoring and control information is periodic and contributes to a regular traffic pattern in the communication networks of the Smart Grid, though during power system faults and contingencies there will be a very large number of messages. Any form of interruption resulting from security issues is likely to have serious effects on the reliable and safe operation of the Smart Grid. Therefore, protecting the whole communication system which is the core of the smart grid is one of the primary goals.

Smart Grid Privacy Concerns

The Deployment of smart grids have raised many privacy concerns. Smart meters installed inside the houses communicate detailed consumption data to different central units and utilities. These detailed data may leak personal information. These information facilitate the creation of user lifestyle profiles, including house occupancy, meal times, working hours, vacation days ... and hence threatens user privacy.

Malicious attacks against the power grid can directly affect everybody's lifestyle. Public bodies and C-level staff of the utilities operating the distribution and transmission networks, as well as electricity marketers and generation organisations should be aware of this situation. Mechanisms to improve the security posture of the current networks are being put in place where cyber security is included as a primary objective of the smart grids.

Researchers have designed protocols that allow for price billing of consumption while not revealing any consumption information to the providers [56]. Meters send readings at different time slots to customer devices that compute a billing message alongside a mathematical proof that a fee is correct. The provider receives the payment message with the proof to check whether the bill is correct without knowing any information about individual meter readings. This protocol uses encryption techniques to send these data, and the user devices perform the billing computation after decrypting the reading using the encryption key. In our thesis, our focus will be on information theoretic security instead.

Providing information security has been a common need of ICT systems since the Internet became the main mode of communication. Thus there are well-established mechanisms to provide information

security against possible threats. Existing security approaches based on cryptography may happen to be difficult to apply in this context, because of the very large dimensions involved. In fact, cryptography heavily relies on secret keys and the dissemination of these keys is often the weak point of the system. If the key remains the same for a very long time, this leaves room for possible adversaries to crack the system, and changing frequently the keys may be either resource consuming or difficult to implement on a very large and distributed system. It is therefore the right time to check the applicability of recently proposed techniques not relying on cryptography. Information-theoretic results prove the existence of physical-layer coding that guarantees a level of secrecy against eavesdroppers, by harnessing the statistical asymmetries inherently in data (sources, side informations) and communication media (interference, noise), without consuming additional resources or requiring pre-shared secret keys. The application of information-theoretic style approaches to secure Smart Grids is extremely attractive, not only because statistical asymmetries between nodes are abundantly available in such scenarios but specially because information theory has the mathematical tools to characterize the fundamental limits, optimal trade-off between reliable estimation/detection and what can be guaranteed in terms of secure information, no matter the way in which the malicious eavesdroppers process the information.

Unlike cyber security, physical layer security has been hardly addressed in the literature of the smart grid to date. In [57], the author proposed a way to encode the redundant measurement at a bit rate below its entropy, so that it cannot be decoded from the encoded bits alone. In this way, he guarantees information-theoretic confidentiality, regardless of the computational power of an eavesdropper. Redundant metering is frequently used to verify the integrity of billing data reported by advanced metering infrastructure, but the redundant measurement introduces a potential confidentiality leak.

The key idea is to compress the redundant measurement to a rate below its entropy, so that it cannot be recovered from just the encoded bits. But the redundant measurement can be recovered in conjunction with the reported measurement, as long as the compression rate is greater than the conditional entropy of the redundant measurement given the reported measurement. Unlike encryption, this method guarantees confidentiality regardless of the computational capability of the eavesdropper

Encryption appears to be the typical approach [58, 59]; however, such approach does not have a robust theoretical basis for both privacy and detection/estimation performance. Such a basis is important for several reasons. First, a theoretical abstraction allows us to recast the problem in a technology-independent manner – we need a privacy framework that not only addresses the capabilities of current techniques but is also extensible to future ones. Second, a theoretical framework enables us to examine the costs of lost privacy against the benefits of data dissemination, namely, the trade-off between privacy and decoder performance. It would be desirable to have the ability to decide that trade-off. Finally, a theoretical framework for privacy and performance may expose points of trade-off that are unexpected.

PART I

Secure Distributed Binary Detection

CHAPTER 2

HYPOTHESIS TESTING WITH COMMUNICATION AND SECURITY CONSTRAINTS

Abstract — *A receiver "Bob" is interested in detecting the joint probability distribution of 2 remotely located data. "Bob" has access to the encoded version of some observations of a random variable X; collected and encoded with a rate R by an encoder "Alice" and sent over a public noise-less channel; and some directly available observations of another random variable Y. At the mean time, an eavesdropper "Eve" has access to these encoded data. It is required to find the maximum error exponent, i.e., the minimum type II probability of error for a type I probability of error being at most ε and a certain equivocation rate, i.e., determine a trade-off between the rate, the error exponent, and the equivocation rate. First, we investigate the case of general hypothesis testing where under both hypotheses the joint distribution could be anything. Then, we investigate the special case of testing against independence where under the alternate hypothesis, the joint distribution refers to the case where X and Y are independent, and we determine the optimal single letter rate-exponent-equivocation and provide examples to Gaussian and Binary sources.*

Keywords — *Hypothesis testing, communication constraints, security constraints, error exponent.*

2.1 Introduction

This chapter studies the problem of hypothesis testing (HT) in which data is compressed and sent to a detector that seeks to decide between two possible distributions. The aim is to characterize all achievable rates of data and equivocation, and the maximal exponent of the Type II error probability when the Type I error probability is at most a fixed value. The conventional HT problem is to decide between two alternative distributions from the observed data which is available at the statistician, i.e., $H_0 : p_X$ versus $H_1 : p_{\bar{X}}$. The optimal error exponent was characterized by Stein [37, 38]:

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n = -D(p_X||p_{\bar{X}})$$

with β_n being the Type II probability of error, subject to the Type I probability of error $\alpha_n \leq \varepsilon$.

The problem of multiterminal hypothesis testing (HT) under communication constraints has first been investigated by Berger [39]. Distributed HT was first introduced in [40] where the authors studied

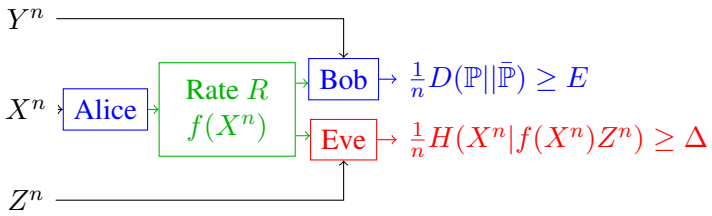


Figure 2.1: HT with communication and privacy constraints.

this problem in the presence of communication constraints. A single-letter characterization is given for testing against independence while partial results are obtained for the general HT problem. Later on [60], a lower bound on the optimal error exponent was proposed based on the exponent of Wyner-Ziv coding [61]. Further results are reported in [62], where based on a sophisticated binning scheme, a novel lower bound is derived. Although for related source coding problems the schemes based on random binning perform well and often optimal, the use of binning in HT is more involved due to the fact that the overall error probability may be dominated by errors in decoding the bin indices. More recently, the authors in [63] studied HT without security constraints and showed that binning is optimal for the type of problems in which the main purpose is to test against conditional independence.

In [64], it was shown that the rate-exponent region of the successive refinement testing against independence and rate regions of the successive refinement lossless one-helper problem are congruent to each other. A strong converse for the successive refinement testing against independence problem is proved which shows the optimal type II error exponents under constant type I error constraints are independent of the exact values of those constants. In [65], they study the vector Gaussian versions of the two problems; hypothesis testing under a communication constraint and the lossy one-helper problem.

In [66], Lower bounds on the error exponent for type II probability of error are presented subject to the exponential-type constraint $\alpha_n \leq \exp(-nr)$ on the error probability of the first kind. On the other hand, [67] studies the problem of interactive testing against independence with communication constraints, where a computable characterization is provided for the optimal trade-off between the communication rates in two-round interaction and the testing performance measured by the error exponent. In [68], they study the case where the two nodes interactively communicate with each other in q rounds to perform a simple binary hypothesis testing on whether the observed random sequences at the nodes are generated independently or not.

The optimal trade-off between distortion and equivocation is well-known in the conventional rate-distortion setup [69], but is not when the target function is the error exponent instead of an average per-letter distortion. More recently in [70], They establish inner and outer bounds on the rate-distortion-equivocation region for the lossy source coding problem with secrecy constraints in which a remote information source should be transmitted to a single destination via multiple agents in the presence of a passive eavesdropper. The eavesdropper, with access to side information correlated to the source, is able to listen in on one of the links from the agents to the destination in order to obtain as much information as possible about the source. On the other hand, the optimal rate-exponent-distortion region for the case of joint estimation and detection against independence was also determined [71]. However, the

problem of HT between two general joint probability distributions is still open and appears to be of a formidable mathematical complexity [72]. Privacy constraints were investigated in the case of parallel distributed detection systems within the Bayesian framework in [73] while differential privacy was addressed in [74].

This chapter addresses the scenario of HT under both communication and privacy constraints. The corresponding model is shown in Fig. 2.1 where Alice maximizes not only the exponent but also the equivocation rate –average uncertainty– at an eavesdropper, denoted as Eve. Eve is assumed to perfectly observe the information bits sent by Alice and has access to a correlated i.i.d. string $Z^n = (Z_1, \ldots, Z_n)$ (possibly the same as Bob). We study the minimum amount of data that Alice needs to communicate to Bob to guarantee the desired exponent while satisfying the average uncertainty requirement at Eve. The optimal tradeoff between distortion and equivocation is well-known in the conventional rate-distortion setup [78], but is not when the target function is the error exponent instead of an average per-letter distortion. The standard source coding problem and the HT problem without privacy constraints appear to be fundamental different from each other.

The chapter is divided into two parts. In the first part, we study the general HT problem and derive an achievable region. The region we provide uses a double layer scheme based on binning and generalizes that provided in [60] by taking into consideration both testing and decoding errors which takes place during the binning process. The use of the double layer based on binning made the problem very complicated but provided a better overall error exponent and was needed to improve the security in the region. In [62], a similar while not equivalent lower bound on the error exponent was proposed but unfortunately the proof is not available. In the second part, we study the special case of testing against independence where we assume that under both hypothesis the probability distributions have equal marginals. In this case, only one layer was used without binning and an optimal single-letter characterization of the rate-exponent-equivocation region is provided. Applications of our results arise in the context in which data must remain private even from the statistician (see [75] and references therein), the region is also evaluated for Gaussian and binary sources.

The rest of this chapter is organized as follows. Section II provides notations and problem definitions. Section III gives the main results of the general hypothesis testing, Section IV demonstrates the result for testing against independence and give an example in which sources are binary and Gaussian. Concluding remarks are given in Section V.

2.2 Problem Definition

2.2.1 Notations

Upper case letters X denote random variables (RVs) with values x in the finite set $\mathcal{X}$. The cardinality of $\mathcal{X}$ is denoted by $\|\mathcal{X}\|$. For RVs X and Y with joint PD p_{XY}, $H(X), H(Y|X)$ denote the entropy and the conditional entropy, respectively, and $h(X)$, $h(Y|X)$ denote the differential entropies, while $I(X;Y)$ denotes the mutual information. The notation $I(p_X; p_{Y|X})$ is also used to denote the mutual

information between X and Y while assuming that the distribution $p_X p_{Y|X}$ governs the pattern. For $p \in [0,1]$, $H_2(p)$ denotes the binary entropy function. For two PDs p, q on $\mathcal{X}$, $D(p\|q)$ is ***Kullback-Leibler divergence***. x_k^n stands for the collection $(x_k, x_{k+1}, \ldots, x_n)$ where we use x^n to denote vectors in $\mathcal{X}^n$ of length n and x_1^n is simply denoted x^n. Let (X, Y, Z) be three RVs such that $P(x|y,z) = P(x|y)$ for each (x, y, z), then $X \multimap Y \multimap Z$ form a Markov Chain.

For a random variable X, we will consider that p_X is the distribution of X while q_X is the empirical distribution or "type" of X as defined in [76]. $T_{q_X}^n$ denotes the set of all sequences x^n having q_X as their type.

For a PD $p_X \in \mathcal{P}(\mathcal{X})$, $T_\varepsilon^\tau(X)$ denotes the set of *strong typical sequences* [77] with constant $\varepsilon > 0$, and for a stochastic mapping $W : \mathcal{X} \mapsto \mathcal{Y}$, $T_\varepsilon^\tau(Y|x^n)$ denotes the set of *conditional strong typical sequences* for a sequence $x^n \in \mathcal{X}^n$ with constant $\varepsilon > 0$. Finally, for $\alpha, \beta \in [0,1]$, $\alpha \star \beta \triangleq \alpha(1-\beta) + (1-\alpha)\beta$.

2.2.2 System Model

The model is shown in Fig. 2.1, where an encoder Alice observes i.i.d realizations of a vector random variable X, and encodes at a rate R. A statistician Bob (the detector), observes the encoded version $f(X^n)$ of $X^n = (X_1, \ldots, X_n)$ and an i.i.d realizations of a vector Y^n directly available, which is arbitrary dependent on X^n by memoryless. An eavesdropper (referred to as Eve) has access to $J = f(X^n)$ and another side information Z^n, also consisting of i.i.d. samples arbitrary dependent on (X^n, Y^n). Alice wishes to communicate the source by using an encoding mapping:

$$f : \mathcal{X}^n \mapsto \{1, \ldots, \|f\|\} \tag{2.1}$$

with coding rate $\log \|f\| \leq nR$. The detector Bob is required to make a decision between two hypotheses:

$$\begin{cases} H_0 : & p_{XY} \ , \\ H_1 : & p_{\bar{X}\bar{Y}} \ . \end{cases} \tag{2.2}$$

It is further assumed that the marginal distributions of X and Y are the same under both hypotheses, i.e., $p_X = p_{\bar{X}}$ and $p_Y = p_{\bar{Y}}$ which does not allow Bob to make the decision without the information sent by Alice. In this setting, Bob has to decide on the basis of the sample Y^n and the message $f(X^n)$ between H_0 and H_1, of which only one is true. For the sake of simplicity, we denote the corresponding probability distributions by $\mathbb{P} = p_{f(X^n)Y^n}$ and $\bar{\mathbb{P}} = p_{f(\bar{X}^n)\bar{Y}}$.

Given $\varepsilon \in (0,1)$, let $\mathcal{A}_n \subset \{1, \ldots, \|f\|\} \times \mathcal{Y}^n$ be the acceptance region for the detector at Bob. The two types of probabilities of error are defined as:

$$\text{Type I} : \quad \alpha_n(f, \mathcal{A}_n) \ = \ \mathbb{P}\left(\mathcal{A}_n^c\right) < \varepsilon \ , \tag{2.3}$$

$$\text{Type II} : \quad \beta_n(f, \mathcal{A}_n) \ = \ \bar{\mathbb{P}}\left(\mathcal{A}_n\right) \ . \tag{2.4}$$

Then, the goal of the detector Bob is to find an encoding function f and an acceptance region $\mathcal{A}_n$ that

minimize the probability (2.4) for a prescribed probability (2.3), and guarantee the equivocation rate (or average uncertainty) at Eve such that:

$$\frac{1}{n}H(X^n|f(X^n)Z^n) \geq \Delta \,, \tag{2.5}$$

or an equivalent definition is the *information leakage*:

$$\frac{1}{n}I(X^n; f(X^n)Z^n) \leq H(X) - \Delta \,. \tag{2.6}$$

For the reasons mentioned above, we change the definition of β_n to include the following:

$$\begin{aligned}\beta_n(f,\varepsilon) \triangleq \min_{\mathcal{A}_n} \Big\{ \bar{\mathbb{P}}(\mathcal{A}_n) | \quad & \mathcal{A}_n \subset f(\mathcal{X}^n) \times \mathcal{Y}^n, \mathbb{P}(\mathcal{A}_n^c) \leq \varepsilon, \\ & \tfrac{1}{n}H(X^n|JZ^n) \geq \Delta \Big\}\end{aligned} \tag{2.7}$$

$$\beta_R(n,\varepsilon) \triangleq \min_f \beta_n(f,\varepsilon) \tag{2.8}$$

According to [40], the asymptotic behavior of the second type probability of error can be described by a parameter $\Theta(R)$ where:

$$\Theta_n(R) \triangleq \sup_f \Big\{ \frac{1}{n} D(\mathbb{P}||\bar{\mathbb{P}}) | \log ||f|| \leq nR \Big\} \tag{2.9}$$

$$\Theta(R) \triangleq \sup_n \Theta_n(R) \tag{2.10}$$

Definition 2.1 (Multi-letter characterization) *A multi-letter characterization of the rate-exponent-equivocation region $\mathcal{R}^*$ is thus given by the set of all tuples $(R, E, \Delta) \in \mathbb{R}_+^3$ such that there exists an encoding function f satisfying:*

$$\frac{1}{n}\log \|f\| \leq R \,, \tag{2.11}$$

$$\frac{1}{n}D(\mathbb{P}||\bar{\mathbb{P}}) \geq E \,, \tag{2.12}$$

$$\frac{1}{n}H(X^n|f(X^n)Z^n) \geq \Delta \,. \tag{2.13}$$

Our objective is to derive an achievable single-letter representation of this region.

2.3 General Hypothesis Testing

In this section we focus on the case where under both hypotheses, the two distributions are general, in other words the two hypotheses are $H_0 : p_{XY}$ and $H_1 : p_{\bar{X}\bar{Y}}$.

2.3.1 Rate-Error-Equivocation Region for the general HT

Our main result for general hypothesis testing is an achievable rate-error-equivocation region and can thus be summarized by the following proposition:

Proposition 2.1 (Achievable rate-error-equivocation region) *A set of all tuples (R, E, Δ) is achievable if the following inequalities are satisfied:*

$$\begin{aligned} R &\geq I(p_X; p_{U|X}) && (2.14)\\ E &\leq \min\{E_1; E_2(R - I(p_X; p_{U|X})); \\ &\qquad E_2(H(p_{X|UZ}|p_{UZ}) - \Delta)\} && (2.15)\\ \Delta &\leq H(p_{X|UZ}|p_{UZ}) && (2.16) \end{aligned}$$

where U and V are auxiliary random variables defined on some finite sets $\mathcal{U}$ and $\mathcal{V}$ such that $U \multimap V \multimap X \multimap (Y, \bar{Y}, Z)$ form a Markov chain. $\bar{U}$ and $\bar{V}$ are also defined such that $(\bar{U}, \bar{V}) \multimap \bar{X} \multimap \bar{Y}$ and $p_{\bar{U}\bar{V}|\bar{X}} = p_{UV|X}$; and E_1 and E_2 are defined by:

$$E_1 = \min_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y} \in \mathscr{L}(UV)} D(p_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}} || p_{\bar{U}\bar{V}\bar{X}\bar{Y}}) \,, \tag{2.17}$$

$$\begin{aligned} \mathscr{L}(UV) = \Big\{ \; & \tilde{U}\tilde{V}\tilde{X}\tilde{Y} : P(\tilde{U}\tilde{V}\tilde{X}) = P(UVX), \\ & P(\tilde{U}\tilde{V}\tilde{Y}) = P(UVY) \Big\} \,, \end{aligned} \tag{2.18}$$

$E_2(R - I(p_X; p_{U|X})) =$

$$\inf_{q_{X|U}} \sup_{q^\star_{V|UX}} \inf_{q_{Y|U}} \inf_{\substack{q_{VXY|U} \\ q_{V|UX} = q^\star_{V|UX}}} G[q_{VXY|U}, R - I(p_X; p_{U|X})] \,, \tag{2.19}$$

$E_2(H(p_{X|UZ}|p_{UZ}) - \Delta) =$

$$\inf_{q_{X|U}} \sup_{q^\star_{V|UX}} \inf_{q_{Y|U}} \inf_{\substack{q_{VXY|U} \\ q_{V|UX} = q^\star_{V|UX}}} G[q_{VXY|U}, H(p_{X|UZ}|p_{UZ}) - \Delta] \,, \tag{2.20}$$

Equations $G[q_{VXY|U}, R - I(p_X; p_{U|X})]$ and $G[q_{VXY|U}, H(p_{X|UZ}|p_{UZ}) - \Delta]$ are defined in (2.21) and (2.22) at the bottom of the next page. This proposition is proved in Appendix A.1.

2.3.2 Degraded Hypothesis with an Arbitrary Large Coding Rate

In this section, we suppose that the alternate hypothesis H_1 is a degraded hypothesis with respect to hypothesis H_0, i.e., $X \multimap Y \multimap \bar{Y}$ form a Markov chain in this order while the rate R is infinity, in other words, the channel is not rate limited. In this particular case, we give an achievable error-equivocation region and show it's optimal. We will consider the case where the conditional probability distribution

of the information available at the eavesdropper Z knowing X is the same as that of side information $\bar{Y}$ at the decoder under the alternate hypothesis H_1. In other words, we will suppose that $p_{Z|X} = p_{\bar{Y}|X}$.

Proposition 2.2 (Degraded Hypothesis) *$\mathcal{R}^*$ is given by the set of all tuples $(E, \Delta) \in \mathbb{R}_+^2$ such that there exists an encoding function f satisfying:*

$$E \leq D(UY||U\bar{Y}) , \tag{2.23}$$

$$\Delta \leq H(X|U\bar{Y}) , \tag{2.24}$$

where the auxiliary random variable U is chosen such that $U \multimap X \multimap Y \multimap \bar{Y}$ form a Markov chain.

The proof of this proposition is given in Appendix A.2. In the proof it can be noticed that binning is not useful in this case to prove the region is optimal and thus is not used.

2.4 Testing Against Independence

In this particular case, the goal is to focus on testing against independence, i.e., between:

$$\begin{cases} H_0 : & p_{XY} , \\ H_1 : & p_X \times p_Y , \end{cases} \tag{2.25}$$

where the two types of error probabilities are in this case defined as the following:

$$\text{Type I} : \quad \alpha_n \triangleq \Pr\left(H_1 | X^n Y^n \sim p_{XY}^n\right) , \tag{2.26}$$

$$\text{Type II} : \quad \beta_n \triangleq \Pr\left(H_0 | X^n Y^n \sim p_X^n \times p_Y^n\right) . \tag{2.27}$$

Bounding the error exponent in this case is much easier due to the fact that:

$$\frac{1}{n} D(p_{f(X^n)Y^n} || p_{f(X^n)} \times p_{Y^n}) = \frac{1}{n} I(J; Y^n) \tag{2.28}$$

$G[q_{VXY|U}, R - I(p_X; p_{U|X})] =$

$$\begin{cases} \min\{D(q_{VXY|U}||p_{XY|U} q_{V|UX}|q_U); D(q_{VXY|U}||p_{\bar{X}\bar{Y}|U} q_{V|UX}|q_U)\} & \\ +[R - I(p_X; p_{U|X}) - I(q_{UX}; q_{V|UX}) + I(q_{UY}; q_{V|UY})]^+ & R < I(p_X; p_{U|X}) + I(q_{UX}; q_{V|UX}) \\ \infty & \text{else.} \end{cases} \tag{2.21}$$

$G[q_{VXY|U}, H(p_{X|UZ}|p_{UZ}) - \Delta] =$

$$\begin{cases} \min\{D(q_{VXY|U}||p_{XY|U} q_{V|UX}|q_U); D(q_{VXY|U}||p_{\bar{X}\bar{Y}|U} q_{V|UX}|q_U)\} & \\ +[H(p_{X|UZ}|p_{UZ}) - I(q_{UX}; q_{V|UX}) + I(q_{UY}; q_{V|UY}) - \Delta]^+ & \Delta > H(p_{X|UZ}|p_{UZ}) - I(q_{UX}; q_{V|UX}) \\ \infty & \text{else.} \end{cases} \tag{2.22}$$

Thus, a multi-letter characterization of the *rate-exponent-equivocation region* $\mathcal{R}^*$ is given by the set of all tuples $(R, E, \Delta) \in \mathbb{R}_+^3$ such that there exists an encoding function f satisfying:

$$\begin{aligned} \frac{1}{n}\log\|f\| &\leq R\,, && (2.29)\\ \frac{1}{n}I(f(X^n);Y^n) &\geq E\,, && (2.30)\\ \frac{1}{n}H(X^n|f(X^n)Z^n) &\geq \Delta\,. && (2.31) \end{aligned}$$

2.4.1 Single-Letter Rate-Error-Equivocation Region

We now state the optimal rate-error-distortion region for testing against independence which provides a single-letter expression.

Proposition 2.3 (Single-letter characterization) *Let $\mathcal{R}^\star$ be the set of all achievable tuples (R, E, Δ), then there exists a random variable U on some finite set $\mathcal{U}$ satisfying:*

$$\begin{aligned} R &\geq I(U;X)\,, && (2.32)\\ E &\leq I(U;Y)\,, && (2.33)\\ \Delta &\leq H(X|UZ)\,, && (2.34) \end{aligned}$$

where $U \multimap X \multimap (Y, Z)$ form a Markov chain.

Note that equivocation is independent of the statistical differences between Y and Z. In contrast to the rate-distortion-equivocation setting previously studied in [78], in the current setting of testing against independence, the analog information Y available at Bob cannot help to improve the equivocation rate at the Eavesdropper. This observation can simply be explained by the fact that in both cases where Y is independent of X or not, then Alice must encode for the worst case and cannot use it to perform binning (e.g. Wyner-Ziv coding).
The proof of both the achievability and converse parts of Proposition 2.3 are given in Appendix A.3.

2.4.2 Eve is the Statistician

We now assume that Eve is an untrusted statistician, i.e., $Y^n = Z^n$. In addition to this, Alice wishes to maximize the equivocation only with respect to a specific part of the source X^n that is denoted by S^n, where $f(X^n) \multimap X^n \multimap (S^n, Y^n)$. Eve observes the same signal Y^n. In this setting, the equivocation rate is given by

$$\frac{1}{n}H(S^n|f(X^n)Y^n) \geq \Delta\,. \quad (2.35)$$

Note that this definition does not affect the coding rate and the exponent and thus the other quantities remains the same.

Proposition 2.4 (Detection under privacy constraints) *Let $\mathcal{R}^\star$ be the set of all achievable tuples (R, E, Δ), then there exists a random variable U on some finite set $\mathcal{U}$ satisfying:*

$$\begin{aligned} R &\geq I(U;X) \,, && (2.36)\\ E &\leq I(U;Y) \,, && (2.37)\\ \Delta &\leq H(S|UY) \,, && (2.38) \end{aligned}$$

where $U \multimap X \multimap (Y, S)$ form a Markov chain.

Proof 2.1 *We first prove the direct part, by replacing X^n with S^n in [78, Lemma 2], we obtain:*

$$H(S^n|rY^n) \geq n\left[H(S|UY) - \eta_n\right] \,, \tag{2.39}$$

provided that n is large enough. In order to show the converse, we set again $U_i = (J, Y^{i-1})$ with $J = f(X^n)$ and thus

$$\begin{aligned} \Delta \leq \frac{1}{n}H(S^n|JY^n) &= \frac{1}{n}\sum_{i=1}^{n} H(S_i|JS^{i-1}Y^{i-1}Y_iY_{i+1}^n) && (2.40)\\ &\leq \frac{1}{n}\sum_{i=1}^{n} H(S_i|U_iY_i) && (2.41)\\ &= H(S|UY) \,. && (2.42) \end{aligned}$$

This concludes the proof of the proposition.

2.4.3 X follows an arbitrary distribution under Hypothesis H_1

In many situations, the marginal distribution of the source X under the alternate hypothesis H_1 could be unknown, and hence the problem should be modified to take this idea into consideration. In this case, our two hypotheses will be defined as follow:

$$\begin{cases} H_0: & p_{XY} \,, \\ H_1: & \tilde{p}_X \times p_Y \,, \end{cases} \tag{2.43}$$

where $\tilde{p}_X$ is any arbitrary distribution. The problem becomes as follow, we need to find the optimal region (R, E, Δ) satisfying the following conditions:

$$\begin{aligned} \frac{1}{n}\log\|f\| &\leq R \,, && (2.44)\\ \frac{1}{n}D(p_{f(X^n)Y^n}||\tilde{p}_{f(X^n)}p_{Y^n}) &\geq E \,, \forall\tilde{p} && (2.45)\\ \frac{1}{n}H(X^n|f(X^n)Z^n) &\geq \Delta \,. && (2.46) \end{aligned}$$

and hence in other words:

$$\begin{aligned} \tfrac{1}{n}\log\|f\| &\leq R\,, \\ \tfrac{1}{n}\min_{\tilde{p}} D(p_{f(X^n)Y^n}||\tilde{p}_{f(X^n)}p_{Y^n}) &\geq E\,, \\ \tfrac{1}{n}H(X^n|f(X^n)Z^n) &\geq \Delta\,. \end{aligned} \tag{2.47}$$

Define $p_X = \sum_x p_{XY}$, then it can be shown that

$$D(p_{XY}||\tilde{p}_X p_Y) = D(p_{XY}||p_X p_Y) + D(p_X||\tilde{p}_X) \tag{2.48}$$

Since $D(p_X||\tilde{p}_X) > 0$, then we can say that:

$$\begin{aligned} \min_{\tilde{p}} D(p_{f(X^n)Y^n}||\tilde{p}_{f(X^n)}p_{Y^n}) &= D(p_{f(X^n)Y^n}||p_{f(X^n)}p_{Y^n}) \\ &= I(f(X^n);Y^n) \end{aligned} \tag{2.49}$$

And hence the solution of the optimal region remains the same.

2.4.4 Testing on Binary Sources

Consider the source model depicted in Fig. 2.2 where the source X is binary and the sources Y and Z are the outputs with input X of a Binary Erasure Channel (BEC) with erasure probability $\epsilon \in [0, 1/2]$ and a Binary Symmetric Channel (BSC) with crossover probability $p \in [0, 1/2]$, respectively.

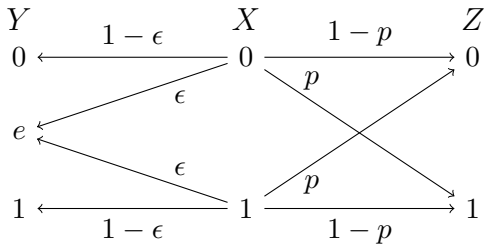


Figure 2.2: Testing on binary sources.

Assume that the source is uniformly distributed , i.e., $\Pr(X = 0) = \Pr(X = 1) = 1/2$.

Proposition 2.5 (Rate-exponent-equivocation region) *The rate-exponent-equivocation region $\mathcal{R}^\star$ is the set of all tuples (R, E, Δ) such that there exist $\alpha \in [0, 1/2]$ satisfying:*

$$R \geq 1 - H_2(\alpha)\,, \tag{2.50}$$

$$E \leq (1-\epsilon)\Big(1 - H_2(\alpha)\Big)\,, \tag{2.51}$$

$$\Delta \leq H_2(p) + H_2(\alpha) - H_2(\alpha \star p)\,. \tag{2.52}$$

This proposition is proved in Appendix A.4.

2.4.5 Testing on Gaussian Sources

Consider the source model depicted in Fig. 2.3 where the source at Alice is a standard Gaussian and the other sources are the outputs of additive white Gaussian noise (AWGN) channels with input X, gains ρ_Y, ρ_Z, and noise powers $1-\rho_Y^2$, $1-\rho_Z^2$, respectively, for some $0 < \rho_Y, \rho_Z < 1$.

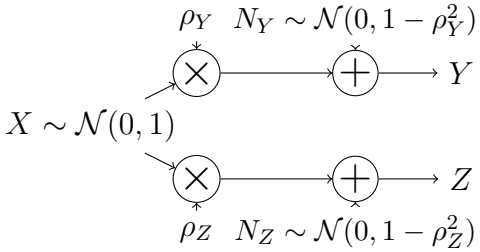


Figure 2.3: Testing on Gaussian sources.

Proposition 2.6 (Inner rate-exponent-equivocation region) *A set of all tuples (R, E, Δ) is achievable if the following inequalities are simultaneously satisfied:*

$$R \geq \frac{1}{2}\log\left(\frac{1}{1-{\rho_U}^2}\right), \tag{2.53}$$

$$E \leq \frac{1}{2}\log\left(\frac{1}{1-{\rho_U}^2{\rho_Y}^2}\right), \tag{2.54}$$

$$\Delta \leq \left[\frac{1}{2}\log 2\pi e\frac{(1-{\rho_U}^2)(1-{\rho_Z}^2)}{1-{\rho_U}^2{\rho_Z}^2}\right]_+ . \tag{2.55}$$

The proof is given in Appendix A.5.

2.4.6 Numerical Results

Based on the results stated in Proposition 2.6, we now plot the error exponent E as a function of the equivocation Δ, for the same string observations at Bob and Eve, i.e., $Y = Z$ with $\rho_Y^2 = \rho_Z^2 = 0.5$ and different coding rates. By observing Fig. 2.4, we notice that the error exponent is decreasing as a function of the equivocation which is an expected outcome since if Alice wishes to increase the exponent (decreasing the error probability) then the desired level of privacy might not be guaranteed anymore. Thus, there exists a tradeoff between the number of encoded bits (from the total amount R of allowed bits) and the equivocation rate. Indeed, for each value R, the exponent E is limited by two factors: the coding rate R, when the equivocation is small, and the desired equivocation rate Δ, when the coding rate is above a certain threshold.

We next plot the same function for different correlation values ρ_Y^2 and an arbitrary large coding rate R, as shown in Fig. 2.5. From this picture, we can observe that as the correlation increases, the error exponent increases as well while the equivocation decreases. This can be explained by the fact that an increase of the correlation implies a higher information leakage at Eve. Otherwise the behavior of the tradeoff between the exponent and the equivocation rate remains unchanged.

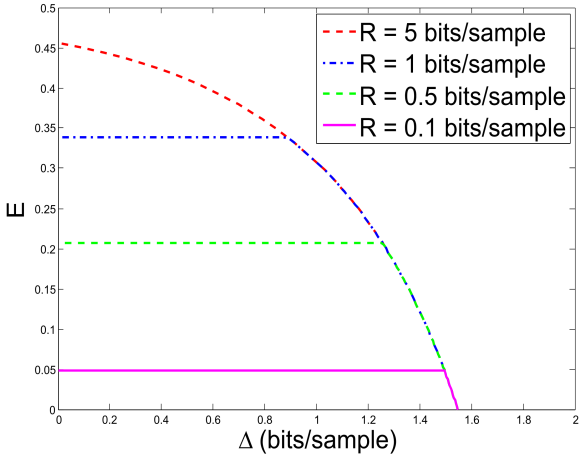


Figure 2.4: Error exponent as a function of the equivocation rate for different coding rates.

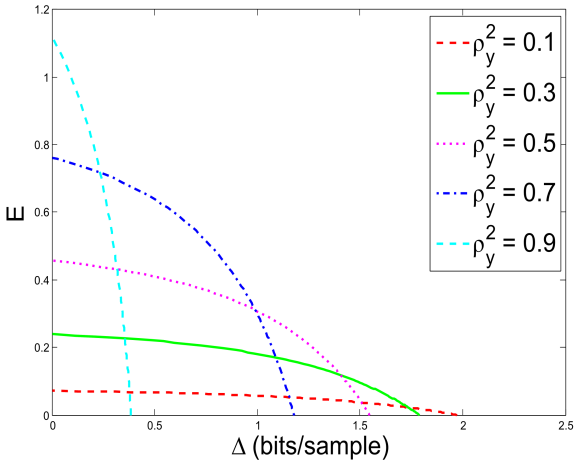


Figure 2.5: Error exponent as a function of the equivocation rate for different correlation values ρ_Y^2 with an arbitrary large coding rate.

2.5 Conclusion

In this chapter, we focused on the problem of hypothesis testing with both communication and privacy constraints. Bob needs to choose between two hypotheses based on the available information and the information remotely communicated by Alice. Indeed, Alice communicate over a public error-free but rate-limited channel. The goal is to guarantee a desired error exponent at Bob for a given rate while satisfying an average equivocation rate at Eve.

When testing against independence, we were able to characterize the optimal trade-off between the rate, the error exponent requirement and the privacy guarantees. Only one encoding layer without binning was needed. In the case of general hypothesis testing, an approach based on the method of types was used in order to derive an achievable rate-error-equivocation region without being able to prove its optimality.

CHAPTER 3

QUANTIZATION LEARNING FOR DISTRIBUTED BINARY DECISION WITH PRIVACY CONSTRAINTS

Abstract — *The design of a quantizer for distributed binary decision in the presence of an eavesdropper (Eve) is investigated. An encoder/quantizer (Alice) observes a source and communicates it via a public noiseless rate-limited channel with the detector (Bob) who has also access to a correlated analog source. Bob can take advantage of both information to perform a binary decision on the joint probability law of these observations. Eve is further assumed to have access to a different correlated analog source and perfectly observe the information bits sent by Alice. This chapter evaluates the various trade-offs between the probabilities of error (on the decision) depending on the amount of information leakage from Alice to Eve in both scalar and vector quantization cases. The particular case of Testing against independence is studied for the case of vector quantization where the error exponent reduces to the mutual information between the two sources, in which case a solution is obtained via an algorithm very similar to that given by Lloyd-Max. Numerical results for memoryless Gaussian sources first demonstrate the performance of the proposed quantization methods, and a practical application involving electric consumption data measured from real houses is then investigated.*

Keywords — *Hypothesis testing; distributed; AUC; error; quantization; secrecy; source coding; distortion measure.*

3.1 Introduction

Many applications require a message to be transmitted from an information source to a desired destination, where decisions are made on the source based on the received data. For instance, data might be sent by an object-detection radar or a video surveillance camera to a monitoring station interested in detecting a specific target or object in the radar's range of vision. In such an application, it is frequently crucial to decrease the rate of transmitted data by encoding the source prior to transmission, and the fact that the receiver has only to make a test, and not to estimate the original signal provides more degrees of freedom to reduce the transmission rate. In our setting, it is even further assumed that the signal which is transmitted, if listened by an eavesdropper, do not provide too much information on the original signal.

Obviously, quantization is the tool of choice for reducing the rate, but, in contrast with the classical situation, distortion is the design criterion for optimizing the quantization. In our case, we are interested in the best possible performance for the test while ensuring that the raw data remain secure w.r.t. an eventual eavesdropper.

Even without any privacy constraint, a first problem is to define an appropriate distortion measure allowing to design a quantizer adapted to hypothesis testing. The mean square error (MSE) is the appropriate distortion measure when the goal is to estimate the source from the corresponding quantized values. However, in detection problems, the main degradation measure should be the type II probability of error (the failure to reject a false null hypothesis) for a prescribed type I probability of error (the incorrect rejection of a true null hypothesis). Minimizing the error probability to determine the optimal quantization scheme is difficult and inconvenient to carry out on many occasions. Consequently, several sub-optimal criteria which are more flexible to estimate and manipulate have been used in this situation.

Contributions on the problem of designing a scalar quantizer for detection or classification systems are numerous. Original work in this area was particularly done by Kassam, Poor, Picinbono and Bucklew; which aim to design a quantizer that optimizes a decision rule based on quantized information. For hypothesis testing, optimal quantizations have been analysed for a variety of quantization schemes and various distortion criteria [79, 80, 81, 82, 83, 84, 85, 86].

The loss induced by quantization in various *Ali-Silvey distances*, measuring the difference between two probability distributions, was investigated by [80] and applied to binary detection problems. A comparison between the properties of the well-known *Kullback-Leibler Divergence* (KLD) with a measure referred to as *Bhattacharyya distance* was first provided by [87]. This is indeed one of the Ali-Silvey distances that have successfully been applied to signal detection and statistical optimization. Subsequently, [88, 89] proposed the generalized f-divergences as distortion measures for an efficient discrimination between statistical hypotheses. Later, [90, 91] used the empirical divergence maximization (EDM), i.e., a proposed strategy to estimate the KLD when maximized over a class of quantization rules.

Some results also addressed distributed detection, such as [92] where authors study one-bit quantization for testing against independence between a remote and an available source. Bayesian decentralized binary decision was investigated in [93], where N sensors observe N random samples whose joint probability law is unknown but can only be one out of two possibilities. Sensors communicate one-bit messages to a fusion center and quantization is based on Chernoff's bound (see also [82]) in the asymptotic regime of N with large number of sensors. The problem of scalar quantization for hypothesis testing has been studied by [84], the encoders were scalar, optimization was based on the Bhattacharyya coefficient, and the decision rule was based on likelihood ratio given by the Neyman-Pearson lemma as the most powerful decision rule [94].

More recently, [95] investigated high rate LLR quantizers of i.i.d. sources for many reconstruction and detection criteria and succeeded in writing a compact expression of the error exponent induced by quantization. Later, [96] extended their results to the case where a sensing unit observes samples of a correlated stationary ergodic multivariate process, i.e. the case of non-independent observations. Their

main contribution includes providing a distortion measure in the high-rate regime i.e., when the rate of quantization tends to infinity.

In our problem, the joint probability distribution of two remote sources is assumed to be known under both the null hypothesis H_0 and the alternative hypothesis H_1, and the most powerful Neyman-Pearson hypothesis test is used as a decision rule (checking the log-likelihood ratio (LLR) against a given threshold T [38]). The likelihood ratio test rejects the null hypothesis if the LLR value of this ratio is too small. How small is too small depends on the significance level of the test, i.e., on what probability of type I error α is considered tolerable. The significance level of the test is determined by the given threshold T. By fixing the type I probability of error to a specified small amount, say ε, the performance of the test can be thus computed in terms of the type II probability of error β.

Detection errors do depend on the quantizer. Therefore, we need to design the quantizer in such a way that the errors are minimized. Unfortunately, such formulations do not lead -in most cases- to computationally tractable algorithms. However, when the number of samples is very large, an asymptotic expression for the probabilities of error can be written using Stein's lemma [77] where the type II probability of error is related to the Kullback-Leibler Divergence referred to also by the error exponent when $\alpha \leq \varepsilon$. In that case the value of β is not affected by α. Based on this result, designing the optimal quantizer is equivalent to maximize the error exponent. However, tuning the quantizer so that the error exponent is maximized can be done explicitly only in the special case of testing against independence. In the general case, other distortion measures have to be considered.

In this chapter, we consider the problem of distributed binary detection operating over insecure links. This mainly consists of testing the joint probability of two distributed sources in the presence of an eavesdropper, as described in Figure 3.1. A similar framework was previously investigated in [97], where binary quantizers are derived in presence of an eavesdropper by maximizing the KLD at the fusion center while constraining the eavesdroppers KLD to a prescribed level. It is worth mentioning that in our setup it is not possible to take the decision at sensor level since it is based on the joint distribution only available at the decision center. Therefore, taking the decision at the sensor level allowed them to consider as distortion metric the KLD between the posterior probabilities of the two hypotheses H_0 and H_1. On the other hand, [98] studied the problem of designing quantizers for a distributed detection network that maximizes the difference in the KLD at the fusion center and the eavesdroppers.

On the other hand, [99] investigated the problem of designing optimal decision rules for a censoring sensor network in the presence of eavesdroppers. In their paper, they assumed that Eve can only determine whether the sensor transmits its decision or not. In reality, Eve can extract much more information than just merely determining the presence or absence of transmission, and hence can make a reasonably good approximation of the source.

In our previous work, [100], we studied the fundamental limits of the problem, i.e. the trade-off between the maximum achievable error exponent at the detector, (i.e., the minimum type II error probability for a fixed type I probability of error), the coding rate at the encoder and the leakage rate at the eavesdropper. In this chapter, we propose practical algorithms to obtain solutions that are as close

as possible to the optimal, which requires the design of optimal quantization in the presence of an eavesdropper.

This chapter makes two contributions: The first part will focus mainly on the scalar quantization where the testing is between two general probability distributions [101]. In this case, the *Bhattacharyya coefficient* is used as a distortion measure since it is much easier to manipulate than the standard divergence and guarantees very close performs to those given by the KLD. The second part focuses on vector quantization for the case of testing against independence. A general algorithm of optimization is given in both cases, and the performance for the of case Gaussian sources is evaluated.

This theoretical results for this scenario would truly come up in practice in many different ways. In this chapter, the application to smart meters will show how testing against independence can be used to test the integrity of the smart devices present at the houses. The joint distribution of some data collected at the smart meter and some other data available at the provider allows the collector to check whether the smart meter behaves properly. Applications of our results arise in these contexts where data must remain private even from the statistician (see [102] and references therein) where any absence of correlation may indicate a smart meter deficiency and hence a fault detection.

