Information and Estimation Theoretic Approaches to Data Privacy

by

Shahab Asoodeh

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
May 2017

Abstract

Warner [145] in 1960s proposed a simple mechanism, now referred to as the randomized response model, as a remedy for what he termed evasive answer bias in survey sampling. The randomized response setting is as follows: n people participate in a survey and a statistician asks each individual a sensitive yes-no question and seeks to find the ratio of ”yes” responses. For privacy purposes, individuals are given a biased coin that comes up heads with probability $a \in (0, \frac{1}{2})$. Each individual flips the coin in private. If it comes up heads, they lie and if it comes up tails, they tell the truth. Warner derived a maximum likelihood unbiased estimator for the true ratio of ”yes” based on the reported responses. Thus the parameter of interest is estimated accurately while preserving the privacy of each user and avoiding survey answer bias.

In this thesis, we generalize Warner's randomized response model in several directions: (i) we assume that the response of each individual consists of private and non-private data and the goal is to generate a response which carries as much ”information” about the non-private data as possible while limiting the ”information leakage” about the private data, (ii) we propose mathematically well founded metrics to quantify the tradeoff between how much the response leaks about the private data and how much information it conveys about the non-private data, (iii) we make no assumptions on the alphabets of the private and non-private data, and (iv) we design optimal response mechanisms which achieve the

fundamental tradeoffs.

Unlike the large body of recent research on privacy which studied the problem of reducing disclosure risk, in this thesis we formulate and study the tradeoff between utility (e.g., statistical efficiency) and privacy (e.g., information leakage). Our approach (which is two-fold: information-theoretic and estimation-theoretic) and results shed light on the fundamental limits of the utility-privacy tradeoff.

To my family for all their love and support

Acknowledgments

First, I would like to thank my advisors Professor Fady Alajaji and Professor Tamás Linder for their generous support, encouragement and enlightening guidance throughout the development of this thesis. It has been an incredible experience being a student of Fady and Tamás the last five years.

Working with Fady has helped me appreciate the necessity of being a good scholar, the importance of strategy while devising research goals, and the need to constantly ask important questions. There was never a time that I did not feel cared for, thanks to his constant support and guidance. I am inspired by his passionate and ethical approaches for true research, and hope to have a similar theme in my future research plans. Just in one sentence: no PhD student can wish for a better advisor than Fady.

Working with Tamás has helped me understand the importance of rigor and clarity in understanding. His deep knowledge of many disciplines and crystal-clear intuitions for complex ideas are things I am only capable of aspiring for, but which I hope to inculcate in to my research style to the extent I can. Having taught a course with him, I have been greatly inspired by his teaching philosophy and lucid presentation style. His emphasis on simplicity and intuition, in teaching as well as in research, will continue to influence me in the years ahead.

It has become clear for me that having two advisors was a necessary thing, if only for their sakes, because it kept me out of the hair of the first one while I bothered the second. Fady and Tamás gave me freedom to come up with my own problems which, albeit hard at first, has forced me to read about many different problems in information theory and statistics. It would not have been possible for me to finish my PhD without their insightful discussions, advice, support, and comments.

I am also very thankful to Professors Glen Takahara, Ashish Khisti, and Mikhail Nediak for agreeing to serve on my thesis committee. I was fortunate to take a course and several seminars with Glen; he is one of the best instructors I have ever had. I am incredibly indebted to Professor Serdar Yüksel. Serdar has been my teacher, instructor, mentor, and more importantly, friend during my years at Queen's. He always had time for me when I needed some advise. Thank you Serdar.

I spent a pleasant 5 years in Jeffery Hall and had opportunity to meet good friends, in particular, Mario Diaz, Yinzheng Gu, Naci Saldi, and Saber Jafarpour. Yes, we did the stereotypical graduate student things, like hunting for free food, drinking unreasonable amounts of coffee, surviving exams and homeworks, and fighting 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 want to give a special thank you to Mario. We had many long discussions about different topics including politics, formula 1, religions, la leyenda del mal, and of course, mathematics. His problem-solving skills and passion for science always cheered me. Some of the results in this dissertation greatly benefited from his smart insights and ideas. Thank you Mario.

I also owe another big thank you to Jennifer. She is an amazing and incredibly caring person in Jeffery Hall who helped all graduate students with a nice smile. There is almost nothing that she cannot help a graduate student with. Thank you Jennifer.

And now I must go back to the beginning, where everything really started, with my family. My parents, Soheila and Nasrollah, are my two main role-models for pretty much everything in life. They encouraged me with their love and patience and gave me every learning opportunity they could think of. I am indebted to them beyond any words. My brothers, Ehsan and Hesam, have been very supportive during the times of my graduate studies away from home.

Contents

List of Figures

Chapter 1

Introduction

With the emergence of modern techniques for data collection – arising from medicine and bioinformatics [97], internet applications such as web search engines [72], social networks, physics and astronomical experiments [73], and mobile data gathering platforms – which has led to a proliferation of large datasets, the need for privacy has become paramount.

As the size of datasets increases with the concomitant amount of information we collect about individuals, it has become more important to maintain privacy of individuals. Statistical studies on privacy date back to Warner's 1960s work on randomized response and survey sampling [145]; however, it has become clear that modern data collection poses new risks of disclosure and privacy breaches. For example, Homer et al. [78] recently showed that it is possible to identify the presence of individual genotypes in high-density SNP arrays, and consequently, it is possible to identify an individual from data obtained from genome-wide association (GWA) studies, which contain a mixture of DNA of thousands of individuals and their genetic fingerprints. This observation has led to the removal of some publicly available genomics data [64] from the US National Institutes of Health (NIH) and the Broad Institute in Cambridge. A major challenge in studies on privacy has thus become characterizing and balancing statistical utility with the privacy of individuals from whom we obtain data [48, 49, 58].

In the large body of research on privacy and statistical inference (e.g., [49, 57, 58, 145]), a major focus has been the problem of reducing ”disclosure risk”, i.e., the probability that certain data of a member of the dataset can be deduced given the released statistics of the dataset. The literature in most cases has stopped short, however, of providing a formal formulation of disclosure risk that would permit information-theoretic and estimation-theoretic tools to be used in characterizing tradeoffs between privacy and the utility associated with an inferential goal. Recently, a formal definition of disclosure risk known as ”differential privacy” was proposed by Dwork and colleagues and extensively studied in the cryptography and theoretical computer science literatures [51, 52, 74]. Differential privacy has strong semantic[1] privacy guarantees that make it a good candidate for declaring a statistical procedure private, and it has been the focus of a growing body of recent work [51, 54, 75, 134, 85].

1.1 Differential Privacy

Dalenius [43] suggested the *ad omnia* privacy desideratum: ”nothing about an individual should be learnable from the database that could not be learned without access to the database”. This requirement is shown to be too strong to be useful in practice. In the absence of a precise mathematical framework for privacy, statisticians have been tempted to use various rules of thumb to maintain privacy, e.g., do not answer any query that requires fewer than k entries [2]. To establish a mathematical framework for privacy, Dwork [50] proposed the notion of differential privacy which, informally speaking, implies that the

[1] Semantically-flavored interpretation of differential privacy: regardless of external knowledge, an adversary with access to the mechanism's output draws the same conclusions whether or not any individual's data is included in the original database. The use of the term ”semantic” dates back to [65] in cryptography.

presence and absence of an individual in the database should not affect in a significant way the probability of obtaining a certain answer for a given query.

Let x denote a given database consisting of n rows, each of which corresponds to the data of each individual. That is, x takes value in $\mathcal{D}^n$, where $\mathcal{D}$ is an abstract set containing all possible values of each row. The answer to any real-valued query $q : \mathcal{D}^n \to \mathbb{R}$ about this database may compromise the privacy of some individuals. Therefore, one needs to design a randomized mechanism $\mathcal{M}$ for the query function q which randomly provides $\mathcal{M}(x)$ with a probability distribution depending on $q(x)$. The mechanism is said to satisfy ε-differential privacy if for all *neighboring* databases, i.e., x and $\tilde{x}$ with Hamming distance $d_{\mathsf{H}}(x, \tilde{x}) = 1$, and any measurable set B we have

$$\sup_{B,x,\tilde{x}} \frac{\Pr(\mathcal{M}(x) \in B)}{\Pr(\mathcal{M}(\tilde{x}) \in B)} \leq e^{\varepsilon}, \tag{1.1}$$

It is shown in [51] and [52] that the requirement (1.1) is achieved by an additive channel $\mathcal{M}(x) = q(x) + N_{\mathsf{L}}$, where the independent noise N_{L} is drawn from a Laplacian distribution with variance depending on ε and also on the L_1 sensitivity of function q (which is the maximum $|q(x) - q(\tilde{x})|$, maximized over all neighboring x and $\tilde{x}$). The constraint (1.1) can be equivalently written in terms of an f-divergence [38], e.g., [103, 131] and also in terms of a mutual information constraint [42].

There are a few works on establishing a connection between differential privacy and information theory. For example, it is shown in [30, 44] that the mutual information $I(X; Z)$ between a database X and the output Z of an ε-differentially private mechanism (i.e., $\mathcal{M}(X) = Z$) can be unbounded. On the other hand, Cuff and Yu [42] showed that when (1.1) is met, then we have $I(X_i; Z|X^{-i}) \leq \varepsilon$, for all $1 \leq i \leq n$, where X_i corresponds to the ith row and X^{-i} denotes all other rows in the database. Furthermore, differential privacy has been compared with the rate-distortion function in [121, 128, 111]. In particular, Mir [111] showed that any mechanism that achieves the rate-distortion function also guarantees a certain level of differential privacy. This result was improved in [144] by bounding the gap between mutual information and differential privacy level for a given Hamming distortion threshold D (for $D \geq 0$ ranging over a certain interval) and it was shown that the gap diminishes if the database is uniformly distributed.

Notice that the definition of differential privacy does not depend on the prior distribution P_X, as it captures the disclosure of information by the mechanism after the adversary observes the mechanism's output (compared to no output). There are some recent studies which take the prior distribution into account and generalize (1.1) to the posterior distribution, e.g., [94, 98, 144].

1.2 Limitations of Differential Privacy

The requirement (1.1) involves the notion of neighboring databases and hence it intuitively implies that the adversary has already learned about all but one entry in the database and wishes to gain extra information about that remaining missed entry. This subtle model for the adversary can be made clear by a recent result of Cuff and Yu [42, Theorem 1]. Having made such a strong assumption on the adversary's knowledge, one expects that differential privacy must yield a very strong privacy guarantee. However, quite surprisingly, as shown in [86], a privacy constraint that limits the inference of the stronger adversary can sometimes leak more sensitive information compared to the privacy constraint designed for the weaker adversaries. Equivalently, a weaker assumption on the knowledge of adversaries might yield a better privacy guarantee. This counter-intuitive phenomenon occurs if the entries of the database are correlated which is quite common in certain applications especially in social networks and networked data. In fact, several examples of privacy breaches of differentially private mechanisms in social networks are presented in [86]. It has thus become clear that differential privacy is not suitable in social networks, see also [76]. Quoting from [158, p.31]:

> *"Publishing complex network data with privacy guarantees remains a challenge. For example, adapting differential privacy to networked data is not straightforward. The development of (possibly new) rigorous privacy definitions which address the complexity of network data is a very important research direction."*

Moreover, differential privacy concerns a scenario where each individual needs to trust the data-collecting statistician or institution, who owns the corresponding database. For example, patients of a certain hospital need to trust the hospital for collecting all their medical records (e.g., HIV status or information about any other critical disease). However, in many practical scenarios, individuals disclose their personal information voluntarily while they do not trust the data-collecting agency and hence prefer to be able to control the level of privacy themselves. In this model, which is often called a *local privacy model*, each individual randomizes his own data using a local mechanism to obtain a report which he sends to an untrusted statistician/agency to be aggregated in a database that can be used to answer queries about the data. Indeed, local privacy is one of the oldest forms of privacy: its essential form dates to Warner [145], who proposed it as a remedy for what he termed "evasive answer bias" in survey sampling. Inspired by seminal work of Warner, Duchi et al. [47] introduced a local version of differential privacy. The definition for ε*-local differential privacy* is analogous to (1.1) except that the condition on neighboring databases is removed.

This privacy constraint is now more suitable for networked data compared to (1.1). This new privacy definition has recently gained interest in the information theory community, e.g., [81, 82, 83, 99, 100, 116] with different utility functions. In particular, Kairouz et al. [82] defined an *information preservation* problem where individuals would like to release an ε-locally differentially private view Z of X that preserves as much information in X as possible.

Although the local definition of differential privacy is shown to have connection with binary hypothesis testing when $|\mathcal{X}| = 2$ [82], it has thus far evaded any operational interpretations. Hence, the quest for an *operational* formulation of privacy that is suitable for correlated data has not ended. In this thesis, we present two operational privacy formulations for any private data X, regardless of the cardinality $\mathcal{X}$, be it finite, countably infinite, or uncountably infinite.

1.3 Information-Theoretic Secrecy Models

Shannon initiated the problem of ”secrecy” in [130] where he mainly used the mathematical tools and insights he already developed to study the fundamental limits of information transmission over noisy channels. Secrecy problems concern the reliable information transmission over a noisy channel subject to the inability of the reliable decoding of the source message (with a finite alphabet) by a third party, the eavesdropper. Shannon studied the existence of a random variable, the so-called *key*, that, when XORing it with the channel input, can completely conceal all information being transmitted over the channel. Thus the eavesdropper who observes the channel is completely ignorant of the message. Shannon proved that if zero information leakage is required (that is the eavesdropper’s observation is statistically independent of the message being transmitted), the entropy of the key must be no less than that of the message.

After almost 25 years, Wyner [154] linked two notions of capacity and secrecy by defining the *secrecy capacity* for the wiretap channel. Wyner studied a model where the transmitter and the legitimate receiver communicate over a discrete memoryless channel (DMC) W_1 while an eavesdropper uses a second DMC W_2 to wiretap the information transfer over the first channel. In this scenario, in addition to making sure that the legitimate receiver can decode the transmitted message M with vanishing error probability, one requires to keep the eavesdropper almost ignorant of M. The logarithm of the maximum number of values for message M that can be reliably transmitted to the legitimate receiver and, at the same time, can be made almost independent of the eavesdropper's observation, is called the *secrecy capacity*. Wyner studied a special case in which W_2 is a degraded version of W_1. He showed that in this case, the secrecy capacity is equal to $\max[I(P_X, \mathsf{W}_1) - I(P_X, \mathsf{W}_2)]$, where the maximum is taken over the input distribution P_X, where $I(\cdot, \cdot)$ denotes mutual information (see Section 1.8 for a detailed definition). The achievability of this result is based on a random coding argument and thus does not give an explicit construction. A similar problem of "secret bit extraction" was studied in [23] which was shown to be efficient and practical from the cryptographic point of view.

Secrecy can also be studied from the source coding point of view, e.g., [53, 68, 69, 119, 139, 143]. In Chapter 2, we present a secret source coding model and compare it with our privacy model (which we describe in the next section): the objective is to encode a pair of source n-tuples (X^n, Y^n) into an index J such that a receiver, knowing J and some external side information Z^n, can losslessly recover Y^n, while any eavesdropper knowing J and possibly a correlated side information E^n can retrieve very little information about X^n. We argue that this model is more realistic in practice than the model used in secrecy problems.

1.4 Our Privacy Model: Generalization of the Local Privacy Model

In this thesis, we adopt the local privacy model where individuals do not have to trust the data-collecting agency and need to design and control their own randomizing mechanisms.

The model assumed in the secrecy problems relies on the fact that the message must be reliably decoded by the ”legitimate” receiver and, at the same time, it must be kept secret from the eavesdropper. However, in some practical scenarios, the information source has some features which are considered to be private and need to be kept private even from the intended receiver.

Example 1.1 ([155]). Consider an information service company which possesses a two-dimensional source with correlated outputs (X, Y). A customer pays a charge to obtain information about Y and then the company supplies $\hat{Y}$ to the customer within a prescribed distortion level. However, since X and Y are correlated, the customer can partially estimate X upon receiving $\hat{Y}$ with some accuracy. The company therefore wishes to keep X private from the intended receiver, because the charge is paid only for Y.

Example 1.2. Sweeney [137], in an interesting experiment, showed that the identity of a US citizen can be determined with high accuracy given gender, birth date, and postal code. Now consider an individual who, in the process of setting up a social network account, voluntarily provides his/her postal code in order to enjoy the customized feeds. According to Sweeney's experiment, the privacy of the individual is compromised as his/her identity can be determined with high accuracy. Thus each individual wishes to provide his/her postal code to the intended receiver to the extent his/her privacy is not compromised. Does the individual have to give up the benefits of customized services in order to maintain privacy? Is there a smart way of revealing postal code such that the identity cannot be efficiently estimated?

Example 1.3. In a survey, an individual is asked about his/her diabetes status. On the one hand, the individual needs to tell the truth, and on the other hand, the truth might reveal partial information about his/her HIV status, as the correlation between diabetes and HIV status is tentatively known [1]. Do individuals have to lie about their diabetes status in order to maintain privacy with respect to their HIV status?

The common theme in all these examples is the existence of two sets of data: non-private data Y and private data X, which is embedded in Y via a fixed channel $P_{Y|X}$ predefined by nature. The user wishes to reveal *only* Y to the intended receiver; however, the correlation between Y and X may disclose partial private information. Therefore, the *utility* is measured with respect to Y and the *privacy leakage* is defined with respect to X, and the goal is to *design* a privacy-preserving mechanism $P_{Z|Y}$ (the so-called *privacy filter*) such that Z carries as much "information" about Y as possible and at the same time infers as little about X as possible, see Fig. 1.1. The mechanism then *displays* Z, which is thus called *displayed data*. To make this goal precise, we need to quantitatively specify the *information efficiency* or utility between Y and Z, denoted by $\mathcal{U}(Y,Z)$, as well as the *privacy leakage* between X and Z, denoted by $\mathcal{P}(X,Z)$. The appropriate $\mathcal{U}$ and $\mathcal{P}$ must satisfy some intuitively clear properties: i) $\mathcal{U}(Y,Z)\geq 0$ with the equality if Y and Z are independent (denoted by $Y \perp\!\!\!\perp Z$), ii) $\mathcal{P}(X,Z)\geq 0$ with equality if and only if Z does not provide any advantage in the inference of X, iii) $\mathcal{P}(X,Z)$ satisfies the data processing inequality,[2] and iv) $\mathcal{U}(Y,Z)$ attains its maximum (which might be ∞) if $Y=Z$.

