Information-Theoretic Metrics for Security and Privacy

by

Flavio du Pin Calmon

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

Author ..... Signature redacted .....
Department of Electrical Engineering and Computer Science
August 14, 2015

Certified by ..... Signature redacted .....
Muriel Médard
Cecil H. Green Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by ..... Signature redacted .....
Yury Polyanskiy
Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ..... Signature redacted .....
Leslie A. Kolodziejski
Professor of Electrical Engineering
Chair, Department Committee on Graduate Theses

Information-Theoretic Metrics for Security and Privacy

by

Flavio du Pin Calmon

Submitted to the Department of Electrical Engineering and Computer Science
on August 14, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

In this thesis, we study problems in cryptography, privacy and estimation through the information-theoretic lens. We introduce information-theoretic metrics and associated results that shed light on the fundamental limits of what can be learned from noisy data. These metrics and results, in turn, are used to evaluate and design both symmetric-key encryption schemes and privacy-assuring mappings with provable information-theoretic security guarantees.

We start by studying information-theoretic properties of symmetric-key encryption in the "small key" regime (i.e. when the key rate is smaller than the entropy rate of the message source). It is well known that security against computationally unbounded adversaries in such settings can only be achieved when the communicating parties share a key that is at least as long as the secret message (i.e. plaintext) being communicated, which is infeasible in practice. Nevertheless, even with short keys, we show that a certain level of security can be guaranteed, albeit not perfect secrecy. In order to quantify exactly how much security can be provided with short keys, we propose a new security metric, called symbol secrecy, that measures how much an adversary that observes only the encrypted message learns about individual symbols of the plaintext. Unlike most traditional rate-based information-theoretic metrics for security, symbol secrecy is non-asymptotic. Furthermore, we demonstrate how fundamental symbol secrecy performance bounds can be achieved through standard code constructions (e.g. Reed-Solomon codes).

While much of information-theoretic security has considered the hiding of the plaintext, cryptographic metrics of security seek to hide functions thereof. Consequently, we extend the definition of symbol secrecy to quantify the information leaked about certain classes of functions of the plaintext. This analysis leads to a more general question: can security claims based on information metrics be translated into guarantees on what an adversary can reliably infer from the output of a security system? On the one hand, information metrics usually quantify how far the probability distribution between the secret and the disclosed information is from the ideal case where independence is achieved. On the other hand, estimation guarantees seek to assure that an adversary cannot significantly improve his estimate of the secret given the information disclosed by the system.

We answer this question in the positive, and present formulations based on rate-distortion theory that allow security bounds given in terms of information metrics to be transformed into bounds on how well an adversary can estimate functions of secret variable. We do this by solving a convex program that minimizes the average estimation error over all possible distributions that satisfy the bound on the information metric. Using this approach, we are able to derive a set of general sharp bounds on how well certain classes of functions

of a hidden variable can(not) be estimated from a noisy observation in terms of different information metrics. These bounds provide converse (negative) results: If an information metric is small, then any non-trivial function of the hidden variable cannot be estimated with probability of error or mean-squared error smaller than a certain threshold.

The main tool used to derive the converse bounds is a set of statistics known as the Principal Inertia Components (PICs). The PICs provide a fine-grained decomposition of the dependence between two random variables. Since there are well-studied statistical methods for estimating the PICs, we can then determine the (im)possibility of estimating large classes of functions by using the bounds derived in this thesis and standard statistical tests. The PICs are of independent interest, and are applicable to problems in information theory, statistics, learning theory, and beyond. In the security and privacy setting, the PICs fulfill the dual goal of providing (i) a measure of (in)dependence between the secret and disclosed information of a security system, and (ii) a complete characterization of the functions of the secret information that can or cannot be reliably inferred given the disclosed information. We study the information-theoretic properties of the PICs, and show how they characterize the fundamental limits of perfect privacy.

The results presented in this thesis are applicable to estimation, security and privacy. For estimation and statistical learning theory, they shed light on the fundamental limits of learning from noisy data, and can help guide the design of practical learning algorithms. Furthermore, as illustrated in this thesis, the proposed converse bounds are particularly useful for creating security and privacy metrics, and characterize the inherent trade-off between privacy and utility in statistical data disclosure problems.

The study of security systems through the information-theoretic lens adds a new dimension for understanding and quantifying security against very powerful adversaries. Furthermore, the framework and metrics discussed here provide practical insight on how to design and improve security systems using well-known coding and optimization techniques. We conclude the thesis by presenting several promising future research directions.

Thesis Supervisor: Muriel Médard
Title: Cecil H. Green Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Yury Polyanskiy
Title: Assistant Professor of Electrical Engineering and Computer Science

Acknowledgments

This thesis would not have been possible without the support, love and kindness of many individuals.

First and foremost, I would like to thank the unwaivering patience, advice and encouragement of my thesis supervisor, Muriel Médard. I was blessed to be Muriel's student and a member of her group. Since the very first day of graduate school, Muriel guided me through the fields of information theory, communication, estimation theory and security, teaching me how to identify and appreciate the theoretical beauty behind important practical engineering problems. With the spirit of an information theorist and the heart of an engineer, Muriel gracefully and brilliantly bridges theory and practice. Her insight and vision had a central role in this research. Moreover, Muriel's impact in my life goes well beyond any specific publication, theorem or lemma. Her kindness and cheer gave me the confidence and determination to pursue a career in research. Thank you, Muriel, for believing in me.

I am also indebted to my co-advisor, Prof. Yury Polyanskiy. If research were like skiing, collaborating with Yury is like going down black slopes: the more you do it, the more your technique improves, the faster you go, and the more exciting it becomes. After a while, other slopes become boring in comparison. Working with Yury over the past couple of years improved the rigor and breadth of my research. He guided me through new topics in information theory, and his insight can be seen throughout this thesis.

I am very grateful to all the member of my thesis committee: Mayank Varia, Nadia Fawaz, Ken R. Duffy, and Shafi Goldwasser. It was an absolute pleasure working with Mayank over the past few years. Mayank's open-mindedness and excitement towards navigating uncharted information-theoretic and cryptographic waters were some of the main driving forces behind this work. Many of the results presented here were only possible due to Mayank's brilliant insights, seemingly unbounded cryptographic knowledge, and constant encouragement. For that I cannot thank him enough.

I would like to thank Nadia for exposing me to the fascinating field of privacy. It was during an internship at Technicolor, under Nadia's supervision, that I started working on information theory and privacy. Her guidance was essential to the approach taken in this thesis. In addition, I would also like to thank Nadia for her support during my job search and beyond: throughout the past few years she never said no when I asked for help, and I asked for help many times. Thank you so much, Nadia.

I was extremely fortunate to have the opportunity to collaborate with Ken R. Duffy throughout graduate school. This collaboration not only positively influenced this thesis, but enabled me to be involved with the fruitful research Ken did with Mark Christiansen on Guesswork. Ken played a crucial role not only as research collaborator, but also as a mentor during my time in graduate school. Ken, I am deeply grateful for your advice and support throughout these years, both in academic life and beyond.

I would also like to thank Shafi Goldwasser for being on my committee. I took Shafi's cryptography class at MIT, and it was through meetings with Shafi that we arrived at some of the core ideas that ignited this research. Shafi has always been an inspiration for me during my graduate career, and I am very thankful for the many enjoyable discussions with her.

I acknowledge and thank MIT Lincoln Laboratory for the financial support throughout the past years. Lincoln Lab enabled me to connect and collaborate with many brilliant researchers who influenced this thesis.

I would like to acknowledge all of the members of the Network Coding and Reliable Communications (NCRC) Group. One of the best things about MIT was knowing that every morning I could leave home and walk to 36-512, where I would be surrounded by some of the smartest, friendliest, and coolest people I have ever met. It was a truly special experience to be a part of this group. Yes, we did the stereotypical graduate student things, like hunting for free food, drinking unreasonable amounts of coffee, surviving quals and pulling all-nighters for deadlines. But what I really appreciate are the moments in between: the heart-to-heart conversations, the mutual encouragement, and the fellowship. You all have made me a better person.

I am thankful for the many people who fostered a stimulating and friendly environment during my time at the NCRC, including Dave Adams, Kerim Fouli, Ulric Ferner, Ullrich Monich, Soheil Feizi, Vitaly Abdrashitov, Jinfeng Du, Shirley Shi, Bernhard Haupler, Ali ParandehGeibi, Coleen Josephson, Amy Zhang, and Muriel Rambeloarison, among others. In particular, I would also like to acknowledge and thank Marie Menshova, Salman Salamatian and Ali Makhdoumi, with whom I collaborated and co-authored papers on several of the research topics presented in this thesis. Perhaps the most important member of our group is Michael Lewy. In addition to being a talented artist, Michael efficiently solves hard problems on a daily basis, always being kind and friendly to all of us. Thank you, Michael, for your help and support.

I arrived at MIT in 2009 together with an amazing group of incoming graduate students. Five of these students ended up joining the NCRC group: Arman Rezaee, Georgios (Gorgeous) Angelopoulos, Jason Cloud, Weifei Zeng and Matt Carrey (honorary member). We fought through classes, finals, quals and research together, and your friendship was one of the best things about graduate school. For that, I am thankful. And a special thanks to Weifei Zeng: from research to soccer to hunting platypus in Tasmania, Weifei has been a great friend. Weifei, you are a gentleman and a scholar.

I am grateful for the support and friendship of Raquel Machado and Michelle Sander. They have been my extended family during my time in Boston, and their kindness was a constant presence throughout my graduate career.

I would like to thank Dr. Christin Sander for being the sunshine in my life during these years. Her cheer and love mean everything to me, and were crucial for the success of this research. Thanks to Christin, I arrive at the end of my Ph.D. having learned how to sail, ski and surf, having hiked in a number of national parks in different continents, and having explored parts of the World I never thought I would see. Christin, I'm excited to go on many, many more adventures with you.

I am infinitely grateful to my twin brother and MIT colleague Andre Calmon. His advice, wisdom, friendship, entrepreneurship, encouragement, and twinship strengthen and inspire me on a daily basis. He was with me on my first and last day of school, always supporting me, always lifting me higher and higher. I am blessed to have such an amazing and brilliant researcher and professor as my brother. Andre, you are my guiding star.

Finally, I would like to thank my parents, Katya and Paulo Calmon. It's a long way from Brasil to MIT, and Andre and I would not have been able to make it without your love, guidance, hard work and sacrifice. You always inspired and encouraged us to be the best that we can be. Mãe e Pai, não tenho palavras para agradecer o carinho e o amor que vocês sempre, sempre, sempre nos deram. Cheguei ao final do doutorado graças ao exemplo que vocês sempre foram na minha vida. Eu sei que vocês nos amam independentemente de qualquer diploma, cargo ou salário. Mesmo assim, dedico esta tese à vocês.

Contents

List of Figures

Chapter 1

Introduction

> *From the point of view of the cryptanalyst, a secrecy system is almost identical with a noisy communication system. The message (transmitted signal) is operated on by a statistical element, the enciphering system, with its statistically chosen key. The result of this operation is the cryptogram (analogous to the perturbed signal) which is available for analysis.*
>
> C. E. Shannon, *Communication Theory of Secrecy Systems*, 1949

Information theory enables us to study seemingly disparate engineering problems through a unified methodological lens. The same information-theoretic framework used to characterize the limits of communication can, for example, be adapted to evaluate the security of cryptographic systems, design privacy-assuring mappings for statistical data disclosure, and determine the boundaries of what can or cannot be reliably estimated from noisy data. What makes information theory so versatile is the fact that problems that involve processing, securing, estimating or transmitting information can be captured through closely related canonical models. Information theorists have successfully proven over the past 70 years that, by studying these canonical models, we can have an impact in fields ranging from wireless communication and cryptography to machine learning and distributed data processing.

In this thesis, we introduce new information-theoretic tools to address challenges in cryptography, privacy and estimation. By studying fundamental models that are common to these fields, we derive information-theoretic metrics and associated results that simultaneously (i) delineate the fundamental limits of estimation and (ii) characterize the security properties of cryptographic and privacy-assuring systems. We present an overview of our contributions towards the end of the chapter. First, we take a step back, and provide a few bits of background to contextualize the approach taken here.

We start by describing in Section 1.1 how the insight of using a common information-theoretic approach to analyze seemingly different application areas dates back to the con-

ception of the field by Claude E. Shannon [1, 2]. Shannon realized that the problems of securing and transmitting information are intertwined, and can be studied using the same set of theoretical tools. We then describe in Section 1.2 two parallel views of secrecy: the information-theoretic and the computational view. In this thesis, we adopt the former. In Section 1.3, we discuss how the canonical model behind communication and cryptography is also behind privacy and estimation. This fact enables us to develop information-theoretic results that are simultaneously applicable to these different fields. In Section 1.4, we describe the specific privacy setting considered in this thesis. Section 1.5 presents a roadmap of the thesis and outlines the main contributions. Finally, we conclude the chapter with an overview of the notation used throughout the thesis in Section 1.7.

1.1 Cryptography and Communication: Similar Problems, Different Goals

Cryptography and communication are closely related fields, but with fundamentally different goals. Cryptographers seek to design systems that protect a secret message (also known as the plaintext) from an eavesdropper. This is usually done by either taking advantage of some information asymmetry between the legitimate parties (e.g. sharing a secret key that the eavesdropper does not have access to), or by demonstrating that successful cryptanalysis is equivalent to solving a problem that is believed to be computationally hard. Communication systems, in turn, seek to protect a message against errors caused by a communication channel. Such systems add redundancy to the message in order to provide resilience against noise and, consequently, increase the probability of successful decoding.

The brilliance of Shannon when studying secrecy systems [2] is reflected, at least in part, by his insight that the tools he created for a mathematical theory of communication [1] were also applicable to the cryptographic setting. As stated in the quote in the beginning of this chapter, Shannon realized that the problem of performing cryptanalysis on a ciphertext is fundamentally the same as decoding a message corrupted by noise. Information theory enabled him to analyze both problems from the same theoretical vantage point. According to Shannon [3], "...(cryptography and communication) are very similar things, in one case trying to conceal information, and in the other case trying to transmit it."

The key problem common to both cryptography and communication is illustrated in Fig. 1-1. In cryptography, X plays the role of the plaintext message that is supposed to remain hidden from an eavesdropper. The random variable Y is the ciphertext, produced from the plaintext through some random mapping (e.g. Y is a function of both X and a randomly selected key). The eavesdropper then observes the ciphertext Y, and will perform cryptanalysis in order to guess not only the plaintext X, but properties of the plaintext (e.g. the first bits of the message), denoted here by S. The engineering goal in this setting is to create a random mapping that thwarts cryptanalysis by minimizing the eavesdropper's

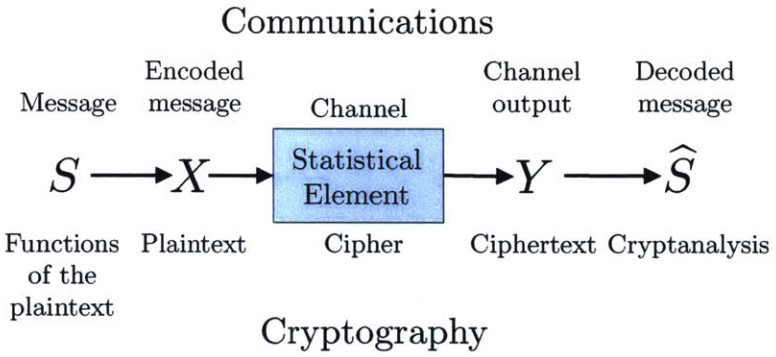

Figure 1-1: The central problem to both cryptography and communication.

ability of estimating S given the ciphertext Y and any other available side information.

In the communication setting, S is the original message, which is channel-encoded into X. The random mapping is given by the communication channel, and will transform the channel input X into the channel output Y. A receiver will then attempt to recover the original message S given an observation of Y. Note that here X may represent a sequence of channel inputs that are transformed into a sequence of channel outputs Y through multiple uses of the channel. The engineering goal is to design a mapping from the message S to the channel input X that maximizes the information that Y carries about X and, equivalently, maximizes that probability that the decoded message $\widehat{S}$ matches the original message S. Observe that here the mapping from S to X is chosen by design, whereas in cryptography the (random) transformation from S to X is given by the distribution of the source of plaintext messages.

Even though cryptography and communication have different design goals, there are fundamental questions that pertain to both settings: How well can X be estimated given an observation of Y? What is the set of functions $S = f(X)$ that can be reliably estimated given Y? How do changes in the random mapping affect the information that Y carries about S? Shannon demonstrated that all of these questions can be addressed using information theory. The setup considered by Shannon for both securing and transmitting information is also at the heart of problems in estimation and privacy, as illustrated in Fig 1-2. We discuss the connection between estimation an privacy in more detail in Section 1.3.

Information theory provides a powerful set of tools for studying problems in communications, cryptography, estimation theory, statistical learning and beyond. By creating a common framework grounded on probability theory, information theory enables us to delineate the fundamental limits of processing, securing and transmitting information. These limits, in turn, have been widely successful as a design driver for practical systems, and have fueled the digital revolution of the last half-century.

In this thesis we follow Shannon's lead, and study problems in cryptography, privacy

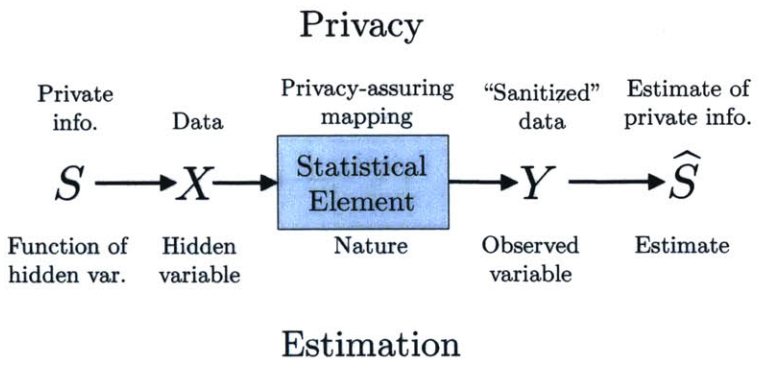

Figure 1-2: The central problem to both estimation and privacy.

and estimation theory through the information-theoretic lens. We introduce information-theoretic metrics and associated results that shed light on the fundamental limits of what can be learned from noisy data. These metrics and results, in turn, are used to evaluate and design both symmetric-key encryption schemes and privacy-assuring mappings with provable information-theoretic security guarantees.

We focus on the question that is central to cryptography, communication, privacy and estimation (illustrated in Figs. 1-1 and 1-2): How well can a random variable S, that is correlated with a hidden variable X, be estimated given an observation of Y? The information-theoretic metrics presented here seek to quantify properties of the random mapping from X to Y that can be translated into bounds on the error of estimating S given an observation of Y. These bounds, often called converse bounds [4], provide universal, algorithm-independent guarantees on what can (or cannot) be learned from Y. With a characterization of these bounds in hand, we then seek to design random mappings that achieve a certain privacy or security goal, usually in terms of how well an adversary can estimate a secret S given the output of the mapping Y. Security provides fertile ground for application of information-theoretic metrics and, as shown in this thesis, it is where converse results carry practical meaning.

Most of the results in this thesis are theoretical in nature, but practical at heart. Our goal is that the many theorems and lemmas presented here serve as a design guideline for practical security and privacy schemes. In the cryptographic setting, we demonstrate how the derived converse bounds can be combined with well-known code constructions to create symmetric-key encryption schemes that provide security guarantees against very powerful eavesdroppers, even when small keys are used. In the context of privacy, we introduce convex formulations that output privacy-assuring mappings that achieve the optimal trade-off between privacy and utility for certain privacy metrics. These mappings, in turn, can be used to create algorithms for distorting data for disclosure with privacy guarantees against statistical inference.

In the next sections, we provide more details on the relationship between information-

theoretic security, modern cryptography, privacy and estimation theory before presenting a detailed overview of the thesis.

1.2 Lunch with Shannon and Turing

In 1943, during the height of the Second World War, Alan Turing visited Bell Labs to consult on secure speech communication projects [5]. Shannon stated in an interview with R. Price in 1982 [6] that Turing and him would frequently have lunch together, but did not discuss cryptography. They usually preferred topics they could discuss freely, such as computing machines and the human brain [3]. Shannon stated that he would often describe to Turing his preliminary notions of information theory. Was the British computer scientist interested in Shannon's seminal ideas? "He was interested," Shannon told Price, "but he didn't believe they were in the right direction. I got a fair amount of negative feedback almost." Despite their disagreement, the ideas of both Shannon and Turing had an enormous impact in cryptography, computing, communication and beyond.

Today, security systems are studied by both information theorists and computer scientists in parallel. However, each community has its own view on secrecy and, in particular, makes different assumptions on the adversarial model[1]. As a result, the security properties of a communication system can, in general, be evaluated from two fundamental perspectives: information-theoretic and computational. The goal of information-theoretic security is to design cryptographic systems with provable security guarantees against adversaries with access to unlimited computing resources and time. Computational security, in turn, seeks to design systems that are secure against adversaries with limited computational resources.

For a noiseless setting, unconditional (i.e. perfect) information-theoretic secrecy can only be attained when the communicating parties share a random key with entropy at least as large as the message itself [2]. Consequently, usual information-theoretic approaches focus on physically degraded models [7], where the goal is to maximize the secure communication *rate* given that the adversary has a noisier observation of the message than the legitimate receiver. On the other hand, computationally secure cryptosystems have thrived from both a theoretical and a practical perspective. Such systems are based on yet unproven hardness assumptions, but nevertheless have led to cryptographic schemes that are widely adopted (for an overview, see [8]). Currently, computationally secure encryption schemes are used millions of times per day, in applications that range from on-line banking transactions to digital rights management.

Computationally secure cryptographic constructions do not necessarily provide an information-theoretic guarantee of security. For example, one-way permutations and public-key encryption cannot be deemed secure against an adversary with unlimited computational resources. This is not to say that such primitives are not secure in practice – real-world adversaries are

[1]We note that these two parallel views are not due to any kind of disagreement between Shannon and Turing, but due to the natural evolution of our understanding of security systems over the past 60 years.

indeed computationally bounded. Alternatively, the traditional[2] "perfect secrecy" definitions adopted in information theory create a rigid framework that usually leads to impractical, or at least unwieldy, security schemes.

In this thesis, and specifically in Chapters 2 and 3, we study information-theoretic properties of symmetric-key encryption schemes in the "small key" regime (i.e. when the key rate is smaller than the entropy rate of the source). In this case, perfect secrecy cannot be attained. Nevertheless, we show that a certain level of information-theoretic security can indeed be guaranteed, albeit that security guarantee is not perfect secrecy. We introduce a new metric to quantify information-theoretic security beyond perfect secrecy, called symbol secrecy, and characterize the class of functions of the plaintext that are information-theoretically hidden for a given level of symbol secrecy. We highlight that, in this analysis, we do not impose computational restrictions on the adversary (the usual approach in modern cryptography), and instead relax the notion of (information-theoretic) security achieved.

The study of cryptographic systems through the information-theoretic lens adds a new dimension for understanding and quantifying security against very powerful adversaries. Furthermore, the framework and metrics discussed here provide insights on how to design and improve security systems using well-known coding techniques. This approach does not seek to replace existing computational security-based methods, but enhance the set of tools available for designing and evaluating security systems.

1.3 Estimation and Security Metrics: Two Sides of the Same Coin

There is a fundamental limit to how much we can learn from data. The problem of determining which functions of a hidden variable can or cannot be estimated from a noisy observation is at the heart of estimation, statistical learning theory [9], and numerous other applications of interest. For example, one of the main goals of prediction is to determine a function of a hidden variable that can be reliably inferred from the output of a system.

Privacy and security applications are concerned with the inverse problem: guaranteeing that a certain set of functions of a hidden variable *cannot* be reliably estimated given the output of a system. Examples of such functions are the identity of an individual whose information is contained in a (supposedly) anonymous dataset [10], sensitive information of a user who joined a database [11, 12], the political preference of a set of users who disclosed their movie ratings [13–15], among others. On the one hand, estimation methods attempt to extract as much information as possible from data. On the other hand, privacy-assuring systems seek to minimize the information about a secret variable that can be reliably estimated from disclosed data. The relationship between privacy and estimation is the same

[2]Perfect secrecy requires independence between the output of a security system Y and the information that is supposed to remain hidden S regardless of the computational resources available to the adversary, and assuming a given threat model in terms of the side information available to the adversary.

as the one noted by Shannon between cryptography and communication: they are connected fields, but with different goals. As illustrated in Fig. 1-2, estimation and privacy are concerned with the same fundamental problem, and can be simultaneously studied through the information-theoretic lens.

Many of the results in this thesis, and particularly Chapters 4 to 7, are situated at the intersection of estimation, privacy and security. We derive a set of general sharp bounds on how well certain classes of functions of a hidden variable can(not) be estimated from a noisy observation. The bounds are expressed in terms of different information metrics of the joint distribution of the hidden and observed variables, and provide converse (negative) results: If an information metric is small, then not only the hidden variable cannot be reliably estimated, but also any non-trivial function of the hidden variable cannot be guessed with probability of error or mean-squared error smaller than a given threshold.

These results are applicable to both estimation and security/privacy. For estimation and statistical learning theory, they shed light on the fundamental limits of learning from noisy data, and can help guide the design of practical learning algorithms. Furthermore, as illustrated in this thesis, the proposed bounds are particularly useful for creating security and privacy metrics, and characterize the inherent trade-off between privacy and utility in statistical data disclosure problems.

The tools used to derive the converse bounds are based on a set of statistics known as the Principal Inertia Components (PICs). The PICs provide a fine-grained decomposition of the dependence between two random variables. Since there are well-studied statistical methods for estimating the PICs [16,17], we can then make claims on the (im)possibility of estimating a large classes of functions by using the bounds derived in this thesis and standard statistical tests. We also demonstrate in Chapter 6 that the PICs play an important role in information theory, and they can be used to characterize the information-theoretic limits of certain estimation problems.

1.4 A Note on Privacy

When referring to privacy in this thesis, we consider the setting studied by Calmon and Fawaz in [18]. Using Fig. 1-2 as reference, we study the problem of disclosing data X to a third-party in order to derive some utility based on X. At the same time, some information correlated with X, denoted by S, is supposed to remain private. The engineering goal is to create a random mapping, called the privacy-assuring mapping, that transforms X into a new data Y that achieves a certain target utility, while minimizing the information revealed about S. For example, X can represent movie ratings that a user intends to disclose to a third-party in order to receive movie recommendations [13–15, 19]. At the same time, the user may want to keep her political preference S secret. We allow the user to distort movie ratings in her data X in order to generate a new data Y. The goal would then be

to find privacy-assuring mappings that minimize the number of distorted entries in Y given a privacy constraint (e.g. the third-party cannot guess S with significant advantage over a random guess). In general, X is not restricted to be the data of an individual user, and can also represent multidimensional data derived from different sources. For more details on designing privacy-assuring mappings and applications with real-world data, we refer the reader to [13–15, 18–20].

From a cryptographic standpoint, this setting can be addressed, at least in part, using secure multi-party computation (MPC) protocols [21]. The goal of MPC protocols is to compute a function over data provided by multiple parties while keeping each individual input private. MPC protocols guarantee that each party learns no more than the output of the computation, and whatever can be inferred about the other parties' inputs given the global output. However, MPC does not guarantee the non-existence of an inference risk from the function output, and individual inputs may be approximately reliably inferred given the output of the function (e.g. when computing a maximum over several inputs, at least one input will be known exactly).

Consequently, unlike MPC, we allow loss in precision of the function computation by permitting the user to distort her data before disclosure. This is particularly relevant in cases where the result of the computation may be inherently tied to the secret information (e.g. documentary recommendations are closely related to the users' political preferences) and is made publicly available. We also do not necessarily assume that the functions of the data that will be computed to provide utility for the user are known a priori (e.g. a company may not want to reveal details about how their recommendation engine works). Instead, we take the information and estimation-theoretic route, and seek to design privacy-assuring mappings that distort the data X in order minimize the information that leaks about S for a given target utility and privacy constraint. Here, the utility constraint acts as a proxy for the functions computed by the service provider, and may be made as specific (or general) as necessary for the application at hand. These privacy-assuring mappings provide privacy guarantees that are independent of computational assumptions. A secondary goal is to characterize the trade-off between privacy and utility (distortion) in this setting. However, our approach does not to subsume MPC, but instead complements the set of tools available for privacy.

Note that this setting is also related to the one studied in the differential privacy literature [11, 12]. Indeed, the traditional differential privacy analysis used in centralized statistical databases can be mapped to this general framework: X can represent a query response over a database, and S a binary variable that indicates if a user joined or not the database. The variable S can also be used to represent a user's individual entry to the database. The goal would then be to distort the query response X (in differential privacy this is often done by adding noise) in order to produce an output Y. We highlight that the setting studied here is more general.

In the next section, we present a more detailed overview of our contributions, and delineate the organization of the rest of the thesis. We then conclude this chapter by introducing the core notation used in the thesis.

1.5 Overview of the Thesis

We start by studying the information-theoretic security properties of symmetric-key encryption schemes with small keys in **Chapter 2**. We consider two communicating parties, namely Alice and Bob, who share a secret key and communicate through a noiseless communication channel. Alice's goal is to securely transmit a plaintext message to Bob. The communication channel is observed by an eavesdropper (Eve), who does not know the key and seeks to recover the plaintext message. Alice and Bob, in turn, wish to minimize the information that Eve gains about the plaintext.

We introduce the concept of list-source codes (LSCs), which are codes that compress a source below its entropy rate. LSCs are a useful tool for understanding how to perform encryption when the length of the randomly selected key is smaller than the entropy of the message source. When the key is small, we use LSC-based analysis to demonstrate how Eve's uncertainty reduces to a near-uniformly distributed list of possible source sequences with an exponential (in terms of the key length) number of elements. We derive fundamental bounds for the rate-list region, and provide code constructions that achieve these bounds. We also illustrate how LSCs can be implemented using standard linear codes.

Furthermore, we present a new information-theoretic metric of security called symbol secrecy, which characterizes the amount of information leaked about specific symbols of the source given an encoded version of the message. We derive fundamental bounds for symbol secrecy, and show how these bounds can be achieved using maximum distance separable (MDS) codes [22] when the source is uniformly distributed.

While symbol secrecy quantifies the information that leaks about individual symbols of the plaintext, most cryptographic metrics seek to characterize the functions of the plaintext that an eavesdropper can (or cannot) reliably estimate. For example, semantic security, introduced in [23], requires that, given an observation of the ciphertext, the eavesdropper cannot guess any function of the plaintext with probability non-negligibly larger than a random guess (i.e. a guess without an observation of the ciphertext). For a precise defintion of semantic security, we refer the reader to [8, Defn. 3.13]. In **Chapter 3**, we extend symbol secrecy to the functional setting by using a rate-distortion-based framework. We first make the key assumption that we know a priori that a certain set of reference functions of the plaintext are hard (or easy) to estimate. Given a target function and the correlation between the target function and the set of reference functions, we then bound the estimation error of the target function in terms of the estimation error of the reference functions.

In the case of symbol secrecy, the set of reference functions are the individual symbols

of the plaintext which, by design, are known to be hard to estimate. We use the aforementioned bound together with Fourier-analytic tools to determine which family of functions of the plaintext can or cannot be reliably estimated. This enables us to map security guarantees in terms of symbol secrecy into guarantees on which functions of the plaintext a computationally unbounded adversary can reliably infer.

The application of the bound derived for the error of estimating a target function given information about a set of reference functions is not restricted to symbol secrecy. We also demonstrate in Chapter 3 that this bound provides insight on how to design symmetric-key encryption schemes that hide specific functions of the plaintext. This approach also sheds light on the fundamental privacy-utility trade-off, described next.

In the privacy setting, the set of reference functions are the features of the data that should be hidden (privacy) or revealed (utility). Here, the bound leads to the following intuitive result: If a certain set of features should be hidden, then all other features of the data that are strongly correlated with it should also be hidden. If the feature that is important for utility is also correlated with another feature that should remain private, then there will be an unfavorable trade-off between privacy and utility. This intuition is captured through the correlation-error product, introduced at the end of the chapter.

The results in Chapter 3 also motivate a more general question: can security claims based on information metrics be translated into guarantees on what an adversary can or cannot reliably infer from the information released by a security system? On the one hand, information metrics usually quantify how far the probability distribution between the secret and the disclosed information is from the ideal case where independence is achieved. On the other hand, estimation guarantees seek to assure that an adversary cannot significantly improve his estimate of the secret given the information disclosed by the system.

This question is answered in **Chapter 4**. We present rate-distortion formulations that allow security bounds based on information metrics to be transformed into bounds on how well an adversary can estimate the secret variable. This is done by solving a convex program that minimizes the average estimation error over all possible distributions that satisfy the bound on the information metric. The solution of this convex program is called the error-rate function. We study extremal properties of error-rate function, and show how to extend the error-rate function to quantify not only the smallest average error of estimating a hidden variable, but also of estimating any function of a hidden variable.

Chapter 4 demonstrates how security guarantees made in terms of an information metric (the usual approach taken in the information-theoretic security literature) can be transformed into guarantees on the adversary's ability to correctly estimate the secret (the desiderata of most modern cryptographic metrics). In **Chapter 5**, we present an information theoretic metric, called the Principal Inertia Components (PICs), that serve both purposes simultaneously.

The PICs give a fine-grained decomposition of the statistical dependence between two

random variables. In the security setting, they provide both (i) a measure of (in)dependence between the secret and disclosed information of a security system, and (ii) a full characterization of the functions of the secret that can or cannot be reliably inferred given the disclosed information. We characterize several information-theoretic properties of the PICs, and derive a converse bound for average estimation error based on the PICs. The PICs also generalize the approach presented in Chapter 3: if the set of reference functions of the plaintext are the singular vectors of the conditional expectation operator between the plaintext and the ciphertext, then the average estimation error (in terms of mean-squared error) of the reference functions is entirely described by the PICs. We also characterize the PICs for a wide range of distributions by analyzing the properties of symmetric functions of sequences of exchangeable random variables.

We study the information-theoretic properties of the PICs in **Chapter 6**, and explore the connection between the PICs and other information-theoretic metrics. We show that, under certain assumptions, the PICs play a central role in estimating a one-bit function of a hidden random variable. This analysis enables us to study and partially resolve an open conjecture on the "most-informative" one-bit function of a uniformly distributed sequence of inputs of an additive binary noise channel [24]. We also show that maximizing the PICs is equivalent to maximizing the first-order term of the Taylor series expansion of certain convex measures of information between the input and the output of a communication channel.

Finally, we apply the PICs to the security and privacy setting in **Chapter 7**. We investigate the problem of intentionally disclosing information about data X (useful information), while guaranteeing that little or no information is revealed about a private variable S (private information). Given that S and X are drawn from finite support sets of the same cardinality, we prove that a non-trivial amount of information about X can be disclosed while not revealing any information about S if and only if the smallest PIC of the joint distribution of S and X is 0. This fundamental result characterizes when useful information can be privately disclosed for any privacy metric based on statistical dependence. We derive sharp bounds for the trade-off between disclosure of useful and private information, and provide explicit constructions of privacy-assuring mappings that achieve these bounds. We conclude Chapter 7 with an example of PIC-based analysis for determining privacy-preserving queries in statistical databases.

1.6 Main Contributions of the Thesis

We believe that the many information-theoretic tools presented here can help guide the design of systems that acquire, process and distribute information while providing reliability, security and privacy guarantees. In particular, we highlight three main contributions of this thesis:

1. **Information-theoretic metrics for secrecy.** We introduce symbol-secrecy and the PICs as information-theoretic metrics for security and privacy. We extend symbol-secrecy to the functional setting by using Fourier-analytic techniques, and derive corresponding fundamental bounds. In addition, we characterize the multiple facets of the PICs, demonstrating how they are the solution of different but related problems in estimation and correlation distillation. We present bounds for the PICs that hold for a wide range of distributions, and introduce an information-theoretic metric based on the PICs called k-correlation.

2. **Bounds on estimation.** We derive converse bounds on estimation error based on the PICs and on symbol secrecy. These results provide lower bounds on (i) the probability of correctly guessing a hidden variable X given an observation Y and (ii) on the minimum mean-squared error of estimating X given Y. These results are stated in terms the PICs between X and Y, and provide universal, algorithm-independent bounds on estimation. We also extend these bounds to the functional setting, and show that the advantage over a random guess of correctly estimating a function of X given an observation of Y is upper-bounded by the largest PIC between X and Y.

3. **Applications to privacy and security.** We apply the proposed security metrics and corresponding converse bounds to symmetric-key encryption and privacy. We demonstrate how symmetric-key encryption schemes that achieve high symbol-secrecy can be created using standard linear code constructions. In addition, we use a PIC-based analysis to characterize the fundamental trade-off between privacy and utility. We show that this analysis, in turn can be used to create privacy-assuring mappings with information-theoretic guarantees. Finally we demonstrate that the smallest PIC determines when perfect privacy can be achieved with non-trivial utility.

1.7 Notation

In this thesis, we adopt the "Just In Time" (JIT) approach for notation, introducing key definitions as they are required in different chapters. Furthermore, we will frequently reintroduce the definition of certain symbols in order to assist the reader. We present the notation that is common to all the chapters of this thesis below.

Capital letters (e.g. X and Y) are used to denote random variables, and calligraphic letters (e.g. $\mathcal{X}$ and $\mathcal{Y}$) denote sets. The exceptions are (i) $\mathcal{I}$, which will be used in Chapter 4 to denote a non-specified information measure, and (ii) T, which will denote the conditional expectation operator (defined below). The support set of random variables X and Y are denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. If X and Y have finite support sets $|\mathcal{X}|< \infty$ and $|\mathcal{Y}|< \infty$, then we denote the joint probability mass function (pmf) of X and Y as $p_{X,Y}$, the conditional pmf of Y given X as $p_{Y|X}$, and the marginal distributions of X and Y as p_X

and p_Y, respectively. We denote the fact that X is distributed according to p_X by $X \sim p_X$. When $p_{X,Y,Z}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y)$ (i.e. X, Y, Z form a Markov chain), we write $X \to Y \to Z$.

For positive integers j, k, n, $j \leq k$, we define $[n] \triangleq \{1, \ldots, n\}$ and $[j,k] \triangleq \{j, j+1, \ldots, k\}$. Matrices are denoted in bold capital letters (e.g. $\mathbf{X}$) and vectors in bold lower-case letters (e.g. $\mathbf{x}$). The (i,j)-th entry of a matrix $\mathbf{X}$ is given by $[\mathbf{X}]_{i,j}$. Furthermore, for $\mathbf{x} \in \mathbb{R}^n$, we let $\mathbf{x} = (x_1, \ldots, x_n)$. We denote by $\mathbf{1}$ the vector with all entries equal to 1, and the dimension of $\mathbf{1}$ will be clear from the context.

A sequence of n random variables $X_1, \ldots, X_n$ is denoted by X^n. Furthermore, for $\mathcal{J} \subseteq [n]$, $X^{\mathcal{J}} \triangleq \left(X_{i_1}, \ldots, X_{i_{|\mathcal{J}|}}\right)$ where $i_k \in \mathcal{J}$ and $i_1 < i_2 < \cdots < i_{|\mathcal{J}|}$. Equivalently, for a vector $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{x}^{\mathcal{J}} \triangleq \left(x_{i_1}, \ldots, x_{i_{|\mathcal{J}|}}\right)$. For two vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$, we denote by $\mathbf{x} \leq \mathbf{z}$ the element-wise set of inequalities $\{x_i \leq z_i\}_{i=1}^n$. We let $\mathcal{P}_t(\mathcal{X})$ be the set of all subsets of $\mathcal{X}$ of size t, i.e. $\mathcal{J} \in \mathcal{P}_t(\mathcal{X}) \Leftrightarrow \mathcal{J} \subseteq \mathcal{X}$ and $|\mathcal{J}| = t$.

For a random variable X with discrete support and $X \sim p_X$, the entropy of X is given by

$$H(X) \triangleq -\mathbb{E}\left[\log\left(p_X(X)\right)\right].$$

If Y has a discrete support set and $X, Y \sim p_{X,Y}$, the mutual information between X and Y is

$$I(X;Y) \triangleq \mathbb{E}\left[\log\left(\frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right)\right].$$

The basis of the logarithm will be clear from the context. The χ^2-information between X and Y is

$$\chi^2(X;Y) \triangleq \mathbb{E}\left[\left(\frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right)\right] - 1.$$

We denote the binary entropy function $h_b : [0,1] \to \mathbb{R}$ as

$$h_b(x) \triangleq -x\log x - (1-x)\log(1-x),$$

where, as usual, $0 \log 0 \triangleq 0$. The inverse of the binary entropy function with input restricted to $[0, 1/2]$ is the mapping $h_b^{-1} : [0, \log 2] \to [0, 1/2]$ where

$$h_b^{-1}(h_b(x)) = \begin{cases} x, & 0 \leq x \leq 1/2 \\ 1-x, & \text{otherwise.} \end{cases}$$

Let X and Y be discrete random variables with finite support sets $\mathcal{X} = [m]$ and $\mathcal{Y} = [n]$, respectively. Then we define the joint distribution matrix $\mathbf{P}$ as an $m \times n$ matrix with $[\mathbf{P}]_{i,j} \triangleq p_{X,Y}(i,j)$. We denote by $\mathbf{p}_X$ (respectively, $\mathbf{p}_Y$) the vector with i-th entry equal to $p_X(i)$ (resp. $p_Y(i)$). $\mathbf{D}_X = \operatorname{diag}(\mathbf{p}_X)$ and $\mathbf{D}_Y = \operatorname{diag}(\mathbf{p}_Y)$ are matrices with diagonal entries equal to $\mathbf{p}_X$ and $\mathbf{p}_Y$, respectively, and all other entries equal to 0. The matrix $\mathbf{P}_{Y|X} \in \mathbb{R}^{m \times n}$ is defined as $[\mathbf{P}_{Y|X}]_{i,j} \triangleq p_{Y|X}(j|i)$. Note that $\mathbf{P} = \mathbf{D}_X\mathbf{P}_{Y|X}$.

For any real-valued random variable X, we denote the L_p-norm of X as

$$\|X\|_p \triangleq \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}.$$

The set of all functions of a random variable $X \sim p_X$ with L_2-norm smaller than 1 is given by

$$\mathcal{L}_2(p_X) \triangleq \{f : \mathcal{X} \to \mathbb{R} \mid \|f(X)\|_2 \leq 1\}.$$

The operators $T_X : \mathcal{L}_2(p_Y) \to \mathcal{L}_2(p_X)$ and $T_Y : \mathcal{L}_2(p_X) \to \mathcal{L}_2(p_Y)$ denote conditional expectation, where

$$(T_X g)(x) = \mathbb{E}\left[g(Y)|X = x\right]$$

and

$$(T_Y f)(y) = \mathbb{E}\left[f(X)|Y = y\right],$$

respectively. We will often omit the subscript X or Y in the definition if the conditioning operation is clear from the context.

For X and Y with discrete support sets, we denote by $P_e(X|Y)$ the smallest average probability of error of estimating X given an observation of Y, defined as

$$P_e(X|Y) = \min_{X \to Y \to \hat{X}} \Pr\{X \neq \hat{X}\},$$

where the minimum is taken over all distributions $p_{\hat{X}|Y}$ such that $X \to Y \to \hat{X}$. The minimum-mean-squared error (MMSE) of estimating X from an observation of Y is given by

$$\mathsf{mmse}(X|Y) \triangleq \min_{X \to Y \to \hat{X}} \mathbb{E}\left[(X - \hat{X})^2\right],$$

where the minimum is again taken over all distributions $p_{\hat{X}|Y}$ such that $X \to Y \to \hat{X}$. Note that, from Jensen's inequality, it is sufficient to consider $\hat{X}$ a deterministic mapping of Y. For any $X \to Y \to g(Y)$ with $\|g(Y)\|_2 = \alpha$ and $\|X\|_2 = \sigma$

$$\begin{aligned}
\mathbb{E}\left[(X - g(Y))^2\right] &= \sigma^2 + \alpha^2 - 2\mathbb{E}\left[Xg(Y)\right] \\
&= \sigma^2 + \alpha^2 - 2\mathbb{E}\left[\mathbb{E}\left[X|Y\right]g(Y)\right] \\
&\geq \sigma^2 + \alpha^2 - 2\|\mathbb{E}\left[X|Y\right]\|_2\|g(Y)\|_2 \\
&= \sigma^2 + \alpha^2 - 2\alpha\|\mathbb{E}\left[X|Y\right]\|_2,
\end{aligned}$$

with equality if and only if $g(Y) \propto \mathbb{E}\left[X|Y\right]$. Minimizing the last expression over all α, we find that the MMSE estimator of X from Y is $g(y) = \mathbb{E}\left[X|Y = y\right]$, and

$$\mathsf{mmse}(X|Y) = \|X\|_2^2 - \|\mathbb{E}\left[X|Y\right]\|_2^2. \tag{1.1}$$

Chapter 2

Symbol Secrecy and List Source Codes

2.1 Overview

We start the study of information-theoretic metrics and their application to security at its historical beginning: symmetric-key encyrption. In this chapter, we present information-theoretic metrics and associated results that seek to characterize the fundamental information-theoretic security properties of symmetric-key encryption schemes when perfect secrecy is not attained. We follow the footsteps of Shannon [2] and Hellman [25], and study symmetric-key encryption with small keys, i.e. when the length of the key is smaller than the length of the message. In this case, the best a computationally unrestricted adversary can do is to decrypt the ciphertext with all possible keys, resulting in a list of possible plaintext messages. The adversary's uncertainty regarding the original message is then represented by a probability distribution over this list. This distribution, in turn, depends on both the distribution of the key and the distribution of the plaintext messages.

Under the assumption that the key is small, perfect secrecy (in the traditional information-theoretic sense) cannot be attained. Consequently, meaningful metrics are required to quantify the level of information-theoretic security provided by the symmetric-key encryption scheme. Towards this goal, we define a new metric for characterizing security, *symbol secrecy*, which quantifies the uncertainty of specific source symbols given an encrypted source sequence. This metric subsumes traditional rate-based information-theoretic measures of secrecy which are generally asymptotic [7]. However, our definition is not asymptotic and, indeed, we provide a construction that achieves fundamental symbol secrecy bounds, based on maximum distance separable (MDS) codes, for finite-length sequences. We note that there has been a long exploration of the connection between coding and cryptography [26], and many of the results presented in this chapter are inscribed in this school of thought.

We also introduce a general source coding framework for analyzing the fundamental

information-theoretic properties of symmetric-key encryption, called *list-source codes* (LSCs). LSCs compress a source sequence *below* its entropy rate and, consequently, a message encoded by an LSC is decoded to a list of possible source sequences instead of a unique source sequence. We demonstrate how any symmetric-key encryption scheme can be cast as an LSC, and prove that the best an adversary can do is to reduce the set of possible messages to an exponentially sized list with certain properties, where the size of the list depends on the length of the key and the distribution of the source. Since the list has a size exponential in the key length, it cannot be resolved in polynomial time in the key length, offering a certain level of computational security. We characterize the achievable ϵ-symbol secrecy of LSC-based encryption schemes, and provide explicit constructions using algebraic coding.

From a practical standpoint, we investigate the problem of secure content caching and distribution. We propose a hybrid encryption scheme based on list-source codes, where a large fraction of the message can be encoded and distributed using a key-independent list-source code. The information necessary to resolve the decoding list, which can be much smaller than the whole message, is then encrypted using a secure method. This scheme allows a significant amount of content to be distributed and cached *before* dealing with key generation, distribution and management issues.

2.2 Main Contributions

We summarize below the main results presented in this chapter.

1. **Symbol secrecy.** We introduce the definitions of absolute and ϵ-symbol secrecy in Section 2.5. Symbol secrecy quantifies the uncertainty that an eavesdropper has about individual symbols of the message.

2. **Encryption with key entropy smaller than the message entropy.** We present the definition of list-source codes (LSCs), together with fundamental bounds, in Section 2.6. Practical code constructions of LSCs are introduced in Section 2.7. We then analyze the symbol secrecy properties of LSCs in Section 2.8.

3. **Applications and practical considerations.** Section 2.9 presents further applications of our results to different security scenarios, together with practical considerations of the proposed secrecy framework. Some of the results presented in the chapter have appeared in [27] and [28].