This chapter is organized as follow. Section 3.2 provides notations and main definitions and Section 3.3 presents the algorithm for scalar quantization of general hypothesis testing and main results. Section 3.4 on the other hand discusses the case of vector quantization for testing against independence and provides an optimization algorithm similar to that of Lloyd-max for the case of detection. Section 3.5 demonstrates the performance for testing against independence in the case of Gaussian sources. Section 3.6 introduces a smart grid application to our problem where the main concern is to detect a smart meter deficiency in the presence of a potential eavesdropper. Concluding remarks are given in Section 3.7.

3.2 Problem Definition

3.2.1 Notations

Random variables are usually written in upper case letters: X, Y, etc. The cardinality of X is denoted by $\|\mathcal{X}\|$ or simply $\mathcal{X}$. Particular realizations of a random variable are written in corresponding lower case letters. For example $x_1, x_2, \ldots, x_t$ could be samples corresponding to the random variable X and a cumulative probability is formally written $P(X > x)$ to differentiate random variable from realization. We use x^n; also $\mathbf{x}$; to denote vectors in $\mathcal{X}^n$ of length n. The *Probability density functions* (pdfs) and the *probability distributions* are denoted by the lower case letter p. The distribution and the joint distribution of X and Y will be denoted p_X, p_Y and p_{XY} respectively.

For RVs X and Y with joint PD p_{XY}, $H(X), H(Y|X)$ denote the *entropy* and the *conditional entropy*, respectively, and $h(X)$, $h(Y|X)$ denote the *differential entropies*, while $I(X;Y)$ denotes the *mutual information*. Let (X, Y, Z) be three RVs such that $P(x|y, z) = P(x|y)$ for each (x, y, z), then $X \multimap Y \multimap Z$ form a Markov Chain.

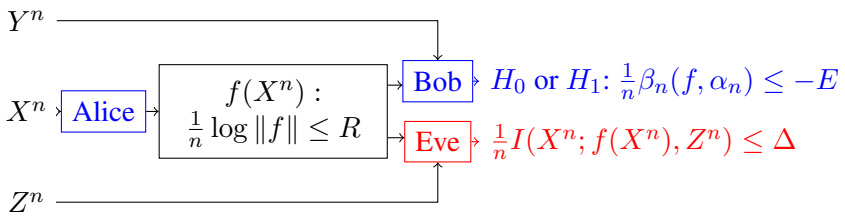


Figure 3.1: Distributed detection under privacy constraints.

3.2.2 System Model

Our model is defined by three nodes as described in Figure 3.1. Alice (the quantizer) observes a sequence of vectors $\mathbf{x}^\tau = (\mathbf{x}_1, \ldots, \mathbf{x}_\tau)$ of i.i.d. samples while Bob (the detector) observes another sequence of i.i.d. samples $\mathbf{y}^\tau = (\mathbf{y}_1, \ldots, \mathbf{y}_\tau)$. These vectors are n-dimensional, i.e.

$$\mathbf{x}_t = (x_{t,1}; x_{t,2}; \ldots; x_{t,n}), \quad t = 1, 2, \ldots, \tau \; . \tag{3.1}$$

τ will be referred to as the number of sample vectors available at both the encoder and the decoder. Although sample vectors are considered to be i.i.d. with respect to each other, each vector can involve samples with memory.

Alice wishes to encode its data with a maximum rate R [bits per dimension] which is accomplished by mapping the inputs into the quantized values $j^\tau \equiv (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_\tau))$, using an n-dimensional vector quantizer:

$$f : \mathcal{X}^n \longrightarrow \mathcal{J} \equiv \{1, \ldots, M\} \; . \tag{3.2}$$

The transformation is done using a codebook of size $||\mathcal{J}|| = M = 2^{nR}$. Each vector is encoded by an index pointing to some codeword from a finite set of vectors, called the codebook. Each codeword; also called reproduction vector is also n-dimensional.

$\mathcal{J}$ is the set of all possible binary words sent through the channel to represent the original vector $\mathbf{x} \in \mathcal{X}^n$. Since the set of indices $\mathcal{J}$ is discrete and the set $\mathcal{X}^n$ is continuous, the mapping function is non-injective. The set of different input vectors producing same output value will be referred to as the quantization region. Let $\mathcal{R}_j$ be the encoding region associated with the index j. $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M\}$ denotes the partition of the space. That is, the regions are disjoint and exhaustive. If the source vector $\mathbf{x}$ is in the encoding region $\mathcal{R}_j$, then its representation is the jth codeword in the codebook:

$$f(\mathbf{x}) = j, \quad \text{if } \mathbf{x} \in \mathcal{R}_j \; . \tag{3.3}$$

The detector Bob receives the message j^τ communicated by Alice and the sequence $\mathbf{y}^\tau$. His goal is to make a decision between two possibilities of the joint probability law of (X^n, Y^n) as it can only be one out of two hypotheses:

$$\begin{cases} H_0 : & (X^n, Y^n) \sim p_{X^n Y^n}(\mathbf{x}, \mathbf{y}) \; , \\ H_1 : & (X^n, Y^n) \sim p_{\bar{X}^n \bar{Y}^n}(\mathbf{x}, \mathbf{y}) \; . \end{cases} \tag{3.4}$$

It is further assumed that the marginal distributions of X^n and Y^n are the same under both hypotheses,

i.e., $p_{X^n} = p_{\bar{X}^n}$ and $p_{Y^n} = p_{\bar{Y}^n}$ which does not allow Bob to make the decision without the information sent by Alice. Let $\hat{H}_0$ and $\hat{H}_1$ be the possible outcomes of the decision process. The two types of error probabilities associated with the given detection problem are given by

$$\begin{array}{lll} \text{Type I :} & \alpha_n \triangleq & \Pr\left(\hat{H}_1 | X^n Y^n \sim p_{X^n Y^n}(\mathbf{x}, \mathbf{y})\right) , \\ \text{Type II :} & \beta_n \triangleq & \Pr\left(\hat{H}_0 | X^n Y^n \sim p_{\bar{X}^n \bar{Y}^n}(\mathbf{x}, \mathbf{y})\right) . \end{array} \tag{3.5}$$

The message $f(\mathbf{x}^\tau)$ is sent to Bob via a public link that is perfectly overheard by Eve (the eavesdropper) who may have access to another sequence of vectors $\mathbf{z}^\tau$. The random variable Z^n is arbitrary correlated to (X^n, Y^n). The goal of this system is then to find a quantization function f and an acceptance region that minimize the type II probability for a prescribed type I probability, and ensures that the information leakage at Eve is bounded, i.e.,

$$\frac{1}{n} I(X^n; f(X^n), Z^n) \leq \Delta . \tag{3.6}$$

In practice, using the probability density functions in order to perform necessary computations is not practical, most often training sets are used instead. This training sequence can be obtained from some large database. For example, if the source is a speech signal, then the training sequence can be obtained by recording several long phone conversations. The training set is assumed to be sufficiently large so that all the statistical properties of the source are captured. We will refer to $\mathcal{T}_{\mathbf{X}}$, $\mathcal{T}_{\mathbf{Y}}$ and $\mathcal{T}_{\mathbf{Z}}$, the training sequences of $X^n \sim P_{X^n}$, $Y^n \sim P_{Y^n}$ and $Z^n \sim P_{Z^n}$ respectively.

The design problem can be thus stated as follows: given the training set, the statistical properties $p_{X^n Y^n}$ and $p_{\bar{X}^n \bar{Y}^n}$, the rate R and the allowed amount of information leakage Δ, find a codebook and a partition $\boldsymbol{\mathcal{R}}$ which result in the smallest average distortion. The main concern would be to find the best distortion measure that represents the hypothesis testing problem in such a way involving the probabilities of errors of the testing problem at Bob, and while also considering the information leakage at Eve.

3.3 Scalar Quantization

The set of inputs of a quantizer can be scalars or vectors. If they are scalars, we call the quantizers scalar quantizers. If they are vectors, we call them vector quantizers. Scalar quantizers have many advantages including the facts that they are the simplest and the cheapest. Furthermore, they are usually memoryless and thus avoid transmission delays caused by complex signal processing at the node. Moreover, scalar quantizers generally perform well in terms of estimation criteria, therefore there's no reason to believe they cannot perform in terms of decision criteria.

In this section, the case of scalar quantization is considered, i.e, the case of $n = 1$. If the initial signal is analog, then the first step to be accomplished prior to quantization is sampling. As stated in the problem definition, we will suppose that after the sampling process we ended up with τ series of scalars available at the level of the quantizer. Since we're using a scalar quantizer, each input scalar is treated separately in producing the output, i.e., the inputs $x^\tau = (x_1, \ldots, x_\tau)$ are mapped to the

quantized values $j^\tau \equiv (f(x_1), \ldots, f(x_\tau))$, using a scalar quantizer:

$$f : \mathcal{X} \longrightarrow \mathcal{J} \equiv \{1, \ldots, M\} , \tag{3.7}$$

with coding rate $||\mathcal{J}|| \leq 2^R$.

3.3.1 Quantizer Design

A scalar quantizer maps each input value to a quantization region by comparing the input value to the quantizer boundary endpoints. Quantization regions are hence reduced into intervals and the quantizer outputs the index of the associated interval. Consider the objective of transmitting a sample x of the source X characterized by a probability density function (pdf) p_X over a channel that can only carry R bits, each time it is used. That is, we can only use R bits to encode each sample of X. Naturally, this restriction implies that we are forced to encode any outcome of X into one of $M = 2^R$ different symbols. The M *decision intervals* are defined by the following $M + 1$ endpoints

$$a_j, j \in \{0, 1, \ldots, M\} , \tag{3.8}$$

These endpoints will also be called *decision boundaries* or *decision thresholds*. A source sample value x is quantized to the quantization index j if and only if x falls into the jth *decision interval*

$$f^{-1}(j) =]a_{j-1}, a_j] , \tag{3.9}$$

such that

$$a_0 = -\infty < a_1 < \cdots < a_{M-1} < a_M = +\infty \tag{3.10}$$

The most powerful decision rule $\gamma_\tau : \mathcal{J}^\tau \times \mathcal{Y}^\tau \rightarrow \{\hat{H}_0, \hat{H}_1\}$ implements the *Neyman-Pearson* lemma. We denote by $p_\theta^\tau(j^\tau, y^\tau)$ a short notation for $p_Y^\tau(y^\tau)P(J^\tau = j^\tau | Y^\tau = y^\tau, H_\theta)$; the probability of the observed outcomes of (J^τ, Y^τ) given the the hypothesis $H_\theta, \quad \theta \in \{0, 1\}$. The notation $p_\theta^\tau(j^\tau, y^\tau)$ is reduced to $p_\theta(j, y)$ when $\tau = 1$. Let $\Lambda(j^\tau, y^\tau)$ be the *Likelihood-Ratio* with j^τ being the indices recieved from the quantizer when τ samples are available at the quantizer [94]:

$$\Lambda(j^\tau, y^\tau) \equiv \frac{p_1^\tau(j^\tau, y^\tau)}{p_0^\tau(j^\tau, y^\tau)} . \tag{3.11}$$

The likelihood ratio expresses how many times more likely the data are under one model than the other. The decision rule consists in accepting the alternative hypothesis when the log-likelihood ratio (LLR) surpasses a given non-negative threshold T. We define the acceptance region for H_0 such that:

$$\mathcal{A}_\tau \equiv \left\{(j^\tau, y^\tau) \in \mathcal{J}^\tau \times \mathcal{Y}^\tau : \Lambda(j^\tau, y^\tau) < T\right\} , \tag{3.12}$$

In other words, the decision rule γ_τ at the level of Bob is defined as:

$$\gamma_\tau(j^\tau, y^\tau) = \begin{cases} \hat{H}_0 & \text{if } (j^\tau, y^\tau) \in \mathcal{A}_\tau \ , \\ \hat{H}_1 & \text{if } (j^\tau, y^\tau) \in \mathcal{A}_\tau^c \ . \end{cases} \tag{3.13}$$

With τ samples sent, the corresponding error probabilities can be computed as:

$$\alpha(f,T) \equiv \Pr\left\{\hat{H}_1 | H_0\right\} = \sum_{j_1=1}^{2^R} \dots \sum_{j_\tau=1}^{2^R} \int_{\Lambda(j^\tau,y^\tau)>T} p_0^\tau(j^\tau, y^\tau) dy^\tau \ , \tag{3.14}$$

$$\beta(f,T) \equiv \Pr\left\{\hat{H}_0 | H_1\right\} = \sum_{j_1=1}^{2^R} \dots \sum_{j_\tau=1}^{2^R} \int_{\Lambda(j^\tau,y^\tau)<T} p_1^\tau(j^\tau, y^\tau) dy^\tau \ . \tag{3.15}$$

The two types probabilities of error have negative correlation with respect T. Tuning threshold T allows us to manipulate the values of the probability of errors as we desire and depending on what type of error we are more interested.

Difficulties arise because of the natural decision criterion to optimize. In this case, optimizing these probabilities does not yield a tractable design procedure. For this reason, we replace the natural criterion by a measure of dissimilarity (or distributional distance) between the distributions under the hypotheses. The KLD is a non-symmetric measure of the difference between two probability distributions. Another quite broad class of distributional distances called *Ali-Silvey distances* which are most frequently used and have found successful application in statistical optimization. In what follows we will focus on the *Bhattacharyya distance* defined as:

$$D_B = -\log\left[E_0\left(\sqrt{\Lambda}\right)\right] \ . \tag{3.16}$$

With Λ being the LLR defined in Equation (3.11), E_0 is the mean with respect to the distribution p_0. We choose to optimize f according to the *Bhattacharyya distance* which is equivalent to minimizing the so called *Bhattacharyya coefficient* d_B given by

$$d_B(p_0^\tau, p_1^\tau) = E_0[\sqrt{\Lambda(J^\tau, Y^\tau)}] \ . \tag{3.17}$$

The sequence x^τ is quantized into the message $j^\tau = (j_1, j_2, \dots, j_\tau)$ and thus $j_t \in \{1, \dots, M\}$ are also i.i.d. for $t = \{1, \dots, \tau\}$. The *Bhattacharyya coefficient* d_B, for τ samples, can be evaluated as:

$$d_B(p_0^\tau, p_1^\tau) = \sum_{j_1=1}^{2^R} \dots \sum_{j_\tau=1}^{2^R} \int_{\mathcal{Y}^\tau} \sqrt{p_0^\tau(j^\tau, y^\tau) p_1^\tau(j^\tau, y^\tau)} dy^\tau \tag{3.18}$$

$$= \left(\sum_{j=1}^{2^R} \int_{\mathcal{Y}} [p_0(j,y) p_1(j,y)]^{1/2} dy \right)^\tau \equiv (d_B(p_0, p_1))^\tau \ . \tag{3.19}$$

In terms of the decision thresholds the Bhattacharyya coefficient per sample can be expressed as:

$$d_B(p_0,p_1)=\sum_{j=1}^{2^R}\int_{\mathcal{Y}}[\int_{a_{j-1}}^{a_j}p_{XY}(x,y)dx\int_{a_{j-1}}^{a_j}p_{\bar{X}\bar{Y}}(x,y)dx]^{1/2}dy\ . \tag{3.20}$$

Optimizing $(d_B(p_0,p_1))^\tau$ is equivalent to optimizing $d_B(p_0,p_1)$. Hence, the optimal quantization thresholds do not depend on τ which was expected as we're dealing with scalar quantization for i.i.d. samples where each sample is processed separately. From now on, the term $d_B(p_0,p_1)$ will be simply denoted d_B. The optimization problem with privacy constraints can thus be seen in the following two different ways. Given a maximum information leakage as a threshold, we would like to maximize the error exponent, thus minimize the *Bhattacharyya coefficient* d_B. Or, given a minimum error exponent threshold at the level of the detector, we aim at minimizing the information leakage. These problems read as follows:

$$\textbf{Problem 1}: \quad \min_{f} \quad d_B \qquad \text{subject to} \quad \tau^{-1}I(X^\tau;J^\tau,Z^\tau)\leq\Delta\ ,$$

$$\textbf{Problem 2}: \quad \min_{f} \quad \tau^{-1}I(X^\tau;J^\tau,Z^\tau) \qquad \text{subject to} \quad -\log d_B\geq E\ .$$

J is another way to write the random variable $f(X)$; a discrete random variable representing the index of the interval being sent by the quantizer. The information leakage based on the mutual information can be expressed by:

$$\begin{aligned}\frac{1}{\tau}I(X^\tau;J^\tau,Z^\tau) &= I(X;J,Z) &(3.21)\\ &= I(X;J)-I(Z;J)+I(X;Z)\ . &(3.22)\end{aligned}$$

Equation (3.21) is due to the memoryless property of $(J^\tau,X^\tau,Y^\tau,Z^\tau)$. Equation (3.22) is obtained using the chain rule and the fact that $J \multimap X \multimap Y$ form a Markov chain. The mutual information $I(X;J)=H(J)$ represents the actual rate of the quantizer. Note that for a given joint distribution of the random variables (X,Z), the value $I(X;Z)$ is constant, hence represents the minimum possible value of the information leakage Δ. An information leakage smaller than $\Delta_{\min}=I(X;Z)$ is not achievable. More specifically, Problem 1 is meaningful only when $\Delta_{\min}=I(X;Z)\leq\Delta<\Delta^\star$. This critical value $\Delta^\star$ is equal to Eve's mutual information $I(X;J,Z)$, which Eve attains when the detector attains the minimum *Bhattacharrya coefficient* $d_B^\star$. This minimum coefficient $d_B^\star$ can be found by solving Problem 1 in the absence of the constraint. A possible formulation of an iterative algorithm to solve the problem 1 is outlined by Algorithm 1.

$$\psi_j^{(k)}(a_j)=\int_{\mathcal{Y}}[\int_{a_{j-1}^{(k)}}^{a_j}p_{XY}(x,y)dx\int_{a_{j-1}^{(k)}}^{a_j}p_{\bar{X}\bar{Y}}(x,y)dx]^{1/2}+[\int_{a_j}^{a_{j+1}^{(k)}}p_{XY}(x,y)dx\int_{a_j}^{a_{j+1}^{(k)}}p_{\bar{X}\bar{Y}}(x,y)dx]^{1/2} \tag{3.26}$$

Algorithm 1 Scalar Quantizer Design for Distributed Binary Decision with Privacy Constraints

1. **Scalar quantizer:** For a given maximum rate R, $M = 2^R$ is the number of *quantization regions*. The designer is looking to find the $M+1$ *decision thresholds* defined by $\{a_0 = -\infty, a_1, \ldots, a_{M-1}, a_M = +\infty\}$ that minimize the *Bhattacharrya coefficient* under a maximum allowed information leakage. The optimal quantization thresholds corresponds to the M indices such that:
$$f^{-1}(j) =]a_{j-1}, a_j] \, . \tag{3.23}$$

2. **Initialization:** Initialize the *decision thresholds* with $\{a_0 = -\infty, a_1^{(0)}, \ldots, a_{M-1}^{(0)}, a_M = +\infty\}$ and calculate the corresponding initial *Bhattacharyya coefficient* $d_B^{(0)}$. Make sure the *decision thresholds* are initialized such that $I(X; J, Z)^{(0)} \leq \Delta$. Set $k = 0$.

3. **Iteration:**

   (a) For $j = \{1, \ldots, M-1\}$: Optimize $a_j^{(k)}$ with respect to $d_B^{(k)}$, while keeping constant $a_0, \ldots, a_{j-1}^{(k)}, a_{j+1}^{(k)}, \ldots, a_M$, The new threshold will be called $a_j^{(k+1)}$. The function being minimized is the sum of M non-negative functions, where two of them only dependent on $a_j^{(k)}$.
   $$a_j^{(k+1)} = \arg\min_{a_j} \psi_j^{(k)}(a_j) \tag{3.24}$$
   $$\textbf{subject to} \quad I(X; J, Z)(a_j) \leq \Delta \, , \tag{3.25}$$
   The expression of ψ_j is shown in Equation (3.26).

   (b) Calculate the corresponding *Bhattacharyya coefficient* and information leakage $d_B^{(k+1)}$ and $I(X; J, Z)^{(k+1)}$ respectively.

   (c) if $\left| d_B^{(k+1)} - d_B^{(k)} \right| \leq \delta$, stop with $a_0^{(k+1)}, \ldots, a_M^{(k+1)}$ the final quantizer's parameters. Else let $k \to k+1$ and return to (a).

Remark 3.1 *Without a privacy constraint, if the marginal distribution of X under both hypotheses H_0 and H_1 is symmetric, then the optimal quantization of X is also symmetric. Therefore, Our problem could be reduced into finding $\frac{M}{2} - 1$ decision thresholds if M is even, or $\frac{M-1}{2}$ decision thresholds if M is odd. Suppose M is even, then $\forall k \geq 0$, we can write*

$$\forall a_j^{(k)} \textit{ such that } \quad j \in \{0, \ldots, \tfrac{M}{2} - 1\}, \quad a_j^{(k)} < 0 \, , \tag{3.27}$$
$$\forall a_j^{(k)} \textit{ such that } \quad j \in \{\tfrac{M}{2} + 1, \ldots, M\}, \quad a_j^{(k)} = -a_{M-j}^{(k)} > 0 \, , \tag{3.28}$$
$$a_{M/2}^{(k)} = 0 \, . \tag{3.29}$$

3.4 Vector Quantization

By combining source inputs together and encoding them as one single block, we can accomplish more efficient compression algorithms. Although vector quantizers are generally more complex and spend

more processing time, the adoption of vector quantization for a given rate will most probably achieve lower errors than when scalar quantization is used at the same rate. Going from one dimension to multi-dimensions is a major step that allows many new ideas, concepts, techniques, and applications to arise that often have no counterpart in the simpler case of scalar quantization. Nevertheless, there are interesting similarities with scalar quantization where many of the design techniques used in VQ are natural generalizations of those in the scalar case [19]. In this particular case, the goal is to focus on testing against independence, i.e., between the given bivariate distribution $p_{X^nY^n}$ against the alternative of independence given by:

$$\begin{cases} H_0: & (X^n, Y^n) \sim p_{X^nY^n}(\mathbf{x}, \mathbf{y}) \ , \\ H_1: & (X^n, Y^n) \sim p_{X^n}(\mathbf{x}) \times p_{Y^n}(\mathbf{y}) \ . \end{cases} \tag{3.30}$$

3.4.1 VQ Algorithm for Detection

Finding an optimal VQ consists of finding the set of regions $\boldsymbol{\mathcal{R}}$ and their corresponding codebook defined in Section 3.2 that minimizes a given objective function which quantizes the quantizer's performance. Even without privacy, defining an appropriate *distortion measure* allowing the design of a quantizer adapted to hypothesis testing is an interesting challenge. The MSE criterion is convenient when the aim is to reconstruct the source X^n. However, estimating the source X^n is not our concern, it will be unreasonable to use such criterion when the problem converts into distributed binary decision as we are only interested in testing whether the two remote sources belong to a certain distribution or another. In the coming section, we aim at finding a suitable distortion measure adequate to this kind of problems. Since each vector is encoded separately, the design of the optimal quantizer does not depend on the number of sample vectors encoded, and consequently τ will not be taken into consideration during the design of the quantizer.

Optimal Quantizer and Distortion Criterion

Quantizer design is typically viewed as a task where the standard objective function is to minimize an expected distortion. We aim at predicting the best distortion that suits our problem most.

A key characteristic of any quantizer is its dimension n, a positive integer. In our setting, $\mathbf{x} = x^n$, $\mathbf{y} = y^n$ are the vectors of dimension n available at Alice and Bob respectively and defined over some alphabet $\mathcal{X}^n, \mathcal{Y}^n \subset \mathbb{R}^n$. Hence, $f : \mathcal{X}^n \mapsto \mathcal{J} = \{1, \ldots, M\}$ is the n-dimensional M-points vector memoryless quantizer (VQ), that is, operating independently on successive vectors. Our goal is to find an encoding function f and an acceptance region that minimizes the errors $\beta_n(\alpha_n)$ as $\alpha_n \to 0$, our objective function becomes:

$$\Gamma(f) = \lim_{\alpha_n \to 0} \frac{1}{n} \log \beta_n(\alpha_n, f) \leq -\frac{1}{n} I(Y^n; f(X^n)) \ , \tag{3.31}$$

The last inequality is proven by [40]. The latter mutual information represents the error exponent for

testing against independence. The objective function can be reduced to:

$$\Gamma(f) \leq \frac{1}{n} h(Y^n|f(X^n)) \leq \frac{1}{2n} \log(2\pi e)^n \, \mathbb{E}(\|Y^n - \mathbb{E}(Y^n|f(X^n))\|^2) \, , \tag{3.32}$$

The last inequality is a result of the Maximum Differential Entropy Lemma [37]. $\mathbb{E}(\|Y^n - \mathbb{E}(Y^n|f(X^n))\|^2)$ is the covariance matrix of the error vector of the minimum mean squared error (MMSE) estimate of Y^n given f. Our problem can be seen as equivalent to minimizing the distortion between Y^n available at the detector and the information carried by the quantization outputs. A reproduction decoder can be thus seen as a mapping $g : \mathcal{J} \mapsto \hat{\mathcal{Y}}^n$ whose reproduction alphabet/codebook is $\hat{\mathcal{Y}}^n = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_M\} \subset \mathbb{R}^n$. Of-course, the reproduction decoder doesn't really exist as we are not looking to reproduce any source at Bob but rather looking to make a decision, hence it's only an intermediate function that allows us to compute the optimal quantization regions $\boldsymbol{\mathcal{R}}$ needed to perform the necessary encoding at Alice. As a result, the objective function can be finally written as

$$\begin{aligned} \mathbb{E}(\|Y^n - \hat{Y}^n(f)\|^2) &= \int_{\mathcal{X}^n} \int_{\mathcal{Y}^n} p_{X^n Y^n}(\mathbf{x}, \mathbf{y}) \|\mathbf{y} - g \circ f(\mathbf{x})\|^2 d\mathbf{x} d\mathbf{y} & (3.33) \\ &= \sum_{j=1}^{M} \int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) \mathbb{E}_{Y^n|X^n}[\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2] d\mathbf{x} \, . & (3.34) \end{aligned}$$

$\hat{\mathbf{y}}_j \in \hat{\mathcal{Y}}^n$ is called the representative of the region/cell $\mathcal{R}_j$. $\|\mathbf{v}\| = \sum_i^n v_i^2$ denotes the usual squared Euclidean norm of some vector $\mathbf{v} \in \mathbb{R}^n$ and $Q(\mathbf{x}) = g \circ f(\mathbf{x}) = g(f(\mathbf{x}))$ is the function that maps $\mathcal{X}^n$ to $\hat{\mathcal{Y}}^n$. The problem of finding the optimal quantizer thus becomes finding the set of optimal regions $\boldsymbol{\mathcal{R}}$ and representatives $\hat{\mathcal{Y}}^n$ that minimize the expected distortion $\mathbb{E}(\|Y^n - \hat{Y}^n(f)\|^2)$ under the condition that the regions form a partition of space.

$$(\boldsymbol{\mathcal{R}}_{\mathbf{opt}}, \hat{\mathcal{Y}}^n_{\mathbf{opt}}) = \underset{\boldsymbol{\mathcal{R}}, \hat{\mathcal{Y}}^n}{\arg\min} \sum_{j=1}^{M} \int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) \, \mathbb{E}_{Y^n|X^n}[\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2] d\mathbf{x} \, . \tag{3.35}$$

Partial Solutions

As it is, it is nearly impossible to find the global minimizer of the expected distortion simultaneously with respect to all the regions and representatives. Nevertheless, the optimization problem defined in Equation (3.35) can be divided into two partial solutions. In the first step, we will focus on finding the optimal representatives $\hat{\mathcal{Y}}^n_{\mathbf{opt}} = \{\hat{\mathbf{y}}_1^\star, \ldots, \hat{\mathbf{y}}_M^\star\}$ given the quantization regions $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_1, \ldots, \mathcal{R}_M\}$

$$\hat{\mathcal{Y}}^n_{\mathbf{opt}} = \underset{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_M}{\arg\min} \sum_{j=1}^{M} \int_{\mathcal{Y}^n} \int_{\mathcal{R}_j} p_{X^n Y^n}(\mathbf{x}, \mathbf{y}) \|\mathbf{y} - \hat{\mathbf{y}}_j\|^2 d\mathbf{x} d\mathbf{y} \, , \tag{3.36}$$

The Euclidean distance in this case can be written as $\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2 = (\mathbf{y} - \hat{\mathbf{y}}_j)^\intercal (\mathbf{y} - \hat{\mathbf{y}}_j)$. The function being minimized is the sum of M non-negative functions, each one of them only dependent on one of

the $\hat{\mathbf{y}}_j$. The problem can be decoupled into M independent problems:

$$\hat{\mathbf{y}}_j^\star = \arg\min_{\hat{\mathbf{y}}_j} \int_{\mathcal{Y}^n} \int_{\mathcal{R}_j} p_{X^nY^n}(\mathbf{x},\mathbf{y})(\mathbf{y}-\hat{\mathbf{y}}_j)^\intercal(\mathbf{y}-\hat{\mathbf{y}}_j)d\mathbf{x}d\mathbf{y} \ , \tag{3.37}$$

The minimum is found by computing the gradient with respect to $\hat{\mathbf{y}}_j$ and equating to zero. $\forall j \in \mathcal{J}$, the expression of each representative can be thus written as:

$$\hat{\mathbf{y}}_j^\star = \frac{\int_{\mathcal{Y}^n}\int_{\mathcal{R}_j} p_{X^nY^n}(\mathbf{x},\mathbf{y})\mathbf{y}d\mathbf{x}d\mathbf{y}}{\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x})d\mathbf{x}} \ . \tag{3.38}$$

Similarly, we now try finding the optimal quantization regions $\boldsymbol{\mathcal{R}}_{\mathbf{opt}} = \{\mathcal{R}_1^\star, \ldots, \mathcal{R}_M^\star\}$ given the representatives $\hat{\mathcal{Y}}^n = \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_M\}$

$$\boldsymbol{\mathcal{R}}_{\mathbf{opt}} = \arg\min_{\mathcal{R}_1,\ldots,\mathcal{R}_M} \sum_{j=1}^{M} \int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) \int_{\mathcal{Y}^n} p_{Y^n|X^n}(\mathbf{y}|\mathbf{x})\|\mathbf{y}-\hat{\mathbf{y}}_j\|^2 d\mathbf{y}d\mathbf{x} \ . \tag{3.39}$$

The problem can be also decoupled into M independent problems and each region can be computed separately. Let

$$d(\mathbf{x},\hat{\mathbf{y}}_j) = \int_{\mathcal{Y}^n} p_{Y^n|X^n=\mathbf{x}}(\mathbf{y}|\mathbf{x})\|\mathbf{y}-\hat{\mathbf{y}}_j\|^2 d\mathbf{y} \ , \tag{3.40}$$

$\forall j \in \{1,\ldots,M\}$ the optimal region can be defined as:

$$\mathcal{R}_j^\star = \left\{\mathbf{x} | d(\mathbf{x},\hat{\mathbf{y}}_j) < d(\mathbf{x},\hat{\mathbf{y}}_{j'}), j' \neq j\right\} \ . \tag{3.41}$$

Iterative Algorithm

An iterative algorithm can thus be derived by iterating between the two partial solutions described above and is given by Algorithm 2. In this algorithm, the output of the quantizer must be carrying as much information about Y^n and not X^n and that's what makes the essential difference between this algorithm and the standard k means algorithm.

3.4.2 VQ Algorithm under Privacy Constraints

In the previous section, we have presented a new technique for finding the optimal quantizer for a particular type of detection but the original system had been designed without considering the possible security threats. A potential eavesdropper Eve can extract information about the source, and hence can make a reasonably good decision regarding the initial source based on its receptions. In this setting, the message sent to Bob via the public rate-limited link is perfectly overheard by Eve (the eavesdropper) who may as well have access to a vector Z^n arbitrary correlated with (X^n, Y^n) as described in the problem definition.

Algorithm 2 Vector Quantizer Design for Distributed Binary Decision

1. Initialize $\{\hat{\mathbf{y}}_1^{(0)}, \hat{\mathbf{y}}_2^{(0)}, \ldots, \hat{\mathbf{y}}_M^{(0)}\}$.
2. Loop
   (a) Given the set of representatives $\{\hat{\mathbf{y}}_1^{(k)}, \hat{\mathbf{y}}_2^{(k)}, \ldots, \hat{\mathbf{y}}_M^{(k)}\}$, calculate the set of regions $\boldsymbol{\mathcal{R}}^{(k)} = \{\mathcal{R}_1^{(k)}, \mathcal{R}_2^{(k)}, \ldots, \mathcal{R}_M^{(k)}\}$ such that:
   $$\mathcal{R}_j^{(k)} = \left\{\mathbf{x} | d(\mathbf{x}, \hat{\mathbf{y}}_j^{(k)}) < d(\mathbf{x}, \hat{\mathbf{y}}_{j'}^{(k)}), j' \neq j\right\}, \tag{3.42}$$
   (b) Given the set of regions $\boldsymbol{\mathcal{R}}^{(k)} = \{\mathcal{R}_1^{(k)}, \mathcal{R}_2^{(k)}, \ldots, \mathcal{R}_M^{(k)}\}$, calculate the set of representatives $\{\hat{\mathbf{y}}_1^{(k+1)}, \hat{\mathbf{y}}_2^{(k+1)}, \ldots, \hat{\mathbf{y}}_M^{(k+1)}\}$ such that
   $$\hat{\mathbf{y}}_j^{(k+1)} = \frac{\int_{\mathcal{Y}^n} \int_{R_j^{(k)}} p_{X^n Y^n}(\mathbf{x}, \mathbf{y}) \mathbf{y} d\mathbf{x} d\mathbf{y}}{\int_{R_j^{(k)}} p_{X^n}(\mathbf{x}) d\mathbf{x}} . \tag{3.43}$$
3. Repeat until $\max_j \|\hat{\mathbf{y}}_j^{(k+1)} - \hat{\mathbf{y}}_j^{(k)}\| \leq \delta$.
4. Return $\{\mathcal{R}_1^\star, \mathcal{R}_2^\star, \ldots, \mathcal{R}_M^\star\}$.

Distorton Criterion

The new goal becomes minimizing the errors $\beta_n(\alpha_n)$ as well as the information leakage at Eve. Our new objective function that we're looking forward to minimize thus becomes:

$$\Gamma(f, \lambda) \leq -\frac{1}{n} I(Y^n; f(X^n)) + \lambda \frac{1}{n} I(X^n; f(X^n), Z^n) , \tag{3.44}$$

Γ is now a Lagrangian cost function formed in the manner to incorporate the separate costs of the two costs; the detection distortion and the information leakage. λ is a variable representing a compromise between the level of privacy and detection errors. The information leakage can be decomposed as $\frac{1}{n} I(X^n; f(X^n), Z^n) = \frac{1}{n} H(f|Z^n) + \frac{1}{n} I(X^n; Z^n)$. The term $I(X^n; Z^n)$ is constant and independent of the quantization function f and depends on the correlation between the two sources X^n and Z^n. In this case, we call Δ_{min} the constant value given by $\Delta_{min} = \frac{1}{n} I(X^n; Z^n)$ which represents the minimum information leakage achievable at Eve. Any lower information leakage is not attainable. The objective function thus doesn't have to include such a constant value and can be furthermore reduced to be written as

$$\begin{aligned} \Gamma(f, \lambda) &\leq \frac{1}{n} h(Y^n|f) + \lambda \frac{1}{n} H(f|Z^n) && (3.45) \\ &\leq \frac{1}{2n} \log(2\pi e)^n + \frac{1}{2n} \log \mathbb{E}(\|Y^n - \hat{Y}^n(f)\|^2) + \lambda \frac{1}{n} H(J|Z^n) . && (3.46) \end{aligned}$$

Remark 3.2 *In most cases, the optimal quantizer will have to be learned from a training set of samples. Moreover, calculating the metrics defined in Equations (3.46) is impractical and the need to use the Monte-Carlo technique over the training set is inevitable especially due to integrating over a complicated domain. In Monte Carlo, the final outcome is an approximation of the correct value with respective error bars.*

$$\mathbb{E}(\|Y^n - \hat{Y}^n(f)\|^2) = \sum_{j\in\mathcal{J}} \int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) \int_{\mathcal{Y}^n} p_{Y^n|X^n}(\mathbf{y}|\mathbf{x}) \|\mathbf{y} - \hat{\mathbf{y}}_j\|^2 d\mathbf{x} d\mathbf{y} \tag{3.47}$$

$$\approx \sum_{j\in\mathcal{J}} \frac{1}{N_X} \sum_{\mathbf{x}\in\mathcal{T}_{\mathbf{X}}} \Big(\frac{1}{N_{Y|X}} \sum_{\mathbf{y}\in\mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}} \|\mathbf{y} - \hat{\mathbf{y}}_j\|^2\Big) \mathbb{1}_{\mathcal{R}_j}(\mathbf{x}) \,, \tag{3.48}$$

$$H(J|Z^n) \approx \frac{1}{N_Z} \sum_{\mathbf{z}\in\mathcal{T}_{\mathbf{Z}}} H(J|Z^n = \mathbf{z}) \approx -\frac{1}{N_Z} \sum_{\mathbf{z}\in\mathcal{T}_{\mathbf{Z}}} \sum_{j\in\mathcal{J}} P(j|\mathbf{z}) \log_2 P(j|\mathbf{z}) \tag{3.49}$$

$$\tag{3.50}$$

$\mathbb{1}_{\mathcal{R}_j}(\mathbf{x})$ *is* 1 *if* $\mathbf{x} \in \mathcal{R}_j$ *and* 0 *otherwise is called the indicator function of region* $\mathcal{R}_j$. $\mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ *is a set of* $\mathbf{y}$ *vectors such that* $\mathbf{y} \sim p_{Y^n|X^n=\mathbf{x}}$ *of size* $N_{Y|X}$.

The problem of finding the optimal set of regions $\boldsymbol{\mathcal{R}}$ and representatives $\hat{\mathcal{Y}}^n$ becomes:

$$(\boldsymbol{\mathcal{R}}_{\mathbf{opt}}, \hat{\mathcal{Y}}^n_{\mathbf{opt}}) = \underset{\boldsymbol{\mathcal{R}}, \hat{\mathcal{Y}}^n}{\arg\min} \left\{ \frac{1}{2} \log \mathbb{E}(\|Y^n - \hat{Y}^n\|^2) + \lambda H(J|Z^n) \right\} . \tag{3.51}$$

Iterative Algorithm

In parallel to Section 3.4.1, we essentially repeat all the concepts and derivations adapted to the case of the presence of an eavesdropper. The optimization problem defined in Equation (3.51) can be divided into two partial solutions likewise: finding the optimal representatives $\hat{\mathcal{Y}}^n$ given the quantization regions $\boldsymbol{\mathcal{R}}$ and finding the optimal quantization regions $\boldsymbol{\mathcal{R}}$ given the representatives $\hat{\mathcal{Y}}^n$. As a first step, given the quantization regions $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_1, \ldots, \mathcal{R}_M\}$, we can write:

$$\{\hat{\mathbf{y}}_1^\star, \ldots, \hat{\mathbf{y}}_M^\star\} = \underset{\hat{\mathbf{y}}_1,\ldots,\hat{\mathbf{y}}_M}{\arg\min} \left\{ \frac{1}{2} \log \sum_{j\in\mathcal{J}} \int_{\mathcal{R}_j} \int_{\mathcal{Y}^n} p_{X^nY^n}(\mathbf{x},\mathbf{y}) \|\mathbf{y} - \hat{\mathbf{y}}_j\|^2 d\mathbf{x} d\mathbf{y} + \lambda H(J|Z^n) \right\}$$

$$= \underset{\hat{\mathbf{y}}_1,\ldots,\hat{\mathbf{y}}_M}{\arg\min} \sum_{j\in\mathcal{J}} \int_{\mathcal{R}_j} \int_{\mathcal{Y}^n} p_{X^nY^n}(\mathbf{x},\mathbf{y}) \|\mathbf{y} - \hat{\mathbf{y}}_j\|^2 d\mathbf{x} d\mathbf{y} \,. \tag{3.52}$$

The fact that $H(J|Z^n)$ is independent of all representatives $\hat{\mathcal{Y}}^n$, the solution of the optimal representatives is hence the same as before.

An iterative algorithm for finding the optimal regions $\boldsymbol{\mathcal{R}}$ and the representation points $\hat{\mathcal{Y}}^n$ to meet the above necessary conditions is outlined in Algorithm 3.

Algorithm 3 Vector Quantizer Design for Distributed Binary Decision with Privacy Constraints

1. Given a training set $\Big\{(\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1), (\mathbf{x}_2, \mathbf{y}_2, \mathbf{z}_2), \dots\Big\} \sim p_{X^n Y^n Z^n}$,
2. Initialize $\{\mathcal{R}_1^{(0)}, \mathcal{R}_2^{(0)}, \dots, \mathcal{R}_M^{(0)}\}$ by assigning the the vectors $\mathbf{x}$ to the regions randomly,
3. Loop
   (a) For a given $\boldsymbol{\mathcal{R}}^{(k)}$, calculate the set of representatives $\{\hat{\mathbf{y}}_1^{(k)}, \hat{\mathbf{y}}_2^{(k)}, \dots, \hat{\mathbf{y}}_M^{(k)}\}$ such that

$$\hat{\mathbf{y}}_j^{(k)} = \frac{\int_{\mathcal{Y}^n} \int_{\mathcal{R}_j^{(k)}} p_{X^n Y^n}(\mathbf{x}, \mathbf{y}) \mathbf{y} d\mathbf{x} d\mathbf{y}}{\int_{\mathcal{R}_j^{(k)}} p_{X^n}(\mathbf{x}) d\mathbf{x}} \approx \frac{\sum\limits_{\mathbf{x} \in \mathcal{T}_{\mathbf{X}}} (\frac{1}{N_{Y|X}} \sum\limits_{\mathbf{y} \in \mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}} \mathbf{y}) \mathbb{1}_{\mathcal{R}_j}^{(k)}(\mathbf{x})}{\sum\limits_{\mathbf{x} \in \mathcal{T}_{\mathbf{X}}} \mathbb{1}_{\mathcal{R}_j}^{(k)}(\mathbf{x})} . \tag{3.53}$$

   (b) For a given set of representatives $\{\hat{\mathbf{y}}_1^{(k)}, \hat{\mathbf{y}}_2^{(k)}, \dots, \hat{\mathbf{y}}_M^{(k)}\}$, calculate $\boldsymbol{\mathcal{R}}^{(k+1)} = \{\mathcal{R}_1^{(k+1)}, \mathcal{R}_2^{(k+1)}, \dots, \mathcal{R}_M^{(k+1)}\}$ by looping over all vectors $\mathbf{x} \in \mathcal{T}_{\mathbf{X}}$ and moving each vector over all set of regions $\mathcal{R}^{(k)}$ and then calculate the corresponding $\Gamma(f, \lambda)$ function given by:

$$\begin{aligned} \Gamma(f, \lambda) \;&\approx\; \frac{1}{2} \log \left\{ \sum_{j \in \mathcal{J}} \frac{1}{N_X} \sum_{\mathbf{x} \in \mathcal{T}_{\mathbf{X}}} (\frac{1}{N_{Y|X}} \sum_{\mathbf{y} \in \mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}} \|\mathbf{y} - \hat{\mathbf{y}}_j\|^2) \mathbb{1}_{\mathcal{R}_j}(\mathbf{x}) \right\} \\ &-\; \lambda \sum_{j \in \mathcal{J}} \left\{ \frac{1}{N_Z} \sum_{\mathbf{z} \in \mathcal{T}_{\mathbf{Z}}} \Big(\frac{1}{N_X} \sum_{\mathbf{x} \in \mathcal{T}_{\mathbf{X}}} \mathbb{p}(\mathbf{x}, \mathbf{z}) \mathbb{1}_{\mathcal{R}_j}(\mathbf{x}) \Big) \log_2 \Big(\frac{1}{N_X} \sum_{\mathbf{x} \in \mathcal{T}_{\mathbf{X}}} \mathbb{p}(\mathbf{x}, \mathbf{z}) \mathbb{1}_{\mathcal{R}_j}(\mathbf{x}) \Big) \right\} , \end{aligned} \tag{3.54}$$

   Update $\mathbb{1}_{\mathcal{R}_j}(\mathbf{x})^{(k+1)}$ at each time after assigning $\mathbf{x}$ to the region having the minimum $\Gamma(f, \lambda)$. $\mathbb{p}(\mathbf{x}, \mathbf{z}) = p_{X^n Z^n}(\mathbf{x}, \mathbf{z}) / p_{X^n}(\mathbf{x}) p_{Z^n}(\mathbf{z})$ is called observed to expected ratio.
4. Repeat until $\max\limits_j \|\hat{\mathbf{y}}_j^{(k+1)} - \hat{\mathbf{y}}_j^{(k)}\| \leq \delta$.
5. Return $\{\mathcal{R}_1^\star, \mathcal{R}_2^\star, \dots, \mathcal{R}_M^\star\}$.