After $\mathcal{U}(Y,Z)$ and $\mathcal{P}(X,Z)$ are specified, the goal can then be quantified precisely by

[2]We say that $\mathcal{P}$ satisfies the data processing inequality if we have $\mathcal{P}(X,Z)\leq\mathcal{P}(X,Y)$ for random variables X,Y and Z satisfying $X \multimap Y \multimap Z$ (see Section 1.8 for the definition).

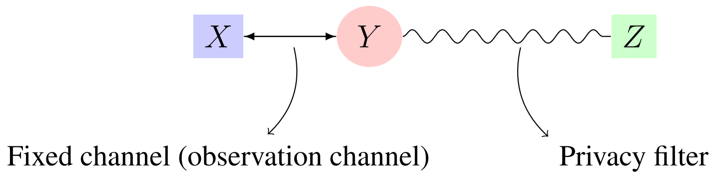


Figure 1.1: Our privacy model: for a given pair of random variables (X, Y), representing private and non-private data, respectively, the goal is to generate the displayed data Z which maximizes $\mathcal{U}(Y, Z)$ while limiting the privacy leakage $\mathcal{P}(X, Z)$. Due to the correlation between X and Y, maximizing $\mathcal{U}(Y, Z)$ and minimizing $\mathcal{P}(X, Z)$ conflict with each other. To quantify the tradeoff, we therefore need to introduce $\mathcal{G}(P_{XY}, \varepsilon)$ as in (1.2) for an $\varepsilon \geq 0$.

introducing the privacy-constrained utility efficiency function

$$\mathcal{G}(P_{XY}, \varepsilon) \coloneqq \sup_{\substack{P_{Z|Y}: X -\!\circ\!- Y -\!\circ\!- Z \\ \mathcal{P}(X,Z) \leq \varepsilon}} \mathcal{U}(Y, Z), \tag{1.2}$$

where P_{XY} is the joint distribution of (X, Y), $\varepsilon \geq 0$ specifies the privacy level, and $X -\!\circ\!- Y -\!\circ\!- Z$ means X, Y, and Z form Markov chain in this order (i.e., X and Z are conditionally independent given Y).

Yamamoto [155] studied this problem from a lossy source coding standpoint where privacy and utility are measured in terms of mutual information and a given distortion function, respectively. This model has recently gained interest in the information theory literature (see, e.g., [6, 14, 22, 90, 112, 120, 126]). In Chapter 2, we study a similar privacy model with the lossless reconstruction of Y with and without eavesdropper's side information. The model studied in Chapter 2 has been recently generalized to the privacy-aware remote source coding in [90].

In this thesis, we adopt information and estimation theoretic approaches to propose appropriate $\mathcal{U}(Y, Z)$ and $\mathcal{P}(X, Z)$.

1.5 Information-Theoretic Approaches in Privacy

As Shannon indicated in [130], from the point of view of the cryptanalyst, a secrecy system is almost identical to a noisy communication system and hence the information-theoretic tools developed to study the fundamental limits of information transfer over a noisy channel can be used to model, describe and analyze a secrecy or privacy system.

Following Shannon's lead, one can measure privacy leakage $\mathcal{P}(X, Z)$ via the mutual information $I(X;Z)$. Consequently, a mechanism $P_{Z|X}$ is said to be ε-private if $I(X;Z) \leq \varepsilon$ for the given prior distribution P_X and ε. Although this choice of $\mathcal{P}(X, Z)$ is not operationally well motivated, it has been used extensively in several papers as an appropriate secrecy metric [39, 53, 66, 92, 108, 119, 139, 144]. In secrecy systems, the main motivation for using mutual information as the secrecy metric is the similarity between the deciphering task for the eavesdropper in the secrecy model and the decoding task in the standard noisy communication setting. According to this similarity, $I(X;Z)$ only provides a lower bound on the exponent of the list size that the eavesdropper needs to make to reliably include the source sequence.

Motivated by a seminal work of Massey [107] on guessing, Merhav and Arikan [109] proposed another metric: the expected number of guesses, given observation Z, that the adversary needs to make to find out the correct value of the source X. The problem with this notion of secrecy is that the adversary needs to possess a testing system by which he can check whether or not his guess is correct. However, any practical system only allows a certain number of incorrect inputs. Moreover, it can be shown that there exists highly insecure systems which are labeled secure under this metric.

More recently, Issa and Wagner [80] proposed to measure leakage in terms of the probability of a successful guess by the adversary which is defined as the probability that the distortion the adversary incurs is below a given level. It seems an appropriate measure of secrecy; however, it is very hard to deal with especially in single-shot settings.

We use mutual information in Chapter 3 as a privacy metric. The main motivation for this choice is that mutual information is well-studied and it turns out the corresponding utility-privacy tradeoff, that we define in Chapter 3, can be used to bound the other better-justified utility-privacy tradeoff in the subsequent chapters. We also use mutual information as the information efficiency metric. In particular, we study the following question: Given discrete correlated random variables $X \in \mathcal{X} = \{1, \ldots, M\}$ and $Y \in \mathcal{Y} = \{1, \ldots, N\}$, how much information can maximally be extracted from Y while revealing a limited amount of information about X?

Since X and Y are correlated, disclosing Y completely compromises the privacy of X. Using the functional representation lemma [87, p. 626], there exists a random variable V with bounded cardinality such that $H(Y|X,V) = 0$ and $V \perp\!\!\!\perp X$. If V satisfies $X -\!\!\circ\!\!- Y -\!\!\circ\!\!- V$, then the channel $P_{V|Y}$ can be chosen as a privacy filter and thus $Z = V$ is displayed. Hence, no information about X is revealed, leading to *perfect privacy*. In this case, the utility is $H(Y) - I(X;Y|V)$. Now suppose that we allow the displayed data Z to reveal at most ε bits of information about X. To quantify the maximum utility, we introduce the so-called *rate-privacy function* $g(P_{XY}, \varepsilon)$ in Chapter 3 as the maximum number of bits that one can transmit about Y which reveals ε bits of information about X:

$$g(P_{XY}, \varepsilon) \coloneqq \sup_{\substack{P_{Z|Y}: X -\!\circ\!- Y -\!\circ\!- Z, \\ I(X;Z) \leq \varepsilon}} I(Y;Z). \tag{1.3}$$

This quantity has a similar form as the well-studied *information bottleneck* function [140]

$$\mathsf{IB}(P_{XY}, R) := \sup_{\substack{P_{Z|Y}: X -\!\circ\!- Y -\!\circ\!- Z, \\ I(Y;Z) \leq R}} I(X;Z). \tag{1.4}$$

Interestingly, the quantity $\mathsf{IB}(P_{XY}, \cdot)$ appears as a solution to several problems in information theory, including: lossy source coding with logarithmic loss distortion [35], a generalization of Mrs. Gerber's Lemma [153] in [147], lossless source coding with one helper [5, 152], and also the strong data processing inequality [4, 11]. The quantity $g(P_{XY}, \cdot)$ was recently shown in [96] to be closely related to the extended Gray-Wyner network [67]. Although $\varepsilon \mapsto g(P_{XY}, \varepsilon)$ and $R \mapsto \mathsf{IB}(P_{XY}, R)$ share several similar properties (e.g., both are strictly increasing, concave and also $\varepsilon \mapsto \frac{g(P_{XY}, \varepsilon)}{\varepsilon}$ and $R \mapsto \frac{\mathsf{IB}(P_{XY}, R)}{R}$ are decreasing), it is important to note that they are fundamentally different.

The functional dual of $g(P_{XY}, \cdot)$ can be defined [105] as

$$t(P_{XY}, R) := \inf_{\substack{P_{Z|Y}: X -\!\circ\!- Y -\!\circ\!- Z, \\ I(Y;Z) \geq R}} I(X;Z), \tag{1.5}$$

which minimizes the privacy level such that Z carries at least R bits of information about Y. It is insightful to notice that the graph of $t(P_{XY}, \cdot)$ and $\mathsf{IB}(P_{XY}, \cdot)$ are, respectively, the lower and upper boundaries of the two-dimensional convex set $\{(I(Y;Z), I(X;Z)) : X -\!\circ\!- Y -\!\circ\!- Z, (X,Y) \sim P_{XY}\}$ [29]. This alternative characterization allows us to study $t(P_{XY}, \cdot)$ and $\mathsf{IB}(P_{XY}, \cdot)$ from the geometric and convex analytic points of view (especially using results such as Carathéodory-Fenchel and Dubin's theorems).

1.6 Estimation-Theoretic Approaches in Privacy

In the main body of recent research activities on privacy, the major focus has been on the disclosure risk and how to limit it. Direct connections between *statistical efficiency* and privacy, however, have been somewhat more challenging to make. With recent issues in data collection, it is becoming more important to understand quantitative tradeoffs between privacy and statistical efficiency, especially in our model where privacy and statistical efficiency need to be defined for different correlated random variables X and Y.

Our goal here is to take a fully inferential point of view on privacy by bringing privacy into contact with estimation theory. Our focus is on the fundamental limits of privacy-aware estimation in both discrete and continuous cases. To this end, we need an appropriate estimation-theoretic privacy leakage function.

It was long believed that adding random independent noise to the private database preserves privacy of each individual. For example, each individual may perturb his/her own data x_i as $z_i = x_i + N_i$ and send it to the data collecting agency, where N_i is a random variable independent of x_i that is uniformly distributed over an interval [3]. However, it is easy to see that the privacy of individuals is not maintained by this perturbation. For example, if N_i is drawn uniformly from interval $[-50, 50]$ and $z_i = 120$ is observed, then it is easy to conclude that $x_i \geq 70$.

To make sure privacy is preserved, the notion of *privacy breach* was introduced in [54, 55]: Given $0 < \rho_1 < \rho_2 < 1$, we say that there is a ρ_1-to-ρ_2 privacy breach with respect to function f if for some $z \in \mathcal{Z}$, we have $\Pr(f(X) = x) \leq \rho_1$ and $\Pr(f(X) = x|Z = z) \geq \rho_2$, or equivalently, if the event $\{f(X) = x\}$ has a jump of magnitude $\rho_2 - \rho_1$ from its prior to posterior after $Z = z$ is observed. Evfimievski et al. [54] proposed a necessary condition on $P_{Z|X}$ which guarantees that ρ_1-to-ρ_2 privacy breach does not happen with respect to any deterministic function.

More recently, a new line of research called *quantitative information flow* has been studying the quantitative measures of information leakage in a private system, e.g., [9, 135, 129, 106]. The most widely used measure of information leakage, for X and Z being defined on countable alphabets, is based on the so-called *min-entropy*, defined as

$$H_\infty(X) \coloneqq H_\infty(P_X) = \min_{x \in \mathcal{X}} \left[-\log P_X(x) \right],$$

which is, in fact, the Rényi entropy [124] of order ∞. Letting $\mathsf{P_c}(X)$ be the probability of correctly guessing X without any side information in a one-try attempt; i.e., $\mathsf{P_c}(X) = \max_{x \in \mathcal{X}} P_X(x)$, we can write $H_\infty(X) = -\log \mathsf{P_c}(X)$. In the presence of side information Z, we let $\mathsf{P_c}(X|Z)$ denote the average probability of correctly guessing X using the ”optimal” strategy of maximum a posterior (MAP) decoding, given by

$$\mathsf{P_c}(X|Z) \coloneqq \sum_{z \in \mathcal{Z}} P_Z(z) \max_{x \in \mathcal{X}} P_{X|Z}(x|z).$$

A natural definition for a conditional Rényi entropy was given by Arimoto [56] as $H_\infty(X|Z) \coloneqq -\log \mathsf{P_c}(X|Z)$. The quantity $I_\infty(X;Z) \coloneqq H_\infty(X) - H_\infty(X|Z) = \log \frac{\mathsf{P_c}(X|Z)}{\mathsf{P_c}(X)}$, which we refer to as *Arimoto's mutual information of order infinity*, therefore measures the ”advantage” of Z in guessing X in a one-try attempt. A major body of research in this area focuses on the connection between $I_\infty(X;Z)$ and the differential privacy level, e.g., [8, 9, 26, 33, 106, 129]. However, since differential privacy does not depend on prior distribution P_X these bounds are usually loose unless P_X is uniform [9, Proposition 1].

More recently Issa et al. [79] proposed another operational measure of privacy leakage for the discrete case. According to their definition, the mechanism $P_{Z|X}$ is said to be ε-private if $I_\infty(U;Z) \le \varepsilon$ for any random variable U satisfying[3] $U \multimap X \multimap Z$. This metric therefore ensures that even randomized functions of X cannot be efficiently guessed from Z. This privacy metric is very stringent and does not depend on the prior distribution.

In this thesis, we consider two quantities as estimation-theoretic measures of privacy leakage: maximal correlation $\rho_m(X,Z)$ [122], and $I_\infty(X;Z)$. The significance of $\rho_m(X,Z)$ as a privacy leakage function is that it is well-defined for the both discrete and continuous cases and it yields a strong semantic privacy guarantee. In particular, the condition $\rho_m^2(X;Z) \le \varepsilon$ implies the following:

- If both X and Z are discrete random variables, then $\mathsf{P_c}(f(X)|Z) - \mathsf{P_c}(f(X)) \le \varepsilon\sqrt{1-\sum_i P^2_{f(X)}(i)}$ [28, Theorem 5.6] for any deterministic function f. Hence, for small $\varepsilon \ge 0$, it is nearly as hard for an adversary observing Z to guess *any* deterministic function of X as it is without Z.

- If both X and Z are absolutely continuous random variables, then we have

$$(1-\varepsilon)\mathsf{var}(f(X)) \le \mathsf{mmse}(f(X)|Z) \le \mathsf{var}(f(X)), \tag{1.6}$$

  for any non-constant real-valued measurable function f, where mmse and var denote the minimum mean-squared estimation error (MMSE) and variance, respectively. The relation (1.6) states that for small $\varepsilon \ge 0$, no function of the private data X can be efficiently estimated given observation Z.

[3]It can be shown that $\sup_{U:U \multimap X \multimap Z} I_\infty(U;Z) = I_\infty^{\mathsf{S}}(X;Z)$, where $I_\infty^{\mathsf{S}}(X;Z)$ is the so-called *Sibson*'s mutual information of order infinity [133, 142]. Since $I_\infty(X;Z) \le I_\infty^{\mathsf{S}}(X;Z)$, the metric $I_\infty^{\mathsf{S}}(X;Z)$ results in a more stringent privacy guarantee as expected from the operational interpretation.

These implications are similar, in essence, to the cryptographic *semantic security* principle [65], which states that an adversary must have a negligible advantage in guessing any function of the input (i.e., plaintext) given an observation of the mechanism's output (i.e., ciphertext).

We see that the constraint $\rho_m^2(X,Z) \leq \varepsilon$ yields a privacy guarantee which is, in some sense, similar to what Dalenius [43] suggested as the privacy desideratum: *almost* nothing about the private data should be learnable from the displayed data that could not be learned without access to it.

To have a fully inferential utility-privacy tradeoff, we also measure the information efficiency in terms of $I_\infty(Y;Z)$ for discrete Y and $\frac{\mathsf{var}(Y)}{\mathsf{mmse}(Y|Z)}$ for continuous Y. Therefore, we propose the following functions as quantitative utility-privacy tradeoffs

$$g^\infty(P_{XY},\varepsilon) \coloneqq \sup_{\substack{P_{Z|Y}: X \multimap Y \multimap Z, \\ I_\infty(X;Z) \leq \varepsilon}} I_\infty(Y;Z), \tag{1.7}$$

and

$$\mathsf{sENSR}(P_{XY},\varepsilon) \coloneqq \inf_{\substack{P_{Z|Y}: X \multimap Y \multimap Z, \\ \rho_m^2(X,Z) \leq \varepsilon}} \frac{\mathsf{mmse}(Y|Z)}{\mathsf{var}(Y)}. \tag{1.8}$$

Motivated by the problem of "secret bit extraction" [23], we also introduce the function $\hat{g}(P_{XY},\cdot)$ as the maximal information extraction from Y under the inferential privacy constraint dictated by maximal correlation:

$$\hat{g}(P_{XY},\varepsilon) \coloneqq \sup_{\substack{P_{Z|Y}: X \multimap Y \multimap Z, \\ \rho_m^2(X,Z) \leq \varepsilon}} I(Y;Z). \tag{1.9}$$

It is interesting to mention that the function $\hat{g}(P_{XY},\cdot)$ has recently been modified in [156] to define a general notion of common information which subsumes Wyner's [151] and

Gács-Körner's [59] notions of common information.

1.7 Contributions and Organization of the Thesis

1.7.1 Chapter 2

We start by studying a simple and practical, yet unexplored, private lossless compression model and comparing it with a well-studied secrecy model. Specifically, motivated by the standard lossless source coding problem with one helper [152] and Yamamoto's lossy privacy model [155], we study in this chapter the fundamental information theoretic limits of recovering Y losslessly with the help of coded side information at the decoder while the mutual information between the message being transmitted over the channel and X is negligible. This model subsumes the well-studied secrecy models by setting $X = Y$. The results of this chapter have appeared in [14].

1.7.2 Chapter 3

In this chapter, we introduce the rate-privacy function $g(P_{XY}, \cdot)$ for a pair of correlated random variables (X, Y) defined over finite alphabets. After proving some fundamental properties of the map $\varepsilon \mapsto g(P_{XY}, \varepsilon)$, we derive a tight lower bound and show that this lower bound is achieved by an erasure mechanism. One difficulty we face in evaluating $g(P_{XY}, \cdot)$ is the cardinality of the auxiliary random variable Z. Using the standard convex cover method [87, 40] and the Fenchel-Eggleston-Carathéodory theorem, one can show that $|\mathcal{Z}| = |\mathcal{Y}|+1$ is sufficient in general. Hence even in the simplest case of $|\mathcal{X}| = |\mathcal{Y}| = 2$, we have four variables in the defining non-convex optimization problem in (1.3) which makes it intractable. We invoke Dubin's Theorem [46] to show that the cardinality bound of Z can be strengthened from $|\mathcal{Y}| + 1$ to $|\mathcal{Y}|$ if the map $\varepsilon \mapsto g(P_{XY}, \varepsilon)$ is strictly concave.

We also formulate a coding problem, the so-called *dependence dilution problem*, and show that $g(P_{XY}, \cdot)$ is a boundary point of the achievable rate region of the dependence dilution problem.