2.3 Related Work

Shannon's seminal work [2] introduced the use of statistical and information-theoretic metrics for analyzing secrecy systems. Shannon characterized several properties of conditional

entropy (equivocation) as a metric for security, and investigated the effect of the source distribution on the security of a symmetric-key cipher. Shannon also considered the properties of "random ciphers", and showed that, for short keys and sufficiently long, non-uniformly distributed messages, the random cipher is (with high probability) breakable: only one message is very likely to have produced a given ciphertext. Shannon defined the length of the message required for a ciphertext to be uniquely produced by a given plaintext as the *unicity distance.*

Hellman extended Shannon's approach to cryptography [25] and proved that Shannon's random cipher model is conservative: A randomly chosen cipher is likely to have small unicity distance, but does not preclude the existence of other ciphers with essentially infinite unicity distance (i.e. the plaintext cannot be uniquely determined from the ciphertext). Indeed, Hellman argued that carefully designed ciphers that match the statistics of the source can achieve high unicity distance. Ahlswede [29] also extended Shannon's theory of secrecy systems to the case where the exact source statistics are unknown.

The problem of quantifying not only an eavesdropper's uncertainty of the entire message but of individual symbols of the message was studied by Lu in the context of additive-like instantaneous block ciphers (ALIB) [30–32]. The results presented here are more general since we do not restrict ourselves to ALIB ciphers. More recently, the design of secrecy systems with distortion constraints on the adversary's reconstruction was studied by Schieler and Cuff [33]. We adopt here an alternative approach, quantifying the information an adversary gains on average about the individual symbols of the message, and investigate which functions of the plaintext an adversary can reconstruct. Our results and definitions also hold for the finite-blocklength regime.

Tools from algebraic coding have been widely used for constructing secrecy schemes [26]. In addition, the notion of providing security by exploiting the fact that the adversary has incomplete access to information (in our case, the key) is also central to several secure network coding schemes and wiretap models. Ozarow and Wyner [34] introduced the wiretap channel II, where an adversary can observe a set k of his choice out of n transmitted symbols, and proved that there exists a code that achieves perfect secrecy. A generalized version of this model was investigated by Cai and Yeung in [35], where they introduce the related problem of designing an information-theoretically secure linear network code when an adversary can observe a certain number of edges in the network. Their results were later extended in [36–39]. A practical approach was presented by Lima *et al.* in [40]. For a survey on the theory of secure network coding, we refer the reader to [41].

The list-source code framework introduced here is related to the wiretap channel II in that a fraction of the source symbols is hidden from a possible adversary. Oliveira *et al.* investigated in [42] a related setting in the context of data storage over untrusted networks that do not collude, introducing a solution based on Vandermonde matrices. The MDS coding scheme introduced in this paper is similar to [42], albeit the framework developed

here is more general.

List decoding techniques for channel coding were first introduced by Elias [43] and Wozencraft [44], with subsequent work by Shannon *et al.* [45,46] and Forney [47]. Later, algorithmic results for list decoding of channel codes were discovered by Gurusuwami and Sudan [48]. We refer the reader to [49] for a survey of list decoding results. List decoding has been considered in the context of source coding in [50]. The approach is related to the one presented here, since we may view a secret key as side information, but [50] did not consider source coding and list decoding together for the purposes of security.

2.4 Communication and Threat Model

In this chapter, we consider a transmitter (Alice) who wishes to transmit confidentially to a legitimate receiver (Bob) a sequence of length n produced by a discrete source X with alphabet $\mathcal{X}$ and probability distribution p_X. We assume that the communication channel shared by Alice and Bob is noiseless, but is observed by a passive, computationally unbounded eavesdropper (Eve). Both Alice and Bob have access to a shared secret key K drawn from a discrete alphabet $\mathcal{K}$, such that $H(K) < H(X^n)$, and encryption/decryption functions $\mathsf{Enc} : \mathcal{X}^n \times \mathcal{K} \to \mathcal{M}$ and $\mathsf{Dec} : \mathcal{M} \times \mathcal{K} \to \mathcal{X}^n$, where $\mathcal{M}$ is the set encrypted messages. Alice observes the source sequence X^n, and transmits an encrypted message $M = \mathsf{Enc}(X^n, K)$. Bob then recovers X^n by decrypting the message using the key, producing $\hat{X}^n = \mathsf{Dec}(M, K)$. The communication is successful if $\hat{X}^n = X^n$. We consider that the encryption is closed [2, pg. 665], so $\mathsf{Dec}(c, k_1) \neq \mathsf{Dec}(c, k_2)$ for $k_1, k_2 \in \mathcal{K}$, $k_1 \neq k_2$. We assume Eve knows the functions Enc and Dec, but does not know the secret key, K. Eve's goal is to gain knowledge about the original source sequence.

2.5 Symbol Secrecy

In this section we define ϵ-symbol secrecy, an information-theoretic metric for quantifying the information leakage from security schemes that do not achieve perfect secrecy. Given a source sequence X^n and a random variable Z dependent of X^n, ϵ-symbol secrecy is the largest fraction t/n such that, given Z, at most ϵ bits can be learned *on average* from any t-symbol subsequence of X^n. We also prove an ancillary lemma that bounds the average mutual information between X^n and Z in terms of symbol secrecy.

Definition 2.1. Let X^n be a random variable with support $\mathcal{X}^n$, and Z be the information that leaks from a security system (e.g. the ciphertext). Denoting $X^{\mathcal{J}} = \{X_i\}_{i \in \mathcal{J}}$, we say that $p_{X^n,Z}$ achieves an ϵ-symbol secrecy of $\mu_\epsilon(X^n|Z)$ if

$$\mu_\epsilon(X^n|Z) \triangleq \max\left\{ \frac{t}{n} \,\middle|\, \frac{I(X^{\mathcal{J}}; Z)}{|\mathcal{J}|} \leq \epsilon \;\; \forall \mathcal{J} \subseteq [n], 0 < |\mathcal{J}| \leq t \right\}. \tag{2.1}$$

In particular, the *absolute symbol secrecy* of X^n from Y is given by

$$\mu_0(X^n|Z) \triangleq \max\left\{\frac{t}{n}\,\middle|\, I(X^{\mathcal{J}};Z)=0 \;\; \forall \mathcal{J}\subseteq[n], 0<|\mathcal{J}|\leq t\right\}. \tag{2.2}$$

We also define the dual function of symbol-secrecy for X^n and Z as:

$$\epsilon_t^*(X^n|Z) \triangleq \inf\left\{\epsilon\geq 0 \,|\, \mu_\epsilon(X^n|Z)\geq t/n\right\}. \tag{2.3}$$

The next examples illustrate a few use cases of symbol secrecy.

Example 2.1. Symbol secrecy encompasses other definitions of secrecy, such as weak secrecy [51], strong secrecy [52] and perfect secrecy. For example, given two sequences of random variables X^n and Z^n, if $\mu_\epsilon(X^n|Z^n)\to 1$ for all $\epsilon>0$, then $\frac{I(X^n;Z^n)}{n}\to 0$. The converse is not true, as demonstrated in Example 2.3 below. Furthermore, $I(X^n;Z^n)=0$ if and only if $\mu_0(X^n|Z^n)=1$. Finally, the reader can verify that $I(X^n;Z^n)\to 0$ if and only if there exists a sequence $\epsilon_n=o(n)$ such that $\mu_{\epsilon_n}(X^n|Z^n)\to 1$.

Example 2.2. Consider the case where $\mathcal{X}=\{0,1\}$, X^n is uniformly drawn from $\mathcal{X}^n$, and Z is the result of sending X^n through a discrete memoryless erasure channel with erasure probability α. Then, for any $\mathcal{J}\subseteq[n]$, $\mathcal{J}\neq\varnothing$,

$$\frac{I(X^{\mathcal{J}};Z)}{|\mathcal{J}|}=(1-\alpha),$$

and, consequently,

$$\mu_\epsilon(X^n|Z)=\begin{cases}0, & \text{for } 0\leq\epsilon<1-\alpha,\\ 1, & \epsilon\geq 1-\alpha.\end{cases}$$

Example 2.3. Now assume again that X^n is a uniformly distributed sequence of n bits, but now $Z=X_1$. This corresponds to the case where one bit of the message is always sent in the clear, and all the other bits are hidden. Then, for any $\mathcal{J}\subseteq[n]$ such that $\{1\}\in\mathcal{J}$,

$$I(X^{\mathcal{J}};Z)=1,$$

and, for $0\leq\epsilon<1$,

$$\mu_\epsilon(X^n|Z)=0.$$

Consequently, a non-trivial symbol-secrecy cannot be achieved for $\epsilon<1$. In general, if a symbol X_i is sent in the clear, then a non-trivial symbol secrecy cannot be achieved for $\epsilon<H(X_i)$. Note that $I(X^n;Z)/n\to 0$, so weak secrecy is achieved.

Example 2.4. We now illustrate how symbol secrecy does not necessarily capture the information that leaks about functions of X^n. We address this issue in more detail in

Chapter 3. Still assuming that X^n is a uniformly distributed sequence of n bits, let Y be the parity bit of X^n, i.e. $Z = \prod_{i=1}^{n}(-1)^{X_i}$. Then, for any $\mathcal{J} \subsetneq [n]$,

$$I(X^{\mathcal{J}}; Z) = 0,$$

and, for $0 \leq \epsilon < 1$,

$$\mu_\epsilon(X^n|Z) = \frac{n-1}{n},$$

and, for $\epsilon \geq 1$, $\mu_\epsilon(X^n|Z) = 1$.

The following lemma provides an upper bound for $I(X^n; Z)$ in terms of $\mu_\epsilon(X^n|Z)$ when X^n is the output of a discrete memoryless source.

Lemma 2.1. *Let X^n be the output of a discrete memoryless source X, and Z a noisy observation of X^n. For any ϵ such that $0 \leq \epsilon \leq H(X)$, if $\mu_\epsilon(X^n|Z) = u^*$, then*

$$\frac{1}{n} I(X^n; Z) \leq H(X) - u^*(H(X) - \epsilon). \tag{2.4}$$

Proof. Let $\mu_\epsilon(X^n|Z) = u^* \triangleq t/n$, $\mathcal{J} \in \mathcal{P}_t([n])$ and $\bar{\mathcal{J}} = [n] \backslash \mathcal{J}$. Then

$$\begin{aligned}
\frac{1}{n} I(X^n; Z) &= \frac{1}{n} I(X^{\mathcal{J}}; Z) + \frac{1}{n} I(X^{\bar{\mathcal{J}}}; Z|X^{\mathcal{J}}) \\
&\leq \frac{t}{n} \left(\epsilon + \frac{1}{t} I(X^{\bar{\mathcal{J}}}; Z|X^{\mathcal{J}}) \right) \\
&\leq u^* \epsilon + \frac{(n-t)}{n} H(X) \\
&= H(X) - u^*(H(X) - \epsilon),
\end{aligned}$$

where the first inequality follows from the definition of symbol secrecy, and the second inequality follows from the assumption that the source is discrete and memoryless and, consequently, $I(X^{\bar{\mathcal{J}}}; Z|X^{\mathcal{J}}) \leq H(X^{\bar{\mathcal{J}}}|X^{\mathcal{J}}) = (n-t)H(X)$. □

The previous result implies that when $\mu_\epsilon(X^n|Z)$ is large, only a small amount of information about X^n can be gained from Z on average. However, even if $I(X^n; Z)$ is large, as long as $\mu_\epsilon(X^n|Z)$ is non-zero, the uncertainty about X^n given Z will be spread throughout the individual symbols of the source sequence. This property is desirable for symmetric-key encryption and, as we shall show in Chapter 3, can be extended to determine which functions of X^n can or cannot be reliably inferred from Z. Furthermore, in Section 2.8 we introduce explicit constructions for symmetric-key encryption schemes that achieve a provable level of symbol secrecy using the list-source code framework introduced next.

2.6 LSCs

In this section we present the definition of LSCs and derive fundamental bounds. We also demonstrate how any symmetric-key encryption scheme can be mapped to a corresponding list-source code.

2.6.1 Definition and Fundamental Limits

The definition of list-source codes is given below.

Definition 2.2. A $(2^{nR}, |\mathcal{X}|^{nL}, n)$-LSC $(f_n, g_{n,L})$ consists of an encoding function $f_n : \mathcal{X}^n \mapsto \{1, \ldots, 2^{nR}\}$ and a list-decoding function $g_{n,L} : \{1, \ldots, 2^{nR}\} \mapsto \mathcal{P}(\mathcal{X}^n) \backslash \varnothing$, where $\mathcal{P}(\mathcal{X}^n)$ is the power set of $\mathcal{X}^n$ and $|g_{n,L}(w)| = |\mathcal{X}|^{nL}$ $\forall w \in \{1, \ldots, 2^{nR}\}$. The value R is that *rate* of the LSC, L is the *normalized list size*, and $|\mathcal{X}|^{nL}$ is the *list size*.

Note that $0 \leq L \leq 1$. From an operational point of view, L is a parameter that determines the size of the decoded list. For example, $L = 0$ corresponds to traditional lossless compression, i.e., each source sequence is decoded to a unique sequence. Furthermore, $L = 1$ represents the trivial case when the decoded list corresponds to $\mathcal{X}^n$.

For a given LSC, an error is declared when a string generated by a source is not contained in the corresponding decoded list. The average error probability is given by

$$\mathsf{e}(f_n, g_{n,L}) \triangleq \Pr(X^n \notin g_{n,L}(f_n(X^n))). \tag{2.5}$$

Definition 2.3. For a given discrete memoryless source X, the rate list size pair (R, L) is said to be *achievable* if for every $\delta > 0$, $0 < \epsilon < 1$ and sufficiently large n there exists a sequence of $(2^{nR_n}, |\mathcal{X}|^{nL_n}, n)$-list-source codes $\{(f_n, g_{n,L_n})\}_{n=1}^{\infty}$ such that $R_n < R + \delta$, $|L_n - L| < \delta$ and $\mathsf{e}(f_n, g_{n,L_n}) \leq \epsilon$. The *rate list region* is the closure of all rate list pairs (R, L).

Definition 2.4. The *rate list function* $R(L)$ is the infimum of all rates R such that (R, L) is in the rate list region for a given normalized list size $0 \leq L \leq 1$.

Theorem 2.1. *For any discrete memoryless source X, the rate list function is given by*

$$R(L) = H(X) - L \log|\mathcal{X}| \ . \tag{2.6}$$

Proof. Let $\delta > 0$ be given and $\{(f_n, g_{n,L_n})\}_{n=1}^{\infty}$ be a sequence of codes with (normalized) list size L_n such that $L_n \to L$ and for any $0 < \epsilon < 1$ and n sufficiently large $0 \leq \mathsf{e}(f_n, g_{n,L_n}) \leq \epsilon$. Then

$$\Pr\left(X^n \in \bigcup_{w \in \mathcal{W}^n} g_{n,L_n}(w)\right) \geq \Pr\left(X^n \in g_{n,L_n}(f_n(X^n))\right) \geq 1 - \epsilon \tag{2.7}$$

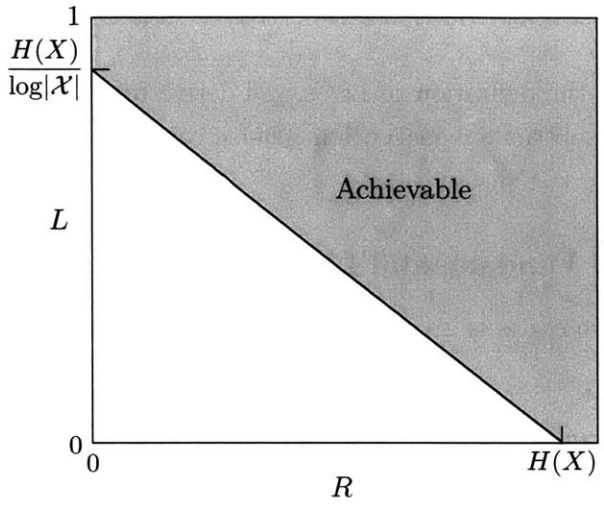

Figure 2-1: Rate list region for normalized list size L and code rate R.

where $\mathcal{W}^n = [2^{nR_n}]$ and R_n is the rate of the code (f_n, g_{n,L_n}). There exists $n_0(\delta, \epsilon, |\mathcal{X}|)$ where if $n \geq n_0(\delta, \epsilon, |\mathcal{X}|)$, then

$$\begin{aligned} R_n + L_n \log|\mathcal{X}| &= \frac{1}{n} \log\left(2^{nR_n}|\mathcal{X}|^{nL_n}\right) \\ &= \frac{1}{n} \log\left(\sum_{w \in \mathcal{W}^n} |g_{n,L_n}(w)|\right) \\ &\geq \frac{1}{n} \log\left|\bigcup_{w \in \mathcal{W}^n} g_{n,L_n}(w)\right| \\ &\geq H(X) - \delta, \end{aligned} \tag{2.8}$$

where the last inequality follows from [53, Lemma 2.14]. Since this holds for any $\delta > 0$, it follows that $R(L) \geq H(X) - L\log|\mathcal{X}|$ for all n sufficiently large.

We prove achievability next. Let $0 < L < 1$ be given, and let $L_n \triangleq \lfloor nL \rfloor / n$. Furthermore, let X^n be a sequence of n source symbols, and denote X^{nL_n} the first nL_n source symbols and $X^{[nL_n+1,n]}$ the last $n(1-L_n)$ source symbols where we assume, without loss of generality, that nL is an integer. Then, from standard source coding results [4, pg. 552], for any $\epsilon > 0$ and n sufficiently large, and denoting $\alpha_n \triangleq \lceil nL_n(H(X)+\epsilon)\rceil/n$, $\beta_n \triangleq \lceil n(1-L_n)(H(X)+\epsilon)\rceil/n$, there are (surjective) encoding functions

$$f^1_{nL} : \mathcal{X}^{nL_n} \to [2^{n\alpha_n}] \text{ and } f^2_{n(1-L_n)} : \mathcal{X}^{n(1-L_n)} \to [2^{n\beta_n}],$$

and corresponding (injective) decoding functions

$$g^1_{n,1} : [2^{n\alpha_n}] \to \mathcal{X}^{nL_n} \text{ and } g^2_{n,1} : [2^{n\beta_n}] \to \mathcal{X}^{nL_n}$$

such that $\Pr(g^1_{n,1}(f^1_{nL_n}(X^{nL_n})) \neq X^{nL_n}) \leq O(\epsilon)$ and $\Pr(g^2_{n,1}(f^2_{n(1-L_n)}(X^{(1-L_n)n})) \neq X^{(1-L_n)n}) \leq O(\epsilon)$.

For $w \in [2^{n\beta_n}]$ and $\mathbf{x} \in \mathcal{X}^n$, let the list-source coding and decoding functions be given by $f_n(\mathbf{x}) \triangleq f^2_{n(1-L_n)}(\mathbf{x}^{[nL_n+1,n]})$ and

$$g_{n,\widetilde{L}_n}(w) \triangleq \{\mathbf{x} \in \mathcal{X}^n : \exists v \in [2^{n\alpha_n}] \text{ such that } (f^1_{nL}(\mathbf{x}^{[nL]}), f^2_{n(1-L)}(\mathbf{x}^{[nL+1,n]})) = (v, w)\},$$

respectively. Then

$$\begin{aligned}\Pr\left(X^n \in g_{n,\widetilde{L}_n}(f_n(X^n))\right) &\geq \Pr\left(g^1_{n,1}(f^1_{nL}(X^{Ln})) = X^{Ln} \wedge g^2_{n,1}(f^2_{n(1-L)}(X^{(1-L)n})) = X^{(1-L)n}\right) \\ &\geq 1 - O(\epsilon).\end{aligned}$$

Observe that the rate-list pair achieved by $(f_n, g_{n,\widetilde{L}_n})$ is $(R_n, \widetilde{L}_n) = (\beta_n, \alpha_n/\log|\mathcal{X}|))$. Consequently,

$$\begin{aligned}R_n &\leq (1 - L_n)(H(X) + \epsilon) + n^{-1} \\ &\leq H(X) + \epsilon - \alpha_n \\ &= H(X) + \epsilon - \widetilde{L}_n \log|\mathcal{X}|,\end{aligned}$$

where the second inequality follows from $\alpha_n \leq L_n(H(X) + \epsilon) + n^{-1}$. Observe that $R_n \to n(1 - L)H(X) + \epsilon \triangleq R$. Since $\widetilde{L}_n \to L(H(X) + \epsilon)/\log|\mathcal{X}| \triangleq \widetilde{L}$ as $n \to \infty$, by choosing n sufficiently large the rate-list pair $(R, \widetilde{L})$ can be achieved, where R and $\widetilde{L}$ satisfy

$$R \leq H(X) + \epsilon - \widetilde{L} \log|\mathcal{X}|.$$

Since ϵ is arbitrary and $\widetilde{L}$ can span any value in $[0, H(X)/\log|\mathcal{X}|]$, it follows that $R(L) \leq H(X) - L\log|\mathcal{X}|$. □

2.6.2 Symmetric-Key Ciphers as LSCs

Let $(\mathsf{Enc}, \mathsf{Dec})$ be a symmetric-key cipher where, without loss of generality, $\mathcal{M} = [2^{nR}]$ and $\mathsf{Enc} : \mathcal{X}^n \times \mathcal{K} \to \mathcal{M}$ and $\mathsf{Dec} : \mathcal{M} \times \mathcal{K} \to \mathcal{X}^n$. Then an LSC can be designed based on this cipher by choosing k' from $\mathcal{K}$ and setting the encoding function $f_n(\mathbf{x}) = \mathsf{Enc}(\mathbf{x}, k')$, where $\mathbf{x} \in \mathcal{X}^n$, and

$$g_{n,L}(f_n(\mathbf{x})) = \{\mathbf{z} \in \mathcal{X}^n : \exists k \in \mathcal{K} \text{ such that } \mathsf{Enc}(\mathbf{z}, k) = f_n(\mathbf{x})\},$$

where L satisfies $|\mathcal{K}| = |\mathcal{X}|^{nL}$. If the key is chosen uniformly from $\mathcal{K}$ then the decoded list corresponds set of possible source sequences that could have generated the ciphertext. The adversary's uncertainty will depend on the distribution of the source sequence X^n.

Alternatively, symmetric-key ciphers can also be constructed based on an $(2^{nR}, |\mathcal{X}|^{nL}, n)$-

list-source code. Let $(f_n, g_{n,L})$ be the corresponding encoding/decoding function of the LSC, and assume that the key is drawn uniformly from $\mathcal{K} = [|\mathcal{X}|^{nL}]$, where the normalized list size L determines the length of the key. Without loss of generality, we also assume that Alice and Bob agree on an ordering of $\mathcal{X}$ and, consequently, $\mathcal{X}^n$ can be ordered using the corresponding dictionary ordering. We denote $\mathsf{pos}(\mathbf{x})$ the position of the source sequence $\mathbf{x} \in \mathcal{X}$ in the corresponding list $g_{n,L}(f_n(\mathbf{x}))$, where $\mathsf{pos} : \mathcal{X}^n \to [|\mathcal{X}|^{nL}]$.

The cipher can then be constructed by letting the message set be $\mathcal{M}' = [2^{nR}] \times [|\mathcal{X}|^{nL}]$ and, for $\mathbf{x} \in \mathcal{X}^n$ and $k \in \mathcal{K}$,

$$\mathsf{Enc}(\mathbf{x}, k) = (f_n(\mathbf{x}), (\mathsf{pos}(\mathbf{x}) + k) \mod |\mathcal{K}|).$$

For $(a, b) \in \mathcal{M}'$, the decryption function is given by

$$\mathsf{Dec}((a, b), k) = \{\mathbf{x} : f_n(\mathbf{x}) = a, \mathsf{pos}(\mathbf{x}) = (b - k) \mod |\mathcal{K}|\}.$$

In this case, an eavesdropper that does not know the key k cannot recover the function $\mathsf{pos}(\mathbf{x})$ and, consequently, her uncertainty will correspond to the list $g_{n,L}(f_n(\mathbf{x}))$.

2.7 LSC Design

In this section we discuss how to construct LSCs that achieve the rate-list tradeoff (2.6) in the finite block length regime. As shown below, an LSC that achieves good rate-list tradeoff does not necessarily lead to good symmetric-key encryption schemes. This naturally motivates the constructions of LSCs that achieve high symbol secrecy.

2.7.1 Necessity for Code Design

Assume that the source X is uniformly distributed in $\mathbb{F}_q$, i.e., $\Pr(X = x) = 1/q \ \forall x \in \mathbb{F}_q$. In this case $R(L) = (1 - L) \log q$. A trivial scheme for achieving the list-source boundary is the following. Consider a source sequence $X^n = (X^p, X^s)$, where X^p denotes the first $p = n - \lfloor Ln \rfloor$ symbols of X^n and X^s denotes the last $s = \lfloor Ln \rfloor$ symbols. Encoding is done by discarding X^s, and mapping the prefix X^p to a binary codeword Y^{nR} of length $nR = \lceil n - \lfloor Ln \rfloor \log q \rceil$ bits. This encoding procedure is similar to the achievability scheme used in the proof of Theorem 2.1.

For decoding, the codeword Y^{nR} is mapped to X^p, and the scheme outputs a list of size q^s composed by X^p concatenated with all possible combinations of suffixes of length s. Clearly, for n sufficiently large, $R \approx (1 - L) \log q$, and we achieve the optimal list-source size tradeoff.

The previous scheme is inadequate for security purposes. An adversary that observes the codeword Y^{nR} can uniquely identify the first p symbols of the source message, and the uncertainty is concentrated over the last s symbols. Assuming that all source symbols are of

equal importance, we should spread the uncertainty over all symbols of the message. Given the encoding $f(X^n)$, a sensible security scheme would provide $I(X_i; f(X^n)) \leq \epsilon \ll \log q$ for $1 \leq i \leq n$. We can naturally extend this notion for groups of symbols or functions over input symbols, which is what symbol secrecy captures.

2.7.2 A Construction Based on Linear Codes

Let X be an i.i.d. source with support $\mathcal{X}$ and entropy $H(X)$, and (s_n, r_n) a source code for X with encoder $s_n : \mathcal{X}^n \to \mathbb{F}_q^{m_n}$ and decoder $r_n : \mathbb{F}_q^{m_n} \to \mathcal{X}^n$. Furthermore, let $\mathcal{C}$ be a (m_n, k_n, d) linear code[1] over $\mathbb{F}_q$ with an $(m_n - k_n) \times m_n$ parity check matrix $\mathbf{H}_n$ (i.e. $\mathbf{c} \in \mathcal{C} \Leftrightarrow \mathbf{H}_n\mathbf{c} = 0$). Consider the following scheme, where we assume

$$k_n \triangleq nL_n \log|\mathcal{X}|/\log q$$

is an integer, $0 \leq L_n \leq 1$ and $L_n \to L$ as $n \to \infty$.

Scheme 2.1. *Encoding*: Let $\mathbf{x}_n \in \mathcal{X}^n$ be an n-symbol sequence generated by the source. Compute the syndrome $\boldsymbol{\sigma}_n$ through the matrix multiplication

$$\boldsymbol{\sigma}_n \triangleq \mathbf{H}_n s_n(\mathbf{x}_n)$$

and map each syndrome to a distinct sequence of $nR = \lceil (m_n - k_n) \log q \rceil$ bits, denoted by $\mathbf{y}_{nR}$.

Decoding: Map the binary codeword $\mathbf{y}_{nR}$ to the corresponding syndrome $\boldsymbol{\sigma}_n$. Output the list

$$g_{n,L_n}(\boldsymbol{\sigma}_n) = \left\{ r_n(\mathbf{z}) \middle| \mathbf{z} \in \mathbb{F}_q^{m_n},\ \boldsymbol{\sigma}_n = \mathbf{H}_n\mathbf{z} \right\}.$$

Theorem 2.2. *If a sequence of source codes $\{(s_n, r_n)\}_{n=1}^{\infty}$ is asymptotically optimal for source X, i.e. $m_n/n \to H(X)/\log q$ with vanishing error probability, scheme 2.1 achieves the rate list function $R(L)$ for source X.*

Proof. Since the cardinality of each coset corresponding to a syndrome $\boldsymbol{\sigma}_n$ is exactly

$$|g_{n,L_n}(\boldsymbol{\sigma}_n)| = q^{k_n},$$

the normalized list size is

$$L_n = \log_{|\mathcal{X}|} q^{k_n} = (k_n \log q)/(n \log|\mathcal{X}|).$$

By assumption, $L_n \to L$ as $n \to \infty$. Denoting $m_n/n = H(X)/\log q + \delta_n$, where $\delta_n \to 0$ since the source code is assumed to be asymptotically optimal, it follows that the rate of

[1] For an overview of linear codes and related terminology, we refer the reader to [22].

the LSC is

$$\begin{aligned} R_n &= \lceil (m_n - k_n) \log q \rceil / n \\ &= \lceil (H(X) + \delta_n \log q) n - L_n n \log |\mathcal{X}| \rceil / n \\ &\to H(X) - L \log |\mathcal{X}|, \end{aligned}$$

which is arbitrarily close to the rate in (2.6) for sufficiently large n. □

The source coding scheme used in the proof of Theorem 2.2 can be any asymptotically optimal scheme. Note that if the source X is uniformly distributed in $\mathbb{F}_q$, then $L_n = k_n/n$ and any message in the coset indexed by $\boldsymbol{\sigma}_n$ is equally likely. Hence, $R_n = (n-k)\log q/n = H(X) - L\log q$, which matches the upper bound in (2.6). Scheme 2.1 provides a constructive way of hiding information, and we can take advantage of the properties of the underlying linear code to make precise assertions regarding the security of the scheme.

With the syndrome in hand, how can we recover the rest of the message? One possible approach is to find a $k_n \times n$ matrix $\mathbf{D}_n$ that has full rank such that the rows of $\mathbf{D}_n$ and $\mathbf{H}_n$ form a basis of $\mathbb{F}_q^{m_n}$. Such a matrix can be easily found, for example, using the Gram-Schmidt process with the rows of $\mathbf{H}_n$ as a starting point. Then, for a source sequence $\mathbf{x}_n$, we simply calculate $\mathbf{t}_n = \mathbf{D}_n \mathbf{x}_n$ and forward $\mathbf{t}_n$ to the receiver through a secure channel. The receiver can then invert the system

$$\begin{pmatrix} \mathbf{H}_n \\ \mathbf{D}_n \end{pmatrix} \mathbf{x}_n = \begin{pmatrix} \boldsymbol{\sigma}_n \\ \mathbf{t}_n \end{pmatrix}, \tag{2.9}$$

and recover the original sequence $\mathbf{x}_n$. This property allows list-source codes to be deployed in practice using well known linear code constructions, such as Reed-Solomon [22, Chap. 5] or Random Linear Network Codes [54, Chap. 2].

Remark 2.1. This approach is valid for general linear spaces, and holds for any pair of full rank matrices $\mathbf{H}_n$ and $\mathbf{D}_n$ with dimensions $(m_n - k_n) \times m_n$ and $k_n \times m_n$, respectively, such that $\text{rank}([\mathbf{H}_n^T \ \mathbf{D}_n^T]^T) = m_n$. However, here we adopt the nomenclature of linear codes since we make use of known code constructions to construct LSCs with provable symbol secrecy properties in the next section.

Remark 2.2. The LSC described in scheme 2.1 can be combined with other encryption methods, providing, for example, an additional layer of security in probabilistic encryption schemes ([8,23]). A more detailed discussion of practical applications is presented in Section 2.9.

2.8 Symbol Secrecy of LSCs

We next present fundamental bounds for the amount of symbol secrecy achievable by any LSC considering a discrete memoryless source. Since any encryption scheme can be cast as an LSC, these results quantify the amount of symbol secrecy achievable by any symmetric-key encryption scheme that encrypts a discrete memoryless source.

Lemma 2.2. *Let $\{(f_n, g_n)\}_{n=1}^{\infty}$ be a sequence of list-source codes that achieves a rate-list pair (R, L) and an ϵ-symbol secrecy of $\mu_\epsilon\left(X^n|Y^{nR_n}\right) \to \mu_\epsilon$ as $n \to \infty$. Then $0 \leq \mu_\epsilon \leq \min\left\{\frac{L\log|\mathcal{X}|}{H(X)-\epsilon}, 1\right\}$.*

Proof. We denote $\mu_\epsilon(X^n|Y^{nR}) = \mu_{\epsilon,n}$. Note that, for $\mathcal{J} \subseteq [n]$ and $|\mathcal{J}| = n\mu_{\epsilon,n}$,

$$\begin{aligned} I(X^{\mathcal{J}}; Y^{nR_n}) &= H(X^{\mathcal{J}}) - H(X^{\mathcal{J}}|Y^{nR_n}) \\ &= n\mu_{\epsilon,n}H(X) - H(X^{\mathcal{J}}|Y^{nR_n}) \\ &\leq n\mu_{\epsilon,n}\epsilon, \end{aligned}$$

where the last inequality follows from the definition of symbol secrecy and $I(X^{\mathcal{J}}; Y^{nR_n}) \leq |\mathcal{J}|\epsilon = n\mu_{\epsilon,n}\epsilon$. Therefore

$$\begin{aligned} \mu_{\epsilon,n}(H(X) - \epsilon) &\leq \frac{1}{n}H(X^{\mathcal{J}}|Y^{nR_n}) \\ &\leq L_n \log|\mathcal{X}|. \end{aligned}$$

The result follows by taking $n \to \infty$. □

The previous result bounds the amount of information an adversary gains about particular source symbols by observing a list-source encoded message. In particular, for $\epsilon = 0$, we find a meaningful bound on what is the largest fraction of input symbols that is *perfectly* hidden.

The next theorem relates the rate-list function with ϵ-symbol secrecy through the upper bound in Theorem 2.2.

Theorem 2.3. *If a sequence of list-source codes $\{(f_n, g_{n,L_n})\}_{n=1}^{\infty}$ achieves a point (R', L) with $\mu_\epsilon(X^n|Y^{nR_n}) \to \frac{L\log|\mathcal{X}|}{H(X)-\epsilon} \triangleq c_\epsilon$ for some ϵ, where $R' = \lim_{n\to\infty}\frac{1}{n}H(Y^{nR_n})$, then $R' = R(L)$.*

Proof. Assume that $\{(f_n, g_{n,L_n})\}_{n=1}^{\infty}$ satisfies the conditions in the theorem and $\delta > 0$ is given. Then for n sufficiently large, we have from (2.4):

$$\begin{aligned} \frac{1}{n}H(Y^{nR_n}) &= \frac{1}{n}I(X^n; Y^{nR_n}) \\ &\leq H(X) - c_\epsilon(H(X) - \epsilon) + \delta \\ &= H(X) - L\log|\mathcal{X}| + \delta. \end{aligned}$$

Since this holds for any δ, then $R' \leq H(X) - L\log|\mathcal{X}|$. However, from Theorem 2.1, $R' \geq H(X) - L\log|\mathcal{X}|$, and the result follows. □

2.8.1 A Scheme Based on MDS Codes

We now prove that for a uniform i.i.d. source X in $\mathbb{F}_q$, using scheme 2.1 with an MDS parity check matrix $\mathbf{H}$ achieves μ_0. Since the source is uniform and i.i.d., no source coding is used.

Proposition 2.1. *If* $\mathbf{H}$ *is the parity check matrix of an* (n, k, d) *MDS code and the source* X^n *is uniform and i.i.d., then Scheme 2.1 achieves the upper bound* $\mu_0 = L$*, where* $L = k/n$*.*

Proof. Let $\mathcal{C}$ be the set of codewords of an $(n, k, n-k+1)$ MDS code over $\mathbb{F}_q$ with parity matrix $\mathbf{H}$, and let $\mathbf{x} \in \mathcal{C}$. Fix a set $\mathcal{J} \in \mathcal{P}_k([n])$ of k positions of $\mathbf{x}$, denoted $\mathbf{x}^{\mathcal{J}}$. Since the minimum distance of $\mathcal{C}$ is $n-k+1$, for any other codeword in $\mathbf{z} \in \mathcal{C}$ we have $\mathbf{z}^{\mathcal{J}} \neq \mathbf{x}^{\mathcal{J}}$. Denoting by $\mathcal{C}^{\mathcal{J}} = \{\mathbf{x}^{\mathcal{J}} \in \mathbb{F}_q^k : x \in \mathcal{C}\}$, then $|\mathcal{C}^{\mathcal{J}}| = |\mathcal{C}| = q^k$. Therefore, $\mathcal{C}^{\mathcal{J}}$ contains all possible combinations of k symbols. Since this property also holds for any coset of $\mathbf{H}$, the result follows. □

We present next a general description of a two-phase secure communication scheme for the threat model described in Section 2.4, presented in terms of the list-source code constructions derived using linear codes. Note that this scheme can be easily extended to any list-source code by using the corresponding encoding/decoding functions instead of multiplication by parity check matrices.

2.9 Discussion

In this section we discuss the application of our results to different security settings.

2.9.1 A Secure Communication Scheme Based on List-Source Codes

We assume that Alice and Bob have access to a symmetric-key encryption/decryption scheme $(\mathsf{Enc}', \mathsf{Dec}')$ that is used with the shared secret key K and is sufficiently secure against the adversary. This scheme can be, for example, a one-time pad. The encryption/decryption procedure is performed as follows, and will be used as components of the overall encryption scheme $(\mathsf{Enc}, \mathsf{Dec})$ described below.

Scheme 2.2. *Input*: The source encoded sequence $\mathbf{x} \in \mathbb{F}_q^n$, parity check matrix $\mathbf{H}$ of a linear code in $\mathbb{F}_q^n$, a full-rank $k \times n$ matrix $\mathbf{D}$ such that $\text{rank}([\mathbf{H}^T\ \mathbf{D}^T]) = n$, and encryption/decryption functions $(\mathsf{Enc}', \mathsf{Dec}')$. We assume both Alice and Bob share a secret key K.

Encryption (Enc):

Phase I (pre-caching): Alice generates $\boldsymbol{\sigma} = \mathbf{Hx}$ and sends to Bob.[2]

Phase II (send encrypted data): Alice generates $\mathbf{e} = \mathsf{Enc}'(\mathbf{Dx}, K)$ and sends to Bob.

Decryption (Dec): Bob calculates $\mathsf{Dec}'(\mathbf{e}, K) = \mathbf{Dx}$ and recovers $\mathbf{x}$ from $\boldsymbol{\sigma}$ and $\mathbf{Dx}$.

Assuming that $(\mathsf{Enc}', \mathsf{Dec}')$ is secure, the information-theoretic security of Scheme 2.2 reduces to the security of the underlying list-source code (i.e. Scheme 2.1). In practice, the encryption/decryption functions $(\mathsf{Enc}', \mathsf{Dec}')$ may depend on a secret or public/private key, as long as it provide sufficient security for the desired application. In addition, assuming that the source sequence is uniform and i.i.d. in F_q^n, we can use MDS codes to make strong security guarantees, as described in the next section. In this case, an adversary that observes $\boldsymbol{\sigma}$ cannot infer *any* information about any set of k symbols of the original message.

Note that this scheme has a *tunable* level of secrecy: The amount of data sent in phase I and phase II can be appropriately selected to match the properties of the encryption scheme available, the size of the key length, and the desired level of secrecy. Furthermore, when the encryption procedure has a higher computational cost than the list-source encoding/decoding operations, list-source codes can be used to reduce the total number of operations required by allowing encryption of a smaller portion of the message (phase II).

The protocol outline presented in Scheme 2.2 is useful in different practical scenarios, which are discussed in the following sections. Most of the advantages of the suggested scheme stem from the fact that list-source codes are key-independent, allowing content to be distributed when a key distribution infrastructure is not yet established, and providing an additional level of security if keys are compromised before phase II in Scheme 2.2.

2.9.2 Content Pre-Caching

As hinted earlier, list-source codes provide a secure mechanism for content pre-caching when a key infrastructure has not yet been established. A large fraction of the data can be list-source coded and securely transmitted before the termination of the key distribution protocol. This is particularly significant in large networks with hundreds of mobile nodes, where key management protocols can require a significant amount of time to complete [55]. Scheme 2.2 circumvents the communication delays incurred by key compromise detection, revocation and redistribution by allowing data to be efficiently distributed concurrently with the key distribution protocol, while maintaining a level of security determined by the underlying list-source code.

[2]Here, Alice can use message authentication codes and public key encryption to augment security. Furthermore, the list-source coding scheme can be used as an additional layer of security with information-theoretic guarantees in symmetric-key ciphers. Since we are interested in the information-theoretic security properties of the scheme, we will not go into further details. We recognize that in order to use this scheme in practice additional steps are needed to meet modern cryptographic standards.

2.9.3 Application to Key Distribution Protocols

List-source codes can also provide additional robustness to key compromise. If the secret key is compromised before phase II of Scheme 2.2, the data will still be as secure as the underlying list-source code. Even if a (computationally unbounded) adversary has perfect knowledge of the key, until the last part of the data is transmitted the best he can do is reduce the number of possible inputs to an exponentially large list. In contrast, if a stream cipher based on a pseudo-random number generator were used and the initial seed was leaked to an adversary, all the data transmitted up to the point where the compromise was detected would be vulnerable. The use of list-source codes provide an additional, information-theoretic level of security to the data up to the point where the last fraction of the message is transmitted. This also allows decisions as to which receivers will be allowed to decrypt the data can be delayed until the very end of the transmission, providing more time for detection of unauthorized receivers and allowing a larger flexibility in key distribution.

In addition, if the level of security provided by the list-source code is considered sufficient and the key is compromised before phase II, the key can be redistributed *without the need of retransmitting the entire data.* As soon as the keys are reestablished, the transmitter simply encrypts the remaining part of the data in phase II with the new key.

2.9.4 Additional Layer of Security

We also highlight that list-source codes can be used to provide an additional layer of security to the underlying encryption scheme. The message can be list-source coded after encryption and transmitted in two phases, as in Scheme 2.2. As argued in the previous point, this provides additional robustness against key compromise, in particular when a compromised key can reveal a large amount of information about an incomplete message (e.g. stream ciphers). Consequently, list-source codes are a simple, practical way of augmenting the security of current encryption schemes.

One example application is to combine list-source codes with stream ciphers. The source-coded message can be initially encrypted using a pseudorandom number generator (PRG) initialized with a randomly selected seed, and then list-source coded. The initial random seed would be part of the encrypted message sent in the final transmission phase. This setup has the advantage of augmenting the security of the underlying stream cipher, and provides randomization to the list-source coded message. In particular, if the LSC is based on MDS codes and assuming that the distribution of the plaintext is nearly uniform, strong information-theoretic symbol secrecy guarantees can be made about the transmitted data, as discussed in Section 2.5. Even if the underlying PRG is compromised, the message would still be secure.

2.9.5 Tunable Level of Secrecy

List-source codes provide a tunable level of secrecy, i.e. the amount of security provided by the scheme can be adjusted according to the application of interest. This can be done by appropriately selecting the size of the list (L) of the underlying code, which determines the amount of uncertainty an adversary will have regarding the input message. In the proposed implementation using linear codes, this corresponds to choosing the size of the parity check matrix $\mathbf{H}$, or, analogously, the parameters of the underlying error-correcting code. In terms of Scheme 2.2, a larger (respectively smaller) value of L will lead to a smaller (larger) list-source coded message in phase I and a larger (smaller) encryption burden in phase II.

2.10 Prologue to Chapter 3

While much of information-theoretic security has considered the hiding of the plaintext, cryptographic metrics of security seek to hide also functions thereof [23]. More specifically, cryptographic metrics characterize how well an adversary can (or cannot) infer functions of a hidden variable, and are stated in terms of lower bounds for average estimation error probability. This contrasts with standard information-theoretic metrics of security, which are concerned with the average number of bits that an adversary learns about the plaintext. Nevertheless, as we will show next, restrictions on the average mutual information can be mapped to lower bounds on average estimation error probability through rate-distortion formulations.

In the next chapter, we use a rate-distortion based approach to extend the definition of symbol secrecy in order to quantify not only the information that an adversary gains about individual symbols of the source sequence, but also the information gained about functions of the encrypted source sequence. We prove that ciphers with high symbol secrecy guarantee that certain functions of the plaintext are provably hidden regardless of computational assumptions. In particular, we show that certain one-bit function of the plaintext (i.e. predicates) cannot be reliably inferred by the adversary. The estimation-theoretic approach that we use naturally leads to a more fundamental information-theoretic quantity called principal inertia components, which is studied in detail in the latter chapters of this thesis.

Chapter 3

A Rate-Distortion View of Symbol Secrecy

3.1 Overview

Symbol secrecy provides a fine-grained metric for quantifying the amount of information that leaks from a security system. However, standard cryptographic definitions of security are concerned not only with what an eavesdropper learns about individual symbols of the plaintext, but also which *functions* of the plaintext an adversary can reliably infer. In order to derive analogous information-theoretic metrics for security, in this chapter we take a step back from the symmetric-key encryption setup and study the general estimation problem of inferring properties of a hidden variable X from an observation Y. More specifically, we derive lower bounds for the error of estimating functions of X from an observation of Y. By using standard converse results (e.g. Fano's inequality [4, Chap. 2]), symbol secrecy guarantees are then translated to guarantees on how well certain functions of the plaintext can or cannot be estimated.

We first derive converse bounds for the minimum-mean-squared-error (MMSE) of estimating a function ϕ of the hidden variable X given Y. We assume that the MMSE of estimating a set of functions $\Phi \triangleq \{\phi_j(X)\}_{i=1}^m$ given Y is known, as well as the correlation between $\phi_j(X)$ and $\phi(X)$. Bounds for the MMSE of $\phi(X)$ are then expressed in terms of the MMSE of each $\phi_j(X)$ and the correlation between $\phi(X)$ and $\phi_j(X)$. We also apply this result to the setting where ϕ and ϕ_j are binary functions, and present bounds for the probability of correctly guessing $\phi(X)$ given Y. These results are of independent interest, and are particularly useful in the security setting considered here.

The set of functions Φ can be used to model known properties of a security system. For example, when X is a plaintext and Y is a ciphertext, the functions ϕ_j may represent certain predicates of X that are known to be hard to infer given Y. In privacy systems, X may be a user's data and Y a distorted version of X generated by a privacy preserving mechanism.

The set Φ could then represent a set of functions that are known to be easy to infer from Y due to inherent utility constraints of the setup. In particular, in Section 3.7 we consider the set Φ as subsets of symbols of the plaintext. In this case, the results introduced in this chapter are used to derive bounds on the MMSE of reconstructing a target function of the plaintext in terms of the symbol-secrecy achieved by the underlying list-source code given by the encryption scheme.

We illustrate the application of our results both for hiding the source data and functions thereof. We provide an extension of the one-time pad [2] to a functional setting, demonstrating how certain classes of functions of the plaintext can be hidden using a short key. We also consider the privacy against statistical inference setup studied in [18], and show how the analysis introduced here sheds light on the fundamental privacy-utility tradeoff.

The results presented in this chapter can also be viewed through the linear operator theory lens used for characterizing the Principal Inertia Components (PICs) of X and Y, introduced in Chapter 5. Indeed, one of the main purposes of this chapter is to introduce the estimation (MMSE)-based view of security and privacy, which will then be extended in the latter chapters of this thesis. More specifically, when the set of functions Φ form a basis of the space of functions of X with finite second moment and, in addition, correspond to the singular vectors of the conditional expectation operator, then the analysis presented here reduces to the PIC-based analysis explained in Chapter 5. Nevertheless, the results in this chapter are of independent interest and serve as a motivation for the analysis introduced in Chapter 5 onwards.

3.2 Main Contributions

We summarize below the main contributions of this chapter to the thesis. Several of the results in this chapter have appeared in [28] and [56].

1. **MMSE bounds.** We introduce a lower bound for the MMSE of estimating a target function ϕ of a hidden variable X given that a certain set of functions Φ are known to be easy or hard to infer in Theorem 3.1. This bound is based on Lemma 3.1, and extended to bound the probability of estimating one-bit functions of X in Corollary 3.1. This analysis is later generalized using the PICs in Chapter 5.

2. **Extending symbol secrecy.** We translate bounds on symbol secrecy into upper bounds on the mutual information between functions of a plaintext and the ciphertext in Theorem 3.4. The proof of this result makes use of Fourier analytic techniques for Boolean functions. These techniques will also be applied in Chapter 6.

3. **Applications: Generalization of the one-time pad and the correlation-error product.** The more practical-facing results in this chapter are (i) a generalization

of the one-time pad for small keys, presented in Theorem 3.3 and (ii) the correlation-error product, introduced in Section 3.8. In the former, we show that a wide range of functions can be information-theoretically hidden using a one-time pad like scheme for "short" keys (i.e. keys with entropy rate significantly smaller than the plaintext source rate) by appropriately choosing the distribution of the key. This result is also proved using Fourier-analytic techniques. In the latter, we use the MMSE-based analysis to introduce the correlation-error product, which is a key component for understanding the fundamental tradeoff between privacy and utility. The correlation-error product will be studied again in Chapter 7 using the PICs.

3.3 Related Work

The use of rate-distortion formulations in security and privacy settings was studied by Yamamoto [57] and Reed [58]. Information-theoretic approaches to privacy that take distortion into account were also considered in [59–61]. We will make use of the privacy against statistical inference framework at the end of this chapter, introduced in [18] and later extended in [14] and [20]. For additional references on information-theoretic security, we refer the reader to Section 2.3.