3.4.3 Probabilities of Error

In source coding, the term "error" arises in many different ways. In estimation theory, it's usually the distortion measure between the initial source vector and the decoded version. In the context of decision making, the probability of error is considered as being the probability of making a wrong decision and which would have a different value for each type of error. There are also the errors that may occur during transmission, but these will be ignored in our study. The probabilities of error are certainly the most important criteria to evaluate the performance or "goodness" of our test.

No hypothesis test is 100% certain. Because the test is based on probabilities, there is always a possibility of drawing an false decision. In hypothesis testing in statistics, two types of error are distinguished. These two errors are inversely related and determined by the level of significance of the test which is determined by the previously defined threshold T. Consequently, the designer must decide

which error has more severe outcome for the situation before specifying the risks.

In order to compare the performance of our test for the various scenarios studied, we need to plot the graph of β_n versus α_n for different rates, privacy constraints and VQ dimensions. The expressions of α_n and β_n shown in Equation (3.5) cannot be calculated directly, and for that reason we use *Monte Carlo integration* along with the training set. Just as in the case of scalar quantization, we use $\mathbf{p}_\theta(j, \mathbf{y})$ as a short notation for $p_{Y^n}(\mathbf{y})P(J = j|Y = \mathbf{y}, H_\theta), \theta = \{0, 1\}$. Let $N = N_X = N_Y$, for the hypotheses defined in Equation (3.30), we write

$$\mathbf{p}_0(j, \mathbf{y}) \approx \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{T}_\mathbf{X}} p_{Y^n|X^n}(\mathbf{y}|\mathbf{x}) \mathbb{1}_{\mathcal{R}_j}(\mathbf{x}) \,, \tag{3.55}$$

$$\mathbf{p}_1(j, \mathbf{y}) \approx \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{T}_\mathbf{X}} p_{Y^n}(\mathbf{y}) \mathbb{1}_{\mathcal{R}_j}(\mathbf{x}) \,. \tag{3.56}$$

Unlike the optimal quantization function, the acceptance region depends on the number of sample vectors τ. The likelihood ratio for a collection of statistically independent observations factors into a product of individual likelihood ratios. The LLR for τ sample vectors can be simplified using:

$$\Lambda(j^\tau, \mathbf{y}^\tau) = \prod_{t=1}^{\tau} \Lambda(j_t, \mathbf{y}_t) = \prod_{t=1}^{\tau} \frac{\mathbf{p}_1(j_t, \mathbf{y}_t)}{\mathbf{p}_0(j_t, \mathbf{y}_t)} \,. \tag{3.57}$$

For a given quantization function f, the probabilities of error can be computed as follow:

$$\alpha_n(T, \tau) = \sum_{j_1=1}^{M} \cdots \sum_{j_\tau=1}^{M} \int_{\Lambda(j^\tau, \mathbf{y}^\tau) > T} \mathbf{p}_0^\tau(j^\tau, \mathbf{y}^\tau) d\mathbf{y}^\tau \tag{3.58}$$

$$= \sum_{j_1=1}^{M} \cdots \sum_{j_k=1}^{M} \frac{\tau}{N} \sum_{(\mathbf{x}^\tau, \mathbf{y}^\tau) \in \mathcal{T}_{\mathbf{XY}}} \Big(\prod_{t=1}^{\tau} \mathbb{1}_{\mathcal{R}_j}(\mathbf{x}_t) \mathbb{1}[\prod_{t=1}^{\tau} \Lambda(j_t, \mathbf{y}_t) > T] \Big) \,, \tag{3.59}$$

$$\beta_n(T, \tau) = \sum_{j_1=1}^{M} \cdots \sum_{j_\tau=1}^{M} \int_{\Lambda(j^\tau, \mathbf{y}^\tau) < T} \mathbf{p}_1^\tau(j^\tau, \mathbf{y}^\tau) d\mathbf{y}^\tau \tag{3.60}$$

$$= \sum_{j_1=1}^{M} \cdots \sum_{j_t=1}^{M} (\frac{\tau}{N})^2 \sum_{\mathbf{x}^\tau \in \mathcal{T}_\mathbf{X}} \Big(\prod_{t=1}^{\tau} \mathbb{1}_{\mathcal{R}_j}(\mathbf{x}_t) \Big) \sum_{\mathbf{y}^\tau \in \mathcal{T}_\mathbf{Y}} \Big(\mathbb{1}[\prod_{t=1}^{\tau} \Lambda(j_t, \mathbf{y}_t) < T] \Big) \,. \tag{3.61}$$

3.5 Testing Against Independence with Memoryless Gaussian Sources

In this section, we will consider the special case of memoryless Gaussian sources. The HT problem can be thus represented by the following observation model:

$$\begin{cases} H_0 : & (x, y) \sim \mathcal{N}(0, \mathbf{\Sigma}_0) \,, \\ H_1 : & (x, y) \sim \mathcal{N}(0, \mathbf{\Sigma}_1) \,. \end{cases} \tag{3.62}$$

(X^n, Z^n) are also i.i.d. such that $(X, Z) \sim \mathcal{N}(0, \mathbf{\Sigma}_Z)$ with

$$\mathbf{\Sigma}_0 = \begin{pmatrix} \sigma_Y^2 & \rho_Y \sigma_Y^2 \\ \rho_Y \sigma_Y^2 & \sigma_Y^2 \end{pmatrix}, \quad \mathbf{\Sigma}_1 = \begin{pmatrix} \sigma_Y^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}, \quad \mathbf{\Sigma}_Z = \begin{pmatrix} \sigma_Z^2 & \rho_Z \sigma_Z^2 \\ \rho_Z \sigma_Z^2 & \sigma_Z^2 \end{pmatrix}. \tag{3.63}$$

ρ_Z denotes the correlation coefficient between X and the side information Z at Eve. In what follows, we will take $\sigma_Y = \sigma_Z = 1$, $\rho_Y = 0.8$. Note that the marginal distributions of X and Y under both hypotheses follow the standard normal distribution $\mathcal{N}(0, 1)$. In the coming sections, we will study both scalar and vector quantization for this special application of Gaussian sources, plot performance curves under various privacy constraints for both cases and compare.

3.5.1 Scalar Quantization of τ samples

In this section we will consider two cases; the case where there's no side information available at Eve, and the case where some side information Z normally correlated with X is available at Eve. Notice that having no side information Z at Eve, is equivalent to setting $\rho_Z = 0$. In this case, the quantity $I(X; J, Z)$ reduces to $I(X; J)$ when X and Z are independent.

No Side Information Available at Eve

Table 3.1 shows various designs and optimization results under different values of τ and R. For the case of $R = 1$, the decision threshold to be optimized is thus reduced to a single variable a. When $\Delta = 0.25$, the *Bhattacharyya coefficient* is minimized when $a = \pm 1.73$ which will represent the optimal decision threshold in this case. Repeating the same procedure for $\tau = 2, 3, 4$ shows that the best decision threshold remains $a = \pm 1.73$ for any τ. It is important to mention that under no privacy constraints, i.e. $\Delta \geq \Delta^\star$, since the marginal distribution of X under both H_0 and H_1 is symmetric in our example, the optimal quantization of X is thus symmetric as well.

	Rate R	Information Leakage Δ	Decision Threshold(s) $\{a_1, \ldots, a_{M-1}\}$	Bhattacharyya Coefficient d_B
$\tau = 1$	1	0.25	± 1.73	0.980
		0.50	± 1.22	0.958
		0.75	± 0.79	0.937
		$\Delta^\star = 1$	0	0.916
	2	1.489	-0.89, 0.45, 3.34	0.888
		$\Delta^\star = 1.935$	-0.89, 0, 0.89	0.871
$\tau = 2$	1	0.25	± 1.73	0.960
$\tau = 10$		0.25	± 1.73	0.816

Table 3.1: Optimal decision threshold(s) for different information leakages at $R = 1$, $R = 2$. The values of a_0 and a_M are excluded from the table. $\Delta^\star$ is the critical information leakage.

Figure 3.2a represents a trade-off between privacy and performance. Since Δ is the tolerable limit on the performance of Eve, the greater the information leakage, the better the performance of the distributed detection network which is represented by an increase in the *Bhattacharyya distance* D_B. Note that, beyond a certain value of $\Delta^\star$, the maximum D_B gets saturated to the optimal D_B at the detector in the absence of Eve. This saturation level for this example is $\Delta^* = 1, \ \forall \tau \geq 1$. Note also that when $\Delta = 0$, the network achieves perfect privacy. But, this also forces the network to go blind in that $d_B \to 1$, or $D_B \to 0$, a zero actual rate, yielding maximal error at the detector level. On the other extreme, consider a scenario where $\Delta \geq \Delta^\star = 1$ in the case of $R = 1$. This is equivalent to the case where there is no eavesdropper present in the network, and the optimal quantizer for the case of $R = 1$ is obviously given by the decision threshold $a^\star = 0$; the value that minimizes the *Bhattacharyya coefficient* d_B under no privacy constraints.

The performance of our test can be plotted using the expression of the two probabilities of error obtained in Section 3.4.3. Figure 3.3a shows the impact of the information leakage constraint on the performance for $R = 1$ and $R = 2$. Figure 3.3b shows the impact of adjusting the privacy constraint for $\tau = 10$, it is clearly seen that the cost is much higher now with a large value of τ. This clearly makes sense: we are deliberately using "out of tune" quantizers, so that the information leakage remains small. As a consequence, when this allowed leakage is small, the test cannot remain really accurate unless one increases significantly τ.

Side Information Available at Eve

In this section, we study the impact of side information at Eve, with different values of ρ_Z and evaluate the impact of this correlation on the information leakage and detection performance. Therefore, we consider the cases of $\rho_Z = \{0.3, 0.6, 0.9\}$ and compare them to the scenario without side information at Eve. In this case, the minimum information leakage for a given ρ_Z is $\Delta_{\min} = I(X;Z) = -\frac{1}{2}\log_2(1-\rho_Z^2)$.

Obviously, for a correlation $\rho_Z > 0$, perfect privacy is no longer admissible even for a zero rate quantizer. The minimum information leakage that can be achieved becomes $\Delta_{\min}$. Table 3.2 shows various scenarios for different correlations between the side information at Eve and the source X, their corresponding minimum and critical information leakage, and the optimal quantization for different privacy constraints.

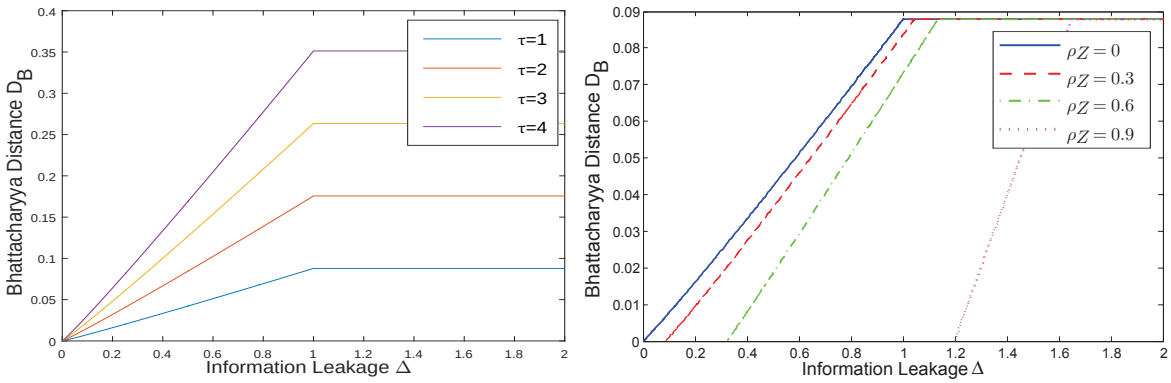
Figure 3.3c shows type II probability of error vs. type I probability of error for different correlations ρ_Z between X and Z for $\tau = 10$ and different information leakages Δ. Note that for the same amount of admissible information leakage and a higher correlation ρ_Z, the errors at Bob increase as the allowed entropy rate decreases.

For any finite $\Delta > 0$, we numerically investigate the trade-off between privacy and performance for various values of ρ_Z. This trade-off is captured by Figure 3.2b, where the maximum D_B at the level

of the detector in the presence of a constrained Eve increases with increasing Δ. Note that, beyond a certain value of $\Delta^\star$, the maximum D_B gets saturated to the optimal D_B at the detector in the absence of Eve. The critical information leakage is no longer $\Delta^\star = 1$ for the case of having a correlated side information at the eavesdropper. In fact, the critical leakage increases with higher ρ_Z, and information leakage for a given quantization scheme becomes higher. The saturation level of D_B for this example occurs at $\Delta^* = 1$ for $\rho_Z = 1$ and $\Delta^\star = 1.13$ for $\rho_Z = 0.6$.

Correlation ρ_Z	Minimum Leakage Δ_{min}	Critical Leakage $\Delta^\star$	Information Leakage Δ	Decision Threshold(s)	Bhattacharyya Coefficient d_B
0.3	0.068	1.025	0.25	± 1.89	0.9848
			0.50	± 1.31	0.9626
			1	± 0.24	0.9181
0.6	0.322	1.13	0.50	± 1.79	0.9818
			0.75	± 1.16	0.9554
			1	± 0.61	0.9293
0.9	1.198	1.648	1.25	± 2.09	0.9897
			1.50	± 0.9	0.9425

Table 3.2: Optimal decision thresholds for different information leakages and correlations for $R = 1$.



(a) Impact of τ

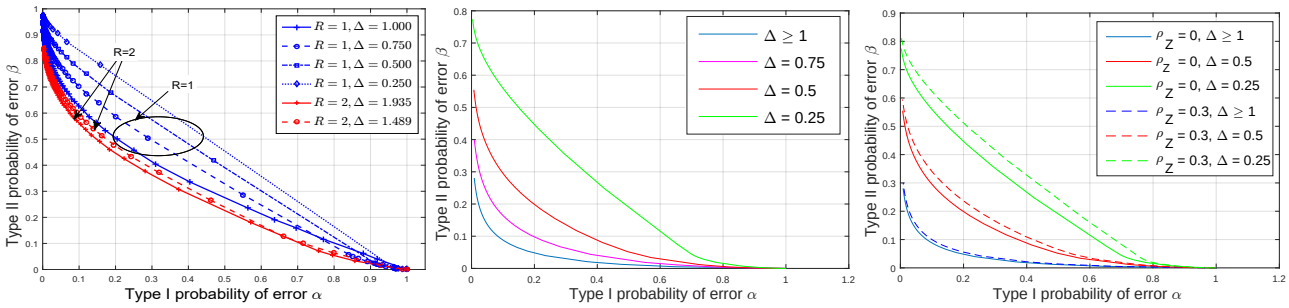
(b) Impact of ρ_Z

Figure 3.2: In (a), we show *Bhattacharyya distance* versus information leakage for different values of τ at rate $R = 1$. In (b), we consider different values of ρ_Z.

3.5.2 Vector Quantization

In this particular example, X^n, Y^n, Z^n denote a sequence of i.i.d. random variables from (X, Y, Z). For the Gaussian sources given in Equation (3.62), we can then write the conditional probability of $Y_i|X_i = x_i \ , i = \{1, \ldots, n\}$ as

$$Y_i|X_i = x_i \sim \mathcal{N}(\rho_Y x_i, (1 - \rho_Y^2)) \ . \tag{3.64}$$



(a) Impact of R (b) Impact of Δ (c) Impact of ρ_Z

Figure 3.3: Type II vs type I probability of error. In (a), different combinations of rate-leakage scenarios are considered at $\tau = 1$. (b) shows the impact of the allowed information leakage on performance when $\tau = 10$. (c) shows the impact of the side information at Eve on performance-privacy trade-off.

The conditional mean of the distance $\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2$ with respect to $X^n = \mathbf{x}$ is then $\mathbb{E}_{Y^n|X^n}[\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2] = \|\rho_Y \mathbf{x} - \hat{\mathbf{y}}_j\|^2 + n(1 - \rho_Y^2)$. Then the distortion measure reduces to a much smoother form for this. Equation (3.33) can be rephrased in much similar way:

$$\mathbb{E}(\|Y^n - \hat{Y}^n(f)\|^2) = \sum_{j \in \mathcal{J}} \int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) \|\rho_Y \mathbf{x} - \hat{\mathbf{y}}_j\|^2 d\mathbf{x} + n(1 - \rho_Y^2) \, . \tag{3.65}$$

In the absence of any privacy constraint, the partial solutions can be simplified. Given the quantization regions $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_1, \ldots, \mathcal{R}_M\}, \forall j \in \mathcal{J}$ the optimal representatives can be calculated using:

$$\hat{\mathbf{y}}_j^\star = \frac{\rho_Y \int_{\mathcal{R}_j} \mathbf{x} p_{X^n}(\mathbf{x}) d\mathbf{x}}{\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) d\mathbf{x}} \, , \tag{3.66}$$

And given the representatives $\hat{\mathcal{Y}}^n = \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_M\}, \forall j \in \mathcal{J}$ the optimal regions can be defined as:

$$\mathcal{R}_j^\star = \left\{ \mathbf{x} | \ \|\rho_Y \mathbf{x} - \hat{\mathbf{y}}_j\|^2 < \|\rho_Y \mathbf{x} - \hat{\mathbf{y}}_{j'}\|^2, j' \neq j \right\} . \tag{3.67}$$

Results show that regions remain unchanged with different values of ρ_Y whereas the values of the centroids given by $\hat{\mathcal{Y}}^n$ are scaled by ρ_Y. This can be explained theoretically. Let $\hat{\mathcal{X}}^n = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_M\}$ be the set of centroids/representatives of our quantization scheme when $\rho_Y = 1$, then $\forall j \in \{1, \ldots, M\}$, $\hat{\mathbf{x}}_j$ is given using the following expression

$$\hat{\mathbf{x}}_j^\star = \frac{\int_{R_j} \mathbf{x} p_{X^n}(\mathbf{x}) d\mathbf{x}}{\int_{R_j} p_{X^n}(\mathbf{x}) d\mathbf{x}} \, , \tag{3.68}$$

The set $\hat{\mathcal{X}}^n$ represents the quantization centroids given by Lloyd-max. The set of representatives $\hat{\mathcal{Y}}^n$ is related to $\hat{\mathcal{X}}^n$ by $\hat{\mathbf{y}}_j^\star = \rho_Y \hat{\mathbf{x}}_j^\star \ \ \forall j \in \{1, \ldots, M\}$. On the other hand, the optimal regions defined in

Equation (3.67) become

$$\mathcal{R}_j^\star = \left\{\mathbf{x}|\ \|\mathbf{x}-\frac{\hat{\mathbf{y}}_j}{\rho_Y}\|^2 < \|\mathbf{x}-\frac{\hat{\mathbf{y}}_{j'}}{\rho_Y}\|^2, j'\neq j\right\} \tag{3.69}$$

$$= \left\{\mathbf{x}|\ \|\mathbf{x}-\hat{\mathbf{x}}_j\|^2 < \|\mathbf{x}-\hat{\mathbf{x}}_{j'}\|^2, j'\neq j\right\}. \tag{3.70}$$

This shows us that under no privacy constraint, the quantization regions do not change with ρ_Y for this particular example of standard Gaussian sources. Figures 3.4a to 3.4d show the optimal quantization regions under different scenarios.

Figure 3.5a shows that increasing the quantization dimension n gives better performance even for smaller rates. This demonstrates the advantage of vector quantization on the performance of testing. On the other hand, even without imposing any privacy constraint, decreasing the rate R is equivalent to increasing the level of security of the system. We have seen before that the information leakage $\frac{1}{n}H(J^n)$ is nothing but the actual coding rate in the absence of side information at the eavesdropper, hence, a rate constraint is also a privacy constraint. The figure shows that a higher VQ dimension n leads to better performance and probably better privacy.

Figure 3.5b demonstrates the impact of varying the quantization dimension n but keeping constant $n\times\tau$ for a given rate of $R=0.5$. You can see that for a higher n, the performance is slightly enhanced although typically the same number of scalar samples is available even if each scalar input is i.i.d. with the other as in our case. This shows that given a specific sequence of scalar samples available, the most efficient performance-wise technique is to increase n as much as possible.

To elaborate, Table 3.3 shows the maximum value of the type II probability of error when $\varepsilon=0.01$ and $\alpha_n\leq\varepsilon$ for the studied senarios. The lower $\beta_n(\alpha_n)$ demonstrates the better detection performance and the importance of using a higher VQ dimension.

R	$n\times\tau$	n	τ	Type II error β_n
		2	8	0.3603
0.5	16	4	4	0.3230
		8	2	0.2859

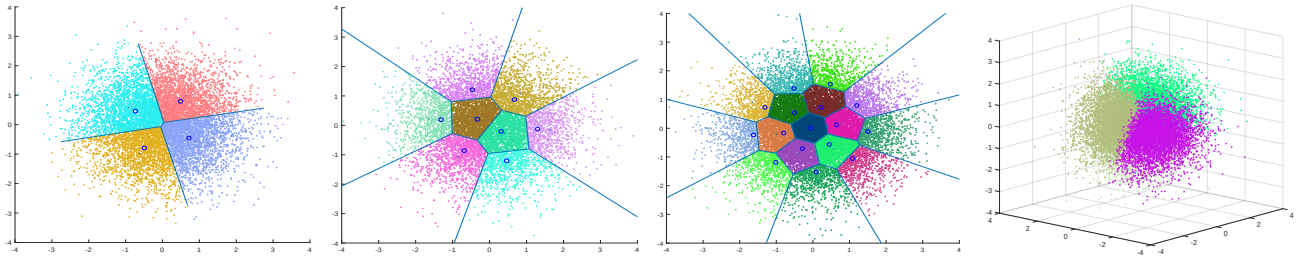
Table 3.3: Minimum β_n for type I error $\alpha_n\leq 0.01$ as n varies.

In order to study VQ with privacy, we need to examine how quantization cells change with different values of λ. For this objective, Figures 3.4e to 3.4h consider the case of VQ of dimension $n=2$, a maximum quantization cells of $M=4$, and a correlation of $\rho_Y=0.8$ under p_0, whereas Figures 3.4i and 3.4j and Figures 3.4k and 3.4l show VQ for $M=8$ and $M=32$ respectively. All these figures demonstrate how quantization cells change when decreasing the allowed level of information leakage gradually by increasing the value of λ. It is noticed that the quantizer might not be using all number of available quantization cells in order to ensure a higher privacy level.

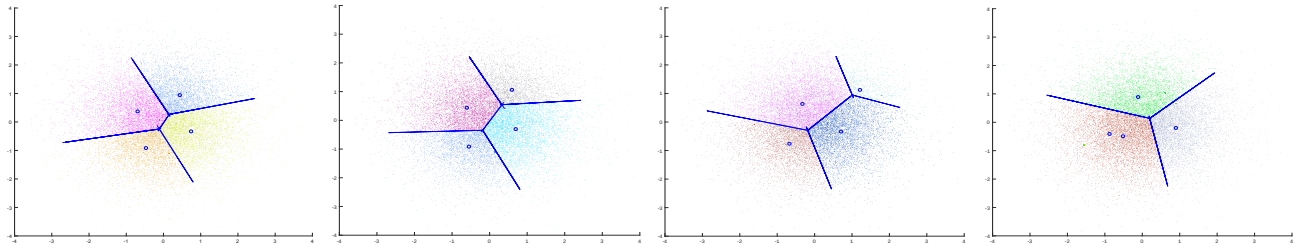
In order to demonstrate the importance of vector quantization over scalar quantization, let's consider a fixed maximum rate R and plot the performance curve β_n vs. α_n for different values of quantization

dimension. Let the fixed rate be $R = 1$ bits/dimension. Figure 3.5c shows how detection performance changes with VQ. A significant difference is not present, but an advantage slightly starts to appear for high values of n.

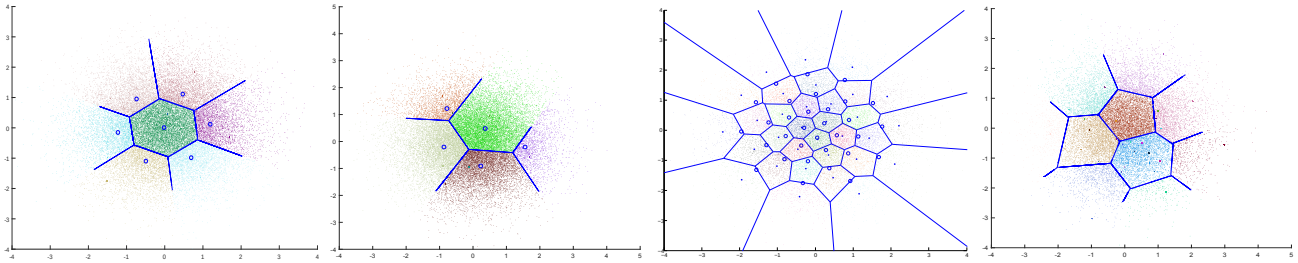
Finally, in Figure 3.5d, we compare vector quantization to scalar quantization when $R = 1$ and when the maximum admissible information leakage is Δ. For VQ, we consider the case of sending 1 vector of dimension $n = 4$, while for SQ, we send $\tau = 4$ samples of the scalar x. We then compare the performance for $\Delta = 0.85$ and $\Delta = 0.4$. The corresponding figure shows the important advantage of VQ over SQ in terms of performance especially under privacy constraints.



(a) $n = 2, M = 4$ (b) $n = 2, M = 8$ (c) $n = 2, M = 16$ (d) $n = 3, M = 4$

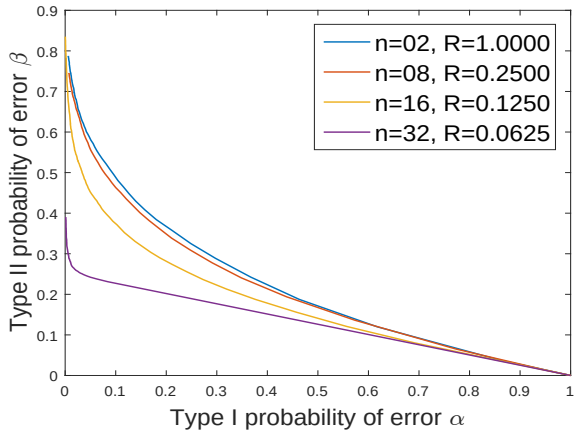


(e) $\lambda = 0.06$ (f) $\lambda = 0.12$ (g) $\lambda = 0.18$ (h) $\lambda = 0.185$

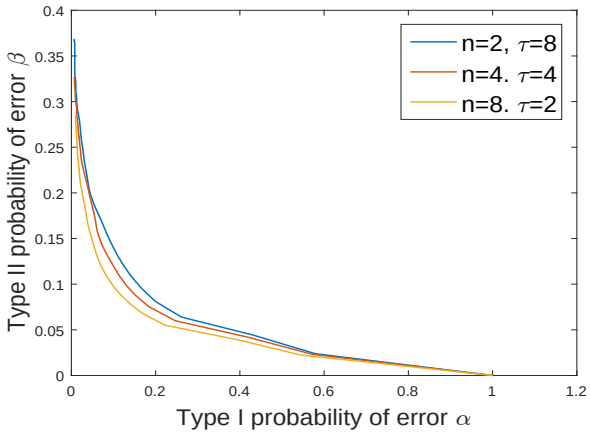


(i) $\lambda = 0.08$ (j) $\lambda = 0.14$ (k) $\lambda = 0$ (l) $\lambda = 0.1$

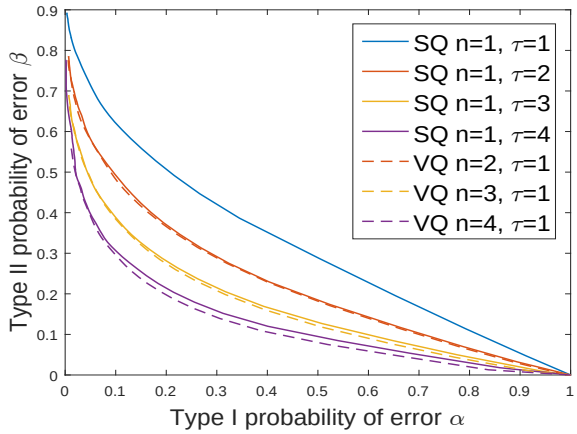
Figure 3.4: Results for VQ optimal regions. First row shows optimal quantization regions under no privacy constraints when $\rho_Y = 0.8$. Second row shows optimal quantization regions for $n = 2, M = 4$ when $\rho_Y = 0.8$ for various values of λ. In the third row, (i) and (j) show the optimal quantization regions for $n = 2, M = 8, \rho_Y = 0.8$ for various values of λ while (k) and (l) show that of $M = 32$. Figures represent 2-dimensional or 3-dimensional space. Codewords are marked with circles, and the optimal regions are separated with boundary lines. The training vectors are assigned to regions using different colors.



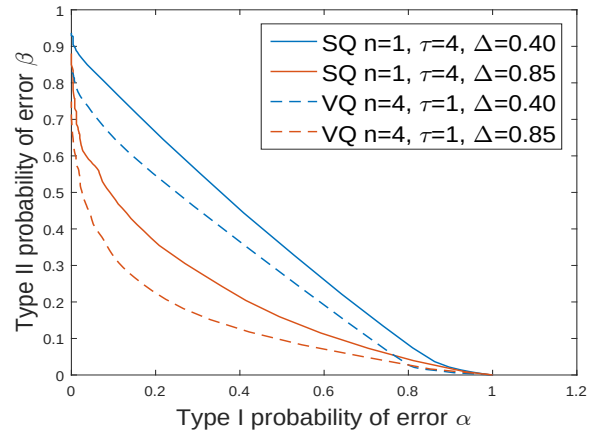
(a) $R < 1$



(b) $n \times \tau =$ constant



(c) Without Privacy Constraints



(d) With Privacy Constraints

Figure 3.5: Performance-privacy trade-offs for different VQ scenarios. Row 1 shows detection performance under no privacy constraints. (a) plots errors at small rates R whereas (b) compares errors when $R = 0.5$ for the same $n \times \tau$ and a variable VQ dimension. Row 2 compares VQ and SQ at rate $R = 1$ when $n \times \tau$ is also the same under both methods. (c) compares the 2 methods under no privacy constraints while (d) compares them when a maximum information leakage value is given.

3.6 Pratical Application: Smart Meter Fault Detection under Privacy Contraints

Communication plays a key component of the smart grid. Signal processing, distributed estimation and detection, machine learning, etc., are important tools to many significant features of the smart grid, such as demand response, distribution automation, auto-recovery, improved reliability and privacy, etc.

A smart meter is one of the most successful key parts of the smart grid. Smart meters transmit up-to-one-minute consumption data, remarkably reducing the need for visits from a meter reader. However, over the past few years, smart meters have created a growing array of serious problems needing immediate actions including meter failure. As any other devices, meters could face failure problems causing them to communicate false and unreliable data to the central unit. A smart meter false reading could be due to incorrect installation of the device or due to a transmission failure, etc. On the other hand, data sent by the smart meter can be manipulated by owners before being sent to the central unit in order to cut power bills which may lead to huge financial losses or probably disrupt the whole power grid. In such cases, the term Fraud detection can be used instead of Fault detection.

Another major concern related to smart meters is the privacy invasion. “The smart electric grid may be just a little too smart”. A smart meter can gather much more data than just how much electricity a home is consuming. It can tell how many people are living in the house, when they get up, when they go to sleep and when they are or not home etc. Each appliance; the refrigerator, microwave, toaster, washer/dryer, etc.; has its own energy fingerprint or "appliance load signature" hidden in the aggregate data consumption. Anyone who gets hold of this data is able to reveal what type of appliances in use and how often they are being used. Many consumers are worried that such smart devices would make them vulnerable to thefts, annoying marketers, or police investigations. In what follows, we will propose a model that enables us to detect faults at the level of the smart meter reading taking into consideration privacy issues.

Consider a neighborhood area network made up of some number of residential users where each user is equipped with a smart meter that records real-time data about electricity use. The network of electricity users perform data aggregation via other users, for example, and report the real-time aggregated data to the Central Unit (CU) via the local gateway/data aggregation unit (DAU) as shown in Figure 3.6. On receiving the reports from the DAU, the CU can estimate the average real-time electricity use of the area.

The average aggregate data of the residential area will be referred to as Y^n. On the other hand, another data X^n from a separate house in the neighbourhood area network is encoded with a rate R and sent via a public rate-limited channel to the central unit that is looking to detect any fault or fraud at the smart meter of that particular house. Fault Detection can then be done by testing whether that two sets of data X^n and Y^n are correlated or not. A non-correlation could signify that the data readings of the smart meter are erroneous or manipulated. Therefore, we would like to make a decision between one

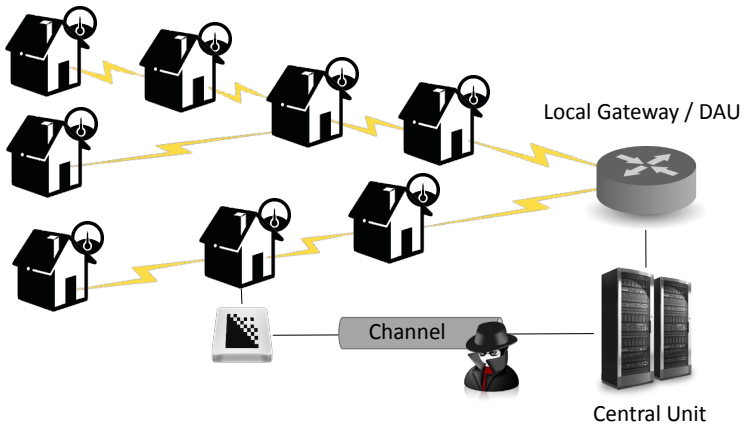


Figure 3.6: An illustration of the studied model.

of the following two hypotheses

$$\begin{cases} H_0: & \text{Smart meter's sending flawless data ,} \\ H_1: & \text{Smart meter's sending flawed data .} \end{cases} \tag{3.71}$$

In what follows, we will ignore how data aggregation is performed and rather focus on how the detection at the level of the central unit takes place, and what would be the optimal encoder f needed to perform the detection such that the errors are minimized and information leakage about X^n is bounded. We will assume there's no side information available at Eve.

Our model can be hence seen exactly as the problem described in Figure 3.1. Similar to the problem of testing against independence defined before, the goal is to make a decision between two possibilities of the probability law of (X^n, Y^n) as it can only be one out of two hypotheses:

$$\begin{cases} H_0: & (X^n, Y^n) \sim p_{X^nY^n}(\mathbf{x}, \mathbf{y}) \,, \\ H_1: & (X^n, Y^n) \sim \tilde{p}_{X^n}(\mathbf{x}) \times p_{Y^n}(\mathbf{y}) \,. \end{cases} \tag{3.72}$$

$\tilde{p}_{X^n}$ is any arbitrary distribution, and p_{Y^n} is the marginal distribution of Y^n under H_0. In other words, the problem is seen as follow, H_0 represents the case where the smart meter is functioning properly and the joint distribution of X^n and Y^n at the normal state would be $p_{X^nY^n}$ which could be constructed using historical data. Under H_1, X^n could be coming from any unknown distribution and that's why the distribution of X^n under H_1 is an arbitrary distribution. H_1 represents the hypothesis where a fault should be detected at the level of the smart meter. For this particular scenario, our optimization function therefore becomes:

$$\Gamma(f, \lambda) = \lim_{\alpha \to 0} \frac{1}{n} \log \beta_n(\alpha_n, f) + \lambda \frac{1}{n} I(X^n; f(X^n)) \tag{3.73}$$

$$\leq -\frac{1}{n} D_{KL}(p_{f(X^n)Y^n} || \tilde{p}_{f(X^n)} p_{Y^n}) + \lambda \frac{1}{n} I(X^n; f(X^n)) \,, \tag{3.74}$$

The last inequality is also proven by [40]. However, the latter expression can be decomposed as follow:

$$D_{KL}(p_{f(X^n)Y^n}||\tilde{p}_{f(X^n)}p_{Y^n}) = D_{KL}(p_{f(X^n)Y^n}||p_{f(X^n)}p_{Y^n}) + D_{KL}(p_{f(X^n)}||\tilde{p}_{f(X^n)}) \quad (3.75)$$

$$= I(f(X^n);Y^n) + D(p_{f(X^n)}||\tilde{p}_{f(X^n)}) \,, \quad (3.76)$$

p_{X^n} being the marginal distribution of X^n under H_0. Since $D(p_{f(X^n)}||\tilde{p}_{f(X^n)}) \geq 0$, the objective function is thus upper bounded by the distortion provided in Equation (3.51) when $Z = \emptyset$ whose solution would be feasible to this particular scenario.

3.6.1 Training Set

A free data set containing detailed power consumption from six different houses is provided by Reference Energy Disaggregation Data Set (REDD) [103]. The data contains power consumption for the total electricity usage as well as for each separate appliance in the house for a number of real houses over several months' time. The data is recorded at a frequency of once a second for the aggregate load and once every three seconds for the appliances.

From now on, we will use the total electricity consumption in one of the six houses to represent our source X^n, while the source Y^n will be the mean of the total consumption of all available six houses. The data will be split into a training set and a test set. In the first part, the optimal quantizer will have to be learned from these data samples, while the data extracted from the second part will be used to evaluate our quantizer performance and system security.

The quantization design given in Algorithm 3 requires a training sequence $\mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \;\; \forall \; \mathbf{x}$. Consequently, we would like to generate sample vectors of $Y^n \sim p_{Y^n|X^n=\mathbf{x}}$ to allow the quantization design to be possible. When direct sampling is difficult, a common algorithm for obtaining a sequence of observation samples from a multivariate probability distribution is known as Gibb's sampling. However, This algorithm is unsuccessful when the multivariate distribution contains islands of high-probability states, with no paths between them. In order to avoid such islands present in the empirical distribution, an appropriate distribution can be fitted closely to the observed frequency of the data. A distribution giving a close fit is supposed to lead to good predictions.

Probability Distribution Fitting

There is generally no unique distribution type suitable for modeling household electricity use. However, there are benefits in not using for example the normal distribution, since it extends to negative electric consumption use values whereas for example the Weibull distribution and Log-Normal distribution do not. Upon inspection the histograms of the data sets, several distributions can be proposed, like Log-normal, Weibull, Gamma, etc. which appear to capture the essential random features of the data sets. [104] presents models of electricity use in an average household based on fit to Log-normal and Weibull probability distributions.

To estimate the goodness of fit between original data sets and proposed distributions, several tests

exist. The Kolmogorov-Smirnov statistic allows us to measure the maximum difference in value between the empirical cumulative distribution and the proposed to be fitted distribution. Although more specialised tests have been developed, the K-S test remains widely used. Comparing the K-S statistic for several distribution densities allows us to choose those who fit our model the most. For a given cumulative distribution function $F(\mathbf{x})$,

$$D_{K-S} = \sup_{\mathbf{x}} |\hat{F}(\mathbf{x}) - F(\mathbf{x})| \,, \tag{3.77}$$

$\hat{F}(\mathbf{x})$ is the empirical cumulative distribution function. Table 4.3 shows different values of the K-S statistic for different distributions for both the univariate and the multivariate case when $n = 1$. The parameters of each distribution are calculated using the maximum likelihood estimator. After inspecting the table, we choose to model the consumption in each household using the log-normal distribution.

	Univariate Model		Multivariate Model
	X	Y	(X, Y)
Normal	0.1799	0.1508	0.2742
Log Normal	0.0708	0.0414	0.1009
Weibull	0.1199	0.1011	
Logistic	0.1782	0.1260	
loglogistic	0.0804	0.0407	
Gamma	0.1164	0.0709	
tlocationscale	0.2030	0.1200	
T copula			0.0942

Table 3.4: K-S statistic for different distribution laws for the univariate case.

Maximum Likelihood Estimator for Log-Normal Distibution

In probability theory, a log-normal (or lognormal) distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. A random variable which is log-normally distributed takes only positive real values. On the other hand, the multivariate log-normal distribution can take several forms. One form of the multivariate log-normal can be written as

$$\ln \mathcal{N}(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|} \prod_{i=1}^{n} x_i} \exp(-\frac{1}{2}(\ln \mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\ln \mathbf{x} - \boldsymbol{\mu})) \,. \tag{3.78}$$

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be estimated using the maximum likelihood estimation (MLE) while only knowing the consumption in a certain period of time. In general, for a fixed set of data and underlying statistical model, the method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function. For determining the maximum likelihood estimators of the log-normal distribution parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we can use the same procedure as for the

normal distribution, the estimated $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are thus,

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{\mathbf{x}} \ln \mathbf{x} \,, \tag{3.79}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\mathbf{x}} (\ln \mathbf{x} - \widehat{\boldsymbol{\mu}})(\ln \mathbf{x} - \widehat{\boldsymbol{\mu}})^{\intercal} \,. \tag{3.80}$$

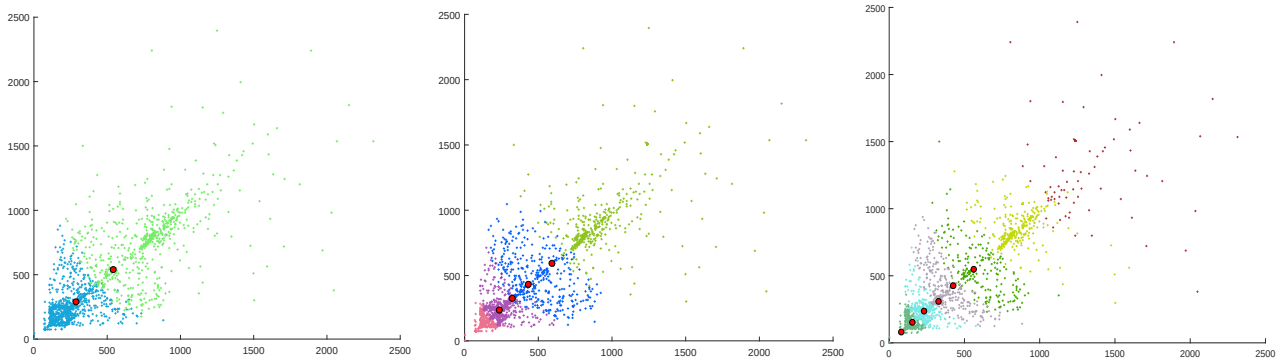
After estimating the parameters, we can now plot and compare the initial empirical distribution and the proposed log-normal. Figures 3.8c to 3.8e show both the empirical distribution and the estimated lognormal when $n = 1$. Finally, To generate more sample vectors, we will be using the estimated log-normal joint distribution. Simulating log-normal random outcomes is nothing more than exponentiating simulated normal random multivariates of the same parameters $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$.

3.6.2 Results

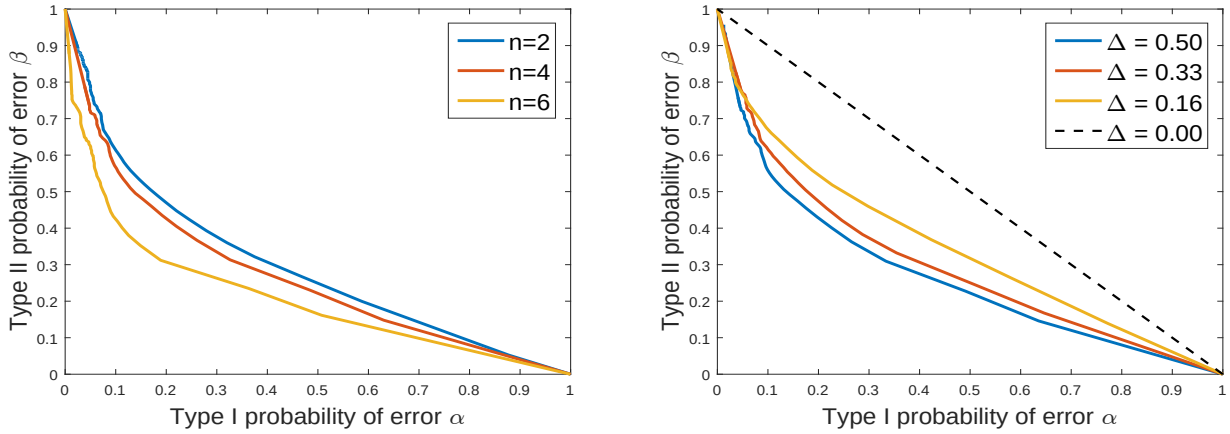
Figures 3.7a to 3.7c show how samples of the testing set are distributed over the optimal quantization regions when $n = 2$ for different values of the rate R.