We also revisit the convex-analytic approach that Witsenhausen and Wyner developed to generalize Mrs. Gerber's lemma [153] and show that this approach can be modified to yield a closed form expression for $g(P_{XY}, \varepsilon)$ when $P_{X|Y}$ is either a binary symmetric channel (BSC) or a binary erasure channel (BEC).

In the second part of this chapter, we focus on a particular family of joint distributions P_{XY}: Y is binary and $P_{X|Y}$ is a binary input symmetric output (BISO) channel. For this family of joint distributions, we show that the general lower bound is achieved if and only if Y is uniform and hence establish the optimality of the erasure mechanism in this case.[4]

A closed form expression for $g(P_{XY}, \varepsilon)$, for any ε in its domain, is given in the following cases:

- If $Y \sim \mathsf{Bernoulli}(q)$ with $0 \leq q \leq \frac{1}{2}$ and $P_{X|Y}$ is either BSC or BEC or Z-channel,
- If $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$ and $P_{X|Y}$ is a BISO,
- If $P_{Y|X}$ is an erasure channel.

The results of this chapter have partially appeared in [19, 13, 18].

1.7.3 Chapter 4

In this chapter, we study the same problem of information extraction under an information theoretic privacy constraint except that here we assume that X and Y are absolutely continuous random variables. To make the problem tractable, we focus on a particular practical

[4]This optimal mechanism cannot be locally differentially private, which shows that the set of information-theoretically private mechanisms is much larger than the set of differentially private mechanisms.

privacy filter of interest which acts in two stages: first Gaussian noise is added and then the resulting random variable is quantized using an M-bit accuracy uniform scalar quantizer $\mathcal{Q}_M$ (for some positive integer $M \in \mathbb{N}$). Specifically, in this chapter $g_M(P_{XY}, \varepsilon)$ is defined as the maximum $I(Y; Z_\gamma^M)$, where $Z = Z_\gamma^M \coloneqq \mathcal{Q}_M(\sqrt{\gamma}Y + N_{\mathsf{G}})$ and N_{G} is standard Gaussian random variable independent of (X, Y) and the maximization is taken over all $\gamma \geq 0$ such that $I(X; Z_\gamma) \leq \varepsilon$. We show that $g_M(P_{XY}, \varepsilon) \to g(P_{XY}, \varepsilon)$ as $M \to \infty$ where $g(P_{XY}, \varepsilon)$ is defined similarly to (1.3) where $Z = Z_\gamma \coloneqq \sqrt{\gamma}Y + N_{\mathsf{G}}$. This result leads us to evaluating $g(P_{XY}, \varepsilon)$ for any given $\varepsilon \geq 0$. Although characterizing the exact value of $g(P_{XY}, \varepsilon)$ seems very difficult, we utilize the I-MMSE relationship [70, 71, 150] to derive a second-order approximation for $g(P_{XY}, \varepsilon)$ when $\varepsilon > 0$ is sufficiently small (the *almost* perfect privacy regime). Interestingly, $\mathsf{sENSR}(P_{XY}, \cdot)$ on this range of ε is closely related to the largely-ignored Rényi's one-sided maximal correlation [122]. The results of this chapter have appeared in [19, 16, 15].

1.7.4 Chapter 5

This chapter is a natural continuation of Chapter 3 where the information-theoretic privacy constraint $I(X; Z) \leq \varepsilon$ is replaced by the inferential constraint $\rho_m^2(X, Z)$ leading to introducing the quantity $\hat{g}(P_{XY}, \cdot)$, defined in (1.9), as the corresponding utility-privacy tradeoff. This function seems to be the first operational utility-privacy tradeoff in the literature where utility and privacy are defined with respect to different sources.

Evaluating this function appears to be difficult; however, it is shown that $g(P_{XY}, \cdot)$ provides a tight upper bound for $\hat{g}(P_{XY}, \cdot)$. Since maximal correlation and mutual information share many similar properties, the techniques developed in Chapter 3 to study $g(P_{XY}, \cdot)$ can be used, mutatis mutandis, to study $\hat{g}(P_{XY}, \cdot)$. In particular, again the erasure mechanism yields the lower bound for $\hat{g}(P_{XY}, \cdot)$ and this lower bound is shown to be optimal for $P_{X|Y}$ being BISO, only if Y is uniformly distributed. The results of this chapter have appeared in [19].

1.7.5 Chapter 6

In this chapter, we assume X and Y are discrete random variables and introduce a parametric family of utility-privacy functions $g^{(\nu,\mu)}(P_{XY}, \varepsilon)$ for any $\nu, \mu \geq 1$ as the maximum $I_\mu(Y; Z)$, where $I_\mu(Y; Z)$ is Arimoto's mutual information of order $\mu \geq 1$ [142, 12] and the maximization is taken over all mechanisms $P_{Z|Y}$ such that $I_\nu(X; Z) \leq \varepsilon$. We show that $g(P_{XY}, \cdot)$ and $g^\infty(P_{XY}, \cdot)$, defined respectively in (1.3) and (1.7), are extreme members of this family when both ν and μ approach one and infinity, respectively. We then argue that $g^\infty(P_{XY}, \cdot)$ can be used to upper and lower bound $g^{(\nu,\mu)}(P_{XY}, \cdot)$ for any $\nu, \mu > 1$ which allows us to focus on evaluating $g^\infty(P_{XY}, \varepsilon)$. We show that for binary X and Y the function $g^\infty(P_{XY}, \varepsilon)$ admits a simple closed form expression for all ε in its entire domain. The optimal mechanism in this case is a simple Z-channel which, as before, is not a differentially private mechanism.

Recall that $I_\infty(X; Z)$ is in a one-to-one relationship with $\mathsf{P_c}(X|Z)$, thus $g^\infty(P_{XY}, \cdot)$ can be studied by introducing the *privacy-constrained guessing probability* function $\hbar(P_{XY}, \varepsilon)$ as the maximum $\mathsf{P_c}(Y|Z)$, where the maximization is taken over all privacy filters $P_{Z|Y}$ such that $\mathsf{P_c}(X|Z) \leq \varepsilon$. A remarkable property of $\hbar(P_{XY}, \cdot)$, proved in this chapter, is that it is piecewise linear. This property is instrumental in deriving an expression for $g^\infty(P_{XY}, \cdot)$ in the binary case.

In the second part of this chapter, we make a simplifying, yet practically convenient, assumption that $\mathcal{Z} = \mathcal{Y}$. Even with this assumption, the computation of the corresponding

$\hbar(P_{XY}, \varepsilon)$ is rather complicated. We show that in this case $\hbar(P_{XY}, \varepsilon)$ and $g^{\infty}(P_{XY}, \varepsilon)$ admit simple closed-form expressions only for sufficiently large, but nontrivial, values of ε.

Finally, we generalize the convex-analytic approach, developed in Chapter 3, to the Arimoto's mutual information thereby computing $g^{(\nu,\nu)}(P_{XY}, \varepsilon)$ for any ε in its domain and $\nu \geq 2$, when $P_{X|Y}$ is BSC. The results of this chapter have appeared in [21, 20].

1.7.6 Chapter 7

As Verdú stated in [142], a shortcoming of Arimoto's mutual information is that its generalization to continuous random variables is not self-evident. Hence, $g^{(\nu,\mu)}$ cannot be adapted for absolutely continuous X and Y. In this chapter, we introduce the so-called *estimation-noise-to-signal ratio*, $\mathsf{sENSR}(P_{XY}, \cdot)$, as an operational utility-privacy tradeoff, given in (1.8). In fact, $\mathsf{sENSR}(P_{XY}, \varepsilon)$ quantifies the attainable minimum quadratic error in estimating Y from the observation Z that satisfies the condition $\rho_m^2(X, Z) \leq \varepsilon$ which leads to a strong semantic privacy guarantee.

Since X and Y are continuous random variables, characterizing the map $\varepsilon \mapsto \mathsf{sENSR}(P_{XY}, \varepsilon)$ seems complicated. Motivated by differential privacy as in Chapter 4, we focus on the independent additive Gaussian filter[5] $Z = Z_\gamma := \sqrt{\gamma}Y + N_{\mathsf{G}}$. With this privacy filter at disposal, $\mathsf{sENSR}(P_{XY}, \varepsilon)$ indeed corresponds to the smallest variance of the Gaussian noise for which the privacy constraint $\rho_m^2(X, Z_\gamma) \leq \varepsilon$ is satisfied. We obtain upper and lower bounds for $\mathsf{sENSR}(P_{XY}, \varepsilon)$ and establish another extremal property of jointly Gaussian distributions: among all (X, Y) with identical maximal correlation, the jointly Gaussian $(X_{\mathsf{G}}, Y_{\mathsf{G}})$ yields the largest $\mathsf{sENSR}(P_{XY}, \varepsilon)$.

We also derive a tight bound for $\mathsf{sENSR}(P_{XY}, \varepsilon)$ in the almost perfect privacy regime

[5] The assumption that the noise is independent of the input is not very restrictive. For example, Geng and Viswanath [62] showed that in the context of differential privacy the assumption of independent noise does not result in any loss of optimality.

(i.e., $\varepsilon \ll 1$) and establish a connection between $\mathsf{sENSR}(P_{XY}, \varepsilon)$ and $g(P_{XY}, \varepsilon)$. The results of this chapter have appeared in [21, 17]

1.8 Notation

Capital letters (e.g., U and V) are used to denote random variables, and calligraphic letters (e.g., $\mathcal{U}$ and $\mathcal{V}$) denote sets. The supports of random variables U and V are denoted by $\mathcal{U}$ and $\mathcal{V}$, respectively. We say X and Y are discrete random variables if X and Y have countable supports, e.g., when $|\mathcal{X}| < \infty$ and $|\mathcal{Y}| < \infty$. We denote the joint probability mass function (pmf) of discrete random variables X and Y by $P_{XY} = \{P_{XY}(x,y) : \ x \in \mathcal{X}, y \in \mathcal{Y}\}$, the conditional pmf of Y given X by $P_{Y|X}$, and the marginal distributions of X and Y by $P_X = \{P_X(x) : \ x \in \mathcal{X}\}$ and $P_Y = \{P_Y(y), \ y \in \mathcal{Y}\}$, respectively. We assume that $P_X(x) > 0$ and $P_Y(y) > 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. We use $\mathsf{Bernoulli}(p)$ to denote the distribution with support $\mathcal{X} = \{0, 1\}$ and $P_X(1) = p$. We use $X \sim P_X$ to denote that X is distributed according to P_X and similarly $(X, Y) \sim P_{XY}$ to denote that X and Y are jointly distributed according to P_{XY}. We say X, Y and Z form a Markov chain $X \multimap Y \multimap Z$, if their joint distribution satisfies $P_{XYZ}(x,y,z) = P_X(x)P_{Y|X}(y|x)P_{Z|Y}(z|y)$. We use $X \perp\!\!\!\perp Y$ to mean that X and Y are independent. An n-tuple random vector $(X_1, \dots, X_n)$ is denoted by X^n. We also let $[M]$ denote $\{1, \dots, M\}$. The set of all functions of a random variable $X \sim P_X$ with finite second moment is denoted by $\mathcal{L}^2(P_X)$. The set of all probability distributions supported over a set $\mathcal{X}$ is denoted by $\mathcal{P}_{\mathcal{X}}$.

For real-valued X and Y, we say X is absolutely continuous random variable if there exists a nonnegative function f, called probability density function (pdf), on $\mathbb{R}$ such that $\Pr(X \le x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$, for any $x \in \mathbb{R}$. We use the notation $X \sim f_X$ to specify the pdf of random variable X, for example, $X \sim \mathcal{N}(0,1)$ means that X is a standard Gaussian random variable.

For a discrete random variable $X \sim P_X$ and absolutely continuous random variable $Y \sim f_Y$, the entropy of X and Y are given by $H(X) \coloneqq -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x)$ and $h(Y) \coloneqq -\int_{\mathcal{Y}} f_Y(t)\log f_Y(t)\mathrm{d}t$, respectively. The base of the logarithm will be clear from the context. If $X \sim \mathsf{Bernoulli}(p)$, then $h_{\mathsf{b}}(p) \coloneqq H(X)$. If X and Y are either both discrete according to P_{XY} or both absolutely continuous according to f_{XY}, then the mutual information between X and Y is defined as $I(X;Y) \coloneqq H(X)+H(Y)-H(X,Y)$ or $I(X;Y) \coloneqq h(X)+h(Y)-h(X,Y)$, respectively. Note that the mutual information between two random variables is a function of the marginal distribution of one variable and the conditional distribution of the other variable given the former one. For example, when X and Y are discrete and have a joint distribution given by $P_X P_{Y|X}$, then we can write $I(X;Y) = I(P_X, P_{Y|X})$ to emphasize the functional dependence of mutual information on these distributions. Similarly, we can denote, $H(X)$ by $H(P_X)$ and $h(Y)$ by $h(f_Y)$.

For $a \in [0,1]$, we let $\bar{a} \coloneqq 1-a$. We frequently use three channels in this thesis. A channel from X to Y is called the binary symmetric channel with crossover probability α, denoted by $\mathsf{BSC}(\alpha)$, if $\mathcal{Y} = \mathcal{X} = \{0,1\}$ and $P_{Y|X}(1|0) = P_{Y|X}(0|1) = \alpha$. A channel from X to Y is called the binary erasure channel with erasure probability δ, denoted by $\mathsf{BEC}(\delta)$, if $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{0,\mathsf{e},1\}$ and $P_{Y|X}(y|x) = \bar{\delta}$ for $x=y$ and $P_{Y|X}(\mathsf{e}|x) = \delta$ for $x \in \mathcal{X}$. A channel from X to Y is called the Z channel with crossover probability β, denoted by $\mathsf{Z}(\beta)$, if $\mathcal{Y} = \mathcal{X} = \{0,1\}$ and $P_{Y|X}(0|0) = 1$ and $P_{Y|X}(0|1) = \beta$.

Given two probability distributions P and Q supported over a finite alphabet $\mathcal{U}$, the Kullback-Leibler divergence is defined as

$$D(P||Q) \coloneqq \sum_{u\in\mathcal{U}} P(u)\log\left(\frac{P(u)}{Q(u)}\right). \tag{1.10}$$

The minimum mean-squared error (MMSE) of estimating Y from an observation Z is given by

$$\mathsf{mmse}(Y|Z) := \min_{P_{\hat{Y}|Z}: Y -\!\circ\!- Z -\!\circ\!- \hat{Y}} \mathbb{E}[(Y - \hat{Y})^2] = \mathbb{E}[(Y - \mathbb{E}[Y|Z])^2] = \mathbb{E}[\mathsf{var}(Y|Z)],$$

where the first equality is due to the orthogonality principle which implies that the optimal estimator is a deterministic function of Z.

Chapter 2

Information-Theoretic Secrecy vs. Privacy: Yamamoto's Lossless Secure Source Coding

2.1 Overview

The secure source coding models concern a tradeoff between utility (i.e., reconstruction distortion) and privacy (i.e., the amount of information about the source leaking over the channel). Given a source Y^n, the goal is to transmit this source securely and reliably over a noiseless public channel which might be perfectly observed by a passive adversary. The utility is defined as the accuracy in the recovering of Y^n by a remote receiver and the privacy is defined as the uncertainty of the source given the message sent over the channel. However, in some cases, it may be desirable to define utility and privacy for two different sources, that is, we want the receiver to know Y^n with some level of accuracy while revealing very little information about a correlated source X^n, which we refer to as the private source.

To motivate this setting, consider the following example. Suppose that Y denotes an attribute of a bank customer that a trusted advertising company would like to target and X denotes another, more sensitive, attribute of the customer. The bank has database (X^n, Y^n)

corresponding to n different users. The company pays the bank to receive Y^n as accurately as possible. However, some governing laws prohibit the database X^n from being revealed too extensively over public communication channels. Consequently, the data given to the company must be chosen so that at most a prescribed amount of information is revealed about X^n over the communication channel while the recovery of Y^n by the company satisfies some level of quality.

Inspired by Yamamoto [155], where a lossy source coding problem is studied under a privacy constraint, we consider a secure lossless source coding model in which an encoder (Alice) encodes a two-dimensional source (X^n, Y^n) such that the receiver (Bob) is able to reconstruct Y^n correctly with high probability and the leakage of information (the information obtained by an eavesdropper, Eve) about X^n is no more than $\Delta \geq 0$. It is clear that no non-trivial level of privacy can be obtained if no side information is available to Bob. Hence, we assume Bob has access to some correlated side information and after observing the channel output wants to recover Y^n with asymptotically vanishing error probability. We study this problem in terms of the compression rate and also the information leakage about X^n (or equivalently the equivocation between the compressed and the private data). We give converse results for different cases including when Bob has coded or uncoded side information, when Eve has uncoded side information, or when the private source, X^n, is hidden even from Alice.

When $X = Y$, the problem we consider here reduces to a well-known model which has been extensively studied e.g., [68, 69, 119, 143, 139]. In particular, Prabhakaran and Ramchandran [119] considered a similar secure lossless setting with $X = Y$ and Bob and Eve having correlated uncoded side information. They focused on the best achievable information leakage rate when the public channel does not have rate limit. Gündüz et al. [69],

[68] gave converse and achievability bounds for a similar setting for both compression rate and information leakage which do not necessarily match. Tandon et al. [139] considered a simpler case where Eve has no side information. In this setting, they gave a single letter characterization of the optimal rates and information leakage and showed that a simple coding scheme based on binning, similar to the one proposed by Wyner in [152], is indeed optimal with and without the privacy constraint. Our results recover all these results in the special case $X = Y$.

The rest of this chapter is organized as follows. In Section 2.2, we formally define our problem and state an outer bound which is our main result. In Section 2.3, we consider a more general model where Eve has side information and present another outer bound. We then present a coding scheme which is shown to be optimal in some special cases. We complete this chapter with some concluding remarks in Section 2.4.