Bellare *et al.* [62] considered the standard wiretap setting [7], and proved the equivalence between semantic security and minimizing the maximum mutual information over all possible input message distributions. Since semantic security [23] is achieved only when an adversary's advantage of correctly computing a function of the hidden variable given an observation of the output is negligibly small, the results in [62] are closely related to the ones presented here and in Chapter 4.

3.4 Lower Bounds for MMSE

The results introduced in this chapter are based on the following Lemma.

Lemma 3.1. *Let* $z_n : (0,\infty)^n \times [0,1]^n \to \mathbb{R}$ *be given by*

$$z_n(\mathbf{a},\mathbf{b}) \triangleq \max\left\{\mathbf{a}^T\mathbf{y} \middle| \mathbf{y} \in \mathbb{R}^n, \|y\|_2 \leq 1, \mathbf{y} \leq \mathbf{b}\right\}. \tag{3.1}$$

Let π *be a permutation of* $(1,2,\ldots,n)$ *such that* $b_{\pi(1)}/a_{\pi(1)} \leq \ldots \leq b_{\pi(n)}/a_{\pi(n)}$. *If* $b_{\pi(1)}/a_{\pi(1)} \geq 1$, $z_n(\mathbf{a},\mathbf{b}) = \|\mathbf{a}\|_2$. *Otherwise,*

$$z_n(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{k^*} a_{\pi(i)}b_{\pi(i)} + \sqrt{\left(\|\mathbf{a}\|_2^2 - \sum_{i=1}^{k^*} a_{\pi(i)}^2\right)\left(1 - \sum_{i=1}^{k^*} b_{\pi(i)}^2\right)}, \tag{3.2}$$

where

$$k^* \triangleq \max \left\{ k \in [n] \;\middle|\; \frac{b_{\pi(k)}}{a_{\pi(k)}} \leq \sqrt{\frac{\left(1 - \sum_{i=1}^{k-1} b_{\pi(i)}^2\right)^+}{\|\mathbf{a}\|_2^2 - \sum_{i=1}^{k-1} a_{\pi(i)}^2}} \right\}. \tag{3.3}$$

Proof. The proof is given in the appendix. □

Throughout this section we assume $\Phi \subseteq \mathcal{L}_2(p_X)$ and $\mathbb{E}[\phi_i(X)\phi_j(X)] = 0$ for $i \neq j$. Furthermore, let Y be an observed variable that is dependent of X, and for a given ϕ_i the inequality

$$\max_{\psi \in \mathcal{L}_2(p_Y)} \mathbb{E}[\phi_i(X)\psi(Y)] = \|\mathbb{E}[\phi_i(X)|Y]\|_2 \leq \lambda_i$$

is satisfied, where $0 \leq \lambda_i \leq 1$. This is equivalent to $\mathsf{mmse}(\phi_i(X)|Y) \geq 1 - \lambda_i^2$. Recall that we define the operator $T_X : \mathcal{L}_2(p_Y) \to \mathcal{L}_2(p_X)$ (respectively $T_Y : \mathcal{L}_2(p_X) \to \mathcal{L}_2(p_Y)$) as the conditional expectation operator given X (resp. given Y) that maps $\psi(y) \to \mathbb{E}[\psi(y)|X = x]$ (resp. $\phi(x) \to \mathbb{E}[\psi(Y)|X = x]$).

Theorem 3.1. *Let* $|\mathbb{E}[\phi(X)\phi_i(X)]| = \rho_i > 0$. *Denoting* $\boldsymbol{\rho} \triangleq (|\rho_1|, \ldots, |\rho_m|)$, $\boldsymbol{\lambda} \triangleq (\lambda_1, \ldots, \lambda_m)$, $\rho_0 \triangleq \sqrt{1 - \sum_{i=1}^{k} \rho_i^2}$, $\lambda_0 = 1$ $\boldsymbol{\rho}_0 \triangleq (\rho_0, \boldsymbol{\rho})$ *and* $\boldsymbol{\lambda}_0 \triangleq (\lambda_0, \boldsymbol{\lambda})$, *then*

$$\|\mathbb{E}[\phi(X)|Y]\|_2 \leq B_{|\Phi|}(\boldsymbol{\rho}_0, \boldsymbol{\lambda}_0), \tag{3.4}$$

where

$$B_{|\Phi|}(\boldsymbol{\rho}_0, \boldsymbol{\lambda}_0) \triangleq \begin{cases} z_{|\Phi|+1}(\boldsymbol{\rho}_0, \boldsymbol{\lambda}_0), & \text{if } \rho_0 > 0, \\ z_{|\Phi|}(\boldsymbol{\rho}, \boldsymbol{\lambda}), & \text{otherwise.} \end{cases} \tag{3.5}$$

and z_n *is given in* (3.1). *Consequently,*

$$\mathsf{mmse}(\phi(X)|Y) \geq 1 - B_{|\Phi|}(\boldsymbol{\rho}_0, \boldsymbol{\lambda}_0)^2. \tag{3.6}$$

Proof. Let $h(X) \triangleq \rho_0^{-1}(\phi(X) - \sum_i \rho_i \phi_i(X))$ if $\rho_0 > 0$, otherwise $h(X) = 0$. Note that $h(X) \in \mathcal{L}_2(p_X)$. Then for $\psi \in \mathcal{L}_2(p_Y)$

$$\begin{aligned} |\mathbb{E}[\phi(X)\psi(Y)]| &= \left| \rho_0 \mathbb{E}[h(X)\psi(Y)] + \sum_{i=1}^{m} \rho_i \mathbb{E}[\phi_i(X)\psi(Y)] \right| \\ &\leq \rho_0 |\mathbb{E}[h(X)\psi(Y)]| + \sum_{i=1}^{m} |\rho_i \mathbb{E}[\phi_i(X)\psi(Y)]| \\ &= \rho_0 |\mathbb{E}[h(X)(T_X\psi)(X)]| + \sum_{i=1}^{m} |\rho_i \mathbb{E}[\phi_i(X)(T_X\psi)(X)]|. \end{aligned}$$

Denoting $|\mathbb{E}[h(X)(T_X\psi)(X)]| \triangleq x_0$, $|\mathbb{E}[\phi_i(X)(T_X\psi)(X)]| \triangleq x_i$, $\mathbf{x} \triangleq (x_0, x_1, \ldots, x_m)$, and

$\boldsymbol{\rho} \triangleq (\rho_0, |\rho_1|, \ldots, |\rho_m|)$, the last inequality can be rewritten as

$$|\mathbb{E}[\phi(X)\psi(Y)]| \leq \boldsymbol{\rho}_0^T \mathbf{x}. \tag{3.7}$$

Observe that $\|\mathbf{x}\|_2 \leq 1$ and $x_i \leq \lambda_i$ for $i = 0, \ldots, m$, and the right hand side of (3.7) can be maximized over all values of $\mathbf{x}$ that satisfy these constraints. We assume, without loss of generality, that $\rho_0 > 0$ (otherwise set $x_0 = 0$). The left-hand side of (3.7) can be further bounded by

$$|\mathbb{E}[\phi(X)\psi(Y)]| \leq z_{m+1}(\boldsymbol{\rho}_0, \boldsymbol{\lambda}_0), \tag{3.8}$$

where $\boldsymbol{\lambda} = (1, \lambda_1, \ldots, \lambda_m)$ and z_{m+1} is defined in (3.1). The result follows directly from Lemma 3.1 and noting that $\max_{\psi \in \mathcal{L}_2(p_Y)} \mathbb{E}[\phi(X)\psi(Y)] = \|\mathbb{E}[\phi(X)|Y]\|_2$. □

Denote $\psi_i(y) \triangleq \mathbb{E}[\phi_i(X)|Y = y] / \|\mathbb{E}[\phi_i(X)|Y]\|_2$ and $\phi_0(X) \triangleq (\phi(X) - \sum_{i=1}^m \rho_i \phi_i(X))/\rho_0^{-1}$. The previous bound can be further improved when $\mathbb{E}[\psi_i(Y)\phi_j(X)] = 0$ for $i \neq j$, $j \in \{0, \ldots, m\}$.

Theorem 3.2. *Let* $|\mathbb{E}[\phi(X)\phi_i(X)]| = \rho_i > 0$ *for* $\phi_i \in \Phi$. *In addition, assume* $\mathbb{E}[\psi_i(Y)\psi_j(Y)] = 0$ *for* $i \neq j$, $i \in [t]$ *and* $j \in \{0, \ldots, |\Phi|\}$, *where* $0 \leq t \leq |\Phi|$. *Then*

$$\|\mathbb{E}[\phi(X)|Y]\|_2 \leq \sqrt{\sum_{k=1}^{t} \lambda_i^2 \rho_i^2 + B_{|\Phi|-t}\left(\widetilde{\boldsymbol{\rho}}, \widetilde{\boldsymbol{\lambda}}\right)^2}, \tag{3.9}$$

where $\widetilde{\boldsymbol{\rho}} = (\rho_0, \rho_t, \ldots, \rho_m)$, $\widetilde{\boldsymbol{\lambda}} = (1, \lambda_t, \ldots, \lambda_m)$ *and* B_m *is defined in* (3.5) *(considering* $B_0 = 0$*). In particular, if* $t = m$,

$$\|\mathbb{E}[\phi(X)|Y]\|_2 \leq \sqrt{\rho_0^2 + \sum_{k=1}^{|\Phi|} \lambda_i^2 \rho_i^2}, \tag{3.10}$$

and this bound is tight when $\rho_0 = 0$. *Furthermore,*

$$\mathsf{mmse}(\phi(X)|Y) \geq 1 - \sum_{k=1}^{t} \lambda_i^2 \rho_i^2 - B_{|\Phi|-t}\left(\widetilde{\boldsymbol{\rho}}, \widetilde{\boldsymbol{\lambda}}\right)^2. \tag{3.11}$$

Proof. For any $\psi \in \mathcal{L}_2(p_Y)$, let $\alpha_i \triangleq \mathbb{E}[\psi(Y)\psi_i(Y)]$ and $\psi_0(Y) \triangleq (\psi(Y) - \sum_{i=1}^t \alpha_i \psi_i(Y))\alpha_0^{-1}$, where $\alpha_0 = (1 - \sum_{i=i}^t \alpha_i^2)^{-1/2}$. Observe that $\psi_0 \in \mathcal{L}_2(p_Y)$ and $\mathbb{E}[\phi_i(X)\psi_j(Y)] = \mathbb{E}[\psi_i(Y)\psi_j(Y)] = 0$ for $i \neq j$, $i \in \{0, \ldots, |\Phi|\}$ and $j \in [t]$. Consequently

$$\begin{aligned}
\mathbb{E}[\phi(X)\psi(Y)] &= \mathbb{E}\left[\left(\sum_{i=0}^{|\Phi|} \rho_i \phi_i(X)\right)\left(\sum_{j=0}^{t} \alpha_j \psi_j(Y)\right)\right] \\
&= \sum_{i=0}^{|\Phi|} \sum_{j=0}^{t} \rho_i \alpha_j \mathbb{E}[\phi_i(X)\psi_j(Y)]
\end{aligned}$$

$$\leq \left| \alpha_0 \sum_{i=0, i\notin[n]}^{|\Phi|} \rho_i \mathbb{E}\left[\phi_i(X)\psi_0(Y)\right] \right| + \sum_{i=1}^{t} |\lambda_i \rho_i \alpha_i|$$

$$\leq |\alpha_0| B_{|\Phi|-t}\left(\widetilde{\boldsymbol{\rho}}, \widetilde{\boldsymbol{\lambda}}\right) + \sum_{i=1}^{t} |\lambda_i \rho_i \alpha_i| \tag{3.12}$$

$$\leq \sqrt{\sum_{i=1}^{t} \lambda_i^2 \rho_i^2 + B_{|\Phi|-t}\left(\widetilde{\boldsymbol{\rho}}, \widetilde{\boldsymbol{\lambda}}\right)^2}. \tag{3.13}$$

Inequality (3.12) follows from the bound (3.4), and (3.13) follows by observing that $\sum_{i=0}^{t} \alpha_i^2 = 1$ and applying the Cauchy-Schwarz inequality.

Finally, when $\rho_0 = 0$, (3.13) can be achieved with equality by taking $\psi = \sum_i \frac{\lambda_i \rho_i}{\sqrt{\sum_i \lambda_i^2 \rho_i^2}} \psi_i$. □

Remark 3.1. The previous theorem and, in particular, (3.10) and (3.11) forshadows the Principal Inertia Component-based analysis that will be introduced in Chapter 5. More specifically, when the functions ϕ_i for a basis for $\mathcal{L}_2(p_X)$ and are the singular vectors of the operator T_X (with corresponding adjoint operator T_Y), then the values λ_i^2 in (3.10) are exactly the PICs of the joint distribution $p_{X,Y}$.

The following three, diverse examples illustrate different usage cases of Theorems 3.1 and 3.2. Example 3.1 illustrates Theorem 3.2 for the binary symmetric channel. In this case, the basis Φ can be conveniently expressed as the parity bits of the input to the channel. Example 3.2 illustrates how Theorem 3.2 can be applied to the q-ary symmetric channel, and demonstrates that bound (3.10) is sharp. Finally, Example 3.3 then illustrates Theorem 3.1 for the specific case where all the values ρ_i and λ_i are equal.

Example 3.1 (Binary Symmetric Channel). Let $\mathcal{X} = \{-1, 1\}$ and $\mathcal{Y} = \{-1, 1\}$, and Y^n be the result of passing X^n through a memoryless binary symmetric channel with crossover probability ϵ. We also assume that X^n is composed by n uniform and i.i.d. bits. For $\mathcal{S} \subseteq [n]$, let $\chi_{\mathcal{S}}(X^n) \triangleq \prod_{i\in\mathcal{S}} X_i$. Any function $\phi : \mathcal{X} \to \mathbb{R}$ can then be decomposed in terms of the basis of functions $\chi_{\mathcal{S}}(X^n)$ as [63]

$$\phi(X^n) = \sum_{\mathcal{S}\subseteq[n]} c_{\mathcal{S}} \chi_{\mathcal{S}}(X^n),$$

where $c_{\mathcal{S}} = \mathbb{E}\left[\phi(X^n)\chi_{\mathcal{S}}(X^n)\right]$. Furthermore, since $\mathbb{E}\left[\chi_{\mathcal{S}}(X^n)|Y^n\right] = (1-2\epsilon)^{|\mathcal{S}|}$, it follows from Theorem 3.2 that

$$\mathsf{mmse}(\phi(X^n)|Y^n) = 1 - \sum_{\mathcal{S}\subseteq[n]} c_{\mathcal{S}}^2 (1-2\epsilon)^{2|\mathcal{S}|}. \tag{3.14}$$

This result can be generalized for the case where $X^n = Y^n \otimes Z^n$, where the operation $\otimes$ denotes bit-wise multiplication, Z^n is drawn from $\{-1, 1\}^n$ and X^n is uniformly distributed.

In this case

$$\mathsf{mmse}(\phi(X^n)|Y^n) = 1 - \sum_{\mathcal{S}\subseteq[n]} c_{\mathcal{S}}^2 \mathbb{E}\left[\chi_{\mathcal{S}}(Z^n)\right]^2. \tag{3.15}$$

This example will be revisited in Section 3.6, where we restrict ϕ to be a binary function.

Example 3.2 (q-ary symmetric channel). For $\mathcal{X} = \mathcal{Y} = [q]$, an (ϵ, q)-ary symmetric channel is defined by the transition probability

$$p_{Y|X}(y|x) = (1-\epsilon)\mathbb{1}_{y=x} + \epsilon/q. \tag{3.16}$$

Any function $\phi_i \in \mathcal{L}_2(p_X)$ such that $\mathbb{E}\left[\phi_i(X)\right] = 0$ satisfies

$$\psi_i(Y) = T_Y\phi(X) = (1-\epsilon)\phi(Y),$$

and, consequently, $\|T_Y\phi(X)\|_2 = (1-\epsilon)$. We shall use this fact to show that the bound (3.10) is sharp in this case.

Observe that for $\phi_i, \phi_j \in \mathcal{L}_2(p_X)$, if $\mathbb{E}\left[\phi_i(X)\phi_j(X)\right] = 0$ then $\mathbb{E}\left[\psi_i(Y)\psi_j(Y)\right] = 0$. Now let $\phi \in \mathcal{L}_2(p_X)$ satisfy $\mathbb{E}\left[\phi(X)\right] = 0$ and $\mathbb{E}\left[\phi(X)\phi_i(X)\right] = \rho_i$ for $\phi_i \in \Phi$, where $|\Phi| = m$, Φ satisfies the conditions in Theorem 3.2, and $\sum_i \rho_i^2 = 1$. In addition, $\|\psi_i\|_2 = (1-\epsilon) = \lambda_i$. Then, from (3.10),

$$\|T_Y\phi(X)\|_2 \leq \sqrt{\sum_{i=1}^{m} \lambda_i^2\rho_i^2} = (1-\epsilon)\sqrt{\sum_i \rho_i^2} = 1-\epsilon,$$

which matches $\|T_Y\phi(X)\|_2$, and the bound is tight in this case.

Example 3.3 (Equal MMSE and correlation). We now turn our attention to Theorem 3.1. Consider the case when the correlations of ϕ with the references functions ϕ_i are all the same, and each ϕ_i can be estimated with the same MMSE, i.e. $\lambda_1 = \cdots = \lambda_m = \lambda$ and $\rho_1^2 = \cdots = \rho_m^2 = \rho^2$, $\rho \geq 0$ and $\lambda^2 \leq \rho^2 \leq 1/m$. Then bound (3.4) becomes

$$\|\mathbb{E}\left[\phi(X)|Y\right]\|_2 \leq m\lambda\rho + \sqrt{(1-m\rho^2)(1-m\lambda^2)}.$$

3.5 One-Bit Functions

Let X be a hidden random variable and Y be a noisy observation of X. Here we denote $\Phi = \{\phi_i\}_{i=1}^m$ a collection of m predicates of X, where $F_i = \phi_i(X)$, $\phi_i : \mathcal{X} \to \{-1, 1\}$ for $i \in [m]$ and, without loss of generality $\mathbb{E}\left[F_i\right] = b_i \geq 0$.

We denote by $\hat{F}_i$ an estimate of F_i given an observation of Y, where $F_i \to X \to Y \to \hat{F}_i$. We assume that for any $\hat{F}_i$

$$\left|\mathbb{E}[F_i\hat{F}_i]\right| \leq 1 - 2\alpha_i$$

for some $0 \leq \alpha_i \leq (1 - b_i)/2 \leq 1/2$. This condition is equivalent to imposing that $\Pr\{F_i \neq \hat{F}_i\} \geq \alpha_i$, since

$$\mathbb{E}\left[F_i \hat{F}_i\right] = \Pr\{F_i = \hat{F}_i\} - \Pr\{F_i \neq \hat{F}_i\} = 1 - 2\Pr\{F_i \neq \hat{F}_i\}.$$

In particular, this captures how well F_i can be guessed based solely on an observation of Y.

Now assume there is a bit $F = \phi(Y)$ such that $\mathbb{E}[FF_i] = \rho_i$ for $i \in [m]$ and $\mathbb{E}[F_i F_j] = 0$ for $i \neq j$. We can apply the same method used in the proof of Theorem 3.1 to bound the probability of F being guessed correctly from an observation of Y.

Corollary 3.1. *For* $\lambda_i = 1 - 2\alpha_i$,

$$\Pr(F \neq \hat{F}) \geq \frac{1}{2}\left(1 - B_{|\Phi|}(\boldsymbol{\rho}, \boldsymbol{\lambda})\right). \tag{3.17}$$

Proof. The proof follows the same steps as Theorem 3.1, $\phi(Y) \in \mathcal{L}_2(p_Y)$. □

In the case $m = 1$, we obtain the following simpler bound, presented in Proposition 3.1, which depends on the following Lemma.

Lemma 3.2. *For any random variables* A, B *and* C

$$\Pr(A \neq B) \leq \Pr(A \neq C) + \Pr(B \neq C).$$

Proof.

$$\begin{aligned}
\Pr(A \neq B) &= \Pr(A \neq B \wedge B = C) + \Pr(A \neq B \wedge B \neq C) \\
&= \Pr(A \neq C \wedge B = C) + \Pr(B \neq C)\Pr(A \neq B | B \neq C) \\
&\leq \Pr(A \neq C) + \Pr(B \neq C).
\end{aligned}$$

□

Proposition 3.1. *If* $\Pr(F_1 \neq \hat{F}_1) \geq \alpha$ *for all* $\hat{F}_1$ *and* $\mathbb{E}[FF_1] = \rho \geq 0$. *Then for any estimator* $\hat{F}$

$$\Pr(F \neq \hat{F}) \geq \left(\frac{1 - \rho}{2} - \alpha\right)^+. \tag{3.18}$$

Proof. From Lemma 3.2:

$$\begin{aligned}
\Pr(F \neq \hat{F}) &\geq \left(\Pr(F_1 \neq F) - \Pr(F_1 \neq \hat{F})\right)^+ \\
&\geq \left(\frac{1 - \rho}{2} - \alpha\right)^+.
\end{aligned}$$

□

3.6 One-Time Pad Encryption of Functions with Boolean Inputs

We return to the setting where a legitimate transmitter (Alice) wishes to communicate a plaintext message X^n to a legitimate receiver (Bob) through a channel observed by an eavesdropper (Eve). Both Alice and Bob share a secret key K that is not known by Eve. Alice and Bob use a symmetric key encryption scheme determined by the pair of encryption and decryption functions $(\mathsf{Enc}, \mathsf{Dec})$, where $Y^n = \mathsf{Enc}(X^n, K)$ and $X^n = \mathsf{Dec}(Y^n, K)$. Here we assume that both the ciphertext and the plaintext have the same length.

We use the results derived in the previous section to assess the security properties of the one-time pad with non-uniform key distribution when no assumptions are made on the computational resources available to Eve. In this case, perfect secrecy (i.e. $I(X^n;Y^n) = 0$) can only be achieved when $H(K) \geq H(X^n)$ [2], which, in turn, is challenging in practice. Nevertheless, as we shall show in this section, information-theoretic security claims can still be made in the short key regime, i.e. $H(K) < H(X^n)$. We first prove the following ancillary result.

Lemma 3.3. *Let F be a Boolean random variable and $F \to X \to Y \to \hat{F}$, where $|\mathcal{Y}| \geq 2$. Furthermore, $\Pr\{F \neq \hat{F}\} \geq \alpha$ for all $Y \to \hat{F}$. Then $I(F;Y) \leq 1 - 2\alpha$.*

Proof. The result is a direct consequence of the fact that the channel with binary input and finite output alphabet that maximizes mutual information for a fixed error probability is the erasure channel, proved next. Assume, without loss of generality, that $\mathcal{Y} = [m]$ and $p_{F,Y}(-1,y) \geq p_{F,Y}(1,y)$ for $y \in [k]$ and $p_{F,Y}(-1,y) \leq p_{F,Y}(1,y)$ for $y \in \{k+1,\ldots,m\}$, where $k \in [m]$. Now let $\widetilde{Y}$ be a random variable that takes values in $[2m]$ such that

$$p_{F,\widetilde{Y}}(b,y) = \begin{cases} p_{F,Y}(b,y) - p_{F,Y}(1,y) & y \in [k], \\ p_{F,Y}(b,y) - p_{F,Y}(-1,y) & y \in \{k+1,\ldots,m\}, \\ p_{F,Y}(1,y) & y - m \in [k], \\ p_{F,Y}(-1,y) & y - m \in \{k+1,\ldots,m\}. \end{cases}$$

Note that $F \to \widetilde{Y} \to Y$, since $Y = \widetilde{Y} - m\mathbb{1}_{\{\widetilde{Y}>m\}}$ and, consequently, $I(F;\widetilde{Y}) \geq I(F;Y)$. Furthermore, the reader can verify that

$$P_e(F|Y) = P_e\left(F|\widetilde{Y}\right) = \alpha.$$

In particular, given the optimal estimator $\widetilde{Y} \to \hat{F}$, a detection error can only occur when $\widetilde{Y} \in \{k+1,\ldots,m\}$, in which case $\hat{F} = F$ with probability $1/2$.

Finally,

$$\begin{aligned} H(F|\widetilde{Y}) &= -\sum_{\substack{b\in\{-1,1\}\\ y\in[2m]}} p_{\widetilde{Y}}(y)p_{F|\widetilde{Y}}(b|y)\log p_{F|\widetilde{Y}}(b|y) \\ &= \sum_{y\in\{m+1,2m\}} p_{\widetilde{Y}}(y) \\ &\geq 2\alpha. \end{aligned}$$

Consequently, $I(F;\widetilde{Y}) = H(F) - H(F|\widetilde{Y}) \leq 1 - 2\alpha$. The result follows. □

Let X^n be a plaintext message composed by a sequence of n bits drawn from $\{-1,1\}^n$. The plaintext can be perfectly hidden by using a one-time pad: A ciphertext Y^n is produced as $Y^n = X^n \otimes Z^n$, where the key $K = Z^n$ is a uniformly distributed sequence of n i.i.d. bits chosen independently from X^n. The one-time pad is impractical since, as mentioned, it requires Alice and Bob to share a very long key.

Instead of trying to hide the entire plaintext message, assume that Alice and Bob wish to hide only a set of functions of the plaintext from Eve. In particular, we denote this set of functions as $\Phi = \{\phi_1, \ldots, \phi_m\}$ where $\phi_i : \{-1,1\}^n \to \{-1,1\}$, $\mathbb{E}[\phi_i(X^n)] = 0$ and $\mathbb{E}[\phi_i(X^n)\phi_j(X^n)] = 0$. The set of functions Φ is said to be hidden $I(\phi_i(X^n);Y^n) = 0$ for all $\phi_i \in \Phi$. Can this be accomplished with a key that satisfies $H(K) \ll H(X^n)$?

The answer is positive, but it depends on Φ. We denote the Fourier expansion of $\phi_i \in \Phi$ as

$$\phi_i = \sum_{\mathcal{S}\subseteq[n]} \rho_{i,\mathcal{S}}\chi_{\mathcal{S}}.$$

The following result shows that ϕ_i is perfectly hidden from Eve if and only if $I(\chi_{\mathcal{S}}(X^n);Y^n) = 0$ for all $\chi_{\mathcal{S}}$ such that $\rho_{i,\mathcal{S}} > 0$.

Lemma 3.4. *If $I(\phi_i(X^n);Y^n) = 0$ for all $\phi_i \in \Phi$, then $I(\chi_{\mathcal{S}}(X^n);Y^n) = 0$ for all $\mathcal{S}$ such that $\rho_{i,\mathcal{S}} > 0$ for some $i \in [m]$.*

Proof. Assume that $I(\chi_{\mathcal{S}}(X^n);Y^n) > 0$ for a given $\rho_{i,\mathcal{S}} > 0$. Then there exists $b : \mathcal{Y}^n \to \{-1,1\}$ such that $\mathbb{E}[b(Y^n)\chi_{\mathcal{S}}(X^n)] = \lambda > 0$. Consequently, from (3.10), $\mathbb{E}[b(Y^n)\phi_1(X^n)] \geq \lambda\rho_{i,\mathcal{S}} > 0$, and $\phi_1(X^n)$ is not independent of Y^n. □

The previous result shows that hiding a set of functions perfectly, or even a single function, might be as hard as hiding X^n. Indeed, if there is a $\phi_i \in \Phi$ such that $\mathbb{E}[\phi_i(X^n)\mathcal{X}_{\mathcal{S}}(X^n)] > 0$ for all $\mathcal{S} \subseteq [n]$ where $|\mathcal{S}| = 1$, then perfectly hiding this set of functions can only be accomplished by using a one-time pad. Nevertheless, if we step back from perfect secrecy, a large class of functions can be hidden with a comparably small key, as in the next example.

Example 3.4 (BSC revisited)**.** Let Z^n be a sequence of n i.i.d. bits such that $\Pr\{Z_i =$

$-1\} = \epsilon$, and consider once again the one-time pad $Y^n = X^n \otimes Z^n$. Furthermore, denote

$$\Phi_k = \{\phi : \{-1,1\}^n \to \{-1,1\} \mid \mathbb{E}\left[\phi(X^n)\chi_{\mathcal{S}}(X^n)\right] = 0 \ \forall |\mathcal{S}| < k\}.$$

Let $\phi \in \Phi_k$ and $\phi(X^n) = \sum_{\mathcal{S}:|\mathcal{S}|\geq k} \rho_{\mathcal{S}}\chi_{\mathcal{S}}(X^n)$. Then, from Theorem 3.2 and Corollary 3.1, for any $\hat{b} : \mathcal{Y}^n \to \{-1,1\}$,

$$\begin{aligned}\Pr\{\phi(X^n) \neq \hat{b}(Y^n)\} &\geq \frac{1}{2}\left(1 - \sqrt{\sum_{|\mathcal{S}|>T} \rho_{\mathcal{S}}^2 (1-2\epsilon)^{2|\mathcal{S}|}}\right) \\ &\geq \frac{1}{2}\left(1 - (1-2\epsilon)^k\right).\end{aligned}$$

Consequently, from Lemma 3.3, $I(\phi(X^n); Y^n) \leq (1-2\epsilon)^k$ for all $\phi \in \Phi_k$. Note that $H(Z^n) = nh(\epsilon)$, which can be made very small compared to n. Therefore, even with a small key, a large class of functions can be almost perfectly hidden from the eavesdropper through this simple one-time pad scheme. The BSC setting discussed in Example 3.1 is generalized in the following theorem which, in turn, is a particular case of the analysis presented in Chapter 6.

Theorem 3.3 (Generalized One-time Pad). *Let $Y^n = X^n \otimes Z^n$, $X^n \perp\!\!\!\perp Z^n$, X^n be uniformly distributed, $\phi : \{-1,1\}^n \to \{-1,1\}$ and $\phi(X^n) = \sum_{\mathcal{S}\subseteq[n]} \rho_{\mathcal{S}}\chi_{\mathcal{S}}(X^n)$. We define $c_{\mathcal{S}} \triangleq \mathbb{E}\left[\chi_{\mathcal{S}}(Z^n)\right]$ for $\mathcal{S} \subseteq [n]$. Then*

$$I(\phi(X^n); Y^n) \leq \sqrt{\sum_{\mathcal{S}\subseteq[n]} (c_{\mathcal{S}}\rho_{\mathcal{S}})^2}. \tag{3.19}$$

In particular, $I(\phi(X^n); Y^n) = 0$ if and only if $c_{\mathcal{S}} = 0$ for all $\mathcal{S}$ such that $\rho_{\mathcal{S}} \neq 0$.

Proof. Let $\psi : \{-1,1\}^n \to \{-1,1\}$ and $\psi(Y^n) = \sum_{\mathcal{S}\subseteq[n]} d_{\mathcal{S}}\chi_{\mathcal{S}}(Y^n)$. Note that $\sum_{\mathcal{S}\subseteq[n]} d_{\mathcal{S}}^2 = 1$. Then

$$\begin{aligned}\mathbb{E}\left[\phi(X^n)\psi(Y^n)\right] &= \mathbb{E}\left[\phi(X^n)\mathbb{E}\left[\psi(Y^n)|X^n\right]\right] \\ &= \mathbb{E}\left[\phi(X^n)\sum_{\mathcal{S}\subseteq[n]} d_{\mathcal{S}}\mathbb{E}\left[\chi_{\mathcal{S}}(Y^n)|X^n\right]\right] \\ &= \mathbb{E}\left[\phi(X^n)\sum_{\mathcal{S}\subseteq[n]} d_{\mathcal{S}}\mathbb{E}\left[\chi_{\mathcal{S}}(X^n \otimes Z^n)|X^n\right]\right] \\ &= \mathbb{E}\left[\phi(X^n)\sum_{\mathcal{S}\subseteq[n]} d_{\mathcal{S}}\mathbb{E}\left[\chi_{\mathcal{S}}(X^n)\chi_{\mathcal{S}}(Z^n)|X^n\right]\right] \\ &= \sum_{\mathcal{S}\subseteq[n]} d_{\mathcal{S}}\mathbb{E}\left[\phi(X^n)\chi_{\mathcal{S}}(X^n)\right]\mathbb{E}\left[\chi_{\mathcal{S}}(Z^n)\right]\end{aligned}$$

$$= \sum_{\mathcal{S}\subseteq[n]} d_{\mathcal{S}}\rho_{\mathcal{S}}c_{\mathcal{S}} \tag{3.20}$$

$$\leq \sqrt{\sum_{\mathcal{S}\subseteq[n]} (c_{\mathcal{S}}\rho_{\mathcal{S}})^2}, \tag{3.21}$$

where (3.21) follows from the Cauchy-Schwarz inequality. The inequality (3.19) then follows from Lemma 3.3. Finally, assume there exists $\mathcal{S} \subseteq [n]$ such that both $c_{\mathcal{S}} \neq 0$ and $\rho_{\mathcal{S}} \neq 0$. Then setting $\psi(Y^n) = \chi_{\mathcal{S}}(Y^n)$, it follows from (3.20) that $\mathbb{E}[\phi(X^n)\psi(Y^n)] = \rho_{\mathcal{S}}c_{\mathcal{S}} \neq 0$ and, consequently, $I(\phi(X^n); Y^n) > 0$. □

3.7 From Symbol Secrecy to Function Secrecy

Symbol secrecy captures the amount of information that an encryption scheme leaks about individual symbols of a message. A given encryption scheme can achieve a high level of (weak) information-theoretic security, but low symbol secrecy. As illustrated in Section 2.7.1, by sending a constant fraction of the message in the clear, the average amount of information about the plaintext that leaks relative to the length of the message can be made arbitrarily small, nevertheless the symbol secrecy performance is always constant (i.e. does not decrease with message length).

When X is uniformly drawn from $\mathbb{F}_q$ for which an $(n, k, n-k+1)$ MDS code exists, then an absolute symbol secrecy of k/n can always be achieved using the encryption scheme suggested in Proposition 2.1. If X is a binary random variable, then we can map sequences of plaintext bits of length $\lfloor \log_2 q \rfloor$ to an appropriate symbol in $\mathbb{F}_q$, and then use the parity check matrix of an MDS code to achieve a high symbol secrecy. Therefore, we may assume without loss of generality that X^n is drawn from $\{-1, 1\}^n$. We also make the assumption that X^n is uniformly distributed. This can be regarded as an approximation for the distribution of X^n when it is, for example, the output of an optimal source encoder with sufficiently large blocklength.

Theorem 3.4. *Let X^n be a uniformly distributed sequence of n bits, $Y = \mathsf{Enc}_n(X^n, K)$, and u_ϵ and ϵ_t^* the corresponding symbol secrecy and dual symbol secrecy of Enc_n, defined in (2.1) and (2.3), respectively. Furthermore, for $\phi : \{-1,1\}^n \to \{-1,1\}$ and $\mathbb{E}[\phi(X^n)] = 0$, let $\phi(X^n) = \sum_{\mathcal{S}\subseteq[n]} \rho_{\mathcal{S}}\chi_{\mathcal{S}}(X^n)$. Then for any $\hat{\phi} : \mathcal{Y} \to \{-1,1\}$*

$$P_e(\phi(X^n)|Y) \geq \frac{1}{2}\left(1 - B_{|\Phi|}(\boldsymbol{\rho}, \boldsymbol{\lambda})\right), \tag{3.22}$$

where $\Phi = \{\chi_{\mathcal{S}} : \rho_{\mathcal{S}} \neq 0\}$, $\lambda(t) \triangleq h_b^{-1}((1-\epsilon_t^* t)^+)$, $\boldsymbol{\lambda} = \{\lambda(|\mathcal{S}|)\}_{\mathcal{S}\subseteq[n]}$ and $\boldsymbol{\rho} = \{|\rho_{\mathcal{S}}|\}_{\mathcal{S}\subseteq[n]}$. In particular,

$$P_e(\phi(X^n)|Y) \geq \frac{1}{2}\left(1 - \sqrt{\sum_{|\mathcal{S}|>n\mu_0} \rho_{\mathcal{S}}^2}\right). \tag{3.23}$$

Proof. From the definition of symbol secrecy, for any $\mathcal{S} \subseteq [n]$ with $|\mathcal{S}| = t$

$$I(\chi_{\mathcal{S}}(X^n); Y) \leq I(X^{\mathcal{S}}; Y) \leq \epsilon_t^* t,$$

and, consequently,

$$H(\chi_{\mathcal{S}}(X^n)|Y) \geq (1 - \epsilon_t^* t)^+.$$

From Fano's inequality, for any binary $\hat{F}$ where $Y \to \hat{F}$

$$\Pr\{\chi_{\mathcal{S}}(X^n) \neq \hat{F}\} \geq h_b^{-1}((1 - \epsilon_t^* t)^+),$$

where $h_b^{-1} : [0, 1] \to [0, 1/2]$ is the inverse of the binary entropy function. In particular, from the definition of absolute symbol secrecy, if $\epsilon_t^* = 0$, then

$$\Pr\{\chi_{\mathcal{S}}(X^n) \neq \hat{F}\} = 1/2 \ \forall |\mathcal{S}| \leq n\mu_0.$$

The result then follows directly from Theorem 3.2, the fact that $\phi(X^n) = \sum_{\mathcal{S} \subseteq [n]} \rho_{\mathcal{S}} \chi_{\mathcal{S}}(X^n)$ and letting $\lambda(t) \triangleq h_b^{-1}((1 - \epsilon_t^* t)^+)$. □

3.8 The Correlation-Error Product

We momentarily diverge from the cryptographic setting and introduce the correlation-error product for the privacy setting considered by Calmon and Fawaz in [18] and describe in Section 1.4. Let S and X be two random variables with joint distribution $p_{S,X}$. S represents a variable that is supposed to remain private, while X represents a variable that will be released to an untrusted data collector in order to receive some utility based on X. The goal is to design a randomized mapping $p_{Y|X}$, called the privacy assuring mapping, that transforms X into an output Y that will be disclosed to a third party.

The goal of a privacy assuring mechanism is to produce an output Y, derived from X according to the mapping $p_{Y|X}$, that will be released to the data collector in the place of X. The released variable Y is chosen such that S cannot be inferred reliably given an observation of Y. Simultaneously, given an appropriate distortion metric, X should be close enough to Y so that a certain level of utility can still be provided. For example, S could be a user's political preference, and X a set of movie ratings released to a recommender system in order to receive movie recommendations. Y is chosen as a perturbed version of the movie recommendations so that the user's political preference is obscured, while meaningful recommendations can still be provided.

Given $S \to X \to Y$ and $p_{S,X}$, a privacy assuring mapping is given by the conditional distribution $p_{Y|X}$. The choice of $p_{Y|X}$ determines the tradeoff between privacy and utility. If $p_{Y|X} = p_Y$, then perfect privacy is achieved (i.e. S and Y are independent), but no utility

can be provided. Conversely, if $p_{Y|X}$ is the identity mapping, then no privacy is gained, but the highest level of utility can be provided.

When $S = \phi(X)$ where $\phi \in \mathcal{L}_2(p_X)$, the bounds from Section 3.4 shed light on the fundamental privacy-utility tradeoff. Returning to the notation of Section 3.4, let $S = \phi(X)$ be correlated with a set of functions $\Phi = \{\phi_i\}_{i=1}^m$. The next result is a direct corollary of Theorem 3.2.

Corollary 3.2. *Let* $\mathbb{E}[S\phi_i(X)] = \rho_i$, $\sum_{i=1}^{|\Phi|} \rho_i^2 = 1$, $\psi_i(Y) = \mathbb{E}[\phi_i(X)|Y]$ *and, for* $i \neq j$, $\mathbb{E}[\phi_i(X)\phi_j(X)] = 0$ *and* $\mathbb{E}[\psi_i(Y)\psi_j(Y)] = 0$. *Then*

$$\mathsf{mmse}(S|Y) = \sum_{i=1}^{|\Phi|} \mathsf{mmse}(\phi_i(Y)|X)\rho_i^2. \tag{3.24}$$

We call the product $\mathsf{mmse}(\phi_i(Y)|X)\rho_i^2$ the *correlation-error* product. The secret variable S cannot be estimated with low MMSE from Y if and only if the functions ϕ_i that are strongly correlated with S (i.e. large ρ_i^2) cannot be estimated reliably. Consequently, if ρ_i is large and ϕ_i is relevant for the utility provided by the data collector, privacy cannot be achieved without a significant loss of utility: $\mathsf{mmse}(\phi_i(X)|Y)$ is necessarily large if $\mathsf{mmse}(S|Y)$ is large. Conversely, in order to hide S, it is sufficient to hide the functions $\phi_i(X)$ that are strongly correlated with $\phi(X)$. This no-free-lunch result is intuitive, since one would expect that privacy cannot be achieved if utility is based on data that is strongly correlated with the private variables. The results presented here prove that this is indeed the case. The correlation-error product will be studied again using a PIC-based analysis in Section 7.8.

3.9 Prologue to Chapters 4 and 5

3.9.1 Transforming Information Guarantees into Estimation Guarantees

In this chapter, symbol-secrecy was extended to the functional setting by (i) mapping symbol secrecy guarantee's into a bound on the error probability of estimating individual bits of the message, (ii) using the MMSE and Fourier-based analysis to transform the bound on estimating individual bits into a bound on estimating a target function of the plaintext, and (iii) mapping this new bound on estimating a target function into a bound on mutual information through Fano's inequality.

However, the inverse approach is arguably more interesting, where bounds on information measures are mapped to estimation restrictions. Guarantees that an adversary cannot infer certain functions of the plaintext reliably are indeed the focus of most modern cryptographic security metrics [23]. This is perhaps one of the crucial differences between cryptographic and information-theoretic metrics for security: information-theoretic metrics guarantee approximate independence (e.g. small mutual information), whereas cryptographic metrics reduce to estimation guarantees (e.g. negligible advantage over a random guess). Of course, both

approaches are related, but the problem remains of translating information-based security guarantees into restrictions on what the adversary can or cannot learn given an observation of the information leaked by a security system. The MMSE-based analysis presented in this chapter takes the first steps towards this direction.

Transforming information guarantees into estimation guarantees is the focus of the next chapter. We introduce a convex program based on rate-distortion formulations to transform bounds on an information measure into guarantees on the (average) error of estimating properties of the plaintext by an adversary. This raises the question: What is the right security metric that captures both information and estimation guarantees? We propose a metric that lives in the intersection of both worlds, called the Principal Inertia Components, in Chapter 5

3.9.2 MMSE-Based Analysis

As discussed above, security guarantees given in terms of an information metric (e.g. upper-bound on mutual information) should be translated into actual guarantees in terms of how well an adversary can (or cannot) estimate functions of the plaintext. The MMSE-based analysis presented in this chapter enabled us to extend symbol-secrecy to this functional setting, and quantified how well a target function of the plaintext can be estimated given knowledge of the MMSE of estimating a set of reference functions. These results also led to the correlation-error product analysis for the privacy against statistical inference framework.

The central setup behind the MMSE-based analysis assumed that certain functions of the plaintext are known to be hard or easy to infer. A natural next step is to investigate this setup when the functions form a basis for the space of $\mathcal{L}_2$ functions of the plaintext. As will be shown in Chapter 5, this assumption naturally leads to information metrics based on Principal Inertia Components (PICs) of the joint distribution of the plaintext and the disclosed data. The PICs are the spectrum of the conditional expectation operator, and are explained in greater detail in Chapter 5.

Chapter 4

From Information Measures to Estimation Guarantees

In this chapter, we establish lower bounds for the average estimation error of a hidden variable X and a function of the hidden variable $f(X)$ given an observation of Y. These bounds depend only on certain measures of information between X and Y and the marginal distribution of X. The results hold for any estimator, and they shed light on the fundamental limits of what can be inferred about a hidden variable from a noisy measurement. The bounds derived here are similar in nature to Fano's inequality [4], and can be characterized as the solution of a convex program which, in turn, is the optimization problem behind rate-distortion formulations.

4.1 Overview

Information-theoretic security metrics, including the ones introduced in the previous chapters, quantify the information (usually measured in terms of mutual information) that leaks from a security system. For example, symbol secrecy, introduced in Chapter 2, captures how much information an adversary learns about individual symbols of the plaintext message. In addition, we demonstrated how symbol secrecy can be extended to quantify the information that leaks about functions of the plaintext in Chapter 3.

In order to have a full picture of what an adversary can or cannot learn from the information that leaks from a security system, it is necessary to map security guarantees in terms of information metrics into guarantees on how well the adversary can or cannot *estimate* functions of the plaintext message. For example, assume that we have an encryption mechanism in place that guarantees that the mutual information between an individual symbol of the plaintext and a computationally unbounded adversary's observation is at most 0.01 bits. Is it possible to bound how well an adversary can guess that symbol? How does this result depend on the distribution of the plaintext source? Are there other information measures

besides mutual information for deriving such bounds on estimation?

In this chapter, we take steps towards answering these questions and present a general framework for mapping a security guarantee given in terms of an information measure (e.g. upper bound on mutual information) into a guarantee on the adversary's estimation capability (e.g. lower bound on estimation error). The proposed approach is closely related to rate-distortion theory, and is equivalent to computing the distortion-rate function [4, Chapter 10] for a given information measure and source distribution. We note that the results introduced in this chapter are not restricted to mutual information and, in the next chapter, will be applied to an information measure based on the principal inertia components.

The approach presented here is related to the one in Chapter 3, where symbol-secrecy was extended to the functional setting by (i) translating symbol secrecy guarantee's into a bound on the error probability of estimating individual bits of the message, (ii) using the MMSE and Fourier-based analysis to translate the bound on estimating individual bits into a bound on estimating a target function of the plaintext, and (iii) mapping this new bound on estimating a target function into a bound on mutual information through Fano's inequality. Here we take the inverse approach, where bounds on information measures are translated into estimation restrictions.

4.2 Main Contributions

Estimation-theoretic setup for security

Throughout the rest of the chapter, we take a step back from security and consider the underlying estimation-theoretic problem: Given an observation of a random variable Y, what can we learn about a correlated, hidden variable X? For example, in the symmetric-key encryption setup considered in the previous chapters, X can be the plaintext message, and Y the ciphertext and any additional side information available to an adversary. The results presented here assume that the joint distribution of X and Y is known a priori.

If the joint distribution between X and Y is known, the probability of error of estimating X given an observation of Y can be calculated exactly. However, in most practical settings, this joint distribution is unknown. Nevertheless, it may be possible to estimate certain correlation measures of X and Y reliably, such as maximal correlation, χ^2-statistic or mutual information. For example, by using the symmetric-key encryption scheme described in Scheme 2.1, the mutual information between X and Y is provably bounded, even though the distribution $p_{X,Y}$ may be difficult to compute exactly.

Given an upper bound θ on a certain information measure $\mathcal{I}$, denoted by $\mathcal{I}(X;Y) \leq \theta$, is it possible to determine a lower bound for the average error of estimating X from Y over all possible estimators? We answer this question in the affirmative. In particular, the problem of computing such bound for a given distribution p_X and θ is equivalent to computing a distortion-rate function, introduced in Definition 4.2. When the estimation

metric is error probability, we call the corresponding distortion-rate function the *error-rate function*, denoted by $e_{\mathcal{I}}(p_X, \theta)$ and described in Definition 4.3. In the context of security, this bound characterizes the best estimation of the plaintext that a (computationally unbounded) adversary can make given an observation of the output of the system. We note that some of the results in this chapter appeared in [64].

Bounding the estimation error of functions

Owing to the nature of the joint distribution, it may be infeasible to estimate X from Y with small estimation error. However, it is possible that a non-trivial function $f(X)$ exists that is of interest to a learner and can be estimated reliably from Y. If f is the identity function, this reduces to the standard problem of estimating X from Y. Determining if such a function exists is relevant to several applications in learning, privacy, security and information theory. In particular, this setting is related to the information bottleneck method [65] and functional compression [66], where the goal is to compress X into Y such that Y still preserves information about $f(X)$.

For most security applications, minimizing the average error of estimating a hidden variable X from an observation of Y is insufficient. As argued in previous chapters, cryptographic definitions of security, and in particular semantic security [23], require that an adversary has negligible advantage in guessing any function of the input given an observation of the output. In light of this, we present bounds for the best possible average error achievable for estimating functions of X given an observation of Y.

Still assuming that $p_{X,Y}$ is unknown, p_X is given and a bound $\mathcal{I}(X;Y) \leq \theta$ is known (where $\mathcal{I}$ is not restricted to being mutual information), we present in Theorem 4.2 a method for adapting bounds for error probability into bounds for the average estimation error of functions of X given Y. This method depends on a few technical assumptions on the information measure (stated in Definition 4.1 and in Theorem 4.2), foremost of which is the existence of a lower bound for the error-rate function that is Schur-concave[1] in p_X for a fixed θ. Theorem 4.2 then states that, under these assumptions, for any deterministic, surjective function $f : \mathcal{X} \to \{1, \ldots, M\}$, we can obtain a lower bound for the average estimation error of f by computing $e_{\mathcal{I}}(p_U, \theta)$, where U is a random variable that is a function X.