In order to compare the performance of the central unit performing the fault detection, we plot the type II vs. type I probability of error for various scenarios. In Figure 3.7d, we fix $R = 0.5$ and we vary the dimension of n, it can be seen how the performance gets better and better with the increase of n. Of course in reality, we can increase n much more as we won't be having any computation limitations and hence obtain a much better performance. On the other hand, Figure 3.7e shows the variation of the probabilities of error for a fixed $n = 4$ and $R = 0.5$ for different values of the privacy constraint Δ. An increase in Δ shows an increase in privacy but the performance is affected. This change in performance is represented by an increase in errors. Note also that when $\Delta \to 0$, the network achieves perfect privacy. But, this also forces the network to achieve maximum detection errors.

It's noted that the performance in this case isn't quite as good as for the case of memoryless Gaussian sources discussed earlier, even though in this case there is memory in the sources. This is basically because the correlation between X^n and Y^n is much lower now. In the previous section, we have considered a correlation of 0.8, whereas the correlation here is said to be something between 0.3 and 0.4. Therefore, unless the observations are highly correlated, different solutions are recommended in order to improve the performance. The designer can eventually increase the rate, the number of vector samples or the VQ dimension in order to achieve higher detection performance. The role of the quantization rate is more important: passing from 1 bit/dimension to 2 bits/dimension provides significant performance improvement. However, this could be useful in systems not taking into account security. In our scenario, increasing the rate would enormously increase the information leakage at the level of the eavesdropper as a result of the direct relation between the two measures.



(a) $R = 0.5$ (b) $R = 1$ (c) $R = 1.3$



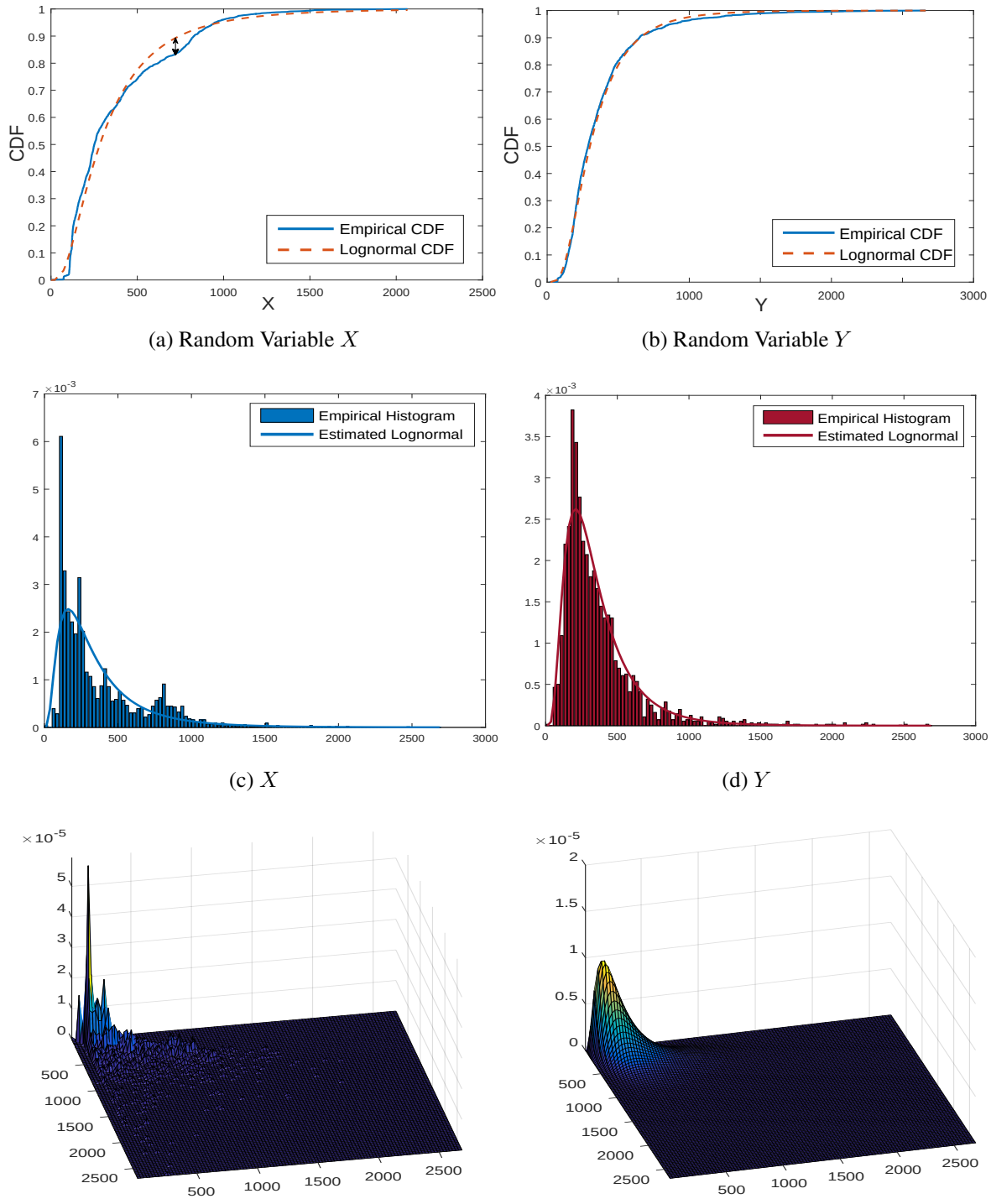
(d) Impact of n (e) Impact of Δ

Figure 3.7: Row 1: optimal quantization regions for $n = 2$ (x and y axis are amplitudes 1 and 2 of the vector $\mathbf{x}$ respectively). Row 2: type II vs type I probabilities of error. In (d), we study the impact of the VQ dimension n when $R = 0.5$ and $\Delta \geq R$. In (e), we study the impact of the allowed level of information leakage on performance when $n = 4$ and $R = 0.5$. In all of these figures, $\tau = 5$.

3.7 Conclusion

In this chapter, we derived an algorithm for binary detection with quantized samples between a remote and a directly available source in the presence of privacy constraints for both scalar and vector quantization. In the VQ part, we considered only the special case of testing against independence. In SQ, our algorithm uses the *Bhattacharyya distance* as a distortion measure and an optimization criterion. While in VQ, we were able to use the asymptotic error exponent as it reduces to the mutual information between the encoder output and the remote source, an iterative algorithm that calculates the quantization regions and the corresponding representatives was then derived.

The algorithm was applied to the special case of testing against independence with Gaussian sources. Numerical results show that the greater the information leakage we can tolerate, the better the detection performance, this remains true until certain value of the information leakage known by the critical leakage, beyond which the *Distortion Measure* gets saturated to the optimal distance in the absence of Eve. In presence of correlated side information at Eve, perfect privacy is not possible. Globally,



(a) Random Variable X

(b) Random Variable Y

(c) X

(d) Y

Figure 3.8: Results for probability distribution fitting with lognormal. First row: illustration of the Kolmogorov-Smirnov statistic. Red line is the estimated lognormal CDF, blue line is ECDF, and the black arrow is the K-S statistic. Second row: empirical univariate distribution vs. fitted log-normal distribution of X and Y for $n = 1$. Third row: joint empirical distribution vs. fitted joint log-normal distribution of (X, Y) for $n = 1$.

increasing the sample vectors size has no impact on the privacy level but does improve the detection performance. For a large sample size, the cost of privacy becomes the central bottleneck in terms of test performance. Finally, results show that VQ doesn't have an important advantage over SQ when

no privacy constraints are present, but appears to be much more efficient in the presence of privacy constraints especially when n gets larger and larger.

Passing from scalar quantization to vector quantization, may not appear to be conceptually hard at the beginning; apparently, the significant modification of the SQ procedure would be replacing the scalar sources with vectors; but turned out to be more complex both analytically and numerically.

The chapter also addressed an important application in the context of smart grids where HT can be used by collectors (the statistician) to test the integrity of the smart devices present at the houses while keeping private the meters measurements from the collectors.

One of the most important aspect of this application is the adaptation to unknown or varying observation statistics. When no observation model is available, we extended the design algorithm to use with training sequences developed under each hypothesis. Of course, during the design procedure we needed distribution fitting in order to generate the sample vectors of $\mathcal{T}_{\mathbf{Y}|\mathbf{X}}$. Nevertheless, if data were large enough, we could have used Gibb's sampling directly from the empirical distribution.

Increasing VQ dimension provides significant performance improvement. However, the role of the quantization rate is more important: passing from a small rate to a high rate provides much more improvement. Increasing the rate could be interesting in systems not taking security into account but is highly undesirable in the case where security plays an essential role considering the fact that information leakage is directly related to the rate.

In our further studies, we will extend the VQ scheme of testing against independence to the case of general hypothesis testing. We will also address decentralized communication network when both sources are remotely located with respect to the detector. The later scenario should be interesting especially for the smart meter application studied in this chapter. The decentralized system could also be extended to include N different sources requiring the design of N corresponding quantizers.

PART II

Secure Lossy Estimation

CHAPTER 4

QUANTIZATION LEARNING FOR A UTILITY-PRIVACY FRAMEWORK

Abstract — *Consider a source coding problem in presence of two dependent with memory sources (X, Y), for which only X is available at the encoder (referred to Alice). We first study the design of vector quantization for the situation where one of the source outputs, i.e., X, must be transmitted to the receiver (referred to Bob) within a prescribed distortion tolerance as in ordinary source coding. On the other hand, the other source, i.e., Y, has to be kept as secret as possible from the receiver or wiretappers. We next consider the opposite case where Y represents a relevant utility sequence to be reconstructed at Bob while trying to keep information about X secret from an eventual eavesdropper. Numerical results for memoryless Gaussian sources demonstrate the performance of the proposed quantization methods. A practical application involving electric consumption data measured from real houses is finally investigated.*

Keywords — *Lossy source coding; vector quantization; distortion measure; algorithm; privacy; information leakage.*

4.1 Introduction

It is well-known that real sources cannot be transmitted without distortion over a finite capacity channel, and therefore must be sampled and compressed before transmission. Even if the sampled measurements take values in a discrete alphabet, there may be a need to distort the data in some way to guarantee a certain level of privacy. However, such a distortion should not affect the system's reliability.

Traditionally, the majority of studies focused only on problems of reliable communication. Recently, extensive research is concentrating on secure communication, i.e., when the goal is to design a communication system that is both reliable and secure. In earlier researches, the focus in secure systems was on cryptography and security was only taken into consideration in the application layer of the OSI model. The information-theoretic study of physical-layer security was pioneered by [44] for the case of the Wiretap channel. Secrecy at the eavesdropper is measured through the conditional uncertainty of the source given the message taking into account the different characteristics possessed by the eavesdropper and the legitamate reciever. Wyner has shown that perfect security is attainable as long as the channel of the eavesdropper is a degraded version of the legitimate user's one. Later, [46] extended this result to the general broadcast channels. Several extensions to different multi-user settings can be

found in [42]. It is important to emphasize that in all of the above mentioned works, the eavesdropper is assumed to observe a noisy (degraded) version of that of the legitimate receiver via a wiretap or a broadcast channel. On the contrary, this chapter assumes that communication is performed over a public channel where the eavesdropper has access to the same information as that of the legitimate receiver.

Source coding with security constraints has been first studied in [105, 106]. A source is composed of two subsets, a decoder must guarantee a distortion that is not larger than D_1 for the utility subset using a certain distortion measure and at the same time cannot be smaller than D_2 for the secret subset using another distortion in conjunction with any other decoder. The authors in [107] formulate the problem that involves minimization of distortion measure by the encoder and the decoder subject to a entropy based privacy constraint. They characterize the achievable distortion-privacy region with or without compression at the transmitter. They also study the impact of the presence of an average power constrained Gaussian communication channel on the privacy-distortion trade-off. More recently, source coding with side information subject to a privacy constraint was also studied. As we said before, security of the source may depend on the different characteristics (side information) which are available at the legitimate receiver and the eavesdropper. Secure lossless distributed source coding was mainly studied by [108, 109, 110], whereas the lossy case was considered by [111, 78].

In our previous works [100, 101], we investigated secure quantization schemes for the problem of detection at the receiver, i.e., when the receiver aims at making decisions on the source based on the encoded data. In this chapter, we investigate two scenarios. The general problem of secure lossy source coding in the presence of an eavesdropper, who observes the information bits and is looking to reveal a private uncoded correlated sequence. In the second scenario, the opposite case is considered. The decoder wishes to recover the uncoded correlated sequence; also known as relevant utility in this case; while the eavesdropper aims at revealing information about the coded source itself. It is assumed that all links between encoders and decoders are public and noiseless so that they cannot provide any advantage to increase secrecy. The key aspect of this model is that the message J produced by the encoder could play a double role depending on the desired setting. In the first setting, it needs to carry the description of the source X itself while still preserving privacy of a relevant data Y unavailable at the encoder. In the other setting, it needs to carry information about a relevant data Y aimed at enabling the decoder to reproduce this correlated data in the best possible way while keeping data about the source X private from an eventual eavesdropper.

Extracting relevant aspects of complex data as in the second scenario is a fundamental task in signal processing and statistics. The problem is often that the data contains many structures in which one or some of them might be relevant. For example, speech signals may be characterized by their volume level, pitch, or content; pictures can be represented by their level of luminosity, color saturation, etc. This problem was stated in an information-theoretic way by the information bottleneck (IB) method [112]. Given the joint distribution of a source variable X and another relevance variable Y, IB operates to compress X, while preserving maximum information about Y. The variable Y thus implicitly defines what is relevant in X. A practical application involving electric consumption data measured

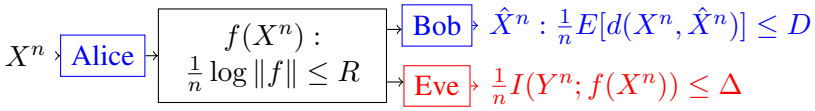


Figure 4.1: Lossy source coding with a privacy constraint.

from real houses is finally investigated.

Such scenario may appear eccentric, but could have relevant applications in the context of smart grids which involve features as demand response and load shifting. As a matter of fact, many new studies are carried out to allow peak energy demands to be shifted to off-peak hours, thus reducing costs for power generation facilities as well as households consumptions. In some communities, capacity to generate power is limited, and when consumers use large amounts of power simultaneously, "brownouts" may occur.

During a peak load period (e.g. a specific day or hour of the day), we can estimate the consumption of a certain appliance (heater, washer, etc.) in a particular neighbourhood network from the correlated total consumption available at the source, to check if it forms a significant proportion of the total peak demand. If that's the case, work could be oriented in such a way to perform load shifting per device depending on its impact on the peak demand.

In this chapter, we will rather focus on an application that depends only on consumption in one household as we do not have enough information regarding a given appliance in a neighbourhood area network.

Notations

Random variables (RVs) are denoted by upper case letters. The cardinality of X is denoted by $\|\mathcal{X}\|$ or simply $\mathcal{X}$. Particular realizations of a random variable are written in corresponding lower case letters while samples are denoted with small letters. We use x^n and $\mathbf{x}$ to denote vectors samples from $\mathcal{X}^n$ of length n. The *Probability density functions* (pdfs) are denoted by the lower case letter p. The pdfs and the joint pdf of X and Y will be denoted p_X, p_Y and p_{XY}, respectively. For RVs X and Y with joint pdf p_{XY}, $h(X), h(Y|X)$ denote the *entropy* and the *conditional entropy*, respectively, while $I(X;Y)$ denotes the *mutual information*.

4.2 Lossy Source Coding with a Privacy Constraint

4.2.1 Problem Definition

In this section, we give a more rigorous formulation of the context depicted in Figure 4.1. The primary challenge in characterizing the privacy-utility trade-off is finding the appropriate quantitative measures of both the utility retained as well as the amount of information leaked.

Alice (the quantizer) observes a source of $X^n = (X_1, \ldots, X_n)$ with memory. Alice wishes to encode its data with a maximum rate R [bits per dimension] which is accomplished by mapping each

vector to the index $j \equiv f(x^n) \in \mathcal{J}$ using an n dimensional vector quantizer:

$$f : \mathcal{X}^n \longrightarrow \mathcal{J} \equiv \{1, \ldots, M\} \tag{4.1}$$

$M \leq 2^{nR}$. Bob is looking recover the source X^n from the received compressed version. The aim of Alice is to transmit the source to Bob in such a way resulting in a minimum expected distortion at Bob. Bob then computes an output sequence $\hat{X}^n = (\hat{X}_1, \ldots, \hat{X}_n)$ using the decoding function

$$g_X : \mathcal{J} \mapsto \hat{\mathcal{X}}^n \tag{4.2}$$

The encoder is chosen such that the input and output sequences achieve a desired utility given by an expected distortion constraint

$$D_X = \frac{1}{n}\,\mathbb{E}[d(X^n, g_X(f(X^n)))] = \frac{1}{n}\,\mathbb{E}[d(X^n, \hat{X}^n)] \tag{4.3}$$

$\hat{\mathcal{X}}^n = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_M\}$ is also denoted as the reproduction alphabet in this setting such that $g_X(j) = \hat{\mathbf{x}}_j$. $\mathbb{E}$ is the expectation and $d : \mathcal{X}^n \times \hat{\mathcal{X}}^n \mapsto \mathbb{R}^+$ is a distortion measure.

The message $f(\mathbf{x})$ is sent to Bob via a public link that is perfectly overheard by Eve (could be the collector itself). It is assumed that the eavesdropper has full knowledge of all system properties: the source statistics, the encoder f, and the legitimate decoder g_X. We assume a mutual information rate as a metric for privacy leakage; however, we allow the fact that an inference can be made from the source X^n about another sequence. We model the inferred data as a random variable Y^n correlated with the measurement variable X^n according to the joint distribution $p_{X^nY^n}$. Thus, the privacy leakage is the mutual information between Y^n and the message $f(X^n)$.

$$L_Y = \frac{1}{n} I(Y^n; f(X^n)) \ . \tag{4.4}$$

4.2.2 Optimal Quantizer Design

For a given rate R, the optimal encoder f is obtained from the optimization of 2 types of secure lossy compression problems. Given a maximum information leakage $L_Y \leq \Delta$, minimize the distortion D_X. Or, given a distortion at Bob $D_X \leq D$, minimize the information leakage at Eve L_Y.

Instead of using a cost function D_X, with a constrained L_Y, or L_Y, with a constrained D_X, we use the unconstrained Lagrangian cost function

$$\Gamma(f, g_X, \lambda) = D_X + \lambda L_Y \ , \tag{4.5}$$

Where $\lambda > 0$ is the Lagrange multiplier. The distortion $d(\mathbf{x}, \hat{\mathbf{x}})$ between an input $\mathbf{x}$ and the decoder output $\hat{\mathbf{x}} = g_X(f(\mathbf{x}))$ is assumed to be the squared error distortion. The MSE is well-known distortion measure usually used when estimating the source from the corresponding quantized values.

The set of different input vectors producing same output value will be referred to as the quantization

region. Let $\mathcal{R}_j$ be the encoding region associated with the index j. $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M\}$ denotes the partition of the space defining the mapping $f(\mathbf{x}) = j, \text{ if } \mathbf{x} \in \mathcal{R}_j$. The expected distortions can be thus written as

$$\begin{aligned} D_X &= \frac{1}{n}\int_{\mathcal{X}^n} p_{X^n}(\mathbf{x})\|\mathbf{x} - g_X \circ f(\mathbf{x})\|^2 d\mathbf{x} && (4.6)\\ &= \frac{1}{n}\sum_{j\in\mathcal{J}}\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x})\|\mathbf{x} - \hat{\mathbf{x}}_j\|^2 d\mathbf{x} && (4.7)\\ & && (4.8)\end{aligned}$$

The information leakage can be written as

$$L_Y = \frac{1}{n}I(Y^n; f(X^n)) = \frac{1}{n}H(f(X^n)) - \frac{1}{n}H(f(X^n)|Y^n) \tag{4.9}$$

Remark 4.1 *In most cases, what we have available is a set of sample vectors given by the training sequence instead of the pdfs. In such scenario, the optimal quantizer will have to be learned from this training sequence. We will refer to $\mathcal{T}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{Y}}$, the training sequences of $X^n \sim P_{X^n}$ and $Y^n \sim P_{Y^n}$ of sizes N_X and N_Y respectively. Define*

$$K_j(C_j(\mathbf{x})) = \frac{1}{N_X}\sum_{\mathbf{x}\in\mathcal{T}_{\mathbf{X}}} C_j(\mathbf{x})\mathbb{1}_{\mathcal{R}_j}(\mathbf{x}) \tag{4.10}$$

With $C_j(\mathbf{x})$ being any function of $\mathbf{x}$ which may also depend on j. The expression K_j is only used as an intermediate to reduce equations' sizes. The expected distortion and the information leakage can be thus estimated as follow

$$\begin{aligned} nD_X &= \sum_{j\in\mathcal{J}} K_j(\|\mathbf{x} - \hat{\mathbf{x}}_j\|^2) && (4.11)\\ nL_Y &= -\sum_{j\in\mathcal{J}} K_j(1)\log K_j(1) + \frac{1}{N_Y}\sum_{\mathbf{y}\in\mathcal{T}_{\mathbf{Y}}}\sum_{j\in\mathcal{J}} K_j(\mathbb{p}(\mathbf{x},\mathbf{y}))\log_2 K_j(\mathbb{p}(\mathbf{x},\mathbf{y})) && (4.12)\end{aligned}$$

$\mathbb{1}_{\mathcal{R}_j}(\mathbf{x})$ *is* 1 *if* $\mathbf{x} \in \mathcal{R}_j$ *and* 0 *otherwise is called the indicator function of region* $\mathcal{R}_j$. $\mathbb{p}(\mathbf{x},\mathbf{y}) = p_{X^nY^n}(\mathbf{x},\mathbf{y})/p_{X^n}(\mathbf{x})p_{Y^n}(\mathbf{y})$.

The problem of finding the optimal set of regions $\boldsymbol{\mathcal{R}}$ and representatives $\hat{\mathcal{X}}^n$ becomes:

$$(\boldsymbol{\mathcal{R}}_{\mathbf{opt}}, \hat{\mathcal{X}}^n_{\mathbf{opt}}) = \underset{\boldsymbol{\mathcal{R}},\hat{\mathcal{X}}^n}{\arg\min}\left\{D_X(\boldsymbol{\mathcal{R}}, \hat{\mathcal{X}}^n) + \lambda L_Y(\boldsymbol{\mathcal{R}})\right\}, \tag{4.13}$$

$D_X(\boldsymbol{\mathcal{R}}, \hat{\mathcal{X}}^n)$ and $L_Y(\boldsymbol{\mathcal{R}})$ are used to emphasize that the expected distortion of the quantizer f depends on the quantization regions $\boldsymbol{\mathcal{R}}$ and their representatives $\hat{\mathcal{X}}^n$ while the information leakage only depends on the quantization regions.

As it is, it is nearly impossible to find the global minimizer of the expected distortion simultaneously with respect to all the regions and representatives. Nevertheless, the optimization problem defined in

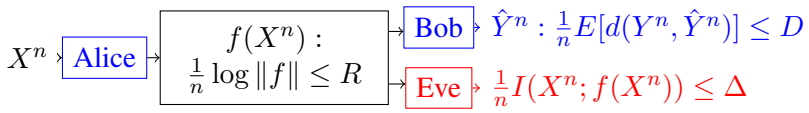


Figure 4.2: Secure lossy coding of a correlated relevant information with privacy constraints.

Equation (4.13) can be divided into two partial solutions. In the first step, we will focus on finding the optimal representatives $\hat{\mathcal{X}}^n_{\mathbf{opt}} = \{\hat{\mathbf{x}}^\star_1, \ldots, \hat{\mathbf{x}}^\star_M\}$ given the quantization regions $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_1, \ldots, \mathcal{R}_M\}$. We can write:

$$\{\hat{\mathbf{x}}^\star_1, \ldots, \hat{\mathbf{x}}^\star_M\} = \underset{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_M}{\arg\min} \{D_X(\boldsymbol{\mathcal{R}}, \hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_M) + \lambda L_Y(\boldsymbol{\mathcal{R}})\} \tag{4.14}$$

$$= \underset{\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_M}{\arg\min} D_X(\boldsymbol{\mathcal{R}}, \hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_M)\ . \tag{4.15}$$

The fact that $L_Y(\boldsymbol{\mathcal{R}})$ is independent of all representatives $\hat{\mathcal{X}}^n$, the solution of the optimal representatives is hence the same as those obtained by Lloyd-max

$$\hat{\mathbf{x}}^\star_j = \frac{\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x})\mathbf{x}d\mathbf{x}}{\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x})d\mathbf{x}} \quad \forall j \in \{1, \ldots, M\} \tag{4.16}$$

Similarly, we now try finding the optimal quantization regions $\boldsymbol{\mathcal{R}}_{\mathbf{opt}} = \{\mathcal{R}^\star_1, \ldots, \mathcal{R}^\star_M\}$ given the representatives $\hat{\mathcal{X}}^n = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_M\}$

$$\{\mathcal{R}^\star_1, \ldots, \mathcal{R}^\star_M\} = \underset{\mathcal{R}_1,\ldots,\mathcal{R}_M}{\arg\min} \left\{D_X(\mathcal{R}_1, \ldots, \mathcal{R}_M, \hat{\mathcal{X}}^n) + \lambda L_Y(\mathcal{R}_1, \ldots, \mathcal{R}_M,)\right\} \tag{4.17}$$

As you can see here, the problem is more complicated and cannot be decoupled into M different problems. An iterative algorithm for finding the optimal regions $\boldsymbol{\mathcal{R}}$ and the representation points $\hat{\mathcal{X}}^n$ to meet the above necessary conditions is outlined in Algorithm 4.

4.3 Lossy Source Coding of a Correlated Relevant Information with a Privacy Constraint

4.3.1 Problem Definition

In this section, we study the setting depicted in Figure 4.2. The decoder is not concerned in recovering X^n anymore, but the correlated relevant information Y^n from the received message $f(X^n)$, while keeping the source private from the eavesdropper. For this purpose, we define the following decoding function:

$$g_Y : \mathcal{J} \mapsto \hat{\mathcal{Y}}^n \tag{4.19}$$

Algorithm 4 Vector Quantizer Design

1. Given a training set $\Big\{(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2),\dots\Big\} \sim p_{X^nY^n}$,
2. Initialize $\{\mathcal{R}_1^{(0)},\mathcal{R}_2^{(0)},\dots,\mathcal{R}_M^{(0)}\}$ by assigning the vectors $\mathbf{x}$ to the regions randomly,
3. Loop
   (a) For a given $\boldsymbol{\mathcal{R}}^{(k)}$, calculate the set of representatives $\{\hat{\mathbf{x}}_1^{(k)},\hat{\mathbf{x}}_2^{(k)},\dots,\hat{\mathbf{x}}_M^{(k)}\}$ such that
   $$\hat{\mathbf{x}}_j^{(k)} = \frac{\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x})\mathbf{x}d\mathbf{x}}{\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x})d\mathbf{x}} \approx \frac{\sum\limits_{\mathbf{x}\in\mathcal{T}_X} \mathbb{1}_{\mathcal{R}_j}^{(k)}(\mathbf{x})\mathbf{x}}{\sum\limits_{\mathbf{x}\in\mathcal{T}_X} \mathbb{1}_{\mathcal{R}_j}^{(k)}(\mathbf{x})}\,. \tag{4.18}$$
   (b) For a given set of representatives $\{\hat{\mathbf{x}}_1^{(k)},\hat{\mathbf{x}}_2^{(k)},\dots,\hat{\mathbf{x}}_M^{(k)}\}$, calculate $\boldsymbol{\mathcal{R}}^{(k+1)} = \{\mathcal{R}_1^{(k+1)},\mathcal{R}_2^{(k+1)},\dots,\mathcal{R}_M^{(k+1)}\}$ by looping over all vectors $\mathbf{x}$ and moving each vector over all set of regions in $\boldsymbol{\mathcal{R}}^{(k)}$ and then calculate the corresponding $\Gamma(f,g_X,\lambda)$ function given by:
   $$\begin{aligned} n\Gamma(f,g_X,\lambda) &= \sum_{j\in\mathcal{J}} K_j(\|\mathbf{x}-\hat{\mathbf{x}}_j\|^2) \\ &- \lambda\sum_{j\in\mathcal{J}}\left\{K_j(1)\log K_j(1) - \frac{1}{N_Y}\sum_{\mathbf{y}\in\mathcal{T}_\mathbf{Y}} K_j(\mathbb{p}(\mathbf{x},\mathbf{y}))\log_2 K_j(\mathbb{p}(\mathbf{x},\mathbf{y}))\right\}, \end{aligned}$$
   Moving a vector $\mathbf{x}$ from $\mathcal{R}_j$ to $\mathcal{R}_{j'}$ can be simply done by updating $K_j(C(\mathbf{x}))$ with $K_j(C(\mathbf{x}))-\frac{1}{N_X}C(\mathbf{x})$ and that of $K_{j'}(C(\mathbf{x}))$ with $K_{j'}(C(\mathbf{x}))+\frac{1}{N_X}C(\mathbf{x})$. Update $\mathbb{1}_{\mathcal{R}_j}(\mathbf{x})^{(k+1)}$ at each time after assigning $\mathbf{x}$ to the region having the minimum $\Gamma(f,g_X,\lambda)$.
4. Step 4: Repeat until $\max\limits_j \|\hat{\mathbf{x}}_j^{(k+1)} - \hat{\mathbf{x}}_j^{(k)}\| \leq \delta$.
5. Step 5: Return $\{\mathcal{R}_1^\star,\mathcal{R}_2^\star,\dots,\mathcal{R}_M^\star\}$.

where $\hat{\mathcal{Y}}^n = \{\hat{\mathbf{y}}_1,\dots,\hat{\mathbf{y}}_M\}$ is the reproduction alphabet in this setting such that $g_Y(j) = \hat{\mathbf{y}}_j$. The average distortion of the code is given by

$$D_Y = \frac{1}{n}\,\mathbb{E}[d(Y^n, g_Y(f(X^n)))] = \frac{1}{n}\,\mathbb{E}[d(Y^n,\hat{Y}^n)] \tag{4.20}$$

where $d:\mathcal{Y}^n\times\hat{\mathcal{Y}}^n \mapsto \mathbb{R}^+$ is a distortion measure. On the other hand, the information leakage at the eavesdropper becomes:

$$L_X = \frac{1}{n} I\Big(X^n; f(X^n)\Big). \tag{4.21}$$

4.3.2 Quantizer Design

Similarly, the objective function to be optimized can be written

$$\Gamma(f, g_Y, \lambda) = D_Y + \lambda L_X \tag{4.22}$$

The distortion measure $d(y^n, \hat{y}^n)$ between y^n and the decoder output $\hat{y}^n = g_Y(f(x^n))$ is also assumed be the squared error distortion.

$$\begin{aligned} D_Y &= \frac{1}{n}\int_{\mathcal{X}^n}\int_{\mathcal{Y}^n} p_{X^nY^n}(\mathbf{x},\mathbf{y})\|\mathbf{y} - g_Y \circ f(\mathbf{x})\|^2 d\mathbf{x}d\mathbf{y} && (4.23)\\ &= \frac{1}{n}\sum_{j\in\mathcal{J}}\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x})\int_{\mathcal{Y}^n} p_{Y^n|X^n}(\mathbf{y}|\mathbf{x})\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2 d\mathbf{x}d\mathbf{y} && (4.24)\\ &= \frac{1}{n}\sum_{j\in\mathcal{J}} K_j\Big(\frac{1}{N_{Y|X}}\sum_{\mathbf{y}\in\mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}}\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2\Big) && (4.25)\end{aligned}$$

$\mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ is a set of $\mathbf{y}$ vectors such that $\mathbf{y} \sim p_{Y^n|X^n=\mathbf{x}}$ of size $N_{Y|X}$. The information leakage can be reduced to $nL_X = H(f(X^n)) = -\sum_{j\in\mathcal{J}} K_j(1)\log K_j(1)$. This is due to the fact that knowing X^n, the uncertainty about $f(X^n)$ is null; $H(f(X^n)|X^n) = 0$. The problem of finding the optimal set of regions $\boldsymbol{\mathcal{R}}$ and the corresponding representatives $\hat{\mathcal{Y}}^n$ become

$$(\boldsymbol{\mathcal{R}}_{\mathbf{opt}}, \hat{\mathcal{Y}}^n_{\mathbf{opt}}) = \arg\min_{\boldsymbol{\mathcal{R}},\hat{\mathcal{Y}}^n}\Big\{D_Y(\boldsymbol{\mathcal{R}},\hat{\mathcal{Y}}^n) + \lambda L_X(\boldsymbol{\mathcal{R}})\Big\} \tag{4.26}$$

The design of the optimal quantizer is hence outlined in Algorithm 5.

4.4 Application: Memoryless Gaussian Sources

Consider the special case of memoryless Gaussian sources represented by the following observation model:

$$(X,Y) \sim \mathcal{N}(0,\boldsymbol{\Sigma}) \quad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y\\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix} \tag{4.29}$$

In this particular example, X^n, Y^n denote a sequence of i.i.d. random variables from (X, Y). In what follows, we will take $\sigma_X = \sigma_Y = 1$. Note that the marginal distributions of X and Y under both hypotheses follow the standard normal distribution $\mathcal{N}(0, 1)$.

We can hence write the conditional probability of $Y_i|X_i = x_i \ , i = \{1, \ldots, n\}$ as

$$Y_i|X_i = x_i \sim \mathcal{N}(\rho x_i, (1-\rho^2)) \ . \tag{4.30}$$

The expected distortion D_Y when the decoder aims at recovering Y^n can be written:

$$D_Y = \frac{1}{n}\int_{\mathcal{X}^n} p_{X^n}(\mathbf{x})\mathbb{E}_{Y^n|X^n}[\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2]d\mathbf{x} \tag{4.31}$$

Algorithm 5 Vector Quantizer Design for Distributed Binary Decision with Privacy Constraints

1. Given a the training set $\Big\{(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2),\dots\Big\} \sim p_{X^nY^n}$,
2. Initialize $\{\mathcal{R}_1^{(0)},\mathcal{R}_2^{(0)},\dots,\mathcal{R}_M^{(0)}\}$ by assigning the the vectors $\mathbf{x}$ to the regions randomly,
3. Loop
   (a) For a given $\boldsymbol{\mathcal{R}}^{(k)}$, calculate the set of representatives $\{\hat{\mathbf{y}}_1^{(k)},\hat{\mathbf{y}}_2^{(k)},\dots,\hat{\mathbf{y}}_M^{(k)}\}$ such that
   $$\hat{\mathbf{y}}_j^{(k)} = \frac{\int_{\mathcal{Y}^n}\int_{\mathcal{R}_j^{(k)}} p_{X^nY^n}(\mathbf{x},\mathbf{y})\mathbf{y}d\mathbf{x}d\mathbf{y}}{\int_{\mathcal{R}_j^{(k)}} p_{X^n}(\mathbf{x})d\mathbf{x}} \approx \frac{\sum\limits_{\mathbf{x}\in\mathcal{T}_{\mathbf{X}}}\big(\frac{1}{N_{Y|X}}\sum\limits_{\mathbf{y}\in\mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}}\mathbf{y}\big)\mathbb{1}_{\mathcal{R}_j}^{(k)}(\mathbf{x})}{\sum\limits_{\mathbf{x}\in\mathcal{T}_{\mathbf{X}}}\mathbb{1}_{\mathcal{R}_j}^{(k)}(\mathbf{x})}\,. \tag{4.27}$$
   (b) For a given set of representatives $\{\hat{\mathbf{y}}_1^{(k)},\hat{\mathbf{y}}_2^{(k)},\dots,\hat{\mathbf{y}}_M^{(k)}\}$, calculate $\boldsymbol{\mathcal{R}}^{(k+1)} = \{\mathcal{R}_1^{(k+1)},\mathcal{R}_2^{(k+1)},\dots,\mathcal{R}_M^{(k+1)}\}$ by looping over all vectors $\mathbf{x}$ in the training set and moving each vector over all set of regions in $\boldsymbol{\mathcal{R}}^{(k)}$ and then calculate the corresponding $\Gamma(f,g_Y,\lambda)$ function given by:
   $$n\Gamma(f,g_Y,\lambda) = \sum_{j\in\mathcal{J}}\Big\{K_j\big(\frac{1}{N_{Y|X}}\sum_{\mathbf{y}\in\mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}}\|\mathbf{y}-\hat{\mathbf{y}}_j\|^2\big) - \lambda K_j(1)\log K_j(1)\Big\} \tag{4.28}$$
   Update $\mathbb{1}_{\mathcal{R}_j}(\mathbf{x})^{(k+1)}$ at each time after assigning $\mathbf{x}$ to the region having the minimum $\Gamma(f,g_Y,\lambda)$.
4. Step 4: Repeat until $\max\limits_j\|\hat{\mathbf{y}}_j^{(k+1)}-\hat{\mathbf{y}}_j^{(k)}\| \le \delta$.
5. Step 5: Return $\{\mathcal{R}_1^\star,\mathcal{R}_2^\star,\dots,\mathcal{R}_M^\star\}$.

The conditional mean of the distance $\|\mathbf{y}-\hat{\mathbf{y}}_j\|^2$ with respect to $X^n=\mathbf{x}$ is then $\mathbb{E}_{Y^n|X^n}[\|\mathbf{y}-\hat{\mathbf{y}}_j\|^2] = \|\rho\mathbf{x}-\hat{\mathbf{y}}_j\|^2 + n(1-\rho^2)$. Then the expected distortion measure reduces to a much smoother form for this case

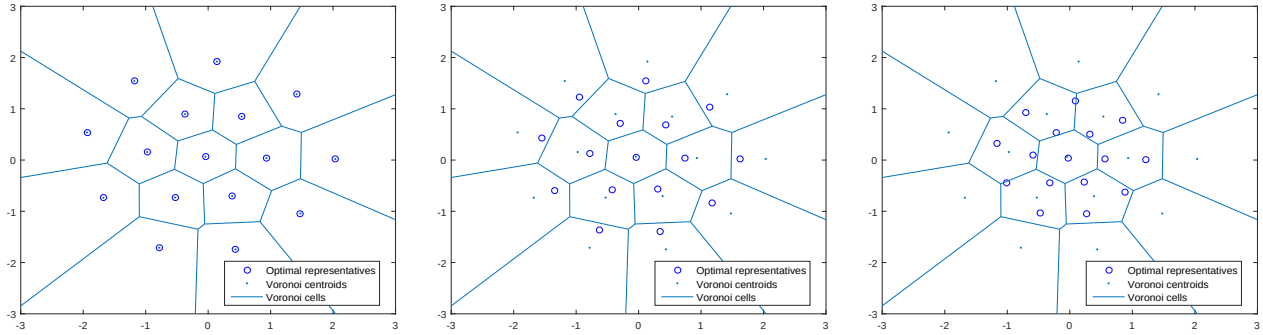
$$D_Y = \frac{1}{n}\sum_{j\in\mathcal{J}}\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x})\|\rho\mathbf{x}-\hat{\mathbf{y}}_j\|^2 d\mathbf{x} + (1-\rho^2)\,. \tag{4.32}$$

It can be noticed that D_X is a particular solution of D_Y when $\rho=1$.

In the absence of any privacy constraint, the algorithm minimizing D_θ, $\theta=\{X,Y\}$ proceeds by alternating between two steps:

$$\begin{aligned}
\text{if } \theta=X:\ \hat{\mathbf{x}}_j &= \frac{\int_{\mathcal{R}_j}\mathbf{x}p_{X^n}(\mathbf{x})d\mathbf{x}}{\int_{\mathcal{R}_j}p_{X^n}(\mathbf{x})d\mathbf{x}} \quad \mathcal{R}_j = \Big\{\mathbf{x}|\ \|\mathbf{x}-\hat{\mathbf{y}}_j\|^2 < \|\mathbf{x}-\hat{\mathbf{y}}_{j'}\|^2, j'\neq j\Big\}\,,\\
\text{if } \theta=Y:\ \hat{\mathbf{y}}_j &= \frac{\rho\int_{\mathcal{R}_j}\mathbf{x}p_{X^n}(\mathbf{x})d\mathbf{x}}{\int_{\mathcal{R}_j}p_{X^n}(\mathbf{x})d\mathbf{x}} \quad \mathcal{R}_j = \Big\{\mathbf{x}|\ \|\rho\mathbf{x}-\hat{\mathbf{y}}_j\|^2 < \|\rho\mathbf{x}-\hat{\mathbf{y}}_{j'}\|^2, j'\neq j\Big\}\,.
\end{aligned} \tag{4.33}$$

Results show that regions remain unchanged in the two settings and with different values of ρ whereas



(a) $\underset{\boldsymbol{\mathcal{R}},\hat{\mathcal{X}}^n}{\arg\min}\, D_X$ (b) $\underset{\boldsymbol{\mathcal{R}},\hat{\mathcal{Y}}^n}{\arg\min}\, D_Y\ ,\ \rho = 0.8$ (c) $\underset{\boldsymbol{\mathcal{R}},\hat{\mathcal{Y}}^n}{\arg\min}\, D_Y\ ,\ \rho = 0.6$

Figure 4.3: Results for VQ optimal regions when $R = 2$. (a) shows optimal quantization regions and representatives under no privacy constraints when the decoder aims at recovering X^n. (b) and (c) show optimal quantization regions and representatives under no privacy constraints when the decoder aims at recovering Y^n when $\rho = 0.8$ and $\rho = 0.6$ respectively. Circles show optimal centroids and dots show Voronoi centroids.

the values of the centroids given by $\hat{\mathcal{Y}}^n$ are scaled by ρ. This can be explained theoretically,

$$\begin{aligned}\left\{\mathbf{x}|\ \|\rho\mathbf{x} - \hat{\mathbf{y}}_j\|^2 < \|\rho\mathbf{x} - \hat{\mathbf{y}}_{j'}\|^2, j' \neq j\right\} &= \left\{\mathbf{x}|\ \|\mathbf{x} - \frac{\hat{\mathbf{y}}_j}{\rho}\|^2 < \|\mathbf{x} - \frac{\hat{\mathbf{y}}_{j'}}{\rho}\|^2, j' \neq j\right\} \\ &= \left\{\mathbf{x}|\ \|\mathbf{x} - \hat{\mathbf{x}}_j\|^2 < \|\mathbf{x} - \hat{\mathbf{x}}_{j'}\|^2, j' \neq j\right\}. \quad (4.34)\end{aligned}$$

Figures 4.3a to 4.3c show the optimal quantization regions for different scenarios for the given Gaussian sources when the rate $R = 2$.

The performance of VQ is typically given in terms of the signal-to-distortion ratio (SDR):

$$SDR = 10\log_{10}\frac{\sigma_\theta^2}{D_\theta}\ (\text{in dB}), \quad \theta = \{X, Y\} \quad (4.35)$$

where σ_θ^2 is the variance of the source and D_θ is the expected distortion. Table 4.1 shows the performance of the VQ for the memoryless Gaussian sources when the decoder aims at decoding X^n while Table 4.2 shows the performance when the decoder aims at decoding Y^n. It can be noticed the decline in performance from the first to the second scenario. In other words, in order to recover Y^n with the same distortion of recovering X^n, we might need 4 times the rate.

Rate	SDR (in dB)							
(bits/dimesion)	n=1	n=2	n=3	n=4	n=5	n=6	n=8	n=10
1	4.3441	4.4313	4.4861	4.7496	4.8392	4.9890	5.4295	6.2376
2	9.3180	9.6905	10.1266	10.6654	11.6294	-	-	-
3	14.6554	15.3717	16.3239	-	-	-	-	-
4	20.2535	20.6972	-	-	-	-	-	-
5	25.2260	26.0711	-	-	-	-	-	-

Table 4.1: The performance of the VQ for the memoryless Gaussian source when the decoder aims at recovering X^n, i.e., $SDR = 10\log_{10}(\sigma_X^2/D_X)$.