2.2 Yamamoto's Lossless Source Coding: Coded Side Information at Bob

Yamamoto [155] considered a lossy source coding scheme with a privacy constraint at the legitimate decoder as well as the eavesdropper. This is contrasted with the typical information-theoretic secrecy models in which the privacy is defined as the uncertainty of the source against a passive eavesdropper. In Yamamoto's model, having observed (X^n, Y^n), the encoder $\varphi : \mathcal{X}^n \times \mathcal{Y}^n \to [2^{nR}]$, transmits a message to the decoder, $\psi : [2^{nR}] \to \hat{\mathcal{Y}}^n$, which is required to recover Y^n within some distortion D while revealing little information about X^n. More precisely, for a given distortion measure $d : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}^+$, we require $\frac{1}{n}\sum \mathbb{E}[d(Y_i, \hat{Y}_i)] \leq D$ while the normalized uncertainty about X^n at the decoder is lower-bounded, i.e., $\frac{1}{n}H(X^n|\varphi(X^n, Y^n)) \geq E$ for a non-negative $E \leq H(X)$. This requirement is different from the privacy constraint usually considered in information-theoretic secrecy (e.g., [69], [53], [139], and [143]), in that here the utility and privacy are measured with respect to two different sources Y and X, respectively. In this sense, X and Y correspond to the private and non-private sources, respectively. The correlation between X and Y makes the utility and privacy constraints contradicting.

We study a similar model as Yamamoto's but for *lossless* compression. Clearly, if no side information is available to the decoder, then the eavesdropper can obtain as much information about X^n as the legitimate decoder and hence only trivial levels of privacy can be achieved when lossless compression of Y is required. We, therefore, assume that side information is provided at the decoder, as depicted in Fig. 2.1.

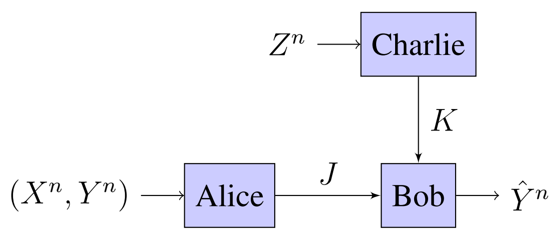


Figure 2.1: Yamamoto's lossless source coding.

A $(2^{nR_A}, 2^{nR_C}, n)$ code for private lossless compression in this setup is composed of two encoding functions at Alice and Charlie, respectively, $f_A : \mathcal{X}^n \times \mathcal{Y}^n \to [2^{nR_A}]$ and $f_C : \mathcal{Z}^n \to [2^{nR_C}]$, and a decoder at Bob, $f_B : [2^{nR_A}] \times [2^{nR_C}] \to \hat{\mathcal{Y}}^n$, where (X^n, Y^n, Z^n) are n independent and identically distributed (i.i.d.) copies of (X, Y, Z) with joint distribution $P(x, y, z)$. We assume that both encoders communicate to Bob over noiseless channels; however, the channel between Alice and Bob is subject to eavesdropping and hence a passive party can have access to the message J transmitted over this channel. A triple $(R_A, R_C, \Delta) \in \mathbb{R}^3_+$ is said to be achievable if for any $\varepsilon > 0$, there exists a $(2^{nR_A}, 2^{nR_C}, n)$

code for n large enough such that

$$\begin{aligned} \Pr(f_B(J,K) \neq Y^n) &< \varepsilon, & (2.1) \\ \frac{1}{n} H(X^n|J) &\geq \Delta - \varepsilon, & (2.2) \end{aligned}$$

where $J := f_A(X^n, Y^n)$ and $K := f_C(Z^n)$. We denote the set of all achievable triples (R_A, R_C, Δ) by $\mathcal{R}$. One special case of interest is when J contains absolutely no information about the private source, that is, when J is independent of X^n, which is called perfect privacy.

We note that for a special case of $X = Y$, inner and outer bounds on the achievable region were initially presented in [68, Theorem 3.1], although these bounds do no match in general. Tight bounds were then given in [139, Theorem 1] whose achievability resembles the binning scheme proposed by Wyner [152] for standard source coding with coded side information at the decoder. This therefore shows that the privacy constraint (2.2) does not change the optimal scheme.

Theorem 2.1. *For any achievable triple* $(R_A, R_C, \Delta) \in \mathcal{R}$ *we have*

$$\begin{aligned} R_A &\geq H(Y|V), \\ R_C &\geq I(Z;V), \\ \Delta &\leq I(X,Y;V) + H(X|U) - H(Y|U), \end{aligned}$$

for some auxiliary random variables $V \in \mathcal{V}$ *and* $U \in \mathcal{U}$ *such that* $P(x,y,z,u,v) = P(x,y,z)P(v|z)P(u|x,y)$ *with* $|\mathcal{U}| \leq |\mathcal{X}| \times |\mathcal{Y}| + 1$ *and* $|\mathcal{V}| \leq |\mathcal{Z}| + 2$.

Remark 2.2. It can be shown that the bound for Δ is maximized when $U = Y$. It is because we can write $H(X|U) - H(Y|U) \leq H(XY|U) - H(Y|U) = H(X|UY) \leq H(X|Y)$.

Proof. First note that Bob is required to reconstruct Y^n losslessly given J and K, and thus by Fano's inequality we have

$$H(Y^n|J, K) \leq n\varepsilon_n, \tag{2.3}$$

where $\varepsilon_n \to 0$ as $n \to \infty$.

We start by obtaining a lower bound for R_A as follows:

$$\begin{aligned}
nR_A &\geq H(J) \geq H(J|K) = H(Y^n, J|K) - H(Y^n|J, K) \\
&\overset{(a)}{\geq} H(Y^n, J|K) - n\varepsilon_n \geq H(Y^n|K) - n\varepsilon_n = \sum_{i=1}^{n} H(Y_i|Y^{i-1}, K) - n\varepsilon_n \\
&\geq \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^{i-1}, K) - n\varepsilon_n \\
&\overset{(b)}{=} \sum_{i=1}^{n} H(Y_i|V_i) - n\varepsilon_n \overset{(c)}{=} nH(Y_Q|V_Q, Q) - n\varepsilon_n \overset{(d)}{=} nH(Y|V) - n\varepsilon_n,
\end{aligned}$$

where (a) follows from (2.3), and (b) is due to the definition $V_i := (Y^{i-1}, X^{i-1}, K)$. In (c) we have introduced a time-sharing random variable Q which is distributed uniformly over $\{1, 2, \ldots, n\}$ and is independent of (X^n, Y^n, Z^n). In (d) we have defined $V := (V_Q, Q)$ and used the fact that Y_Q has the distribution of Y and hence we can replace Y_Q with Y. Next we obtain a lower bound on R_C:

$$\begin{aligned}
nR_C &\geq H(K) = I(Z^n; K) = \sum_{i=1}^{n} I(Z_i; K|Z^{i-1}) \\
&\overset{(a)}{=} \sum_{i=1}^{n} I(Z_i; K, Z^{i-1}) \overset{(b)}{=} \sum_{i=1}^{n} I(Z_i; K, Z^{i-1}, X^{i-1}, Y^{i-1}) \\
&\geq \sum_{i=1}^{n} I(Z_i; K, X^{i-1}, Y^{i-1}) = nI(Z_Q; V_Q, Q) = nI(Z; V),
\end{aligned}$$

where (a) is due to the fact that Z_i is independent of Z^{i-1} for each i and (b) follows from the Markov chain relation $Z_i \multimap\!\!- (K, Z^{i-1}) \multimap\!\!- (Y^{i-1}, X^{i-1})$.

We now upper bound the equivocation that any asymptotically lossless scheme produces. First we show the following identity which expresses $H(X^n|J)$ in terms of $H(Y^n|J)$ and some auxiliary terms:

$$H(X^n|J) - H(Y^n|J) = \sum_{i=1}^{n} [H(X_i|U_i) - H(Y_i|U_i)], \tag{2.4}$$

where $U_i := (X_{i+1}^n, Y^{i-1}, J)$. We will prove a general version of this identity later in Lemma 2.5. The equivocation can then be upper bounded as

$$\begin{aligned}
n(\Delta - \varepsilon) &\le H(X^n|J) \\
&\overset{(a)}{=} H(Y^n|J) + \sum_{i=1}^{n} [H(X_i|U_i) - H(Y_i|U_i)] \\
&= H(Y^n|K, J) + I(Y^n; K|J) + \sum_{i=1}^{n} [H(X_i|U_i) - H(Y_i|U_i)] \\
&\le n\varepsilon_n + I(K; Y^n, X^n|J) + \sum_{i=1}^{n} [H(X_i|U_i) - H(Y_i|U_i)] \\
&\overset{(b)}{\le} n\varepsilon_n + I(K; X^n, Y^n) + \sum_{i=1}^{n} [H(X_i|U_i) - H(Y_i|U_i)] \\
&= n\varepsilon_n + \sum_{i=1}^{n} I(K; X_i, Y_i|X^{i-1}, Y^{i-1}) + \sum_{i=1}^{n} [H(X_i|U_i) - H(Y_i|U_i)] \\
&= n\varepsilon_n + \sum_{i=1}^{n} I(K, X^{i-1}, Y^{i-1}; X_i, Y_i) + \sum_{i=1}^{n} [H(X_i|U_i) - H(Y_i|U_i)] \\
&= n\varepsilon_n + \sum_{i=1}^{n} I(V_i; X_i, Y_i) + \sum_{i=1}^{n} [H(X_i|U_i) - H(Y_i|U_i)] \\
&= n\varepsilon_n + nI(V_Q; X_Q, Y_Q|Q) + n[H(X_Q|U_Q, Q) - H(Y_Q|U_Q, Q)]
\end{aligned}$$

$$\stackrel{(c)}{=} n\varepsilon_n + nI(V_Q, Q; X_Q, Y_Q) + n[H(X_Q|U_Q, Q) - H(Y_Q|U_Q, Q)]$$
$$\stackrel{(d)}{=} n\varepsilon_n + n[I(V; X, Y) + H(X|U) - H(Y|U)],$$

where (a) follows from (2.4), (b) follows from the Markov chain relation $J \multimap\!\!- (X^n, Y^n) \multimap\!\!- K$ and hence $I(X^n, Y^n; K|J) \leq I(X^n, Y^n; K)$, (c) is due to the fact that Q is independent of (X_Q, Y_Q) and in (d) we have introduced $U := (U_Q, Q)$.

We note that by definitions of U and V, the Markov chain conditions $(X, Y) \multimap\!\!- Z \multimap\!\!- V$ and $Z \multimap\!\!- (X, Y) \multimap\!\!- U$ are satisfied. The cardinality bounds given in the statement of the theorem can be proved using the Support Lemma [40]. □

Remark 2.3. As mentioned earlier, the special case $X = Y$ is studied in [139] where it is shown that for any achievable triple (R_A, R_C, Δ), the optimal equivocation satisfies $\Delta \leq I(Y; V)$. We see that Theorem 2.1 yields the same result and thus gives a tight bound in this special case.

In practice, the private source X might not be directly available to Alice. In this case, her mapping is $f_A : \mathcal{Y}^n \to \{1, 2, \dots, 2^{nR_A}\}$ and the above theorem reduces to the following corollary.

Corollary 2.4. *When the source X^n is not available to Alice, any achievable triple (R_A, R_C, Δ) satisfies*

$$\begin{aligned} R_A &\geq H(Y|V), \\ R_C &\geq I(Z; V), \\ \Delta &\leq I(Y; V) + H(X|U) - H(Y|U), \end{aligned}$$

for some $U \in \mathcal{U}$ and $V \in \mathcal{V}$ such that $P(x, y, z, u, v) = P(x, y, z)P(v|z)P(u|y)$ and

$|\mathcal{U}| \leq |\mathcal{Y}| + 1$ *and* $|\mathcal{V}| \leq |\mathcal{Z}| + 2$.

Proof. The proof follows easily from the proof of Theorem 2.1. In particular, introducing $V_i := (Y^{i-1}, K)$ and $U_i := (X_{i+1}^n, Y^{i-1}, J)$, we can follow easily the chain of inequalities given for the equivocation analysis with appropriate modifications. Since now $J = f_A(Y^n)$, we have $(X_i, Z_i) -\!\!\circ\!\!- Y_i -\!\!\circ\!\!- U_i$. $\square$

2.3 Yamamoto's Lossless Source Coding: Uncoded Side Information at Eve

We now turn our focus to the case where there is an eavesdropper, Eve, with perfect access to the channel from Alice to Bob and also side information E^n. Unlike in the last section, in this model the achievable (R_A, R_C, Δ) has not been fully characterized in the case of $X = Y$. However, Gündüz et al. [69] and Probhakaran and Ramchandran [119] showed that if $R_C > H(Z)$, that is uncoded side information is available at Bob, then (R_A, Δ) is an achievable pair if and only if $R_A \geq H(Y|Z)$ and $\Delta \leq \max[I(Y;Z|U) - I(Y;E|U)]$ where the maximization is taken over U that satisfies $Z -\!\!\circ\!\!- Y -\!\!\circ\!\!- U$, thus providing a full single-letter characterization of the achievable rate-equivocation region. In this section, we assume coded side information is available at Bob and Eve has uncoded side information E^n, as depicted in Fig. 2.2. In the following, we assume that the Eve's side information E^n forms the Markov chain $X^n -\!\!\circ\!\!- Y^n -\!\!\circ\!\!- E^n$.

2.3.1 A Converse Result

The achievable (R_A, R_C, Δ) in this model is defined similarly as before with the utility constraint (2.1) and the privacy constraint

$$\frac{1}{n} H(X^n | E^n, J) \geq \Delta - \varepsilon. \tag{2.6}$$

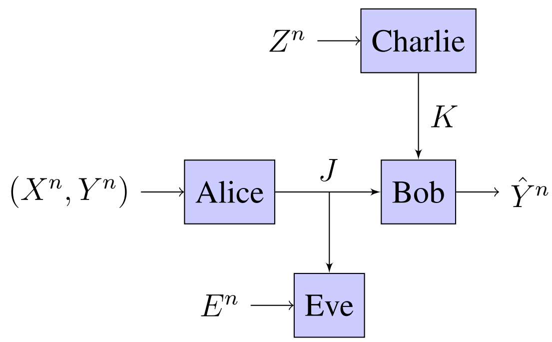


Figure 2.2: Yamamoto's lossless source coding with eavesdropper having side information.

Before we get to an outer bound for the achievable region of this model, we need to state the following lemma which is a generalization of identity (2.4) that we used in the proof of Theorem 2.1.

Lemma 2.5. *Let* (J, X^n, Y^n, E^n) *be jointly distributed according to* $P(j, x^n, y^n, e^n)$*. Then we can write:*

$$H(X^n|E^n, J) - H(Y^n|E^n, J) = \sum_{i=1}^{n} [H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)]$$

where $U_i := (X_{i+1}^n, Y^{i-1}, E^{-i}, J)$ *for each* $i \in [n]$ *and* $E^{-i} := (E^{i-1}, E_{i+1}^n)$.

Proof. We can write

$$\begin{aligned}
0 &\overset{(a)}{=} \sum_{i=1}^{n} I(Y_i, E_i; X_{i+1}^n, E_{i+1}^n|J, Y^{i-1}, E^{i-1}) - I(Y^{i-1}, E^{i-1}; X_i, E_i|J, X_{i+1}^n, E_{i+1}^n) \\
&= H(Y^n, E^n|J) - H(X^n, E^n|J) \\
&\quad - \sum_{i=1}^{n} [H(Y_i, E_i|X_{i+1}^n, Y^{i-1}, E^{-i}, J) - H(X_i, E_i|X_{i+1}^n, Y^{i-1}, E^{-i}, J)] \\
&= H(Y^n|E^n, J) - H(X^n|E^n, J) \\
&\quad - \sum_{i=1}^{n} [H(Y_i|E_i, X_{i+1}^n, Y^{i-1}, E^{-i}, J) - H(X_i|E_i, X_{i+1}^n, Y^{i-1}, E^{-i}, J)]
\end{aligned}$$

$$\stackrel{(b)}{=} \quad H(Y^n|E^n, J) - H(X^n|E^n, J) - \sum_{i=1}^{n}[H(Y_i|E_i, U_i) - H(X_i|E_i, U_i)],$$

where (a) follows from Ciszár sum identity [87, page 25], and in (b) we used the definition of U_i. $\square$

Theorem 2.6. *The set of all achievable triples $(R_A, R_C, \mathit{\Delta})$ for this model when Eve is provided with side information E^n and $E^n -\!\!\circ\!\!- Y^n -\!\!\circ\!\!- X^n$, satisfies*

$$\begin{aligned} R_A &\geq H(Y|V), \\ R_C &\geq I(Z; V), \\ \mathit{\Delta} &\leq I(X, Y; V) - I(X, Y; E|U) + H(X|E, U) - H(Y|E, U), \end{aligned}$$

for some U and V which form $(Z, E) -\!\!\circ\!\!- (X, Y) -\!\!\circ\!\!- U$ and $(X, Y, E) -\!\!\circ\!\!- Z -\!\!\circ\!\!- V$.

Proof. The lower bounds for both R_A and R_C follow along the same lines as in the proof of Theorem 2.1. We shall show the upper bound for the equivocation. We note that since Bob is required to reconstruct Y^n losslessly, Fano's inequality implies that

$$H(Y^n|J, K) \leq n\varepsilon_n \tag{2.7}$$

for $\varepsilon_n \to 0$ as $n \to \infty$. As before, let $J = f_A(X^n, Y^n)$ and $K = f_C(Z^n)$.