Note that Schur-concavity is crucial for this result. In Theorem 4.1, we show that this condition is always satisfied when $\mathcal{I}(X;Y)$ is concave in p_X for a fixed $p_{Y|X}$, convex in $p_{Y|X}$ for a fixed p_X, and satisfies the Data Processing Inequality. This generalizes a result by Ahlswede [67] on the extremal properties of rate-distortion functions. Consequently, Fano's inequality can be adapted in order to bound the average estimation error of functions, as shown in Corollary 4.1. By observing that a particular form of the bound stated in Theorem 5.4 is Schur-concave, we present in the next chapter a bound for the error probability of

[1] A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *Schur-concave* if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ where $\mathbf{x}$ is majorized by $\mathbf{y}$, then $f(\mathbf{x}) \geq f(\mathbf{y})$.

estimating functions in terms of the maximal correlation, stated in Corollary 5.4.

4.3 A Convex Program for Mapping Information Guarantees to Bounds on Estimation

Throughout the rest of this chapter, we let X and Y be two random variables drawn from finite sets $\mathcal{X}$ and $\mathcal{Y}$. We have the following definition.

Definition 4.1. We say that a function $\mathcal{I}$ that maps any joint probability mass function (pmf) to a non-negative real number is an *information measure* (equivalently *measure of information*) if for any discrete random variables W, X, Y and Z (i) $\mathcal{I}(p_{X,Y})$ is convex in $p_{Y|X}$ for a fixed p_X, (ii) $\mathcal{I}$ satisfies the data processing inequality, i.e. if $X \to Y \to Z$ then $\mathcal{I}(p_{X,Z}) \leq \mathcal{I}(p_{X,Y})$, and (iii) if W is a one-to-one mapping of Y and Z is a one-to-one mapping of X, then $\mathcal{I}(p_{W,Z}) = \mathcal{I}(p_{X,Y})$ (invariance property). We overload the notation of $\mathcal{I}$ and let $\mathcal{I}(p_{X,Y}) = \mathcal{I}(p_X, p_{Y|X})$ in order to make the dependence on the marginal distribution and the channel clear. Furthermore, we also denote $\mathcal{I}(p_{X,Y}) = \mathcal{I}(X;Y)$ when the distribution is clear from the context. Examples of information measure includes maximal correlation, defined in (5.1), and mutual information.

Now assume the standard estimation setup where a hidden variable X should be estimated from an observed random variable Y. We assume that the joint distribution between $p_{X,Y}$ is not known, but the marginal distribution p_X is known, and that a security constraint $\mathcal{I}(p_{X,Y}) \leq \theta$ is given for an information measure $\mathcal{I}$. Since $\mathcal{I}$ satisfies the Data Processing Inequality, for any estimate $\hat{X}$ of X such that $X \to Y \to \hat{X}$ we have $\mathcal{I}(X;\hat{X}) \leq \mathcal{I}(X;Y) \leq \theta$. The problem of translating a security guarantee in terms of $\mathcal{I}$ into a constraint on how well an adversary can estimate (on average) the hidden variable X given an estimation metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be approximated by solving the optimization problem

$$\inf_{p_{\hat{X}|X}} \quad \mathbb{E}\left[d(X,\hat{X})\right] \tag{4.1}$$

$$\text{s.t.} \quad \mathcal{I}(X;\hat{X}) \leq \theta. \tag{4.2}$$

This motivates the following definition

Definition 4.2. We denote the smallest (average) estimation error $D_{\mathcal{I},d}$ for a given information measure $\mathcal{I}$ and estimation cost function d as

$$D_{\mathcal{I},d}(p_X,\theta) \triangleq \inf_{p_{\hat{X}|X}} \left\{ \mathbb{E}\left[d(X,\hat{X})\right] \mid \mathcal{I}(p_X, p_{\hat{X}|X}) \leq \theta \right\}, \tag{4.3}$$

where the infimum is over all conditional distributions $p_{\hat{X}|X}$.

Observe that if $p_{Y|X}$ were known

$$D_{\mathcal{I},d}(p_X,\theta) \leq \inf_{p_{\hat{X}|Y}} \left\{ \mathbb{E}\left[d(X,\hat{X})\right] \mid \mathcal{I}(p_X,p_{Y|X}) \leq \theta, X \to Y \to \hat{X} \right\},$$

since, by the assumption that $\mathcal{I}$ satisfies the DPI, $\mathcal{I}(X;\hat{X}) \leq \mathcal{I}(X;Y) \leq \theta$. When $\mathcal{I}(X;Y) = I(X;Y)$, $D_{I,d}(p_X,\theta)$ is the distortion-rate function [4, pg. 306]. When the distortion function d is the Hamming distortion, $D_{\mathcal{I},d}(p_X,\theta)$ gives the smallest probability of error for estimating X given an observation Y that satisfied $\mathcal{I}(X;Y) \leq \theta$. This case will be of particular interest in this chapter, motivating the next definition.

Definition 4.3. Denoting the Hamming distortion metric as

$$d_H(x,y) \triangleq \begin{cases} 0, & x = y, \\ 1, & \text{otherwise}, \end{cases}$$

we define the *error-rate function* for the information measure $\mathcal{I}$ as

$$e_{\mathcal{I}}(p_X,\theta) \triangleq D_{\mathcal{I},d_H}(p_X,\theta).$$

The definition of error-rate function directly leads to the following simple lemma.

Lemma 4.1. *For a given information measure $\mathcal{I}$ and any fixed $p_{X,Y}$ such that $\mathcal{I}(p_{X,Y}) \leq \theta$*

$$P_e(X|Y) \geq e_{\mathcal{I}}(p_X,\theta).$$

Proof. Observe that $P_e(X|Y) = \min_{X \to Y \to \hat{X}} \mathbb{E}\left[d_H(X,\hat{X})\right]$, where the minimum is over all distributions $p_{\hat{X}|X}$ that satisfy the Markov constraint. Since $\mathcal{I}$ satisfies the DPI, then $\mathcal{I}(X;\hat{X}) \leq I(X;Y) \leq \theta$, and the result follows from Definition 4.2. □

The previous lemma shows that the characterization of $e_{\mathcal{I}}(p_X,\theta)$ for different measures of information $\mathcal{I}$ is particularly relevant for applications in privacy and security, where X is a variable that should remain hidden (e.g. plaintext) and Y is an adversary's observation(e.g. ciphertext). Knowing e_I allows us to translate an upper bound $\mathcal{I}(X;Y) \leq \theta$ into an estimation guarantee: regardless of an adversary's computational resources, given only access to Y he will not be able to estimate X with an average error probability $P_e(X|Y)$ smaller than $e_{\mathcal{I}}(p_X,\theta)$. Therefore, by simply estimating θ and calculating $e_{\mathcal{I}}(p_X,\theta)$ we are able to evaluate the security threat incurred by an adversary that has access to Y.

Example 4.1 (Error-rate function for mutual information.)**.** Using the expression for the rate-distortion function under Hamming distortion for mutual information ([68, (9.5.8)]), for $\mathcal{I}(X;Y) = I(X;Y)$ and $\mathcal{X} = [m]$, the error-rate function is given by $e_I(p_X,\theta) = d^*$, where d^* is the solution of

$$h_b(d^*) + d^* \log(m-1) = H(X) - \theta, \tag{4.4}$$

and $h_b(x) \triangleq -x\log x - (1-x)\log(1-x)$. Denoting $X \to Y \to \hat{X}$ and $p_e \triangleq P_e(X|Y)$, note that (4.4) implies Fano's inequality [4, 2.140]:

$$h_b(p_e) + p_e \log(m-1) \geq H(X) - I(X;Y) = H(X|Y). \tag{4.5}$$

We present next results on the extremal properties of the error-rate function. This analysis will be particularly useful for determining how to bound the probability of error of estimating functions of a random variable.

4.3.1 Extremal Properties of the Error-Rate Function

Due to convexity of $\mathcal{I}(p_X, p_{\hat{X}|X})$ in $p_{\hat{X}|X}$, it follows directly that $e_{\mathcal{I}}(p_X, \theta)$ is convex in θ for a fixed p_X. We will now prove that, for a fixed θ, $e_{\mathcal{I}}(p_X, \theta)$ is *Schur-concave* in p_X if $\mathcal{I}(p_X, p_{\hat{X}|X})$ is concave in p_X for a fixed $p_{\hat{X}|X}$. Ahlswede [67, Theorem 2] proved this result for the particular case where $\mathcal{I}(X;Y) = I(X;Y)$ by investigating the properties of the explicit characterization of the rate-distortion function under Hamming distortion. The proof presented here is considerably simpler and more general, and is based on a proof technique used by Ahlswede in [67, Theorem 1].

Theorem 4.1. *If $\mathcal{I}(p_X, p_{\hat{X}|X})$ is concave in p_X for a fixed $p_{\hat{X}|X}$, then $e_{\mathcal{I}}(p_X, \theta)$ is Schur-concave in p_X for a fixed θ.*

Proof. Consider two probability distributions p_X and q_X defined over $\mathcal{X} = \{1, \ldots, m\}$, and assume that p_X majorizes q_X, i.e. $\sum_{i=1}^{k} q_X(i) \leq \sum_{i=1}^{k} p_X(i)$ for $1 \leq k \leq m$. Therefore q_X is a convex combination of permutations of p_X [69], and can be written as $q_X = \sum_{i=1}^{l} a_i (p_X \circ \pi_i)$ for some $l \geq 1$, where $a_i \geq 0$, $\sum a_i = 1$ and π_i is a permutation of p_X, i.e. $p_X \circ \pi_i = p_{\pi_i(X)}$. Hence, for a fixed $p_{\hat{X}|X}$:

$$\begin{aligned}
\mathcal{I}(q_X, p_{\hat{X}|X}) &= \mathcal{I}\left(\sum_{i=1}^{l} a_i (p_X \circ \pi_i), p_{\hat{X}|X}\right) \\
&\geq \sum_{i=1}^{l} a_i \mathcal{I}(p_X \circ \pi_i, p_{\hat{X}|X}), \\
&= \sum_{i=1}^{l} a_i \mathcal{I}(p_X, \pi_i \circ p_{\hat{X}|X}),
\end{aligned} \tag{4.6}$$

where the inequality follows from the concavity assumption and the last equality from $\mathcal{I}(X, \hat{X})$ being invariant to one-to-one mappings of X and $\hat{X}$. Consequently, from Definition 4.3,

$$e_{\mathcal{I}}(q_X, \theta) = \inf_{p_{\hat{X}|X}} \left\{ \sum_{x,x' \in [m]} d_H(x,x') q_X(x) p_{\hat{X}|X}(x'|x) \middle| \mathcal{I}(q_X, p_{\hat{X}|X}) \leq \theta \right\}$$

$$
\begin{aligned}
&\overset{(a)}{\geq} \inf_{p_{\hat{X}|X}} \left\{ \sum_{i\in[l]} a_i \sum_{x,x'\in[m]} d_H(\pi_i(x),x') p_X(x) p_{\hat{X}|X}(x'|\pi_i(x)) \middle| \sum_{i\in[l]} a_i \mathcal{I}(p_X, \pi_i \circ p_{\hat{X}|X}) \leq \theta \right\} \\
&\overset{(b)}{=} \inf_{p_{\hat{X}|X}} \left\{ \sum_{i\in[l]} a_i \sum_{x,x'\in[m]} d_H(\pi_i(x),\pi_i(x')) p_X(x) p_{\hat{X}|X}(\pi_i(x')|\pi_i(x)) \right. \\
&\qquad\qquad \left. \middle| \sum_{i\in[l]} a_i \mathcal{I}(p_X, \pi_i \circ p_{\hat{X}|X} \circ \pi_i) \leq \theta \right\} \\
&\overset{(c)}{\geq} \inf_{p^1_{\hat{X}|X},\ldots,p^l_{\hat{X}|X}} \left\{ \sum_{i\in[l]} a_i \sum_{x,x'\in[m]} d_H(x,x') p_X(x) p^i_{\hat{X}|X}(x|x') \middle| \sum_{i\in[l]} a_i \mathcal{I}(p_X, p^i_{\hat{X}|X}) \leq \theta \right\} \\
&\overset{(d)}{=} \inf_{\theta_1,\ldots,\theta_l\geq 0} \left\{ \sum_{i=1}^{l} a_i e_{\mathcal{I}}(p_X,\theta_i) \middle| \sum_{i=1}^{l} a_i\theta_i = \theta \right\} \\
&\overset{(e)}{\geq} \inf_{\theta_1,\ldots,\theta_l\geq 0} \left\{ e_{\mathcal{I}}\left(p_X, \sum a_i\theta_i\right) \middle| \sum_{i=1}^{l} a_i\theta_i = \theta \right\} \\
&= e_{\mathcal{I}}(p_X,\theta),
\end{aligned}
$$

where inequality (a) follows from (4.6), (b) follows from the fact that the infimum is taken over all mapping $p_{\hat{X}|X}$ and that $\mathcal{I}(X;\hat{X})$ is invariant to one-to-one mappings of X and $\hat{X}$, (c) follows by allowing an arbitrary mapping $p^i_{\hat{X}|X}$ to be chosen for each i, (d) is obtained by noting that the optimal coice of $p^i_{\hat{X}|X}$ is the one that minimizes the Hamming distortion d_H for a given upperbound on $\mathcal{I}(p_X, p^i_{\hat{X}|X})$, and(e) follows from the convexity of $e_I(p_X,\theta)$ in θ. Since this holds for any q_X that is majorized by p_X, $e_{\mathcal{I}}(p_X,\theta)$ is Schur-concave. □

4.4 Bounding the Estimation Error of Functions of a Hidden Random Variable

For any function $f : \mathcal{X} \to \mathcal{U}$, we denote by $\hat{f}$ the maximum a posteriori (MAP) estimator of $f(X)$ given an observation of Y. For a given integer $1 \leq M \leq |\mathcal{X}|$, we define

$$
\mathcal{F}_M \triangleq \{f : \mathcal{X} \to \mathcal{U} \mid f \text{ is surjective and } |\mathcal{U}| = M\} \tag{4.7}
$$

and

$$
P_{e,M}(X|Y) \triangleq \min_{f\in\mathcal{F}_M} P_e(f(X)|Y). \tag{4.8}
$$

$P_{e,|\mathcal{X}|}(X|Y)$ is simply the error probability of estimating X from Y, i.e. $P_{e,|\mathcal{X}|}(X|Y) = P_e(X|Y)$.

The next theorem shows that a lower bound for $P_{e,M}$ can be derived for any information measures $\mathcal{I}$ as long as $e_{\mathcal{I}}(p_X,\theta)$ or a lower bound for $e_{\mathcal{I}}(p_X,\theta)$ is Schur-concave in p_X.

Theorem 4.2. *For a given M, $1 \leq M \leq m$, and p_X, let $U = g_M(X)$, where $g_M : \{1,\ldots,m\} \to \{1,\ldots,M\}$ is defined as*

$$g_M(x) \triangleq \begin{cases} 1 & 1 \leq x \leq m-M+1 \\ x-m+M & m-M+2 \leq x \leq m \end{cases}.$$

Let p_U be the marginal distribution[2] of U. Assume that, for a given information measure $\mathcal{I}$, there exists a function $L_{\mathcal{I}}(\cdot,\cdot)$ such that for all distributions q_X and any θ $e_{\mathcal{I}}(q_X,\theta) \geq L_{\mathcal{I}}(q_X,\theta)$. If $L_{\mathcal{I}}(p_X,\theta)$ is Schur-concave in p_X, then for $X \sim p_X$ and $\mathcal{I}(X;Y) \leq \theta$,

$$P_{e,M}(X|Y) \geq L_{\mathcal{I}}(p_U,\theta). \tag{4.9}$$

In addition[3], for any $S \to X \to Y$ such that p_U majorizes p_S,

$$P_e(S|Y) \geq L_{\mathcal{I}}(p_U,\theta). \tag{4.10}$$

Proof. The result follows from the following chain of inequalities:

$$\begin{aligned} P_{e,M}(X|Y) &\overset{(a)}{\geq} \min_{f\in\mathcal{F}_M,\tilde{\theta}} \left\{ e_{\mathcal{I}}\left(p_{f(X)},\tilde{\theta}\right) : \tilde{\theta} \leq \theta \right\} \\ &\geq \min_{f\in\mathcal{F}_M} \left\{ e_{\mathcal{I}}\left(p_{f(X)},\theta\right) \right\} \\ &\overset{(b)}{\geq} \min_{f\in\mathcal{F}_M} \left\{ L_{\mathcal{I}}\left(p_{f(X)},\theta\right) \right\} \\ &\overset{(c)}{\geq} L_{\mathcal{I}}(p_U,\theta), \end{aligned}$$

where (a) follows from the Data Processing Inequality, (b) follows from $e_{\mathcal{I}}(q_X,\theta) \geq L_{\mathcal{I}}(q_X,\theta)$ for all q_X, and θ and (c) follows from the Schur-concavity of the lower bound and by observing that p_U majorizes $p_{f(X)}$ for every $f \in \mathcal{F}_M$. In the case of $P_e(S|X)$, the same inequalities hold with S playing the role of $f(X)$ in (a) and (b), and the last inequality also following from Schur-concavity of $L_{\mathcal{I}}(p_S,\theta)$ in p_S. □

The following two corollaries illustrate how Theorem 4.2 can be used for different measures of information, namely mutual information and maximal correlation.

Corollary 4.1. *Let $I(X;Y) \leq \theta$. Then*

$$P_{e,M}(X|Y) \geq d^*$$

[2]The pmf of U is $p_U(1) = \sum_{i=1}^{m-M+1} p_X(i)$ and $p_U(k) = p_X(m-M+k)$ for $k = 2,\ldots,M$.

[3]We thank Dr. Nadia Fawaz (nadia.fawaz@technicolor.com) for pointing out this extension.

where d^* is the solution of

$$h_b(d^*) + d^* \log(m-1) = \min\{H(U) - \theta, 0\},$$

and $h_b(\cdot)$ is the binary entropy function.

Proof. $R_I(p_X, \delta)$ is the well known rate-distortion function under Hamming distortion, which satisfies ([68, (9.5.8)]) $R_I(p_X, \delta) \geq H(X) - h_b(d^*) - d^* \log(m-1)$. The result follows from Theorem 4.1, since mutual information is concave in p_X. □

4.4.1 A Conjecture on the Schur-Concavity of Error-Rate Functions

The authors have not yet managed to prove or disprove the following conjecture.

Conjecture 4.1. *For any information measure $\mathcal{I}$, $e_{\mathcal{I}}(p_X, \theta)$ is Schur-concave in p_X for a fixed θ.*

If conjecture 4.1 is true, then the bounding procedure presented in theorem 4.2 can be applied to a very broad set of measures of information, including k-correlation, by simply lower bounding $e_{\mathcal{I}}(p_u, \theta)$. Indeed, even if a closed form solution for the error-rate function is not known, $P_{e,M}$ can be lower bounded by numerically solving (4.3) in terms of p_U and θ using widely available convex solvers.

4.5 Final Remarks

In this chapter, we characterized properties of the error-rate function e_I. Assuming that X and Y are discrete random variables with support $\mathcal{X} = [m]$ and $\mathcal{Y} = [n]$, the joint pmf $p_{X,Y}$ can be displayed as the entries of a matrix $\mathbf{P} \in \mathbb{R}^{m \times n}$, where $[\mathbf{P}]_{i,j} = p_{X,Y}(i,j)$. The problem of determining the estimator $\hat{X}$ of X given an observation of Y then reduces to finding a row-stochastic matrix $\mathbf{P}_{\hat{X}|Y} \in \mathbb{R}^{n \times m}$ that is the solution of

$$P_e(X|Y) = \min_{\mathbf{P}_{\hat{X}|Y}} 1 - \operatorname{tr}\left(\mathbf{P} \times \mathbf{P}_{\hat{X}|Y}\right). \tag{4.11}$$

Note that the previous minimization is a linear program, and by taking its dual the reader can verify that the optimal $\mathbf{P}_{\hat{X}|Y}$ is the maximum-a posteriori (MAP) estimator, as expected.

As discussed in the introduction of this chapter, the joint distribution matrix $\mathbf{P}$ may not be known exactly – only a given information measure $\mathcal{I}(p_{X,Y})$ may be known. Equation (4.11) indicates that possible information measures that lead to sharp lower bounds for error probability may be those somehow related to the spectrum of $\mathbf{P}$. Indeed, the trace of the product of two matrices is closely related to their spectra (cf. Von Neumman's trace inequality). This motivates the following question: Are there information measures that capture the spectrum of a joint distribution matrix $\mathbf{P}$? In the next chapter, we answer this

question in the positive by introducing information measures and lower bounds on estimation error based the Principal Inertia Components (PICs), which, in turn, are connected to the spectrum of $\mathbf{P}$. As discussed in Section 3.9, the PICs are also related to the MMSE-based analysis presented in the previous chapter. We study the PICs in detail next.

Chapter 5

Principal Inertia Components

We introduce in this section the Principal Inertia Components (PICs) of the joint distribution of two random variables X and Y. The PICs provide a fine-grained decomposition of the statistical dependence between X and Y, and lead to information measures that lie in the intersection of information and estimation theory. The PICs possess several desirable information-theoretic properties (e.g. satisfy the Data Processing Inequality, convexity, tensorization, etc.), and describe which functions of X can or cannot be reliably inferred given an observation of Y. As we demonstrate in this chapter and in Chapters 6 and 7, the PICs are powerful tools for evaluating and designing security and privacy systems.

5.1 Overview

Let X and Y be two discrete random variables with finite support $\mathcal{X}$ and $\mathcal{Y}$, respectively. Assume that the joint distribution $p_{X,Y}$ is unknown, but that the marginal distribution p_X is given. In Chapter 4, we introduced a framework for deriving bounds for the average error of estimating X from Y given that a certain information measure $\mathcal{I}(X;Y)$ between X and Y is bounded above by θ, i.e. $\mathcal{I}(X;Y) \leq \theta$. In practice, the value of θ and p_X could be determined, for example, from multiple i.i.d. samples drawn according to $p_{X,Y}$. The number of samples available might be insufficient to characterize $p_{X,Y}$, but enough to estimate θ and p_X reliably. Under these assumptions, what can be said about the smallest probability of error of estimating X or a function of X given an observation of Y?

If $\mathcal{I}(X;Y) = I(X;Y)$, where $I(X;Y)$ is the mutual information between X and Y, then Fano's inequality (4.4) provides a lower bound for the probability of error $P_e(X|Y)$ of guessing X given Y. However, in practice, several other statistics are used in addition to mutual information in order to capture the information (correlation) between X and Y. In this chapter, we focus on one particular metric, the principal inertia components of $p_{X,Y}$, denoted by the vector $(\lambda_1(X;Y), \ldots, \lambda_d(X;Y))$, where $d = \min\{m-1, n-1\}$, and $\lambda_1(X;Y) \geq \lambda_2(X;Y) \geq \ldots \geq \lambda_d(X;Y)$. The exact definition of the PICs is presented in Section 5.4

The PICs are information theoretic and statistical metrics that are particularly well suited for deriving bounds for the average estimation error. As will be shown in this chapter, they naturally appear as the solution of different but interconnected problems in estimation, maximization of correlation, and analysis of the conditional expectation operator. Furthermore, the largest PIC, which is equivalent to the maximal correlation coefficient of two random variables, is a meaningful security metric.

A Geometric Interpretation of the PICs

The PICs also posses an intuitive geometric interpretation, described next. Let X and Y be related through a conditional distribution (channel), denoted by $p_{Y|X}$. For each $x \in \mathcal{X}$, $p_{Y|X}(\cdot|x)$ will be a vector on the $|\mathcal{Y}|$-dimensional simplex, and the position of these vectors on the simplex will determine the nature of the relationship between X and Y. If $p_{Y|X}$ is fixed, what can be learned about X given an observation of Y, or the degree of accuracy of what can be inferred about X *a posteriori*, will then depend on the marginal distribution p_X. The value $p_X(x)$, in turn, ponderates the corresponding vector $p_{Y|X}(\cdot|x)$ akin to a mass. As a simple example, if $|\mathcal{X}|= |\mathcal{Y}|$ and the vectors $p_{Y|X}(\cdot|x)$ are located on distinct corners of the simplex, then X can be perfectly learned from Y. As another example, assume that the vectors $p_{Y|X}(\cdot|x)$ can be grouped into two clusters located near opposite corners of the simplex. If the sum of the masses induced by p_X for each cluster is approximately $1/2$, then one may expect to reliably infer on the order of 1 unbiased bit of X from an observation of Y.

The above discussion naturally leads to considering the use of techniques borrowed from classical mechanics. For a given inertial frame of reference, the mechanical properties of a collection of distributed point masses can be characterized by the moments of inertia of the system. The moments of inertia measure how the weight of the point masses is distributed around the center of mass. An analogous metric exists for the distribution of the vectors $p_{Y|X}$ and masses p_X in the simplex, and it is the subject of study of a branch of applied statistics called *correspondence analysis* ([16, 70]). In correspondence analysis, the joint distribution $p_{X,Y}$ is decomposed in terms of the PICs, which, in some sense, are analogous to the moments of inertia of a collection of point masses. In mathematical probability, the study of principal inertia components dates back to Hirschfeld [71], Gebelein [72], Sarmanov [73] and Rényi [74], and similar analysis have also recurrently appeared in the information theory and applied probability literature. We present a short review of the relevant literature in Section 5.3.

5.2 Main Contributions

This chapter has four main contributions, listed below. Some of the results presented in the chapter appeared in [64].

1. We present an overview of the PICs and their different interpretations, summarized in Theorem 5.1.

2. We analyze properties of a measure of information (correlation) between X and Y based on the PICs of the joint distribution of X and Y. The estimation of the PICs is widely studied in the field of correspondence analysis, and is used in practice to analyze categorical data. The metric we propose, called *k-correlation*, is defined as the sum of the k largest PICs, which, in turn, are the singular values of a particular decomposition of the joint distribution matrix of X and Y. We show that k-correlation generalizes both the maximal correlation and the χ^2 measures of correlation. We also prove that k-correlation satisfies two key properties for information measures: (i) the Data Processing Inequality and (ii) convexity in the conditional probabilities $p_{Y|X}$.

3. We derive a family of lower bounds for the error probability of estimating X given Y based on the PICs of $p_{X,Y}$ and the marginal distribution of X in Theorems 5.4 and 5.6. By applying the techniques derived in Chapter 4, we then extend these bounds for the probability of correctly estimating a function of the hidden variable X given an observation of Y.

4. We characterize the PICs for a wide range of distributions, namely when X is a sequence of n i.i.d. random variables and Y is a symmetric function of X, presented in Theorem 5.7. This result is also extended to exchangeable random variables in Theorem 5.9.

The PICs generalize other measures that are used in information theory. In particular, $\lambda_1 = \rho_m(X;Y)^2$, where $\rho_m(X;Y)$ is the *maximal correlation* between X and Y, defined as [74]

$$\rho_m(X;Y) \triangleq \max_{\substack{\mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0 \\ \mathbb{E}[f(X)^2]=\mathbb{E}[g(Y)^2]=1}} \mathbb{E}\left[f(X)g(Y)\right]. \tag{5.1}$$

In Section 5.4, we discuss how to compute the PICs and provide alternative characterizations. Compared to mutual information, the PICs provide a finer-grained decomposition of the correlation between X and Y.

We propose a metric of information called k-correlation, defined as $\mathcal{J}_k(X;Y) \triangleq \sum_{i=1}^{k} \lambda_i(X;Y)$. This metric satisfies two key properties:

- Convexity in $p_{Y|X}$ (Theorem 5.2);
- Data Processing Inequality (Theorem 5.3). This is also satisfied by $\lambda_1(X;Y), \ldots, \lambda_d(X;Y)$ individually,

and, consequently, is an information measure as per Def. 4.1.

By making use of the fact that the principal inertia components satisfy the Data Processing Inequality, we are able to derive a family of bounds for $P_e(X|Y)$ in terms of p_X and $\lambda_1(X:Y),\ldots,\lambda_d(X;Y)$, described in Theorem 5.4. This result sheds light on the relationship of $P_e(X|Y)$ with the principal inertia components.

One immediate consequence of Theorem 5.4 is a useful scaling law for $P_e(X|Y)$ in terms of the largest principal inertia (i.e. maximal correlation). Let $X = 1$ be the most likely outcome for X. Corollary 5.3 proves that the advantage an adversary has of guessing X, over the trivial solution of simply guessing the most likely outcome of X (say $X = 1$), satisfies

$$\mathsf{Adv}(X|Y) \triangleq |1 - p_X(1) - P_e(X|Y)| \leq O\left(\sqrt{\lambda_1(X;Y)}\right). \tag{5.2}$$

5.2.1 Organization of the Chapter

The rest of this chapter is organized as follows. Section 5.3 presents an overview of related work. Section 5.4 presents the definition an alternative characterizations of the PICs. Section 5.5 introduces the k-correlation metric of information, and proves that it is convex in the transition probability $p_{Y|X}$ and satisfies the Data Processing Inequality. Section 5.6 presents a Fano-like inequality based on the PICs and the marginal distribution p_X, and in Section 5.7 we present a convex program for calculating the error-rate function for k-correlation. Finally, Section 5.8 characterizes the PICs between a symmetric function of a set of samples and a subset of these samples.

5.3 Related Work

The joint distribution matrix $\mathbf{P}$ can be viewed as a contingency table and decomposed using standard techniques from correspondence analysis [16,75]. For an overview of correspondence analysis, we refer the reader to [70]. The term "principal inertia components", used here, is borrowed from the correspondence analysis literature [16]. However, the study of the principal inertia components of the joint distribution of two random variables or, equivalently, the spectrum of the conditional expectation operator, predates correspondence analysis, and goes back to the work of Hirshfield [71], Gebelein [72], Sarmanov [73] and Rényi [74], having appeared in the work of Witsenhausen [76], Ahlswede and Gács [77] and, more recently, Anantharam *et al.* [78], Polyanskiy [79], Raginsky [80] and Calmon *et al.* [64], among others.

The largest principal inertia of $\mathbf{P}$ is equal to $\rho_m(X;Y)^2$, where $\rho_m(X;Y)$ is the *maximal correlation* between X and Y. Maximal correlation has been widely studied in the information theory and statistics literature (e.g [73, 74]). Ahslwede and Gács studied maximal correlation in the context of contraction coefficients in strong data processing inequalities [77], and more recently Anantharam *et al.* present in [78] an overview of different characterizations of maximal correlation, as well as its application in information theory. Estimating the maximal correlation is also the goal of the ACE algorithm introduced by Breiman and

Friedman [17], and further analyzed by Buja [81].

The Data Processing Inequality for the principal inertias was shown by Kang and Ulukus in [82, Theorem 2] in a different setting than the one considered here. Kang and Ulukus make use of the decomposition of the joint distribution matrix to derive outer bounds for the rate-distortion region achievable in certain distributed source and channel coding problems.

Lower bounds on the average estimation error can be found using Fano-type inequalities. Recently, Guntuboyina *et al.* ([83, 84]) presented a family of sharp bounds for the minmax risk in estimation problems involving general f-divergences. These bounds generalize Fano's inequality and, under certain assumptions, can be extended in order to lower bound $P_e(X|Y)$.

Most information-theoretic approaches for estimating or communicating functions of a random variable are concerned with properties of specific functions given i.i.d. samples of the hidden variable X, such as in the functional compression literature [66,85]. These results are rate-based and asymptotic, and do not immediately extend to the case where the function $f(X)$ can be an arbitrary member of a class of functions, and only a single observation is available.

More recently, Kumar and Courtade [24] investigated Boolean functions in an information-theoretic context. In particular, they analyzed which is the most informative (in terms of mutual information) 1-bit function (i.e. $M = 2$) for the case where X is composed by n i.i.d. Bernoulli(1/2) random variables, and Y is the result of passing X through a discrete memoryless binary symmetric channel. Even in this simple case, determining the most informative function is non-trivial. We study this problem in Chapter 6.

5.4 Definition and Characterizations of the PICs

We start with the definition of principle inertia components for discrete random variables. The definition and results presented in the section can be extended to general probability measures.

Definition 5.1. Let $\mathbf{P} \in \mathbb{R}^{m\times n}$ be a matrix with entries $[\mathbf{P}]_{i,j} = p_{X,Y}(i,j)$, and $\mathbf{D}_X \in \mathbb{R}^{m\times m}$ and $\mathbf{D}_Y \in \mathbb{R}^{n\times n}$ be diagonal matrices with diagonal entries $[\mathbf{D}_X]_{i,i} = p_X(i)$ and $[\mathbf{D}_Y]_{j,j} = p_Y(j)$, respectively, where $i \in [m]$ and $j \in [n]$. We define

$$\mathbf{Q} \triangleq \mathbf{D}_X^{-1/2}\mathbf{P}\mathbf{D}_Y^{-1/2}. \tag{5.3}$$

Denoting the singular value decomposition [86] of $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, then

$$\mathbf{\Sigma}^2 = \operatorname{diag}\left(1, \lambda_1(X;Y), \ldots, \lambda_d(X;Y)\right), \tag{5.4}$$

where $d \triangleq \min(m,n) - 1$. The values $\lambda_1(X;Y), \ldots, \lambda_d(X;Y)$ are called the *principal inertia components* (PICs) of $p_{X,Y}$. We consider, without loss of generality, $\lambda_1(X;Y) \geq \lambda_2(X;Y) \geq$

$\ldots \geq \lambda_d(X;Y)$. The columns of the matrices $\mathbf{D}_X^{-1/2}\mathbf{U}$ and $\mathbf{D}_Y^{-1/2}\mathbf{V}$ are called the *principal functions* of X and Y.

The fact that $\sigma_1(\mathbf{Q}) = 1$ can be explained as follows. Let $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^n$ such that $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$, then we can define two functions $f : [m] \to \mathbb{R}$ and $g : [n] \to \mathbb{R}$ where $f(i) = \mathbf{u}_i/\sqrt{p_X(i)}$ and $g(j) = \mathbf{v}_j/\sqrt{p_Y(j)}$. Then

$$\|f(X)\|_2^2 = \sum_{i\in[m]} \left(\frac{\mathbf{u}_i}{\sqrt{p_X(i)}}\right)^2 p_X(i) = 1$$

since $\|\mathbf{u}\|_2 = 1$, and, equivalently $\|g(Y)\|_2 = 1$. Consequently

$$\mathbf{u}^T\mathbf{Q}\mathbf{v} = \mathbb{E}\left[f(X)g(Y)\right] \leq \|f(X)\|_2\|g(Y)\|_2 = 1,$$

where equality is achieved when $\mathbf{u}_i = \sqrt{p_X(i)}$ and $\mathbf{v}_j = \sqrt{p_Y(j)}$.

The next theorem states for equivalent characterizations of the PICs.

Theorem 5.1. *The following characterizations of the PICs $\lambda_1(X;Y), \ldots, \lambda_d(X;Y)$ are equivalent:*

(1) $\sqrt{\lambda_k(X;Y)}$ is the $(k+1)$-st largest singular value of $\mathbf{Q}$.

(2) Let $f_0 : \mathcal{X} \to \mathbb{R}$ and $g_0 : \mathcal{X} \to \mathbb{R}$ be the constant functions $f_0(x) = 1$ and $g_0(y) = 1$ for $x \in [m]$ and $y \in [n]$. Then for $k \in [d]$ and

$$\lambda_k(X;Y) = \max\Big\{\mathbb{E}\left[f(X)g(Y)\right]^2 \Big| f \in \mathcal{L}_2(p_X), g \in \mathcal{L}_2(p_Y), \mathbb{E}\left[f(X)f_j(X)\right] = 0, \mathbb{E}\left[g(Y)g_j(Y)\right] = 0, j \in \{0,\ldots,k-1\}\Big\}, \tag{5.5}$$

where

$$(f_k, g_k) \triangleq \operatorname{argmax}\Big\{\mathbb{E}\left[f(X)g(Y)\right]^2 \Big| f \in \mathcal{L}_2(p_X), g \in \mathcal{L}_2(p_Y), \mathbb{E}\left[f(X)f_j(X)\right] = 0, \mathbb{E}\left[g(Y)g_j(Y)\right] = 0, j \in \{0,\ldots,k-1\}\Big\}. \tag{5.6}$$

Furthermore, $g_k(Y) = \frac{\mathbb{E}[f_k(X)|Y]}{\|\mathbb{E}[f_k(X)|Y]\|_2}$, and $\lambda_k(X;Y) = \|\mathbb{E}\left[f_k(X)|Y\right]\|_2^2$.

(3) Consider the conditional expectation operator $T : \mathcal{L}_2(p_X) \to \mathcal{L}_2(p_Y)$, given by

$$Tf(y) = \mathbb{E}\left[f(X)|Y = y\right]. \tag{5.7}$$

Then $(1, \sqrt{\lambda_1(X;Y)}, \ldots, \sqrt{\lambda_d(X;Y)})$ is the spectrum of T.

(4) For $k \in [d]$:

$$1 - \lambda_k(X;Y) = \min\left\{\mathsf{mmse}(f(X)|Y)\middle| h \in \mathcal{L}_2(p_X), \mathbb{E}\left[f(X)f_j(X)\right] = 0, j \in \{0,\ldots,k-1\}\right\}, \tag{5.8}$$

where

$$h_k \triangleq \operatorname{argmin}\left\{\mathsf{mmse}(h(X)|Y)\middle| f \in \mathcal{L}_2(p_X), \mathbb{E}\left[f(X)f_j(X)\right] = 0, j \in \{0,\ldots,k-1\}\right\}. \tag{5.9}$$

If $\lambda_k(X;Y)$ is unique, then $h_k = f_k$ given in (5.6).

Proof. We will first prove that (1) $\iff$ (2), and then show that (2) $\iff$ (3) and (2) $\iff$ (4).

- (1) $\iff$ (2). Let $f \in \mathcal{L}_2(p_X)$ and $g \in \mathcal{L}_2(p_Y)$. Define the column-vectors $\mathbf{f} \triangleq (f(1),\ldots,f(m))^T$ and $\mathbf{g} \triangleq (g(1),\ldots,g(n))^T$. Then

  $$\mathbb{E}\left[f(X)g(Y)\right] = \mathbf{f}^T\mathbf{P}\mathbf{g}$$

  and

  $$\mathbf{f}^T\mathbf{D}_X\mathbf{f} = \mathbf{g}^T\mathbf{D}_Y\mathbf{g} = 1.$$

  For $\mathbf{Q} = \mathbf{D}_X^{-1/2}\mathbf{P}\mathbf{D}_Y^{-1/2} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, put $\mathbf{u} \triangleq \mathbf{U}^T\mathbf{D}_X^{1/2}\mathbf{f}$ and $\mathbf{v} \triangleq \mathbf{V}\mathbf{D}_Y^{1/2}\mathbf{g}$. Then $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$, and

  $$\mathbb{E}\left[f(X)g(Y)\right] = \mathbf{u}^T\mathbf{\Sigma}\mathbf{v}.$$

  The result then follows directly from the variational characterization of singular values.

  Note that the column-vectors $(\mathbf{f}_0,\mathbf{f}_1,\ldots,\mathbf{f}_d)$ corresponding to the functions $(f_0,f_1,\ldots,f_d)$ are the first $d+1$ columns of $\mathbf{D}_X^{-1/2}\mathbf{U}$, and the column-vectors $(\mathbf{g}_0,\mathbf{g}_1,\ldots,\mathbf{g}_d)$ corresponding to the functions $(g_0,g_1,\ldots,g_d)$ are the first $d+1$ of $\mathbf{D}_Y^{-1/2}\mathbf{V}$. In addition, let $\mathbf{z}_k \in \mathbb{R}^n$ be the column vector with entries $\mathbb{E}\left[f_k(X)|Y=j\right]$. Then

  $$\mathbf{z}_k = \mathbf{f}^T\mathbf{P}\mathbf{D}_Y^{-1} = \mathbf{f}_k^T\mathbf{D}_X^{1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{D}_Y^{-1/2} = \sqrt{\lambda_k(X;Y)}\mathbf{g}_k,$$

  so $\lambda_k(X;Y) = \|\mathbb{E}\left[f_k(X)|Y\right]\|_2^2$ and, consequently, $g_k(Y) = \frac{\mathbb{E}[f_k(X)|Y]}{\|\mathbb{E}[f_k(X)|Y]\|_2}$.

- (2) $\iff$ (3). The equivalence follows by noting that

  $$\begin{aligned}\sqrt{\lambda_1(X;Y)} &= \max_{\substack{\mathbb{E}[f(X)]=\mathbb{E}[g(X)]=0\\ \|f(X)\|_2=\|g(Y)\|_2=1}} \mathbb{E}\left[f(X)g(Y)\right] \\ &= \max_{\substack{\mathbb{E}[f(X)]=\mathbb{E}[g(X)]=0\\ \|f(X)\|_2=\|g(Y)\|_2=1}} \mathbb{E}\left[\mathbb{E}\left[g(Y)f(X)|Y\right]\right]\end{aligned}$$

$$= \max_{\substack{\mathbb{E}[f(X)]=0 \\ \|f(X)\|_2=1}} \|\mathbb{E}\left[f(X)|Y\right]\|_2,$$

where the last equality follows from $g_1(Y) = \frac{\mathbb{E}[f_1(X)|Y]}{\|\mathbb{E}[f_1(X)|Y]\|_2}$. Since this last expression is exactly the second largest term of the spectrum of the conditional expectation operator T (the largest being 1), the result follows for $\lambda_1(X;Y)$. The equivalent result for the other PICs by adding orthogonality constraints.

- (2) $\iff$ (4). The result follows directly from $\lambda_k(X;Y) = \|\mathbb{E}\left[f_k(X)|Y\right]\|_2^2$ and (1.1).

□

The previous theorem provides different operational characterization of the PICs. Characterization (1) lends itself to the geometric interpretation discussed in Section 5.1. Characterization (2) implies that the principal functions of X and Y are the solution to the following problem: Consider two parties, namely Alice and Bob, where Alice has access to an observation of X and Bob has access to an observation Y. Alice and Bob's goal is to produce zero-mean, unit variance functions $f(X)$ and $g(Y)$, respectively, that maximizes the correlation $\mathbb{E}\left[f(X)g(Y)\right]$ without communicating. The optimal choice of functions is exactly f_1 and g_1, given in the theorem. This also implies that

$$\lambda_1(X;Y) = \rho_m(X;Y)^2.$$

Characterization (4) above proves that the the PICs are the solution to another related question: Given a noisy observation Y of a hidden variable X, what is the unit-variance, zero-mean function of X that can be estimated with the smallest mean-squared error? It follows directly from (5.8) that the function is $f_1(X)$, and the minimum MMSE is $1 - \lambda_1(X;Y)$. Indeed, the principal functions form a basis for $\mathcal{L}_2(p_X)$, and are closely related to the MMSE analysis presented in Chapter 3. We will return to the connection of the PICs and MMSE in chapter 7.

Tensorization of the PICs

The next result states the well-known tensorization property the PICs between sequences of independent random variables. We present a proof of the discrete case here for the sake of completeness.

Lemma 5.1. *Let* $(X_1, Y_1) \perp\!\!\!\perp (X_2, Y_2)$, $d_1 = \min\{|\mathcal{X}_1|, |\mathcal{Y}_1|\} - 1$ *and* $d_2 = \min\{|\mathcal{X}_2|, |\mathcal{Y}_2|\} - 1$. *Then the PICs of* $p_{(X_1,X_2),(Y_1,Y_2)}$ *are* $\lambda_i(X_1, Y_1)\lambda_j(X_2, Y_2)$ *for* $(i, j) \in [0, d_1] \times [0, d_2]$, *where* $\lambda_0(X_1, Y_1) = \lambda_0(X_2, Y_2) = 1$. *Furthermore, the principal functions* (X_1, Y_1) *by* f_i *and of* (X_2, Y_2) *by* $\widetilde{f}_j$, *then the PICs of* $p_{(X_1,X_2),(Y_1,Y_2)}$ *are of the form* $(x_1, x_2) \mapsto f_i(x_1)\widetilde{f}_j(x_2)$. *In particular*

$$\lambda_1((X_1, X_2); (Y_1, Y_2)) = \max\{\lambda_1(X_1; Y_1), \lambda_1(X_2; Y_2)\}.$$

Proof. Let $[\mathbf{Q}_1]_{i,j} = \frac{p_{X_1,Y_1}(i,j)}{\sqrt{p_{X_1}(i)p_{Y_1}(j)}}$ and $[\mathbf{Q}_2]_{i,j} = \frac{p_{X_2,Y_2}(i,j)}{\sqrt{p_{X_2}(i)p_{Y_2}(j)}}$. Denoting by $\mathbf{Q}$ the decomposition in Definition 5.1 of $p_{(X_1,X_2),(Y_1,Y_2)}$ then, due to the independence assumption, $\mathbf{Q} = \mathbf{Q}_1 \otimes \mathbf{Q}_2$, where $\otimes$ is the Kronecker product. The result follows directly from the fact that the singular values of the Kronecker product of two matrices is the Kronecker product of the singular values (and equivalently for the singular vectors). □

5.5 A Measure of Information Based on the PICs

In this section we introduce the k-correlation measure, which is equivalent to the sum of the k largest PICs. We prove that k-correlation is convex in $p_{Y|X}$ and satisfies the Data Processing Inequality, being an information measure according to Def. 4.1.

Throughout this section, we denote $\lambda_k(X;Y) = \lambda_k$ for short. Consider the matrix $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ given in (5.3), and define

$$\widetilde{\mathbf{A}} \triangleq \mathbf{D}_X^{1/2}\mathbf{U}, \ \widetilde{\mathbf{B}} \triangleq \mathbf{D}_Y^{1/2}\mathbf{V}.$$

Then

$$\mathbf{P} = \widetilde{\mathbf{A}}\mathbf{\Sigma}\widetilde{\mathbf{B}}^T, \tag{5.10}$$

where $\widetilde{\mathbf{A}}^T\mathbf{D}_X^{-1}\widetilde{\mathbf{A}} = \widetilde{\mathbf{B}}^T\mathbf{D}_Y^{-1}\widetilde{\mathbf{B}} = \mathbf{I}$.

It follows directly from Theorem 5.1 that $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$ and $\mathbf{\Sigma}$ have the form

$$\widetilde{\mathbf{A}} = [\mathbf{p}_X \ \ \mathbf{A}], \ \widetilde{\mathbf{B}} = [\mathbf{p}_Y \ \mathbf{B}], \ \mathbf{\Sigma} = \operatorname{diag}\left(1, \sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}\right), \tag{5.11}$$

and, consequently, the joint distribution can be written as

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) + \sum_{k=1}^{d} \sqrt{\lambda_k} b_{y,k} a_{x,k}, \tag{5.12}$$

where $a_{x,k}$ and $b_{y,k}$ are the entries of $\mathbf{A}$ and $\mathbf{B}$ in (5.11), respectively.

Using this decomposition of the joint distribution matrix, we define below a measure of information between X and Y based on the PICs.

Definition 5.2. Let $\|\mathbf{C}\|_k$ denote the k-th Ky Fan norm[1] ([86, Example 7.4.8]) of a matrix $\mathbf{C}$. For $1 \leq k \leq d$, we define the *k-correlation* between X and Y as

$$\mathcal{J}_k(X;Y) \triangleq \|\mathbf{Q}\mathbf{Q}^T\|_k - 1 \tag{5.13}$$

$$= \sum_{i=1}^{k} \lambda_i. \tag{5.14}$$

[1] For $\mathbf{C} \in \mathbb{R}^{m\times n}$, $\|\mathbf{C}\|_k = \sum_{i=1}^{k} \sigma_i(\mathbf{C})$.

Note that

$$\mathcal{J}_1(X;Y) = \rho_m(X;Y)^2,$$

and

$$\mathcal{J}_d(X;Y) = \mathbb{E}\left[\frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right] - 1 = \chi^2(X;Y).$$

We now show that k-correlation and, consequently, maximal correlation, is convex in $p_{Y|X}$ for a fixed p_X and satisfies the Data Processing Inequality.

Convexity in $p_{Y|X}$

We use the next lemma to prove convexity of $\mathcal{J}_k(X;Y)$ in the transition probability $p_{Y|X}$.

Lemma 5.2. *For $\mathbf{W} \in \mathcal{P}_{++}^m$ and $1 \leq k \leq m$, the function $h_k : \mathbb{R}^{m\times n} \times \mathcal{P}_{++}^m \to \mathbb{R}$ defined as*

$$h_k(\mathbf{C}, \mathbf{W}) \triangleq \|\mathbf{C}\mathbf{W}^{-1}\mathbf{C}^T\|_k \tag{5.15}$$

is convex.