Rate	SDR (in dB)							
(bits/dimesion)	n=1	n=2	n=3	n=4	n=5	n=6	n=8	n=10
1	2.2805	2.2324	2.3085	2.4365	2.4364	2.4988	2.6164	2.8307
2	3.5558	3.6680	3.7776	3.8190	3.9164	-	-	-
3	4.1403	4.2951	4.3355	-	-	-	-	-
4	4.4074	4.3622	-	-	-	-	-	-
5	4.4260	4.4766	-	-	-	-	-	-

Table 4.2: The performance of the VQ for the memoryless Gaussian source when the decoder aims at recovering Y^n, i.e., $SDR = 10\log_{10}(\sigma_Y^2/D_Y)$ when $\rho = 0.8$.

Figure 4.4a demonstrate the rate-distortion function when $\theta = X$ and $\theta = Y$ at 2 different correlations $\rho = 0.8, 0.9$. It can be noticed that the $R(D)$ function is better when the problem is recovering X^n from $f(X^n)$. A higher ρ also signifies a better rate-distortion function. This is well demonstrated in Figure 4.4b. The variation of the $R(D)$ function with respect to n is not significant because the source is memoryless. A significant improvement would have taken place if the source was not memoryless.

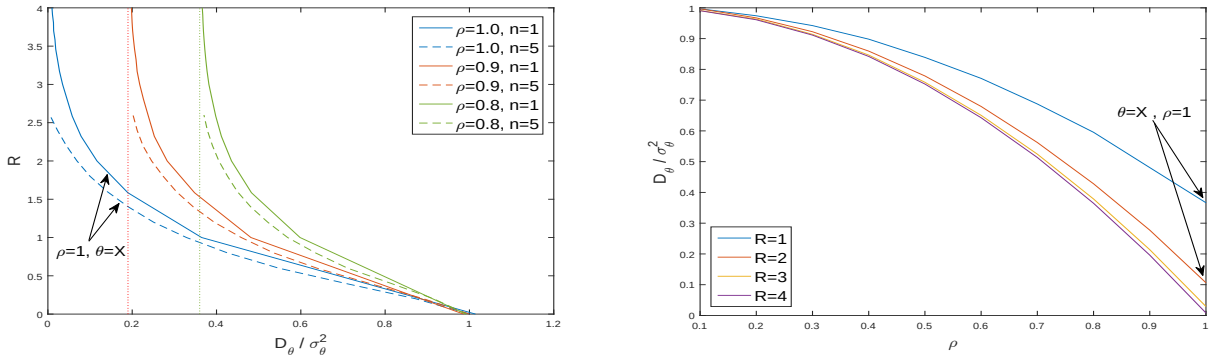
In order to study the performance of VQ with privacy, we need to investigate the variation of information leakage vs. distortion for a given rate. The expression of the information leakage in Equation (4.12) is a time consuming execution. For the memoryless Gaussian sources defined in Equation (4.29), we can use the fact that $Y^n = \rho I X^n + \mathcal{N}(0, (1-\rho^2)I)$ when $\sigma_X = \sigma_Y = 1$ and the entropy power inequality to find a feasible solution with much lower execution time. I being the identity matrix of size n.

$$L_Y = \frac{1}{n}(h(Y^n) - h(Y^n|f)) = \frac{1}{2}\log_2(2\pi e) - \frac{1}{n}h(Y^n|f) \tag{4.36}$$

Using the entropy power inequality, a lower bound of $h(Y^n|f)$ is given

$$h(Y^n|f) \geq \frac{n}{2}\log_2\left(\rho^2 2^{\frac{2}{n}h(X^n|f)} + (2\pi e)(1-\rho^2)\right) \tag{4.37}$$

Using the fact that $\frac{1}{n}h(X^n|f) = \frac{1}{n}h(X^n) - \frac{1}{n}H(J) = \frac{1}{2}\log_2(2\pi e) - \frac{1}{n}H(J)$, $H(Y^n|f)$ can be hence



(a) D_θ/σ_θ^2 vs. R (b) D_θ/σ_θ^2 vs. ρ

Figure 4.4: $\theta \in \{X, Y\}$. (a) shows the rate-distortion function $R(D)$ when $n = 1$ and $n = 5$. Blue curves show $R(D_X)$ when the main problem is to recover X^n. Red and green curves show $R(D_Y)$ when the main problem is to recover Y^n, the vertical doted lines in this case represent the minimum attainable distortion, $D_{Y(min)} = 1 - \rho^2$ at each case. (b) show distortion D_θ/σ_θ^2 vs. correlation ρ, a correlation of $\rho = 1$ corresponds to the case of $\theta = X$.

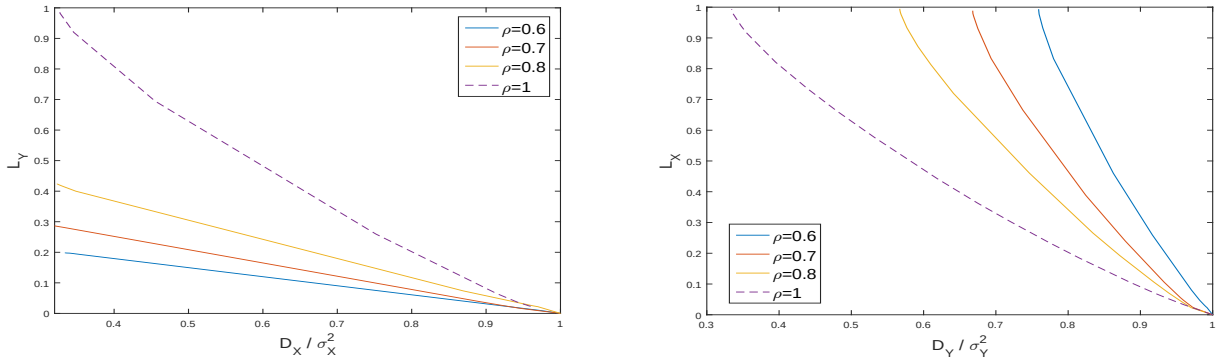
expressed in terms of $H(J)$ only

$$
\begin{aligned}
L_Y &\leq \frac{1}{2}\log_2(2\pi e) - \frac{1}{2}\log_2\left((2\pi e)\rho^2 2^{-\frac{2}{n}h(J)} + (2\pi e)(1-\rho^2)\right) && (4.38)\\
&= -\frac{1}{2}\log_2\left(\rho^2 2^{-\frac{2}{n}h(J)} + (1-\rho^2)\right) && (4.39)\\
&= -\frac{1}{2}\log_2\left(\rho^2 2^{-2L_X} + (1-\rho^2)\right) && (4.40)
\end{aligned}
$$

The objective functions of the two settings can be written

$$
\begin{aligned}
(\boldsymbol{\mathcal{R}}_{\mathbf{opt}}, \hat{\mathcal{X}}^n_{\mathbf{opt}}) &= \underset{\boldsymbol{\mathcal{R}}, \hat{\mathcal{X}}^n}{\arg\min} \left\{ \frac{1}{n}\sum_{j\in\mathcal{J}} \int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) \|\mathbf{x} - \hat{\mathbf{x}}_j\|^2 d\mathbf{x} - \lambda \frac{1}{2}\log_2\left(\rho^2 2^{-\frac{2}{n}H(J)} + (1-\rho^2)\right)\right\}\\
(\boldsymbol{\mathcal{R}}_{\mathbf{opt}}, \hat{\mathcal{Y}}^n_{\mathbf{opt}}) &= \underset{\boldsymbol{\mathcal{R}}, \hat{\mathcal{Y}}^n}{\arg\min} \left\{ \frac{1}{n}\sum_{j\in\mathcal{J}} \int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) \|\rho\mathbf{x} - \hat{\mathbf{y}}_j\|^2 d\mathbf{x} + (1-\rho^2) + \lambda \frac{1}{n} H(J)\right\}
\end{aligned}
$$

$\frac{1}{n}H(J)$ is nothing but the actual rate of the encoder. Figure 4.5a studies the first setting, i.e., L_X vs. D_Y when $R = 1$ and $n = 4$ for different correlation values ρ between X and Y. A higher correlation value indicates a better distortion-leakage curve. On the other hand, Figure 4.5a studies the second setting, i.e., L_Y vs. D_X when for different correlation values ρ between X and Y. Contrarily, A lower correlation value indicates a better distortion-leakage curve.



(a) D_X/σ_X^2 vs. L_Y (b) D_Y/σ_Y^2 vs. L_X

Figure 4.5: (a) shows the rate-leakage function $L_Y(D_X)$. (b) shows the rate-leakage function $L_X(D_Y)$. In both cases, we assume $n = 4$ and $R = 1$.

4.5 Pratical Application: Smart Meter Device Consumption Recovery under Privacy Contraints

Smart meters read aggregate power consumption in the house and communicates it to the central unit in the aim of providing a certain utility for the user. However, the collection of such data at a fine-grained time-scale presents privacy risks to the household occupants. Aggregate loads can be disaggregated and that the per-device consumption at every instant can be retrieved which may reveal private information, including household occupancy, sleeping and eating patterns of its members, etc.

Experimental setup. The scenario of this experiment involves a smart meter in a household and a service provider. The smart meter reads the aggregate consumption and communicates data to the service provider. Exactly as before, two scenarios can be studied. In the first scenario, the service provider is looking to recover the aggregate consumption X^n in the household. The household is willing to give the aggregate load X^n to the service provider, but wishes to keep the information leakage regarding a certain appliance Y^n; say the washing machine; bounded. In the second scenario, the service provider is looking to recover the consumption of a certain appliance Y^n in the household; say also the washing machine; to provide utility to the user, for instance automated control of the washer/dryer. The household is also willing to give the encoded version of the aggregate load X^n to the service provider, but wishes to keep the information leakage regarding the source X^n bounded.

The Reference Energy Disaggregation Data Set (REDD), is an available data set containing detailed power usage information from several real households [103]. The data contains power consumption for the whole house as well as for each individual device in the house recorded during a couple of months. From now on, we will use the total electrical consumption in one of the houses to represent our source X^n, while the sequence Y^n will be the consumption of the washer/dryer available at the same time.

Equation (4.28) requires a training sequence $\mathcal{T}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\ \forall\, \mathbf{x}$. Consequently, we would like to generate sample vectors of $Y^n|X^n = \mathbf{x}$ to allow the quantization design to be possible. In order to do that, an appropriate distribution can be fitted closely to the observed data. A distribution giving a close fit is

supposed to lead to good predictions.

There are many probability distributions of which some can be fitted more closely to the observed data than others, depending on the characteristics of the data. A copula; used to describe the dependence between random variables; can also be adopted to generate pseudo-random samples from general classes of multivariate probability distributions.

To estimate the goodness of fit between original data sets and proposed distributions, several tests exist. The Kolmogorov-Smirnov statistic allows us to measure the maximum difference in value between the empirical cumulative distribution and the proposed to be fitted distribution. Comparing this value for several distribution densities allows us to choose those who fit our model the most. The Kolmogorov-Smirnov statistic for a given cumulative distribution function $F(\mathbf{u})$ is

$$D_{K-S} = \sup_{\mathbf{u}} |\hat{F}(\mathbf{u}) - F(\mathbf{u})| \,, \tag{4.41}$$

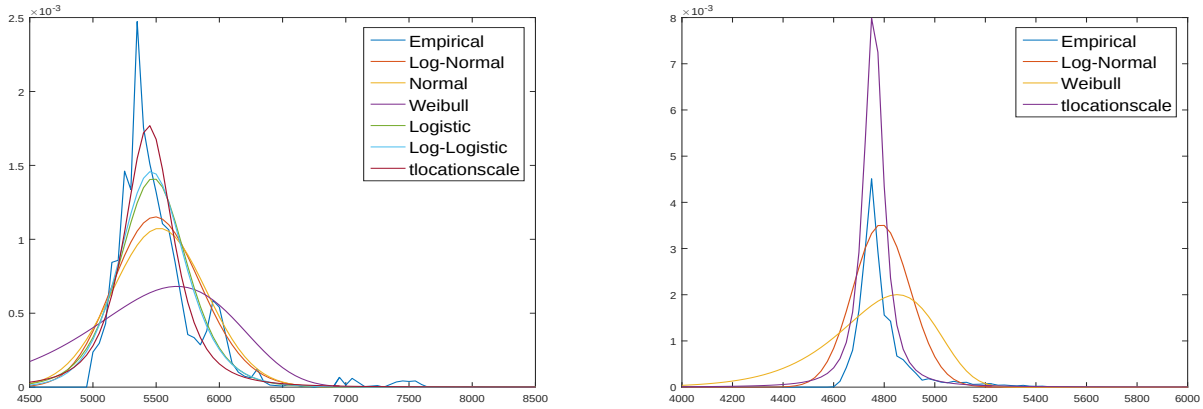
$\hat{F}(\mathbf{u})$ is the empirical cumulative distribution function. Table 4.3 shows different values of the K-S statistic for different distributions for both the univariate and the multivariate case when $n = 1$ and compare with a T copula. It can be noticed that using the T copula in order to generate samples has an advantage. Figure 4.6 shows the empirical distribution and the different distribution fittings of the univariate model when $n = 1$.

	Univariate Model		Multivariate Model
	X	Y	(X, Y)
Normal	0.1335	0.1876	0.2431
Log Normal	0.1195	0.1833	0.2361
Weibull	0.2237	0.2631	
Logistic	0.0734	0.1050	
loglogistic	0.0624	0.1025	
Gamma	0.1243	0.1848	
tlocationscale	0.0751	0.0831	
T copula			0.0539

Table 4.3: K-S statistic for different distribution laws when $n = 1$.

Results

Figures 4.7a to 4.7e show how the optimal quantization regions are distributed over the given testing set of the REDD data when $n = 2$ for both scenarios. For the first scenario; i.e. when the decoder aims at recovering X^n, the regions are nothing but the Voronoi regions. Figure 4.8 show performance with and without privacy constraint. It can be noticed the impact of an increase in n is much more significant now that the sources are no longer memoryless.



(a) X (b) Y

Figure 4.6: Empirical univariate distribution versus various proposed distribution fittings

Tables 4.4 and 4.5 show performance of the two scenario using the SDR metric for different values of the rate and the VQ dimension. The distortion in scenario 2 is very high, this is basically due to the low correlation between X^n and Y^n; the correlation between the aggregate consumption and that of the washer/dryer is lower than 0.2. On the other hand, much better performance is achieved when the problem becomes recovering the main source while keeping secret a lowly correlated sequence.

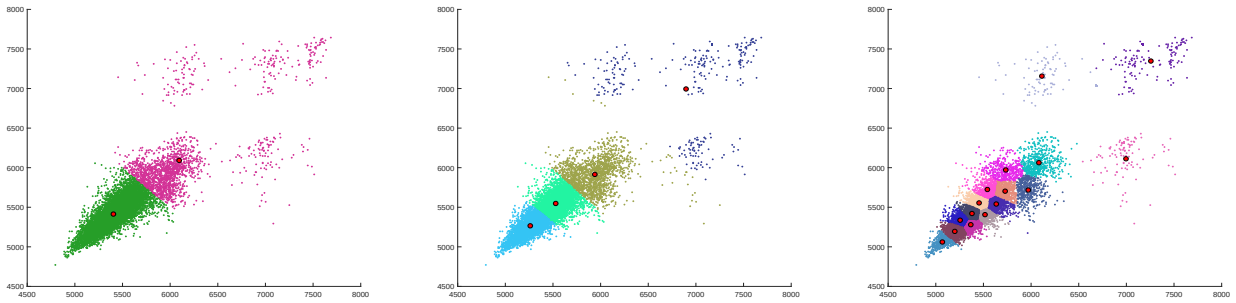
Different approaches can be done in order to improve the signal to distortion ratio. The designer can eventually increase the rate or the VQ dimension in order to achieve better performance. The role of the quantization rate appears to be more efficient. However, this could be useful in systems not taking into account security. In our scenario, increasing the rate would enormously increase the information leakage at the level of the eavesdropper as a result of the direct relation between the two measures.

VQ	SDR (in dB)				
dimesion	R=1	R=2	R=3	R=4	R=5
n=2	7.3266	14.0087	18.6309	22.7004	27.1631
n=3	9.4814	14.5406	19.6680	-	-
n=4	10.4940	15.7922	-	-	-

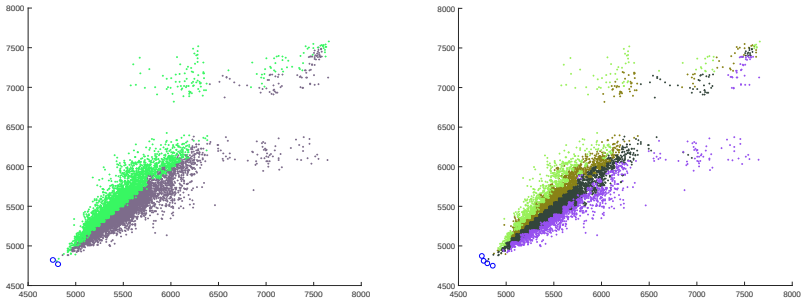
Table 4.4: The performance of the VQ when the decoder aims at recovering X^n, i.e., $SDR = 10\log_{10}(\sigma_X^2/D_X)$.

VQ	SDR (in dB)				
dimesion	R=0.5	R=1	R=2	R=3	R=4
n=2	1.1981	2.2097	2.7715	2.8534	2.9307
n=4	1.9523	3.0755	3.3738	-	-

Table 4.5: The performance of the VQ when the decoder aims at recovering Y^n, i.e., $SDR = 10\log_{10}(\sigma_Y^2/D_Y)$.



(a) $R = 0.5$ (b) $R = 1$ (c) $R = 2$



(d) $R = 0.5$ (e) $R = 1$

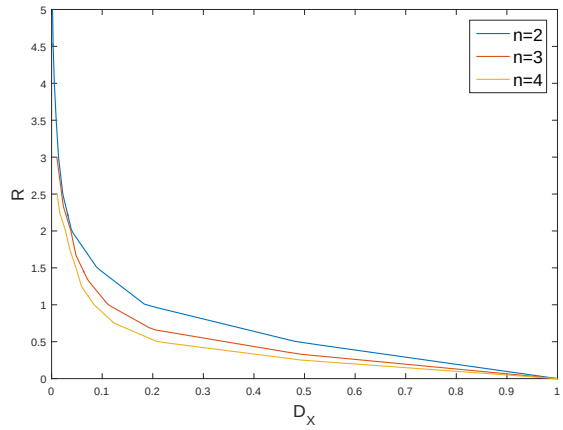
Figure 4.7: Optimal quantization regions for $n = 2$. First row shows regions optimizing D_X. Second row shows regions optimizing D_Y. x and y axis are amplitudes 1 and 2 of the vector $\mathbf{x}$ respectively. Codewords are marked with circles.

4.6 Conclusion

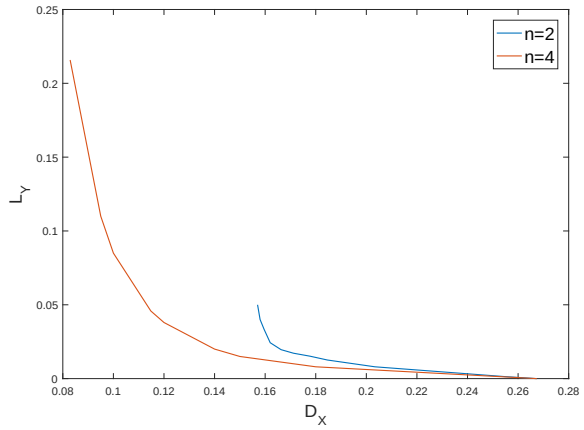
The theoretical framework that we have developed here allows us to precisely quantify the utility-privacy trade-off problem in two cases. Given a series of measurements X^n, we reveal a perturbation $\hat{X}^n$ that allows us to guarantee a measure of both privacy about a correlated sequence Y^n and utility in X^n in the first case. In the second scenario, we reveal a perturbation $\hat{Y}^n$ that allows us to guarantee a measure of both privacy about the source X^n and utility about the correlated sequence Y^n. The utility guarantee comes from the upper bound on the MSE distance while the privacy guarantee comes from the bound on information leakage while

The algorithm was applied to the spacial case of memoryless Gaussian sources. Numerical results show that the greater the information leakage we can tolerate, the better the signal to distortion ratio. In the case where the objective is to recover the correlated sequence, zero distortion is not possible even at infinite rate unless the two sequences are perfectly correlated. In memoryless sources, increasing the vectors' size has no impact on the privacy level but does improve the decoder performance. The improvement would be much more efficient for systems with memory.

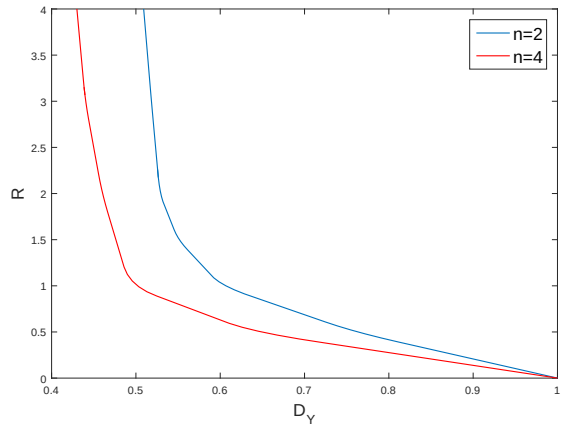
The chapter also addressed an important application in the context of smart grids where both scenarios can be used depending on the utility needed. The two experiments; carried out on real smart-meter dataset; show that information leakage can be bounded over time while maintaining utility of the



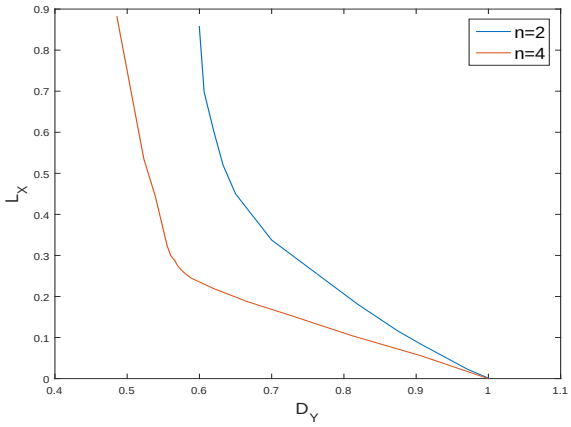
(a) R vs. D_X



(b) L_Y vs. D_X



(c) R vs. D_Y



(d) L_X vs. D_Y

Figure 4.8: First row shows performance when the decoder aims at recovering X^n. (a) shows the smart meter performance under no privacy constraint. (b) shows leakage vs. distortion in the presence of privacy when $R = 1$. Second row shows performance when the decoder aims at recovering Y^n. (c) shows the smart meter performance under no privacy constraint. (d) shows leakage vs. distortion in the presence of privacy when $R = 1$.

distorted data. The importance of such application is the ability to adapt to an unknown or a variable observation statistics. When the observation model was not available, we were able to extend the design algorithm to use with the available training sequences.

CHAPTER 5

CONCLUSION

5.1 General Comments

In this thesis, we derived some fundamental results on secure binary detection with side information in Chapters 2 and 3, and secure estimation (utility-privacy framework) in Chapter 4. In Chapter 2, we provided several new results of optimality and single-letter characterization of the achievable rate-error-equivocation region, whereas Chapters 3 and 4 provided an iterative algorithm for the design of the optimal quantizers that correspond to the corresponding specific scenario.

In Chapter 2, we were able to derive an achievable rate-error-equivocation region for the case of distributed general hypothesis problem when the side information is directly available at the detector, but was only able to prove the region's optimality for the particular case of testing against independence. This represents a main limitation of our results.

Similar limitations were also present in Chapter 3, where the design of the optimal vector quantization scheme was also restricted to the case of testing against independence as the error exponent was reduced to a smoother mutual information between the two remote sources, whereas the general hypothesis testing was applicable only when quantization was considered to be scalar.

Another limiting factor in Chapter 2 is the fact that both the encoder and the detector were assumed to observe vectors of i.i.d. realisations of X and Y respectively. The assumption that observations be i.i.d. tends to simplify the underlying mathematics of the statistical methods. However, in the practical applications considered in this thesis, the assumption cannot be considered realistic. We cannot assume that the voltages/frequencies or the electric consumptions measured at different adjacent time-slots to be independent. Such a restriction was solved in Chapter 3 where vectors were no longer i.i.d. when a vector quantization scheme was to be designed. This is considered as a huge advantage.

In spite of the aforementioned limitations, several possible extensions of this work can be identified, especially when taking into consideration possible applications in the context of smart grid.

5.2 Further Directions in the Detection Framework

5.2.1 M-ary Hypothesis Testing

In many problems, one might be required to distinguish between more than two hypotheses where signal detection problems can be casted in the framework of M-ary hypothesis testing. In such scenario, we wish to decide among M possible situations from some observations (data). These problems

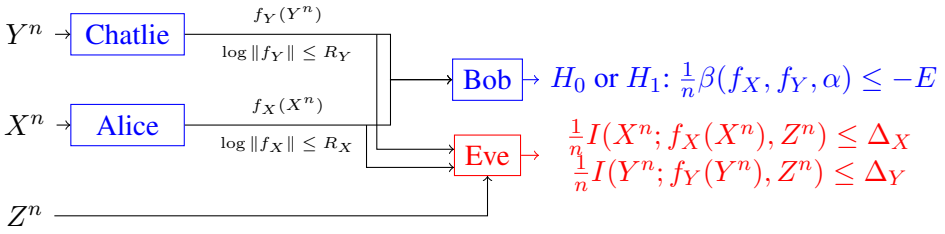


Figure 5.1: Distributed detection with coded side information under privacy constraints.

frequently occurs in pattern recognition or in communication when one of M signals is to be detected.

In both of Chapters 2 and 3, our problem only takes into consideration binary hypothesis testing, but what if there are multiple hypotheses that we need to choose from, for example M possible probability density functions of the source. This could also be an interesting problem with many applications in the domain of Smart Grids.

Although the Neyman Pearson decision rule adopted in our work can be extended for the M-ary hypothesis testing, it is often not used in practice. More commonly the minimum probability of error criterion or its generalization, the Bayes risk, is employed.

5.2.2 Hypothesis Testing with Coded Side Information

We can also think about an extension of the binary detection with side information at the detector to the case where the side information is no longer directly available but rather coded and sent through another noiseless channel. This could be critical when the data known by the "side information" is also remotely located with respect to the detector and perhaps susceptible to eavesdropping.

When coded, the side information is no longer referred to as side information especially when the goal remains to test between $H_0 : p_{XY}$ and $H_1 : p_{\bar{X}\bar{Y}}$. X^n and Y^n become 2 sources encoded separately and tested jointly. An (n, R_X, R_Y)-code for source coding in this setup is defined by

- An encoding function at the source X^n denoted by $f_X : \mathcal{X}^n \mapsto \{1, 2, \ldots, 2^{nR_X}\}$.
- An encoding function at the source Y^n denoted by $f_Y : \mathcal{Y}^n \mapsto \{1, 2, \ldots, 2^{nR_Y}\}$.
- A decision rule at the detector denoted by $\gamma : \{1, 2, \ldots, 2^{nR_X}\} \times \{1, 2, \ldots, 2^{nR_Y}\} \mapsto \{H_0, H_1\}$.

Eve could also be monitoring the channel between Charlie (the encoder of the side information) and the detector as seen in Figure 5.1. In that case, another security constraint must be taken into consideration. The new goal becomes finding the two optimal encoders f_X and f_Y that secure the desired trade-off between the HT performance at Bob and the privacy level at Eve, i.e. minimizing the errors $\beta(f_X, f_Y, \alpha)$ as well as the two information leakages at Eve. For the case of testing against independence, the new objective function for this scenario becomes:

$$\Gamma(f_X, f_Y, \lambda_X, \lambda_Y) \leq -\frac{1}{n}I(f_X(X^n); f_Y(Y^n)) + \lambda_X \frac{1}{n}I(X^n; f_X(X^n), Z^n) + \lambda_Y \frac{1}{n}I(Y^n; f_Y(Y^n), Z^n)\,, \quad (5.1)$$

$\frac{1}{n}I(f_X(X^n); f_Y(Y^n))$ represents the error exponent of the test at the detector Bob for the case of multi-terminal testing against independence [60]. The eavesdropper could be intercepting both channel simultaneously or could be intercepting only one of the channels at a time and at that time, is only interested in obtaining information about the related source of that channel. In both cases, the expression of the information leakage composite won't be affected. This is mainly due to the Markov chains given by $f_X(X^n) \multimap X^n \multimap (Y^n, Z^n)$ and $f_Y(Y^n) \multimap Y^n \multimap (X^n, Z^n)$.

If only one of the channel is being monitored, this could be seen as a particular case of this scenario taking either $\lambda_X = 0$ or $\lambda_Y = 0$. The main problem studied in Chapters 2 and 3; an important subsystem of the above general system; is also a special case with $R_Y \geq \frac{1}{n}\log\|\mathcal{Y}^n\|$, which means that Y^n is fully observed by the decoder (system with full side information).

To find the optimal vector quantizers, we can essentially repeat all the concepts adapted to the case of the presence of an uncoded side information.

$$h(f_Y(Y^n)|f_X(X^n)) \leq \int_{\mathcal{X}^n}\int_{\mathcal{Y}^n} p_{X^nY^n}(\mathbf{x},\mathbf{y})\|Q_Y(\mathbf{y}) - Q_X(\mathbf{x})\|^2 d\mathbf{x}d\mathbf{y} \tag{5.2}$$

$$= \sum_{j=1}^{2^{nR_X}}\sum_{j'=1}^{2^{nR_Y}} \int_{\mathcal{R}_j}\int_{\mathcal{S}_{j'}} p_{X^nY^n}(\mathbf{x},\mathbf{y})\|\hat{\mathbf{y}}_j - \hat{\mathbf{x}}_j\|^2 d\mathbf{x}d\mathbf{y}\,. \tag{5.3}$$

$Q_X : \mathcal{X}^n \mapsto \hat{\mathcal{X}}^n$ and $Q_Y : \mathcal{Y}^n \mapsto \hat{\mathcal{Y}}^n$ are the vector quantizers corresponding to the encoders f_X and f_Y of dimension n and size $M_X = 2^{nR_X}$ and $M_Y = 2^{nR_Y}$ respectively. $\hat{\mathcal{X}}^n = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_{M_X}\} \subset \mathbb{R}^n$ and $\hat{\mathcal{Y}}^n = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_{M_Y}\} \subset \mathbb{R}^n$ are their respective reproduction alphabets, whereas $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_1, \ldots, \mathcal{R}_{M_X}\}$ and $\boldsymbol{\mathcal{S}} = \{\mathcal{S}_1, \ldots, \mathcal{S}_{M_Y}\}$ are their respective quantization regions.

The smart grid application considered in section 3.6 could also make sense in the following framework. A coded side information means that the average electric consumption of the neighbourhood area network is no longer fully observed by the detector, but rather encoded by a separate encoder and sent over another rate-limited channel.

5.2.3 Decentralized Hypothesis Testing Network

Recently, there is an increasing interest in using decentralized systems with distributed encoders. One example of the decentralized system is shown in Figure 5.2. This system consists of k nodes, a detector and an eavesdropper. Correlated observation vectors X_i^n are collected by the nodes. Each node processes the observation by mapping it into the discrete variable $f_i(X_i^n)$; that is transmitted to the detector. The detector makes a decision on the joint distribution of the k sources based on the information it receives and a possible side information Y^n.

The goal is therefore to test between $H_0 : p_{X_1X_2\ldots X_k}$ vs. $H_1 : p_{\bar{X}_1\bar{X}_2\ldots\bar{X}_k}$. This could be very useful especially if you would like to know the probability distribution profile of a large consumption region or multiple production units.

The biggest issue we may face in the design of encoders of such scenario is that there's no explicit expression of the error exponent. On the other hand, the information leakage could be simply written the same way as before, and the method of Lagrange multipliers can be extended to solve problems with multiple constraints using a

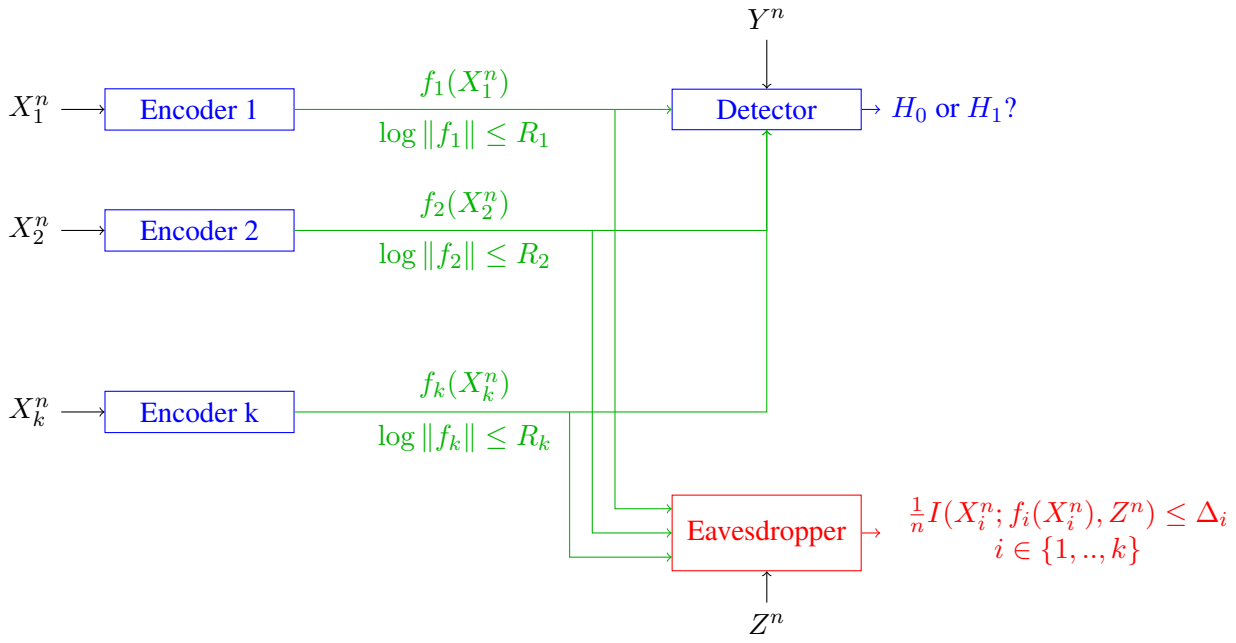


Figure 5.2: Decentralized detection under privacy constraints.

similar argument.

$$\Gamma(f_1, \ldots, f_k, \lambda_1, \ldots, \lambda_k) = \frac{1}{n} \log \beta(f_1, \ldots, f_k, \alpha) + \frac{1}{n} \sum_{i=1}^{k} \lambda_i I(X_i^n; f_i(X_i^n), Z^n) \tag{5.4}$$

5.3 Further Directions in the Estimation Framework

The decentralized network defined in the previous section could also be seen in an estimation lossy source coding framework where the decoder is interested in estimating the sum, for example, of the k separate sources $S^n = X_1^n + X_2^n + \cdots + X_k^n$. Of-course this could be done by decoding each source separately and then adding the k different estimated codewords obtained, however, this is not the most efficient way to estimate the sum.

The overall measure of performance at Bob is then the average distortion between the sum and the estimated sum, which is actually the output of the decoder g, over all possible inputs.

$$\begin{aligned} \mathbb{E}[d(\hat{S}^n, S^n)] = \int_{\mathcal{X}_1^n} \int_{\mathcal{X}_2^n} \cdots \int_{\mathcal{X}_k^n} \int_{\mathcal{Y}^n} & p_{X_1^n X_2^n \ldots X_k^n, Y^n}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k, \mathbf{y}) \\ & \|(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_k) - g(f_1(\mathbf{x}_1), f_2(\mathbf{x}_2), \ldots, f_k(\mathbf{x}_k), \mathbf{y})\|^2 d\mathbf{x}_1 d\mathbf{x}_2 \ldots d\mathbf{x}_k \end{aligned} \tag{5.5}$$

The decoder function at Bob is thus denoted by

$$g : \{1, 2, \ldots, 2^{nR_1}\} \times \{1, 2, \ldots, 2^{nR_2}\} \cdots \times \{1, 2, \ldots, 2^{nR_k}\} \times \mathcal{Y}^n \mapsto \hat{\mathcal{S}}^n \tag{5.6}$$

In this scenario, we can explore the above mentioned problem, where k nodes must separately compress their sources in such a way that the eavesdropper with side information learns as little as possible about them. In this model, a general secure distributed compression problem is considered in which the transmitters, with correlated observations, intend to send information to a receiver, Bob, over noiseless channels with limited capacity in such a way that he can reconstruct the arithmetic mean of the k source reliably. Also, there is an eavesdropper, Eve,

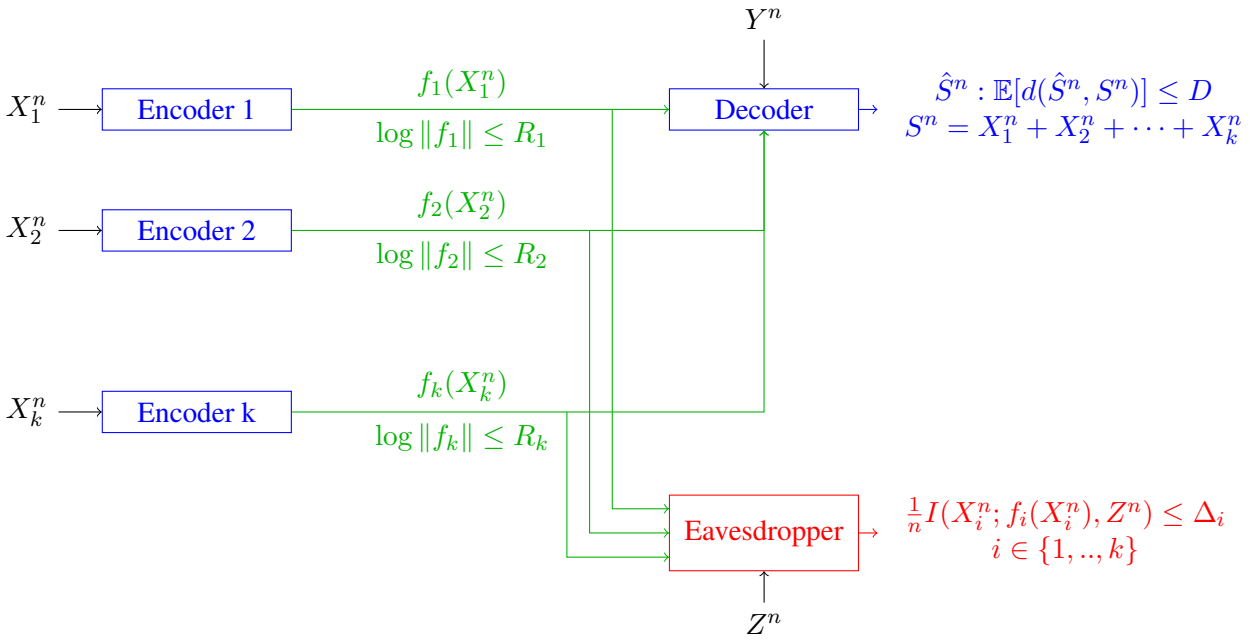


Figure 5.3: Decentralized estimation under privacy constraints.

who's listening to each channel, one at a time, and when she listens to each of the sender's channel, she is only interested in learning information about that sender's source. . Bob and Eve have their own side information correlated to encoders' observations. In a problem where it's required to design k vector quantizer optimal to such scenario, the new objective function becomes

$$\Gamma(f_1, \ldots, f_k, \lambda_1, \ldots, \lambda_k) = \frac{1}{n}\,\mathbb{E}[d(\hat{S}^n, S^n)] + \frac{1}{n}\sum_{i=1}^{k} \lambda_i I(X_i^n; f_i(X_i^n), Z^n) \tag{5.7}$$

The fundamental limits of the problem, i.e. the trade-off between the coding rates, the distortion at Bob, and the equivocations at Eve can also be studied. Suppose $k = 2$, the new problem can then be defined as follow, we are required to find a region $(R_1, R_2, D, \Delta_1, \Delta_2)$ such that:

$$\begin{aligned} R_1 &\geq \tfrac{1}{n}||f_1(X_1^n)||, \\ R_2 &\geq \tfrac{1}{n}||f_2(X_2^n)||, \\ D &\leq \tfrac{1}{n}E[(\hat{S}^n, S^n)], \text{ where } S^n = X_1^n + X_2^n \\ \Delta_1 &\geq \tfrac{1}{n}I(X_1^n; f_1(X_1^n), Z^n) \\ \Delta_2 &\geq \tfrac{1}{n}I(X_2^n; f_2(X_2^n), Z^n) \end{aligned} \tag{5.8}$$

This extension could also be interesting as a smart grid application when the purpose is to estimate the total consumption of a neighbourhood area network. Monitoring the total electric consumption in a residential area is very important for features of the smart grid as "demand response". Such applications improve the ability of electricity producers and consumers to communicate with one another and make decisions about how and when to produce and consume electrical power, allowing a change in the power consumption of an electric utility customer to better match the demand for power with the supply without threatening the privacy of the clients.

Appendices

APPENDIX A

AUXILIARY PROOFS OF CHAPTER 2

A.1 Proof of Proposition 2.1

We now prove the achievability of the region given in Proposition 2.1 for General Hypothesis Testing under both communication and Privacy constraints.

Codebook generation

Randomly pick 2^{nR_1} sequences $u^n(r_1)$ from $T_\varepsilon^n(U)$. For each joint type q_{UX}, fix a conditional type $q^\star_{V|UX}(q_{UX})$ and randomly choose a set of codewords $B^n(q_{UX})$ drawn uniformly with replacement from the marginal type class $T^n_{q_V}$ induced by q_{UX} and $q^\star_{V|UX}$. The size of the codebook $B^n(q_{UX})$ is an integer satisfying:

$$\begin{aligned}&\exp(nI(q_{UX};q^\star_{V|UX}) + (|\mathcal{U}|\,|\mathcal{V}|\,|\mathcal{X}|+2)\log(n+1))\\ &\leq |B^n(q_{UX})|\\ &\leq \exp(nI(q_{UX};q^\star_{V|UX}) + (|\mathcal{U}|\,|\mathcal{V}|\,|\mathcal{X}|+4)\log(n+1))\,.\end{aligned}$$

Encoding

Assume that a sequence x^n is produced at Alice. Look for the codeword $u^n(r_1)$ such that $x^n \in T_\varepsilon^n(X|U)(u^n)$. [OR $(u^n,x^n) \in T_\varepsilon^n(U;X)$]. As a second step, for each $u^n(r_1)$, look for the sequence $v^n(r_1,r_2)$ having a conditional type $q^\star_{V|UX}$. The encoder sends the positions of the sequence u^n and v^n denoted by r_1 and r_2 and the joint type k of (u^n,x^n). The encoder message is set to $\mathcal{J}_1\times\mathcal{J}_2\times\mathcal{K}$:

$$\begin{aligned}\mathcal{J}_1 &= \{1,2,...,M_1 \triangleq \exp[nR_1]\}\,,\\ \mathcal{J}_2 &= \{1,2,...,M_2 \triangleq \exp[nR_2]\}\,,\\ \mathcal{K} &= \{1,2,...,(n+1)^{|\mathcal{U}||\mathcal{X}|}\}\,.\end{aligned} \tag{A.1}$$

The first encoding step requires that $R_1 > I(p_X;p_{U|X})$ to succeed with a probability higher than $1-\delta$. There are two cases to be considered:

1. $\log|B^n(q_{UX})| \leq nR_2$, we can map each member of $|B^n(q_{UX})|$ to an element of $\mathcal{J}_2$ in a one-to-one manner.

2. $\log|B^n(q_{UX})| \geq nR_2$, we assign each distinct member of $|B^n(q_{UX})|$ to $\mathcal{J}_2$ uniformly at random.

Decoding

In the first decoding step, Bob looks for the unique codeword $u^n(r_1)$ such that $(u^n(r_1),y^n) \in T_\varepsilon^n(U;Y)$. The second decoding step is also composed of 2 cases:

1. $\log|B^n(q_{UX})| \leq nR_2$, decoder decodes with no error.