The upper bound for the equivocation is obtained as follows:

$$\begin{aligned} H(X^n|E^n, J) &\stackrel{(a)}{=} H(Y^n|E^n, J) + \sum_{i=1}^{n}[H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\ &= H(Y^n|J, K) + I(Y^n; K|J) - I(Y^n; E^n|J) \\ &\quad + \sum_{i=1}^{n}[H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} \quad n\varepsilon_n + I(X^n, Y^n; K|J) - I(Y^n; E^n|J) \\
&\qquad + \sum_{i=1}^{n} [H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\overset{(c)}{\leq} \quad n\varepsilon_n + I(X^n, Y^n; K) - I(Y^n; E^n) + I(E^n; J) \\
&\qquad + \sum_{i=1}^{n} [H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\overset{(d)}{=} \quad n\varepsilon_n + \sum_{i=1}^{n} [I(X_i, Y_i; K, X^{i-1}, Y^{i-1}) - I(Y_i, X_i; E_i) + I(E_i; J, E^{i-1}) \\
&\qquad + H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\overset{(e)}{\leq} \quad n\varepsilon_n + \sum_{i=1}^{n} [I(X_i, Y_i; V_i) - I(Y_i, X_i; E_i) + I(E_i; U_i)] \\
&\qquad + \sum_{i=1}^{n} [H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\overset{(f)}{=} \quad n\varepsilon_n + \sum_{i=1}^{n} [I(X_i, Y_i; V_i) - I(Y_i, X_i; E_i|U_i)] \\
&\qquad + \sum_{i=1}^{n} [H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\overset{(g)}{=} \quad n\varepsilon_n + I(X_Q, Y_Q; V_Q, Q) - I(Y_Q, X_Q; E_Q|U_Q, Q) \\
&\qquad + H(X_Q|E_Q, U_Q, Q) - H(Y_Q|E_Q, U_Q, Q)],
\end{aligned}
$$

where (a) follows from Lemma 2.5 and (b) is due to (2.7). Since $K \multimap (X^n, Y^n) \multimap J$ and $E^n \multimap Y^n \multimap J$, we have $I(X^n, Y^n; K|J) \leq I(X^n, Y^n; K)$ and $I(Y^n; E^n|J) = I(Y^n; E^n) - I(E^n; J)$ and hence (c) follows. We again used the Markov chain relation $E^n \multimap Y^n \multimap X^n$ in (d). The definition $V_i := (K, X^{i-1}, Y^{i-1})$ and the fact that $I(E_i; J, E^{i-1}) \leq I(E_i; U_i)$ are used in (e). Note that since $U_i \multimap (X_i, Y_i) \multimap E_i$ we have in (f) that $I(X_i, Y_i; E_i|U_i) = I(X_i, Y_i; E_i) - I(E_i; U_i)$. The proof completes by introduction of a time sharing random variable Q uniformly distributed over $\{1, 2, \ldots, n\}$

and independent of (X^n, Y^n, Z^n, E^n) and letting $X = X_Q, Y = Y_Q$, $E = E_Q$, $V = (V_Q, Q)$ and $U = (U_Q, Q)$. $\square$

Remark 2.7. Setting $E^n = \emptyset$ and thus removing the eavesdropper's side information, Theorem 2.6 yields $\Delta \leq I(X, Y; V) + H(X|U) - H(Y|U)$ and hence Theorem 2.6 subsumes Theorem 2.1.

In the simple case of $X = Y$, the optimal scheme when coded side information is available at Bob and $E^n = \emptyset$ is proposed in [139] which is shown to resemble the binning scheme of Wyner in [152]. Although, a tight bound for the equivocation when E^n is available is not yet known, Theorem 2.6, specialized to $X = Y$, implies

$$\Delta \leq I(Y; V) - I(Y; E|U),$$

for auxiliary random variables U and V which form Markov chains $V -\!\!\circ\!\!- Z -\!\!\circ\!\!- (Y, E)$ and $U -\!\!\circ\!\!- Y -\!\!\circ\!\!- (Z, E)$.

2.3.2 A Coding Scheme When Bob Has Uncoded Side Information

As a special case, we consider the case where Alice does not see the private source and also $R_C > H(Z)$ (i.e., Bob has uncoded side information). In this case, Theorem 2.6 implies that the best achievable equivocation is upper bounded by

$$\max[I(Y; Z) - I(Y; E|U) + H(X|E, U) - H(Y|E, U)],$$

where the maximization is taken over U which forms the Markov chain relation $U -\!\!\circ\!\!- Y -\!\!\circ\!\!- (Z, E, X)$. In the following, we give a simple coding scheme which incurs a smaller equivocation and is thus suboptimal. In fact, if the above maximization results in a

U which is independent of Z, then the following coding scheme is optimal. On the other hand, if the maximization results in a U which is constant, then it implies that Slepian-Wolf binning is optimal, because if Alice uses Slepian-Wolf binning then the equivocation is equal to $H(X|E) - H(Y|Z)$, as observed in [119].

Theorem 2.8. *When X^n is not given to Alice and Bob observes side information Z^n, then $(R_A, \varDelta)$ which satisfies*

$$\begin{aligned} R_A &\geq H(Y|Z), \\ \varDelta &\leq I(Y;Z|U) - I(Y;E|U) + H(X|E,U) - H(Y|E,U), \end{aligned}$$

is achievable where the auxiliary random variable U forms the Markov chain $(X, Z, E) -\!\circ\!- Y -\!\circ\!- U$.

Proof. Our scheme is similar to the ones proposed in [69] and [41]. Given Y^n, we generate $2^{n(I(Y;U)+\varepsilon)}$ independent codewords of length n, $U^n(w)$, $w \in \{1, 2, \ldots, 2^{nI(Y;U)+\varepsilon}\}$ according to $\prod_{i=1}^n P(u_i)$. We then uniformly bin all the U^n sequences into $2^{n(I(Y;U)-I(U;Z))}$ bins. Let $B(i)$ be the indices assigned to bin i. There are approximately $2^{nI(U;Z)}$ indices in each bin. We also uniformly bin Y^n sequences into $2^{n(H(Y|U,Z)+\varepsilon)}$ bins and let $C(k)$ be the set of sequences Y^n in bin k. Alice adopts a two-part encoding scheme. Given Y^n, Alice, in the first part, looks for a codeword $U^n(w)$ such that $(Y^n, U^n(w)) \in \mathcal{A}_{YU}^n$, where $\mathcal{A}_{YU}^n$ denotes the set of all strongly typical $(y^n, u^n) \in \mathcal{Y}^n \times \mathcal{U}^n$ with respect to the distribution $P(y, u)$. She then reveals the bin index J_1 such that $w \in B(J_1)$. In the second part, she reveals J_2 such that $Y^n \in C(J_2)$.

Given J_1, J_2 and Z^n, Bob can find, with high probability, $U^n(w)$ such that $w \in B(J_1)$ and $(U^n(w), Z^n) \in \mathcal{A}_{ZU}^n$. It is then clear from the Slepian-Wolf theorem that Bob can recover Y^n with high probability given $U^n(w)$, Z^n, and J_2. The rate of this encoder is clearly equal to $H(Y|U,Z)+I(Y;U)-I(U;Z)=H(Y|Z)$.

The equivocation for this scheme can be found as

$$\begin{aligned}
H(X^n|J_1,J_2,E^n) &= H(X^n|J_1,E^n)-I(X^n;J_2|J_1,E^n)\\
&\geq H(X^n|U^n,E^n)-H(J_2)\\
&\overset{(a)}{\geq} H(X^n|U^n,E^n)-nH(Y|U,Z)\\
&\overset{(b)}{\geq} n[H(X|U,E)-H(Y|U,Z)]\\
&= n[H(X|E,U)-H(Y|E,U)+I(Y;Z|U)-I(Y;E|U)],
\end{aligned}$$

where (a) follow from the fact that J_2 is a random variable over a set of size $2^{nH(Y|U,Z)}$ and (b) is proved as follows:

$$\begin{aligned}
H(X^n|U^n,E^n) &= \sum_{(u^n,e^n)\in\mathcal{U}^n\times\mathcal{E}^n} P(u^n,e^n)H(X^n|U^n=u^n,E^n=e^n)\\
&\geq \sum_{(u^n,e^n)\in\mathcal{T}^n_{U,E}} P(u^n,e^n)H(X^n|U^n=u^n,E^n=e^n)\\
&= \sum_{(u^n,e^n)\in\mathcal{T}^n_{U,E}} P(u^n,e^n)\left[-\sum_{x^n\in\mathcal{X}^n} P(x^n|u^n,e^n)\log(P(x^n|u^n,e^n))\right]\\
&\geq \sum_{(u^n,e^n)\in\mathcal{T}^n_{U,E}} P(u^n,e^n)\left[-\sum_{x^n\in\mathcal{T}^n_{X|u^n,e^n}} P(x^n|u^n,e^n)\log(P(x^n|u^n,e^n))\right]\\
&\overset{(c)}{\geq} n(H(Y|U,E)-\delta_n)\sum_{(u^n,e^n)\in\mathcal{T}^n_{U,E}} P(u^n,e^n)\left[\sum_{x^n\in\mathcal{T}^n_{X|u^n,e^n}} P(x^n|u^n,e^n)\right]\\
&= n(H(Y|U,E)-\delta_n)\sum_{(u^n,e^n)\in\mathcal{T}^n_{U,E}} P(u^n,e^n)\left[\Pr\{(u^n,e^n,X^n)\in\mathcal{T}^n_{X|u^n,e^n}\}\right]\\
&\overset{(d)}{\geq} n(H(Y|U,E)-\delta_n)(1-\delta'_n),
\end{aligned}$$

where $\mathcal{T}^n_{U,E}$ denotes the set of typical sequences (u^n, e^n) and (c) is due to the property of typical sequences; in particular for typical x^n sequence with respect to $P(x^n|u^n, e^n)$ for $(u^n, e^n) \in \mathcal{T}^n_{U,E}$ we have $P(x^n|u^n, e^n) \leq 2^{-(n(H(X|U,E)-\delta(n)))}$ for $\delta_n \to 0$ as $n \to \infty$. We invoked Markov lemma [87, Lemma 12.1] in (d) to conclude that for the Markov chain relation $(X, E) -\!\circ\!- Y -\!\circ\!- U$ we have $(x^n, y^n, e^n, u^n) \in \mathcal{T}^n_{X,Y,E,U}$ and hence $\Pr\{(u^n, e^n, X^n) \in \mathcal{T}^n_{U,E,X}\} > 1 - \delta'_n$ for each pair $(u^n, e^n) \in (u^n, e^n) \in \mathcal{T}^n_{U,E}$ and $\delta'_n \to 0$ as $n \to \infty$. $\square$

2.4 Concluding Remarks

Having combined the idea of compression of private and non-private sources of Yamamoto [155] with secure source coding problem (e.g. [69], [139] and [119]), we introduced a lossless source coding problem in which, given a two-dimensional source (X^n, Y^n), the encoder must compress the source into an index J with rate R_A such that the receiver recovers Y^n losslessly and simultaneously reveals only little information about X^n. This model differs from typical information-theoretic secrecy models in that the utility and privacy constraints are defined for two different sources and thus provides a more general utility-equivocation tradeoff.

We gave converse results for compression rates and also the information leakage rate (or equivocation) which reduce to known results in the special case of $X = Y$. In particular, with this simplifying assumption, Theorem 2.1 and Theorem 2.8 reduce to [139, Theorem 1] and [69, Corollary 3.2].

However, it is not clear at the moment that the bounds are tight in general. Constructing an achievability scheme for the most general case (i.e., the setting of Theorem 2.6) is the subject of our future studies.

Chapter 3

Information Extraction Under an Information-Theoretic Privacy Constraint: Discrete Case

3.1 Overview and Motivation

With the emergence of user-customized services, there is an increasing desire to balance between the need to share data and the need to protect sensitive and private information. For example, individuals who join a social network are asked to provide information about themselves which might compromise their privacy. However, they agree to do so, to some extent, in order to benefit from the customized services such as recommendations and personalized searches. Hence, each user needs to make a balance between the benefit he receives from the customized services and the level of privacy they wish to maintain. As another example, suppose a software company wants to gather statistical information on how people use its software. Since many users might have used the software to handle some personal or sensitive information -for example, a browser for anonymous web surfing or a financial management software- they may not want to share their data with the company. On the other hand, the company cannot legally collect the raw data either, so it needs to entice its users. In all these situations, a tradeoff in a conflict between utility advantage and privacy breach is required and the question is how to achieve this tradeoff. For example, how can a company collect high-quality aggregate information about users while strongly guaranteeing to its users that it is not storing user-specific information?

To deal with such privacy considerations, Warner [145] proposed the *randomized response model*[1] in which each individual user randomizes his own data using a local randomizer (i.e., a noisy channel) before sharing the data to an untrusted data collector to be aggregated. As opposed to *conditional security*, e.g. [25, 45, 125], the randomized response model assumes no limit on the computational capability of the adversary and thus it provides *unconditional* privacy. This model, in which the control of private data remains in the users' hands, has recently regained attention after Warner within the information theory [79, 83, 116, 126, 127] and the computer science communities [47, 82].

There have been several studies on the tradeoff between privacy and utility for different examples of randomized response models with different choices of utility and privacy measures. For instance, Duchi et al. [47] studied the optimal ε-locally differentially private mechanism (defined in Section 1.2) $\mathcal{M} : X \to Z$ which minimizes the risk of estimation of a parameter θ related to P_X. Kairouz et al. [82] studied an optimal ε-locally differentially private mechanism in the sense of mutual information, where an individual would like to release an ε-locally differentially private version Z of X that preserves as much information about X as possible. Calmon et al. [32] proposed a novel privacy measure (which includes maximal correlation and chi-square correlation) between X and Z and studied the optimal privacy mechanism (according to their privacy measure) which minimizes the error probability $\Pr(\hat{X}(Z) \neq X)$ for any estimator $\hat{X} : Z \to X$.

In all above examples of randomized response models, given a private source, denoted by X, the mechanism generates Z which can be publicly displayed without breaching the

[1] For obvious reasons, this model is recently referred to as the *local privacy model* [47].

desired privacy level. However, in a more realistic model of privacy, we can assume that for any given private data X, nature generates Y, via a fixed channel $P_{Y|X}$ (See Fig. 1.1). Now we aim to release a public display Z of Y such that the amount of information in Y is preserved as much as possible while Z satisfies a privacy constraint with respect to X. To motivate this model, consider two communicating agents Alice and Bob. Alice possesses Y and ultimately wants to reveal it to Bob in order to receive a payoff. However, she is worried about her private data, represented by X, which is correlated with Y. For instance, X might represent her precise location and Y represents measurement of traffic load of a route she has taken. She wants to reveal these measurements to an online road monitoring system to receive some utility. However, she does not want to reveal too much information about her exact location. In such situations, utility is measured with respect to Y and privacy is measured with respect to X. Our goal is to characterize the maximum payoff that Alice can get from Bob (by revealing Z to him) without compromising her privacy. Thus, it is of interest to characterize such competing objectives in the form of a quantitative tradeoff. Such a characterization provides a controllable balance between utility and privacy.

3.2 Main Contributions

The main contributions of this chapter are as follows:

- Using mutual information as a measure of both utility and privacy, we formulate the corresponding utility-privacy tradeoff for discrete random variables X and Y via the rate-privacy function, $g(P_{XY}, \cdot)$. If $g(P_{XY}, \varepsilon) = R$, then mutual information $I(Y; Z)$ is maximally equal to R among all channels $P_{Z|Y}$ satisfying $I(X; Z) \leq \varepsilon$. We obtain a necessary and sufficient condition for $g(P_{XY}, 0) = 0$. Assuming this property, we show that

$$\lim_{\varepsilon \downarrow 0} \frac{g(P_{XY}, \varepsilon)}{\varepsilon} = \sup_{\substack{P_{Z|Y}: X -\!\!\circ\!\!- Y -\!\!\circ\!\!- Z, \\ I(X;Z)>0}} \frac{I(Y;Z)}{I(X;Z)}, \tag{3.1}$$

which establishes a connection between $g(P_{XY}, \varepsilon)$ and a "reverse" notion of strong data processing inequality [4, 10, 11]. This connection is recently studied in [31] to mirror all the results of [10] in the context of privacy. In Chapter 4, we derive some results for continuous random variables which accentuate this connection.

- Inspired by (3.1), we focus on the rate of increase $g'(P_{XY}, 0)$ of $g(P_{XY}, \varepsilon)$ at $\varepsilon = 0$ and show that this rate characterizes the behavior of $g(P_{XY}, \varepsilon)$ for any $\varepsilon \geq 0$ provided that $g(P_{XY}, 0) = 0$. In particular, we show that

$$g'(P_{XY}, 0) \geq \max_{y \in \mathcal{Y}} \frac{-\log P_Y(y)}{D(P_{X|Y}(\cdot|y) \| P_X(\cdot))},$$

which leads to a lower bound to the reverse strong data processing coefficient in (3.1). The same connection can be established for the strong data processing inequality [10, 11]. Letting

$$\Gamma(R) := \max_{\substack{P_{Z|Y}: X -\!\!\circ\!\!- Y -\!\!\circ\!\!- Z \\ I(Y;Z) \leq R}} I(X;Z),$$

one can easily show that

$$\Gamma'(0) = \lim_{R \to 0} \frac{\Gamma(R)}{R} = \sup_{\substack{P_{Z|Y}: \\ X -\!\!\circ\!\!- Y -\!\!\circ\!\!- Z}} \frac{I(X;Z)}{I(Y;Z)},$$

and hence the rate of increase of $\Gamma(R)$ at $R = 0$ characterizes the strong data processing coefficient. Note that here we have always $\Gamma(0) = 0$.

- We derive lower and upper bounds of $g(P_{XY}, \varepsilon)$ for any ε in its domain. In particular, we show that $g(P_{XY},\varepsilon) \geq \frac{\varepsilon H(Y)}{I(X;Y)}$. We then obtain conditions on P_{XY} such that this bound is tight. For example, we show that if the channel from Y to X satisfies a certain notion of symmetry, then $g(P_{XY},\varepsilon) = \frac{\varepsilon H(Y)}{I(X;Y)}$, if and only if $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$. This now implies that in this case, we have

$$\sup_{\substack{P_{Z|Y}: \\ X -\!\!\circ\!\!- Y -\!\!\circ\!\!- Z}} \frac{I(Y;Z)}{I(X;Z)} = \frac{1}{I(X;Y)}.$$

We also show that $g(P_{XY},\varepsilon) \leq H(Y|X) + \varepsilon$, where the equality holds if Y is an erased version of X, or equivalently, $P_{Y|X}$ is an erasure channel.

- We propose an information-theoretic setting, the so-called "dependence dilution" coding problem, in which $g(P_{XY},\cdot)$ appears as a natural upper-bound for the achievable rate. Specifically, we examine the joint-encoder version of an *amplification-masking tradeoff*, a setting recently introduced by Courtade [36], and we show that the dual of $g(P_{XY},\cdot)$ upper bounds the masking rate.

3.3 Problem Formulation

Consider two random variables X and Y, defined over alphabets $\mathcal{X} = [M]$ and $\mathcal{Y} = [N]$, respectively, with a fixed joint distribution $P_{XY} = \mathsf{P}$. Let X represent the *private data* and let Y be the *observable data*, correlated with X and generated by the channel $P_{Y|X}$ predefined by nature, which we call the *observation channel*. Suppose that there exists a channel $P_{Z|Y}$ such that Z, the *displayed data* made available to public users, has a small mutual information with X. Such a channel is called the *privacy filter*. This setup is shown in Fig. 1.1. The objective is then to find a privacy filter which gives rise to the highest mutual information between Y and Z. To quantify this goal, we introduce the *rate-privacy* *function*[2] as

$$g(\mathsf{P}, \varepsilon) := \sup_{P_{Z|Y} \in \mathcal{D}_\varepsilon(\mathsf{P})} I(Y; Z), \tag{3.2}$$

where

$$\mathcal{D}_\varepsilon(\mathsf{P}) := \{P_{Z|Y} : X \multimap Y \multimap Z, I(X; Z) \leq \varepsilon\}. \tag{3.3}$$

Equivalently, we refer to $g(\mathsf{P}, \varepsilon)$ as the *privacy-constrained information extraction function*, as Z can be thought of as the extracted information from Y under privacy constraint $I(X; Z) \leq \varepsilon$.