Proof. Let $\mathbf{Y} \triangleq \mathbf{C}\mathbf{W}^{-1}\mathbf{C}^T$. Since $\mathbf{Y}$ is symmetric and positive semidefinite, $\|\mathbf{Y}\|_k$ is the sum of the k largest eigenvalues of $\mathbf{Y}$, and can be written as [87,88]:

$$h_k(\mathbf{C}, \mathbf{W}) = \|\mathbf{Y}\|_k = \max_{\mathbf{Z}^T\mathbf{Z}=\mathbf{I}_k} \operatorname{tr}\left(\mathbf{Z}^T\mathbf{Y}\mathbf{Z}\right), \tag{5.16}$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix. Let $\mathbf{Z}$ be fixed and $\mathbf{Z}^T\mathbf{Z} = \mathbf{I}_k$, and denote the i-th column of $\mathbf{Z}$ by $\mathbf{z}_i$. Note that $g(\mathbf{a}, \mathbf{W}) \triangleq \mathbf{a}^T\mathbf{W}^{-1}\mathbf{a}$ is convex [89, Example 3.4] and, consequently, $g(\mathbf{C}^T\mathbf{z}_i, \mathbf{W})$ is also convex in $\mathbf{C}$ and $\mathbf{W}$. Since the sum of convex functions is itself convex, then $\operatorname{tr}\left(\mathbf{Z}^T\mathbf{Y}\mathbf{Z}\right) = \sum_{i=1}^m g(\mathbf{C}^T\mathbf{z}_i, \mathbf{W})$ is also convex in $\mathbf{C}$ and bW. The result follows by noting that the pointwise supremum over an infinite set of convex functions is also a convex function [89, Sec. 3.2.3]. □

Theorem 5.2. *For a fixed p_X, $\mathcal{J}_k(X;Y)$ is convex in $p_{Y|X}$.*

Proof. Note that $\mathcal{J}_k(X;Y) = h_k(\mathbf{D}_X\mathbf{P}_{Y|X}, \mathbf{D}_Y) - 1$, where h_k is defined in equation (5.15). For a fixed p_X, $\mathbf{D}_Y$ is a linear combination of $p_{Y|X}$. Therefore, since h_k is convex (Lemma 5.2), and composition with an affine mapping preserves convexity, the result follows. □

A Data Processing Result

The following lemma will be used to prove that the PICs satisfy the Data Processing Inequality.

Lemma 5.3 (DPI for MMSE). *For $X \to Y \to Z$ and any $f \in \mathcal{L}_2(p_X)$, $\mathbb{E}[f(X)] = 0$,*

$$\|\mathbb{E}[f(X)|Z]\|_2^2 \leq \lambda_1(Y;Z)\|\mathbb{E}[f(X)|Y]\|_2^2. \tag{5.17}$$

Consequently, $\mathsf{mmse}(f(X)|Y) \leq \mathsf{mmse}(f(X)|Z)$.

Proof. Let $f \in \mathcal{L}_2(p_X)$, $\mathbb{E}[f(X)] = 0$ and $g \in \mathcal{L}_2(Z)$, $\mathbb{E}[g(Z)] = 0$, $\|g(Z)\|_2 = 1$. Then

$$\begin{aligned}
\mathbb{E}[f(X)g(Z)] &= \mathbb{E}[\mathbb{E}[f(X)g(Z)|Y]] \\
&\overset{(a)}{=} \mathbb{E}[\mathbb{E}[f(X)|Y]\,\mathbb{E}[g(Z)|Y]] \\
&\overset{(b)}{\leq} \|\mathbb{E}[f(X)|Y]\|_2 \|\mathbb{E}[g(Z)|Y]\|_2 \\
&\overset{(c)}{\leq} \sqrt{\lambda_1(Z;Y)}\|\mathbb{E}[f(X)|Y]\|_2,
\end{aligned}$$

where (a) follows from the assumption that $X \to Y \to Z$, (b) follows from the Cauchy-Schwarz inequality, and (c) follows from characterization (3) in Theorem 5.1. The result then follows by choosing $g(z) = \mathbb{E}[f(X)|Z = z]\,/\|\mathbb{E}[f(X)|Z]\|_2$. □

Lemma 5.3 leads to the following theorem.

Theorem 5.3 (DPI for the PICs). *Assume that $X \to Y \to Z$. Then $\lambda_k(X;Z) \leq \lambda_1(Y;Z)\lambda_k(X;Y)$ for all k.*

Remark 5.1. This data processing result was also proved by Kang and Ulukus in [82, Theorem 2], even though they do not make the explicit connection with maximal correlation and PICs. A weaker form of Theorem 5.3 can be derived using a clustering result presented in [16, Sec. 7.5.4] and originally due to Deniau *et al.* [90]. We use a different proof technique from the one in [16, Sec. 7.5.4] and [82, Theorem 2] to show result stated in the theorem, and present the proof here for completeness. Finally, a related data processing result was stated in [79].

Proof. A direct consequence of Theorem 5.1 is that for any two random variables X, Y

$$\lambda_k(X;Y) = \min_{\{f_i\}_{i=1}^k \subseteq \mathcal{L}_2(p_X)} \max_{\substack{f \in \mathcal{L}_2(p_X) \\ \mathbb{E}[f(X)f_i(X)]=0}} \|\mathbb{E}[f(X)|Y]\|_2^2,$$

and equivalently for $\lambda_k(X;Z)$. The result then follows directly from (5.17). □

The next corollary is a direct consequence of the previous theorem.

Corollary 5.1. *For $X \to Y \to Z$ forming a Markov chain, $\mathcal{J}_k(X;Z) \leq \lambda_1(Y;Z)\mathcal{J}_k(X;Y)$.*

5.6 A Lower Bound for the Estimation Error Probability in Terms of the PICs

Throughout the rest of the chapter, we assume without loss of generality that p_X is sorted in decreasing order, i.e. $p_X(1) \geq p_X(2) \geq \ldots \geq p_X(m)$.

Definition 5.3. Let $\mathbf{\Lambda}(p_{X,Y})$ denote the vector of PICs of a joint distribution $p_{X,Y}$ sorted in decreasing order, i.e. $\mathbf{\Lambda}(p_{X,Y}) = (\lambda_1(X;Y), \ldots, \lambda_d(X;Y))$. We denote $\mathbf{\Lambda}(p_{X,Y}) \leq \widetilde{\boldsymbol{\lambda}} \triangleq (\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_d)$ if $\lambda_1(X;Y) \leq \widetilde{\lambda}_1, \ldots, \lambda_d(X;Y) \leq \widetilde{\lambda}_d$ and

$$\mathcal{R}(q, \widetilde{\boldsymbol{\lambda}}) \triangleq \left\{ p_{X,Y} | p_X = q \text{ and } \mathbf{\Lambda}(p_{X,Y}) \leq \widetilde{\boldsymbol{\lambda}} \right\}. \tag{5.18}$$

In the next theorem we present a Fano-like bound for the estimation error probability of X that depends on the marginal distribution p_X and on the principal inertias.

Theorem 5.4. *For $\widetilde{\boldsymbol{\lambda}} = (\lambda_1, \ldots, \lambda_d)$ and fixed p_X, define $c_0 = \sum_{i \in [m]} p_X(i)^2$,*

$$k^* \triangleq \max \left\{ k \in [m] \;\middle|\; p_X(k) - c_0 \geq 0 \right\}, \tag{5.19}$$

$$f_0^*(p_X, \boldsymbol{\lambda}) \triangleq \sum_{i=1}^{k^*} \lambda_i p_X(i) + \sum_{i=k^*+1}^{m} \lambda_{i-1} p_X(i) - \lambda_{k^*} c_0 ,$$

$$\begin{aligned} g_0(\beta, p_X, \boldsymbol{\lambda}) &\triangleq f_0^*(p_X, \boldsymbol{\lambda}) + \sum_{i=1}^{m} \left([p_X(i) - \beta]^+ \right)^2, \\ U_0(\beta, p_X, \boldsymbol{\lambda}) &\triangleq \beta + \sqrt{g_0(\beta, p_X, \boldsymbol{\lambda})}, \\ U_1(p_X, \boldsymbol{\lambda}) &\triangleq \min_{0 \leq \beta \leq p_X(2)} U_0(\beta, p_X, \boldsymbol{\lambda}). \end{aligned}$$

Then for any joint pmf $q_{X,Y} \in \mathcal{R}(px, \boldsymbol{\lambda})$,

$$P_e(X|Y) \geq 1 - U_1(p_X, \boldsymbol{\lambda}). \tag{5.20}$$

Proof. The proof of the theorem is presented in the appendix. □

Remark 5.2. If $\lambda_i = 1$ for all $1 \leq i \leq d$, (5.20) reduces to $P_e(X|Y) \geq 0$. Furthermore, if $\lambda_i = 0$ for all $1 \leq i \leq d$, (5.20) simplifies to $P_e(X|Y) \geq 1 - p_X(1)$.

We now present a few direct but powerful corollaries of the result in Theorem 5.4. We note that a bound similar to (5.21) below has appeared in the context of bounding the minmax decision risk in [91, (3.4)]. However, the proof technique used in [91] does not lead to the general bound presented in Theorem 5.4.

Corollary 5.2. *If X is uniformly distributed in $[m]$, then*

$$P_e(X|Y) \geq 1 - \frac{1}{m} - \frac{\sqrt{(m-1)\chi(X;Y)^2}}{m} \,. \tag{5.21}$$

Furthermore, if only a bound on the maximal correlation $\rho_m(X;Y) = \sqrt{\lambda_1}$ is given, then

$$\begin{aligned} P_e(X|Y) &\geq 1 - \frac{1}{m} - \sqrt{\lambda_1}\left(1 - \frac{1}{m}\right) \\ &= 1 - \frac{1}{m} - \rho_m(X;Y)\left(1 - \frac{1}{m}\right). \end{aligned}$$

Corollary 5.3. *For any pair of variables (X,Y) with marginal distribution in X equal to p_X and maximal correlation (largest principal inertia) $\rho_m^2(X;Y) = \lambda_1$, we have for all $\beta \geq 0$*

$$P_e(X|Y) \geq 1 - \beta - \sqrt{\lambda_1\left(1 - \sum_{i=1}^{m} p_X^2(i)\right) + \sum_{i=1}^{m}\left([p_X(i) - \beta]^+\right)^2} \,. \tag{5.22}$$

In particular, setting $\beta = p_X(2)$,

$$\begin{aligned} P_e(X|Y) &\geq 1 - p_X(2) - \sqrt{\lambda_1\left(1 - \sum_{i=1}^{m} p_X^2(i)\right) + \left(p_X(1) - p_X(2)\right)^2} \\ &\geq 1 - p_X(1) - \rho_m(X;Y)\sqrt{\left(1 - \sum_{i=1}^{m} p_X^2(i)\right)} \,. \end{aligned} \tag{5.23}$$

Remark 5.3. The bounds (5.22) and (5.23) are particularly insightful in showing how the error probability scales with the input distribution and the maximal correlation. For a given $p_{X,Y}$, recall that $\mathsf{Adv}(X|Y)$, defined in (5.2), is the advantage of correctly estimating X from an observation of Y over a random guess of X when Y is unknown. Then, from equation (5.23)

$$\begin{aligned} \mathsf{Adv}(X|Y) &\leq \rho_m(X;Y)\sqrt{\left(1 - \sum_{i=1}^{m} p_X^2(i)\right)} \\ &\leq \rho_m(X;Y). \end{aligned}$$

Therefore, the advantage of estimating X from Y decreases at least linearly with the maximal correlation between X and Y.

5.7 The Error-Rate Function for k-Correlation

We present in this section a convex program for lower-bounding the error-rate function for k-correlation $e_{\mathcal{J}_k}(p_X, \theta)$ (cf. Definition 4.3). For $\mathcal{I} = \mathcal{J}_k$, the convex program (4.3) under Hamming distortion may be difficult to compute due to the constraint on the sum of the singular values. The next theorem presents a convex program that evaluates a lower bound for $e_{\mathcal{J}_k}(p_X, \theta)$ and can be solved using standard methods.

Theorem 5.5. *For $\mathcal{X} = [m]$ and p_X given,*

$$
\begin{aligned}
e_{\mathcal{J}_k}(p_X, \theta) \geq \min_{\mathbf{P}_{\hat{X}|X}} \; & 1 - \operatorname{tr}\left(\mathbf{D}_X \mathbf{P}_{\hat{X}|X}\right) \\
\text{s.t.} \; & \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{p_X(i) p^2_{\hat{X}|X}(j|i)}{y_j} \leq \theta + 1, && (5.24) \\
& \mathbf{P}_{\hat{X}|X} \text{ is row stochastic}, && (5.25) \\
& [\mathbf{P}_{\hat{X}|X}]_{i,j} = p_{\hat{X}|X}(j|i), \\
& \sum_{j=1}^{m} p_X(i) p_{\hat{X}|X}(j|i) = y_j, \; 1 \leq j \leq m.
\end{aligned}
$$

Proof. We prove that the previous optimization program is convex and lower bounds $e_{\mathcal{J}_k}(p_X, \theta)$. Put $\mathbf{F} \triangleq \mathbf{D}_X^{-1/2} \mathbf{P} \mathbf{D}_Y^{-1/2}$. Then

$$
\mathcal{J}_k(X;Y) = \|\mathbf{F}\mathbf{F}^T\|_k - 1.
$$

Let

$$
c_i \triangleq \sum_{j=1}^{m} \frac{p_X(i) p^2_{\hat{X}|X}(j|i)}{y_j}
$$

be the i-th diagonal entry of $\mathbf{F}\mathbf{F}^T$. By using the fact that the eigenvalues majorize the diagonal entries of a Hermitian matrix ([86, Theorem 4.3.45]), we find

$$
\sum_{i=1}^{k} c_i \leq \|\mathbf{F}\mathbf{F}^T\|_k,
$$

and the result follows. Note that convexity of the constraint (5.24) follows from the fact that the perspective of a convex function is convex [89, Sec. 2.3.3]. □

In addition, inequality (5.23) can be used to bound the probability of correctly guessing any function of X from an observation of Y, as shown next

Lemma 5.4. *Let* $\mathcal{J}_1(X;Y) = \rho_m(X;Y) \leq \theta$. *Then*

$$P_{e,M}(X|Y) \geq 1 - p_U(1) - \theta\sqrt{\left(1 - \sum_{i=1}^{M} p_U^2(i)\right)},$$

where $P_{e,M}(X|Y)$ *is defined in* (4.8) *and* U *is defined as in Theorem 4.2.*

Proof. The proof follows directly from Theorems 5.2, 5.3 and Corollary 5.3, by noting that (5.23) is Schur-concave in p_X. □

The previous Lemma leads to the following powerful theorem, which states that the probability of guessing *any* function of a hidden variable X from an observation Y is upper bounded by the maximal correlation of X and Y.

Theorem 5.6. *Let* p_X *be fixed,* $|\mathcal{X}| \leq \infty$ *and* $\mathcal{F}_M$ *be given in* (4.7). *Define*

$$\mathsf{Adv}_M(X|Y) \triangleq \max\left\{1 - \max_{k\in[M]} p_{f(X)}(k) - P_e(f(X)|Y) \mid f \in \mathcal{F}_M\right\}.$$

Then

$$\mathsf{Adv}_M(X|Y) \leq \rho_m(X;Y)\sqrt{1 - \frac{1}{M}} \leq \rho_m(X;Y). \tag{5.26}$$

Proof. For $f \in \mathcal{F}_M$

$$\begin{aligned}
\mathsf{Adv}(f(X)|Y) &\leq \rho(f(X);Y)\sqrt{1 - \sum_{i\in[M]} p_{f(X)(i)^2}} \\
&\leq \rho(X;Y)\sqrt{1 - \sum_{i\in[M]} p_{f(X)(i)^2}} \\
&\leq \rho(X;Y)\sqrt{1 - \frac{1}{M}},
\end{aligned}$$

where the first inequality follows from (5.23), the second inequality follows from the DPI for the PICs (Theorem 5.3), and the last inequality follows from the fact that $\sum_{i\in[M]} p_{f(X)}(i)^2$ is minimized when $p_{f(X)}$ is uniform. The result follows by maximizing over all $f \in \mathcal{F}_M$. □

5.8 Additional Results for the PICs[2]

In this section we derive bounds for the PICs between a set of i.i.d. samples of a discrete random variable and a symmetric function of these samples by studying the properties of the conditional expectation operator $T : \mathcal{L}_2(p_Y) \to \mathcal{L}_2(p_Y)$ (introduced earlier in this chapter in Theorem 5.1), where $(Tf)(x) \triangleq \mathbb{E}[f(Y)|X = x]$. We revisit next a few properties of T.

[2]The author acknowledges and is indebted to Prof. Y. Polyanskiy (yp@mit.edu) for his contributions in this Section, in particular in Lemma 5.5.

Recall that, from Jensen's inequality, the largest singular value of T is 1, achieved when f, with corresponding right eigenvector $f(y) = 1$. Furthermore, from Theorem 5.1 the PICs of X and Y are the square of the singular values of T. To see why this the case, note that for any $f \in \mathcal{L}(p_X)$ and $g \in \mathcal{L}(p_Y)$ with zero mean, we have

$$\begin{aligned}\mathbb{E}\left[f(X)g(Y)\right] &= \mathbb{E}\left[f(X)\mathbb{E}\left[g(Y)|X\right]\right] \\ &= \mathbb{E}\left[f(X)(Tg)(X)\right] \\ &\leq \|Tg(X)\|_2,\end{aligned}$$

where the last inequality follows directly by applying Cauchy-Schwarz. Consequently, $\mathbb{E}\left[f(X)g(Y)\right]$ is maximized when g maximizes $\|Tg(X)\|_2$, which, in turn, results in the largest singular value of the operator T. Consequently, as stated in Theorem 5.1, the square of the singular values of T or the PICs of $p_{X,Y}$.

Rather surprisingly, we can characterize the singular values of T for a wide range of probability distributions using the Efron-Stein decomposition [92]. For this, we use a proof technique similar to the one used by Dembo *et al.* [93] to characterize the maximal correlation between the sum of i.i.d. random variables. The following analysis follows similar steps of the proof of [93, Theorem 1].

Let $Y_1, \ldots, Y_n$ independent random variables and $Y^n = (Y_1, \ldots, Y_n)$. Then any function $f : \mathcal{Y}_1 \times \ldots \times Y_n \to \mathbb{R}$ can be decomposed into its "higher order interactions" [92] as

$$f(Y^n) = \sum_{\mathcal{S} \subseteq [n]} f_{\mathcal{S}}(Y_{\mathcal{S}}), \tag{5.27}$$

where $[n] = \{1, \ldots, n\}$, $Y_{\mathcal{S}} = (Y_i : i \in \mathcal{S})$ and $f_\emptyset = \mathbb{E}\left[f(Y^n)\right]$. Each coefficient $f_{\mathcal{S}}$ depends only on $Y_{\mathcal{S}}$, and, for $\mathcal{S} \not\subseteq \mathcal{S}'$, it holds that $\mathbb{E}\left[f_{\mathcal{S}}(Y_{\mathcal{S}})|Y_{\mathcal{S}'}\right] = 0$. The function $f_{\mathcal{S}}$ is unique and is given by

$$f_{\mathcal{S}}(y_{\mathcal{S}}) = \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S} \setminus \mathcal{S}'|} \mathbb{E}\left[f(Y^n)|Y_{\mathcal{S}'} = y_{\mathcal{S}'}\right].$$

Observe that if $\mathcal{S} \neq \mathcal{S}'$, then $\mathcal{S} \not\subseteq \mathcal{S}'$ and/or $\mathcal{S} \not\subseteq \mathcal{S}'$. Assuming the former, we have

$$\mathbb{E}\left[f_{\mathcal{S}}(Y_{\mathcal{S}}) f_{\mathcal{S}'}(Y_{\mathcal{S}'})\right] = \mathbb{E}\left[f_{\mathcal{S}'}(Y_{\mathcal{S}'})\mathbb{E}\left[f_{\mathcal{S}}(Y_{\mathcal{S}})|Y_{\mathcal{S}'}\right]\right] = 0.$$

Therefore $\mathbb{E}\left[f(Y^n)^2\right] = \sum_{\mathcal{S} \in [n]} \mathbb{E}\left[f_{\mathcal{S}}(Y_{\mathcal{S}})^2\right]$.

Assume henceforth that $f : \mathcal{Y}^n \to \mathbb{R}$ is a symmetric function (i.e. invariant to permutations in its arguments) and $Y_1, \ldots, Y_n$ are identically distributed. Then $f_{\mathcal{S}}$ depends only on the cardinality of $\mathcal{S}$. In this case

$$f(Y^n) = \sum_{\mathcal{S} \subseteq [n]} f_{|\mathcal{S}|}(Y_{\mathcal{S}}). \tag{5.28}$$

and

$$\mathbb{E}\left[f(Y^n)^2\right] = \sum_{i=0}^{n} \binom{n}{i} \mathbb{E}\left[f_i(Y^i)^2\right]. \tag{5.29}$$

Then, for $m \leq n$,

$$\begin{aligned}(Tf)(Y^m) &= \mathbb{E}\left[f(Y)|Y^m\right] \\ &= \sum_{\mathcal{S}\subseteq[n]} \mathbb{E}\left[f_{|\mathcal{S}|}(Y_{\mathcal{S}})|Y^m\right] \\ &= \sum_{\mathcal{S}\subseteq[m]} f_{|\mathcal{S}|}(Y_{\mathcal{S}}).\end{aligned}$$

Observe that $(Tf)(Y^m)$ is also symmetric. In order to simplify notation, we denote $(Tf)(Y^m)$ by Tf and $f(Y^n)$ by f. The second moment of Tf is given by

$$\|(Tf)\|_2^2 = \sum_{i=0}^{m} \binom{m}{i} \mathbb{E}\left[f_i(Y^i)^2\right].$$

Therefore,

$$\frac{\|Tf\|_2^2}{\|f\|_2^2} = \frac{\sum_{i=0}^{m} \binom{m}{i} \mathbb{E}\left[f_i(Y^i)^2\right]}{\sum_{i=0}^{n} \binom{n}{i} \mathbb{E}\left[f_i(Y^i)^2\right]}. \tag{5.30}$$

In particular, if $f(Y^n)$ has zero mean and unit variance, then

$$\mathbb{E}\left[(Tf)^2\right] = \frac{\sum_{i=1}^{m} \binom{m}{i} \mathbb{E}\left[f_i(Y^i)^2\right]}{\sum_{i=1}^{n} \binom{n}{i} \mathbb{E}\left[f_i(Y^i)^2\right]}. \tag{5.31}$$

We can maximize (5.31) by solving the following linear program:

$$\begin{aligned}\max_{a\in\mathbb{R}^m} \quad & \sum_{i=1}^{m} \binom{m}{i} a_i \\ \text{s.t.} \quad & \sum_{i=1}^{m} \binom{n}{i} a_i = 1, \ a \geq 0,\end{aligned} \tag{5.32}$$

which is equivalent to

$$\begin{aligned}\max_{c\in\mathbb{R}^m} \quad & \sum_{i=1}^{m} \frac{\binom{m}{i}}{\binom{n}{i}} c_i \\ \text{s.t.} \quad & c_i = 1, \ c \geq 0,\end{aligned}$$

Note that $\binom{m}{i}/\binom{n}{i} = (m)_i/(n)_i$, where $(\cdot)_i$ is the Pochhammer symbol[3]. Since for any $i \geq 1$

[3] $(m)_k \triangleq m(m-1)\cdots(m-k+1)$.

$(m)_i/(n)_i \leq (m)_1/(n)_1 = m/n$, it follows directly that

$$\|Tf\|_2^2 \leq \frac{m}{n}. \tag{5.33}$$

Remark 5.4. Observe that if we restrict the symmetric function f to satisfy $f_i = 0$ for $i = 1, \ldots, k-1$, then the previous analysis leads to $\|Tf\|_2^2 \leq (m)_k/(n)_k$.

Let $\mathcal{F}_{\text{sym}}(\mathcal{Y}^n)$ denote the set of all real valued symmetric functions over $\mathcal{Y}^n$, and the operator $T_S : \mathcal{F}_{\text{sym}}(\mathcal{Y}^n) \to \mathcal{F}_{\text{sym}}(\mathcal{Y}^m)$ be the Markov operator restricted to symmetric functions, where $(T_S f)(y) = \mathbb{E}[f(Y^n)|Y^m = y]$. This leads to the following generalization of [93, Theorem 1].

Theorem 5.7. *Let $Y^n = (Y_1, \ldots, Y_n)$ be a collection of n i.i.d. random variables and $f \in \mathcal{F}_{sym}(\mathcal{Y}^n)$ satisfy $0 < \|f(Y^n)\|_2 < \infty$. Then $\lambda_1(f(Y^n); Y^m) \leq m/n$. Furthermore, if $|\mathcal{Y}| = d$, then for $i \in [n]$ and $k \in [d^{i-1}, \ldots d^i - 1]$*

$$\lambda_k(f(Y^n); Y^m) \leq \frac{(m)_i}{(n)_i}. \tag{5.34}$$

Proof. The bound $\lambda_1(f(Y^n); Y^m) \leq m/n$ follows directly from (5.33). Now let $|\mathcal{Y}| = d + 1$, and define

$$\lambda_1(\mathcal{F}_{\text{sym}}(Y^n); Y^m) \triangleq \max_{\substack{f \in \mathcal{F}_{\text{sym}}(\mathcal{Y}^n) \\ \mathbb{E}[f(Y^n)]=0, \|f(Y^n)\|_2=1}} \|Tf\|_2^2,$$

and, denoting the argument that maximizes the right-hand side of the previous expression as f_1^*, we define recursively,

$$\lambda_k(\mathcal{F}_{\text{sym}}(Y^n); Y^m) \triangleq \max_{\substack{f \in \mathcal{F}_{\text{sym}}(\mathcal{Y}^n) \\ \mathbb{E}[f(Y^n)]=0, \|f(Y^n)\|_2=1 \\ \mathbb{E}[f(Y^n) f_i^*(Y^n)]=0, i \in [k-1]}} \|Tf\|_2^2.$$

Since the linear space of functions with zero mean and non-zero variance in $\mathcal{L}_2(p_Y)$ has dimension $d-1$, there exists $d-1$ uncorrelated symmetric functions in $\mathcal{F}_{\text{sym}}(\mathcal{Y}^n)$ of the form $f(Y^n) = \sum_{i \in [n]} f_1(Y^n)$ (cf. decomposition (5.29)). It follows from (5.31) and (5.33) that these functions achieve $\|Tf(Y^m)\|_2^2 = m/n$, proving that $\lambda_1(\mathcal{F}_{\text{sym}}(Y^n); Y^m) = \cdots = \lambda_{d-1}(\mathcal{F}_{\text{sym}}(Y^n); Y^m) = m/n$. Denote the corresponding functions that achieve the maximum $f_1^* \ldots f_{d-1}^*$.

Now in order to characterize the next values of $\lambda_k(\mathcal{F}_{\text{sym}}(Y^n); Y^m)$ observe that any symmetric function f that has zero-mean and is orthogonal (uncorrelated) to $f_1^*, \ldots, f_{d-1}^*$ must have the "low-order interaction" terms f_1 in (5.29) equal to zero. Consequently, we must consider the functions in $\mathcal{L}_2(p_{Y^2})$ that are orthogonal to linear combinations of functions in $\mathcal{L}_2(p_Y)$. Consequently since the dimension of $\mathcal{L}_2(p_{Y^2})$ is d^2 and the dimension of $\mathcal{L}_2(p_Y)$ is d, the subspace of functions in $\mathcal{L}_2(p_{Y^2})$ that satisfied the orthogonality constraints is $d^2 - d = d(d-1)$. Using the exact same analysis that led to (5.33) then from Remark 5.4

it follows that for $k \in [d, d(d-1)]$

$$\lambda_k(\mathcal{F}_{\text{sym}}(Y^n); Y^m) = \frac{(m)_2}{(n)_2}.$$

The result then follows by proceeding inductively and noting that for any $f \in \mathcal{F}_{\text{sym}}(\mathcal{Y}^n)$ we have $\lambda_k(f(Y^n); Y^m) \leq \lambda_k(\mathcal{F}_{\text{sym}}(Y^n); Y^m)$. □

5.8.1 Functions that are Invariant Under Transitive Group Actions

Let $\mathcal{S}$ be a subset of $[m]$, and let G be a group that acts transitively on $[n]$ ($m \leq n$). Let $a_{\mathcal{S}}$ be the number of images of $\mathcal{S}$ under the action of G which is inside $[m]$ and $b_{\mathcal{S}}$ the total number of images of $\mathcal{S}$. We have the following lemma.

Lemma 5.5.

$$\frac{a_{\mathcal{S}}}{b_{\mathcal{S}}} \leq \frac{m}{n}. \tag{5.35}$$

Proof. Let g be a uniformly chosen element of G and let $\mathcal{S}_g$ be the image of $\mathcal{S}$ under the action of the group element g. Let X be a uniformly chosen element from $\mathcal{S}_g$. By transitivity of the action of G the distribution of X is uniform on $[n]$. Then we have

$$\frac{a_{\mathcal{S}}}{b_{\mathcal{S}}} = \Pr\{\mathcal{S}_g \subseteq [m]\} \leq \Pr\{X \in [m]\} = \frac{m}{n}.$$

□

The previous lemma combined with the Efron-Stein decomposition allows us to upper-bound the maximal correlation for a large class of functions that are invariant under a given transitive group action.

Theorem 5.8. *Let G be a finite group acting transitively on $[n]$, and let Y^n be i.i.d. Then for any function f that is G-invariant (e.g. if $f = \sum_{g \in G} f_1(Y^n \circ g)$ and f_1 is arbitrary) and has finite second-moment, then the maximal correlation ρ^* satisfies*

$$\lambda_1(Y^m; f(Y^n)) \leq \frac{m}{n}. \tag{5.36}$$

Proof. We assume, without loss of generality, $\mathbb{E}[f(Y^n)] = 0$ and $\|f\|_2 = 1$. We denote as before the conditional expectation operator as $Tf(Y^n) \triangleq \mathbb{E}[f(Y^n)|Y^m]$. Let $f(Y^n) = \sum_{\mathcal{S} \subseteq [n]} f_{\mathcal{S}}(Y_{\mathcal{S}})$ be the Efron-Stein decomposition of f. Since f is G-invariant and Y^n is i.i.d., $f_{\mathcal{S}} = f_{\mathcal{S} \circ g}$ for all $g \in G$. Let $\mathcal{F} \subseteq \mathcal{P}([n])$ be the smallest set of subsets of $[n]$ such that for every $\mathcal{S} \subseteq [n]$ there exists $\mathcal{S}' \in \mathcal{F}$ such that $f_{\mathcal{S}} = f_{\mathcal{S}'}$. Then

$$\|f\|_2^2 = \sum_{\mathcal{S} \in \mathcal{F}} b_{\mathcal{S}} \|f_{\mathcal{S}}\|_2^2,$$

and

$$\|Tf\|_2^2 = \sum_{\mathcal{S} \in (\mathcal{F} \cap \mathcal{P}([m]))} a_{\mathcal{S}} \|f_{\mathcal{S}}\|_2^2,$$

where $a_{\mathcal{S}}$ is the size of the image of $\mathcal{S}$ under the action of G which is inside $[m]$ and $b_{\mathcal{S}}$ is the total number of distinct images of $\mathcal{S}$ under G (as in the previous lemma). Then

$$\frac{\|Tf\|_2^2}{\|f\|_2^2} \leq \max_{\mathcal{S} \subseteq [m]} \frac{a_{\mathcal{S}}}{b_{\mathcal{S}}} \leq \frac{m}{n},$$

where the first inequality follows from the linear fractional optimization done in (5.32), and the second inequality follows from Lemma 5.5. The result follows directly. □

5.8.2 Exchangeable Random Variables

We now assume that the random variables $Y^\infty = (Y_1, \ldots, Y_n, \ldots)$ are exchangeable instead of i.i.d. and possess a finite support set. We say that Y^∞ is exchangeable if for every $\mathcal{S}, \mathcal{S}' \subseteq [n]$ such that $|\mathcal{S}| = |\mathcal{S}'|$ we have $p_{Y^{\mathcal{S}}} = p_{Y^{\mathcal{S}'}}$. It turns out that, in this case, the results presented in this section still hold, as shown in the next theorem.

Theorem 5.9. *Let $Y^\infty = (Y_1, \ldots, Y_n, \ldots)$ be a collection of exchangeable random variables. Then for any $f \in \mathcal{F}_{\text{sym}}(\mathcal{Y}^n)$*

$$\lambda_1(f(Y^n); Y^m) \leq \frac{m}{n}. \tag{5.37}$$

Proof. It follows from de Finetti's representation theorem [94] that there exists a random variable A such that, conditioned on A, $Y_1, Y_2, \ldots$ are independent. Therefore

$$\begin{aligned}
\sup_{f \in \mathcal{F}(Y^n)} \mathbb{E}\left[\mathbb{E}\left[f(Y^n)|Y^m\right]^2\right] &= \sup_{f \in \mathcal{F}(Y^n)} \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[f(Y^n)|Y^m\right]^2 \middle| A\right]\right] \\
&\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}(Y^n)} \mathbb{E}\left[\mathbb{E}\left[f(Y^n)|Y^m\right]^2 \middle| A\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\frac{m}{n} \middle| A\right]\right] \\
&= \frac{m}{n},
\end{aligned}$$

where the first inequality follows from Jensen's inequality, and the second follows from Theorem 5.7 and the fact that $(Y^1, \ldots, Y^n)$ are conditionally independent given A. □

5.9 Prologue to Chapters 6 and 7

The results presented in this chapter demonstrate that the PICs are a useful information measure that shed light on the fundamental limits of estimation. The PICs provide a full characterization of the functions of a hidden variable that can (or cannot) be estimated with

small mean-squared error. In addition Theorem 5.6 connects the largest PIC, namely the maximal correlation, with the probability of correctly guessing *any* function of a hidden, discrete random variable. This approach extends the results presented in Chapter 3. The PICs between a set of i.i.d. samples and symmetric functions can also be characterized explicitly, as shown in Theorem 5.7.

In the final chapters of this thesis, we demonstrate the application of the results and bounds based on the PICs to problems in information theory, privacy and security. The PICs play a central role in estimating one-bit functions of a hidden variable, and characterize the fundamental limits of perfect privacy. In addition, privacy-assuring mechanisms with strong privacy and utility guarantees can be found by solving convex programs based on the PICs. We describe these applications in detail next.

Chapter 6

Applications of the Principal Inertia Components to Information Theory

6.1 Overview

In this chapter, we present results that connect the principal inertia components with other information-theoretic metrics. As seen in Chapter 5, the distribution of the vectors $p_{Y|X}$ in the simplex or, equivalently, the PICs of the joint distribution of X and Y, is inherently connected to how an observation of Y is statistically related to X. In this chapter, we explore this connection within an information theoretic framework. We show that, under certain assumptions, the PICs play an important part in estimating a one-bit function of X, namely $f(X)$ where $f : \mathcal{X} \to \{0, 1\}$, given an observation of Y: they can be understood as the filter coefficients in the linear transformation of $p_{f(X)|X}$ into $p_{f(X)|Y}$. Alternatively, the PICs can bear an interpretation as the transform of the distribution of the noise in certain additive-noise channels, in particular when X and Y are binary strings. We also show that maximizing the PICs is equivalent to maximizing the first-order term of the Taylor series expansion of certain convex measures of information between $f(X)$ and Y. We conjecture that, for symmetric distributions of X and Y and a given upper bound on the value of the largest PIC, $I(f(X);Y)$ is maximized when all the principal inertia components have the same value as the largest principal inertia component. This is equivalent to Y being the result of passing X through a q-ary symmetric channel. This conjecture, if proven, would imply the conjecture made by Kumar and Courtade in [24].

Finally, we study the Markov chain $B \to X \to Y \to \widehat{B}$, where B and $\widehat{B}$ are binary random variables, and the role of the principal inertia components in characterizing the relation between B and $\widehat{B}$. We show that this relation is linked to solving a non-linear maximization problem, which, in turn, can be solved when $\widehat{B}$ is an unbiased estimate of B, the joint distribution of X and Y is symmetric and $\Pr\{B = \widehat{B} = 0\} \geq \mathbb{E}[B]^2$. We illustrate this result for the setting where X is a binary string and Y is the result of sending X through

a memoryless binary symmetric channel. We note that this is a similar setting to the one considered by Anantharam *et al.* in [95].

The rest of the chapter is organized as follows. Section 6.1.1 presents additional notation and definitions used in this chapter. Section 6.3 introduces the notion of conforming distributions and ancillary results. Section 6.4 presents results concerning the role of the principal inertia components in inferring one-bit functions of X from an observation of Y, as well as the linear transformation of p_X into p_Y in certain symmetric settings. We argue that, in such settings, the principal inertia components can be viewed as filter coefficients in a linear transformation. In particular, results for binary channels with additive noise are derived using techniques inspired by Fourier analysis of Boolean functions. Furthermore, Section 6.4 also introduces a conjecture that encompasses the one made by Kumar and Courtade in [24]. Finally, Section 6.5 provides further evidence for this conjecture by investigating the Markov chain $B \to X \to Y \to \widehat{B}$ where B and $\widehat{B}$ are binary random variables.

6.1.1 Notation

For a given joint distribution $p_{X,Y}$ and corresponding joint distribution matrix $\mathbf{P}$, the set of all vectors contained in the unit cube in $\mathbb{R}^n$ that satisfy $\|\mathbf{Px}\|_1 = a$ is given by

$$\mathcal{C}^n(a, \mathbf{P}) \triangleq \{\mathbf{x} \in \mathbb{R}^n | 0 \leq \mathbf{x}_i \leq 1, \|\mathbf{Px}\|_1 = a\}. \tag{6.1}$$

In this chapter, we represent the set of all $m \times n$ probability distribution matrices by $\mathcal{P}_{m,n}$.

For $x^n \in \{-1, 1\}^n$ and $\mathcal{S} \subseteq [n]$, $\chi_{\mathcal{S}}(x^n) \triangleq \prod_{i \in \mathcal{S}} x_i$ (we consider $\chi_{\emptyset}(x) = 1$). For $y^n \in \{-1, 1\}^n$, $a^n = x^n \oplus y^n$ is the vector resulting from the entrywise product of x^n and y^n, i.e. $a_i = x_i y_i$, $i \in [n]$.

Given two probability distributions p_X and q_X and $f(t)$ a smooth convex function defined for $t > 0$ with $f(1) = 0$, the f-divergence is defined as [96]

$$D_f(p_X || q_X) \triangleq \sum_x q_X(x) f\left(\frac{p_X(x)}{q_X(x)}\right). \tag{6.2}$$

The f-information is given by

$$I_f(X;Y) \triangleq D_f(p_{X,Y} || p_X p_Y). \tag{6.3}$$

When $f(x) = x \log(x)$, then $I_f(X;Y) = I(X;Y)$. A study of information metrics related to f-information was given in [97] in the context of channel coding converses.

6.2 Main Contributions

This chapter presents different applications of the PICs to information theory, and is based on the author's work in [98]. In Section 6.4, we demonstrate that the PICs have a role

in channel transformations very similar to that of filter coefficients in a linear filter. Here, functions of a hidden variable are analogous to the input signals of the filter. This is illustrated through an example in binary additive noise channels, where we argue that the binary symmetric channel is somewhat equivalent to a "low-pass filter". We then use this interpretation to study the "one-bit function conjecture" [24]. We present further evidence for that conjecture here, and introduce another, related conjecture based on the discussion.

The new conjecture (cf. Conjecture 6.1), albeit not proved, is more general than the "one-bit function conjecture". It states that, given a symmetric distribution $p_{X,Y}$, if we generate a new distribution $q_{X,Y}$ by making all the PICs of $p_{X,Y}$ equal to the largest one, then the new distribution is "more informative" about bits of X. By "more informative", we mean that, for any 1-bit function b, $I(b(X);Y)$ is larger under $q_{X,Y}$ than under $p_{X,Y}$. Indeed, from the previous discussion and from an estimation-theoretic perspective, any function of X can be estimated with smaller MMSE when considering $q_{X,Y}$ than $p_{X,Y}$. Furthermore, in this case, we show that $q_{X,Y}$ is a q-ary symmetric channel. This conjecture, if proven, would imply as a corollary the original one-bit function conjecture.

We resolve the one-bit function conjecture in a specific setting in Section 6.5. Instead of considering the mutual information between $b(X)$ and Y, we study the mutual information between $b(X)$ and a one-bit estimator $\hat{b}(Y)$. We show in Theorem 6.3 that, when $\hat{b}(Y)$ is an unbiased estimator, the information that $\hat{b}(Y)$ carries about $b(X)$ can be upper-bounded for a wide range of information measures. This result also leads to bounds on estimation error probability that recovers a particular form of the result stated in Theorem 5.4.

6.3 Conforming distributions

In this chapter we shall focus on probability distributions that meet the following definition.

Definition 6.1. A joint distribution $p_{X,Y}$ is said to be *conforming* if the corresponding matrix $\mathbf{P}$ satisfies $\mathbf{P} = \mathbf{P}^T$ and $\mathbf{P}$ is positive-semidefinite.

Conforming distributions are particularly interesting since they are closely related to symmetric channels[1]. In addition, if a joint distribution is conforming, then its eigenvalues are equivalent to its principal inertia components when X is uniform. We shall illustrate this relation in the following two lemmas and in Section 6.4.

Remark 6.1. If X and Y have a conforming joint distribution, then they have the same marginal distribution. Consequently, $\mathbf{D} \triangleq \mathbf{D}_X = \mathbf{D}_Y$, and $\mathbf{P} = \mathbf{D}^{1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T\mathbf{D}^{1/2}$.

Lemma 6.1. *If* $\mathbf{P}$ *is conforming, then the corresponding conditional distribution matrix* $\mathbf{P}_{Y|X}$ *is positive semi-definite. Furthermore, for any symmetric channel* $\mathbf{P}_{Y|X} = \mathbf{P}_{Y|X}^T$, *there is an input distribution* p_X *(namely, the uniform distribution) such that the principal*

[1] We say that a channel is symmetric if $\mathbf{P}_{Y|X} = \mathbf{P}_{Y|X}^T$.

inertia components of $\mathbf{P} = \mathbf{D}_X \mathbf{P}_{Y|X}$ *correspond to the square of the eigenvalues of* $\mathbf{P}_{Y|X}$. *In this case, if* $\mathbf{P}_{Y|X}$ *is also positive-semidefinite, then* $\mathbf{P}$ *is conforming.*

Proof. Let $\mathbf{P}$ be conforming and $\mathcal{X} = \mathcal{Y} = [m]$. Then $\mathbf{P}_{Y|X} = \mathbf{D}^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T\mathbf{D}^{1/2} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^{-1}$, where $\mathbf{Q} = \mathbf{D}^{-1/2}\mathbf{U}$. It follows that $\operatorname{diag}(\mathbf{\Sigma})$ are the eigenvalues of $\mathbf{P}_{Y|X}$, and, consequently, $\mathbf{P}_{Y|X}$ is positive semi-definite.

Now let $\mathbf{P}_{Y|X} = \mathbf{P}_{Y|X}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. The entries of $\mathbf{\Lambda}$ here are the eigenvalues of $\mathbf{P}_{Y|X}$ and not necessarily positive. Since $\mathbf{P}_{Y|X}$ is symmetric, it is also doubly stochastic, and for X uniformly distributed Y is also uniformly distributed. Therefore, $\mathbf{P}$ is symmetric, and $\mathbf{P} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T/m$. It follows directly that the principal inertia components of $\mathbf{P}$ are exactly the diagonal entries of $\mathbf{\Lambda}^2$, and if $\mathbf{P}_{Y|X}$ is positive-semidefinite then $\mathbf{P}$ is conforming.

□

The q-ary symmetric channel, defined below, is of particular interest to some of the results derived in the following sections.

Definition 6.2. The q-ary symmetric channel with crossover probability $\epsilon \leq 1 - q^{-1}$, also denoted as (ϵ, q)-SC, is defined as the channel with input X and output Y where $\mathcal{X} = \mathcal{Y} = [q]$ and

$$p_{Y|X}(y|x) = \begin{cases} 1 - \epsilon & \text{if } x = y \\ \frac{\epsilon}{q-1} & \text{if } x \neq y. \end{cases}$$

In the rest of this section, we assume that X and Y have a conforming joint distribution matrix with $\mathcal{X} = \mathcal{Y} = [q]$ and PICs $\lambda_k(X;Y) = \sigma_k^2$ for $k \in [d-1]$. The following lemma shows that conforming $\mathbf{P}$ can be transformed into the joint distribution of a q-ary symmetric channel with input distribution p_X by setting $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_{q-1}^2$, i.e. making all principal inertia components equal to the largest one.

Lemma 6.2. *Let* $\mathbf{P}$ *be a conforming joint distribution matrix of* X *and* Y, *with* X *and* Y *uniformly distributed,* $\mathcal{X} = \mathcal{Y} = [q]$, $\mathbf{P} = q^{-1}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ *and* $\mathbf{\Sigma} = \operatorname{diag}(1, \sigma_1, \ldots, \sigma_d)$. *For* $\widetilde{\mathbf{\Sigma}} = \operatorname{diag}(1, \sigma_1, \ldots, \sigma_1)$, *let* X *and* $\widetilde{Y}$ *have joint distribution* $\widetilde{\mathbf{P}} = q^{-1}\mathbf{U}\widetilde{\mathbf{\Sigma}}\mathbf{U}^T$. *Then,* $\widetilde{Y}$ *is the result of passing* X *through a* (ϵ, q)*-SC, with*

$$\epsilon = \frac{(q-1)(1 - \rho_m(X;Y))}{q}. \tag{6.4}$$

Proof. The first column of $\mathbf{U}$ is $\mathbf{p}_X^{1/2}$ and, since X is uniformly distributed, $\mathbf{p}_X^{1/2} = q^{-1/2}\mathbf{1}$. Therefore

$$\begin{aligned} \widetilde{\mathbf{P}} &= q^{-1}\mathbf{U}\widetilde{\mathbf{\Sigma}}\mathbf{U}^T \\ &= q^{-1}\sigma_1\mathbf{I} + q^{-2}(1 - \sigma_1)\mathbf{1}\mathbf{1}^T. \end{aligned} \tag{6.5}$$

Consequently, $\widetilde{\mathbf{P}}$ has diagonal entries equal to $(1+(q-1)\sigma_1)/q^2$ and all other entries equal to $(1-\sigma_1)/q^2$. The result follows by noting that $\sigma_1 = \rho_m(X;Y)$. □

Remark 6.2. For X, Y and $\widetilde{Y}$ given in the previous lemma, a natural question that arises is whether Y is a degraded version of $\widetilde{Y}$, i.e. $X \to \widetilde{Y} \to Y$. Unfortunately, this is **not true** in general, since the matrix $\mathbf{U}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}\mathbf{U}^{\mathbf{T}}$ does not necessarily contain only positive entries, although it is doubly-stochastic. However, since the PICs of X and $\widetilde{Y}$ upper bound the PICs of X and Y, it is natural to expect that, at least in some sense, $\widetilde{Y}$ is more informative about X than Y. This intuition is indeed correct for certain estimation problems where a one-bit function of X is to be inferred from a single observation Y or $\widetilde{Y}$, and will be investigated in the next section. In addition, using the characterization of the PICs in Theorem 5.1, it follows that *any* function of X can be inffered with smaller MMSE from $\widetilde{Y}$ than from Y.

6.4 One-bit Functions and Channel Transformations

Let $B \to X \to Y$, where B is a binary random variable. When X and Y have a conforming probability distribution, the PICs of X and Y have a particularly interesting interpretation: they can be understood as the filter coefficients in the linear transformation of $p_{B|X}$ into $p_{B|Y}$. In order to see why this is the case, consider the joint distribution of B and Y, denoted here by $\mathbf{Q}$, given by

$$\mathbf{Q} = [\mathbf{f} \;\; 1-\mathbf{f}]^T\mathbf{P} = [\mathbf{f} \;\; 1-\mathbf{f}]^T\mathbf{P}_{X|Y}\mathbf{D}_Y = [\mathbf{g} \;\; 1-\mathbf{g}]^T\mathbf{D}_Y, \tag{6.6}$$

where $\mathbf{f} \in \mathbb{R}^m$ and $\mathbf{g} \in \mathbb{R}^n$ are column-vectors with $\mathbf{f}_i = p_{B|X}(0|i)$ and $\mathbf{g}_j = p_{B|Y}(0|j)$. In particular, if B is a deterministic function of X, $\mathbf{f} \in \{0,1\}^m$.