2. $\log|B^n(q_{UX})| \geq nR_2$, the decoder having Y as side information receives r_2 as bin index, looks for the sequence v^n having the minimum $H(v^n|u^n, y^n)$ from $Bin(r_2)$ containing sequences $v^n \in B^n(q_{UX})$ and having r_2 as index.

Let $\mathscr{C}_{r_1r_2} \subset T_\mu^n(X|u^n(r_1)v^n(r_1, r_2))$, the decoder then sets the acceptance region $\mathcal{A}_n$ for H_0 to:

$$\mathcal{A}_n = \bigcup_{r_1=1}^{M_1} \bigcup_{r_2=1}^{M_2} (\mathscr{C}_{r_1r_2} \times T_\mu^n(Y|u^n(r_1)v^n(r_1, r_2)) \,. \tag{A.2}$$

Equivocation Rate

Denote by the event E : "An error occurred during the encoding or the decoding step" and let $\delta = \varepsilon/(H(X|UZ) - R_2)$. Using [78, Lemma 2], the equivocation rate at Eve can be lower bounded as follow:

$$\begin{aligned}
\frac{1}{n}H(X^n|f(X^n)Z^n) &\geq \frac{\Pr\left\{\bar{E}\right\}}{n}H(X^n|r_1r_2kZ^n, \bar{E}) && \text{(A.3)}\\
&\geq \frac{1-\delta}{n}\big[H(X^n) - I(X^n; r_1Z^n) - I(X^n; r_2k|r_1Z^n)\big] && \text{(A.4)}\\
&\geq \frac{1-\delta}{n}\big[H(X^n) - I(X^n; U^nZ^n) - H(r_2) - H(k)\big] && \text{(A.5)}\\
&\geq \big[H(X|UZ) - R_2\big] - \varepsilon \,. && \text{(A.6)}
\end{aligned}$$

The last line follows from the i.i.d assumption and the fact that $r_2 \in \{1, \ldots, 2^{nR_2}\}$ and hence $H(r_2) \leq \log(2^{nR_2}) = nR_2$. The disappearance of the term $H(k)$ is due to the fact that $k \in \mathcal{K} = \{1, 2, \ldots, (n+1)^{|\mathcal{U}||\mathcal{X}|}\}$ and hence $H(k) \leq \log(n+1)^{|\mathcal{U}||\mathcal{X}|}$ and $\frac{1}{n}\log(n+1)^{|\mathcal{U}||\mathcal{X}|} \to 0$ as $n \to \infty$. Thus, each $\Delta \leq H(X|UZ) - R_2$ is achievable.

Error Exponent

The probability of error can be written as

$$P_e \leq P_d^n + P_t^n \,, \tag{A.7}$$

where P_d^n represents the probability of decoding errors, while P_t^n is the probability of testing errors.

Calculation of P_d^n

We now evaluate the decoding errors, i.e., the errors occurring when the wrong sequence v^n is selected, where the definitions of the decoding errors are the same as in the method of types [113]. The proof uses similar techniques to those in [114]. Let $\hat{v}^n$ be the chosen encoded sequence in our scheme, and let F be the event:

$$F = \{\exists \tilde{v}^n \neq \hat{v}^n | H(\tilde{v}^n|u^n, y^n) \leq H(\hat{v}^n|u^n, y^n)\} \,. \tag{A.8}$$

Consider the following subsets of the sequence space:

$$\begin{aligned}\xi_1 &= \left\{(u^n, v^n, x^n, y^n): \quad v^n \in T^n_{q^\star_{V|UX}}, \log|B^n(q_{UX})| \geq nR_2\right\}, \\ \xi_2 &= \left\{(u^n, v^n, x^n, y^n): \quad v^n \notin T^n_{q^\star_{V|UX}}\right\},\end{aligned} \tag{A.9}$$

where ξ_1 corresponds to binning errors while ξ_2 corresponds to the covering errors. Equivalently, we can view these error events as properties of the joint type, so we can define:

$$\begin{aligned}\mathcal{D}_1 &= \left\{q_{VXY|U}: \quad q_{V|UX} = q^\star_{V|UX}, \log|B^n(q_{UX})| \geq nR_2\right\}, \\ \mathcal{D}_2 &= \left\{q_{VXY|U}: \quad q_{V|UX} \neq q^\star_{V|UX}\right\}.\end{aligned} \tag{A.10}$$

In order to calculate this error exponent, we need to use the following lemmas. In this appendix, we evaluate the error exponent of the decoding errors, i.e., we prove:

$$E_2 = \inf_{q_X} \sup_{q^\star_{V|UX}} \inf_{q_Y} \inf_{\substack{q_{VXY|U} \\ q_{V|UX}=q^\star_{V|UX}}} G[q_{VXY|U}, R_2].$$

In order to calculate this error exponent, we need to use the following lemmas.

Lemma A.1 *For all strings u^n, v^n, x^n such that $v^n \in T^n_{q_V}$*

$$\Pr\{v^n \in B^n(q_{UX})\}$$

$$\leq (n+1)^{|\mathcal{V}|(1+|\mathcal{U}||\mathcal{X}|)+4} \times \exp[n(I(q_{UX}; q^\star_{V|UX}) - H(q_V))].$$

Proof A.1 *During the construction process of the code-book $B^n(q_{UX})$, each of the codewords is chosen with replacement from the set $T^n_{q_V}$. We make $|B^n(q_{UX})|$ such choices with a probability $\left|T^n_{q_V}\right|^{-1}$. Using the fact that $\left|T^n_{q_X}\right| \geq (n+1)^{-|\mathcal{X}|}\exp[nH(q_X)]$, and the bounds given on $|B^n(q_{UX})|$, the lemma can be proved.*

Lemma A.2 *For any pair of strings x^n and y^n, let:*

$$\mathcal{F}(x^n, y^n) = \{\tilde{x}^n, \tilde{y}^n | H(\tilde{x}^n, \tilde{y}^n) \leq H(x^n, y^n)\}, \tag{A.11}$$

then

$$|\mathcal{F}(x^n, y^n)| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp[nH(x^n, y^n)]. \tag{A.12}$$

Proof A.2 *For all q_{XY} such that $H(q_{XY}) \leq H(x^n, y^n)$, we have:*

$$\begin{aligned}|\mathcal{F}(x^n, y^n)| &= \sum_{q_{XY}} \left|T^n_{q_{XY}}\right| \\ &\leq \sum_{q_{XY}} \exp[nH(q_{XY})] \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp[nH(x^n, y^n)].\end{aligned} \tag{A.13}$$

Lemma A.3 *For any pair of strings x^n and y^n, let:*

$$\mathcal{F}(x^n|y^n) = \{\tilde{x}^n | H(\tilde{x}^n|y^n) \leq H(x^n|y^n)\}, \tag{A.14}$$

then

$$|\mathcal{F}(x^n|y^n)| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp[nH(x^n|y^n)] \,. \tag{A.15}$$

Proof A.3 *The proof of this lemma is very identical to that of lemma A.2 and will thus be ommited.*

Lemma A.4 *Suppose that* $(u^n, v^n, x^n, y^n) \in (\xi_2)^c$*, i.e., that* $v^n \in T^n_{q^\star_{V|UX}}$*. Then,*

$$\Pr\{V^n = v^n, X^n = x^n, Y^n = y^n | U^n = u^n\} \leq p^n_{XY|U}(x^n, y^n|u^n) \frac{1}{\left|T^n_{q^\star_{V|UX}}\right|} \,. \tag{A.16}$$

Proof A.4 *Define the following event* $G = \left\{\exists \tilde{v}^n \in B^n(q_{UX}) : \tilde{v}^n \in T^n_{q^\star_{V|UX}}\right\}$*. For* (u^n, v^n, x^n, y^n) *in this lemma,* $\{U^n = u^n, V^n = v^n, X^n = x^n, Y^n = y^n\}$ *implies that the event G has occurred:*

$\Pr\{V^n = v^n, X^n = x^n, Y^n = y^n | U^n = u^n\}$

$$\begin{aligned}
&= \Pr\{V^n = v^n, X^n = x^n, Y^n = y^n, G | U^n = u^n\} && \text{(A.17)}\\
&= \Pr\{X^n = x^n, Y^n = y^n | U^n = u^n\} \\
&\quad \times \Pr\{G | U^n = u^n, X^n = x^n, Y^n = y^n\} \\
&\quad \times \Pr\{V^n = v^n | U^n = u^n, X^n = x^n, Y^n = y^n, G\} && \text{(A.18)}\\
&\leq p^n_{XY|U}(x^n, y^n|u^n) \\
&\quad \times \Pr\{V^n = v^n | U^n = u^n, X^n = x^n, Y^n = y^n, G\} && \text{(A.19)}\\
&= p^n_{XY|U}(x^n, y^n|u^n) \frac{1}{\left|T^n_{q^\star_{V|UX}}\right|} && \text{(A.20)}
\end{aligned}$$

where in the last line, we used the fact that V^n *is uniformly distributed over* $T^n_{q^\star_{V|UX}}$*.*

Lemma A.5 *Suppose that the sequence* (u^n, v^n, x^n, y^n) *is in* ξ_2*. Then*

$$\Pr\{V^n = v^n, X^n = x^n, Y^n = y^n | U^n = u^n\} \leq \exp[-(n+1)^2] \tag{A.21}$$

Proof A.5 *For* (u^n, v^n, x^n, y^n) *in this lemma,* $\{U^n = u^n, V^n = v^n, X^n = x^n, Y^n = y^n\}$ *implies that the event* G^c *has occurred. Thus*

$\Pr\{V^n = v^n, X^n = x^n, Y^n = y^n | U^n = u^n\}$

$$\begin{aligned}
&= \Pr\{V^n = v^n, X^n = x^n, Y^n = y^n, G^c | U^n = u^n\} && \text{(A.22)}\\
&= p^n_{X|U}(x^n|u^n) \Pr\{G^c | U^n = u^n, X^n = x^n\} \\
&\quad \times \Pr\{Y^n = y^n | U^n = u^n, X^n = x^n, G^c\} \\
&\quad \times \Pr\left\{V^n = v^n | U^n = u^n, X^n = x^n, \bar{Y}^n = y^n, G^c\right\} && \text{(A.23)}\\
&\leq \Pr\{G^c | U^n = u^n, X^n = x^n\} && \text{(A.24)}
\end{aligned}$$

$\Pr\{G^c|U^n = u^n, X^n = x^n\}$ *is the probability that there is no* $\tilde{v}^n \in B^n(q_{UX})$ *such that* $\tilde{v}^n \in T^n_{q^\star_{V|UX}}$*. Let*

$m_2 = |B^n(q_{UX})|$ and $B^n(q_{UX})[r_2]$ be the r_2-th codeword in the set $B^n(q_{UX})$. Then

$$\begin{aligned}\Pr\{G^c|U^n = u^n, X^n = x^n\} &= \prod_{r_2=1}^{m_2} \Pr\left\{B^n(q_{UX})[r_2] \notin T^n_{q^\star_{V|UX}}\right\} \\ &= \prod_{r_2=1}^{m_2} [1 - \Pr\left\{B^n(q_{UX})[r_2] \in T^n_{q^\star_{V|UX}}\right\}] \\ &= (1 - \frac{\left|T^n_{q^\star_{V|UX}}\right|}{|T^n_{q_V}|})^{m_2} \\ &\leq \exp[-\frac{\left|T^n_{q^\star_{V|UX}}\right|}{|T^n_{q_V}|} m_2]\end{aligned} \tag{A.25}$$

The last line follows by applying the inequality $(1-t)^m \leq \exp(-tm)$. By using the following properties:

$$\left|T^n_{q_V}\right| \leq \exp[nH(q_V)] \tag{A.26}$$

$$\left|T^n_{q_{V|UX}}\right| \geq (n+1)^{-|\mathcal{U}||\mathcal{V}||\mathcal{X}|} \exp(nH(q_{V|UX}|q_{UX})) \tag{A.27}$$

That allows us to write:

$$\begin{aligned}\Pr\{G^c|U^n = u^n, X^n = x^n\} &\leq \exp\left[-m_2(n+1)^{-|\mathcal{U}||\mathcal{V}||\mathcal{X}|} \exp[-nI(q^\star_{V|UX}; q_{UX})]\right] \\ &\leq \exp[-(n+1)^2]\end{aligned}$$

where the last step follows from the fact the $m_2 \geq \exp(nI(q_{UX}; q^\star_{V|UX}) + (|\mathcal{U}|\,|\mathcal{V}|\,|\mathcal{X}| + 2)\log(n+1))$

Lemma A.6 *Let $(u^n, v^n, x^n, y^n) \in \xi_1$. Then*

1. *If $\log|B^n(q_{UX})| \leq nR_2$ then*

   $\Pr\{F|U^n = u^n, V^n = v^n, X^n = x^n, Y^n = y^n\} = 0$

2. *If $\log|B^n(q_{UX})| \geq nR_2$*

   $\Pr\{F|U^n = u^n, V^n = v^n, X^n = x^n, Y^n = y^n\}$

   $$\leq \exp[-n((R_2 - I(q_{UX}; q^\star_{V|UX}) + I(q_{UY}; q_{V|UY}) - \delta^n_b)^+)]$$

where $\delta^n_b = \frac{1}{n}\log(n+1)^{|\mathcal{V}|(1+|\mathcal{U}||\mathcal{X}|+|\mathcal{U}||\mathcal{Y}|)+4}$

Proof A.6 *If q_{UX} is such that $\log|B^n(q_{UX})| \leq nR_2$, then the decoder decodes with no error. On the other hand, if $\log|B^n(q_{UX})| \geq nR_2$, then for the given sequence (u^n, v^n, x^n, y^n), the probability $Pr\{F|U^n = u^n, V^n = v^n, X^n = x^n, Y^n = y^n\}$ is bounded as in (A.28).*

(a) follows from lemma A.1, and (b) follows from lemma A.3.

Using lemma A.4, lemma A.5 and lemma A.6, the probability of error P^n_d can thus be bounded as in (A.29). The last term can be omitted and thus by using inequality (A.30) and the following inequalities:

$$\begin{aligned}\left|T_{q_{V|UX}}\right| &\geq (n+1)^{|\mathcal{U}||\mathcal{V}||\mathcal{X}|} \exp[nH(q_{V|UX}|q_{UX})] \\ |T_{q_{UVXY}}| &\leq \exp[nH(q_{UVXY})] \leq (n+1)^{|\mathcal{U}||\mathcal{V}||\mathcal{X}||\mathcal{Y}|}\end{aligned}$$

we can write:

$$P_d^n \leq \sum_{q_{X|U}} \sum_{q_{Y|U}} \sum_{q_{VXY|U} \in \mathcal{D}_1} \min \{D(q_{VXY|U}||p_{XY|U} q_{V|UX}|q_U); D(q_{VXY|U}||p_{\bar{X}\bar{Y}|U} q_{V|UX}|q_U)\} + [R_2 - I(q^\star_{V|UX}; q_{UX}) + I(q_{V|UY}; q_{UY}) - \delta_b^n]^+$$

The summations can be upper bounded by maximizing over the types and optimizing over the choice of the channel $q^\star_{V|UX}$.

$$P_d^n \leq |\mathcal{P}^n(\mathcal{X})|\,|\mathcal{P}^n(\mathcal{Y})|\,|\mathcal{P}^n(\mathcal{V}\times\mathcal{X}\times\mathcal{Y})| \exp\big(-n(\min_{q_{X|U}} \max_{q^\star_{V|UX}} \min_{q_{Y|U}} \min_{\substack{q_{VXY|U} \\ q_{V|UX}=q^\star_{V|UX}}} G^n[q_{VXY|U}, p_{XY|U}, n, R_2])\big)$$

$G^n[q_{VXY|U}, p_{XY|U}, n, R_2]$ is as defined by (A.49).

$$\begin{aligned}
&\Pr\{F|U^n=u^n, V^n=v^n, X^n=x^n, Y^n=y^n\} \\
&\leq \sum_{\substack{\tilde{v}^n \in \mathcal{F}(v^n|u^n,y^n) \\ \tilde{v}\neq v}} \Pr\{\tilde{v}^n \in B^n(q_{UX}), m_2(\tilde{v}^n)=m_2(v^n)|U^n=u^n, V^n=v^n, X^n=x^n, Y^n=y^n\} \\
&\leq \sum_{\substack{\tilde{v}^n \in \mathcal{F}(v^n|u^n,y^n) \\ \tilde{v}\neq v}} \Pr\{\tilde{v}^n \in B^n(q_{UX})|U^n=u^n, X^n=x^n, Y^n=y^n\} \\
&\quad \times \Pr\{m_2(\tilde{v}^n)=m_2(v^n)|\tilde{v}^n \in B^n(q_{UX})\} \\
&\overset{(a)}{\leq} \sum_{\substack{\tilde{v}^n \in \mathcal{F}(v^n|u^n,y^n) \\ \tilde{v}\neq v}} (n+1)^{|\mathcal{V}|(1+|\mathcal{U}||\mathcal{X}|)+4} \exp[n(I(q_{UX}; q^\star_{V|UX}) - H(q_V)]\tfrac{1}{M_2} \\
&\overset{(b)}{\leq} (n+1)^{|\mathcal{V}|(1+|\mathcal{U}||\mathcal{X}|+|\mathcal{U}||\mathcal{Y}|)+4} \exp[nH(q_{V|UY}|q_{UY})] \exp[n(I(q_{UX}; q^\star_{V|UX}) - H(q_V)]\tfrac{1}{M_2} \\
&\leq (n+1)^{|\mathcal{V}|(1+|\mathcal{U}||\mathcal{X}|+|\mathcal{U}||\mathcal{Y}|)+4} \exp[-n(R_2 - I(q_{UX}; q^\star_{V|UX}) + I(q_{UY}; q_{V|UY}))] \\
&= \exp[-n((R_2 - I(q_{UX}; q^\star_{V|UX}) + I(q_{UY}; q_{V|UY}) - \delta_b^n)^+)]
\end{aligned} \tag{A.28}$$

$$\begin{aligned}
P_d^n &\leq \sum_{\xi_1} \Pr\{F|U^n=u^n, V^n=v^n, X^n=x^n, Y^n=y^n\} \\
&\quad \times \Pr\{V^n=v^n, X^n=x^n, Y^n=y^n|U^n=u^n\} \\
&+ \sum_{\xi_2} \Pr\{V^n=v^n, X^n=x^n, Y^n=y^n|U^n=u^n\} \\
&\leq \sum_{\xi_1} \Big[\exp[-n((R_2 - I(q_{UX}; q^\star_{V|UX}) + I(q_{UY}; q_{V|UY}) - \delta_b^n)^+)] \times p^n_{XY|U}(x^n, y^n|u^n) \tfrac{1}{\left|T_{q^\star_{V|UX}}\right|}\Big] \\
&+ \sum_{\xi_2} \exp[-(n+1)^2] \\
&= \sum_{q_{X|U}} \sum_{q_{Y|U}} \Big[\Big(\sum_{q_{VXY|U} \in \mathcal{D}_1} \sum_{(v^n,x^n,y^n) \in T_{q_{VXY|U}}} p^n_{XY}(x^n, y^n) \tfrac{1}{\left|T_{q^\star_{V|UX}}\right|} \\
&\times \exp[-n((R_2 - I(q_{UX}; q^\star_{V|UX}) + I(q_{UY}; q_{V|UY}) - \delta_b^n)^+)]\Big) \\
&+ \sum_{q_{VXY|U} \in \mathcal{D}_2} \sum_{(v^n,x^n,y^n) \in T_{q_{VXY|U}}} \exp[-(n+1)^2]\Big]
\end{aligned} \tag{A.29}$$

$$\begin{aligned}
p^n_{XY|U}(x^n, y^n|u^n) \leq \max\Big\{ &\exp[-n(D(q_{XY|U}||p_{XY|U}|q_U) + H(q_{XY|U}|q_U))]; \\
&\exp[-n(D(q_{XY|U}||p_{\bar{X}\bar{Y}|U}|q_U) + H(q_{XY|U}|q_U))]\Big\}
\end{aligned} \tag{A.30}$$

The size of the set of all empirical distributions of X and of length n can be bounded by:

$$|\mathcal{P}^n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|} \tag{A.31}$$

Similarly, the sets $|\mathcal{P}^n(\mathcal{X})|$, $|\mathcal{P}^n(\mathcal{V}\times\mathcal{X}\times\mathcal{Y})|$ can be bounded in an identical manner. Thus the cardinalities can be absorbed inside the exponent and thus become insignificant for a sufficiently large n. Using continuity arguments as used in [114, Lemma 14], the probability of error will be written as: $\liminf_{n\to\infty} -\frac{1}{n}\log P_d^{(n)}$

$$\begin{aligned}
&\geq \liminf_{n\to\infty} \min_{q_{X|U}} \max_{q^\star_{V|UX}} \min_{q_{Y|U}} \inf_{\substack{q_{VXY|U} \\ q_{V|UX}=q^\star_{V|UX}}} \\
&\qquad G^n[q_{VXY|U}, p_{XY|U}, R_2, n] \\
&\geq \inf_{q_X} \sup_{q^\star_{V|UX}} \inf_{q_Y} \inf_{\substack{q_{VXY|U} \\ q_{V|UX}=q^\star_{V|UX}}} G[q_{VXY|U}, R_2]
\end{aligned}$$

where $G[q_{VXY|U}, R_2]$ is as defined as in (A.48). This allows us to write the error exponent of the decoding errors as:

$$E_2 = \inf_{q_{X|U}} \sup_{q^\star_{V|UX}} \inf_{q_{Y|U}} \inf_{\substack{q_{VXY|U} \\ q_{V|UX}=q^\star_{V|UX}}} G[q_{VXY|U}, R_2]\ , \tag{A.32}$$

We try to eliminate R_2 using Fourier Motzkin elimination, we have:

$$\begin{aligned}
R &\geq I(p_X; p_{U|X}) + R_2 && \text{(A.33)}\\
E &\leq E_1 && \text{(A.34)}\\
E &\leq E_2(R_2) && \text{(A.35)}\\
\Delta &\leq H(p_{X|UZ}|p_{UZ}) - R_2 && \text{(A.36)}\\
R_2 &\geq 0 && \text{(A.37)}
\end{aligned}$$

Using the fact that E_2 is an increasing function in R_2 we can write:

$$\begin{aligned}
R_2 &\leq R - I(p_X; p_{U|X}) && \text{(A.38)}\\
E &\leq E_1 && \text{(A.39)}\\
R_2 &\geq E_2^{-1}(E) && \text{(A.40)}\\
R_2 &\leq H(p_{X|UZ}|p_{UZ}) - \Delta && \text{(A.41)}\\
R_2 &\geq 0 && \text{(A.42)}
\end{aligned}$$

Using the obtained upper and lower bounds on R_2 we can write:

$$\begin{aligned}
R &\geq I(p_X; p_{U|X}) && \text{(A.43)}\\
\Delta &\leq H(p_{X|UZ}|p_{UZ})) - R_2 && \text{(A.44)}\\
E &\leq E_1 && \text{(A.45)}\\
E &\leq E_2(R - I(p_X; p_{U|X})) && \text{(A.46)}\\
E &\leq E_2(H(p_{X|UZ}|p_{UZ}) - \Delta) && \text{(A.47)}
\end{aligned}$$

In other words we can thus say that the achievable region can be now written in the following way:

$$\begin{aligned} R &\geq I(p_X; p_{U|X}) && \text{(A.50)} \\ \Delta &\leq H(p_{X|UZ}|p_{UZ}) && \text{(A.51)} \\ E &\leq \min\{E_1; E_2(R - I(p_X; p_{U|X})); \\ &\qquad E_2(H(p_{X|UZ}|p_{UZ}) - \Delta)\} && \text{(A.52)} \end{aligned}$$

where $E_2(R - I(p_X; p_{U|X}))$ and $E_2(H(p_{X|UZ}|p_{UZ}) - \Delta)$ are defined in proposition 2.1.

Calculation of P_t^n

The exponent of testing is studied under the assumption that no decoding errors have occurred, i.e., under the assumption that the correct codeword has been found in the bin. In order to bound P_t^n, we can proceed similar to the proof done to evaluate the lower bound of the error exponent in [60]. The definition of the acceptance region in this case changes to include the two layers as seen in (A.2). Then, following similar steps with the assumption that $M_1 + M_2 = \exp[n(I(UV; X) + \eta)]$, it is easy to check that [60]:

$$P_t^n \leq (n+1)^{|\mathcal{U}||\mathcal{V}||\mathcal{X}||\mathcal{Y}|} \max_{q_{UVXY}} \exp[-n(d(UVXY) - 2\mu - \eta)] \,,$$

where:

$$d(UVXY) = D(p_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}}||p_{\bar{U}\bar{V}\bar{X}\bar{Y}}) + \delta' \,. \tag{A.53}$$

Therefore, the error exponent E_1 is now shown, which concludes the proof of Proposition 2.1.

A.2 Proof of Proposition 2.2

We first show the achievability and then the converse of the region provided in proposition 2.2.

$G[q_{VXY|U}, R_2] =$

$$\begin{cases} \min\{D(q_{VXY|U}||p_{XY|U}q_{V|UX}|q_U); D(q_{VXY|U}||p_{\bar{X}\bar{Y}|U}q_{V|UX}|q_U)\} \\ +[R_2 - I(q_{UX}; q_{V|UX}) + I(q_{UY}; q_{V|UY})]^+ & R_2 < I(q_{UX}; q_{V|UX}) \\ \infty & \text{else.} \end{cases} \tag{A.48}$$

$G^n[q_{VXY|U}, p_{XY|U}, R_2, n] =$

$$\begin{cases} \min\{D(q_{VXY|U}||p_{XY|U}q_{V|UX}|q_U); D(q_{VXY|U}||p_{\bar{X}\bar{Y}|U}q_{V|UX}|q_U)\} \\ +[R_2 - I(q^\star_{V|UX}; q_{UX}) + I(q_{V|UY}; q_{UY}) - \delta_b^n]^+ & R_2 < I(q_{V|UX}, q_{UX}) \\ \infty & \text{else.} \end{cases} \tag{A.49}$$

A.2.1 Direct Part

The proof of the direct part can be derived from the region proposed in Proposition 2.1. Taking $R_2 = 0$ is sufficient to prove the achievability of the rate and the equivocation. Our main concern lies in proving the error exponent $D(UY||\bar{U}\bar{Y})$ is achievable. In this case, the new codebook is generated only with one layer U without binning and thus all decoding errors are vanished and only detecting errors are taken into considerations. Hence, it is required to prove show:

$$D(\tilde{U}\tilde{X}\tilde{Y}||\bar{U}\bar{X}\bar{Y}) \geq D(\tilde{U}\tilde{Y}||\bar{U}\bar{Y}) \tag{A.54}$$

$$\stackrel{(a)}{=} D(UY||\bar{U}\bar{Y}) \,, \tag{A.55}$$

where (a) follows from the fact that $p_{\tilde{U}\tilde{V}\tilde{Y}} = p_{UVY}$ and hence $p_{\tilde{U}\tilde{Y}} = p_{UY}$. This yields the achievability of the error exponent, i.e., $E \leq D(UY||\bar{U}\bar{Y})$.

A.2.2 Converse Part

$J = f(X^n)$ being the message sent by the encoder "Alice" to the decoder "Bob".The following Markov chain holds for each subset $\mathcal{G} \subset \{1, \ldots, n\}$:

$$(J, X_{\mathcal{G}}, Y_{\mathcal{G}}, \bar{Y}_{\mathcal{G}}) \multimap X_{\mathcal{G}^c} \multimap Y_{\mathcal{G}^c} \multimap \bar{Y}_{\mathcal{G}^c} \tag{A.56}$$

By setting $U_i = JY^{i-1}\bar{Y}^{i-1}\bar{Y}_{i+1}^n$, the Markov chain $U_i - X_i - Y_i - \bar{Y}_i$ is satisfied.

Error Exponent

$$\begin{aligned}
nE &\leq D(p_{JY^n}||p_{J\bar{Y}^n}) && \text{(A.57)}\\
&= D(p_{Y^n}p_{J|Y^n}||p_{\bar{Y}^n}p_{J|\bar{Y}^n}) && \text{(A.58)}\\
&= nD(p_Y||p_{\bar{Y}}) + D(p_{J|Y^n}||p_{J|\bar{Y}^n}) && \text{(A.59)}\\
&\stackrel{(a)}{=} D(p_{J|Y^n}p_{Y^n|\bar{Y}^n}||p_{J|\bar{Y}^n}p_{Y^n|\bar{Y}^n}) && \text{(A.60)}\\
&= D(p_{JY^n|\bar{Y}^n}||p_{J|\bar{Y}^n}p_{Y^n|\bar{Y}^n}) && \text{(A.61)}\\
&= I(J;Y^n) - I(J;\bar{Y}^n) && \text{(A.62)}\\
&\stackrel{(b)}{=} \sum_{i=1}^{n} I(JY^{i-1};Y_i) - I(J\bar{Y}^{i-1};\bar{Y}_i) && \text{(A.63)}\\
&\stackrel{(c)}{=} \sum_{i=1}^{n} \left[I(JY^{i-1};Y_i) - I(J\bar{Y}^{i-1};\bar{Y}_i) + I(\bar{Y}_{i+1}^n;Y_i|JY^{i-1}) - I(Y^{i-1};\bar{Y}_i|J\bar{Y}_{i+1}^n)\right] && \text{(A.64)}\\
&\stackrel{(d)}{=} \sum_{i=1}^{n} I(JY^{i-1}\bar{Y}^{i-1}\bar{Y}_{i+1}^n;Y_i) - I(JY^{i-1}\bar{Y}^{i-1}\bar{Y}_{i+1}^n;\bar{Y}_i) && \text{(A.65)}\\
&= \sum_{i=1}^{n} I(U_i;Y_i) - I(U_i;\bar{Y}_i) && \text{(A.66)}
\end{aligned}$$

Where (a) follows from the fact that the marginal distributions p_Y and $p_{\bar{Y}}$ are equal under both hypotheses. (b) is due to chain rule and (c) is due to Csiszár and Körner's Equality [**?**, Lemma 7]. (d) is due to the markov chain

defined in (A.56).

Define a time-sharing RV Q uniformly distributed over $\{1, 2, \ldots, n\}$ and $X = X_Q, Y = Y_Q, \bar{Y} = \bar{Y}_Q, U = U_Q$. Thus, we can say that

$$\begin{aligned} E &\leq \frac{1}{n} I(U_Q; Y_Q|Q=i) - I(U_Q; \bar{Y}_Q|Q=i) && \text{(A.67)} \\ &= I(U;Y) - I(U;\bar{Y}) && \text{(A.68)} \\ &= D(UY||U\bar{Y}) \,, && \text{(A.69)} \end{aligned}$$

Equivocation Rate

$$\begin{aligned} n\Delta &\leq H(X^n|J\bar{Y}^n) && \text{(A.70)} \\ &\overset{(a)}{=} \sum_{i=1}^{n} H(X_i|JX^{i-1}\bar{Y}_i\bar{Y}^{i-1}\bar{Y}_{i+1}^n) && \text{(A.71)} \\ &\leq \sum_{i=1}^{n} H(X_i|JX^{i-1}\bar{Y}_i\bar{Y}_{i+1}^n) && \text{(A.72)} \\ &\overset{(b)}{=} \sum_{i=1}^{n} H(X_i|JX^{i-1}\bar{Y}_i\bar{Y}_{i+1}^n Y^{i-1}\bar{Y}^{i-1}) && \text{(A.73)} \\ &\leq \sum_{i=1}^{n} H(X_i|J\bar{Y}_i\bar{Y}_{i+1}^n Y^{i-1}\bar{Y}^{i-1}) && \text{(A.74)} \\ &= H(X_i|U_i\bar{Y}_i) && \text{(A.75)} \end{aligned}$$

Where (a) is due to chain rule and (b) uses the similar Markov chain used before. From the same definition of the time-sharing random variable, we can conclude that $\Delta \leq H(X|U\bar{Y})$. This concludes the proof of proposition 2.2.

A.3 Proof of Proposition 2.3

In this proof, we show the achievability and the converse proof of the rate-error-equivocation region proposed in Proposition 2.3

A.3.1 Proof of the Achievability Part

The proof of the direct part can be derived from the region proposed in Proposition 2.1. Taking $R_2 = 0$ is sufficient to prove the achievability of the rate and the equivocation. Our main concern lies in proving the error exponent $I(U;Y)$ is achievable. In this case, the new codebook is generated only with one layer U without binning and thus all decoding errors are vanished and only detecting errors are taken into considerations. In the encoding step, The encoder looks for a codeword $u^n(r)$ such that $(u^n(r), x^n) \in T_\varepsilon^\tau(UX)$ and sends the message $f(X^n) = r$ on the error free channel while the detector has access to $u^n(r)$ and the analog observation y^n. The definition of the acceptance region becomes:

$$\mathcal{A}_n = \bigcup_{r \in [1:2^{nR}]} \{u^n(r)\} \times T_\varepsilon^\tau(Y|u^n(r)) \subset T_\varepsilon^\tau(UY) \,. \tag{A.76}$$

It is required to prove show:

$$D(\tilde{U}\tilde{X}\tilde{Y}||\bar{U}\bar{X}\bar{Y}) \geq D(\tilde{U}\tilde{Y}||\bar{U}\bar{Y}) \tag{A.77}$$

$$\stackrel{(a)}{=} D(UY||\bar{U}\bar{Y}) \tag{A.78}$$

$$\stackrel{(b)}{=} I(U;Y)\,, \tag{A.79}$$

where (a) follows from the fact that $p_{\tilde{U}\tilde{V}\tilde{Y}} = p_{UVY}$ and hence $p_{\tilde{U}\tilde{Y}} = p_{UY}$, and (b) follows from the fact that $\bar{U}$ and $\bar{Y}$ are independent, $p_U = p_{\bar{U}}$, and $p_Y = p_{\bar{Y}}$. This yields the achievability of the error exponent, i.e., $E \leq I(U;Y)$. This concludes the proof of the achievability of Proposition 2.3.

A.3.2 Proof of the Converse Part

Let $J = f(X^n)$ be the message, the following Markov chain holds: $(J, X^{i-1}, X_{i+1}^n, Y^{i-1}, Y_{i+1}^n, Z^{i-1}, Z_{i+1}^n) \multimap X_i \multimap (Y_i, Z_i)$ for all $i = \{1, \ldots, n\}$. From the memoryless assumption of the sources, we have: $J \multimap X^{i-1} \multimap Y^{i-1}$ and thus by setting $U_i = (J, Y^{i-1})$ then $U_i \multimap X_i \multimap (Y_i, Z_i)$.

Coding rate

$$\begin{aligned}
nR &\geq I(J;X^n) && \text{(A.80)}\\
&= \sum_{i=1}^{n} I(J;X_i|X^{i-1}) && \text{(A.81)}\\
&\stackrel{(a)}{=} \sum_{i=1}^{n} \left[H(J|X^{i-1}Y^{i-1}) - H(J|X_iX^{i-1}Y^{i-1})\right] && \text{(A.82)}\\
&= \sum_{i=1}^{n} I(J;X_i|X^{i-1}Y^{i-1}) && \text{(A.83)}\\
&\stackrel{(b)}{=} \sum_{i=1}^{n} I(JX^{i-1}Y^{i-1};X_i) && \text{(A.84)}\\
&\geq \sum_{i=1}^{n} I(U_i;X_i)\,, && \text{(A.85)}
\end{aligned}$$

where (a) follows from the fact that $J \multimap X^n \multimap Y^n$ and (b) follows from the memoryless assumption of the sources. Define a time-sharing RV Q uniformly distributed over $\{1, 2, \ldots, n\}$ and $X = X_Q, U = U_Q$. Then, $U \multimap X \multimap (Y, Z)$ and

$$R \geq \frac{1}{n}\sum_{i=1}^{n} I(U_Q;X_Q|Q=i) = I(U;X)\,. \tag{A.86}$$

Equivocation rate

$$\begin{aligned}
\Delta \leq \frac{1}{n}H(X^n|JZ^n) &= \frac{1}{n}\sum_{i=1}^{n}H(X_i|JX^{i-1}Z^{i-1}Z_{i+1}^{n}Z_i) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}H(X_i|JX^{i-1}Z_i) && \text{(A.87)} \\
&\overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n}H(X_i|JX^{i-1}Y^{i-1}Z_i) && \text{(A.88)} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}H(X_i|JY^{i-1}Z_i) && \text{(A.89)} \\
&= \frac{1}{n}\sum_{i=1}^{n}H(X_i|U_iZ_i) && \text{(A.90)} \\
&= H(X|UZ)\,, && \text{(A.91)}
\end{aligned}$$

where (a) follows from the fact that $(J, X_i, Z_i) \multimap X^{i-1} \multimap Y^{i-1}$ and the last step from the time-sharing argument.

Error exponent

$$\begin{aligned}
E \leq \frac{1}{n}I(J;Y^n) &= \frac{1}{n}\sum_{i=1}^{n}I(J;Y_i|Y^{i-1}) && \text{(A.92)} \\
&= \frac{1}{n}\sum_{i=1}^{n}I(U_i;Y_i) && \text{(A.93)} \\
&= \sum_{i=1}^{n}\frac{1}{n}I(U_Q;Y_Q|Q=i) && \text{(A.94)} \\
&= I(U;Y)\,. && \text{(A.95)}
\end{aligned}$$

It is worth mentioning that another optimal choice is $U_i = (J, X^{i-1})$. This concludes the proof of the converse of Proposition 2.3.

A.4 Proof of Proposition 2.5

The achievability part is a direct application of Proposition 2.3. Define the auxiliary RV U obtained as the output of a Degraded Binary Symmetric Channel with input X and crossover probability $\alpha \in [0, 1]$, as depicted in Fig. A.1. The rate inequality reads:

$$R \geq I(U;X) = 1 - H_2(\alpha) \tag{A.96}$$

and the exponent inequality reads:

$$E \leq I(U;Y) = (1-\epsilon)\big(1 - H_2(\alpha)\big)\,. \tag{A.97}$$

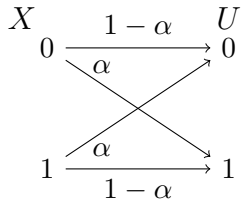


Figure A.1: Binary auxiliary RV.

The equivocation is given by

$$\Delta \leq H(X|UZ) = H_2(\alpha) + H_2(p) - H_2(\alpha \star p) \, . \tag{A.98}$$

To show the converse, i.e., the optimality of the specific choice of our previous auxiliary RV U, we need to show that there exists an $\alpha \in [0, \frac{1}{2}]$ such that:

$$\begin{aligned} I(U;X) &\geq 1 - H_2(\alpha) \, , && \text{(A.99)} \\ I(U;Y) &\leq (1-\epsilon)\big(1 - H_2(\alpha)\big) \, , && \text{(A.100)} \\ H(X|UZ) &\leq H_2(p) + H_2(\alpha) - H_2(\alpha \star p) \, . && \text{(A.101)} \end{aligned}$$

First, we have that $I(U;X) = 1 - H(X|U)$. By noticing that the binary entropy function is a continuous mapping from $[0, 1/2]$ to $[0, 1]$, and since $0 \leq H(X|U) \leq H(X) = 1$, we have that there exists $\alpha \in [0, \frac{1}{2}]$ such that $H(X|U) = H_2(\alpha)$. From this, we can write:

$$I(U;X) = 1 - H_2(\alpha) \, . \tag{A.102}$$

By using the same α, i.e., $H_2(\alpha) = H(X|U)$, we obtain:

$$\begin{aligned} I(U;Y) &= (1-\epsilon)(1 - H(X|U)) && \text{(A.103)} \\ &= (1-\epsilon)\big(1 - H_2(\alpha)\big) \, . && \text{(A.104)} \end{aligned}$$

Similarly,

$$\begin{aligned} H(X|UZ) &= H(X|U) - I(X;Z|U) && \text{(A.105)} \\ &= H(X|U) + H(Z|X) - H(Z|U) && \text{(A.106)} \\ &= H_2(\alpha) + H_2(p) - H(Z|U) \, . && \text{(A.107)} \end{aligned}$$

Since Z is the output of a BSC with input X and and crossover probability p, then Z can be written as $Z = X \oplus \tilde{Z}$ with $\tilde{Z} \backsim \text{Bern}(p)$. Using *Mrs. Gerber's Lemma* [37], we can write:

$$H(Z|U) \geq H_2(H_2^{-1}(H(X|U)) \star p) \, . \tag{A.108}$$

Since α is chosen such that $H_2(\alpha) = H(X|U)$, then $\alpha = H_2^{-1}(H(X|U))$ and hence we can state that:

$$H(X|UZ) \leq H_2(\alpha) + H_2(p) - H_2(\alpha \star p) \, . \tag{A.109}$$

A.5 Proof of Proposition 2.6

We restrict our attention to an auxiliary RV U obtained as the output of AWGN channel with input X and noise power $(1-\rho_U^2)$ with $0 < \rho_U < 1$. The rate must satisfy

$$\begin{aligned} R \geq I(U;X) &= \frac{1}{2}\log 2\pi e - h(\rho_U X + N_U|X) && \text{(A.110)} \\ &= \frac{1}{2}\log 2\pi e - h(N_U) && \text{(A.111)} \\ &= \frac{1}{2}\log\left(\frac{1}{1-{\rho_U}^2}\right) . && \text{(A.112)} \end{aligned}$$

The exponent satisfies:

$$E \leq I(U;Y) = \frac{1}{2}\log\left(\frac{1}{1-[\mathrm{Cov}(U,Y)]^2}\right) , \tag{A.113}$$

where

$$\begin{aligned} \mathrm{Cov}(U,Y) &= \mathbb{E}(UY) - \mathbb{E}(U)\mathbb{E}(Y) && \text{(A.114)} \\ &\overset{(a)}{=} \rho_U\rho_Y\mathbb{E}(X^2) + \rho_U\mathbb{E}(N_Y)\mathbb{E}(X) \\ &+ \rho_Y\mathbb{E}(N_U)\mathbb{E}(X) + \mathbb{E}(N_U)\mathbb{E}(N_Y) && \text{(A.115)} \\ &= \rho_U\rho_Y , && \text{(A.116)} \end{aligned}$$

where (a) follows the fact that $U = \rho_U X + N_U$ and $Y = \rho_U Y + N_Y$, which proves expression (2.54). The equivocation reads as:

$$\begin{aligned} \Delta \leq h(X|UZ) &= h(X) - I(U;X) + I(U;Z) \\ &- h(Z) + h(Z-X|UX) && \text{(A.117)} \\ &= -\frac{1}{2}\log\left(\frac{1}{1-{\rho_U}^2}\right) + \frac{1}{2}\log\left(\frac{1}{1-\rho_U^2\rho_Z^2}\right) \\ &+ \frac{1}{2}\log 2\pi e(1-\rho_Z^2) && \text{(A.118)} \\ &= \frac{1}{2}\log 2\pi e\frac{(1-{\rho_U}^2)(1-{\rho_Z}^2)}{1-{\rho_U}^2{\rho_Z}^2} . && \text{(A.119)} \end{aligned}$$

Finally, we conclude the proof by observing that the equivocation rate must be positive.

PART III

Synthèse de la Thèse, en Français

CHAPTER 6

SYNTHÈSE DE LA THÈSE, EN FRANÇAIS

6.1 Introduction

6.1.1 Une Revue Historique

Les débuts de ce que nous savons maintenant comme la communication numérique moderne proviennent de l'œuvre de Nyquist (1924) [2].

À la lumière des travaux de Nyquist, Hartley (1928) [3] a étudié la quantité d'informations qui peuvent être transmises de façon fiable sur un canal à bande limitée lorsque plusieurs niveaux d'amplitude sont autorisés. Un autre progrès significatif dans le développement des communications a été le travail de Wiener (1942), qui a examiné le problème de l'estimation d'une forme d'onde de signal souhaitée $s(t)$ en présence de bruit additif $n(t)$, Sur la base de l'observation du signal reçu: $r(t) = s(t) + n(t)$.

Les résultats de Hartley et Nyquist sur le taux de transmission maximum de l'information numérique ont été des précurseurs de l'œuvre de Claude Shannon; Souvent appelé le père de l'ère numérique. Au début de son article, Shannon a reconnu le travail accompli avant lui par des pionniers tels que Nyquist et Hartley chez Bell Labs. Bien que leur influence ait été profonde, c'est Shannon qui a révolutionné la communication et défini un nouveau champ de recherche sur la communication que nous connaissons maintenant comme théorie de l'information. L'un de ces concepts clés était sa définition de la limite de capacité de canal.