Note that by the Markov condition $X \multimap Y \multimap Z$, we can always restrict $\varepsilon \geq 0$ to only $0 \leq \varepsilon < I(X; Y)$, because by the data processing inequality we have $I(X; Z) \leq I(X; Y)$ and hence for $\varepsilon \geq I(X; Y)$ the privacy constraint is always satisfied by setting $Z = Y$, which yields $g(\mathsf{P}, \varepsilon) = H(Y)$. Note also that using the Support Lemma [40, 87], one can readily show that it suffices to consider the random variable Z that is supported on an alphabet $\mathcal{Z}$ with cardinality $|\mathcal{Z}| \leq N + 1$. Moreover, the continuity of $P_{Z|Y} \to I(X; Z)$ implies that $\mathcal{D}_\varepsilon(\mathsf{P})$ is compact, and hence the supremum in (3.2) is indeed a maximum. We will show later that $g(\mathsf{P}, \cdot)$ is concave and strictly increasing on $[0, I(X; Y)]$ and hence the continuity of $I(Y; Z)$ and $I(X; Z)$ in $P_{Z|Y}$ implies that the feasible set $\mathcal{D}_\varepsilon(\mathsf{P})$ in (3.2) can be replaced by $\{P_{Z|Y} : X \multimap Y \multimap Z, I(X; Z) = \varepsilon\}$. For the sake of brevity, we denote $g(\mathsf{P}, \varepsilon)$ by $g(\varepsilon)$ when this does not cause confusion.

A dual representation of $g(\varepsilon)$, the so called *Privacy Funnel*, is introduced in [105] and [31] as the least information leakage about X such that the communication rate is greater

[2] Since mutual information is adopted for utility, the privacy-utility tradeoff characterizes the optimal *rate* for a given privacy level, where rate indicates the precision of the displayed data Z with respect to the observable data Y for a privacy filter, which suggests the name.

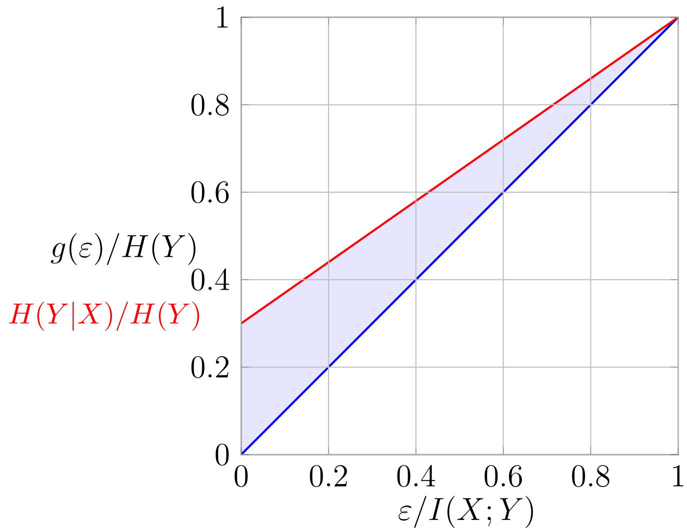


Figure 3.1: The region of $g(\varepsilon)$ in terms of $\frac{\varepsilon}{I(X;Y)}$ specified by (3.5), where the upper and lower bounds are straight lines of slopes $\frac{I(X;Y)}{H(Y)}$ and 1, respectively.

than a positive constant, i.e.,

$$t(\mathsf{P}, R) := \min_{\substack{P_{Z|Y}: X -\!\!\circ\!\!- Y -\!\!\circ\!\!- Z \\ I(Y;Z) \geq R}} I(X;Z). \tag{3.4}$$

Note that $t(\mathsf{P}, R) = \varepsilon$ if and only if $g(\varepsilon) = R$ for $R, \varepsilon > 0$ and also $g(0) = \max\{R : t(\mathsf{P}, R) = 0\}$.

3.4 Properties

Given $\varepsilon_1 < \varepsilon_2$ and a joint distribution P, we have $\mathcal{D}_{\varepsilon_1}(\mathsf{P}) \subset \mathcal{D}_{\varepsilon_2}(\mathsf{P})$, thus $g(\cdot)$ is non-decreasing, i.e., $g(\varepsilon_1) \leq g(\varepsilon_2)$. Using a similar technique as in [132, Lemma 1], Calmon et al. [31] showed that the mapping $R \mapsto \frac{t(\mathsf{P},R)}{R}$ is non-decreasing for $R > 0$. This, in fact, implies that $\varepsilon \mapsto \frac{g(\varepsilon)}{\varepsilon}$ is non-increasing for $\varepsilon > 0$. This observation leads to a lower bound for the rate-privacy function $g(\varepsilon)$ as described in the following lemma which demonstrates the possible range of the map $\varepsilon \mapsto g(\varepsilon)$ is as depicted in Fig. 3.1.

Lemma 3.1 ([31]). *The mapping $\varepsilon \mapsto \frac{g(\varepsilon)}{\varepsilon}$ is non-increasing on $(0, \infty)$. Moreover, $g(\varepsilon)$*

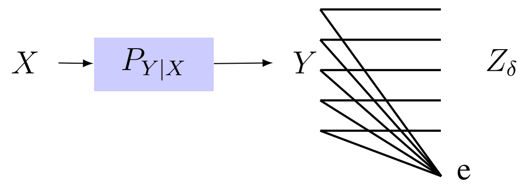


Figure 3.2: Privacy filter that achieves the lower bound in (3.5) where Z_δ is the output of an erasure privacy filter with erasure probability specified in (3.6).

lies between two straight lines as follows:

$$\varepsilon \frac{H(Y)}{I(X;Y)} \leq g(\varepsilon) \leq H(Y|X) + \varepsilon, \tag{3.5}$$

for $\varepsilon \in (0, I(X;Y))$.

Although the proof of the bounds in (3.5) is given in [31], in the following we give an insightful and conceptual proof of the lower bound. Consider a simple erasure channel $\mathsf{W}_\delta : Y \to Z_\delta$, shown in Fig. 3.2, with erasure probability $0 \leq \delta \leq 1$. It is easy to see that $I(X;Z_\delta) = \bar{\delta} I(X;Y)$ and $I(Y;Z_\delta) = \bar{\delta} H(Y)$ and consequently $\mathsf{W}_\delta \in \mathcal{D}_\varepsilon(\mathsf{P})$ if $\bar{\delta} = \frac{\varepsilon}{I(X;Y)}$. Hence for this particular choice of δ, we have $g(\varepsilon) \geq I(Y;Z_\delta) = \bar{\delta} H(Y)$ which proves the lower bound in (3.5). This observation shows that the lower bound in (3.5) is achieved by W_δ, illustrated in Fig. 3.2, with the erasure probability

$$\delta = 1 - \frac{\varepsilon}{I(X;Y)}. \tag{3.6}$$

We next show that $\varepsilon \mapsto g(\varepsilon)$ is concave and continuous.

Lemma 3.2. *The mapping* $\varepsilon \mapsto g(\varepsilon)$ *is concave.*

Proof. It suffices to show that for any $0 \le \varepsilon_1 < \varepsilon_2 < \varepsilon_3 \le I(X;Y)$, we have

$$\frac{g(\varepsilon_3) - g(\varepsilon_1)}{\varepsilon_3 - \varepsilon_1} \le \frac{g(\varepsilon_2) - g(\varepsilon_1)}{\varepsilon_2 - \varepsilon_1}, \tag{3.7}$$

which, in turn, is equivalent to

$$\left(\frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_3 - \varepsilon_1}\right) g(\varepsilon_3) + \left(\frac{\varepsilon_3 - \varepsilon_2}{\varepsilon_3 - \varepsilon_1}\right) g(\varepsilon_1) \le g(\varepsilon_2). \tag{3.8}$$

Let $P_{Z_1|Y} : Y \to Z_1$ and $P_{Z_3|Y} : Y \to Z_3$ be two optimal privacy filters in $\mathcal{D}_{\varepsilon_1}(\mathsf{P})$ and $\mathcal{D}_{\varepsilon_3}(\mathsf{P})$ with disjoint output alphabets $\mathcal{Z}_1$ and $\mathcal{Z}_3$, respectively. We introduce an auxiliary binary random variable $U \sim \mathsf{Bernoulli}(\lambda)$, independent of (X,Y), where $\lambda := \frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_3 - \varepsilon_1}$ and define the following random privacy filter $P_{Z_\lambda|Y}$: We pick $P_{Z_3|Y}$ if $U = 1$ and $P_{Z_1|Y}$ if $U = 0$, and let Z_λ be the output of this random channel which takes values in $\mathcal{Z}_1 \cup \mathcal{Z}_3$. Note that $(X,Y) \multimap\!\!- Z_\lambda \multimap\!\!- U$. Then we have

$$I(X;Z_\lambda) = I(X;Z_\lambda, U) = I(X;Z_\lambda|U) = \lambda I(X;Z_3) + (1-\lambda)I(X;Z_1) \le \varepsilon_2,$$

which implies that $P_{Z_\lambda|Y} \in \mathcal{D}_{\varepsilon_2}(\mathsf{P})$. On the other hand, we have

$$\begin{aligned} g(\varepsilon_2) \ge I(Y;Z_\lambda) &= I(Y;Z_\lambda,U) = I(Y;Z_\lambda|U) = \lambda I(Y;Z_3) + (1-\lambda)I(Y;Z_1), \\ &= \left(\frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_3 - \varepsilon_1}\right) g_{\varepsilon_3}(X;Y) + \left(\frac{\varepsilon_3 - \varepsilon_2}{\varepsilon_3 - \varepsilon_1}\right) g_{\varepsilon_1}(X;Y) \end{aligned}$$

which, according to (3.8), completes the proof. □

Remark 3.3. By the concavity of $\varepsilon \mapsto g(\varepsilon)$, we can show that $g(\varepsilon)$ is a *strictly* increasing function of $\varepsilon \le I(X;Y)$. To see this, assume there exists $\varepsilon_1 < \varepsilon_2 \le I(X;Y)$ such that

$g(\varepsilon_1) = g(\varepsilon_2)$. Since $\varepsilon \mapsto g(\varepsilon)$ is concave, then it follows that for all $\varepsilon \geq \varepsilon_2$, $g(\varepsilon) = g(\varepsilon_2)$ and since for $\varepsilon = I(X;Y)$, $g(I(X;Y)) = H(Y)$, it implies that for any $\varepsilon \geq \varepsilon_2$, we must have $g(\varepsilon) = H(Y)$ which contradicts the upper bound shown in (3.5).

The following is a direct implication of Lemma 3.2.

Corollary 3.4. *The mapping $\varepsilon \mapsto g(\varepsilon)$ is continuous for $\varepsilon \geq 0$.*

Remark 3.5. Using the concavity of the map $\varepsilon \mapsto g(\varepsilon)$, we can provide an alternative proof for the lower bound in (3.5). Note that point $(I(X;Y), H(Y))$ is on the curve $g(\cdot)$, and hence by concavity, the straight line $\varepsilon \mapsto \varepsilon \frac{H(Y)}{I(X;Y)}$ lies below the lower convex envelop of $g(\varepsilon)$, i.e., the chord connecting $(0, g(0))$ to $(I(X;Y), H(Y))$, and hence $g(\varepsilon) \geq \varepsilon \frac{H(Y)}{I(X;Y)}$. In fact, this chord yields a better lower bound for $g(\varepsilon)$ on $\varepsilon \in [0, I(X;Y]$ as

$$g(\varepsilon) \geq \varepsilon \frac{H(Y)}{I(X;Y)} + g(0) \left[1 - \frac{\varepsilon}{I(X;Y)}\right], \tag{3.9}$$

which reduces to the lower bound in (3.5) only if $g(0) = 0$.

3.5 Geometric Interpretation of $g(\varepsilon)$

Witsenhausen and Wyner [147] generalized Mrs. Gerber's Lemma [153]. In what follows, we describe their model, briefly illustrate their approach, and then we connect this approach to $g(\varepsilon)$. Before we describe their model, we need the following theorem.

Theorem 3.6 (Dubin's Theorem [46])**.** *If $\mathcal{C}$ is a compact and convex subset of a finite dimensional*[3] *vector space V and $\mathcal{C}'$ is the intersection of $\mathcal{C}$ with k hyperplanes, then every extreme point of $\mathcal{C}'$ can be written as a convex combination of $(k+1)$ extreme points of $\mathcal{C}$.*

[3]The original proof in [46] assumed general vector spaces and a linearly bounded and linearly closed convex set $\mathcal{C}$, see [146] for more details.

Given $(X,Y)\sim \mathsf{P}$ with marginals p_X and q_Y over $\mathcal{X}=[M]$ and $\mathcal{Y}=[N]$, respectively, let the (backward) channel from Y to X be denoted by T. The main question studied in [147] is to characterize $F_{\mathsf{T}}(\mathsf{q}_Y,\,\cdot\,):[0,H(Y)]\to[0,H(X)]$, defined as

$$F_{\mathsf{T}}(\mathsf{q}_Y,\mathit{\Delta}) \coloneqq \min_{\substack{P_{Z|Y}:X -\!\!\circ\!\!- Y -\!\!\circ\!\!- Z,\\ H(Y|Z)\geq \mathit{\Delta}}} H(X|Z). \tag{3.10}$$

As before, the Support Lemma implies that it is sufficient to consider Z supported over $\mathcal{Z}$ with cardinality $|\mathcal{Z}|=N+1$. Now consider $\mathcal{S}\subseteq\mathbb{R}^{N+1}$ given by $\mathcal{S}=\{(\mathsf{q},H(\mathsf{q}),H((\mathsf{T}\mathsf{q})_{\mathcal{X}})) : \mathsf{q}\in\mathcal{P}_{\mathcal{Y}}\}$, where $(\mathsf{T}\mathsf{q})_{\mathcal{X}}\in\mathcal{P}_{\mathcal{X}}$ is the marginal distribution of X when $Y\sim\mathsf{q}$. Clearly, setting $\mathsf{q}=\mathsf{q}_Y$, we have $H(\mathsf{q})=H(Y)$ and $H((\mathsf{T}\mathsf{q})_{\mathcal{X}})=H(X)$. Let $\mathcal{C}$ be the convex hull of $\mathcal{S}$. By definition, any point in $\mathcal{C}$ can be written as $\sum_{i=1}^{N+1}\omega_i(\mathsf{q}_i,H(\mathsf{q}_i),H((\mathsf{T}\mathsf{q}_i)_{\mathcal{X}}))$, where $\sum_{i=1}^{N+1}\omega_i=1$, $\omega_i\geq 0$, and $\mathsf{q}_i\in\mathcal{P}_{\mathcal{Y}}$ for $i\in[N+1]$. Consequently, $\mathcal{C}$ can be written as

$$\mathcal{C}=\left\{(\mathsf{q},H(Y'|Z),H(X'|Z)) : Y'\sim\mathsf{q}, X'\sim(\mathsf{T}\mathsf{q})_{\mathcal{X}}, P_Z(i)=\omega_i, P_{Y'|Z}(\cdot|i)=\mathsf{q}_i, i\in[N+1]\right\}.$$

Clearly, we have $(X',Y')\sim\mathsf{P}$ if and only if $\mathsf{q}=\mathsf{q}_Y$. We have

$$F_{\mathsf{T}}(\mathsf{q}_Y,\mathit{\Delta})=\min\{\eta:\ (\mathsf{q}_Y,\mathit{\Delta},\eta)\in\mathcal{C}\}=\min\{\eta:\ (\mathit{\Delta},\eta)\in\mathcal{C}_Y\}, \tag{3.11}$$

where $\mathcal{C}_Y\coloneqq\mathcal{C}\cap\{\mathsf{q}=\mathsf{q}_{\mathsf{Y}}\}$. This implies that the graph of $F_{\mathsf{T}}(\mathsf{q}_Y,\cdot)$ coincides with the lower boundary of the convex set $\mathcal{C}_Y$ and thus it is convex (more specifically, $F_{\mathsf{T}}(\mathsf{q}_Y,\mathit{\Delta})$ is jointly convex in $(\mathsf{q}_Y,\mathit{\Delta})$).

We note that $\mathcal{C}_Y$ is the intersection of $\mathcal{C}$ with a plane described by $\{\mathsf{q}=\mathsf{q}_Y\}$, which can be viewed as an intersection of $(N-1)$ hyperplanes. Therefore, Dubin's Theorem 3.6 can be invoked to show that the extreme points of $\mathcal{C}_Y$ can be written as convex combinations of at most N points of $\mathcal{S}$. Consequently, if $F_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ is strictly convex, all points of its graph are extreme points of $\mathcal{C}_Y$, and hence $F_{\mathsf{T}}(\mathsf{q}_Y, \varDelta)$ is achievable by Z with $|\mathcal{Z}| \le N$.

To evaluate $F_{\mathsf{T}}(\mathsf{q}_Y, \varDelta)$, Witsenhausen and Wyner suggested to study its conjugate function $F^*_{\mathsf{T}}(\mathsf{q}_Y, \cdot\,) : \mathbb{R} \to \mathbb{R}$, defined as

$$\begin{aligned} F^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda) &\coloneqq \min_{0 \le \varDelta \le H(Y)} F_{\mathsf{T}}(\mathsf{q}_Y, \varDelta) - \lambda\varDelta = \min\{\eta - \lambda\varDelta :\ (\varDelta, \eta) \in \mathcal{C}_Y\} \\ &= \min\{\eta - \lambda\varDelta :\ (\mathsf{q}_Y, \varDelta, \eta) \in \mathcal{C}\}. \end{aligned} \tag{3.12}$$

It is worth noting that $F^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda)$ determines a support line of slope $\lambda \in \mathbb{R}$ for $\mathcal{C}_Y$, or equivalently, the line $\lambda x + F^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda)$ is a support line of slope λ for the graph of $F_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$. This observation implies that $F_{\mathsf{T}}(\mathsf{q}_Y, \varDelta)$ can be recovered from F^*_{T} as

$$F_{\mathsf{T}}(\mathsf{q}_Y, \varDelta) = \max_{\lambda \in \mathbb{R}} \left[F^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda) + \lambda\varDelta\right]. \tag{3.13}$$

Since $F_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ is increasing, we can assume $\lambda \in \mathbb{R}^+$ in (3.13). On the other hand, the data processing inequality shows that $F_{\mathsf{T}}(\mathsf{q}_Y, \varDelta) \ge \varDelta + H(X) - H(Y)$, and thus, a line of slope 1 supports the graph of $F_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ at point $\varDelta = H(Y)$. This in turn implies that we can, without loss of generality, assume that $\lambda \le 1$ in (3.13).