If $\mathbf{P}$ is conforming and $\mathcal{X} = \mathcal{Y} = [m]$, then $\mathbf{P} = \mathbf{D}^{1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T\mathbf{D}^{1/2}$, where $\mathbf{D} = \mathbf{D}_X = \mathbf{D}_Y$. Assuming $\mathbf{D}$ fixed, the joint distribution $\mathbf{Q}$ is entirely specified by the linear transformation of $\mathbf{f}$ into $\mathbf{g}$. Denoting $\mathbf{T} \triangleq \mathbf{U}^T\mathbf{D}^{1/2}$, this transformation is done in three steps:

1. (Linear transform) $\widehat{\mathbf{f}} \triangleq \mathbf{T}\mathbf{f}$,
2. (Filter) $\widehat{\mathbf{g}} \triangleq \mathbf{\Sigma}\widehat{\mathbf{f}}$, where the diagonal of $\mathbf{\Sigma}^2$ are the PICs of X and Y,
3. (Inverse transform) $\mathbf{g} = \mathbf{T}^{-1}\widehat{\mathbf{g}}$.

Note that $\widehat{\mathbf{f}}_1 = \widehat{\mathbf{g}}_1 = 1 - \mathbb{E}[B]$ and $\widehat{\mathbf{g}} = \mathbf{T}\mathbf{g}$. Consequently, the PICs of X and Y bear an interpretation as the filter coefficients in the linear transformation of $p_{B|X}(0|\cdot)$ into $p_{B|Y}(0|\cdot)$.

A similar interpretation can be made for symmetric channels, where $\mathbf{P}_{Y|X} = \mathbf{P}_{Y|X}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ and $\mathbf{P}_{Y|X}$ acts as the matrix of the linear transformation of $\mathbf{p}_X$ into $\mathbf{p}_Y$. Note that $\mathbf{p}_Y = \mathbf{P}_{Y|X}\mathbf{p}_X$, and, consequently, $\mathbf{p}_X$ is transformed into $\mathbf{p}_Y$ in the same three steps as before:

1. (Linear transform) $\widehat{\mathbf{p}_X} = \mathbf{U}^T\mathbf{p}_X$,

2. (Filter) $\widehat{\mathbf{p}_Y} = \mathbf{\Lambda}\widehat{\mathbf{p}_X}$, where the diagonal of $\mathbf{\Lambda}^2$ are the PICs of X and Y in the particular case when X is uniformly distributed (Lemma 6.1),

3. (Inverse transform) $\mathbf{p}_Y = \mathbf{U}\widehat{\mathbf{p}_Y}$.

From this perspective, the vector $\mathbf{z} = \mathbf{U}\mathbf{\Lambda}\mathbf{1}m^{-1/2}$ can be understood as a proxy for the "noise effect" of the channel. Note that $\sum_i \mathbf{z}_i = 1$. However, the entries of $\mathbf{z}$ are not necessarily positive, and $\mathbf{z}$ might not be a *de facto* probability distribution.

We now illustrate these ideas by investigating binary channels with additive noise in the next section, where $\mathbf{T}$ will correspond to the well-known Walsh-Hadamard transform matrix.

6.4.1 Example: Binary Additive Noise Channels

In this example, let $\mathcal{X}^n, \mathcal{Y}^n \subseteq \{-1,1\}^n$ be the support sets of X^n and Y^n, respectively. We define two sets of channels that transform X^n into Y^n. In each set definition, we assume the conditions for $p_{Y^n|X^n}$ to be a valid probability distribution (i.e. non-negativity and unit sum).

Definition 6.3. The set of *parity-changing channels* of block-length n, denoted by $\mathcal{A}_n$, is defined as:

$$\mathcal{A}_n \triangleq \left\{p_{Y^n|X^n} \mid \forall \mathcal{S} \subseteq [n],\ \exists c_{\mathcal{S}} \in [-1,1] \text{ s.t. } \mathbb{E}\left[\chi_{\mathcal{S}}(Y^n)|X^n\right] = c_{\mathcal{S}}\chi_{\mathcal{S}}(X^n)\right\}. \tag{6.7}$$

The set of all *binary additive noise channels* is given by

$$\mathcal{B}_n \triangleq \left\{p_{Y^n|X^n} \mid \exists Z^n \text{ s.t. } Y^n = X^n \oplus Z^n,\ \text{supp}(Z^n) \subseteq \{-1,1\}^n, Z^n \perp\!\!\!\perp X^n\right\}.$$

The definition of parity-changing channels is inspired by results from the literature on Fourier analysis of Boolean functions. As in Chapter 3, for an overview of the topic we refer the reader to the survey [63]. The set of binary additive noise channels, in turn, is widely used in the information theory literature. The following theorem shows that both characterizations are equivalent.

Theorem 6.1. $\mathcal{A}_n = \mathcal{B}_n$.

Proof. Let $Y^n = X^n \oplus Z^n$ for some Z^n distributed over $\{-1,1\}^n$ and independent of X^n. Thus

$$\begin{aligned}\mathbb{E}\left[\chi_{\mathcal{S}}(Y^n)|X^n\right] &= \mathbb{E}\left[\chi_{\mathcal{S}}(Z^n \oplus X^n) \mid X^n\right] \\ &= \mathbb{E}\left[\chi_{\mathcal{S}}(X^n)\chi_{\mathcal{S}}(Z^n) \mid X^n\right] \\ &= \chi_{\mathcal{S}}(X^n)\mathbb{E}\left[\chi_{\mathcal{S}}(Z^n)\right],\end{aligned}$$

where the last equality follows from the assumption that $X^n \perp\!\!\!\perp Z^n$. By letting $c_{\mathcal{S}} = \mathbb{E}\left[\chi_{\mathcal{S}}(Z^n)\right]$, it follows that $p_{Y^n|X^n} \in \mathcal{A}_n$ and, consequently, $\mathcal{B}_n \subseteq \mathcal{A}_n$.

Now let y_n be fixed and $\delta_{y^n} : \{-1,1\}^n \to \{0,1\}$ be given by

$$\delta_{y^n}(x^n) = \begin{cases} 1, & x^n = y^n, \\ 0, & \text{otherwise.} \end{cases}$$

Since the function δ_{y^n} has Boolean inputs, it can be expressed in terms of its Fourier expansion [63, Prop. 1.1] as

$$\delta_{y^n}(x^n) = \sum_{\mathcal{S}\subseteq[n]} \widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(x^n).$$

Now let $p_{Y^n|X^n} \in \mathcal{A}_n$. Observe that $p_{Y^n|X^n}(y^n|x^n) = \mathbb{E}\left[\delta_{y^n}(Y^n) \mid X^n = x^n\right]$ and, for $z^n \in \{-1,1\}^n$,

$$\begin{aligned}
p_{Y^n|X^n}(y^n \oplus z^n | x^n \oplus z^n) &= \mathbb{E}\left[\delta_{y^n \oplus z^n}(Y^n) \mid X^n = x^n \oplus z^n\right] \\
&= \mathbb{E}\left[\delta_{y^n}(Y^n \oplus z^n) \mid X^n = x^n \oplus z^n\right] \\
&= \mathbb{E}\left[\sum_{\mathcal{S}\subseteq[n]} \widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(Y^n \oplus z^n) \mid X^n = x^n \oplus z^n\right] \\
&= \mathbb{E}\left[\sum_{\mathcal{S}\subseteq[n]} \widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(Y^n)\chi_{\mathcal{S}}(z^n) \mid X^n = x^n \oplus z^n\right] \\
&\overset{(a)}{=} \sum_{\mathcal{S}\subseteq[n]} c_{\mathcal{S}}\widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(x^n \oplus z^n)\chi_{\mathcal{S}}(z^n) \\
&= \sum_{\mathcal{S}\subseteq[n]} c_{\mathcal{S}}\widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(x^n) \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{\mathcal{S}\subseteq[n]} \widehat{d}_{\mathcal{S}}\chi_{\mathcal{S}}(Y^n) | X^n = x^n\right] \\
&= \mathbb{E}\left[\delta_{y^n}(Y^n) \mid X^n = x^n\right] \\
&= p_{Y^n|X^n}(y^n|x^n).
\end{aligned}$$

Equalities (a) and (b) follow from the definition of $\mathcal{A}_n$. By defining the distribution of Z^n as $p_{Z^n}(z^n) \triangleq p_{Y^n|X^n}(z^n|\mathbf{1}^n)$, where $\mathbf{1}^n$ is the vector with all entries equal to 1, it follows that $Z^n = X^n \oplus Y^n$, $Z^n \perp\!\!\!\perp X^n$ and $p_{Y^n|X^n} \subseteq \mathcal{B}_n$.

□

The previous theorem suggests that there is a correspondence between the coefficients $c_{\mathcal{S}}$ in (6.7) and the distribution of the additive noise Z^n in the definition of $\mathcal{B}_n$. The next result shows that this is indeed the case and, when X^n is uniformly distributed, the coefficients $c_{\mathcal{S}}^2$ correspond to the principal inertia components between X^n and Y^n.

Theorem 6.2. *Let $p_{Y^n|X^n} \in \mathcal{B}_n$, and $X^n \sim p_{X^n}$. Then $\mathbf{P}_{X^n,Y^n} = \mathbf{D}_{X^n}\mathbf{H}_{2^n}\mathbf{\Lambda}\mathbf{H}_{2^n}$, where $\mathbf{H}_l$ is the $l \times l$ normalized Hadamard matrix (i.e. $\mathbf{H}_l^2 = \mathbf{I}$). Furthermore, for $Z^n \sim p_{Z^n}$, $\operatorname{diag}(\mathbf{\Lambda}) = 2^{n/2}\mathbf{H}_{2^n}\mathbf{p}_{Z^n}$, and the diagonal entries of $\mathbf{\Lambda}$ are equal to $c_{\mathcal{S}}$ in (6.7). Finally, if X is uniformly distributed, then $c_{\mathcal{S}}^2$ are the principal inertia components of X^n and Y^n.*

Proof. Let $p_{Y^n|X^n} \in \mathcal{A}_n$ be given. From Theorem 6.1 and the definition of $\mathcal{A}_n$, it follows that $\chi_{\mathcal{S}}(Y^n)$ is a right eigenvector of $p_{Y^n|X^n}$ with corresponding eigenvalue $c_{\mathcal{S}}$. Since $\chi_{\mathcal{S}}(Y^n)2^{-n/2}$ corresponds to a row of $\mathbf{H}_{2^n}$ for each $\mathcal{S}$ (due to the Kronecker product construction of the Hadamard matrix) and $\mathbf{H}_{2^n}^2 = \mathbf{I}$, then $\mathbf{P}_{X^n,Y^n} = \mathbf{D}_{X^n}\mathbf{H}_{2^n}\mathbf{\Lambda}\mathbf{H}_{2^n}$. Finally, note that $\mathbf{p}_Z^T = 2^{-n/2}\mathbf{1}^T\mathbf{\Lambda}\mathbf{H}_{2^n}$. From Lemma 6.1, it follows that $c_{\mathcal{S}}^2$ are the principal inertia components of X^n and Y^n if X^n is uniformly distributed. □

Remark 6.3. Theorem 6.2 indicates that one possible method for estimating the distribution of the additive binary noise Z^n is to estimate its effect on the parity bits of X^n and Y^n. In this case, we are estimating the coefficients $c_{\mathcal{S}}$ of the Walsh-Hadamard transform of p_{Z^n}. This approach was studied by Raginsky *et al.* in [99].

Theorem 6.2 illustrates the filtering role of the principal inertia components, discussed in the beginning of this section. If X^n is uniform, and using the same notation as in (6.6), then the vector of conditional probabilities $\mathbf{f}$ is transformed into the vector of *a posteriori* probabilities $\mathbf{g}$ by: (i) taking the Hadamard transform of $\mathbf{f}$, (ii) filtering the transformed vector according to the coefficients $c_{\mathcal{S}}$, where $\mathcal{S} \in [n]$, and (iii) taking the inverse Hadamard transform. The same rationale applies to the transformation of $\mathbf{p}_X$ into $\mathbf{p}_Y$ in binary additive channels.

6.4.2 Quantifying the Information of a Boolean Function of the Input of a Noisy Channel

We now investigate the connection between the principal inertia components and f-information in the context of one-bit functions of X. Recall from the discussion in the beginning of this section and, in particular, equation (6.6), that for a binary B and $B \to X \to Y$, the distribution of B and Y is entirely specified by the transformation of $\mathbf{f}$ into $\mathbf{g}$, where $\mathbf{f}$ and $\mathbf{g}$ are vectors with entries equal to $p_{B|X}(0|\cdot)$ and $p_{B|Y}(0|\cdot)$, respectively.

For $\mathbb{E}[B] = 1 - a$, the f-information between B and Y is given by[2]

$$I_f(B;Y) = \mathbb{E}\left[af\left(\frac{\mathbf{g}_Y}{a}\right) + (1-a)f\left(\frac{1-\mathbf{g}_Y}{1-a}\right)\right].$$

For $0 \leq r, s \leq 1$, we can expand $f\left(\frac{r}{s}\right)$ around 1 as

$$f\left(\frac{r}{s}\right) = \sum_{k=1}^{\infty} \frac{f^{(k)}(1)}{k!}\left(\frac{r-s}{r}\right)^k.$$

[2]Note that here we assume that $\mathcal{Y} = [n]$, so there is no ambiguity in indexing $p_{B|Y}(0|Y)$ by $\mathbf{g}_Y$.

Denoting

$$c_k(\alpha) \triangleq \frac{1}{a^{k-1}} + \frac{(-1)^k}{(1-a)^{k-1}},$$

the f-information can then be expressed as

$$I_f(B;Y) = \sum_{k=2}^{\infty} \frac{f^{(k)}(1)c_k(a)}{k!} \mathbb{E}\left[(\mathbf{g}_Y - a)^k\right]. \tag{6.8}$$

Similarly to [96, Chapter 4], for a fixed $\mathbb{E}[B] = 1-a$, maximizing the principal inertia components between X and Y will always maximize the first term in the expansion (6.8). To see why this is the case, observe that

$$\begin{aligned}\mathbb{E}\left[(\mathbf{g}_Y - a)^2\right] &= (\mathbf{g}-a)^T \mathbf{D}_Y (\mathbf{g}-a) \\ &= \mathbf{g}^T \mathbf{D}_Y \mathbf{g} - a^2 \\ &= \mathbf{f}^T \mathbf{D}_X^{1/2} \mathbf{U} \mathbf{\Sigma}^2 \mathbf{U}^T \mathbf{D}_X^{1/2} \mathbf{f} - a^2. \end{aligned} \tag{6.9}$$

For a fixed a and any $\mathbf{f}$ such that $\mathbf{f}^T\mathbf{1} = a$, (6.9) is non-decreasing in the diagonal entries of $\mathbf{\Sigma}^2$ which, in turn, are exactly the principal inertia components of X and Y. Equivalently, (6.9) is non-decreasing in the χ^2-divergence between $p_{X,Y}$ and $p_X p_Y$.

However, we do note that increasing the principal inertia components **does not** increase the f-information between B and Y in general. Indeed, for a fixed $\mathbf{U}$, $\mathbf{V}$ and marginal distributions of X and Y, increasing the principal inertia components might not even lead to a valid probability distribution matrix $\mathbf{P}$.

Nevertheless, if $\mathbf{P}$ is conforming and X and Y are uniformly distributed over $[q]$, as shown in Lemma (6.2), by increasing the principal inertia components we can define a new random variable $\widetilde{Y}$ that results from sending X through a (ϵ, q)-SC, where ϵ is given in (6.4). In this case, the f-information between B and Y has a simple expression when B is a function of X.

Lemma 6.3. *Let $B \to X \to \widetilde{Y}$, where $B = h(X)$ for some $h : [q] \to \{0,1\}$, $\mathbb{E}[B] = 1-a$ where aq is an integer, X is uniformly distributed in $[q]$ and $\widetilde{Y}$ is the result of passing X through a (ϵ, q)-SC with $\epsilon \leq (q-1)/q$. Then*

$$I_f(B;\widetilde{Y}) = a^2 f(1+\sigma_1 c) + 2a(1-a) f(1-\sigma_1) + (1-a)^2 f\left(1+\sigma_1 c^{-1}\right) \tag{6.10}$$

where $\sigma_1 = \rho_m(X;\widetilde{Y}) = 1 - \epsilon q (q-1)^{-1}$ and $c \triangleq (1-a)a^{-1}$. In particular, for $f(x) = x\log x$, then $I_f(X;\widetilde{Y}) = I(X;\widetilde{Y})$, and for $\sigma_1 = 1-2\delta$

$$\begin{aligned} I(B;\widetilde{Y}) &= h_b(a) - \alpha H_b(2\delta(1-a)) - (1-a) H_b(2\delta a) && (6.11)\\ &\leq 1 - H_b(\delta), && (6.12)\end{aligned}$$

where $H_b(x) \triangleq -x\log(x) - (1-x)\log(1-x)$ *is the binary entropy function.*

Proof. Since B is a deterministic function of X and aq is an integer, $\mathbf{f}$ is a vector with aq entries equal to 1 and $(1-a)q$ entries equal to 0. It follows from (6.5) that

$$\begin{aligned} I_f(B;\widetilde{Y}) =& \frac{1}{q}\sum_{i=1}^{q} af\left(\frac{(1-\sigma_1)a + \mathbf{f}_i\sigma_1}{a}\right) + (1-a)f\left(\frac{1-(1-\sigma_1)a - \mathbf{f}_i\sigma_i}{1-a}\right) \\ =& a^2 f\left(1+\sigma_1\frac{1-a}{a}\right) + 2a(1-a)f\left(1-\sigma_1\right) + (1-a)^2 f\left(1+\sigma_1\frac{a}{1-a}\right). \end{aligned}$$

Letting $f(x) = x\log x$, (6.11) follows immediately. Since (6.11) is concave in a and symmetric around $a = 1/2$, it is maximized at $a = 1/2$, resulting in (6.12). □

6.4.3 On the "Most Informative Bit" Conjecture

We now return to channels with additive binary noise, analyzed is Section 6.4.1. Let X^n be a uniformly distributed binary string of length n ($\mathcal{X} = \{-1, 1\}$) and Y^n be the result of passing X^n through a memoryless binary symmetric channel with crossover probability $\delta \leq 1/2$. Kumar and Courtade conjectured [24] that for all binary B and $B \to X^n \to Y^n$ we have

$$I(B;Y^n) \leq 1 - H_b(\delta). \quad \text{(conjecture)} \tag{6.13}$$

It is sufficient to consider B a function of X^n, denoted by $B = h(X^n)$, $h : \{-1,1\}^n \to \{0,1\}$, and we make this assumption henceforth.

From the discussion in Section 6.4.1, for the memoryless binary symmetric channel $Y^n = X^n \oplus Z^n$, where Z^n is an i.i.d. string with $\Pr\{Z_i = 1\} = 1-\delta$, and any $\mathcal{S} \in [n]$,

$$\begin{aligned} \mathbb{E}\left[\chi_{\mathcal{S}}(Y^n)|X^n\right] &= \chi_{\mathcal{S}}(X^n)\left(\Pr\left\{\chi_{\mathcal{S}}(Z^n) = 1\right\} - \Pr\left\{\chi_{\mathcal{S}}(Z^n) = -1\right\}\right) \\ &= \chi_{\mathcal{S}}(X^n)\left(2\Pr\left\{\chi_{\mathcal{S}}(Z^n) = 1\right\} - 1\right) \\ &= \chi_{\mathcal{S}}(X^n)(1-2\delta)^{|\mathcal{S}|}. \end{aligned}$$

It follows directly that $c_{\mathcal{S}} = (1-2\delta)^{|\mathcal{S}|}$ for all $\mathcal{S} \subseteq [n]$. Consequently, from Theorem 6.2, the principal inertia components of X^n and Y^n are of the form $(1-2\delta)^{2|\mathcal{S}|}$ for some $\mathcal{S} \subseteq [n]$. Observe that the principal inertia components act as a low pass filter on the vector of conditional probabilities $\mathbf{f}$ given in (6.6).

Can the noise distribution be modified so that the principal inertia components act as an all-pass filter? More specifically, what happens when $\widetilde{Y}^n = X^n \oplus W^n$, where W^n is such that the principal inertia components between X^n and $\widetilde{Y}^n$ satisfy $\sigma_i = 1 - 2\delta$? Then, from Lemma 6.2, $\widetilde{Y}^n$ is the result of sending X^n through a $(\epsilon, 2^n)$-SC with $\epsilon = 2\delta(1-2^{-n})$. Therefore, from (6.12),

$$I(B;\widetilde{Y}^n) \leq 1 - H_b(\delta).$$

For any function $h : \{-1,1\}^n \to \{0,1\}$ such that $B = h(X^n)$, from standard results in Fourier analysis of Boolean functions [63, Prop. 1.1], $h(X^n)$ can be expanded as

$$h(X^n) = \sum_{\mathcal{S} \subseteq [n]} \hat{h}_{\mathcal{S}} \chi_{\mathcal{S}}(X^n).$$

The value of B is uniquely determined by the action of h on $\chi_{\mathcal{S}}(X^n)$. Consequently, for a fixed function h, one could expect that $\widetilde{Y}^n$ should be more informative about B than Y^n, since the parity bits $\chi_{\mathcal{S}}(X^n)$ are more reliably estimated from $\widetilde{Y}^n$ than from Y^n. Indeed, the memoryless binary symmetric channel attenuates $\chi_{\mathcal{S}}(X^n)$ exponentially in $|\mathcal{S}|$, acting (as argued previously) as a low-pass filter. In addition, if one could prove that for any fixed h the inequality $I(B;Y^n) \leq I(B;\widetilde{Y}^n)$ holds, then (6.13) would be proven true. This motivates the following conjecture.

Conjecture 6.1. *For all* $h : \{-1,1\}^n \to \{0,1\}$ *and* $B = h(X^n)$

$$I(B;Y^n) \leq I(B;\widetilde{Y}^n).$$

We note that Conjecture 6.1 can be false if B is not a deterministic function of X^n. In the next section, we provide further evidence for this conjecture by investigating information metrics between B and an estimate $\widehat{B}$ derived from Y^n.

6.5 One-bit Estimators

Let $B \to X \to Y \to \widehat{B}$, where B and $\widehat{B}$ are binary random variables with $\mathbb{E}[B] = 1-a$ and $\mathbb{E}[\widehat{B}] = 1-b$. We denote by $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ the column vectors with entries $\mathbf{x}_i = p_{B|X}(0|i)$ and $\mathbf{y}_j = p_{\widehat{B}|Y}(0|j)$. The joint distribution matrix of B and $\widehat{B}$ is given by

$$\mathbf{P}_{B,\widehat{B}} = \begin{pmatrix} z & a-z \\ b-z & 1-a-b+z \end{pmatrix}, \tag{6.14}$$

where $z = \mathbf{x}^T \mathbf{P} \mathbf{y} = \Pr\{B = \widehat{B} = 0\}$. For fixed values of a and b, the joint distribution of B and $\widehat{B}$ only depends on z.

Let $f : \mathcal{P}_{2\times 2} \to \mathbb{R}$, and, with a slight abuse of notation, we also denote f as a function of the entries of the 2×2 matrix as $f(a,b,z)$. If f is convex in z for a fixed a and b, then f is maximized at one of the extreme values of z. Examples of such functions f include mutual information and expected error probability. Therefore, characterizing the maximum and minimum values of z is equivalent to characterizing the maximum value of f over all possible mappings $X \to B$ and $Y \to \widehat{B}$. This leads to the following definition.

Definition 6.4. For a fixed $\mathbf{P}$ and given $\mathbb{E}[B] = 1-a$ and $\mathbb{E}[\widehat{B}] = 1-b$, the minimum and

maximum values of z over all possible mappings $X \to B$ and $Y \to \widehat{B}$ are defined as

$$z_l^*(a,b,\mathbf{P}) \triangleq \min_{\substack{\mathbf{x}\in\mathcal{C}^m(a,\mathbf{P}^T)\\ \mathbf{y}\in\mathcal{C}^n(b,\mathbf{P})}} \mathbf{x}^T\mathbf{P}\mathbf{y} \quad \text{and} \quad z_u^*(a,b,\mathbf{P}) \triangleq \max_{\substack{\mathbf{x}\in\mathcal{C}^m(a,\mathbf{P}^T)\\ \mathbf{y}\in\mathcal{C}^n(b,\mathbf{P})}} \mathbf{x}^T\mathbf{P}\mathbf{y},$$

respectively, and $\mathcal{C}^n(a,\mathbf{P})$ is defined in (6.1).

The next lemma provides a simple upper-bound for $z_u^*(a,b,\mathbf{P})$ in terms of the largest principal inertia components or, equivalently, the maximal correlation between X and Y.

Lemma 6.4. $z_u^*(a,b,\mathbf{P}) \le ab + \rho_m(X;Y)\sqrt{a(1-a)b(1-b)}$.

Remark 6.4. An analogous result was derived by Witsenhausen [76, Thm. 2] for bounding the probability of agreement of a common bit derived from two correlated sources.

Proof. Let $\mathbf{x} \in \mathcal{C}^m(a,\mathbf{P}^T)$ and $\mathbf{y} \in \mathcal{C}^n(b,\mathbf{P})$. Then, for $\mathbf{P}$ decomposed as $\mathbf{P} = \mathbf{D}_X^{1/2}\mathbf{Q}\mathbf{D}_Y^{1/2}$ where $\mathbf{Q}$ given in (5.3) and denoting $\mathbf{\Sigma}^- = \operatorname{diag}(0,\sigma_1,\ldots,\sigma_d)$,

$$\begin{aligned}\mathbf{x}^T\mathbf{P}\mathbf{y} &= ab + \mathbf{x}^T\mathbf{D}_X^{1/2}\mathbf{U}\mathbf{\Sigma}^-\mathbf{V}^T\mathbf{D}_Y^{1/2}\mathbf{y}\\ &= ab + \hat{\mathbf{x}}^T\mathbf{\Sigma}^-\hat{\mathbf{y}},\end{aligned} \tag{6.15}$$

where $\hat{\mathbf{x}} \triangleq \mathbf{U}^T\mathbf{D}_X^{1/2}\mathbf{x}$ and $\hat{\mathbf{y}} \triangleq \mathbf{V}^T\mathbf{D}_Y^{1/2}\mathbf{y}$. Since $\hat{\mathbf{x}}_1 = \|\hat{\mathbf{x}}\|_2 = a$ and $\hat{\mathbf{y}}_1 = \|\hat{\mathbf{y}}\|_2 = b$, then

$$\begin{aligned}\hat{\mathbf{x}}^T\mathbf{\Sigma}^-\hat{\mathbf{y}} &= \sum_{i=2}^{d+1}\sigma_{i-1}\hat{\mathbf{x}}_i\hat{\mathbf{y}}_i\\ &\le \sigma_1\sqrt{\left(\|\hat{\mathbf{x}}\|_2^2-\hat{\mathbf{x}}_1^2\right)\left(\|\hat{\mathbf{y}}\|_2^2-\hat{\mathbf{y}}_1^2\right)}\\ &= \sigma_1\sqrt{(a-a^2)(b-b^2)}.\end{aligned}$$

The result follows by noting that $\sigma_1 = \rho_m(X;Y)$. □

We will focus in the rest of this section on functions and corresponding estimators that are (i) unbiased ($a = b$) and (ii) satisfy $z = \Pr\{\hat{B} = B = 0\} \ge a^2$. The set of all such mappings is given by

$$\mathcal{H}(a,\mathbf{P}) \triangleq \left\{(\mathbf{x},\mathbf{y}) \mid \mathbf{x} \in \mathcal{C}^m(a,\mathbf{P}^T), \mathbf{y} \in \mathcal{C}^n(a,\mathbf{P}), \mathbf{x}^T\mathbf{P}\mathbf{y} \ge a^2\right\}.$$

The next results provide upper and lower bounds for z for the mappings in $\mathcal{H}(a,\mathbf{P})$.

Lemma 6.5. *Let $0 \le a \le 1/2$ and $\mathbf{P}$ be fixed. For any $(\mathbf{x},\mathbf{y}) \in \mathcal{H}(a,\mathbf{P})$*

$$a^2 \le z \le a^2 + \rho_m(X;Y)a(1-a), \tag{6.16}$$

where $z = \mathbf{x}^T\mathbf{P}\mathbf{y}$.

Proof. The lower bound for z follows directly from the definition of $\mathcal{H}(a, \mathbf{P})$, and the upper bound follows from Lemma 6.4. □

The previous lemma allows us to provide an upper bound over the mappings in $\mathcal{H}(a, \mathbf{P})$ for the f-information between B and $\widehat{B}$ when I_f is non-negative.

Theorem 6.3. *For any non-negative I_f and fixed a and $\mathbf{P}$,*

$$\sup_{(\mathbf{x},\mathbf{y})\in\mathcal{H}(a,\mathbf{P})} I_f(B;\hat{B}) \leq a^2 f\left(1+\sigma_1 c\right) + 2a(1-a) f\left(1-\sigma_1\right) + (1-a)^2 f\left(1+\sigma_1 c^{-1}\right) \quad (6.17)$$

where here $\sigma_1 = \rho_m(X;\widetilde{Y})$ and $c \triangleq (1-a)a^{-1}$. In particular, for $a = 1/2$,

$$\sup_{(\mathbf{x},\mathbf{y})\in\mathcal{H}(1/2,\mathbf{P})} I_f(B;\hat{B}) \leq \frac{1}{2}\left(f(1-\sigma_1) + f(1+\sigma_1)\right). \quad (6.18)$$

Proof. Using the matrix form of the joint distribution between B and $\widehat{B}$ given in (6.14), for $\mathbb{E}[B] = \mathbb{E}\left[\widehat{B}\right] = 1-a$, the f information is given by

$$I_f(B;\hat{B}) = a^2 f\left(\frac{z}{a^2}\right) + 2a(1-a) f\left(\frac{a-z}{a(1-a)}\right) + (1-a)^2 f\left(\frac{1-2a+z}{(1-a)^2}\right). \quad (6.19)$$

Consequently, (6.19) is convex in z. For $(\mathbf{x},\mathbf{y}) \in \mathcal{H}(a,\mathbf{P})$, it follows from Lemma 6.5 that z is restricted to the interval in (6.16). Since $I_f(B;\hat{B})$ is non-negative by assumption, $I_f(B;\hat{B}) = 0$ for $z = a^2$ and (6.19) is convex in z, then $I_f(B;\hat{B})$ is non-decreasing in z for z in (6.16). Substituting $z = a^2 + \rho_m(X;Y)a(1-a)$ in (6.19), inequality (6.17) follows. □

Remark 6.5. Note that the right-hand side of (6.17) matches the right-hand side of (6.10), and provides further evidence for Conjecture 6.1. This result indicates that, for conforming probability distributions, the information between a binary function and its corresponding unbiased estimate is maximized when all the principal inertia components have the same value.

Following the same approach from Lemma 6.3, we find the next bound for the mutual information between B and $\widehat{B}$.

Corollary 6.1. *For $0 \leq a \leq 1$ and $\rho_m(X;Y) = 1-2\delta$*

$$\sup_{(p_{B|X},p_{\widehat{B}|Y})\in\mathcal{H}(a,\mathbf{P})} I(B;\hat{B}) \leq 1 - H_b(\delta).$$

We now provide a few application examples for the results derived in this section.

6.5.1 Lower Bounding the Estimation Error Probability

For z given in (6.14), the average estimation error probability is given by $\Pr\{B \neq \widehat{B}\} = a + b - 2z$, which is a convex (linear) function of z. If a and b are fixed, then the error

probability is minimized when z is maximized. Therefore

$$\Pr\{B \neq \widehat{B}\} \geq a + b - 2z_u^*(a, b).$$

Using the bound from Lemma 6.4, it follows that

$$\Pr\{B \neq \widehat{B}\} \geq a + b - 2ab - 2\rho_m(X;Y)\sqrt{a(1-a)b(1-b)}. \tag{6.20}$$

The bound (6.20) is exactly the bound derived by Witsenhausen in [76, Thm 2.]. Furthermore, minimizing the right-hand side of (6.20) over $0 \leq b \leq 1/2$, we arrive at

$$\Pr\{B \neq \widehat{B}\} \geq \frac{1}{2}\left(1 - \sqrt{1 - 4a(1-a)(1-\rho_m(X;Y)^2)}\right), \tag{6.21}$$

which is a particular form of the bound derived in Theorem 5.4.

6.5.2 Memoryless Binary Symmetric Channels with Uniform Inputs

We now turn our attention back to the setting considered in Section 6.4.1. Let Y^n be the result of passing X^n through a memoryless binary symmetric channel with crossover probability δ, X^n uniformly distributed, and $B \to X^n \to Y^n \to \widehat{B}$. Then $\rho_m(X^n;Y^n) = 1 - 2\delta$ and, from (6.21), when $\mathbb{E}[B] = 1/2$,

$$\Pr\{B \neq \widehat{B}\} \geq \delta.$$

Consequently, inferring any unbiased one-bit function of the input of a binary symmetric channel is at least as hard (in terms of error probability) as inferring a single output from a single input.

Using the result from Corollary 6.1, it follows that when $\mathbb{E}[B] = \mathbb{E}\left[\widehat{B}\right] = a$ and $\Pr\{B = \widehat{B} = 0\} \geq a^2$, then

$$I(B;\widehat{B}) \leq 1 - H_b(\delta). \tag{6.22}$$

Remark 6.6. Anantharam *et al.* presented in [95] a computer aided proof that the upper bound (6.22) holds for any $B \to X^n \to Y^n \to \widehat{B}$. Nevertheless, we highlight that the methods introduced here allowed an analytical derivation of (6.22) for unbiased estimators.

Chapter 7

Applications of the Principal Inertia Components to Security and Privacy

7.1 Overview

In this chapter, we present a few applications of the principal inertia components to problems in security and privacy. We adopt the privacy against statistical inference framework presented in [18] and discussed in Section 1.4 with the mutual information utility function. This setup, called the *Privacy Funnel*, was introduced in [20]. Consider two communicating parties, namely Alice and Bob. Alice's goal is to disclose to Bob information about a set of measurement points, represented by the random variable X. Alice discloses this information in order to receive some utility from Bob. Simultaneously, Alice wishes to limit the amount of information revealed about a private random variable S that is dependent on X. For example, X may represent Alice's movie ratings, released to Bob in order to receive movie recommendations, whereas S may represent Alice's political preference or yearly income. Bob is honest but curious, and will try to extract the maximum amount of information about S from the data disclosed by Alice.

Instead of revealing X directly to Bob, Alice releases a new random variable, denoted by Y. This random variable is produced from X through a random mapping $p_{Y|X}$, called the *privacy-assuring mapping*. We assume that $p_{S,X}$ is fixed and known by both Alice and Bob, and $S \to X \to Y$. Alice's goal is to find a mapping $p_{Y|X}$ that minimizes $I(S;Y)$, while guaranteeing that the information disclosed about X is above a certain threshold t, i.e. $I(X;Y) \geq t$. We refer to the quantity $I(S;Y)$ as the *disclosed private information*, and $I(X;Y)$ as the *disclosed useful information*. When $I(S;Y) = 0$, we say that *perfect privacy* is achieved, i.e. Y does not reveal any information about S. We consider here the non-interactive, one-shot regime, where Alice discloses information once, and no additional information is released. We also assume that Bob knows the privacy-assuring mapping $p_{Y|X}$ chosen by Alice, and no side information is available to Bob about S besides the value Y.

We present in this chapter necessary and sufficient conditions for achieving perfect privacy while disclosing a non-trivial amount of useful information when both S and X have finite support $\mathcal{S}$ and $\mathcal{X}$, respectively. We prove that the smallest PIC of $p_{S,X}$ plays a central role for achieving perfect privacy: If $|\mathcal{X}| \leq |\mathcal{S}|$, then perfect privacy is achievable with $I(X;Y) > 0$ if and only if the smallest PIC of $p_{S,X}$ is 0. Since $I(S;Y) = 0$ (perfect privacy) if and only if $S \perp\!\!\!\perp Y$, this fundamental result holds for any privacy metric where statistical independence implies perfect privacy. We also provide an explicit lower bound for the amount of useful information that can be released while guaranteeing perfect privacy, and demonstrate how to construct $p_{Y|X}$ in order to achieve this bound.

In addition, we derive general bounds for the minimum amount of disclosed private information $I(S;Y)$ given that, on average, at least t bits of useful information is revealed to Bob, i.e. $I(X;Y) \geq t$. These bounds are sharp, and delimit the achievable privacy-utility region for the considered setting. Adopting an analysis related to the information bottleneck [100] and for characterizing linear contraction coefficients in strong data processing inequalities in [77, 78], we determine the smallest achievable ratio between disclosed private and useful information, i.e. $\inf_{p_{Y|X}} I(S;Y)/I(X;Y)$. We prove that this value is upper-bounded by the smallest PIC, and is zero if and only if the smallest PIC is zero. In this case, we present an explicit construction of a privacy-assuring mapping that discloses a non-trivial amount of useful information while guaranteeing perfect privacy. We also introduce convex programs that can be used to design privacy-assuring mappings based on the PICs.

7.2 Main Contributions

This chapter focuses on applying the PICs to different problems in privacy and, in particular, to the privacy against statistical inference framework described above. Lemmas 7.1 to 7.2 prove different properties of the privacy funnel function. Theorems 7.1 and 7.2 characterize the best tradeoff between disclosed private and useful information in terms of the smallest PIC, leading to Theorem 7.3. Theorem 7.3, in turn, is the highlight of this chapter, stating that if the smallest PIC is zero, not only perfect privacy can be achieved, but an amount of useful information that is *strictly bounded away from zero* can be disclosed with perfect privacy. Some of these results also appear in [101].

We then return to the correlation-error product discussed at the end of Chapter 3 in Section 7.8, and demonstrate how the PICs shed light on the fundamental tradeoff between privacy and utility by decomposing the MMSE of a hidden variable. This analysis leads to Proposition 7.2, which presents a linear program for creating privacy-assuring mappings that provide privacy guarantees in terms of what an adversary can or cannot reliably infer from the disclosed useful information. Finally, we apply the results for the PICs for symmetric functions in a database privacy setup. We show in the last section that answering statistical queries over a randomly select subset of entries of a database is a simple, yet powerful tool

for providing privacy.

One of the main goals of this chapter is to characterize the fundamental limits of privacy. Consequently, the theoretical analysis presented here often assumes knowledge of the joint distribution $p_{S,X}$. However, we highlight that several of the results in sections 7.6 to 7.10 hold without complete knowledge of $p_{S,X}$. For example, the results in Section 7.6 are given in terms of the smallest PIC of $p_{S,X}$, and the subsampling method discussed in Section 7.10 requires only an independence assumption on the entries of a database.

7.2.1 Outline of the Chapter

The rest of the chapter is organized as follows. Section 7.4 introduces the privacy funnel and ancillary results. Section 7.5 relates the smallest achievable ratio between disclosed private and useful information with the principal inertia components. Section 7.6 presents a necessary and sufficient condition for achieving perfect privacy in terms of the smallest principal inertia component and the cardinality of $\mathcal{X}$. Section 7.7 presents an explicit threshold for the amount of useful information that can be disclosed with perfect privacy, and investigates the case where S and X are vectors of i.i.d. random variables.

We then revisit the correlation-error product results from Chapter 3 through the PIC lens in Section 7.8. Section 7.9 presents a convex program that can be used for finding privacy-assuring mappings given constraints that certain functions of the data can or cannot be reliably inferred from the data. Finally, Section 7.10 discusses applications of the PICs to privacy-preserving queries in statistical databases.

7.3 Related Work

Information-theoretic formulations for privacy have appeared in [58, 60, 61, 102, 103]. For an overview, we refer the reader to [18, 61] and the references therein. The privacy against statistical inference framework considered here was further studied in [13, 14]. The results presented in this chapter are closely connected to the study of hypercontractivity coefficients and strong data processing results, such as in [77–80, 104]. PIC-based analysis were used in the context of security in [56, 105]. Extremal properties of privacy were also investigated in [106, 107].

We note that the privacy against statistical inference setting is related to differential privacy [11, 12]. Assuming the classic differential privacy setting in centralized statistical databases, the private variable S can represent an individual user's entry to the database, and the variable X the output of a query over the database. Unlike in differential privacy, here we consider an additional distortion constraint, which can be chosen according to the application at hand. In the privacy funnel setting, the distortion constraint is given in terms of the mutual information between X and the perturbed query output Y.

7.4 The Privacy Funnel

We define next the privacy funnel function, which captures the smallest amount of disclosed private information for a given threshold on the amount of disclosed useful information. We then characterize properties of the privacy funnel function in the rest of this section.

Definition 7.1. For $0 \leq t \leq H(X)$ and a joint distribution $p_{S,X}$ over $\mathcal{S} \times \mathcal{X}$, we define the *privacy funnel function* $G_I(t, p_{S,X})$ as

$$G_I(t, p_{S,X}) \triangleq \inf \{I(S;Y) | I(X;Y) \geq t, S \to X \to Y\}, \tag{7.1}$$

where the infimum is over all mappings $p_{Y|X}$ such that $\mathcal{Y}$ is finite. For a fixed $p_{S,X}$ and $t \geq 0$, the set of pairs $\{(t, G_I(t, p_{S,X}))\}$ is called the *privacy region* of $p_{S,X}$.

7.4.1 Properties of the Privacy Funnel Function

We now enunciate a few useful properties of $G_I(t, p_{S,X})$ and the privacy region.

Lemma 7.1.

$$G_I(t, p_{S,X}) = \min_{p_{Y|X}} \{I(S;Y) | I(X;Y) \geq t, S \to X \to Y, \ |\mathcal{Y}| \leq |\mathcal{X}|+2\}. \tag{7.2}$$

Proof. Let $p_{S,X}$ and $p_{Y|X}$ be given, with $S \to X \to Y$ and $|\mathcal{Y}| > |\mathcal{X}|+2$. Denote by $\mathbf{w}_i$ the vector in the $|\mathcal{X}|$-simplex with entries $p_{X|Y}(\cdot|i)$. Furthermore, let $a_i \triangleq H(X) - H(X|Y=i)$, and $b_i \triangleq H(S) - H(S|Y=i)$. Therefore

$$\sum_{i=1}^{\mathcal{Y}} p_Y(i) [\mathbf{w}_i, a_i, b_i] = [\mathbf{p}_X, I(X;Y), I(S;Y)].$$

Since $\mathbf{w}_i$ belongs to the $|\mathcal{X}|$-simplex, the vector $[\mathbf{w}_i, a_i, b_i]$ is taken from a $|\mathcal{X}|+1$ dimensional space. Then, from Carathéodory's theorem, the point $[\mathbf{p}_X, I(X;Y), I(S;Y)]$ can also be achieved by at most $|\mathcal{X}|+2$ non-zero values of $p_Y(i)$. It follows directly that it is sufficient to consider $|\mathcal{Y}| \leq |\mathcal{X}|+2$ for the infimum (7.1).

The set of all mappings $p_{Y|X}$ for $|\mathcal{Y}| \leq |\mathcal{X}|+2$ is compact, and both $p_{Y|X} \to I(S;Y)$ and $p_{Y|X} \to I(X;Y)$ are continuous and bounded when S, X and Y have finite support. Consequently, the infimum in (7.1) is attainable. □

Lemma 7.2. *For a fixed $p_{S,X}$, the mapping $t \to \frac{G_I(t,p_{S,X})}{t}$ is non-decreasing.*

Proof. For $0 < t \leq H(X)$ and $p_{S,X}$ fixed, let $G_I(t, p_{S,X}) = \alpha$. From Lemma 7.1, there exists $p_{Y|X}$ that achieves $I(S;Y) = \alpha$ for $I(X;Y) \geq t$. Now consider $p_{\widetilde{Y}|X}$ where $\widetilde{\mathcal{Y}} = [|\mathcal{Y}|+1]$ and, for $0 < \lambda \leq 1$,

$$p_{\widetilde{Y}|X}(y|x) = (1-\lambda)\mathbf{1}_{\{y=|\mathcal{Y}|+1\}} + \lambda \mathbf{1}_{\{y \neq |\mathcal{Y}|+1\}} p_{Y|X}(y|x).$$

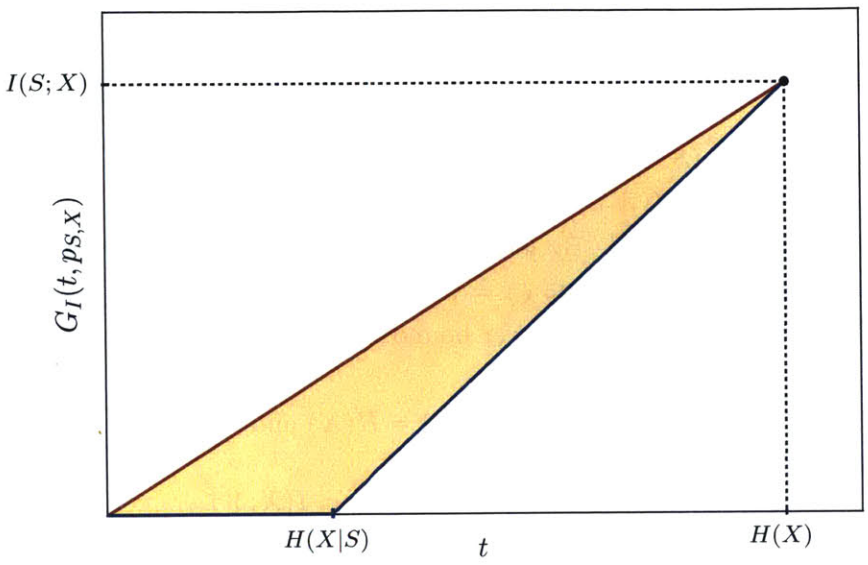

Figure 7-1: For a fixed $p_{S,X}$, the privacy region is contained within the shaded area. The red and the blue lines correspond, respectively, to the upper and lower bounds presented in Lemma 7.3.

Note that $\widetilde{Y}$ is an "erased" version of Y, with the erasure symbol being $|\mathcal{Y}|+1$. It follows directly that $I(S;\widetilde{Y}) = \lambda I(S;Y) = \lambda\alpha$, $I(X;\widetilde{Y}) = \lambda I(X;Y) \geq \lambda t$, and

$$\frac{G_I(\lambda t, P_{S,X})}{\lambda t} \leq \frac{\lambda I(S;Y)}{\lambda t} = \frac{G_I(t, p_{S,X})}{t}.$$

Since this holds for any $0 < \lambda \leq 1$, the result follows. □

Lemma 7.3. *For* $0 \leq t \leq H(X)$,

$$\max\{t - H(X|S), 0\} \leq G_I(t, p_{S,X}) \leq \frac{tI(X;S)}{H(X)}. \tag{7.3}$$

Proof. Observe that $G_I(H(X), p_{S,X}) = I(X;S)$, since $I(X;Y) = H(X)$ implies that $p_{Y|X}$ is a one-to-one mapping of X. The upper bound then follows directly from Lemma 7.2.

Clearly $G_I(t, p_{S,X}) \geq 0$. In addition, for any $p_{Y|X}$,

$$\begin{aligned} I(S;Y) &= I(X;Y) - I(X;Y|S) \\ &\geq I(X;Y) - H(X|S) \\ &\geq t - H(X|S), \end{aligned}$$

proving the lower bound. □

Figure 7-1 illustrates the bounds from Lemma 7.3. The privacy region is contained withing the shaded area. The next two examples illustrate that both the upper bound (red

line) and the lower bound (blue line) of the privacy region can be achieved for particular instances of $p_{S,X}$.

Example 7.1. Let $X = (S, W)$, where $W \perp\!\!\!\perp S$. Then by setting $Y = W$, we have $I(S;Y) = 0$ and $I(X;Y) = H(W) = H(X|S)$. Consequently, from Lemmas 7.2 and 7.3, $G_I(t, p_{S,X}) = 0$ for $t \in [0, H(X|S)]$. By letting $Y = W$ w.p. λ and $Y = (S, W)$ w.p. $1 - \lambda$ for $\lambda \in [0, 1]$, the lower-bound $G_I(t, p_{S,X}) = t - H(X|S)$ can be achieved for $H(X|S) = H(W) \leq t \leq H(X)$. Consequently, the lower bound in (7.3) is sharp.

Example 7.2. Now let $X = f(S)$. Then $I(X;S) = H(X)$ and

$$I(S;Y) = I(X;Y) - I(X;Y|S) = I(X;Y).$$

Consequently, $G_I(t, p_{S,X}) = t$, and the upper bound in (7.3) is sharp.

7.5 The Optimal Privacy-Utility Coefficient and the PICs

We now study the smallest possible ratio between disclosed private and useful information, defined next.