La théorie de l'information est l'un des rares domaines scientifiques à avoir un début identifiable - le papier de Claude Shannon de 1948. L'histoire de l'évolution de la progression d'un seul article théorique à un vaste domaine qui a redéfini notre monde est fascinante. Peut-être pendant les 25 premières années de son existence, la théorie de l'information a servi de riche source de problèmes de recherche universitaire [8, 9, 10, 11], et comme suggestion tentante que ses approches peuvent rendre les systèmes de communication plus efficaces et plus performants. Mis à part de petites expériences et quelques systèmes militaires particuliers, la théorie était considérée comme une belle théorie et avait peu d'interaction avec la pratique. Cependant, au milieu des années 1970, les systèmes de communication ont commencé à mettre en œuvre des idées de théorie de l'information de manière approfondie.

Shannon a formulé le problème fondamental de la transmission fiable de l'information en termes statistiques, en utilisant des modèles probabilistes pour les sources d'information et les canaux de communication. Il a également défini la notion de capacité de canal et fourni un cadre mathématique par lequel on peut le calculer.

6.1.2 Éléments des Systèmes de Communication Pratique

La théorie de l'information que nous connaissons aujourd'hui n'est pas seulement l'œuvre de Claude Shannon, mais le résultat de nombreuses contributions importantes faites par des individus différents, issus de milieux

variés, qui ont pris les idées de Shannon et les ont développées. La diversité et les orientations de leurs objectifs et de leurs intérêts formaient la forme de la théorie de l'information d'aujourd'hui.

Au sens le plus fondamental, la communication implique implicitement la transmission d'un point à un autre à travers une succession de processus. Figure 6.1 illustre le diagramme fonctionnel et les éléments de base d'un système de communication numérique [1].

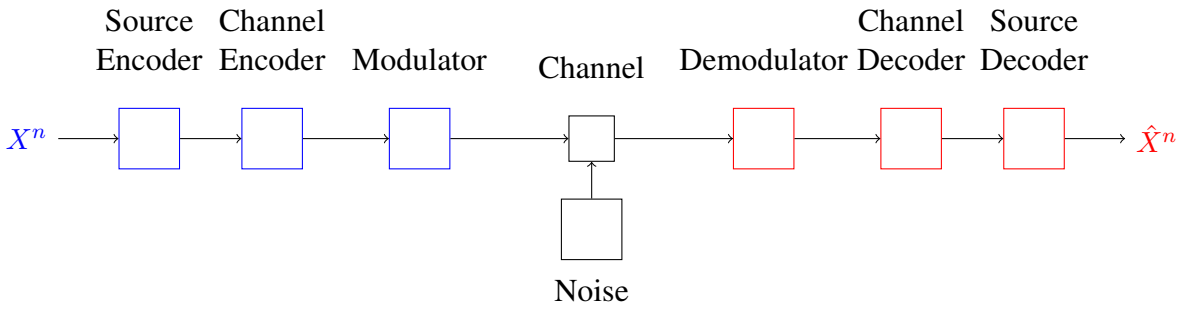


Figure 6.1: Éléments de base d'un système de communication numérique.

6.1.3 Codage Source et Quantification

Dans cette thèse, notre objectif principal sera la composante *code source*. Des systèmes de communication optimum pour la transmission de la source pourraient être construits en concevant séparément des codes sources pour la source et des codes de correction d'erreur pour le canal. Par conséquent, une œuvre telle que celle-ci qui se concentre uniquement sur les aspects de codage de source ou de signal ne signifie pas intrinsèquement une perte de généralité. Un système global efficace peut toujours être construit en mettant en cascade un bon système de codage de signal avec un bon système de contrôle d'erreur. En fait, la plupart des systèmes de communication pratiques pour la transmission de source aujourd'hui sont basés sur la séparation.

La théorie du codage source a d'abord été formulée par Shannon où un taux minimum de compression de données sans perte a été établi. Ce taux est le même que le taux d'entropie H de la source qui a été définie plus tôt. La valeur exacte de ce taux dépend de la source d'information, plus spécifiquement, de la nature statistique de la source. Il est possible de compresser la source, sans perte, avec un taux de compression proche de H.

Shannon a également développé la théorie de compression de données avec perte. Ceci est mieux connu sous le nom de théorie de la distorsion des taux [16]. En compression de données avec perte, les données décompressées ne doivent pas nécessairement être identiques aux données d'origine. Au lieu de cela, une certaine quantité de distorsion, D, est tolérée. Shannon a montré que pour une source donnée (avec toutes ses propriétés statistiques connues) et une mesure de distorsion donnée, il existe une fonction, $R(D)$, appelée fonction de distorsion de débit. La théorie dit que si D est la quantité tolérable de distorsion, alors $R(D)$ est le meilleur taux de compression possible.

Quantification; Une forme de méthode de compression avec perte; Cartographie les valeurs d'amplitude en une plage discrète, de sorte que le signal quantifié ne prend qu'un ensemble de valeurs discret, généralement fini. Par conséquent, la quantification entraîne une perte d'information en introduisant une distorsion dans le signal quantifié qui ne peut être éliminée. L'augmentation du nombre de sorties discrètes d'un quantificateur réduit généralement la distorsion, mais ne peut pas l'éliminer. Le compromis fondamental dans ce choix est la qualité du signal résultant par rapport à la quantité de données nécessaires pour représenter chaque entrée.

L'ensemble des entrées d'un quantificateur peut être des scalaires ou des vecteurs. S'ils sont des scalaires, on appelle les quantificateurs scalaires quantificateurs. S'ils sont des vecteurs, nous les appelons quantificateurs vectoriels. Les quantificateurs scalaires et vectoriels jouent un rôle important dans la compression des données et ont donc fait l'objet d'études approfondies. Alors que la quantification scalaire est utilisée principalement pour la conversion analogique-numérique, la quantification vectorielle est utilisée avec un traitement de signal numérique sophistiqué, où dans la plupart des cas le signal d'entrée a déjà une certaine forme de représentation numérique et la sortie souhaitée est une version compressée du signal numérique . La quantification vectorielle est habituellement, mais non exclusivement, utilisée à des fins de compression de données.

Un vecteur peut être utilisé pour décrire presque n'importe quel type de motif, tel qu'un segment d'une forme d'onde de parole ou une image, simplement en formant un vecteur d'échantillons à partir de la forme d'onde ou de l'image. La quantification vectorielle peut être considérée comme une forme de reconnaissance de motif dans laquelle un motif d'entrée est approché par l'un d'un ensemble prédéterminé de motifs standard, ou dans une autre langue, le motif d'entrée est apparié à l'un d'un ensemble stocké de modèles ou de mots de code. La quantification vectorielle est beaucoup plus qu'une généralisation formelle de la quantification scalaire. Ces dernières années, il est devenu une technique importante dans la reconnaissance de la parole ainsi que dans la compression de la parole et de l'image, et son importance et l'application sont de plus en plus.

Le but de la quantification est de fournir une description de précision limitée d'un vecteur d'entrée précédemment inconnu. C'est seulement parce que l'entrée n'est pas connue à l'avance qu'il est nécessaire de quantifier. Ainsi, l'entrée doit être modélisée comme une variable aléatoire, ayant un caractère statistique spécifique, habituellement spécifié par sa fonction de densité de probabilité. Par conséquent, l'erreur introduite dans la quantification de cette entrée sera également aléatoire. Pour évaluer commodément les performances d'un quantificateur particulier, nous avons besoin d'un seul chiffre qui indique la dégradation globale de la qualité ou la distorsion survenue pendant la durée de vie de son utilisation avec une entrée statistiquement spécifiée. En outre, une certaine mesure globale de la performance, habituellement basée sur la moyenne statistique, est requise qui prend en compte le pdf d'entrée ainsi que la caractéristique de quantificateur spécifique.

Codage Source Distribué

Le problème du codage source devient beaucoup plus intéressant et difficile dans un contexte de réseau. Plusieurs nouveaux scénarios se présentent:

- Des informations différentes de l'information source peuvent être disponibles pour des terminaux de codage séparés qui ne peuvent pas coopérer.

- Les décodeurs peuvent avoir accès à des informations supplémentaires sur les informations source.

Les problèmes de codage source avec des informations côté décodeur sont un cas particulier de problèmes de codage de source distribuée. Le rôle et les avantages potentiels de Side Information (S.I.) dans la compression des données sans perte et avec perte sont un thème central dans la théorie de l'information. D'une manière qui est bien comprise pour divers systèmes de codage source, S.I. peut être une ressource précieuse, entraînant des augmentations de performances significatives par rapport au cas où il est absent.

Codage Source Orienté Détection

La conception du codeur / décodeur source dépend de l'objectif du système de communication. Lloyd, par exemple, visant à concevoir des systèmes pour l'application particulière de la récupération de la source. Dans de nombreuses autres applications, l'objectif est de détecter une source plutôt que de l'estimer. Par exemple, un opérateur de radar doit décider si ce qu'il voit sur l'écran radar indique la présence d'un plan (le signal) ou la présence de parasites (le bruit). Ce type d'applications Était le cadre original de *Signal Detection Theory* (voir le travail fondateur de Green & Swets, 1966) [35].

La détection, la prise de décision et le test d'hypothèse (TH) peuvent parfois se référer à la même chose. La signification a été étendue dans le champ de la communication pour détecter laquelle, parmi un ensemble d'alternatives mutuellement exclusives, est correcte. Ces alternatives mutuellement exclusives sont habituellement appelées hypothèses. Le problème principal que nous considérons dans cette thèse est le problème du test d'hypothèse binaire dans lequel nous supposons qu'il existe deux hypothèses possibles, l'hypothèse nulle H_0 et l'hypothèse alternative H_1.

Un TH peut être erroné de deux manières différentes. Soit le détecteur décide H_1 lorsque H_0 est l'hypothèse vraie, soit décide H_0 quand H_1 est l'hypothèse correcte. Les probabilités de ces deux erreurs, respectivement notées α et β, donnent ensemble la performance d'un détecteur (ou d'un test) [36].

Le test des hypothèses distribuées (TH) a d'abord été étudié par Berger [39] et Ahlswede & Csiszar in [40]. Il s'agissait d'une première étape pour combiner deux problèmes apparemment différents qui ont été étudiés séparément dans les domaines de l'inférence statistique standard et de la théorie de l'information. Ce scénario se compose d'un décodeur (détecteur) qui est nécessaire pour effectuer une décision binaire basée sur des données collectées à distance envoyées par le codeur.

Dans ces problèmes, l'objectif est donc de définir une mesure de distorsion appropriée permettant de concevoir un quantificateur adapté au test d'hypothèse. Le critère MSE est approprié lorsque l'objectif est de reconstituer la source. Toutefois, il peut être déraisonnable lorsque d'autres applications sont concernées. Un quantificateur conçu pour minimiser l'erreur quadratique moyenne peut perdre de façon imprudente les informations nécessaires à une bonne performance de détection. Dans les problèmes de détection, la principale mesure de dégradation devrait être les deux types d'erreur. Cependant, la minimisation des probabilités d'erreur est difficile et incommode à effectuer à de nombreuses reprises. Par conséquent, plusieurs mesures de performance sous-optimales telles que les mesures de dissimilarité qui sont plus faciles à manipuler sont étudiées.

Codage Source Orienté par le Secret

De nos jours, une énorme quantité de flux d'informations dans le réseau. Avec cette énorme quantité d'informations, la principale tâche des concepteurs de réseau est de s'assurer que les données peuvent être transmises de manière fiable et sécurisée à travers le réseau. Cette dernière exigence devient de plus en plus aiguë, surtout lorsque des informations sensibles sont impliquées. Imaginons un réseau dans lequel les informations circulent d'un noeud à un autre à travers un certain nombre de noeuds intermédiaires. La conception du système utilise généralement ces nœuds intermédiaires pour faciliter la transmission. Cependant, ces noeuds peuvent être des terminaux ou des terminaux publics dont nous ne pouvons pas nous fier pleinement pour accéder à des quantités importantes de nos informations. Ce scénario conduit à un compromis naturel entre la coopération et le secret dans le système et motive l'étude de la communication et de la compression sécurisées.

Au début de l'ère de la théorie de l'information, la majorité des études ne portaient que sur les problèmes de

communication fiable. Récemment, des recherches approfondies se concentrent sur la communication sécurisée, c'est-à-dire lorsque l'objectif est de concevoir un système de communication à la fois fiable et sécurisé. Les techniques conventionnelles pour assurer la confidentialité et les réseaux de communication sont basées sur le chiffrement cryptographique où la sécurité n'était prise en compte que dans la couche d'application du modèle OSI. En cryptage, l'émetteur utilise une clé pour chiffrer des informations de source, c'est-à-dire en clair, converties en texte chiffré. Le récepteur souhaité extrait le texte en clair d'un texte chiffré à l'aide d'une clé correspondante. Si une a accès au texte chiffré, mais qu'elle ne connaît pas la clé de décryptage correspondante, elle ne peut pas obtenir les informations sources. Les principales faiblesses de cette notion de sécurité sont les hypothèses mises sur l'attaquant. En pratique, l'attaquant est habituellement supposé avoir un temps limité ou des ressources informatiques calculées limitées à ce qu'il ne peut pas tester toutes les clés possibles pour extraire l'information source.

Shannon a présenté la notion théorique d'information de la sécurité considérée à la couche physique, où le secret à l'eavesdropper est mesuré par l'incertitude conditionnelle de la source donnée le message en tenant compte des différentes caractéristiques possédées par l'eavesdropper et le récepteur légitime . Un système est théoriquement sécurisé en termes d'information si sa sécurité dérive purement de la théorie de l'information [42]. Autrement dit, il ne peut pas être rompu même lorsque l'adversaire dispose d'une puissance de calcul illimitée. L'adversaire n'a tout simplement pas assez d'informations pour briser le cryptage. Par conséquent, la sécurité théorique de l'information ne fait aucune hypothèse de calcul sur l'attaquant, et est acceptée comme la forme la plus stricte de sécurité [43].

Cette thèse (dans Chapters 2 et 3) fournit quelques nouveaux résultats fondamentaux sur la détection et le codage de la source orientée sur le secret en présence d'informations latérales aux terminaux de réception.

Plus précisément, nous étudions le problème du TH multiterminal sécurisé avec des informations secondaires à la fois au détecteur et à l'eavesdropper. Ce scénario se compose de trois nœuds:

- Un codeur principal (appelé Alice), qui observe une source locale,
- Un récepteur légitime (appelé Bob), qui souhaite estimer la distribution conjointe de la source d'Alice et les informations latérales directement disponibles à partir d'une version compressée reçue par un canal (public) à débit limité,
- Un eavesdropper (appelé Eve), qui observe parfaitement les bits d'information envoyés par Alice à Bob, et a également accès à une source corrélée qui peut être utilisée comme information latérale.

Dans le chapitre 2, nous étudions les limites fondamentales du problème, c'est-à-dire le compromis entre l'exposant maximal d'erreur réalisable au détecteur (c'est-à-dire la probabilité d'erreur minimale de type II pour une probabilité fixe d'erreur de type I) , La vitesse de codage au codeur et le taux de fuite à l'eavesdropper. Alors que dans le chapitre 3, nous proposons des algorithmes pratiques pour obtenir des solutions aussi proches que possible de l'optimal, ce qui nécessite la conception d'une quantification optimale (scalaire et vectorielle) en présence d'un eavesdropper.

6.1.4 Codage Sécurisé des Sources Distribuées avec Application aux Compteurs Intelligents

Le Smart Grid nécessite une transmission fiable et sécurisée des informations en temps réel. Par conséquent, la protection de l'ensemble du système de communication qui est au cœur du réseau intelligent est l'un des principaux objectifs.

Contrairement à la cyber-sécurité, la sécurité des couches physiques a été à peine abordée dans la littérature du réseau intelligent à ce jour. Dans [57], l'auteur a proposé un moyen de coder la mesure redondante à un débit inférieur à son entropie, de sorte qu'il ne peut pas être décodé à partir des bits codés seuls. De cette façon, il garantit la confidentialité théorique de l'information, quel que soit le pouvoir de calcul d'un eavesdropper. Le comptage redondant est fréquemment utilisé pour vérifier l'intégrité des données de facturation rapportées par l'infrastructure de comptage avancée, mais la mesure redondante introduit une fuite potentielle de confidentialité.

L'idée clé est de compresser la mesure redondante à un taux inférieur à son entropie, de sorte qu'elle ne peut pas être récupérée à partir des bits codés seulement. Mais la mesure redondante peut être récupérée en conjonction avec la mesure rapportée, tant que le taux de compression est supérieur à l'entropie conditionnelle de la mesure redondante compte tenu de la mesure rapportée. Contrairement au cryptage, cette méthode garantit la confidentialité indépendamment de la capacité de calcul de l'eavesdropper

Le cryptage semble être l'approche typique [58, 59]; Cependant, une telle approche n'a pas une base théorique solide à la fois pour la protection de la vie privée et la performance de détection / estimation. Une telle base est importante pour plusieurs raisons. Tout d'abord, une abstraction théorique nous permet de refondre le problème d'une manière indépendante de la technologie - nous avons besoin d'un cadre de confidentialité qui non seulement traite les capacités des techniques actuelles, mais est également extensible aux futures. Deuxièmement, un cadre théorique nous permet d'examiner les coûts de la vie privée perdue par rapport aux avantages de la diffusion des données, à savoir le compromis entre la confidentialité et la performance du décodeur. Il serait souhaitable d'avoir la capacité de décider de ce compromis. Enfin, un cadre théorique pour la protection de la vie privée et la performance peut exposer des points de compromis qui sont inattendus.

6.2 Test d'Hypothèse avec Contraintes de Communication et de Sécurité

6.2.1 Introduction

Ce chapitre étudie le problème du test d'hypothèses (TH) dans lequel les données sont compressées et envoyées à un détecteur qui cherche à choisir entre deux distributions possibles. L'objectif est de caractériser tous les taux de données réalisables et l'équivocation et l'exposant maximal de la probabilité d'erreur de type II lorsque la probabilité d'erreur de type I est au plus une valeur fixe. Le problème conventionnel TH est de décider entre deux distributions alternatives à partir des données observées qui sont disponibles au statisticien, c'est-à-dire $H_0 : p_X$ versus $H_1 : p_{\bar{X}}$.

Le problème de test d'hypothèse multiterminales (TH) sous contraintes de communication a d'abord été étudié par Berger [39]. Le TH distribué a été introduit dans [40] où les auteurs ont étudié ce problème en présence de contraintes de communication. Une caractérisation d'une seule lettre est donnée pour le test con-

tre l'indépendance tandis que des résultats partiels sont obtenus pour le problème TH général. Plus tard, [60], une limite inférieure sur l'exposant d'erreur optimal a été proposée sur la base de l'exposant de Wyner- Codage de Ziv [61]. D'autres résultats sont rapportés dans [62], où, sur la base d'un schéma de binning sophistiqué, une nouvelle borne inférieure est dérivée. Bien que pour les problèmes de codage source apparentés les schémas basés sur le binning aléatoire fonctionnent bien et souvent optimal, l'utilisation du binning dans TH est plus impliquée du fait que la probabilité globale d'erreur peut être dominée par des erreurs dans le décodage des indices bin. Plus récemment, Les auteurs de [63] ont étudié TH sans contraintes de sécurité et ont montré que le binning est optimal pour le type de problèmes dont le but principal est de tester l'indépendance conditionnelle.

Le compromis optimal entre la distorsion et l'équivocation est bien connu dans la configuration classique de distorsion de débit [69], mais n'est pas lorsque la fonction cible est l'exposant d'erreur au lieu d'une distorsion moyenne par lettre. Plus récemment dans [70], ils établissent des limites internes et externes sur la région de distorsion-distorsion-équivocation pour le problème de codage source avec perte avec des contraintes de secret dans lesquelles une source d'information distante doit être transmise à une destination.

Ce chapitre traite du scénario de TH sous les contraintes de communication et de confidentialité. Le modèle correspondant est montré dans la figure 6.2 où Alice maximise non seulement l'exposant mais aussi le taux d'équivocation - une incertitude moyenne - à un eavesdropper, noté Eve. Eve est supposée observer parfaitement les bits d'information envoyés par Alice et a accès à un i.i.d. String $Z^n = (Z_1, \ldots, Z_n)$ (peut-être le même que Bob). Nous étudions le minimum de données que Alice doit communiquer à Bob pour garantir l'exposant souhaité tout en satisfaisant à l'exigence d'incertitude moyenne à Eve. Le compromis optimal entre la distorsion et l'équivocation est bien connu dans la configuration conventionnelle de distorsion de débit [78], mais n'est pas lorsque la fonction cible est l'exposant d'erreur au lieu d'une distorsion. Le problème de codage source standard et le problème TH sans contraintes de confidentialité semblent être fondamentalement différents l'un de l'autre.

Le chapitre est divisé en deux parties. Dans la première partie, nous étudions le problème TH général et dérivons une région réalisable. La région que nous proposons utilise un schéma à double couche basé sur le binning et généralise celui fourni dans [60] en prenant en considération à la fois les erreurs de test et de décodage qui ont lieu pendant le processus de binning. L'utilisation de la double couche basée sur le binning a rendu le problème très compliqué mais a fourni un meilleur exposant d'erreur globale et était nécessaire pour améliorer la sécurité dans la région. Dans [62], une limite inférieure similaire mais non équivalente sur l'exposant d'erreur a été proposée mais malheureusement la preuve n'est pas disponible. Dans la deuxième partie, nous étudions le cas particulier du test contre l'indépendance où l'on suppose que, dans les deux hypothèses, les distributions de probabilité ont des margaux égaux. Dans ce cas, une seule couche a été utilisée sans binning et une caractérisation optimale d'une seule lettre de la région d'équation de taux-exposant est fournie. Les applications de nos résultats se posent dans le contexte dans lequel les données doivent rester privées même du statisticien (voir [75] et les références qui y figurent), la région est également évaluée pour les sources gaussiennes et binaire.

6.2.2 Définition du Problème

Modèle de Système

Le modèle est représenté sur la figure 6.2, où un codeur Alice observe des réalisations i.i.d d'une variable aléatoire vectorielle X, et code à un taux R. Un statisticien Bob (le détecteur) observe la version encodée $f(X^n)$ de $X^n = (X_1, \ldots, X_n)$ et des réalisations i.i.d d'un vecteur Y^n directement disponible, Qui est arbitrairement dépendante

de X^n sans mémoire. Un eavesdropper (appelé Eve) a accès à $J = f(X^n)$ et une autre information latérale Z^n, également composée de i.i.d. Échantillons arbitraires dépendants de (X^n, Y^n). Alice souhaite communiquer la source en utilisant un mappage d'encodage:

$$f : \mathcal{X}^n \mapsto \{1, \dots, \|f\|\} \tag{6.1}$$

Avec le taux de codage $\log \|f\| \leq nR$. Le détecteur Bob doit prendre une décision entre deux hypothèses:

$$\begin{cases} H_0 : & p_{XY} \ , \\ H_1 : & p_{\bar{X}\bar{Y}} \ . \end{cases} \tag{6.2}$$

On suppose également que les distributions marginales de X et Y sont les mêmes dans les deux hypothèses, c'est-à-dire $p_X = p_{\bar{X}}$ et $p_Y = p_{\bar{Y}}$ qui ne permet pas à Bob de prendre la décision sans les informations envoyées par Alice. Dans ce contexte, Bob doit décider à partir de l'échantillon Y^n et du message $f(X^n)$ entre H_0 et H_1, dont un seul est vrai. Pour simplifier, on note les distributions de probabilité correspondantes par $\mathbb{P} = p_{f(X^n)Y^n}$ et $\bar{\mathbb{P}} = p_{f(\bar{X}^n)\bar{Y}}$.

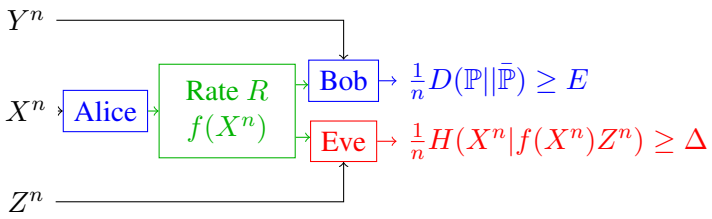


Figure 6.2: Tests d'hypothèses avec contraintes de communication et de confidentialité.

Soit $\varepsilon \in (0, 1)$, $\mathcal{A}_n \subset \{1, \dots, \|F\|\} \times \mathcal{Y}^n$ soit la région d'acceptation pour le détecteur chez Bob. Les deux types de probabilités d'erreur sont définis comme suit:

$$\text{Type I}: \quad \alpha_n(f, \mathcal{A}_n) = \mathbb{P}(\mathcal{A}_n^c) < \varepsilon \ , \tag{6.3}$$

$$\text{Type II}: \quad \beta_n(f, \mathcal{A}_n) = \bar{\mathbb{P}}(\mathcal{A}_n) \ . \tag{6.4}$$

Le but du détecteur Bob est de trouver une fonction de codage f et une région d'acceptation $\mathcal{A}_n$ qui minimisent la probabilité (6.4) pour une probabilité prescrite (6.3), et garantir le taux d'équivocation (ou incertitude moyenne) à Eve tel que:

$$\frac{1}{n} H(X^n | f(X^n) Z^n) \geq \Delta \ , \tag{6.5}$$

ou une définition équivalente est le *de fuite d'information*:

$$\frac{1}{n} I(X^n; f(X^n) Z^n) \leq H(X) - \Delta \ . \tag{6.6}$$

Notre objectif est d'obtenir une représentation d'une seule lettre réalisable de cette région.

6.2.3 Test d'Hypothèse Générale

Dans cette section, nous nous concentrerons sur le cas où, dans les deux hypothèses, les deux distributions sont générales, en d'autres termes, les deux hypothèses sont $H_0 : p_{XY}$ et $H_1 : p_{\bar{X}\bar{Y}}$.

Région de Taux-Erreur-Équivocation pour le TH Général

Notre résultat principal pour le test général d'hypothèse est une région de taux-erreur-équivocation réalisable et peut donc être résumée par la proposition suivante:

Proposition 6.1 (Région taux-erreur-équivocation réalisable) *Un ensemble de tous les uplets (R, E, Δ) est réalisable si les inégalités suivantes sont satisfaites:*

$$\begin{aligned} R &\geq I(p_X; p_{U|X}) && (6.7)\\ E &\leq \min\{E_1; E_2(R - I(p_X; p_{U|X}));\\ &\qquad E_2(H(p_{X|UZ}|p_{UZ}) - \Delta)\} && (6.8)\\ \Delta &\leq H(p_{X|UZ}|p_{UZ}) && (6.9) \end{aligned}$$

Où U et V sont des variables aléatoires auxiliaires définies sur certains ensembles finis $\mathcal{U}$ et $\mathcal{V}$ tels que $U \multimap V \multimap X \multimap (Y, \bar{Y}, Z)$ former une chaîne de Markov. $\bar{U}$ et $\bar{V}$ sont également définis tels que $(\bar{U}, \bar{V}) \multimap \bar{X} \multimap \bar{Y}$ et $P_{\bar{U}\bar{V}|\bar{X}} = p_{UV|X}$; E_1 et E_2 sont définis par:

$$E_1 = \min_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y} \in \mathscr{L}(UV)} D(p_{\tilde{U}\tilde{V}\tilde{X}\tilde{Y}} || p_{\bar{U}\bar{V}\bar{X}\bar{Y}}) \,, \tag{6.10}$$

$$\begin{aligned} \mathscr{L}(UV) = \{ \quad & \tilde{U}\tilde{V}\tilde{X}\tilde{Y} : P(\tilde{U}\tilde{V}\tilde{X}) = P(UVX), \\ & P(\tilde{U}\tilde{V}\tilde{Y}) = P(UVY)\} \,, \end{aligned} \tag{6.11}$$

$E_2(R - I(p_X; p_{U|X})) =$

$$\inf_{q_{X|U}} \sup_{q^\star_{V|UX}} \inf_{q_{Y|U}} \inf_{\substack{q_{VXY|U} \\ q_{V|UX} = q^\star_{V|UX}}} G[q_{VXY|U}, R - I(p_X; p_{U|X})] \,, \tag{6.12}$$

$E_2(H(p_{X|UZ}|p_{UZ}) - \Delta) =$

$$\inf_{q_{X|U}} \sup_{q^\star_{V|UX}} \inf_{q_{Y|U}} \inf_{\substack{q_{VXY|U} \\ q_{V|UX} = q^\star_{V|UX}}} G[q_{VXY|U}, H(p_{X|UZ}|p_{UZ}) - \Delta] \,, \tag{6.13}$$

$G[q_{VXY|U}, R(p_X; p_{U|X})]$ et $G[q_{VXY|U}, H(p_{X|UZ}|p_{UZ}) - \Delta]$ sont définis dans (6.14) et (6.15) au bas de la page suivante.

6.2.4 Test contre l'Indépendance

Dans ce cas particulier, l'objectif est de se concentrer sur le test d'indépendance, c'est-à-dire entre:

$$\begin{cases} H_0: & p_{XY} \ , \\ H_1: & p_X \times p_Y \ , \end{cases} \tag{6.16}$$

Où les deux types de probabilités d'erreur sont dans ce cas définis comme suit:

$$\text{Type I}: \quad \alpha_n \triangleq \Pr\left(H_1 | X^n Y^n \sim p_{XY}^n\right) \ , \tag{6.17}$$

$$\text{Type II}: \quad \beta_n \triangleq \Pr\left(H_0 | X^n Y^n \sim p_X^n \times p_Y^n\right) \ . \tag{6.18}$$

Limiter l'exposant erreur dans ce cas est beaucoup plus facile en raison du fait que:

$$\frac{1}{n} D(p_{f(X^n)Y^n} || p_{f(X^n)} \times p_{Y^n}) = \frac{1}{n} I(J; Y^n) \tag{6.19}$$

Ainsi, une caractérisation multi-lettres de la région *taux-exposant-équivocation* $\mathcal{R}^*$ est donnée par l'ensemble de tous les uples $(R, E, \Delta) \in \mathbb{R}_+^3$ De telle sorte qu'il existe une fonction de codage f satisfaisant:

$$\frac{1}{n} \log \|f\| \leq R \ , \tag{6.20}$$

$$\frac{1}{n} I(f(X^n); Y^n) \geq E \ , \tag{6.21}$$

$$\frac{1}{n} H(X^n | f(X^n) Z^n) \geq \Delta \ . \tag{6.22}$$

Région de Taux-Erreur-Équivocation de Lettre Unique

Nous énonçons maintenant la région optimale de taux-erreur-distorsion pour tester contre l'indépendance qui fournit une expression d'une seule lettre.

Proposition 6.2 (Caractérisation d'une seule lettre) *Soit $\mathcal{R}^\star$ l'ensemble de tous les uples réalisables (R, E, Δ),*

$G[q_{VXY|U}, R - I(p_X; p_{U|X})] =$

$$\begin{cases} \min\{D(q_{VXY|U} || p_{XY|U} q_{V|UX} | q_U); D(q_{VXY|U} || p_{\bar{X}\bar{Y}|U} q_{V|UX} | q_U)\} & \\ +[R - I(p_X; p_{U|X}) - I(q_{UX}; q_{V|UX}) + I(q_{UY}; q_{V|UY})]^+ & R < I(p_X; p_{U|X}) + I(q_{UX}; q_{V|UX}) \\ \infty & \text{else.} \end{cases} \tag{6.14}$$

$G[q_{VXY|U}, H(p_{X|UZ}|p_{UZ}) - \Delta] =$

$$\begin{cases} \min\{D(q_{VXY|U} || p_{XY|U} q_{V|UX} | q_U); D(q_{VXY|U} || p_{\bar{X}\bar{Y}|U} q_{V|UX} | q_U)\} & \\ +[H(p_{X|UZ}|p_{UZ}) - I(q_{UX}; q_{V|UX}) + I(q_{UY}; q_{V|UY}) - \Delta]^+ & \Delta > H(p_{X|UZ}|p_{UZ}) - I(q_{UX}; q_{V|UX}) \\ \infty & \text{else.} \end{cases} \tag{6.15}$$

alors il existe une variable aléatoire U sur un ensemble fini $\mathcal{U}$ satisfaisant :

$$R \geq I(U;X)\,, \tag{6.23}$$

$$E \leq I(U;Y)\,, \tag{6.24}$$

$$\Delta \leq H(X|UZ)\,, \tag{6.25}$$

Où $U \multimap X \multimap (Y,Z)$ forment une chaîne de Markov.

Notez que l'équivocation est indépendante des différences statistiques entre Y et Z. Contrairement au paramètre de distorsion-distorsion-équivocation précédemment étudié dans [78], dans le contexte actuel de test contre l'indépendance, l'information analogique Y disponible à Bob ne peut pas aider à améliorer le taux d'équivocation à l'Eavesdropper. Cette observation peut simplement s'expliquer par le fait que dans les deux cas où Y est indépendant de X ou non, Alice doit encoder pour le pire cas et ne peut pas l'utiliser pour effectuer le binning (par exemple le codage de Wyner-Ziv).

6.2.5 Conclusion

Dans ce chapitre, nous nous sommes concentrés sur le problème de test d'hypothèse avec des contraintes de communication et de confidentialité. Bob doit choisir entre deux hypothèses basées sur les informations disponibles et sur les informations communiquées à distance par Alice. En effet, Alice communique sur un canal public sans erreur mais limité. L'objectif est de garantir un exposant d'erreur voulu à Bob pour un débit donné tout en satisfaisant un taux d'équivocation moyen à Eve.

En testant l'indépendance, nous avons pu caractériser le compromis optimal entre le taux, l'exigence d'exposant d'erreur et les garanties de confidentialité. Une seule couche de codage sans binning était nécessaire. Dans le cas d'essais généraux d'hypothèses, une approche basée sur la méthode des types a été utilisée afin de dériver une région de taux-erreur-équivocation réalisable sans pouvoir prouver son optimalité.

6.3 Apprentissage de Quantification pour une Décision Binaire Distribuée avec Contraintes de Confidentialité

6.3.1 Introduction

De nombreuses applications nécessitent qu'un message soit transmis d'une source d'information à une destination souhaitée, où les décisions sont prises sur la source sur la base des données reçues. Par exemple, des données peuvent être envoyées par un radar de détection d'objet ou une caméra de surveillance vidéo à une station de surveillance intéressée à détecter une cible ou un objet spécifique dans la plage de vision du radar. Dans une telle application, il est souvent crucial de diminuer la vitesse des données transmises en codant la source avant la transmission.

De toute évidence, la quantification est l'outil de choix pour réduire la vitesse, mais, contrairement à la situation classique, la distorsion est le critère de conception pour optimiser la quantification. Dans notre cas,Nous sommes intéressés par la meilleure performance possible pour le test tout en veillant à ce que les données brutes restent sécurisées par rapport à un eavesdropper éventuel.

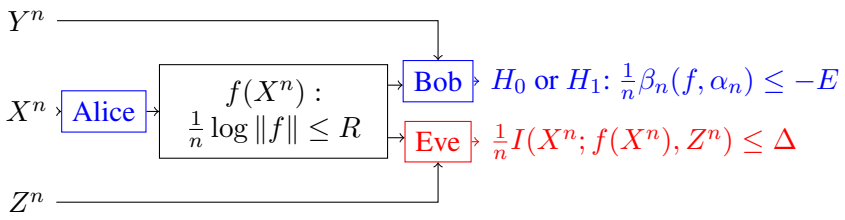


Figure 6.3: Distributed detection under privacy constraints.

Même sans contrainte de confidentialité, un premier problème consiste à définir une mesure de distorsion appropriée permettant de concevoir un quantificateur adapté au test d'hypothèse. L'erreur quadratique moyenne (MSE) est la mesure de distorsion appropriée lorsque l'objectif est d'estimer la source à partir des valeurs quantifiées correspondantes. Cependant, dans les problèmes de détection, la principale mesure de dégradation devrait être la probabilité d'erreur de type II (échec de rejet d'une fausse hypothèse nulle) pour une probabilité d'erreur de type I prescrite (rejet incorrect d'une hypothèse nulle vraie). Minimiser la probabilité d'erreur pour déterminer le schéma de quantification optimal est difficile et incommode à effectuer à de nombreuses reprises. En conséquence, plusieurs critères sous-optimaux qui sont plus souples à estimer et à manipuler ont été utilisés dans cette situation.

Dans notre problème, on suppose que la distribution de probabilité conjointe de deux sources éloignées est connue à la fois sous l'hypothèse nulle H_0 et l'hypothèse alternative H_1 et le test d'hypothèse de Neyman-Pearson le plus puissant est utilisé comme règle de décision [38].

Dans ce chapitre, nous considérons le problème de la détection binaire distribuée opérant sur des liens non sécurisés. Ceci consiste principalement à tester la probabilité conjointe de deux sources distribuées en présence d'un eavesdropper, comme décrit dans la figure 3.1.

Ce chapitre fait deux contributions: La première partie se concentrera principalement sur la quantification scalaire où le test se situe entre deux distributions générales de probabilité [101]. Dans ce cas, le *coefficient Bhattacharyya* est utilisé comme mesure de distorsion car il est beaucoup plus facile à manipuler que la divergence et les garanties standard. La deuxième partie se concentre sur la quantification vectorielle pour le test de l'indépendance. Un algorithme général d'optimisation est donné dans les deux cas, et la performance pour les sources Gaussiennes est évaluée.

Ces résultats théoriques pour ce scénario viendraient vraiment dans la pratique de beaucoup de façons différentes. Dans ce chapitre, l'application aux compteurs intelligents montrera comment tester l'indépendance peut être utilisé pour tester l'intégrité des appareils intelligents présents dans les maisons. La distribution conjointe de certaines données collectées au compteur intelligent et d'autres données disponibles auprès du fournisseur permet au collecteur de vérifier si le compteur intelligent se comporte correctement. Les applications de nos résultats apparaissent dans ces contextes où les données doivent rester privées même du statisticien (voir [102] et références là) où toute absence de corrélation peut indiquer une déficience de compteur intelligent et donc une détection de défaut.

6.3.2 Définition du Problème

Modèle de Système

Notre modèle est défini par trois noeuds comme décrit dans la figure 6.3. Alice (le quantificateur) observe une suite de vecteurs $\mathbf{x}^\tau = (\mathbf{x}_1, \ldots, \mathbf{x}_\tau)$ de i.i.d. Échantillons alors que Bob (le détecteur) observe une autre séquence de i.i.d. Échantillons $\mathbf{y}^\tau = (\mathbf{y}_1, \ldots, \mathbf{y}_\tau)$. Ces vecteurs sont n-dimensionnels, c'est-à-dire,

$$\mathbf{x}_t = (x_{t,1}; x_{t,2}; \ldots; x_{t,n}), \;\; t = 1, 2, \ldots, \tau \; . \tag{6.26}$$

τ sera désigné par le nombre de vecteurs échantillons disponibles à la fois au codeur et au décodeur. Bien que les vecteurs échantillons soient considérés comme i.i.d. l'un par rapport à l'autre, chaque vecteur peut impliquer des échantillons avec mémoire.

Alice veut coder ses données avec un taux maximum R [bits par dimension] qui est accompli en cartographiant les entrées dans les valeurs quantifiées $j^\tau \equiv (f(\mathbf{x}_1), \ldots, F(\mathbf{x}_\tau))$, en utilisant un quantificateur vecteur n-dimensionnel:

$$f : \mathcal{X}^n \longrightarrow \mathcal{J} \equiv \{1, \ldots, M\} \; . \tag{6.27}$$

La transformation se fait à l'aide d'un codebook de taille $||\mathcal{J}|| = M = 2^{nR}$. Chaque vecteur est encodé par un index qui pointe vers un mot de code à partir d'un ensemble fini de vecteurs, appelé codebook. Chaque mot de code; également appelé vecteur de reproduction est n-dimensionnel.

$\mathcal{J}$ est l'ensemble de tous les mots binaires possibles envoyés par le canal pour représenter le vecteur original $\mathbf{x} \in \mathcal{X}^n$. Puisque l'ensemble des indices $\mathcal{J}$ est discret et l'ensemble $\mathcal{X}^n$ est continu, la fonction de mappage est non-injective. L'ensemble des différents vecteurs d'entrée produisant la même valeur de sortie sera désigné sous le nom de région de quantification. Soit $\mathcal{R}_j$ la région de codage associée à l'index j.

$\boldsymbol{\mathcal{R}} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M\}$ indique la partition de l'espace. Autrement dit, les régions sont disjointes et complet. Si le vecteur source $\mathbf{x}$ est dans la région de codage $\mathcal{R}_j$, alors sa représentation est l'indice j de code dans le codebook:

$$f(\mathbf{x}) = j, \;\; \text{if} \;\; \mathbf{x} \in \mathcal{R}_j \; . \tag{6.28}$$

Le détecteur Bob reçoit le message j^τ communiqué par Alice et la séquence $\mathbf{y}^\tau$. Son but est de prendre une décision entre deux possibilités de la loi de probabilité conjointe de (X^n, Y^n) car elle ne peut être qu'une des deux hypothèses:

$$\begin{cases} H_0 : & (X^n, Y^n) \sim p_{X^n Y^n}(\mathbf{x}, \mathbf{y}) \; , \\ H_1 : & (X^n, Y^n) \sim p_{\bar{X}^n \bar{Y}^n}(\mathbf{x}, \mathbf{y}) \; . \end{cases} \tag{6.29}$$

On suppose en outre que les distributions marginales de X^n et Y^n sont les mêmes sous les deux hypothèses, soit $p_{X^n} = p_{\bar{X}^n}$ et $p_{Y^n} = p_{\bar{Y}^n}$.Ce qui ne permet pas à Bob de prendre la décision sans les informations envoyées par Alice. Soit $\hat{H}_0$ et $\hat{H}_1$ les résultats possibles du processus décisionnel. Les deux types de probabilités d'erreur associés au problème de détection sont donnés par:

$$\begin{array}{lrcl} \text{Type I :} & \alpha_n & \triangleq & \Pr\left(\hat{H}_1 | X^n Y^n \sim p_{X^n Y^n}(\mathbf{x}, \mathbf{y})\right) \; , \\ \text{Type II :} & \beta_n & \triangleq & \Pr\left(\hat{H}_0 | X^n Y^n \sim p_{\bar{X}^n \bar{Y}^n}(\mathbf{x}, \mathbf{y})\right) \; . \end{array} \tag{6.30}$$

Le message $f(\mathbf{x}^\tau)$ est envoyé à Bob via un lien public qui est parfaitement entendu par Eve (l'eavesdropper)

qui peut avoir accès à une autre séquence de vecteurs $\mathbf{z}^\tau$. La variable aléatoire Z^n est arbitrairement corrélée à (X^n, Y^n). L'objectif de ce système est alors de trouver une fonction de quantification f et une région d'acceptation qui minimisent la probabilité de type II pour une probabilité de type I prescrite et s'assure que la fuite d'information à Eve est limitée, c'est-à-dire

$$\frac{1}{n} I(X^n; f(X^n), Z^n) \leq \Delta \, . \tag{6.31}$$

6.3.3 Quantification Scalaire

Dans cette section, le cas de la quantification scalaire est considéré, c'est-à-dire le cas de $n = 1$. Comme il est indiqué dans la définition du problème, nous supposerons qu'après le processus d'échantillonnage, nous avons obtenu τ série de scalaires disponibles au niveau du quantificateur. Puisque nous utilisons un quantificateur scalaire, chaque scalaire d'entrée est traité séparément en produisant la sortie, c'est-à-dire que les entrées $x^\tau = (x_1, \ldots, x_\tau)$ sont mappées aux valeurs quantifiées $j^\tau \equiv (f(x_1), \ldots, f(x_\tau))$, en utilisant un quantificateur scalaire:

$$f : \mathcal{X} \longrightarrow \mathcal{J} \equiv \{1, \ldots, M\} \, , \tag{6.32}$$

avec un taux de codage $||\mathcal{J}|| \leq 2^R$.