As suggested by (3.13), in order to characterize $F_{\mathsf{T}}(\mathsf{q}_Y, \varDelta)$, it is sufficient to characterize $F^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda)$. To this end, suppose $\lambda \in [0, 1]$ is fixed and consider the mapping $\phi(\,\cdot\,, \lambda) : \mathcal{P}_{\mathcal{Y}} \to \mathbb{R}$, given by

$$\phi(\mathsf{q}, \lambda) = H((\mathsf{T}\mathsf{q})_{\mathcal{X}}) - \lambda H(\mathsf{q}). \tag{3.14}$$

Let $\mathcal{S}_\lambda$ be the graph of ϕ, i.e.,

$$\mathcal{S}_\lambda := \{(\mathsf{q}, \phi(\mathsf{q}, \lambda)) : \mathsf{q} \in \mathcal{P}_\mathcal{Y}\} = \{(\mathsf{q}, \eta - \lambda \varDelta) : (\mathsf{q}, \varDelta, \eta) \in \mathcal{S}\},$$

and let $\mathcal{C}_\lambda$ be its convex hull. Clearly, $\mathcal{C}_\lambda = \{(\mathsf{q}, \eta - \lambda\varDelta) : (\mathsf{q}, \varDelta, \eta) \in \mathcal{C}\}$. It follows from (3.12) that $F^*_{\mathsf{T}}(\cdot, \lambda)$ can be viewed as the lower boundary of $\mathcal{C}_\lambda$, and thus, as the lower convex envelope of $\phi(\cdot, \lambda)$.

Hence, if for some λ, the pair $(\mathsf{q}_Y, F^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda))$ can be written as a convex combination of the points $(\mathsf{q}_i, \phi(\mathsf{q}_i, \lambda)) \in \mathcal{S}_\lambda$, $i \in [k]$ for some $k \geq 2$ and weights $\omega_i \geq 0$ (i.e., $\sum_{i=1}^k \omega_i = 1$), then $\mathsf{q}_Y = \sum_{i=1}^k \omega_i \mathsf{q}_i$ and the random variable Z, defined by $P_Z(i) = \omega_i$ and $P_{Y|Z}(\cdot|i) = \mathsf{q}_i$, attains the minimum of $H(X|Z) - \lambda H(Y|Z)$. Therefore, the point $\left(\sum_{i=1}^k \omega_i H(\mathsf{q}_i), \sum_{i=1}^k \omega_i H((\mathsf{T}\mathsf{q}_i)_\mathcal{X})\right)$ lies on the lower boundary of $\mathcal{C}_Y$ which implies that $F_{\mathsf{T}}(\mathsf{q}_Y, x) = \sum_{i=1}^k \omega_i H((\mathsf{T}\mathsf{q}_i)_\mathcal{X})$ for $x = \sum_{i=1}^k \omega_i H(\mathsf{q}_i)$ and $\mathsf{q}_Y = \sum_{i=1}^k \omega_i \mathsf{q}_i$, and that the graph of the function $F_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ at this point has a support line of slope λ.

If, on the other hand, $F^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda)$ coincides with $\phi(\mathsf{q}_Y, \lambda)$ for some λ, then we have that: (i) the line $\lambda x + F^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda) = H(X) - \lambda(H(Y) - x)$ supports the graph of $F_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ at $x = H(Y)$ and (ii) the minimum $H(X|Z) - \lambda H(X|Z)$ is attained by a constant Z.

In summary, Witsenhausen and Wyner [147] concluded that *"all the information about the shape of F_{T} is contained in the restriction of F^*_{T} to its domain on which it differs from ϕ"*. Hence, the procedure to characterize $F_{\mathsf{T}}(\mathsf{q}, \varDelta)$ is as follows: (see [147, Thm 4.1])

- Fix $0 \leq \lambda \leq 1$ and compute the lower convex envelope of $\phi(\cdot, \lambda)$ (i.e., $F^*_{\mathsf{T}}(\cdot, \lambda)$),

- If a point of the graph of $F^*_{\mathsf{T}}(\cdot, \lambda)$ can be written as a convex combination of $\phi(\mathsf{q}_i, \lambda)$

with weights ω_i, $i \in [k]$ for some $k \geq 2$, then

$$F_{\mathsf{T}}\left(\sum_{i=1}^{k}\omega_i\mathsf{q}_i, \sum_{i=1}^{k}\omega_i H(\mathsf{q}_i)\right) = \sum_{i=1}^{k}\omega_i H((\mathsf{T}\mathsf{q}_i)_{\mathcal{X}}).$$

- If, for some λ, the function $F_{\mathsf{T}}^{*}(\mathsf{q}_Y, \lambda)$ coincides with $\phi(\mathsf{q}_Y, \lambda)$, then this corresponds to a line of slope λ supporting the graph of $F_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ at its endpoint $\Delta = H(\mathsf{q}_Y)$.

If $\mathsf{T} = \mathsf{BSC}(\alpha)$, then characterizing $F_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ is equivalent to the so-called Mrs. Gerber's Lemma [153]. Thus, this approach gives an easier proof for Mrs. Gerber's Lemma, see [147, IV.A], than the original one given in [153]. Witsenhausen and Wyner also examined $\mathsf{T} = \mathsf{BEC}(\delta)$ and also $\mathsf{T} = \mathsf{Z}(\beta)$ and obtained closed form expressions for $F_{\mathsf{T}}(\mathsf{q}_Y, \Delta)$ in these cases.

It is important to mention that a subtle crucial assumption in the above analysis is that the channels from Z to Y and from Y to X (i.e, T) are independent. However, constraining $I(X;Z) \leq \varepsilon$ makes $P_{Y|Z}$ depend on T and hence $g(\varepsilon)$ cannot be analyzed using a similar technique as above. However, if we instead look at $t(\mathsf{P}, R)$, the dual representation of $g(\varepsilon)$ given in (3.4), then we can use the above argument to obtain a geometric interpretation of $g(\varepsilon)$. First, note that $t(\mathsf{P}, \cdot)$ is strictly increasing on $(0, H(Y))$ and convex[4] and also $t(\mathsf{P}, R) = \varepsilon$ if and only if $g(\varepsilon) = R$. Consequently, $t(\mathsf{P}, \cdot)$ is strictly convex if and only if $\varepsilon \mapsto g(\varepsilon)$ is strictly concave.

We clearly have

$$\begin{aligned} t(\mathsf{P}, R) &= H(X) - \max_{\substack{P_{Z|Y}: X -\!\circ\!- Y -\!\circ\!- Z, \\ H(Y|Z) \leq H(Y) - R}} H(X|Z) \\ &= H(X) - \max\{\eta : (\mathsf{q}_{\mathsf{Y}}, \Delta, \eta) \in \mathcal{C}\} \end{aligned}$$

[4]This can easily be shown using a similar argument as in the proof of Lemma 3.2.

$$=:\quad H(X) - G_{\mathsf{T}}(\mathsf{q}_Y, \mathit{\Delta}), \tag{3.15}$$

where $\mathit{\Delta} := H(Y) - R$. Therefore, the graph of $G_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ is the upper boundary of $\mathcal{C}_Y$ (and thus it is concave which provides an alternative proof for the concavity of $\varepsilon \mapsto g(\varepsilon)$). This analogy between $F_{\mathsf{T}}(\mathsf{q}_Y, \mathit{\Delta})$ and $G_{\mathsf{T}}(\mathsf{q}_Y, \mathit{\Delta})$ allows us to invoke Dubin's theorem, the same way as Witsenhausen and Wyner did in [147], to conclude that if $G_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ is strictly concave, or equivalently, if g is strictly concave, then g is achieved by Z with $|\mathcal{Z}| \leq N$. It is important, however, to mention that g is not in general strictly concave, see for example Lemma 3.36. We obtain a sufficient condition that g is not strictly concave for a large family of channels T in the next section.

Recall that the graphs of $F_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ and $G_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ are the lower and upper boundaries of the compact and convex set $\mathcal{C}_Y$, respectively. Hence similar to [147], we evaluate G_{T} using its conjugate function G^*_{T}. We define $G^*_{\mathsf{T}}(\mathsf{q}_Y, \cdot\,) : \mathbb{R} \to \mathbb{R}$ as

$$G^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda) := \max_{0 \leq \mathit{\Delta} \leq H(Y)} \left[G_{\mathsf{T}}(\mathsf{q}_Y, \mathit{\Delta}) - \lambda \mathit{\Delta}\right] = \max\{\eta - \lambda\mathit{\Delta} :\ (\mathit{\Delta}, \eta) \in \mathcal{C}_Y\}. \tag{3.16}$$

Note that $G^*_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ determines a line of slope $\lambda \in \mathbb{R}$ supporting $\mathcal{C}_Y$ from above, or equivalently, the line $\lambda x + G^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda)$ is a support line of slope λ for the graph of $G_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$. This observation implies that $G_{\mathsf{T}}(\mathsf{q}_Y, \mathit{\Delta})$ can be recovered from G^* as

$$G_{\mathsf{T}}(\mathsf{q}_Y, \mathit{\Delta}) = \min_{\lambda \in \mathbb{R}} \left[G^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda) + \lambda \mathit{\Delta}\right]. \tag{3.17}$$

Since $G_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ is increasing, we can assume $\lambda \in \mathbb{R}^+$ in (3.17). It can be shown that, for a fixed λ, the graph of $G^*_{\mathsf{T}}(\cdot, \lambda)$ constitutes the upper boundary of $\mathcal{C}_\lambda$. Thus, the graph of $G^*_{\mathsf{T}}(\cdot, \lambda)$ coincides with the upper concave envelope of $\phi(\cdot, \lambda)$, defined in (3.14). Hence, we can have the following steps to evaluate[5] $G_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$:

- Fix $\lambda \geq 0$ and compute the upper concave envelope of $\phi(\cdot, \lambda)$(i.e., $G^*_{\mathsf{T}}(\cdot, \lambda)$),
- If a point of the graph of $G^*_{\mathsf{T}}(\cdot, \lambda)$ can be written as a convex combination of $\phi(\mathsf{q}_i, \lambda)$ with weights ω_i, $i \in [k]$ for some $k \geq 2$, then

$$G_{\mathsf{T}}\left(\sum_{i=1}^{k} \omega_i \mathsf{q}_i, \sum_{i=1}^{k} \omega_i H(\mathsf{q}_i)\right) = \sum_{i=1}^{k} \omega_i H((\mathsf{T}\mathsf{q}_i)_{\mathcal{X}}). \tag{3.18}$$

- If, for some q and λ, the function $G^*_{\mathsf{T}}(\mathsf{q}, \lambda)$ coincides with $\phi(\mathsf{q}, \lambda)$, then this corresponds to a support line of slope λ at the point $\Delta = H(\mathsf{q})$.

This observation allows us to derive a closed form expression for $G_{\mathsf{T}}(\mathsf{q}_Y, \cdot)$ when $\mathsf{T} = \mathsf{BEC}(\delta)$ and $\mathsf{q}_Y = \mathsf{Bernoulli}(q)$ with $0 \leq q \leq \frac{1}{2}$.

Theorem 3.7. *Let* $\mathsf{T} = \mathsf{BEC}(\delta)$ *and* $\mathsf{q}_Y = \mathsf{Bernoulli}(q)$ *with* $0 \leq q \leq \frac{1}{2}$. *Then* $G_{\mathsf{T}}(\mathsf{q}_Y, \Delta) = h_{\mathsf{b}}(\delta) + \bar{\delta}\Delta$ *for* $0 \leq \Delta \leq h_{\mathsf{b}}(q)$ *and*

$$g(\varepsilon) = \frac{\varepsilon}{\bar{\delta}},$$

for any $\varepsilon \leq \bar{\delta} h_{\mathsf{b}}(q)$.

Proof. Fix $\mathsf{q} = \mathsf{Bernoulli}(r)$ and $\lambda \geq 0$. In this case, $\phi(\mathsf{q}, \lambda)$ and $G^*_{\mathsf{T}}(\mathsf{q}, \lambda)$ are functions of r, thus we denote them by $\phi(r, \lambda)$ and $G^*_{\mathsf{T}}(r, \lambda)$, respectively. We have $\phi(r, \lambda) = h_{\mathsf{b}}(\delta) + (\bar{\delta} - \lambda) h_{\mathsf{b}}(r)$. Thus, $\phi(\cdot, \lambda)$ is concave for $\lambda \leq \bar{\delta}$ and convex for $\lambda \geq \bar{\delta}$. Letting $G^*_{\mathsf{T}}(q, \lambda)$ denote $G^*_{\mathsf{T}}(\mathsf{q}_Y, \lambda)$, we can write $G^*_{\mathsf{T}}(q, \lambda) = \phi(q, \lambda)$ for $\lambda \leq \bar{\delta}$, and $G^*_{\mathsf{T}}(q, \lambda) = h_{\mathsf{b}}(\delta)$ for

[5] The duality between F_{T} and G_{T} was first observed by F. P. Calmon (flavio@seas.harvard.edu) and led to [18].

$\lambda \geq \bar{\delta}$. In light of (3.17), we conclude that

$$G_{\mathsf{T}}(q, \varDelta) \coloneqq G_{\mathsf{T}}(\mathsf{q}_Y, \varDelta) = \min\{G^*_{\mathsf{T}}(q, \lambda) + \lambda\varDelta : \ \lambda \geq 0\}.$$

Focusing on the domain of G^*_{T} on which it differs from ϕ, we have that $\min_{\lambda\geq\bar{\delta}}[h_{\mathsf{b}}(\delta) + \lambda\varDelta] = h_{\mathsf{b}}(\delta) + \bar{\delta}\varDelta$. Hence, $G_{\mathsf{T}}(q, \varDelta) = h_{\mathsf{b}}(\delta) + \bar{\delta}\varDelta$ for $0 \leq \varDelta \leq h_{\mathsf{b}}(q)$. This then implies that

$$t(\mathsf{P}, R) = H(X) - G_{\mathsf{T}}(q, h_{\mathsf{b}}(q) - R) = \bar{\delta}R,$$

and consequently $g(\varepsilon) = \frac{\varepsilon}{\bar{\delta}}$ for $\varepsilon \leq I(X; Y) = \bar{\delta}h_{\mathsf{b}}(q)$. □

The next theorem provides the values of $G_{\mathsf{T}}(\mathsf{q}_Y, \varDelta)$ for $0 \leq \varDelta \leq H(Y)$ when $\mathsf{T} = \mathsf{BSC}(\alpha)$ and $\mathsf{q}_Y = \mathsf{Bernoulli}(q)$ with $0 \leq \alpha, q \leq \frac{1}{2}$.

Theorem 3.8. [6] *Let* $\mathsf{T} = \mathsf{BSC}(\alpha)$ *with* $0 \leq \alpha \leq \frac{1}{2}$ *and* $\mathsf{q}_Y = \mathsf{Bernoulli}(q)$ *with* $0 \leq q \leq \frac{1}{2}$. *Let also* $\mathcal{G} \coloneqq \{(\varDelta, G_{\mathsf{T}}(\mathsf{q}_Y, \varDelta)) : 0 \leq \varDelta \leq H(Y)\}$. *Then we have,*

$$\mathcal{G} = \left\{\left(\omega h_{\mathsf{b}}\left(\frac{q}{z}\right), \omega h_{\mathsf{b}}\left(\frac{q}{z} * \alpha\right) + \bar{\omega}h_{\mathsf{b}}(\alpha)\right) : \ 0 \leq \omega \leq 1, z = \max\{\omega, 2q\}\right\}.$$

Proof. As before, fix $\mathsf{q} = \mathsf{Bernoulli}(r)$ and $\lambda \geq 0$. In this case, $\phi(\mathsf{q}, \lambda)$ and $G^*_{\mathsf{T}}(\mathsf{q}, \lambda)$ are functions of r, thus we denote them by $\phi(r, \lambda)$ and $G^*_{\mathsf{T}}(r, \lambda)$, respectively. We have $\phi(r, \lambda) = h_{\mathsf{b}}(\alpha * r) - \lambda h_{\mathsf{b}}(r)$. It can be verified that $\phi(\cdot, \lambda)$ is convex for $\lambda \geq (1 - 2\alpha)^2$ and $G^*_{\mathsf{T}}(r, \lambda) = h_{\mathsf{b}}(\alpha)$. For $0 \leq \lambda < (1 - 2\alpha)^2$, the map $\phi(\cdot, \lambda)$ is concave on an interval symmetric about $r = \frac{1}{2}$ and convex elsewhere. Moreover, $\phi(r, \lambda) \leq \phi(\frac{1}{2}, \lambda)$ on the region of concavity. Thus, if $\phi(\frac{1}{2}, \lambda) < h_{\mathsf{b}}(\alpha)$, then $G^*_{\mathsf{T}}(r, \lambda)$ is a convex combination

[6] This theorem was proved in collaboration with F. P. Calmon (flavio@seas.harvard.edu) [18].