Definition 7.2. The *optimal privacy-utility coefficient* for a given distribution $p_{S,X}$ is given by

$$v^*(p_{S,X}) \triangleq \inf_{p_{Y|X}} \frac{I(S;Y)}{I(X;Y)}. \tag{7.4}$$

It follows directly from Lemma 7.2 that

$$v^*(p_{S,X}) = \lim_{t \to 0} \frac{G_I(t, p_{S,X})}{t}. \tag{7.5}$$

We show in Section 7.6 that the value of $v^*(p_{S,X})$ is related to the smallest principal inertia component of $p_{S,X}$ (i.e. the smallest eigenvalue of the spectrum of the conditional expectation operator, defined below). We also prove that $v^*(p_{S,X}) = 0$ is a necessary and sufficient condition for achieving perfect privacy while disclosing a non-trivial amount of useful information. Before introducing these results, we present an alternative characterization of $v^*(p_{S,X})$ (Lemma 7.4), and introduce the principal inertia components (Definition 7.3) and an auxiliary result (Lemma 7.5).

Remark 7.1. The proof of Lemma 7.4 and Theorem 7.1 in this chapter are closely related to [78]. We acknowledge that their proof techniques inspired some of the results presented here.

7.5.1 Characterization of the Optimal Privacy-Utility Coefficient

Lemma 7.4. *Let q_S denote the distribution of S when $p_{S|X}$ is fixed and $X \sim q_X$. Then*

$$v^*(p_{S,X}) = \inf_{q_X \neq p_X} \frac{D(q_S||p_S)}{D(q_X||p_X)}. \tag{7.6}$$

Proof. For fixed $p_{Y|X}$ and $p_{S,X}$

$$\begin{aligned}
\frac{I(S;Y)}{I(X;Y)} &= \frac{\sum_{y\in\mathcal{Y}} p_Y(y) D(p_{S|Y=y}||p_S)}{\sum_{y\in\mathcal{Y}} p_Y(y) D(p_{X|Y=y}||p_X)} \\
&\geq \min_{\substack{y\in\mathcal{Y}: \\ D(p_{X|Y=y}||p_X)>0}} \frac{D(p_{S|Y=y}||p_S)}{D(p_{X|Y=y}||p_X)} \\
&\geq \inf_{q_X \neq p_X} \frac{D(q_S||p_S)}{D(q_X||p_X)}.
\end{aligned}$$

Now let d^* be the infimum in the right-hand side of (7.6), and q_X satisfy

$$\frac{D(q_Y||p_Y)}{D(q_X||p_X)} = d^* + \delta,$$

where $\delta > 0$. For $\epsilon > 0$ and sufficiently small, let $p_{Y|X}$ be such that $\mathcal{Y} = [2]$, $p_Y(1) = \epsilon$, $p_{X|Y}(x|1) = q_X(x)$ and

$$p_{X|Y}(x|2) = \frac{1}{1-\epsilon} p_X(x) - \frac{\epsilon}{1-\epsilon} q_X(x).$$

Since for any distribution r_X with support $\mathcal{X}$ we have $D\left((1-\epsilon)p_X + \epsilon r_X||p_X\right) = o(\epsilon)$, we find

$$\begin{aligned}
I(S;Y) &= \epsilon D(p_{S|Y=1}||p_S) + (1-\epsilon) D(p_{S|Y=0}||p_S) \\
&= \epsilon D(q_S||p_S) + o(\epsilon),
\end{aligned}$$

and equivalently, $I(X;Y) = \epsilon D(q_X||p_X) + o(\epsilon)$. Consequently,

$$\frac{I(S;Y)}{I(X;Y)} = \frac{\epsilon D(q_S||p_S) + o(\epsilon)}{\epsilon D(q_X||p_X) + o(\epsilon)} \to d^* + \delta,$$

where the limit is taken as $\epsilon \to 0$. Since this holds for any $\delta > 0$, then $v^*(p_{S,X}) \leq d^*$, proving the result. □

7.5.2 The Smallest PIC

The smallest PIC is of particular interest for privacy, and upper bounds the value of $v^*(p_{S,X})$. In particular, we will be interested in the coefficient $\delta(p_{S,X})$, defined bellow

Definition 7.3. Let $d \triangleq \min\{|\mathcal{S}|, |\mathcal{X}|\} - 1$, and $\lambda_d(S;X)$ the smallest PIC of $p_{S,X}$. We define

$$\delta(p_{S,X}) \triangleq \begin{cases} \lambda_d(S;X) & \text{if } |\mathcal{X}| \leq |\mathcal{S}|, \\ 0 & \text{otherwise.} \end{cases} \tag{7.7}$$

The following lemma provides a useful characterization of $\delta(p_{S,X})$, related to the interpretation of the PICs as the spectrum of the conditional expectation operator given in Theorem 5.1.

Lemma 7.5. *For a given $p_{S,X}$,*

$$\delta(p_{S,X}) = \min\left\{ \|\mathbb{E}\left[f(X)|S\right]\|_2^2 \middle|\ f : \mathcal{X} \to \mathbb{R}, \mathbb{E}\left[f(X)\right] = 0, \|f(X)\|_2 = 1 \right\}. \tag{7.8}$$

Proof. Let $f : \mathcal{X} \to \mathbb{R}$, $\mathbb{E}\left[f(X)\right] = 0$ and $\|f(X)\|_2^2 = 1$, and $\mathbf{f} \in \mathbb{R}^{|\mathcal{X}|}$ be a vector with entries $f_i = f(i)$ for $i \in \mathcal{X}$. Observe that

$$\begin{aligned} \|\mathbb{E}\left[f(X)|S\right]\|_2^2 &= \sum_{s \in \mathcal{S}} p_S(s) \mathbb{E}\left[f(X)|S = s\right]^2 \\ &= \mathbf{f}^T \mathbf{P}_{X|S}^T \mathbf{D}_S \mathbf{P}_{X|S} \mathbf{f}^T \\ &= \mathbf{f}^T \mathbf{D}_X^{1/2} \mathbf{Q}^T \mathbf{Q} \mathbf{D}_X^{1/2} \mathbf{f} \\ &\geq \delta(p_{S,X}), \end{aligned}$$

where the last inequality follows by noting that $\mathbf{x} \triangleq \mathbf{f}^T \mathbf{D}_X^{1/2}$ satisfies $\|\mathbf{x}\|_2 = 1$ and that $\delta(p_{S,X})$ is the smallest eigenvalue of the positive semi-definite matrix $\mathbf{Q}^T\mathbf{Q}$, where $\mathbf{Q} \triangleq \mathbf{D}_S^{-1/2} \mathbf{P}_{X,S} \mathbf{D}_X^{-1/2}$. □

7.6 Information Disclosure with Perfect Privacy

If $v^*(p_{S,X}) = 0$, then it may be possible to disclose some information about X without revealing any information about S. However, since $G_I(0, p_{X,S}) = 0$, it is not immediately clear that $v^*(p_{S,X}) = 0$ implies that there exists t strictly bounded away from 0 such $G_I(t, p_{X,S}) = 0$. This would represent the ideal privacy setting, since, from Lemma 7.1, there would exist a privacy-assuring mapping that allows the disclosure of some non-negligible amount of useful information for $I(S;Y) = 0$. This, in turn, would mean that perfect privacy is achievable with non-negligible utility *regardless of the specific privacy metric used*, since S and Y would be independent.

In this section, we prove that if the optimal privacy-utility coefficient is 0, then there indeed exists a privacy-assuring mapping that allows the disclosure of a non-trivial amount of useful information while guaranteeing perfect privacy. We also show that the value of $\delta(p_{S,X})$ is closely related to $v^*(p_{S,X})$. This relationship is analogous to the one between the hypercontractivity coefficient s^*, defined in [77] and [108], and the maximal correlation ρ_m.

In particular, as shown in the next two theorems, $v^*(p_{S,X}) \leq \delta(p_{S,X})$ and $v^*(p_{S,X}) = 0 \iff \delta(p_{S,X}) = 0$.

Theorem 7.1. *For any $p_{S,X}$ with finite support $\mathcal{S} \times \mathcal{X}$,*

$$v^*(p_{S,X}) \leq \delta(p_{S,X}). \tag{7.9}$$

Proof. Let $p_{S|X}$ be fixed, and define

$$g_\lambda(p_X) \triangleq H(S) - \lambda H(X),$$

where $H(S)$ and $H(X)$ are the entropy of S and X, respectively, when $(S,X) \sim p_{S|X}p_X$. For $0 < \epsilon \ll 1$, let

$$p_\epsilon(i) \triangleq p_X(i)(1 + \epsilon f(i))$$

be a perturbed version of p_X, where $\mathbb{E}[f(X)] = 0$ and, w.l.o.g., $\|f(X)\|_2 = 1$. The second derivative of $g_\lambda(p_\epsilon)$ at $\epsilon = 0$ is[1]

$$\begin{aligned} \left.\frac{\partial^2 g_\lambda(p_\epsilon)}{\partial \epsilon^2}\right|_{\epsilon=0} &= \log_2(e)\left(-\|\mathbb{E}[f(X)|S]\|_2^2 + \lambda\|f(X)\|_2^2\right) \\ &= \log_2(e)\left(-\|\mathbb{E}[f(X)|S]\|_2^2 + \lambda\right). \end{aligned} \tag{7.10}$$

Thus, from Lemma 7.5, if $\lambda \leq \delta(p_{S,X})$ then for any sufficiently small perturbation of p_X, (7.10) will be non-positive. Conversely, if $\lambda > \delta(p_{S,X})$, then we can find a perturbation $f(X)$ such that (7.10) is positive. Therefore, $g_\lambda(p_X)$ has a negative semi-definite Hessian if and only if $0 \leq \lambda \leq \delta(p_{S,X})$.

For any $S \to X \to Y$, we have $I(S;Y)/I(X;Y) \geq v^*(p_{S,X})$, and, consequently, for $0 \leq \lambda^\dagger \leq v^*(p_{S,X})$,

$$g_{\lambda^\dagger}(p_X) \geq H(S|Y) - \lambda^\dagger H(X|Y),$$

and $g_{\lambda^\dagger}(p_X)$ touches the upper-concave envelope of $g_{\lambda^\dagger}$ at p_X. Consequently, $g_{\lambda^\dagger}$ has a negative semi-definite Hessian at p_X and, from (7.10), $\lambda^\dagger \leq \delta(p_{S,X})$. Since this holds for any $0 \leq \lambda^\dagger \leq v^*(p_{S,X})$, we find $v^*(p_{S,X}) \leq \delta(p_{S,X})$. □

Remark 7.2. For a fixed $p_{S|X}$, the function $g_\lambda(p_X)$ is concave when $\lambda = 0$ and convex when $\lambda = 1$. A consequence of Theorem 7.1 is that the maximum λ for which $g_\lambda(p_X)$ has a negative Hessian at p_X is $\delta(p_{S,X})$. Furthermore, Lemma 7.4 implies that the value of λ for which $g_{\lambda_1}(p_X)$ touches it's lower concave envelope at p_X for all $\lambda_1 \geq \lambda$ is $v^*(p_{S,X})$. Therefore, both $\inf_{p_X} v^*(p_{S,X})$ and $\inf_{p_X} \delta(p_{S,X})$ equal the maximum value of λ such that the function $g_\lambda(p_X)$ is concave at all values of p_X. Therefore, we established that for a given

[1] This was observed in [78] and [108], and follows directly from $-\frac{\partial^2}{\partial \epsilon^2} a(1+b\epsilon)\log_2 a(1+b\epsilon) = -b^2 a \log_2(e)$.

$p_{S|X}$,

$$\inf_{p_X} v^*(p_{S,X}) = \inf_{p_X} \delta(p_{S,X}).$$

The next theorem proves that $\delta(p_{S,X})$ can serve as a proxy for perfect privacy, since the optimal privacy-utility coefficient is 0 if and only if $\delta(p_{S,X})$ is also 0.

Theorem 7.2. *Let $p_{S,X}$ be such that $H(X) > 0$ and $\mathcal{S}$ and $\mathcal{X}$ are finite. Then*[2]

$$v^*(p_{S,X}) = 0 \iff \delta(p_{S,X}) = 0. \tag{7.11}$$

Proof. Theorem 7.1 immediately gives $\delta(p_{S,X}) = 0 \Rightarrow v^*(p_{S,X}) = 0$. Let $v^*(p_{S,X}) = 0$. Then, since $D(q_X||p_X) \leq -\min_{i\in\mathcal{X}} \log_2 p_X(i)$ and $\mathcal{X}$ is finite, Lemma 7.4 implies that for any $\epsilon > 0$ there exists q_X and $0 < \delta \leq -\min_{i\in\mathcal{X}} \log_2 p_X(i)$ such that

$$D(q_X||p_X) \geq \delta > 0$$

and

$$D(q_S||p_S) < \epsilon.$$

We can then construct a sequence $q_X^1, q_X^2, q_X^3, \ldots$ such that $q_X^i \neq p_X$, $D(q_S^k||p_S) \leq \epsilon_k$ and

$$\lim_{k\to\infty} \epsilon_k = 0.$$

Let $\mathbf{q}_S^k$ be a vector whose entries are $q_S^k(\cdot)$. Then, from Pinsker's inequality,

$$\epsilon_k \geq \frac{1}{2}\|\mathbf{q}_S^k - \mathbf{p}_S\|_1^2 \geq \frac{1}{2}\|\mathbf{q}_S^k - \mathbf{p}_S\|_2^2. \tag{7.12}$$

Defining $\mathbf{x}^k = \mathbf{q}_X^k - \mathbf{p}_X$, observe that $0 < \|\mathbf{x}^k\|_2^2 \leq 2$ and, from (7.12), $\|\mathbf{P}_{S|X}\mathbf{x}^k\|_2 \leq \sqrt{2\epsilon_k}$. Hence,

$$\lim_{k\to\infty} \frac{\|\mathbf{P}_{S|X}\mathbf{x}^k\|_2^2}{\|\mathbf{x}^k\|_2^2} = 0. \tag{7.13}$$

In addition, denoting $s_m \triangleq \min_{s\in\mathcal{S}} p_S(s)$ and $x_M \triangleq \min_{x\in\mathcal{X}} p_X(x)$, for each k we have

$$\begin{aligned}
\frac{\|\mathbf{P}_{S|X}\mathbf{x}^k\|_2^2}{\|\mathbf{x}^k\|_2^2} &\geq \min_{\|\mathbf{y}\|_2^2>0} \frac{\|\mathbf{P}_{S|X}\mathbf{y}\|_2^2}{\|\mathbf{y}\|_2^2} \\
&= \min_{\|\mathbf{y}\|_2^2>0} \frac{\|\mathbf{P}_{S,X}\mathbf{D}_X^{-1/2}\mathbf{y}\|_2^2}{\|\mathbf{D}_X^{1/2}\mathbf{y}\|_2^2} \qquad (7.14) \\
&\geq \min_{\|\mathbf{y}\|_2^2>0} \frac{s_m\|\mathbf{D}_S^{-1/2}\mathbf{P}_{S,X}\mathbf{D}_X^{-1/2}\mathbf{y}\|_2^2}{x_M\|\mathbf{y}\|_2^2} \qquad (7.15)
\end{aligned}$$

[2]If S is binary, then (7.11) implies that perfect privacy is achievable iff S and X are independent (since $\delta(p_{S,X}) = \rho_m(S;X)^2$), recovering [107, Thm. 2].

$$= \frac{s_m}{x_M} \min_{\|\mathbf{y}\|_2^2>0} \frac{\|\mathbf{Q}\mathbf{y}\|_2^2}{\|\mathbf{y}\|_2^2} \tag{7.16}$$

$$= \frac{s_m \delta(p_{S,X})}{x_M}. \tag{7.17}$$

In the derivation above, (7.14) follows from $\mathbf{D}_X$ being invertible (by definition), (7.15) is a direct consequence of $\|\mathbf{D}_S^{-1/2}\mathbf{y}\|_2^2 \leq s_m^{-1}\|\mathbf{y}\|_2^2$ and $\|\mathbf{D}_X^{1/2}\mathbf{y}\|_2^2 \leq x_M\|\mathbf{y}\|_2^2$ for any $\mathbf{y}$, and (7.16) and (7.17) follow from the definition of $\mathbf{Q}$ and $\delta(p_{S,X})$, respectively. Combining (7.17) with (7.13), it follows that $\delta(p_{S,X}) = 0$, proving the desired result. □

We are now ready to prove that a non-trivial amount of useful information can be disclosed without revealing any private information if and only if $v^*(p_{S,X}) = 0$ (or equivalently, $\delta(p_{S,X}) = 0$). This result follows naturally from Theorem 7.2, since $v^*(p_{S,X}) = 0$ implies that $\delta(p_{S,X}) = 0$, which means that the matrix $\mathbf{Q}$ and, consequently, $\mathbf{P}_{S|X}$, is either not full rank or has more columns than rows (i.e. $|\mathcal{X}| > |\mathcal{S}|$). This, in turn, can be exploited in order to find a mapping $p_{Y|X}$ such that Y reveals some information about X, but no information about S. This argument is made precise in the next theorem.

Remark 7.3. When $\mathbf{P}_{S|X}$ is not full rank or has more columns than rows, then S and X are weakly independent. As shown in [109, Thm. 4] and [107], this implies that a privacy-assuring mapping that achieves perfect privacy while disclosing a non-trivial amount of useful information can be found. Theorem 7.3 recovers this result in terms of the smallest principal inertia component, and Cor. 7.2 provides an estimate of the amount of useful information that can be revealed with perfect privacy.

Theorem 7.3. *For a given $p_{S,X}$, there exists a privacy-assuring mapping $p_{Y|X}$ such that $S \to X \to Y$, $I(X;Y) > 0$ and $I(S;Y) = 0$ if and only if $\delta(p_{S,X}) = 0$ (equivalently $v^*(p_{S,X}) = 0$). In particular,*

$$\exists t_0 > 0 : G_I(t_0, p_{S,X}) = 0 \iff \delta(p_{S,X}) = 0. \tag{7.18}$$

Proof. The direct part of the theorem follows directly from the definition of $v^*(p_{S,X})$ and Theorem 7.2. Assume that $\delta(p_{S,X}) = 0$. Then, from Lemma 7.5, there exists $f : \mathcal{X} \to \mathbb{R}$ such that $\|f(X)\|_2 = 1$, $\mathbb{E}[f(X) = 0]$, and $\|\mathbb{E}[f(X)|S]\|_2 = 0$. Consequently, $\mathbb{E}[f(X)|S = s] = 0$ for all $s \in \mathcal{S}$.

Fix $\mathcal{Y} = [2]$, and, for $\epsilon > 0$ and ϵ appropriately small,

$$p_{Y|X}(y|x) = \begin{cases} \frac{1}{2} - \epsilon f(x), & y = 1, \\ \frac{1}{2} + \epsilon f(x), & y = 2. \end{cases}$$

Note that it is sufficient to choose $\epsilon = (2\max_{x\in\mathcal{X}}|f(X)|)^{-1}$, so ϵ is strictly bounded away

from 0. In addition, $p_Y(1) = 1/2$. Therefore,

$$I(X;Y) = 1 - \sum_{x \in \mathcal{X}} p_X(x) h_b\left(\frac{1}{2} + \epsilon f(x)\right) > 0 \tag{7.19}$$

where $h_b(x) \triangleq -x \log_2 x - (1-x)\log_2(1-x)$ is the binary entropy function. Since $S \to X \to Y$,

$$\begin{aligned} p_{Y|S}(y|s) &= \sum_{x \in \mathcal{X}} p_{Y|X}(y|x) p_{X|S}(x|s) \\ &= \sum_{x \in \mathcal{X}} \left(\frac{1}{2} + (-1)^y \epsilon f(x)\right) p_{X|S}(x|s) \\ &= 1/2 + (-1)^y \epsilon \mathbb{E}\left[f(X)|S = s\right] \\ &= 1/2, \end{aligned}$$

and, consequently, S and Y are independent. Then $I(S;Y) = 0$, and the result follows. □

The previous result proves that if either $|\mathcal{X}| > |\mathcal{S}|$ or the smallest principal inertia component of $p_{S,X}$ is 0 (i.e. $\delta(p_{S,X}) = 0$), then it is possible to achieve perfect privacy while disclosing some useful information. In particular, the value of t_0 in (7.10) is lower-bounded by the expression in (7.19). We note that this result would not necessarily hold if $\mathcal{S}$ and $\mathcal{X}$ are not finite sets.

The proof of Theorem 7.3 holds for *any* measure of information $\mathcal{I}$ that satisfies $\mathcal{I}(X;Y) = 0$ if and only if X and Y are independent, since it depends solely on the properties of $p_{S,X}$. Examples of $\mathcal{I}$ are maximal correlation or information metrics based on f-divergences [110]. This leads to the following result.

Corollary 7.1. *Let $p_{S,X}$ be given, and $\mathcal{I}$ be a non-negative measure of information (e.g. total variation or maximal correlation) such that for any two random variable A and B $\mathcal{I}(A;B) = 0 \iff A \perp\!\!\!\perp B$. Then there exists $p_{Y|X}$ such that $S \to X \to Y$, $\mathcal{I}(X;Y) > 0$ and $\mathcal{I}(S;Y) = 0$ if and only if $\delta(p_{S,X}) = 0$.*

Proof. This is a direct consequence of Theorem 7.3, since, by assumption, $\mathcal{I}(X;Y) = 0 \iff I(X;Y) = 0$ and $\mathcal{I}(S;Y) = 0 \iff I(S;Y) = 0$. □

Remark 7.4. As long as privacy is measured in terms of statistical dependence (with perfect privacy implying statistical independence) and some utility can be derived when Y is not independent of X, then $\delta(p_{S,X})$ fully characterizes when perfect privacy is achievable with non-trivial utility.

7.7 On the Amount of Useful Information Disclosed with Perfect Privacy

We present next an explicit lower bound for the largest amount of useful information that can be disclosed while guaranteeing perfect privacy. The result follows directly from the construction used in the proof of Theorem 7.3.

Corollary 7.2. *For fixed $p_{S,X}$, let*

$$\mathcal{F}_0 \triangleq \{f : \mathcal{X} \to \mathbb{R} | \mathbb{E}[f(X)] = 0,\ \|f(X)\|_2 = 1,\ \|\mathbb{E}[f(X)|S]\|_2 = 0\} \cup f_0,$$

where f_0 is the trivial function that maps $\mathcal{X}$ to $\{0\}$. Then $G_I(t, p_{S,X}) = 0$ for $t \in [0, t^*]$, where

$$t^* \geq 1 - \max_{f \in \mathcal{F}_0} \mathbb{E}\left[h_b\left(\frac{1}{2} + \frac{f(X)}{2\|f\|_\infty}\right)\right]. \tag{7.20}$$

Furthermore, the lower bound for t^* is sharp when $\delta(p_{S,X}) = 0$, i.e. there exists a $p_{S,X}$ such that $t^* > 0$ and $G_I(t, p_{S,X}) = 0$ if and only if $t \in [0, t^*]$.

Proof. If $\delta(p_{S,X}) = 0$, then the lower bound for t^* follows directly from the proof of Theorem 7.3 and, in particular, (7.18). If $\delta(p_{S,X}) > 0$, then $\mathcal{F}_0 = \{f_0\}$, and the lower bound (7.20) reduces to the trivial bound $t^* \geq 0$.

In order to prove that the lower bound is sharp, consider S being an unbiased bit, drawn from $\{1, 2\}$, and X the result of sending S through an erasure channel with erasure probability $1/2$ and $\mathcal{X} = \{1, 2, 3\}$, with 3 playing the role of the erasure symbol. Let

$$f(x) \triangleq \begin{cases} 1, & x \in \{1, 2\}, \\ -1 & x = 3. \end{cases}$$

Then $f \in \mathcal{F}_0$, $h_b\left(\frac{1}{2} + \frac{f(x)}{2\|f\|_\infty}\right) = 0$ for $x \in \mathcal{X}$ and $t^* = 1$. But, from Lemma 7.3, $t^* \leq H(X|S) = 1$. The result follows. □

The previous bound for t^* can be loose, especially if $|\mathcal{X}|$ is large. In addition, the right-hand side of (7.20) can be made arbitrarily small by decreasing $\min_{x \in \mathcal{X}} p_X(x)$. Nevertheless, (7.20) is an explicit estimate of the amount of useful information that can be disclosed with perfect privacy.

When $S^n = (S_1, \ldots, S_n)$ and $X^n = (X_1, \ldots, X_n)$, where $(S_i, X_i) \sim p_{S,X}$ are i.i.d. random variables, the next proposition states that $\delta(p_{S^n,X^n}) = \delta(p_{S,X})^n$. Consequently, as long as $\delta(p_{S,X}) < 1$, it is possible to disclose a non-trivial amount of useful information while disclosing an arbitrarily small amount of private information by making n sufficiently large.

Proposition 7.1. *Let $S^n = (S_1, \ldots, S_n)$ and $X^n = (X_1, \ldots, X_n)$, where $(S_i, X_i) \sim p_{S,X}$*

are i.i.d. random variables. Then

$$v^*(p_{S^n,X^n}) \leq \delta(p_{S^n,X^n}) = \delta(p_{S,X})^n. \tag{7.21}$$

Proof. The result is a direct consequence of the tensorization property of the principal inertia components, presented in [64, 79, 82]. □

7.8 The Correlation-Error Product Revisted

In this section we revisit the correlation-error product results introduced in Section 3.8. The principal functions (cf. Defn 7.3) provide a basis for for the functions of a random variable, and the PICs measure the MMSE of estimating each of these functions. Consequently, the PICs provide a complete characterization of which functions can or cannot be inferred reliably. This extends the analysis in Chapter 3, where the functions were arbitrary (i.e. not necessarily forming a basis).

We go over again a few definitions that are key to this section and were already introduced previously in the thesis. As usual, let X and Y be two random variables (not necessarily with discrete support sets) with finite second moment and support $\mathcal{X}$ and $\mathcal{Y}$, respectively. For $f : \mathcal{X} \to \mathbb{R}$, recall that $\mathsf{mmse}(f(X)|Y) = \mathbb{E}[f(X)|Y]$ is the minimum mean squared error estimator of $f(X)$ given an realization of Y. Observe that the mean squared error $f(X)$ given Y can be expressed as

$$\begin{aligned}\mathsf{mmse}(f(X)|Y) &= \mathbb{E}\left[f(X)^2 - \mathbb{E}[f(X)|Y]^2\right] \\ &= \|f(X)\|_2^2\left(1 - \frac{\|\mathbb{E}[f(X)|Y]\|_2^2}{\|f(X)\|_2^2}\right),\end{aligned} \tag{7.22}$$

Consequently, the value of the MMSE depends on the spectrum of the conditional expectation operator $Tf(y) \triangleq \mathbb{E}[f(X)|Y = y]$ which, in turn, is composed by the principal inertia components (cf. Theorem 5.1).

In the rest of this section we have two main goals: (i) determine the functions of X that can be inferred with small minimum mean squared error from Y, and (ii) relate these functions to the MMSE estimator of X given Y.

7.8.1 Functions That Can Be Inferred With Small MMSE

We assume henceforth that X has finite second moment and, in order to ignore trivial constant mappings, that $\mathbb{E}[f(X)] = 0$ and $\mathbb{E}\left[f(X)^2\right] = 1$. Under these assumptions, one can determine functions $f_1, f_2, \ldots$ as in Theorem 5.1, such that f_i is given by

$$\begin{aligned}f_i = \operatorname{argmax}\big\{\|\mathbb{E}[f(X)|Y]\|_2^2 \,|\, f : \mathcal{X} \to \mathbb{R},\ \mathbb{E}[f(X)] = 0,\ \mathbb{E}\left[f(X)^2\right] = 1, \\ \mathbb{E}[f(X)f_j(X)] = 0 \text{ for } 1 \leq j \leq i-1\big\}.\end{aligned}$$

Then

$$\mathbb{E}\left[f_i(X)|Y\right] = \lambda_i(X;Y)^2,$$

where $\sigma_1^2, \sigma_2^2 \ldots$ are the PICs of X and Y.

It follows directly that for any non-trivial function $h : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}\left[h(X)\right] = 0$

$$\mathsf{mmse}(h(X)|Y) \geq \|h(X)\|_2^2 \left(1 - \rho_m(X;Y)^2\right), \tag{7.23}$$

with equality if $h(X) = cf_1(X)$, where $c = \|h(X)\|_2$. Therefore, for a fixed variance c, $cf_1(X)$ is the function of X that can be most reliably estimated (in terms of mean-squared error) from all possible mapping $\mathcal{X} \to \mathbb{R}$. Furthermore,

$$\mathsf{mmse}(h(X)|Y) = \|h(X)\|_2^2 \left(1 - \sum_i c_i^2 \lambda_i(X;Y)\right), \tag{7.24}$$

where $c_i = \mathbb{E}\left[h(X)f_i(X)\right] / \|h(X)\|_2$ and $\sum_i c_i^2 = 1$. Consequently, functions that are closely "aligned" with f_i for small i can be inferred with small mean squared error.

This result is at the heart of the correlation-error product discussed in Section 3.8, and reveals the true nature of the fundamental tradeoff between privacy and utility. The principal functions f_i that correspond to smaller PICs are the ones that cannot be reliably inferred from Y. For example, assume that we wish to design a privacy-assuring mapping where the secret $S = (h_1(X), \ldots, h_t(X))$ is composed by a certain set of functions $h_1, \ldots, h_k$ of X that are supposed to remain private. The privacy-assuring mapping $p_{Y|X}$ should then assure that the principal functions that span the subspace formmed by $(h_1(X), \ldots, h_t(X))$ must have small PICs. We show how to design privacy-assuring mappings based on this intuition in Section 7.9. Before that, we present a brief discussion on the connection between the PICs and the MMSE estimator (i.e. estimating the identity function of X given Y).

7.8.2 PICs and the MMSE Estimator

The MMSE is a property of the probability space and the support sets $\mathcal{X}$ and $\mathcal{Y}$ of the random variables, whereas the PICs and the corresponding principal functions are properties solely of the probability space. If one is given only the probability measure and asked to define random variables X and Y so that X can be estimated from Y with the minimum mean squared error and has unit variance, then $f_1(X) = X$. This illustrates why principal inertia components are used in the analysis of categorical data (i.e. data without associated numerical values).

If the random variables X and Y are given and there is some underlying functional dependency between them, is this dependency better captured by the MMSE estimator of X given Y or by f_1? We will first discuss a simple example where Y is a deterministic function of X.

Deterministic Mappings

Let $Y = h(X)$, where X is a random variable with finite second moment and $h : \mathcal{X} \to \mathbb{R}$. We assume, for the sake of illustration, that $|\mathcal{X}| < \infty$. Then for $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f(X)^2] = \mathbb{E}[g(Y)^2] = 1$

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \mathbb{E}[f(X)\mathbb{E}[g(Y)|X]] \\ &= \mathbb{E}[f(X)g(h(X))] \\ &\leq 1,\end{aligned}$$

with equality if and only if $f(X) = g(h(X))$. Therefore, all the PICs are equal to unity, and the corresponding eigenvectors $g_1, \ldots, g_d$, where $d = |\mathcal{Y}| - 1$, is given by *any* set of $d-1$ orthonormal vectors that satisfy the mean and variance constraints. Note that there are *an infinite number* of such vectors, and therefore, an infinite number of basis $f_1, \ldots f_d$. In this case, the eigenvectors in general do not capture in any way capture the dependency between X and Y.

Assuming, without loss of generality, that $\mathbb{E}[h(X)] = 0$, we may set $f(X) = h(X)/\|h(X)\|_2$ and $g(Y) = Y/\|h(X)\|_2$, and the maximal correlation is achieved. However, *any other mapping* $z : \mathcal{X} \to \mathbb{R}$ that satisfies $h^{-1}(x) = z^{-1}(x)$ for all $x \in \mathcal{X}$ and $\mathbb{E}[z(X)] = 0$ would also achieve the maximal correlation, so the relationship is *not* defined by any given basis of eigenvectors.

The eigenvectors give the structure of this dependency in the probability space, and this is of interest if the numerical values of X and Y are irrelevant (such as the case considered in correspondence analysis). However, in order to truly define the nature of the dependency between X and Y as seen in data, it is necessary to take into consideration the random variables. The MMSE estimator of X given an observation of Y also does not necessarily reveal this relationship, at least not in its entirety. As a simple example, let $Y = X^2$ and X have a distribution that is symmetric around 0, then $\mathbb{E}[X|Y] = 0$, and $\mathsf{mmse}(X|Y)$ is trivial. Nevertheless, $\mathsf{mmse}(Y|X)$ perfectly captures the dependency between Y and X, since $\mathbb{E}[Y|X] = h(X)$.

In General: Maximal Correlation and MMSE

In order to relate the MMSE estimator with maximal correlation, we introduce the following definitions.

Definition 7.4. For $\mathrm{Var}(Y) > 0$, the one-sided maximal correlation coefficient between X and Y is given by

$$\rho_s(Y|X) = \sup\left\{\mathbb{E}\left[\frac{f(X)(Y - \mathbb{E}[Y])}{\sqrt{\mathrm{Var}(Y)}}\right] \mid f : \mathcal{X} \to \mathbb{R}, \mathbb{E}[f(X)] = 0, \mathbb{E}[f(X)^2] = 1\right\}. \tag{7.25}$$

The function f that achieves the supremum in (7.25) is denoted by $r_X^*(x)$.

Observe that $\rho_s(Y|X)$ is not symmetric in general, i.e. $\rho_s(Y|X) \neq \rho_s(X|Y)$. Furthermore

$$\rho_m(X;Y) = \sup\left\{\rho_s(g(Y)|X) | g : \mathcal{Y} \to \mathbb{R}, \mathbb{E}\left[g(Y)\right] = 0, \mathbb{E}\left[g(Y)^2\right] = 1\right\}.$$

Consequently $\rho_s(Y|X) \leq \rho_m(X;Y)$.

The next theorem shows that $r_X^*(x)$ matches the MMSE estimator (up to an affine transformation).

Theorem 7.4. *$\rho_s(Y|X) = \|\mathbb{E}\left[Z|X\right]\|_2$, where $Z = (Y - \mathbb{E}\left[Y\right])/\sqrt{\mathrm{Var}(Y)}$. Furthermore, if $\rho_s(Y|X) > 0$, then $r^*(x) = \mathbb{E}\left[Z|X = x\right]/\|Z|X\|_2$.*

Proof. Let $Z = (Y - \mathbb{E}\left[Y\right])/\sqrt{\mathrm{Var}(Y)}$ and $\mathbb{E}\left[f(X)\right] = 0, \mathbb{E}\left[f(X)^2\right] = 1$. Then

$$\begin{aligned}\mathbb{E}\left[f(X)Z\right] &= \mathbb{E}\left[f(X)\mathbb{E}\left[Z|X\right]\right] \\ &\leq \|f(X)\|_2 \|\mathbb{E}\left[Z|X\right]\|_2 \\ &= \|\mathbb{E}\left[Z|X\right]\|_2.\end{aligned}$$

Equality holds if and only if $f(X) = c\mathbb{E}\left[Z|X\right]$ for some constant c. Letting $c = 1/\|\mathbb{E}\left[Z|X\right]\|_2$, the result follows. □

A direct consequence of this characterization and Theorem 5.1is given below.

Corollary 7.3. *Let f, g be the principal functions that maximize correlation, i.e. $\mathbb{E}\left[f(X)g(Y)\right] = \rho_m(X;Y)$. Then*

$$g(y) = \mathbb{E}\left[f(X)|Y = y\right]/\|\mathbb{E}\left[f(X)|Y\right]\|_2$$

and $f(x) = \mathbb{E}\left[g(Y)|X = x\right]/\|\mathbb{E}\left[g(Y)|X\right]\|_2$. In this case, $\rho_s(f(X)|Y) = \rho_s(g(Y)|X) = \rho_m(X;Y)$.

If $\rho_s(X|Y) = \rho_s(Y|X)$, then are they equal to $\rho_m(X;Y)$? Alas, the answer is no in general (but true if either X or Y are binary). To see why this is the case, let $f(X)$ and $g(Y)$ match the principal function corresponding to the second largest PIC of X and Y. However, if X is binary, we have only one such component, and equality holds.

7.9 Privacy-Assuring Mappings with Estimation Constraints

In this section we present a linear program that can be used for deriving privacy-assuring mappings with estimation constraints under certain assumptions. Consider the following setup: We are given a set of n i.i.d. samples $(X_1, \ldots, X_n)$, where $X_i \sim p_X$ and $\mathcal{X} = [m]$ is finite. Our goal is to produce a new set of samples $(X_1', \ldots, X_n')$ where $X_i' \sim p_X$, and X_i' is

produced from X_i through a privacy-assuring mapping $p_{X'|X}$. In addition, we are given the following constraints: For a certain set of functions $\{f_k\}_{k=1}^{t_1}$, $f_k : \mathcal{X} \to \mathbb{R}$,

$$\mathsf{mmse}(f_k(X_i)|X_i') \geq \theta_k,$$

and for another set of functions $\{\widetilde{f}_k\}_{k=1}^{t_2}$, $\widetilde{f}_k : \mathcal{X} \to \mathbb{R}$, we have

$$\mathsf{mmse}(\widetilde{f}_k(X_i)|X_i') \leq \widetilde{\theta}_k.$$

In other words, we wish to produce a new set of samples such that the functions $\{f_k\}_{k=1}^{t_1}$ of the original set of samples cannot be estimated reliable (privacy constraints), whereas the functions $\{\widetilde{f}_k\}_{k=1}^{t_2}$ can be estimated reliably (utility constraints), and both set of samples have the same distribution. How should $p_{X'|X}$ be chosen? We discuss next a procedure for generating a privacy-assuring mapping $p_{X'|X}$ that satisfies these constraints (if they are feasible).

We assume, without loss of generality, that each function f_k and $\widetilde{f}_k$ has zero-mean and unit variance. We also make the additional, crucial assumption that $\{f_1, \ldots, f_{t_1}, \widetilde{f}_1, \ldots, \widetilde{f}_{t_2}\}$ are uncorrelated (and, consequently, $t_1 + t_2 \leq m - 1$). This assumption is crucial for the convex program below to be a linear program.

We can find a candidate $p_{X'|X}$ by solving the linear program described in the next proposition. As usual, $\mathbf{D}_X$ is a diagonal matrix with entries p_X, and we denote by $\mathbf{f}_k$ the column vector with i-th entry equal to $f_k(i)$, and equivalently for $\widetilde{\mathbf{f}}_k$.

Proposition 7.2. *For the constraints described above, a candidate privacy-assuring mapping $p_{X'|X}$ can be found by solving the following linear program:*

$$\min_{\sigma_1,\ldots,\sigma_{t_1+t_2}} \quad \sum_i a_i \sigma_i \tag{7.26}$$

$$\text{s.t.} \quad \mathbf{P}_{X'|X} = \mathbf{F}\mathbf{\Sigma}\mathbf{F}^T\mathbf{D}_X, \tag{7.27}$$

$$\mathbf{P}_{X'|X}\mathbf{1} = \mathbf{1}, \tag{7.28}$$

$$[\mathbf{P}_{X'|X}]_{i,j} \geq 0, \tag{7.29}$$

$$\sigma_i \leq \sqrt{1 + \theta_i}, \ i = 1, \ldots, t_1 \tag{7.30}$$

$$\sigma_j \geq \sqrt{1 + \widetilde{\theta}_{t_1 - j}}, \ j = t_1 + 1, \ldots, t_1 + t_2, \tag{7.31}$$

where $a_i \in \mathbb{R}$ is arbitrary and may depend on the application at hand, $\mathbf{F} \triangleq [\mathbf{1}\ \mathbf{f}_1\ \cdots\ \mathbf{f}_{t_1}\ \widetilde{\mathbf{f}}_1\ \cdots \widetilde{\mathbf{f}}_{t_2}]$ and $\mathbf{\Sigma} \triangleq \operatorname{diag}(1, \sigma_1, \ldots, \sigma_{t_1+t_2})$. The mapping $p_{X'|X}$ is given by the entries of $\mathbf{P}_{X'|X}$.

Proof. From the definition of the PICs (cf. Definition 5.1), and denoting by $\mathbf{P}_{X'|X}$ the matrix with entries $p_{X'|X}$, we have

$$\mathbf{P}_{X'|X} = \mathbf{D}_X^{-1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{D}_X,$$

where the square-root of the PICs are the entries of $\mathbf{\Sigma}$, and $\mathbf{D}_X^{-1/2}\mathbf{U}$ and $\mathbf{D}_X^{-1/2}\mathbf{V}$ are the corresponding principal functions. We then impose that the principal functions of X and X' are given by the functions $\{f_1, \ldots, f_{t_1}, \widetilde{f}_1, \ldots, \widetilde{f}_{t_2}\}$, resulting in constraint (7.27). It follows directly that the MMSE of estimating f_i (resp. $\widetilde{f}_i$) given X' is $\mathsf{mmse}(f_i(X)|X') = \sigma_i^2$ (resp. $\mathsf{mmse}(\widetilde{f}_i(X)|X') = \sigma_{i+t_1}^2$), resulting in constraints (7.30) and (7.31). Constraints (7.28) and (7.29) guarantee that $p_{X'|X}$ is a valid conditional distribution. □

Remark 7.5. The exact objective function in (7.26) can depend on the overall desired privacy or utility goal. If the functions $\{f_1, \ldots, f_{t_1}, \widetilde{f}_1, \ldots, \widetilde{f}_{t_2}\}$ are not orthogonal, then a standard orthogonalization procedure (e.g. Gram-Schmidt process) can be found to find a basis $\{h_1, \ldots, h_d\}$ to be used in the previous linear program. However, if the privacy/utility MMSE constraints cannot be translated directly in terms of individual constraints of $\{h_1, \ldots, h_d\}$ (e.g. the constraints depend on a linear combination of h_i's), then the resulting optimization program may not be convex.

7.10 The Power of Subsampling for Privacy

We conclude this chapter with a high-level discussion of the application of the PICs in database privacy. The goal of this section is not to provide a complete study of the matter, but rather to motivate schemes that use the bounds on estimation results derived here for designing new privacy-assuring mechanisms. We also seek to understand the use of simple techniques, such as subsampling, to guarantee privacy in this context.

Within this section, let $D = \{Y_1, \ldots, Y_n\}$ be a database whose entries are the realization of n random variables $Y_1, \ldots, Y_n$ with probability distribution P_Y^n. We assume that Y_i has a discrete support set $\mathcal{Y}$. A given query over the database is represented by a mapping $f : \mathcal{Y}^n \to \mathcal{C}$ for some set $\mathcal{C}$. It is natural to assume that queries over D are symmetric functions of $Y_1, \ldots, Y_n$, since they are agnostic to the logical ordering of the database entries in memory. This is the case, for example, of SQL queries. Queries are usually of the type "What is the average value of the entries of D" or "How many entries of D satisfy a certain property" and so on.

Assume that a given user requests the answer to a query f computed over D. How can we reply to f while still preserving the privacy of individual inputs of D? We need to first define what we mean by *privacy* within this database setting. We identify three related approaches.

Differential Privacy

Differential privacy attempts to mitigate the privacy threat due to the variation of an individual entry of D. A particular query f is said to be differentially private if for any $d = \{y_1, \ldots, y_k, \ldots, y_n\}$ and $d' = \{y_1, \ldots, y'_k, \ldots, y_n\}$, where $y_k \neq y'_k$, we have $\Pr\{f(d) \in$

$\mathcal{A}\} \approx \Pr\{f(d') \in \mathcal{A}\}$ for all $\mathcal{A} \in \mathcal{C}$ [12]. This can be achieved, for example, by adding noise to the output of the query. We do not consider differential privacy here.

Privacy Against Statistical Inference

This is the privacy framework considered in the privacy funnel in Section 7.4 and introduced in [18]. The goal in this case is to prevent that private information about the individual entries of D from being learned by a third-party. Assume that for each Y_i there is a private variable S_i that represents the secret (private) information of each user. For example, Y_i might be a vector representing items purchased at a supermarket from client i, and S_i a binary variable representing if client i is expecting a child or not. The privacy mechanism would then generate a new database D' such that, for a given distortion metric Δ and query f, $\Delta(f(D'), f(D)) \leq \delta$, and, simultaneously, the values of S_i cannot be reliably inferred from D'.

Database Privacy

The desiderata of database privacy, defined next, is to guarantee that information about the *distribution* of the entries of the database can be learned through queries, but not information regarding individual realizations of the entries of the database. This idea is captured in the following definition, where, as usual, we denote $\mathcal{F}_{\text{sym}}(\mathcal{Y}^n) \triangleq \{f | f : \mathcal{Y}^n \to \mathbb{R}, f \text{ is symmetric}\}$.

Definition 7.5. For a given query f and a distortion metric $\Delta : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, we say that a function $\widetilde{f}$ is a $(\epsilon, \delta, \alpha)$-database private version of a query f if assuming that the entries of D are drawn from a set of exchangeable random variables Y^∞ (cf. Section 5.8.2)

$$\Pr\{\Delta(f(D), \widetilde{f}(D)) \geq \delta\} \leq \alpha \text{ and} \tag{7.32}$$

$$\mathsf{Adv}(g(D)|\widetilde{f}(D)) \leq \epsilon \; \forall g \in \mathcal{F}_{\text{sym}}(\mathcal{Y}^n), \; g \neq f. \tag{7.33}$$

Remark 7.6. The previous privacy definition can be extended to a setting where the adversary performs multiple queries.

Remark 7.7. Database privacy, as defined here, has two key limitations: (i) the assumption that the entries of the database are drawn from an exchangeable distribution and (ii) in general, if the advantage of guessing any symmetric function is small, then probably the query f is not very informative. Nevertheless, the goal of the definition is that queries that are data-base private reveal properties of the *distribution* of the entries in D, i.e. are informative about p_{Y^n}, but do not reveal much information about the individual entries.

7.10.1 Database Privacy and Subsampling

The results introduced in Chapter 5 indicate that answering queries over a randomly selected subsample of entries of D is a powerful method for guaranteeing database privacy. Consider a sequence of functions $f_k : \mathcal{Y}^k \to \mathbb{R}$ such that there exists a constant $c_0 \in \mathbb{R}$ where $f_k(Y^k) \to c_0$ with high probability. Then, for sufficiently large m and n, $m < n$,

$$\Pr\{\Delta(f_m(Y^m), f_n(Y^n)) \geq \delta\} \leq \alpha,$$

where the exact values of δ and α will depend on the nature of the convergence of f_k. For example, if f_k are averages (i.e. $f_k(Y^k) = \sum_{i\in[k]} h(Y_i)/k$), then this convergence is exponentially fast. In addition, from Theorem 5.9 and the DPI for the PICs, for any symmetric function $g \in \mathcal{F}_{\text{sym}}(\mathcal{Y}^n)$

$$\lambda_1(g(Y^n); f_m(Y^m)) \leq \frac{m}{n},$$

and, from bound (5.2),

$$\mathsf{Adv}(g(Y^n)|f_m(Y^m)) \leq \sqrt{\frac{m}{n}}. \tag{7.34}$$

This provides a very simple, yet powerful method for guaranteeing database privacy for statistical queries: simply reply with the query computed over a random subsample of the database, where the exact size of the subsample depends on the privacy and distortion constraints. The probability of the adversary guessing an undesired function of the database will then scale according to (7.34), and accurate query replies can still be given as long as the queries concern aggregate statistics. This method is simple and intuitive, and the results presented in this thesis allow a precise characterization of the level of privacy that can be achieved.

7.11 Final Remarks

The PICs are powerful statistics that provide a fine-grained decomposition of the joint probability distribution of two random variables. As demonstrated in Chapter 6, the PICs play an important role in information theory, and can be used to precisely characterize the effect of a random transformation $p_{Y|X}$ on the functions of a hidden random variable X. This analysis, in turn, sheds light on which functions of a hidden variable X are the most informative given an observation Y for a wide-range of information metrics.

In this chapter, we showed that the PICs are also particularly well suited as security and privacy metrics. The PICs simultaneously (i) determine when useful information can be disclosed with perfect privacy, (ii) characterize which functions of a private variable can or cannot be reliably inferred given the output of a security system, and (iii) provide a benchmark metric against which existing and new privacy metrics can be compared to.

We believe that the study and value of the PICs go well beyond the ones suggested in

this thesis. The theoretical properties of the PICs are of independent interest, and their connection with other information-theoretic metrics is still not fully understood (e.g. in the context of Strong Data Processing Inequalities [104]). Furthermore, we are convinced that the PICs are of value to other applications beyond security and privacy, such as in statistical learning and coding theory. In the final chapter of this thesis, we present some potential future directions and applications of the results presented here.