Conception de Quantification

Des difficultés surgissent du fait du critère naturel de décision à optimiser. Dans ce cas, l'optimisation de ces probabilités ne donne pas une procédure de conception traitable. Pour cette raison, nous remplaçons le critère naturel par une mesure de dissimilarité (ou distance de répartition) entre les distributions sous les hypothèses. Le KLD est une mesure non symétrique de la différence entre deux distributions de probabilité. Une autre classe assez large de distances distributives appelées *distances Ali-Silvey* qui sont les plus fréquemment utilisées et qui ont trouvé une application réussie dans l'optimisation statistique. Dans ce qui suit, nous allons nous concentrer sur la *Bhattacharyya distance* définie comme:

$$D_B = -\log \left[E_0 \left(\sqrt{\Lambda} \right) \right] \, . \tag{6.33}$$

Avec Λ étant le LLR, E_0 est la moyenne par rapport à la distribution p_0. Nous choisissons d'optimiser f selon la *Bhattacharyya distance*.

6.3.4 Quantification Vectorielle

En combinant les entrées sources et en les codant comme un seul bloc, nous pouvons obtenir des algorithmes de compression plus efficaces. Bien que les quantificateurs vectoriels soient généralement plus complexes et passent plus de temps de traitement, l'adoption de la quantification vectorielle pour une vitesse donnée obtiendra très probablement des erreurs plus faibles que lorsque la quantification scalaire est utilisée au même taux. Dans ce cas particulier, l'objectif est de se concentrer sur le test d'indépendance, c'est-à-dire entre la distribution bivariée

donnée $p_{X^nY^n}$ par rapport à l'alternative d'indépendance donnée par:

$$\begin{cases} H_0: & (X^n, Y^n) \sim p_{X^nY^n}(\mathbf{x}, \mathbf{y}) \,, \\ H_1: & (X^n, Y^n) \sim p_{X^n}(\mathbf{x}) \times p_{Y^n}(\mathbf{y}) \,. \end{cases} \tag{6.34}$$

Algorithme QV pour la Détection

La conception du quantificateur est généralement considérée comme une tâche où la fonction objectif standard est de minimiser une distorsion attendue. Nous visons à prédire la meilleure distorsion qui convient le mieux à notre problème.

Une caractéristique clé de tout quantificateur est sa dimension n, un entier positif. Dans notre contexte, $\mathbf{x} = x^n$, $\mathbf{y} = y^n$ sont les vecteurs de dimension n disponibles respectivement chez Alice et Bob et définis sur un alphabet $\mathcal{X}^n, \mathcal{Y}^n \subset \mathbb{R}^n$. Par conséquent, $f : \mathcal{X}^n \mapsto \mathcal{J} = \{1, \ldots, M\}$ est le quantificateur vectorielle de M-points n-dimensionnel sans mémoire, Fonctionnant indépendamment sur des vecteurs successifs. Notre objectif est de trouver une fonction d'encodage f et une région d'acceptation qui minimise les erreurs $\beta_n(\alpha_n)$ comme $\alpha_n \to 0$, notre fonction objective devient:

$$\Gamma(f) = \lim_{\alpha_n \to 0} \frac{1}{n} \log \beta_n(\alpha_n, f) \leq -\frac{1}{n} I(Y^n; f(X^n)) \,, \tag{6.35}$$

La dernière inégalité est prouvée par [40]. Cette dernière information mutuelle représente l'exposant de l'erreur pour tester l'indépendance. La fonction objectif peut être réduite à:

$$\Gamma(f) \leq \frac{1}{n} h(Y^n | f(X^n)) \leq \frac{1}{2n} \log(2\pi e)^n \, \mathbb{E}(\|Y^n - \mathbb{E}(Y^n | f(X^n))\|^2) \,, \tag{6.36}$$

La dernière inégalité est le résultat du Lemme maximum d'entropie différentielle [37]. $\mathbb{E}(\|Y^n - \mathbb{E}(Y^n|f(X^n))\|^2)$ est la matrice de covariance de vecteur d'erreur de l'estimation de l'erreur quadratique moyenne minimale (MMSE) de Y^n. Notre problème peut être considéré comme équivalent à minimiser la distorsion entre Y^n disponible au niveau du détecteur et l'information portée par les sorties de quantification. Un décodeur de reproduction peut donc être considéré comme un mappage $g : \mathcal{J} \mapsto \hat{\mathcal{Y}}^n$ dont l'alphabet de reproduction est $\hat{\mathcal{Y}}^n = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_M\} \subset \mathbb{R}^n$. En conséquence, la fonction objective peut être finalement

$$\mathbb{E}(\|Y^n - \hat{Y}^n(f)\|^2) = \int_{\mathcal{X}^n} \int_{\mathcal{Y}^n} p_{X^nY^n}(\mathbf{x}, \mathbf{y}) \|\mathbf{y} - g \circ f(\mathbf{x})\|^2 d\mathbf{x} d\mathbf{y} \tag{6.37}$$

$$= \sum_{j=1}^{M} \int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) \mathbb{E}_{Y^n|X^n}[\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2] d\mathbf{x} \,. \tag{6.38}$$

$\hat{\mathbf{y}}_j \in \hat{\mathcal{Y}}^n$ est appelé le représentant de la région/cellule $\mathcal{R}_j$. $\|\mathbf{v}\| = \sum_i^n v_i^2$ indique la norme euclidienne carrée habituelle d'un vecteur $\mathbf{v} \in \mathbb{R}^n$ et $Q(\mathbf{x}) = g \circ f(\mathbf{x}) = g(f(\mathbf{x}))$ est la fonction qui mappe $\mathcal{X}^n$ à $\hat{\mathcal{Y}}^n$. Le problème de trouver le quantificateur optimal devient ainsi trouver l'ensemble des régions optimales $\boldsymbol{\mathcal{R}}$ et des représentants $\hat{\mathcal{Y}}^n$ qui minimisent la distorsion attendue $\mathbb{E}(\|Y^n - \hat{Y}^n(f)\|^2)$ sous la condition que les régions forment une partition de l'espace.

$$(\boldsymbol{\mathcal{R}}_{\mathbf{opt}}, \hat{\mathcal{Y}}^n_{\mathbf{opt}}) = \arg\min_{\boldsymbol{\mathcal{R}}, \hat{\mathcal{Y}}^n} \sum_{j=1}^{M} \int_{\mathcal{R}_j} p_{X^n}(\mathbf{x}) \, \mathbb{E}_{Y^n|X^n}[\|\mathbf{y} - \hat{\mathbf{y}}_j\|^2] d\mathbf{x} \,. \tag{6.39}$$

Algorithme QV sous Contraintes de Confidentialité

Le nouvel objectif devient de minimiser les erreurs $\beta_n(\alpha_n)$ ainsi que les fuites d'informations à Eve. Notre nouvelle fonction objective que nous avons hâte de minimiser devient donc:

$$\Gamma(f,\lambda) \leq -\frac{1}{n}I(Y^n; f(X^n)) + \lambda\frac{1}{n}I(X^n; f(X^n), Z^n)\ , \tag{6.40}$$

Γ est maintenant une fonction de coût lagrangienne formée de la manière d'incorporer les coûts séparés des deux coûts; La distorsion de détection et la fuite d'information. λ est une variable représentant un compromis entre le niveau de confidentialité et les erreurs de détection.

6.3.5 Application Pratique: Détection de Pannes Smart Meter sous Contraintes de Confidentialité

Considérons un réseau de quartier composé d'un certain nombre d'utilisateurs résidentiels où chaque utilisateur est équipé d'un compteur intelligent qui enregistre les données en temps réel sur l'utilisation de l'électricité. Le réseau d'utilisateurs d'électricité effectue l'agrégation de données par l'intermédiaire d'autres utilisateurs, par exemple, et rapporte les données agrégées en temps réel à l'Unité Centrale (UC) par l'intermédiaire de l'unité d'agrégation de données (DAU) comme le montre la figure 6.4. À la réception des rapports de la DAU, l'UC peut estimer l'utilisation moyenne en temps réel de l'électricité de la zone.

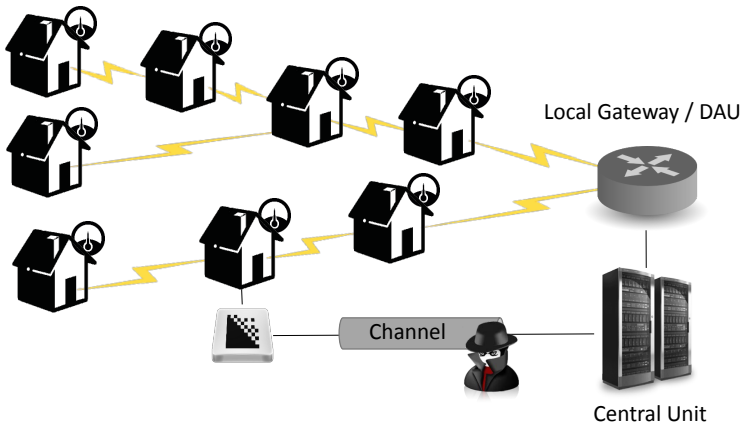


Figure 6.4: Une illustration du modèle étudié.

Les données agrégées moyennes de la zone résidentielle seront appelées Y^n . D'autre part, une autre donnée X^n provenant d'une maison séparée dans le réseau de quartier est codée avec un taux R et envoyée via un canal public à débit limité à l'unité centrale qui cherche à détecter une faute ou fraude au compteur intelligent de cette maison particulière. La détection des défauts peut alors être effectuée en testant si ces deux séries de données X^n et Y^n sont corrélées ou non. Une non corrélation pourrait signifier que les lectures de données du compteur intelligent sont erronées ou manipulées. Par conséquent, nous aimerions prendre une décision entre l'une des deux hypothèses suivantes

$$\begin{cases} H_0: & \text{Compteurs intelligents qui transmettent des données sans faille ,} \\ H_1: & \text{Compteurs intelligents envoyant des données erronées .} \end{cases} \tag{6.41}$$

Notre modèle peut donc être vu exactement comme le problème décrit dans la figure 6.3. Comme pour le problème de test contre l'indépendance définie précédemment, l'objectif est de prendre une décision entre deux possibilités de la loi de probabilité de (X^n, Y^n) car elle ne peut être qu'une sur deux hypothèses:

$$\begin{cases} H_0: & (X^n, Y^n) \sim p_{X^n Y^n}(\mathbf{x}, \mathbf{y}) \ , \\ H_1: & (X^n, Y^n) \sim \tilde{p}_{X^n}(\mathbf{x}) \times p_{Y^n}(\mathbf{y}) \ . \end{cases} \tag{6.42}$$

$\tilde{p}_{X^n}$ est une distribution arbitraire, et p_{Y^n} est la distribution marginale de Y^n sous H_0. En d'autres termes, le problème est considéré comme suit, H_0 représente le cas où le compteur intelligent fonctionne correctement et la distribution conjointe de X^n et Y^n à l'état normal serait $p_{X^n Y^n}$ qui pourrait être construit en utilisant des données historiques. Sous H_1, X^n pourrait provenir de toute distribution inconnue et c'est pourquoi la distribution de X^n sous H_1 est une distribution arbitraire. H_1 représente l'hypothèse où un défaut doit être détecté au niveau du compteur intelligent. Pour ce scénario particulier, notre fonction d'optimisation devient donc:

$$\begin{aligned} \Gamma(f, \lambda) &= \lim_{\alpha \to 0} \frac{1}{n} \log \beta_n(\alpha_n, f) + \lambda \frac{1}{n} I(X^n; f(X^n)) & (6.43) \\ &\leq -\frac{1}{n} D_{KL}(p_{f(X^n)Y^n} || \tilde{p}_{f(X^n)} p_{Y^n}) + \lambda \frac{1}{n} I(X^n; f(X^n)) \ , & (6.44) \end{aligned}$$

La dernière inégalité est également prouvée par [40]. Cependant, cette dernière expression peut être décomposée comme suit:

$$\begin{aligned} D_{KL}(p_{f(X^n)Y^n} || \tilde{p}_{f(X^n)} p_{Y^n}) &= D_{KL}(p_{f(X^n)Y^n} || p_{f(X^n)} p_{Y^n}) + D_{KL}(p_{f(X^n)} || \tilde{p}_{f(X^n)}) & (6.45) \\ &= I(f(X^n); Y^n) + D(p_{f(X^n)} || \tilde{p}_{f(X^n)}) \ , & (6.46) \end{aligned}$$

P_{X^n} étant la distribution marginale de X^n sous H_0. L'ancienne fonction de coût reste donc faisable.

6.3.6 Conclusion

Dans ce chapitre, nous avons dérivé un algorithme de détection binaire avec des échantillons quantifiés entre une source distante et une source directement disponible en présence de contraintes de confidentialité pour la quantification scalaire et vectorielle. Dans la partie QV, nous n'avons considéré que le cas particulier du test contre l'indépendance. Dans QS, notre algorithme utilise la distance de *Bhattacharyya* comme mesure de distorsion et un critère d'optimisation. Alors que dans QV, nous avons pu utiliser l'exposant d'erreur asymptotique comme il se réduit à l'information mutuelle entre la sortie codeur et la source à distance, un algorithme itératif qui calcule les régions de quantification et les représentants correspondants a ensuite été dérivé.

Le chapitre a également abordé une application importante dans le contexte des réseaux intelligents où TH peut être utilisé par les collectionneurs (le statisticien) pour tester l'intégrité des dispositifs intelligents présents dans les maisons tout en gardant privé les mesures des compteurs.

L'un des aspects les plus importants de cette application est l'adaptation à des statistiques d'observation inconnues ou variables. Quand aucun modèle d'observation n'est disponible, nous avons étendu l'algorithme de conception à utiliser avec les séquences d'apprentissage développées dans chaque hypothèse.

6.4 Apprentissage de la Quantification pour un cadre de Utilité-Confidentialité

6.4.1 Introduction

Dans ce chapitre, nous étudions deux scénarios. Le problème général du codage source sécurisé avec perte en présence d'un eavesdropper, qui observe les bits d'information et cherche à révéler une séquence privée non codée et corrélée. Dans le deuxième scénario, le cas contraire est considéré. Le décodeur souhaite récupérer la séquence corrélée non codée; Également appelée utilité pertinente en l'espèce; Tandis que l'eavesdropper cherche à révéler des informations sur la source codée elle-même. Il est supposé que tous les liens entre encodeurs et décodeurs sont publics de sorte qu'ils ne peuvent fournir aucun avantage pour accroître le secret. L'aspect clé de ce modèle est que le message J produit par l'encodeur pourrait jouer un double rôle en fonction du réglage désiré. Dans le premier réglage, il doit porter la description de la source X lui-même tout en conservant la confidentialité d'une donnée pertinente Y indisponible à l'encodeur. Dans l'autre paramètre, il doit porter des informations sur les données pertinentes Y visant à permettre au décodeur de reproduire ces données corrélées de la meilleure façon possible tout en gardant les données sur la source X privées d'un éventuel eavesdropper.

Extraire les aspects pertinents des données complexes comme dans le deuxième scénario est une tâche fondamentale dans le traitement du signal et les statistiques. Le problème est souvent que les données contiennent de nombreuses structures dans lesquelles une ou plusieurs d'entre elles pourraient être pertinentes. Par exemple, les signaux de parole peuvent être caractérisés par leur niveau de volume, leur hauteur ou leur contenu; Les images peuvent être représentées par leur niveau de luminosité, leur saturation en couleur, etc. Étant donné la distribution conjointe d'une variable source X et d'une autre variable de pertinence Y, on opère pour compresser X, tout en conservant des informations maximales sur Y. Une application pratique impliquant des données de consommation électrique mesurées à partir de maisons réelles est enfin étudiée.

6.4.2 Codage de Source à Perte avec une Contrainte de Confidentialité

Définition du Problème

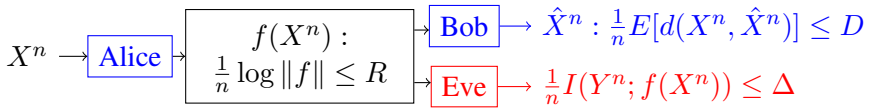


Figure 6.5: Codage de source à perte avec une contrainte de confidentialité.

Dans cette section, nous donnons une formulation plus rigoureuse du contexte représenté dans la figure 6.5. Le principal défi à relever pour caractériser le compromis entre la vie privée et l'utilité est de trouver les mesures quantitatives appropriées à la fois de l'utilité retenue et de la quantité d'information divulguée.

Alice (le quantificateur) observe une source de $X^n = (X_1, \ldots, X_n)$ avec la mémoire. Alice veut coder ses données avec un taux maximum R [bits par dimension] qui est accompli en cartographiant chaque vecteur à l'index $j \equiv f(x^n) \in \mathcal{J}$ en utilisant un quantificateur vectoriel d'une dimension n:

$$f : \mathcal{X}^n \longrightarrow \mathcal{J} \equiv \{1, \ldots, M\} \tag{6.47}$$

$M \leq 2^{nR}$. Bob cherche à récupérer la source X^n à partir de la version compressée reçue. Le but d'Alice est de transmettre la source à Bob de telle manière résultant en une distorsion minimale attendue à Bob. Bob calcule ensuite une séquence de sortie $\hat{X}^n = (\hat{X}_1, \ldots, \hat{X}_n)$ en utilisant la fonction de décodage

$$g_X : \mathcal{J} \mapsto \hat{\mathcal{X}}^n \tag{6.48}$$

Le codeur est choisi de telle sorte que les séquences d'entrée et de sortie atteignent une utilité souhaitée donnée par une contrainte de distorsion attendue

$$D_X = \frac{1}{n}\,\mathbb{E}[d(X^n, g_X(f(X^n)))] = \frac{1}{n}\,\mathbb{E}[d(X^n, \hat{X}^n)] \tag{6.49}$$

$\hat{\mathcal{X}}^n = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_M\}$ est aussi désigné comme l'alphabet de reproduction tel que $g_X(j) = \hat{\mathbf{x}}_j$. $\mathbb{E}$ est l'espérance et $d : \mathcal{X}^n \times \hat{\mathcal{X}}^n \mapsto \mathbb{R}^+$ est une mesure de distorsion.

Le message $f(\mathbf{x})$ est envoyé à Bob via un lien public qui est parfaitement entendu par Eve (pourrait être le collectionneur lui-même). Il est supposé que l'eavesdropper a une connaissance complète de toutes les propriétés du système: les statistiques de la source, le codeur f et le décodeur légitime g_X. Nous supposons un taux d'information mutuelle comme une métrique pour la fuite de confidentialité; cependant, nous permettons le fait qu'une inférence peut être faite à partir de la source X^n sur une autre séquence. Nous modélisons les données inférées comme une variable aléatoire Y^n corrélée à la variable de mesure X^n selon la distribution conjointe $p_{X^nY^n}$. Ainsi, la fuite de confidentialité est l'information mutuelle entre Y^n et le message $f(X^n)$.

$$L_Y = \frac{1}{n} I(Y^n; f(X^n))\ . \tag{6.50}$$

Conception de Quantification

Pour un taux R, le codeur optimal f est obtenu à partir de l'optimisation de 2 types de problèmes de compression avec perte sécurisée. Compte tenu d'une fuite maximale d'information $L_Y \leq \Delta$, minimiser la distorsion D_X. Ou, étant donné une distorsion à Bob $D_X \leq D$, minimiser les fuites d'informations à Eve L_Y.

Au lieu d'utiliser une fonction de coût D_X, avec une contrainte L_Y, ou L_Y, avec une contrainte D_X, nous utilisons la fonction de coût lagrangienne non contrainte

$$\Gamma(f, g_X, \lambda) = D_X + \lambda L_Y\ , \tag{6.51}$$

Où $\lambda > 0$ est le multiplicateur de Lagrange. La distorsion $d(\mathbf{x}, \hat{\mathbf{x}})$ entre une entrée $\mathbf{x}$ et la sortie du décodeur $\hat{\mathbf{x}} = g_X(f(\mathbf{x}))$ est supposée être la distorsion d'erreur au carré. Le MSE est une mesure de distorsion bien connue qui est habituellement utilisée pour estimer la source à partir des valeurs quantifiées correspondantes.

L'ensemble des différents vecteurs d'entrée produisant la même valeur de sortie sera désigné sous le nom de région de quantification. Soit $\mathcal{R}_j$ la région de codage associée à l'index j. $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M\}$ indique la partition de l'espace définissant le mappage $f(\mathbf{x}) = j,\ \text{if}\ \mathbf{x} \in \mathcal{R}_j$.La distorsion attendue peut être ainsi

$$\begin{aligned} D_X &= \frac{1}{n}\int_{\mathcal{X}^n} p_{X^n}(\mathbf{x})\|\mathbf{x} - g_X \circ f(\mathbf{x})\|^2 d\mathbf{x} && (6.52)\\ &= \frac{1}{n}\sum_{j\in\mathcal{J}}\int_{\mathcal{R}_j} p_{X^n}(\mathbf{x})\|\mathbf{x} - \hat{\mathbf{x}}_j\|^2 d\mathbf{x} && (6.53) \end{aligned}$$

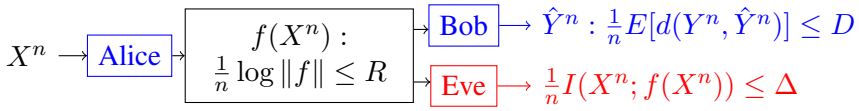


Figure 6.6: Codage Source avec Perte d'une information Corrélée Pertinente avec une Contrainte de Confidentialité.

La fuite de l'information peut être écrite

$$L_Y = \frac{1}{n} I(Y^n; f(X^n)) = \frac{1}{n} H(f(X^n)) - \frac{1}{n} H(f(X^n)|Y^n) \tag{6.54}$$

6.4.3 Codage Source avec Perte d'une Information Corrélée Pertinente avec une Contrainte de Confidentialité

Problem Definition

Dans cette section, nous étudions le cas représenté dans la figure 6.6. Le décodeur n'est pas concerné par la récupération de X^n, mais les informations corrélées pertinentes Y^n du message reçu $f(X^n)$, tout en gardant la source privée de l'eavesdropper. Pour cela, nous définissons la fonction de décodage suivante:

$$g_Y : \mathcal{J} \mapsto \hat{\mathcal{Y}}^n \tag{6.55}$$

où $\hat{\mathcal{Y}}^n = \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_M\}$ est l'alphabet de reproduction dans ce cadre de telle sorte que $g_Y(j) = \hat{\mathbf{y}}_j$. La distorsion moyenne du code est donnée par

$$D_Y = \frac{1}{n}\mathbb{E}[d(Y^n, g_Y(f(X^n)))] = \frac{1}{n}\mathbb{E}[d(Y^n, \hat{Y}^n)] \tag{6.56}$$

Où $d : \mathcal{Y}^n \times \hat{\mathcal{Y}}^n \mapsto \mathbb{R}^+$ est une mesure de distorsion. D'autre part, la fuite d'information à l'eavesdropper devient:

$$L_X = \frac{1}{n} I(X^n; f(X^n)). \tag{6.57}$$

Conception de Quantification

De même, la fonction objective à optimiser peut être écrite

$$\Gamma(f, g_Y, \lambda) = D_Y + \lambda L_X \tag{6.58}$$

La mesure de distorsion $d(y^n, \hat{y}^n)$ entre y^n et la sortie du décodeur $\hat{y}^n = g_Y(f(x^n))$ est également supposée être la distorsion d'erreur au carré.

6.4.4 Application Pratique: Récupération de Consommation de Dispositif Smart Meter sous Contraintes de Confidentialité

Le scénario de cette expérience implique un compteur intelligent dans un ménage et un fournisseur de services. Le compteur intelligent lit la consommation globale et communique les données au fournisseur de services.

Exactement comme auparavant, deux scénarios peuvent être étudiés. Dans le premier scénario, le fournisseur de services cherche à récupérer la consommation totale X^n dans le ménage. Le ménage est disposé à donner la charge globale X^n au fournisseur de services, mais souhaite conserver la fuite d'information concernant un certain appareil Y^n; dire la machine à laver; délimité. Dans le deuxième scénario, le fournisseur de services cherche à récupérer la consommation d'un certain appareil Y^n dans le ménage; dire aussi la machine à laver; pour fournir une utilité à l'utilisateur, par exemple une commande automatisée de la laveuse / sécheuse. Le ménage est également disposé à donner la version encodée de la charge globale X^n au fournisseur de service, mais souhaite conserver la fuite d'information concernant la source X^n bornée.

6.4.5 Conclusion

Le cadre théorique que nous avons développé ici nous permet de quantifier avec précision le problème de l'arbitrage entre l'utilité et la confidentialité dans deux cas. En considérant une série de mesures X^n, nous révélons une perturbation $\hat{X}^n$ qui nous permet de garantir une mesure de la vie privée sur une séquence corrélée Y^n et l'utilité dans X^n dans le premier cas. Dans le second scénario, nous révélons une perturbation $\hat{Y}^n$ qui nous permet de garantir une mesure de la vie privée sur la source X^n et l'utilité sur la séquence corrélée Y^n. La garantie d'utilité vient de la limite supérieure sur la distance de MSE tandis que la garantie de confidentialité vient de la fuite d'information liée.

Le chapitre a également abordé une application importante dans le contexte des réseaux intelligents où les deux scénarios peuvent être utilisés en fonction de l'utilité nécessaire. Les deux expériences; réalisé sur un véritable ensemble de données de compteurs intelligents; montrent que les fuites d'information peuvent être limitées dans le temps tout en maintenant l'utilité. L'importance d'une telle application est la capacité de s'adapter à une statistique d'observation inconnue ou variable. Lorsque le modèle d'observation n'était pas disponible, nous avons pu étendre l'algorithme de conception à utiliser avec les séquences d'entraînement disponibles.

BIBLIOGRAPHY

[1] John G. Proakis. *Digital Communications*. McGraw-Hill, 1995.

[2] H. Nyquist. Certain factors affecting telegraph speed. *Bell System Technical Journal*, 3(2):324–346, 1924.

[3] R. V. L. Hartley. Transmission of information. *Bell System Technical Journal*, 7(3):535–563, 1928.

[4] N. Wiener. *Cybernetics; or, Control and communication in the animal and the machine*. Actualités scientifiques et industrielles. J. Wiley, 1948.

[5] N. Wiener. *Cybernetics Or Control and Communication in the Animal and the Machine*. M.I.T. Pr.paperback.23. M.I.T. Press, 1961.

[6] Claude E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[7] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. Number v. 1 in The Mathematical Theory of Communication. University of Illinois Press, 1949.

[8] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1-4):157–168, 1968.

[9] F.M. Reza. *An Introduction to Information Theory*. Dover Books on Mathematics Series. Dover, 1961.

[10] R.B. Ash. *Information Theory*. Interscience tracts in pure and applied mathematics. Interscience Publishers, 1965.

[11] S. Goldman. *Information Theory*. Dover books on intermediate and advanced mathematics. Dover Publications, 1968.

[12] C. E. Shannon. Communication in the presence of noise. *Proc. Institute of Radio Engineers*, 37(1):10–21, 1949.

[13] R. W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160, 1950.

[14] T.M. Thompson. *From Error-Correcting Codes Through Sphere Packings to Simple Groups*. Number v. 21 in Carus Mathematical Monographs. Mathematical Association of America, 1983.

[15] W. C. Huffman. *Fundamentals of error-correcting codes*. Cambridge University Press, Cambridge, U.K. New York, 2 003.

[16] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall electrical engineering series. Prentice-Hall, 1971.

[17] A. Kolmogorov. On the shannon theory of information transmission in the case of continuous signals. *IRE Transactions on Information Theory*, 2(4):102–108, December 1956.

[18] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE Nat. Conv. Rec., Pt. 4*, pages 142–163. 1959.

[19] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression*. The Springer International Series in Engineering and Computer Science. Springer US, 1991.

[20] H. Abut. *Vector quantization*. IEEE Press selected reprint series. IEEE Press, 1981.

[21] R. M. Gray. Vector quantization. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 75–100. Kaufmann, San Mateo, CA, 1990.

[22] J. C. Kieffer. A survey of the theory of source coding with a fidelity criterion. *IEEE Transactions on Information Theory*, 39(5):1473–1490, Sep 1993.

[23] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, Oct 1998.

[24] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, Mar 1982.

[25] J. Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12, March 1960.

[26] Pier Luigi Dragotti and Michael Gastpar. *Distributed Source Coding: Theory, Algorithms and Applications*. Academic Press, 2009.

[27] B. Girod, A. M. Aaron, S. Rane, and D. Rebollo-Monedero. Distributed video coding. *Proceedings of the IEEE*, 93(1):71–83, Jan 2005.

[28] X. Cao and M. Kuijper. Distributed source coding via linear block codes: A general framework for multiple sources. *IEEE Transactions on Communications*, 60(11):3483–3490, November 2012.

[29] C. Ling, S. Gao, and J. C. Belfiore. Wyner-ziv coding based on multidimensional nested lattices. *IEEE Transactions on Communications*, 60(5):1328–1335, May 2012.

[30] E. Dupraz, A. Roumy, and M. Kieffer. Source coding with side information at the decoder and uncertain knowledge of the correlation. *IEEE Transactions on Communications*, 62(1):269–279, January 2014.

[31] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, Jul 1973.

[32] A. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, Jan 1976.

[33] A. Wyner. On source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 21(3):294–300, May 1975.

[34] R. Zamir. The rate loss in the wyner-ziv problem. *IEEE Transactions on Information Theory*, 42(6):2073–2084, Nov 1996.

[35] DM Green and JA Swets. Signal detection theory and psychophysics. 1966. *New York*, 888:889, 1966.

[36] S.M. Kay. *Fundamentals of Statistical Signal Processing: Detection theory*. Prentice Hall Signal Processing Series. Prentice-Hall PTR, 1998.

[37] Abbas El Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, New York, NY, USA, 2012.

[38] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[39] T. Berger. Decentralized estimation and decision theory. In *presented at theIEEE 7th Spring Workshop Information Theory, Mt. Kisco, NY*, Sept. 1979.

[40] R. Ahlswede and I Csiszar. Hypothesis testing with communication constraints. *Information Theory, IEEE Transactions on*, 32(4):533–542, Jul 1986.

[41] C. E. Shannon. Communication theory of secrecy systems. *The Bell System Technical Journal*, 28(4):656–715, Oct 1949.

[42] Yingbin Liang, H. Vincent Poor, and Shlomo Shamai (Shitz). Information theoretic security. *Found. Trends Commun. Inf. Theory*, 5(4–5):355–580, April 2009.

[43] M. Bloch, J. Barros, M. R.D. Rodrigues, and S. W. McLaughlin. Wireless information-theoretic security. *IEEE Trans. Inf. Theor.*, 54(6):2515–2534, June 2008.

[44] A. D. Wyner. The wire-tap channel. *The Bell System Technical Journal*, 54(8):1355–1387, Oct 1975.

[45] M. Bloch and J. Barros. *Physical-Layer Security: From Information Theory to Security Engineering*. Cambridge University Press, 2011.

[46] I. Csiszar and J. Korner. Broadcast channels with confidential messages. *IEEE Transactions on Information Theory*, 24(3):339–348, May 1978.

[47] J. Ekanayake, N. Jenkins, K. Liyanage, J. Wu, and A. Yokoyama. *Smart Grid: Technology and Applications*. Wiley, 2012.

[48] Greentech Media. PG&E Sued Over Smart Meters, Slows Down Bakersfield Deployment. `http://www.greentechmedia.com/articles/read/pge-sued-over-smartmeters-slows-down-bakersfield-deployment`, 2009.

[49] San Jose Mercury News. PG&E details technical problems with SmartMeters. `http://www.mercurynews.com/ci14963541`, 2010.

[50] Greentech Media. PG&E to Bakersfield: Weather, Not Smart Meters, Cause of Higher Power Bills. `http://www.greentechmedia.com/green-light/post/`

pge-to-bakersfieldweather-not-smart-meters-cause-of-higher-power-bills, 2009.

[51] KGO-TV. Regulators unsupportive of SmartMeters moratorium. http://abclocal.go.com/kgo/story?section=news/7onyourside&id=7410033, 2010.

[52] KGO-TV. Experiment raises questions about SmartMeters. http://abclocal.go.com/kgo/story?section=news/7onyourside&id=7424533, 2010.

[53] Y. Zhang, J. Zhang, J. Ma, and Z. Wang. Fault detection based on discriminant analysis theory in electric power system. In *2009 International Conference on Sustainable Power Generation and Supply*, pages 1–5, April 2009.

[54] Jairo Alonso Ortiz Victor Fidalgo Villar Luis Tarrafeta Elyoenai Egozcue, Daniel Herreras Rodríguez. Smart grid security. Technical report, European Network and Information Security Agency.

[55] Dan Suriyamongkol. Non-technical losses in electrical power systems.

[56] Alfredo Rial and George Danezis. Privacy-preserving smart metering. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*, WPES '11, pages 49–60, New York, NY, USA, 2011. ACM.

[57] D. P. Varodayan and G. X. Gao. Redundant metering for integrity with information-theoretic confidentiality. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 345–349, Oct 2010.

[58] O. Ur-Rehman and N. Zivic. Secure design patterns for security in smart metering systems. In *2015 IEEE European Modelling Symposium (EMS)*, pages 278–283, Oct 2015.

[59] O. Ur-Rehman, N. Zivic, and C. Ruland. Security issues in smart metering systems. In *Smart Energy Grid Engineering (SEGE), 2015 IEEE International Conference on*, pages 1–7, Aug 2015.

[60] Te Han. Hypothesis testing with multiterminal data compression. *IEEE Transactions on Information Theory*, 33(6):759–772, Nov 1987.

[61] S. Jayaraman and T. Berger. An error exponent for lossy source coding with side information at the decoder. In *Information Theory, 1995. Proceedings., 1995 IEEE International Symposium on*, pages 263–, Sep 1995.

[62] Hidetoshi Shimokawa, Te Sun Han, and S.-I Amari. Error bound of hypothesis testing with data compression. In *Information Theory, 1994. Proceedings., 1994 IEEE International Symposium on*, pages 114–, Jun 1994.

[63] M.S. Rahman and AB. Wagner. On the optimality of binning for distributed hypothesis testing. *Information Theory, IEEE Transactions on*, 58(10):6282–6303, Oct 2012.

[64] C. Tian and Jun Chen. Successive refinement for hypothesis testing and lossless one-helper problem. *Information Theory, IEEE Transactions on*, 54(10):4666–4681, Oct 2008.

[65] M.S. Rahman and A.B. Wagner. Vector gaussian hypothesis testing and lossy one-helper problem. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 968–972, June 2009.

[66] Te Han and K. Kobayashi. Exponential-type error probabilities for multiterminal hypothesis testing. *Information Theory, IEEE Transactions on*, 35(1):2–14, Jan 1989.

[67] Yu Xiang and Young-Han Kim. Interactive hypothesis testing with communication constraints. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1065–1072, Oct 2012.

[68] Yu Xiang and Young-Han Kim. Interactive hypothesis testing against independence. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 2840–2844, July 2013.

[69] Joffrey Villard and Pablo Piantanida. Secure lossy source coding with side information at the decoders. *CoRR*, abs/1009.3891, 2010.

[70] F. Naghibi, S. Salimi, and M. Skoglund. The ceo problem with secrecy constraints. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 756–760, June 2014.

[71] Gil Katz, P. Piantanida, R. Couillet, and M. Debbah. Joint estimation and detection against independence. In *52th Annual Allerton Conference on Communication, Control, and Computing*, 2014.

[72] Te Han and S.-I Amari. Statistical inference under multiterminal data compression. *Information Theory, IEEE Transactions on*, 44(6):2300–2324, Oct 1998.

[73] Zuxing Li, T.J. Oechtering, and K. Kittichokechai. Parallel distributed bayesian detection with privacy constraints. In *Communications (ICC), 2014 IEEE International Conference on*, pages 2178–2183, June 2014.

[74] Zuxing Li and T.J. Oechtering. Differential privacy in parallel distributed bayesian detections. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1–7, July 2014.

[75] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local Privacy, Data Processing Inequalities, and Statistical Minimax Rates. *ArXiv e-prints*, February 2013.

[76] Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2nd edition, 2011.

[77] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[78] J. Villard and P. Piantanida. Secure multiterminal source coding with side information at the eavesdropper. *Information Theory, IEEE Transactions on*, 59(6):3668–3692, June 2013.

[79] S.A. Kassam. Optimum quantization for signal detection. *Communications, IEEE Transactions on*, 25(5):479–484, May 1977.

[80] H.V. Poor and John B. Thomas. Applications of ali-silvey distance measures in the design generalized quantizers for binary decision systems. *Communications, IEEE Transactions on*, 25(9):893–900, Sep 1977.

[81] R. R. Tenney and N. R. Sandell. Detection with distributed sensors. *IEEE Transactions on Aerospace and Electronic Systems*, AES-17(4):501–510, July 1981.

[82] G.R. Benitz and J.A. Bucklew. Asymptotically optimal quantizers for detection of i.i.d. *Information Theory, IEEE Transactions on*, 35(2):316–325, Mar 1989.

[83] B. Picinbono and P. Duvaut. Optimum quantization for detection. *IEEE Transactions on Communications*, 36(11):1254–1258, Nov 1988.

[84] M. Longo, T.D. Lookabaugh, and R.M. Gray. Quantization for decentralized hypothesis testing under communication constraints. *Information Theory, IEEE Transactions on*, 36(2):241–255, Mar 1990.

[85] K.R. Varshney and L.R. Varshney. Quantization of prior probabilities for hypothesis testing. *Signal Processing, IEEE Transactions on*, 56(10):4553–4562, Oct 2008.

[86] Y. Zhang, F. Sun, and W. Chen. Optimum quantization for binary hypothesis testing. *IEEE Communications Letters*, 17(8):1501–1504, August 2013.

[87] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *Communication Technology, IEEE Transactions on*, 15(1):52–60, February 1967.

[88] H. V. Poor. A companding approximation for the statistical divergence of quantized data. In *Decision and Control, 1983. The 22nd IEEE Conference on*, pages 697–702, Dec 1983.

[89] H.V. Poor. Fine quantization in signal detection and estimation. *Information Theory, IEEE Transactions on*, 34(5):960–972, Sep 1988.

[90] Michael A. Lexa. Empirical divergence maximization for quantizer design: An analysis of approximation error. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4220–4223, May 2011.

[91] M.A. Lexa. Quantization via empirical divergence maximization. *Signal Processing, IEEE Transactions on*, 60(12):6408–6420, Dec 2012.

[92] Minna Chen, Wei Liu, Biao Chen, and J. Matyjas. Quantization for distributed testing of independence. In *Information Fusion (FUSION), 2010 13th Conference on*, pages 1–5, July 2010.

[93] A. Kashyap, T. Basar, and R. Srikant. Asymptotically optimal quantization for detection in power constrained decentralized sensor networks. In *American Control Conference, 2006*, pages 6 pp.–, June 2006.

[94] B. Lindgren. *Statistical Theory, Fourth Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 1993.

[95] R. Gupta and A.O. Hero. High-rate vector quantization for detection. *Information Theory, IEEE Transactions on*, 49(8):1951–1969, Aug 2003.

[96] J. Villard and P. Bianchi. High-rate vector quantization for the neyman-pearson detection of correlated processes. *IEEE Transactions on Information Theory*, 57(8):5387–5409, Aug 2011.

[97] V. S. S. Nadendla and P. K. Varshney. Design of binary quantizers for distributed detection under secrecy constraints. *IEEE Transactions on Signal Processing*, 64(10):2636–2648, May 2016.

[98] V.S.S. Nadendla, Hao Chen, and P.K. Varshney. Secure distributed detection in the presence of eavesdroppers. In *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*, pages 1437–1441, Nov 2010.

[99] S. Marano, V. Matta, and P.K. Willett. Distributed detection with censoring sensors under physical layer secrecy. *Signal Processing, IEEE Transactions on*, 57(5):1976–1986, May 2009.

[100] Maggie Mhanna and Pablo Piantanida. On secure distributed hypothesis testing. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 1605–1609, June 2015.

[101] Maggie Mhanna, Pablo Piantanida, and Pierre Duhamel. Quantization for distributed binary detection under secrecy constraints. In *Communications (ICC), 2016 IEEE International Conference on*, Kuala Lumpur, Malaysia, Malaysia, May 2016.

[102] Z. Erkin, J.R. Troncoso-Pastoriza, R.L. Lagendijk, and F. Perez-Gonzalez. Privacy-preserving data aggregation in smart metering systems: an overview. *Signal Processing Magazine, IEEE*, 30(2):75–86, March 2013.

[103] J. Zico Kolter and Matthew J. Johnson. REDD: A Public Data Set for Energy Disaggregation Research. In *SustKDD Workshop on Data Mining Applications in Sustainability*, 2011.

[104] Joakim Munkhammar, Jesper Rydén, and Joakim Widén. Characterizing probability density distributions for household electricity load profiles from high-resolution electricity use data. *Applied Energy*, 135(0):382 – 390, 2014.

[105] H. Yamamoto. A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers (corresp.). *IEEE Transactions on Information Theory*, 29(6):918–923, Nov 1983.

[106] H. Yamamoto. A rate-distortion problem for a communication system with a secondary decoder to be hindered. *IEEE Transactions on Information Theory*, 34(4):835–842, Jul 1988.

[107] E. Akyol, C. Langbort, and T. Başar. Privacy constrained information processing. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 4511–4516, Dec 2015.

[108] V. Prabhakaran and K. Ramchandran. On secure distributed source coding. In *Information Theory Workshop, 2007. ITW '07. IEEE*, pages 442–447, Sept 2007.

[109] D. Gunduz, E. Erkip, and H. V. Poor. Lossless compression with security constraints. In *2008 IEEE International Symposium on Information Theory*, pages 111–115, July 2008.

[110] R. Tandon, S. Ulukus, and K. Ramchandran. Secure source coding with a helper. *IEEE Transactions on Information Theory*, 59(4):2178–2187, April 2013.

[111] E. Ekrem and S. Ulukus. Secure lossy source coding with side information. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 1098–1105, Sept 2011.

[112] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[113] I. Csiszar. The method of types [information theory]. *Information Theory, IEEE Transactions on*, 44(6):2505–2523, Oct 1998.

[114] B.G. Kelly and A.B. Wagner. Reliability in source coding with side information. *Information Theory, IEEE Transactions on*, 58(8):5086–5111, Aug 2012.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Apprentissage de Quantificateurs pour la Détection Distribuée Préservant la Confidentialité, avec Application aux Compteurs Intelligents

Mots clés : Détection et Estimation, Communication fiable, Codage de source, Quantification, Réseaux intelligents

Résumé : Cette thèse porte sur quelques problèmes de codage de source dans lesquels on souhaite préserver la confidentialité vis à vis d'une écoute du canal.

Dans la première partie, nous fournissons des nouveaux résultats fondamentaux sur le codage de source pour la détection (utilisateur légitime) et la confidentialité (vis à vis d'une écoute du canal) en présence d'informations secondaires aux terminaux de réception. Nous proposons plusieurs nouveaux résultats d'optimisation de la région de débit-erreur-équivocation réalisable, et proposons des algorithmes pratiques pour obtenir des solutions aussi proches que possible de l'optimal, ce qui nécessite la conception de quantificateurs en présence d'un eavesdropper.

Dans la deuxième partie, nous étudions le problème de l'estimation sécurisée dans un cadre d'utilité-confidentialité où l'utilisateur recherche soit à extraire les aspects pertinents de données complexes ou bien à les cacher vis à vis d'un eavesdropper potentiel.

L'objectif est principalement axé sur l'élaboration d'un cadre général qui combine la théorie de l'information et la théorie de la communication, visant à fournir un nouvel outil pour la confidentialité dans les Smart Grids. D'un point de vue théorique, cette recherche a permis de quantifier les limites fondamentales et donc le compromis entre sécurité et performance (estimation / détection).

Title : Privacy-Preserving Quantization Learning for Distributed Detection with Applications to Smart Meters

Keywords : Detection and Estimation, Privacy, Source coding, Quantization, Smart-Grids

Abstract : This thesis investigates source coding problems in which some secrecy should be ensured with respect to eavesdroppers.

In the first part, we provide some new fundamental results on both detection and secrecy oriented source coding in the presence of side information at the receiving terminals. We provide several new results of optimality and single-letter characterization of the achievable rate-error-equivocation region, and propose practical algorithms to obtain solutions that are as close as possible to the optimal, which requires the design of optimal quantization in the presence of an eavesdropper.

In the second part, we study the problem of secure estimation in a utility-privacy framework where the user is either looking to extract relevant aspects of complex data or hide them from a potential eavesdropper.

The objective is mainly centered on the development of a general framework that combines information theory with communication theory, aiming to provide a novel and powerful tool for security in Smart Grids. From a theoretical perspective, this research was able to quantify fundamental limits and thus the tradeoff between security and performance (estimation/detection).

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France