Chapter 8

Conclusion and Future Work

In this thesis, we studied information-theoretic metrics and their application to security and privacy. The results introduced here lie in the intersection of information theory, estimation theory, and cryptography. We presented converse bounds that shed light on the fundamental limits of what can or cannot be learned from a noisy observation of a hidden variable. In the context of security and privacy, where the hidden variable represents sensitive information, these bounds provide strong security guarantees: regardless of the computational resources available to the adversary, he will not be able to estimate certain properties of the sensitive information with estimation error smaller than the proposed bounds. We then used these bounds to both evaluate and design cryptographic and privacy-assuring systems.

As a first step, we studied the information-theoretic security properties of symmetric-key encryption schemes when the rate of the key is smaller than the rate of the message. We demonstrated how, in this case, the adversary's uncertainty corresponds to a list of all possible plaintext messages that could have been generated from the observed ciphertext. We studied properties of the uncertainty list through a source-coding framework called list-source codes, and introduced fundamental performance limits of LSCs. These limits, in turn, characterize the best trade-off between key length and the adversary's uncertainty list. We then argued that the length of the list is insufficient as a security metric, and introduced a new information-theoretic security metric called symbol secrecy and associated results. Symbol secrecy quantifies the adversary's uncertainty about individual symbols of the message, and encompasses other well studied information-theoretic security metrics.

Second, we extended symbol secrecy to the functional setting through a rate-distortion framework. This was done by making the key assumption that a set of reference functions $\{f_i(X)\}_{i=1}^k$ of the hidden variable X are known to be easy or hard to estimate given an observed variable Y. If the correlations between a target function $f(X)$ and the reference functions are known, we can then bound the estimation error of $f(X)$ given Y. When X is the plaintext, Y the ciphertext, and $\{f_i(X)\}_{i=1}^k$ represent individual symbols of a plaintext message, then this bound was combined with Fourier-analytic tools to characterize the classes of functions of X that cannot be reliably estimated when high symbol-secrecy is achieved.

Third, we introduced a general framework, grounded on rate-distortion formulations, to transform security guarantees in terms of an information metrics between X and Y (i.e. $\mathcal{I}(X;Y) \leq \theta$) into bounds on estimation metrics (i.e. $P_e(X|Y) \geq e_{\mathcal{I}}(p_X, \theta)$). The recipe here is simple: given $\mathcal{I}(X;Y) \leq \theta$, we minimize $P_e(X|Y)$ over all possible distribution $p_{X,Y}$. The result of this optimization problem is the error-rate function $e_{\mathcal{I}}(p_X, \theta)$. We presented results on the extremal properties of $e_{\mathcal{I}}(p_X, \theta)$, and, under certain technical assumptions, show how to extend the error-rate function to bound the probability of error of estimating functions of X.

Fourth, we presented and characterized several properties of the principal inertia components of $p_{X,Y}$. The PICs are powerful information-theoretic metrics that provide both (i) a measure of dependence between X and Y, and (ii) a complete characterization of which functions of X can be reliably estimated given an observation of Y. We derive lower bounds on $P_e(X|Y)$ based on the PICs. In particular, we show that the largest PIC (equivalently, the maximal correlation $\rho_m(X;Y)$) plays a key role in estimation:

$$\mathsf{Adv}(f(X)|Y) \leq \rho_m(X;Y),$$

i.e. the advantage over a random guess of estimating any function of X given Y is at most $\rho_m(X;Y)$. We also present bounds for the PICs for the distribution between a symmetric function of a sequence of i.i.d. random variables Y^n and a subsequence $Y^m \subseteq Y^n$:

$$\lambda_k(f(Y^n), Y^m) \leq \frac{(m)_j}{(n)_j},$$

where j depends on $|\mathcal{Y}|$ and k.

Finally, we analyzed the connection between the PICs and other information-theoretic metrics for security. We show that the PICs play a key role in estimating one-bit functions of a channel input given a channel output, and partially resolve the "most informative one-bit function" conjecture. We also illustrated several applications of the PICs to privacy. In this setting, we proved that perfect privacy can be achieved if and only if the smallest PIC is zero.

While this is the final chapter of the thesis, we are excited that the results presented here point towards several promising future research directions. Similarly to what happened in the field of communication, we believe that the study of problems in security, privacy and statistical learning through the information-theoretic lens can provide powerful practical insight. This insight, in turn, can serve as a design driver for the systems that will enable us to face the data challenges of the future: security, privacy, distributed data processing, content distribution and beyond. We outline next a few future potentials and applications of the topics presented in this thesis.

Data-driven Privacy

There are a multitude of algorithms for data analytics. These algorithms usually seek to discover structure and patterns in the data, which are then used for deriving utility. However, there are few attempts in turning these algorithms around in order to identify and mitigate possible security and privacy threats. One promising direction of research is to apply and adapt standard machine learning tools in order to develop data-driven privacy and anonymization methods. For example, by identifying correlations using principle component or correspondence analysis, it may be possible to pinpoint how sensitive information is related to data that will be disclosed. Alternatively, if standard statistical tests indicate that the points within a dataset are approximately independent and identically distributed, then subsampling can be used as a powerful privacy mechanism with provable privacy guarantees using the results derived in this thesis.

The steps for data-driven privacy methods are closely related to the ones taken in this thesis with the PICs: (i) identify meaningful statistical metrics that can be reliably estimated from data, (ii) derive converse bounds on estimation based on these metrics, and (iii) perform statistical analysis on the input and output of a privacy-assuring system, and, based on the theoretical bounds, provide design feedback. The ultimate goal is to develop efficient privacy algorithms that can leverage the wide availability of data in order to achieve a good trade-off between privacy and utility. Furthermore, by developing converse results similar to the ones introduced in this work, it is possible to provide strong privacy guarantees for such data-driven algorithms. These results have the same flavor as the ones found in universal source coding in information theory.

Interactive Setting

The setting considered in this thesis is non-interactive, in the sense that one party transforms a variable X into the output of a security system Y. An untrusted party then observes Y and attempts to estimate properties of X. In practice, security systems are continuously releasing data, and a malicious adversary may interact with the system in order to gain an advantage in estimating X.

A related line of work on information-theoretic security in an interactive setting was done in the context of Guesswork by Christiansen *et al.* [111–114]. Here, the authors consider the scenario where the adversary can repeatedly probe an oracle with questions about the secret variable X. This captures, for example, the security threat incurred by an adversary that attempts to gain access to a system by guessing a secret string (e.g. a password) several times. We believe that a challenging, yet promising extension of this thesis is the study of the properties of the security metrics introduced here, and in particular the PICs, in this interactive setting.

A General Methodological Framework for Privacy and Anonymity

Privacy threats occur at different interfaces of the data collection, processing and distribution pipeline. Consequently, privacy metrics and mechanisms are usually tailor-made for each specific interface. For example, differential privacy applies to statistical databases (e.g. [11, 12]), k-anonymity applies to the release of datasets (e.g. [10]), risk disclosure methods apply to statistical datasets (e.g., [115–117]), and noise addition applies in datamining (e.g. [118]). This has led to a shattered landscape of privacy models and metrics without a common underlying theory. Nevertheless, these problems are connected, and we are convinced that they can be analyzed through a common methodological lens by using the results introduced in this thesis and, more broadly, tools from information theory, estimation theory and computational statistics.

It would be of both practical and theoretical interest to create a unified framework for studying privacy and anonymity. The results in this thesis focus mostly on privacy. However, anonymity is another major challenge in today's data-driven world. Quantifying and mitigating de-anonymization risks is crucial for storing and releasing datasets for research. This is particularly difficult in large datasets, since high-dimensional data is inherently non-anonymous. For example, most Americans are uniquely identified by their gender, zip code and birth date, [10], and even a subset of a user's movie ratings can serve as a unique identifier [119]. We believe that information-theoretic metrics combined source-coding constructions (e.g. Huffman codes) can be used to create simple yet powerful methods for anonymizing large datasets.

Distributed Data Storage and Processing

Processing Big Data requires many distributed servers, running in parallel, in order to fetch, organize and analyze information. However, making data available across several nodes simultaneously for parallel processing (such as in a MapReduce setting) also presents new reliability and security challenges. Security systems in this computational-intensive, distributed framework must be sufficiently light-weight in order not to hinder performance and scalability. Furthermore, time-sensitive computations should be assigned to nodes with efficiency and reliability in mind. If the available processing power is insufficient, it might also be necessary, for example, to rely on the public cloud for certain computations (e.g. Amazon EC2), further aggravating reliability and security concerns. The fact that the data involved in such systems, due to their considerable size, are often understood only in terms of general statistical parameters, renders the problem particularly challenging.

We believe that many of the results presented here can be applied to develop theory and methods for distributed processing of statistical data. The converse bounds on estimation, and specifically the results in Section 5.8, can be used to study how to assign storage and computation tasks in face of the heterogeneous reliability, performance and security properties of different nodes in the system. Methodologically speaking, theoretical tools

such as the PICs are particularly well suited for quantifying the loss of precision of the final computation when a node fails in a distributed computing setting. The same tools can also be used to quantify the security threat posed if one of the processing nodes is attacked. By extending this approach, it may be possible to develop a Shannon-like, asymptotic theory for information processing in distributed systems with noisy components. This theory will lead to a crisp understanding of the trade-offs involved when acquiring, processing, securing and storing data.

Symmetric-key Encryption Schemes with Provable Information-theoretic Security Properties

In Chapter 2, we presented symmetric-key encryption constructions that provide a provable level of symbol secrecy under certain assumptions on the source distribution. These constructions still need to be significantly improved in order to meet modern cryptographic standards [8] (cf. probabilistic encryption [23]). Nevertheless, the proposed constructions complement other symmetric-key ciphers by adding an additional layer of security with provable information-theoretic properties.

The tools presented in this thesis, and in particular the LSC constructions and the PIC-based analysis, can, at least in theory, help guide the design of S-boxes in modern AES-based ciphers [8]. In addition, the metrics presented here add a new, information-theoretic dimension for evaluating encryption schemes used in practice. The design of practical symmetric-key encryption schemes that meet modern cryptographic standards and simultaneously guarantee a provable level of information-theoretic security is a promising research direction.

Open Questions

Finally, we introduced in this thesis two conjectures that remain unresolved. The proof of Conjecture 4.1 would reveal a fundamental property of information metrics, and enable the application of the method introduced in Section 4.4 to a broader setting. Furthermore, the proof of Conjecture 6.1 would shed light on the performance limits of binary classification for a wide range of probability distributions. We restate the conjectures below, presenting a slightly more general version of Conjecture 6.1.

Conjecture 4.1. *Let $\mathcal{I}(X;Y)$ be an information metric (cf. Definition 4.1), and let $e_{\mathcal{I}}$ be the error-rate function of $\mathcal{I}$ (cf. Definition 4.3). Then for a fixed $\theta \geq 0$, $e_{\mathcal{I}}(p_X, \theta)$ is Schur-concave in p_X.*

Conjecture 6.1. (Restated) *Let X and Y be two discrete random variables with finite support, where X and Y are uniformly distributed. We assume that joint distribution matrix $\mathbf{P}$ corresponding to $p_{X,Y}$ is symmetric ($\mathbf{P} = \mathbf{P}^T$). Then, using the PIC decomposition in Definition 7.3, $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T/|\mathcal{X}|$, where $\mathbf{\Sigma} = \operatorname{diag}\left(1, \sqrt{\lambda_1(X;Y)}, \ldots, \sqrt{\lambda_d(X;Y)}\right)$, $d = |\mathcal{X}|-1$.*

We now define a new random variable $\widetilde{Y}$ produced by making all the PICs of $p_{X,Y}$ equal to the largest one. Consequently, X and $\widetilde{Y}$ will have a joint distribution $p_{X,\widetilde{Y}}$ with corresponding joint distribution matrix $\mathbf{P}' = \mathbf{U}\widetilde{\mathbf{\Sigma}}\mathbf{V}^T/|\mathcal{X}|$, where $\widetilde{\mathbf{\Sigma}} = \mathrm{diag}\left(1, \sqrt{\lambda_1(X;Y)}, \ldots, \sqrt{\lambda_1(X;Y)}\right)$. Then, for any function $b : \mathcal{X} \to \{0,1\}$, $I(b(X);Y) \leq I(b(X);\widetilde{Y})$.

Appendix A

Proof of Lemma 3.1

For fixed $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ where $a_i > 0$ and $b_i \geq 0$, let $z_P : \mathbb{R}^n \to \mathbb{R}$ and $z_D : \mathbb{R}^n \to \mathbb{R}$ be given by

$$z_P(\mathbf{y}) \triangleq \mathbf{a}^T\mathbf{y},$$
$$z_D(\mathbf{u}) \triangleq \mathbf{a}^T\mathbf{b} + \mathbf{u}^T\mathbf{b} + \|\mathbf{u}\|_2.$$

Furthermore, we define $\mathcal{A}(\mathbf{a}) \triangleq \{\mathbf{u} \in \mathbb{R}^n | \mathbf{u} \geq \mathbf{a}\}$ and $\mathcal{B}(\mathbf{b}) \triangleq \{\mathbf{y} \in \mathbb{R}^n \mid \|\mathbf{y}\|_2 \leq 1, \mathbf{y} \leq \mathbf{b}\}$.

The optimal value $z_n(\mathbf{a}, \mathbf{b})$ is given by the following pair of primal-dual convex programs:

$$z_n(\mathbf{a}, \mathbf{b}) = \max_{\mathbf{y} \in \mathcal{B}(\mathbf{b})} z_P(\mathbf{y}) = \min_{\mathbf{u} \in \mathcal{A}(\mathbf{a})} z_D(\mathbf{u}).$$

Assume, without loss of generality, that $b_1/a_1 \leq b_2/a_2 \leq \ldots \leq b_n/a_n$, and let k^* be defined in (3.3).

Let $c_j \triangleq \sqrt{\frac{(1-\sum_{i=1}^{j} b_i^2)}{\|\mathbf{a}\|_2^2 - \sum_{i=1}^{j} a_i^2}}$. Note that since $\sum_{i=1}^{k^*} b_i^2 < 1$, we have $c_{k^*} > 0$. In addition, let

$$\mathbf{y}^* = (b_1, \ldots, b_{k^*}, a_{k^*+1}c_{k^*}, \ldots, a_n c_{k^*})$$

and

$$\mathbf{u}^* = (-b_1/c_{k^*}, \ldots, -b_{k^*}/c_{k^*}, -a_{k^*+1}, \ldots, -a_n).$$

From the definition of k^*, $\mathbf{y}^* \in \mathcal{B}(\mathbf{b})$ and $\mathbf{u}^* \in \mathcal{A}(\mathbf{a})$. Furthermore,

$$\begin{aligned} z_P(\mathbf{y}^*) &= \mathbf{a}^T\mathbf{y}^* \\ &= \sum_{i=1}^{k^*} a_i b_i + \sum_{i=k^*+1}^{n} c_{k^*} a_i^2 \\ &= \sum_{i=1}^{k^*} a_i b_i + \sqrt{\left(\|\mathbf{a}\|_2^2 - \sum_{i=1}^{k^*} a_i^2\right)\left(1 - \sum_{i=1}^{k^*} b_i^2\right)}, \end{aligned} \tag{A.1}$$

and

$$
\begin{aligned}
z_D(\mathbf{u}^*) =& \mathbf{a}^T\mathbf{b} + \mathbf{u}^{*T}\mathbf{b} + \|\mathbf{u}^*\|_2 \\
=& \sum_{i=1}^{k^*} \left(a_i b_i - \frac{b_i^2}{c_{k^*}} \right) \\
& + c_{k^*}^{-1} \sqrt{\sum_{i=1}^{k^*} b_i^2 + c_{k^*}^2 \left(\|\mathbf{a}\|_2^2 - \sum_{i=1}^{k^*} a_i^2 \right)} \\
=& \sum_{i=1}^{k^*} a_i b_i + c_{k^*}^{-1} \left(1 - \sum_{i=1}^{k^*} b_i^2 \right) \\
=& \sum_{i=1}^{k^*} a_i b_i + \sqrt{\left(\|\mathbf{a}\|_2^2 - \sum_{i=1}^{k^*} a_i^2 \right) \left(1 - \sum_{i=1}^{k^*} b_i^2 \right)} \\
=& z_P(\mathbf{y}^*).
\end{aligned}
$$

Since both the primal and the dual achieve the same value at $\mathbf{y}^*$ and $\mathbf{u}^*$, respectively, it follows that the value $z_P(\mathbf{y}^*)$ given in (A.1) is optimal.

Appendix B

Proof of PIC Error Bound

B.1 Proof of Theorem 5.4

Theorem 5.4 follows directly from the next two lemmas.

Lemma B.1. *Let the marginal distribution* $\mathbf{p}_X$ *and the PICs* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ *be given, where* $d = m - 1$. *Then for any* $p_{X,Y} \in \mathcal{R}(\mathbf{p}_X, \boldsymbol{\lambda})$, $0 \leq \alpha \leq 1$ *and* $0 \leq \beta \leq p_X(2)$

$$P_e(X|Y) \geq 1 - \beta - \sqrt{f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) + \sum_{i=1}^{m} \left([p_X(i) - \beta]^+\right)^2},$$

where

$$\begin{aligned} f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) = &\sum_{i=2}^{d+1} p_X(i)(\lambda_{i-1} + c_i - c_{i-1}) \\ &+ p_X(1)(c_1 + \alpha) - \alpha \mathbf{p}_X^T \mathbf{p}_X \ , \end{aligned} \tag{B.1}$$

and $c_i = [\lambda_i - \alpha]^+$ *for* $i = 1, \ldots, d$ *and* $c_{d+1} = 0$.

Proof. Let X and Y have a joint distribution matrix $\mathbf{P}$ with marginal p_X and principal inertias individually bounded by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$. We assume without loss of generality that $d = m - 1$, where $|\mathcal{X}| = m$. This can always be achieved by adding inertia components equal to 0.

Consider $X \to Y \to \hat{X}$, where $\hat{X}$ is the estimate of X from Y. The mapping from Y to $\hat{X}$ can be described without loss of generality by a $|\mathcal{Y}| \times |\mathcal{X}|$ row stochastic matrix, denoted by $\mathbf{F}$, where the (i, j)-th entry is the probability $p_{\hat{X}|Y}(j|i)$. The probability of correct estimation P_c is then

$$P_c = 1 - P_e(X|Y) = \operatorname{tr}\left(\mathbf{P}_{X,\hat{X}}\right),$$

where $\mathbf{P}_{X,\hat{X}} \triangleq \mathbf{PF}$.

The matrix $\mathbf{P}_{X,\hat{X}}$ can be decomposed according to (5.10), resulting in

$$P_c = \mathrm{tr}\left(\mathbf{D}_X^{1/2}\mathbf{U}\widetilde{\mathbf{\Sigma}}\mathbf{V}^T\mathbf{D}_{\hat{X}}^{1/2}\right) = \mathrm{tr}\left(\widetilde{\mathbf{\Sigma}}\mathbf{V}^T\mathbf{D}_{\hat{X}}^{1/2}\mathbf{D}_X^{1/2}\mathbf{U}\right), \tag{B.2}$$

where

$$\begin{aligned}
\mathbf{U} &= \left[\mathbf{p}_X^{1/2} \;\; \mathbf{u}_2 \;\cdots\; \mathbf{u}_m\right], \\
\mathbf{V} &= \left[\mathbf{p}_{\hat{X}}^{1/2} \;\; \mathbf{v}_2 \;\cdots\; \mathbf{v}_m\right], \\
\widetilde{\mathbf{\Sigma}} &= \mathrm{diag}\left(1, \widetilde{\lambda}_1^{1/2}, \ldots, \widetilde{\lambda}_d^{1/2}\right), \\
\mathbf{D}_{\hat{X}} &= \mathrm{diag}\left(\mathbf{p}_{\hat{X}}\right),
\end{aligned}$$

and $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ are orthogonal matrices. The probability of correct detection can be written as

$$\begin{aligned}
P_c &= \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \sum_{k=2}^{m}\sum_{i=1}^{m}\left(\widetilde{\lambda}_{k-1}p_X(i)p_{X'}(i)\right)^{1/2}u_{k,i}v_{k,i} \\
&= \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \sum_{k=2}^{m}\sum_{i=1}^{m}\widetilde{\lambda}_{k-1}^{1/2}\widetilde{u}_{k,i}\widetilde{v}_{k,i}
\end{aligned}$$

where $u_{k,i} = [\mathbf{u}_k]_i$, $v_{k,i} = [\mathbf{v}_k]_i$, $\widetilde{u}_{k,i} = p_X(i)u_{k,i}$ and $\widetilde{v}_{k,i} = p_{X'}(i)v_{k,i}$. Applying the Cauchy-Schwarz inequality twice, we obtain

$$\begin{aligned}
P_c &\le \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \sum_{i=1}^{m}\left(\sum_{k=2}^{m}\widetilde{v}_{k,i}^2\right)^{1/2}\left(\sum_{k=2}^{m}\widetilde{\lambda}_{k-1}\widetilde{u}_{k,i}^2\right)^{1/2} \\
&= \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \sum_{i=1}^{m}\left(p_{X'}(i)(1-p_{X'}(i))\sum_{k=2}^{m}\widetilde{\lambda}_{k-1}\widetilde{u}_{k,i}^2\right)^{1/2} \\
&\le \mathbf{p}_X^T\mathbf{p}_{\hat{X}} + \left(1-\sum_{i=1}^{m}p_{X'}^2(i)\right)^{1/2}\left(\sum_{i=1}^{m}\sum_{k=2}^{m}\widetilde{\lambda}_{k-1}\widetilde{u}_{k,i}^2\right)^{1/2}
\end{aligned} \tag{B.3}$$

Let $\overline{\mathbf{U}} = [\mathbf{u}_2 \cdots \mathbf{u}_m]$ and $\mathbf{\Sigma} = \mathrm{diag}\left(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_d\right)$. Then

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{k=2}^{m}\widetilde{\lambda}_{k-1}\widetilde{u}_{k,i}^2 &= \mathrm{tr}\left(\mathbf{\Sigma}\overline{\mathbf{U}}^T\mathbf{D}_X\overline{\mathbf{U}}\right) \\
&\le \sum_{k=1}^{d}\sigma_k\widetilde{\lambda}_k, \\
&\le \sum_{k=1}^{d}\sigma_k\lambda_k.
\end{aligned} \tag{B.4}$$

where σ_k are the singular values of $\overline{\mathbf{U}}^T \mathbf{D}_X \overline{\mathbf{U}}$. The first inequality follows from the application of Von-Neumman's trace inequality and the fact that $\overline{\mathbf{U}}^T \mathbf{D}_X \overline{\mathbf{U}}$ is symmetric and positive semi-definite. The second inequality follows by observing that the principal inertias satisfy the data processing inequality and, therefore, $\widetilde{\lambda}_k \leq \lambda_k$.

We will now find an upper bound for (B.4) by bounding the eigenvalues σ_k. First, note that $\overline{\mathbf{U}}\ \overline{\mathbf{U}}^T = I - \mathbf{p}_X^{1/2} \left(\mathbf{p}_X^{1/2}\right)^T$ and consequently

$$\begin{aligned} \sum_{k=1}^{d} \sigma_k &= \mathrm{tr}\left(\overline{\mathbf{U}}^T \mathbf{D}_X \overline{\mathbf{U}}\right) \\ &= \mathrm{tr}\left(\mathbf{D}_X \left(I - \mathbf{p}_X^{1/2} \left(\mathbf{p}_X^{1/2}\right)^T\right)\right) \\ &= 1 - \sum_{i=1}^{m} p_X^2(i) \ . \end{aligned} \tag{B.5}$$

Second, note that $\overline{\mathbf{U}}^T \mathbf{D}_X \overline{\mathbf{U}}$ is a principal submatrix of $\mathbf{U}^T \mathbf{D}_X \mathbf{U}$, formed by removing the first row and columns of $\mathbf{U}^T \mathbf{D}_X \mathbf{U}$. It then follows from Cauchy's interlacing theorem that

$$p_X(m) \leq \sigma_{m-1} \leq p_X(m-1) \leq \ldots \leq p_X(2) \leq \sigma_1 \leq p_X(1). \tag{B.6}$$

Combining the restriction (B.5) and (B.6), an upper bound for (B.4) can be found by solving the following linear program

$$\begin{aligned} \max_{\sigma_i} \quad & \sum_{i=1}^{d} \lambda_i \sigma_i \\ \text{subject to} \quad & \sum_{i=1}^{d} \sigma_i = 1 - \mathbf{p}_X^T \mathbf{p}_X, \\ & p_X(i+1) \leq \sigma_i \leq p_X(i), \ i = 1, \ldots, d \ . \end{aligned} \tag{B.7}$$

Let $\delta_i \triangleq p_X(i) - p_X(i+1)$ and $\gamma_i \triangleq \lambda_i p_X(i+1)$. The dual of (B.7) is

$$\begin{aligned} \min_{y_i, \alpha} \quad & \alpha \left(p_X(1) - \mathbf{p}_X^T \mathbf{p}_X\right) + \sum_{i=1}^{m-1} \delta_i y_i + \gamma_i \\ \text{subject to} \quad & y_i \geq [\lambda_i - \alpha]^+ \ , \ i = 1, \ldots, d \ . \end{aligned} \tag{B.8}$$

For any given value of α, the optimal values of the dual variables y_i in (B.8) are

$$y_i = [\lambda_i - \alpha]^+ = c_i, \ i = 1, \ldots, d \ .$$

Therefore the linear program (B.8) is equivalent to

$$\min_{\alpha} f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}), \tag{B.9}$$

where $f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda})$ is defined in the statement of the theorem.

Denote the solution of (B.7) by $f_P^*(\mathbf{p}_X, \boldsymbol{\lambda})$ and of (B.8) by $f_D^*(\mathbf{p}_X, \boldsymbol{\lambda})$. It follows that (B.4) can be bounded

$$\begin{aligned} \sum_{k=1}^{d} \sigma_k \lambda_k &\leq f_P^*(\mathbf{p}_X, \boldsymbol{\lambda}) \\ &= f_D^*(\mathbf{p}_X, \boldsymbol{\lambda}) \\ &\leq f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) \ \forall \ \alpha \in \mathbb{R}. \end{aligned} \tag{B.10}$$

We may consider $0 \leq \alpha \leq 1$ in (B.10) without loss of generality.

Using (B.10) to bound (B.3), we find

$$P_c \leq \mathbf{p}_X^T \mathbf{p}_{\hat{X}} + \left[f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) \left(1 - \sum_{i=1}^{m} p_{X'}^2(i) \right) \right]^{1/2} \tag{B.11}$$

The previous bound can be maximized over all possible output distributions $p_{X'}$ by solving:

$$\begin{aligned} \max_{x_i} \quad & \left[f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) \left(1 - \sum_{i=1}^{m} x_i^2 \right) \right]^{1/2} + \sum_{i=1}^{m} p_X(i) x_i \\ \text{subject to} \quad & \sum_{i=1}^{m} x_i = 1, \\ & x_i \geq 0, i = 1, \ldots, m \ . \end{aligned} \tag{B.12}$$

The dual function of (B.12) over the additive constraint is

$$\begin{aligned} L(\beta) &= \max_{x_i \geq 0} \ \beta + \left[f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) \left(1 - \sum_{i=1}^{m} x_i^2 \right) \right]^{1/2} \\ &\qquad + \sum_{i=1}^{m} (p_X(i) - \beta) x_i \\ &= \beta + \sqrt{f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) + \sum_{i=1}^{m} \left([p_X(i) - \beta]^+ \right)^2}. \end{aligned} \tag{B.13}$$

Since $L(\beta)$ is an upper bound of (B.12) for any β and, therefore, is also an upper bound of

(B.11), then

$$P_c \leq \beta + \sqrt{f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) + \sum_{i=1}^{m} \left([p_X(i) - \beta]^+\right)^2}. \tag{B.14}$$

Note that we can consider $0 \leq \beta \leq p_X(2)$ in (B.14), since $L(\beta)$ is increasing for $\beta > p_X(2)$. Taking $P_e(X|Y) = 1 - P_c$, the result follows. □

The next result tightens the bound introduced in lemma B.1 by optimizing over all values of α.

Lemma B.2. *Let* $f_0^*(\mathbf{p}_X, \boldsymbol{\lambda}) \triangleq \min_\alpha f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda})$ *and* k^* *be defined as in* (3.3). *Then*

$$\begin{aligned} f_0^*(\mathbf{p}_X, \boldsymbol{\lambda}) = &\sum_{i=1}^{k^*} \lambda_i p_X(i) + \sum_{i=k^*+1}^{m} \lambda_{i-1} p_X(i) \\ &- \lambda_{k^*} \mathbf{p}_X^T \mathbf{p}_X \ , \end{aligned} \tag{B.15}$$

where $\lambda_m = 0$.

Proof. Let $\mathbf{p}_X$ and $\boldsymbol{\lambda}$ be fixed, and $\lambda_k \leq \alpha \leq \lambda_{k-1}$, where we define $\lambda_0 = 1$ and $\lambda_m = 0$. Then $c_i = \lambda_i - \alpha$ for $1 \leq i \leq k-1$ and $c_i = 0$ for $k \leq i \leq d$ in (B.1). Therefore

$$\begin{aligned} f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda}) = &\sum_{i=1}^{k-1} \lambda_i p_X(i) + \alpha p_X(k) \\ &+ \sum_{i=k+1}^{m} \lambda_{i-1} p_X(i) - \alpha \mathbf{p}_X^T \mathbf{p}_X \end{aligned} \tag{B.16}$$

Note that (B.16) is convex in α, and is decreasing when $p_X(k) - \mathbf{p}_X^T \mathbf{p}_X \leq 0$ and increasing when $p_X(k) - \mathbf{p}_X^T \mathbf{p}_X \geq 0$. Therefore, $f_0(\alpha, \mathbf{p}_X, \boldsymbol{\lambda})$ is minimized when $\alpha = \lambda_k$ such that $p_X(k) \geq \mathbf{p}_X^T \mathbf{p}_X$ and $p_X(k-1) \leq \mathbf{p}_X^T \mathbf{p}_X$. If $p_X(k) - \mathbf{p}_X^T \mathbf{p}_X \geq 0$ for all k (i.e. p_X is uniform), then we can take $\alpha = 0$. The result follows. □

Bibliography

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[2] ——, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

[3] R. Price, "A conversation with Claude Shannon: one man's approach to problem solving," *Cryptologia*, vol. 9, no. 2, pp. 167–175, Apr. 1985.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*, 2nd ed. Wiley-Interscience, Jul. 2006.

[5] E. M. Guizzo, "The essential message : Claude Shannon and the making of information theory," Thesis, Massachusetts Institute of Technology, 2003.

[6] "Claude E. Shannon: An interview conducted by Robert Price, 28th of July, 1982," IEEE History Center, Interview #423, transcript availabe at http://ethw.org/Oral-History:Claude_E._Shannon.

[7] Y. Liang, H. V. Poor, and S. Shamai (Shitz), "Information theoretic security," *Found. Trends Commun. Inf. Theory*, vol. 5, pp. 355–580, Apr. 2009.

[8] J. Katz and Y. Lindell, *Introduction to Modern Cryptography: Principles and Protocols*, 1st ed. Chapman and Hall/CRC, Aug. 2007.

[9] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AML-Book, Mar. 2012.

[10] L. Sweeney, "K-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[11] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*, 2006.

[12] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*. Springer, 2006, vol. 4052, pp. 1–12.

[13] S. Salamatian, A. Zhang, F. P. Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data," *IEEE GlobalSIP*, 2013.

[14] S. Salamatian, A. Zhang, F. P. Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "Managing your private and public data: Bringing down inference attacks against your privacy," *IEEE J. Sel. Topics Signal Process.*, October 2015.

[15] S. Bhamidipati, N. Fawaz, B. Kveton, and A. Zhang, "PriView: Personalized Media Consumption Meets Privacy against Inference Attacks," *IEEE Software*, vol. 32, no. 4, pp. 53–59, Jul. 2015.

[16] M. J. Greenacre, *Theory and Applications of Correspondence Analysis.* Academic Pr, Mar. 1984.

[17] L. Breiman and J. H. Friedman, "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 580–598, Sep. 1985.

[18] F. P. Calmon and N. Fawaz, "Privacy against statistical inference," in *Proc. 50th Ann. Allerton Conf. Commun., Contr., and Comput.*, 2012, pp. 1401–1408.

[19] A. Zhang, S. Bhamidipati, N. Fawaz, and B. Kveton, "PriView: Media Consumption and Recommendation Meet Privacy Against Inference Attacks," in *IEEE Web 2.0 Security and Privacy Workshop*, 2014.

[20] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Medard, "From the information bottleneck to the privacy funnel," in *IEEE Inf. Theory Workshop (ITW)*, 2014, pp. 501–505.

[21] A. C.-C. Yao, "Protocols for secure computations," in *FOCS*, vol. 82, 1982, pp. 160–164.

[22] R. Roth, *Introduction to Coding Theory.* Cambridge, UK ; New York: Cambridge University Press, Mar. 2006.

[23] S. Goldwasser and S. Micali, "Probabilistic encryption," *Journal of Computer and System Sciences*, vol. 28, no. 2, pp. 270–299, Apr. 1984.

[24] G. R. Kumar and T. A. Courtade, "Which boolean functions maximize information of noisy inputs?" *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4515–4525, Aug. 2014.

[25] M. Hellman, "An extension of the Shannon theory approach to cryptography," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 289–294, May 1977.

[26] R. E. Blahut, D. J. Costello, U. Maurer, and T. Mittelholzer, Eds., *Communications and Cryptography: Two Sides of One Tapestry*, 1st ed. Springer, Jun. 1994.

[27] F. P. Calmon, M. Médard, L. Zeger, J. Barros, M. M. Christiansen, and K. R. Duffy, "Lists that are smaller than their parts: A coding approach to tunable secrecy," in *Proc. 50th Annual Allerton Conf. on Commun., Control, and Comput.*, 2012.

[28] F. P. Calmon, M. Médard, M. Varia, K. R. Duffy, M. M. Christiansen, and L. M. Zeger, "Hiding Symbols and Functions: New Metrics and Constructions for Information-Theoretic Security," *arXiv:1503.08513 [cs, math]*, Mar. 2015.

[29] R. Ahlswede, "Remarks on shannon's secrecy systems," *Problems of Control and Inf. Theory*, vol. 11, no. 4, 1982.

[30] S.-C. Lu, "The existence of good cryptosystems for key rates greater than the message redundancy (corresp.)," *IEEE Trans. Inf. Theory*, vol. 25, no. 4, pp. 475–477, Jul. 1979.

[31] ——, "Random ciphering bounds on a class of secrecy systems and discrete message sources," *IEEE Trans. Inf. Theory*, vol. 25, no. 4, pp. 405–414, Jul. 1979.

[32] ——, "On secrecy systems with side information about the message available to a cryptanalyst (corresp.)," *IEEE Trans. Inf. Theory*, vol. 25, no. 4, pp. 472–475, Jul. 1979.

[33] C. Schieler and P. Cuff, "Rate-distortion theory for secrecy systems," *IEEE Trans. Inf. Theory*, vol. PP, no. 99, 2014.

[34] L. Ozarow and A. Wyner, "Wire-tap channel II," in *Advances in Cryptology*, 1985, pp. 33–50.

[35] N. Cai and R. Yeung, "Secure network coding," in *Proc. IEEE Int. Symp. on Inf. Theory*, 2002.

[36] J. Feldman, T. Malkin, C. Stein, and R. A. Servedio, "On the capacity of secure network coding," in *Proc. 42nd Ann. Allerton Conf. Commun., Contr., and Comput.*, 2004.

[37] A. Mills, B. Smith, T. Clancy, E. Soljanin, and S. Vishwanath, "On secure communication over wireless erasure networks," in *Proc. IEEE Int. Symp. on Inf. Theory*, Jul. 2008, pp. 161 –165.

[38] S. El Rouayheb, E. Soljanin, and A. Sprintson, "Secure network coding for wiretap networks of type II," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1361 –1371, Mar. 2012.

[39] D. Silva and F. Kschischang, "Universal secure network coding via Rank-Metric codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 1124 –1135, Feb. 2011.

[40] L. Lima, M. Médard, and J. Barros, "Random linear network coding: A free cipher?" in *Proc. IEEE Int. Symp. on Inf. Theory*, Jun. 2007, pp. 546 –550.

[41] N. Cai and T. Chan, "Theory of secure network coding," *IEEE Proc.*, vol. 99, no. 3, pp. 421 –437, Mar. 2011.

[42] P. Oliveira, L. Lima, T. Vinhoza, J. Barros, and M. Médard, "Trusted storage over untrusted networks," in *IEEE Global Telecommunications Conference*, Dec. 2010, pp. 1 –5.

[43] P. Elias, "List decoding for noisy channels," Research Laboratory of Electronics, MIT, Technical Report 335, September 1957.

[44] J. M. Wozencraft, "List decoding," Research Laboratory of Electronics, MIT, Progress Report 48, 1958.

[45] C. Shannon, R. Gallager, and E. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. I," *Information and Control*, vol. 10, no. 1, pp. 65–103, Jan. 1967.

[46] ——, "Lower bounds to error probability for coding on discrete memoryless channels. II," *Information and Control*, vol. 10, no. 5, pp. 522–552, May 1967.

[47] G. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. 14, no. 2, pp. 206– 220, Mar. 1968.

[48] V. Guruswami, "List decoding of error-correcting codes," Thesis, MIT, Cambridge, MA, 2001.

[49] ——, "List decoding of binary Codes–A brief survey of some recent results," in *Coding and Cryptology*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2009, vol. 5557, pp. 97–106.

[50] M. Ali and M. Kuijper, "Source coding with side information using list decoding," in *Proc. IEEE Int. Symp. on Inf. Theory.* IEEE, Jun. 2010, pp. 91–95.

[51] A. D. Wyner, "The Wire-Tap Channel," *Bell System Technical Journal*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.

[52] U. Maurer and S. Wolf, "Information-Theoretic Key Agreement: From Weak to Strong Secrecy for Free," in *Advances in Cryptology (EUROCRYPT)*, ser. Lecture Notes in Computer Science, B. Preneel, Ed. Springer Berlin Heidelberg, 2000, no. 1807, pp. 351–368.

[53] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, Aug. 2011.

[54] T. Ho and D. Lun, *Network Coding: An Introduction.* New York: Cambridge University Press, Apr. 2008.

[55] L. Eschenauer and V. D. Gligor, "A key-management scheme for distributed sensor networks," in *Proceedings of the 9th ACM Conference on Computer and Communications Security*, ser. CCS '02. New York, NY, USA: ACM, 2002, pp. 41–47.

[56] F. P. Calmon, M. Varia, and M. Médard, "On information-theoretic metrics for symmetric-key encryption and privacy," in *Proc. 52nd Annual Allerton Conference on Communication, Control, and Computing*, 2014.

[57] H. Yamamoto, "Rate-distortion theory for the Shannon cipher system," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 827–835, May 1997.

[58] I. S. Reed, "Information Theory and Privacy in Data Banks," in *Proc. of the National Computer Conference and Exposition*, ser. AFIPS '73. New York, NY, USA: ACM, June 1973, pp. 581–587.

[59] A. Sarwate and L. Sankar, "A rate-disortion perspective on local differential privacy," in *Proc. 52nd Ann. Allerton Conf. Commun., Contr., and Comput.*, Sep. 2014, pp. 903–908.

[60] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, "From t-Closeness-Like Privacy to Postrandomization via Information Theory," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010.

[61] L. Sankar, S. Rajagopalan, and H. Poor, "Utility-Privacy Tradeoffs in Databases: An Information-Theoretic Approach," *IEEE Trans. on Inf. Forensics and Security*, vol. 8, no. 6, pp. 838–852, Jun. 2013.

[62] M. Bellare, S. Tessaro, and A. Vardy, "Semantic security for the wiretap channel," in *Advances in Cryptology - CRYPTO 2012*, ser. Lecture Notes in Comput. Sci. Springer, Jan. 2012, no. 7417, pp. 294–311.

[63] R. O'Donnell, "Some topics in analysis of boolean functions," in *Proc. 40th ACM Symp. on Theory of Computing*, 2008, pp. 569–578.

[64] F. P. Calmon, M. Varia, M. Médard, M. Christiansen, K. Duffy, and S. Tessaro, "Bounds on inference," in *Proc. 51st Ann. Allerton Conf. Commun., Contr., and Comput.*, Oct. 2013, pp. 567–574.

[65] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv:physics/0004057 [physics.data-an]*, Apr. 2000.

[66] V. Doshi, D. Shah, M. Médard, and M. Effros, "Functional compression through graph coloring," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3901 –3917, Aug. 2010.

[67] R. Ahlswede, "Extremal properties of rate distortion functions," *IEEE Trans. on Info. Theory*, vol. 36, no. 1, pp. 166–171, 1990.

[68] R. G. Gallager, *Information theory and reliable communication.* New York: Wiley, 1968.

[69] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: theory of majorization and its applications.* New York: Springer Series in Statistics, 2011.

[70] M. Greenacre and T. Hastie, "The geometric interpretation of correspondence analysis," *J. Am. Stat. Assoc.*, vol. 82, no. 398, pp. 437–447, Jun. 1987.

[71] H. O. Hirschfeld, "A connection between correlation and contingency," in *Proc. Cambridge Philos. Soc.*, vol. 31, 1935, pp. 520–524.

[72] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *ZAMM-Z. Angew. Math. Me.*, vol. 21, no. 6, pp. 364–379, 1941.

[73] O. Sarmanov, "Maximum correlation coefficient (nonsymmetric case)," *Selected Translations in Mathematical Statistics and Probability*, vol. 2, pp. 207–210, 1962.

[74] A. Rényi, "On measures of dependence," *Acta Math. Hung.*, vol. 10, no. 3-4, pp. 441–451, Sep. 1959.

[75] M. Greenacre, *Correspondence Analysis in Practice, Second Edition*, 2nd ed. Chapman and Hall/CRC, May 2007.

[76] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM J. on Appl. Math.*, vol. 28, no. 1, pp. 100–113, Jan. 1975.

[77] R. Ahlswede and P. Gacs, "Spreading of sets in product spaces and hypercontraction of the markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, Dec. 1976.

[78] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," arXiv e-print 1304.6133, Apr. 2013.

[79] Y. Polyanskiy, "Hypothesis testing via a comparator," in *Proc. 2012 IEEE Int. Symp. on Inf. Theory*, Jul. 2012, pp. 2206–2210.

[80] M. Raginsky, "Logarithmic Sobolev inequalities and strong data processing theorems for discrete channels," in *Proc. 2013 IEEE Int. Symp. on Inf. Theory*, Jul. 2013, pp. 419–423.

[81] A. Buja, "Remarks on Functional Canonical Variates, Alternating Least Squares Methods and Ace," *The Annals of Statistics*, vol. 18, no. 3, pp. 1032–1069, Sep. 1990.

[82] W. Kang and S. Ulukus, "A new data processing inequality and its applications in distributed source and channel coding," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 56–69, 2011.

[83] A. Guntuboyina, "Lower bounds for the minimax risk using -divergences, and applications," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2386–2399, 2011.

[84] A. Guntuboyina, S. Saha, and G. Schiebinger, "Sharp inequalities for f-divergences," *arXiv:1302.0336*, Feb. 2013.

[85] A. Orlitsky and J. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903 –917, Mar. 2001.

[86] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge University Press, Oct. 2012.

[87] K. Fan, "On a theorem of Weyl concerning eigenvalues of linear transformations I," *P. Natl. Acad. Sci. USA*, vol. 35, no. 11, pp. 652–655, Nov. 1949.

[88] M. L. Overton and R. S. Womersley, "On the sum of the largest eigenvalues of a symmetric matrix," *SIAM J. Matrix Anal. A.*, vol. 13, no. 1, pp. 41–45, Jan. 1992.

[89] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK; New York: Cambridge University Press, 2004.

[90] C. Deniau, G. Oppenheim, and J. P. Benzécri, "Effet de l'affinement d'une partition sur les valeurs propres issues d'un tableau de correspondance," *Cahiers de l'analyse des données*, vol. 4, no. 3, pp. 289–297.

[91] A. Guntuboyina, "Minimax lower bounds," Ph.D., Yale University, United States – Connecticut, 2011.

[92] B. Efron and C. Stein, "The jackknife estimate of variance," *The Annals of Statistics*, vol. 9, no. 3, pp. 586–596, May 1981.

[93] A. Dembo, A. Kagan, and L. A. Shepp, "Remarks on the maximum correlation coefficient," *Bernoulli*, vol. 7, no. 2, pp. 343–350, Apr. 2001.

[94] P. Diaconis and D. Freedman, "Finite exchangeable sequences," *Ann. Probab.*, vol. 8, no. 4, pp. 745–764, Aug. 1980.

[95] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and the mutual information between boolean functions," in *Proc. 51st Ann. Allerton Conf. Commun., Contr., and Comput.*, Oct. 2013, pp. 13–19.

[96] I. Csiszár, *Information Theory And Statistics: A Tutorial.* Now Publishers Inc, 2004.

[97] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. 48th Ann. Allerton Conf. Commun., Contr., and Comput.*, 2010, pp. 1327–1333.

[98] F. P. Calmon, M. Varia, and M. Médard, "An exploration of the role of principal inertia components in information theory," in *Information Theory Workshop (ITW), 2014 IEEE*, 2014, pp. 252–256.

[99] M. Raginsky, J. G. Silva, S. Lazebnik, and R. Willett, "A recursive procedure for density estimation on the binary hypercube," *Electron. J. Statist.*, vol. 7, pp. 820–858, 2013.

[100] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Ann. Allerton Conf. Commun., Contr., and Comput.*, 1999, pp. 368–377.

[101] F. P. Calmon, A. Makhdoumi, and M. Médard, "Fundamental limits of perfect privacy," *International Symp. on Info. Theory*, 2015.

[102] R. Tandon, L. Sankar, and H. Poor, "Discriminatory lossy source coding: Side information privacy," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5665–5677, Sep. 2013.

[103] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the twenty-second ACM Symposium on Principles of Database Systems*, New York, NY, USA, 2003, pp. 211–222.

[104] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *arXiv:1405.3629 [cs, math]*, May 2014.

[105] C. T. Li and A. E. Gamal, "Maximal correlation secrecy," *arXiv:1412.5374 [cs, math]*, Dec. 2014.

[106] S. Chakraborty, N. Bitouze, M. Srivastava, and L. Dolecek, "Protecting data against unwanted inferences," in *2013 IEEE Information Theory Workshop (ITW)*, Sep. 2013, pp. 1–5.

[107] S. Asoodeh, F. Alajaji, and T. Linder, "Notes on information-theoretic privacy," in *Proc. 52nd Ann. Allerton Conf. Commun., Contr., and Comput.*, Sep. 2014, pp. 1272–1278.

[108] S. Kamath and V. Anantharam, "Non-interactive simulation of joint distributions: The Hirschfeld-Gebelein-Rényi maximal correlation and the hypercontractivity ribbon," in *Proc. 50th Ann. Allerton Conf. Commun., Contr., and Comput.* IEEE, 2012, pp. 1057–1064.

[109] T. Berger and R. Yeung, "Multiterminal source encoding with encoder breakdown," *IEEE Trans. on Inf. Theory*, vol. 35, no. 2, pp. 237–244, Mar. 1989.

[110] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[111] M. M. Christiansen, K. R. Duffy, F. P. Calmon, and M. Médard, "Brute force searching, the typical set and Guesswork," in *Proc. 2013 IEEE Int. Symp. on Inf. Theory.* IEEE, 2013, pp. 1257–1261.

[112] M. M. Christiansen, K. R. Duffy, F. P. Calmon, and M. Médard, "Guessing a password over a wireless channel (on the effect of noise non-uniformity)," in *Asilomar Conference on Signals, Systems and Computers.* IEEE, 2013, pp. 51–55.

[113] M. M. Christiansen and K. R. Duffy, "Guesswork, large deviations, and Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 796–802, 2013.

[114] M. M. Christiansen, K. R. Duffy, F. P. Calmon, and M. Médard, "Quantifying the computational security of multi-user systems," *arXiv:1405.5024*, 2014.

[115] D. Denning, "Secure statistical databases with random sample queries," *ACM Trans. Database Sys.*, vol. 5, no. 3, pp. 291–315, 1980.

[116] L. Beck, "A security mechanism for statistical database," *ACM Trans. Database Syst.*, vol. 5, no. 3, pp. 316–338, 1980.

[117] J. Domingo-Ferrer, A. Oganian, and V. Torra, "Information-theoretic disclosure risk measures in statistical disclosure control of tabular data," in *Proc. 14th Intl. Conf. Scientific and Statistical Database Management.* IEEE Computer Society, 2002, pp. 227–231.

[118] D. Agrawal and C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. 20th Symp. Principles of Database Systems*, Santa Barbara, CA, May 2001.

[119] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *IEEE Symp. on Security and Privacy*, May 2008, pp. 111–125.