of $(0,\phi(0,\lambda))$ and $(1,\phi(1,\lambda))$ and then again $G^*_{\mathsf{T}}(r,\lambda) = h_{\mathsf{b}}(\alpha)$. If $\phi(\frac{1}{2},\lambda) > h_{\mathsf{b}}(\alpha)$, then by definition of upper concave envelope there must exists $r_\lambda \in [r,\frac{1}{2})$ when $r < \frac{1}{2}$ (resp. $\bar{r}_\lambda \in [r,1)$ when $r > \frac{1}{2}$) such that $(r, G^*_{\mathsf{T}}(r,\lambda)) \in \mathcal{C}_\lambda$ can be written as a convex combination of $(0,\phi(0,\lambda))$ and $(r_\lambda, \phi(r_\lambda,\lambda))$ (resp. $(\bar{r}_\lambda, \phi(\bar{r}_\lambda,\lambda))$ and $(0,\phi(0,\lambda))$). Since by assumption $q \le \frac{1}{2}$, we obtain from (3.18) that

$$G_{\mathsf{T}}(q, \bar{c}h_{\mathsf{b}}(0) + ch_{\mathsf{b}}(r_\lambda)) = \bar{c}h_{\mathsf{b}}(\alpha * 0) + ch_{\mathsf{b}}(\alpha * r_\lambda), \tag{3.19}$$

where $0 \le c \le 1$ and $\bar{c}0 + cr_\lambda = q$. Consequently, we can write $G_{\mathsf{T}}(q, \omega h_{\mathsf{b}}(\frac{q}{\omega})) = \bar{\omega}h_{\mathsf{b}}(\alpha) + \omega h_{\mathsf{b}}(\alpha * \frac{q}{\omega})$, where $\omega := \frac{q}{r_\lambda}$ for $q \le r_\lambda \le \frac{1}{2}$. Finally, if $\phi(\frac{1}{2},\lambda) = h_{\mathsf{b}}(\alpha)$, then $(r, G^*_{\mathsf{T}}(r,\lambda)) \in \mathcal{C}_\lambda$ for any r can be written as a convex combination of points $(0,\phi(0,\lambda))$, $(\frac{1}{2},\phi(\frac{1}{2},\lambda))$, and $(1,\phi(1,\lambda))$. Thus, from (3.18) we obtain

$$G_{\mathsf{T}}(q, c_1h_{\mathsf{b}}(0) + c_2h_{\mathsf{b}}\left(\frac{1}{2}\right) + c_3h_{\mathsf{b}}(1)) = c_1h_{\mathsf{b}}(\alpha*0) + c_2h_{\mathsf{b}}\left(\alpha*\frac{1}{2}\right) + c_3h_{\mathsf{b}}(\alpha*1), \tag{3.20}$$

where $0 \le c_i \le 1$, $1 \le i \le 3$, $\sum_{i=1}^{3} c_i = 1$, and $c_1 0 + c_2\frac{1}{2} + c_3 1 = q$. Consequently, we obtain $G_{\mathsf{T}}(q,\omega) = \bar{\omega}h_{\mathsf{b}}(\alpha) + \omega$ where $\omega = c_2 \le 2q$. Combining (3.19) and (3.20), the result follows. $\square$

Given $\mathcal{G}$ in this theorem, one can characterize the set $\{(R, t(\mathsf{P},R)) : 0 \le R \le H(Y)\}$ by $\{(h_{\mathsf{b}}(q) - \mathit{\Delta}, h_{\mathsf{b}}(q*\alpha) - \eta) :\ (\mathit{\Delta},\eta) \in \mathcal{G}\}$. If $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$ and $\mathsf{T} = \mathsf{BSC}(\alpha)$, then Theorem 3.8 implies that $G_{\mathsf{T}}(\mathsf{q}_Y, \mathit{\Delta}) = h_{\mathsf{b}}(\alpha) + \mathit{\Delta}(1 - h_{\mathsf{b}}(\alpha))$ and hence $g(\varepsilon) = \frac{\varepsilon}{1-h_{\mathsf{b}}(\alpha)}$ for any $\varepsilon \le 1 - h_{\mathsf{b}}(\alpha)$. We will show in Section 3.8.2 that $g(\varepsilon) = \frac{\varepsilon}{I(X;Y)}$ when $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$ and T is any binary input symmetric output channel, which includes $\mathsf{BEC}(\delta)$ and $\mathsf{BSC}(\alpha)$, and thus generalize Theorems 3.7 and 3.8 for the uniform case.

We will generalize this technique to study similar problems for maximal correlation and Arimoto's conditional entropy in Chapters 5 and 6, respectively.

3.6 Non-Trivial Filters For Perfect Privacy

As it becomes clear later, requiring that $g(0) = 0$ is a useful assumption for the analysis of $g(\varepsilon)$. Thus, it is crucial to find a necessary and sufficient condition on the joint distribution P under which $g(0) = 0$. When this holds, we say that perfect privacy implies trivial utility.

This problem is studied in two different cases: (i) when Y is a scalar random variable and (ii) when n i.i.d. copies $Y_1, \ldots, Y_n$ are available and we require $\varepsilon \to 0$ as $n \to \infty$ (i.e., asymptotic perfect privacy).

3.6.1 Scalar Case

Assuming $(X, Y) \sim \mathsf{P}$ is given, we obtain a necessary and sufficient condition on P under which $g(0) > 0$. To do this, we need the following definition.

Definition 3.9 ([24]). *The random variable X is said to be* weakly independent *of Y if the rows of the transition matrix $P_{X|Y}$, i.e., the set of vectors $\{P_{X|Y}(\cdot|y),\ y \in \mathcal{Y}\}$, are linearly dependent.*

Theorem 3.10. *We have $g(0) > 0$ if and only if X is weakly independent of Y.*

Proof. $\Rightarrow$ direction:

The fact that $g(0) > 0$ implies that there exists a random variable Z over an alphabet $\mathcal{Z}$ such that the Markov condition $X \multimap Y \multimap Z$ is satisfied and $Z \perp\!\!\!\perp X$ while $I(Y; Z) > 0$. Hence, for any z_1 and z_2 in $\mathcal{Z}$, we must have $P_{X|Z}(x|z_1) = P_{X|Z}(x|z_2)$ for all $x \in \mathcal{X}$, which implies that

$$\sum_{y\in\mathcal{Y}} P_{X|Y}(x|y)P_{Y|Z}(y|z_1) = \sum_{y\in\mathcal{Y}} P_{X|Y}(x|y)P_{Y|Z}(y|z_2)$$

and hence

$$\sum_{y\in\mathcal{Y}} P_{X|Y}(x|y)\left[P_{Y|Z}(y|z_1) - P_{Y|Z}(y|z_2)\right] = 0.$$

Since Y is not independent of Z, there exist z_1 and z_2 such that $P_{Y|Z}(y|z_1) \neq P_{Y|Z}(y|z_2)$ and hence the above shows that the set of vectors $P_{X|Y}(\cdot|y)$, $y \in \mathcal{Y}$ is linearly dependent.

$\Leftarrow$ direction:

Berger and Yeung [24, Appendix II], in a completely different context, showed that if X is weakly independent of Y, one can always construct a binary random variable Z correlated with Y which satisfies $X \multimap Y \multimap Z$ and $X \perp\!\!\!\perp Z$, and thus $g(0) > 0$. $\square$

Remark 3.11. Theorem 3.10 first appeared in [13]. However, Calmon et al. [31], in the study of the Privacy Funnel $t(\mathsf{P}, R)$, showed an equivalent necessary and sufficient condition for the non-trivial utility in case of perfect privacy. In fact, they showed that for a given P, one can always generate Z such that $I(X;Z) = 0$, $I(Y;Z) > 0$ and $X \multimap Y \multimap Z$, or equivalently $g(0) > 0$, if and only if the smallest singular value of the conditional expectation operator $f \mapsto \mathbb{E}[f(X)|Y]$ is zero. This condition can, in fact, be shown to be equivalent to X being weakly independent of Y.

Remark 3.12. Recalling that $\mathcal{X} = [M]$ and $\mathcal{Y} = [N]$, it is clear from Definition 3.9 that X is weakly independent of Y if $N > M$. Hence, Theorem 3.10 implies that $g(0) > 0$ if Y has strictly larger alphabet than X.

In light of the above remark, in the most common case $N = M$, one might have $g(0) = 0$, which corresponds to the most conservative scenario as no privacy leakage implies no broadcasting of observable data. In such cases, the rate of increase of $g(\varepsilon)$ at $\varepsilon = 0$, that is $g'(0) := \frac{\mathrm{d}}{\mathrm{d}\varepsilon} g(\varepsilon)|_{\varepsilon=0}$, which corresponds to the initial efficiency of privacy-constrained information extraction, proves to be very important in characterizing the behavior of $g(\varepsilon)$ for all $\varepsilon \geq 0$. This is because, for example, by concavity of $\varepsilon \mapsto g(\varepsilon)$, the slope of $g(\varepsilon)$ is maximized at $\varepsilon = 0$ and so

$$g'(0) = \lim_{\varepsilon \to 0} \frac{g(\varepsilon)}{\varepsilon} = \sup_{\varepsilon > 0} \frac{g(\varepsilon)}{\varepsilon},$$

and hence $g(\varepsilon) \leq \varepsilon g'(0)$ for all $\varepsilon \leq I(X;Y)$. Also the lower bound in (3.5) implies that $g'(0) \geq \frac{H(Y)}{I(X;Y)}$, for any pair of discrete random variables (X, Y).

It is easy to show that X is weakly independent of binary Y if and only if X and Y are independent, thus the following corollary immediately follows.

Corollary 3.13. *Let Y be a non-constant binary random variable correlated with X. Then $g(0) = 0$.*

The following examples show that if $P_{Y|X}$ is an erasure channel (even binary erasure channel) then $g(0) > 0$.

Example 3.14. Suppose $X \sim \mathsf{Bernoulli}(p)$ for $0 \leq p \leq \frac{1}{2}$ and $P_{Y|X} = \mathsf{BEC}(\delta)$, i.e., $P_{Y|X}(x|x) = 1 - \delta$ and $P_{Y|X}(\mathsf{e}|x) = \delta$ for $x \in \{0, 1\}$, where e denotes the erasure. Let $Z = 1$ when $Y \in \{0, 1\}$ and $Z = 0$ when $Y = \mathsf{e}$. It is clear to see that $Z \sim \mathsf{Bernoulli}(\delta)$ and hence $I(Y;Z) = h_{\mathsf{b}}(\delta)$. On the other hand, $P_{Z|X}(z|0) = P_{Z|X}(z|1)$ and hence $Z \perp\!\!\!\perp X$. Note that $H(Y|X) = h_{\mathsf{b}}(\delta)$, thus according to the upper bound in (3.5), $g(0) = h_{\mathsf{b}}(\delta)$.

Example 3.15. A discrete memoryless channel W with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, is called *generalized erasure* if $\mathcal{Y}$ can be decomposed as $\mathcal{Y}_0 \cup \mathcal{Y}_1$ such that $\mathsf{W}(y|x)$ does not depend on x whenever $y \in \mathcal{Y}_0$. Suppose $|\mathcal{Y}_0| = k$ and $p_j^0 := P_Y(j)$,

$j \in \mathcal{Y}_0$. Using the similar argument as Example 3.14, it is straightforward to show that $g(0) \geq H\left(p_1^0, \ldots, p_k^0, 1-\sum_{j=1}^k p_j^0\right)$.

3.6.2 Vector Case

Next, we study the same problem in the vector case, i.e., what is a condition on the joint distribution for which $g(0) > 0$, when n i.i.d. copies of Y are available? More precisely, let $P_{Y|X}(\cdot|x)$ be the distribution over $\mathcal{Y}$ induced by $x \in \mathcal{X}$ and $Y_1, \ldots, Y_n$ be n i.i.d. samples drawn from the parametric distribution $P_{Y|X}(\cdot|X)$, where the parameter X has prior p_X. Let the simplified version $\tilde{g}_n(\varepsilon)$ of the rate-privacy function be defined as

$$\tilde{g}_n(\varepsilon) := \sup_{f\colon\, I(f(Y^n);X)\leq\varepsilon} H(f(Y^n)), \tag{3.21}$$

where the maximization is taken over deterministic function $f : \mathcal{Y}^n \to \mathcal{Z}$ such that the privacy constraint $I(f(Y^n); X) \leq \varepsilon$ is satisfied. The next theorem gives an asymptotic lower bound for the normalized $\frac{1}{n}\tilde{g}_n(\varepsilon)$ in the limit when $\varepsilon \to 0$.

Theorem 3.16. *For any pair of discrete random variables $(X, Y) \sim \mathsf{P}$, we have*

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\tilde{g}_n(\varepsilon) \geq H_\infty^*(Y|X),$$

where $H_\infty^*(Y|X) := \min_{x\in\mathcal{X}}\min_{y\in\mathcal{Y}}\left(-\log P_{Y|X}(y|x)\right)$.

For the proof, we need the following lemma which relates the difference of the entropies of two distributions P and Q supported over a set $\mathcal{U}$ with their total variational distance $\mathsf{TV}(P, Q) := \sum_{u\in\mathcal{U}} |P(u) - Q(u)|$.

Lemma 3.17 ([40]). *If P and Q are two distributions with total variational distance $\mathsf{TV}(P,Q)$, then*

$$|H(P) - H(Q)| \leq \mathsf{TV}(P,Q)\log(|\mathcal{X}|-1) + h_{\mathsf{b}}(\mathsf{TV}(P,Q)).$$

The lemma implies that there exists a function $\delta : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$ and $|H(P) - H(Q)| \leq \delta(\mathsf{TV}(P,Q))$. Note that assuming $P = \mathsf{P}$ and $Q = \mathsf{p}_X\mathsf{q}_Y$, then this lemma exhibits an upper bound for $I(X;Y)$ in terms of $\mathsf{TV}(P,Q)$. Now we are in position to give the proof of the theorem.

Proof of Theorem 3.16. Recall that $|\mathcal{X}| = M$. Let $P_j^n(y^n) := P_{Y^n|X}(y^n|x_j) = \prod_{k=1}^n P_{Y|X}(y_k|x_j)$ be the distribution over $\mathcal{Y}^n$ that each x_j, $j \in [M]$ induces. Given these M distributions, we construct nearly equiprobable bins $K_j^n(i) \subset \mathcal{Y}^n$ for $i \in [2^r]$, (with r to be determined later), such that $P_j^n(K_j^n(i)) := \sum_{y^n \in K_j^n(i)} P_j^n(y^n)$ is close to 2^{-r} for each $j \in [M]$ and $i \in [2^r]$. Let U^r denote the uniform distribution over $\{0,1\}^r$.

Recalling the definition of $H_\infty^*(Y|X)$, we can write

$$P_j^n(y^n) \leq 2^{-nH_\infty^*(Y|X)}, \qquad j \in [M]. \tag{3.22}$$

We start the construction of the bins $K_j^n(1), K_j^n(2), \ldots, K_j^n(J_j)$, for each $j \in [M]$, where $J_j \leq 2^r - 1$ is the number of bins for each j. The first bin is constructed as follows. We agglomerate the minimal number of mass points of P_j^n into $K_j^n(1)$ as needed to make sure

$$P_j^n(K_j^n(1)) \geq 2^{-r} - 2^{-s}, \tag{3.23}$$

for some $s < nH^*_\infty(Y|X)$. This together with (3.22) shows that

$$P^n_j(K^n_j(1)) < 2^{-r} - 2^{-s} + 2^{-nH^*_\infty(Y|X)}, \tag{3.24}$$

which can be simplified as

$$P^n_j(K^n_j(1)) < 2^{-r}, \tag{3.25}$$

because $s < nH^*_\infty(Y|X)$.

Once condition (3.23) is met, the construction for the first bin is completed and we move on to the second bin. This procedure can go on until either we run out of mass points or the restriction $J_j \le 2^r - 1$ is violated. In the latter case, we set $J_j = 2^r - 1$ and then collect all mass points left into the bin $K^n_j(J_j + 1)$. The former happens if the total probability of the left-over is strictly less than $2^{-r} - 2^{-s}$ so that we cannot meet the requirement (3.23) which yields

$$P^n_j\left(\bigcup_{i=1}^{J_j} K^n_j(i)\right) > 1 - 2^{-r} + 2^{-s}. \tag{3.26}$$

On the other hand, we know from (3.25) that $P^n_j\left(\bigcup_{i=1}^{J_j} K^n_j(i)\right) < J_j 2^{-r}$ which, together with (3.26), implies

$$1 - 2^{-r} + 2^{-s} < P^n_j\left(\bigcup_{i=1}^{J_j} K^n_j(i)\right) < J_j 2^{-r}, \tag{3.27}$$

leading to a lower bound for the number of bins in this case

$$J_j > 2^r + 2^{r-s} - 1, \tag{3.28}$$

which is greater than the allowable upper-bound $2^r - 1$. We hence conclude that with s that satisfies $s < nH_\infty^*$, the procedure stops only when the restriction $J_j \leq 2^r - 1$ is violated, and therefore, we assume $J_j = 2^r - 1$ in what follows.

As specified earlier, we construct the last bin $K_j^n(J_j + 1)$ by including all the leftover mass there. We therefore have

$$K_j^n(J_j + 1) = \text{supp}(P_j^n) - \bigcup_{i=1}^{J_j} K_j^n(i), \tag{3.29}$$

where $\text{supp}(P_j^n)$ denotes the support of P_j^n. Since each bin has probability lower-bounded by (3.23), it follows from (3.29) that

$$P_j^n(K_j^n(J_j + 1)) = 1 - \sum_{i=1}^{J_j} P_j^n(K_j^n(i)) \leq 1 - J_j \left(2^{-r} - 2^{-s}\right), \tag{3.30}$$

which, after substituting $J_j = 2^r - 1$, is simplified as

$$P_j^n(K_j^n(J_j + 1)) \leq 2^{r-s} + 2^{-r} - 2^{-s}. \tag{3.31}$$

We have thus far constructed $M \times 2^r$ bins, namely 2^r bins for each P_j^n, $j \in [M]$. Consider now the deterministic mapping $g_n : \mathcal{Y}^n \times \mathcal{X} \to [2^r]$ defined as follows:

$$g_n(y^n, x_j) = i \qquad \text{if} \qquad y^n \in K_j^n(i).$$

This mapping requires x_j because for each $j \in [M]$ the corresponding bins are disjoint. However, we know that by using a proper channel encoding and decoding, ϕ_n and ψ_n, respectively, one can decode Y^n to obtain $\psi_n(Y^n)$ such that $\Pr(X \neq \psi_n(Y^n))$ decays exponentially. So, we can have a deterministic function which acts only on Y^n from which

x_j is obtained with probability exponentially close to one. Hence our sequence of deterministic mappings is:

$$f_n(y^n) \coloneqq g_n(y^n, \psi_n(y^n)) = i \qquad \text{if} \qquad y^n \in K_j^n(i).$$

where j is the index of the decoded symbol, that is the j such that $\psi_n(y^n) = x_j$.

Now let us look at the total variation distance between $\tilde{P}_j^n \coloneqq f_n \circ P_j^n$ and U^r.

$$\begin{aligned}
\mathsf{TV}\left(\tilde{P}_j^n, U^r\right) &= \sum_{i=1}^{2^r} |2^{-r} - P_j^n(K_j^n(i))| \\
&\overset{(a)}{=} \sum_{i=1}^{J_j} \left(2^{-r} - P_j^n(K_j^n(i))\right) + |2^{-r} - P_j^n(K_j^n(J_j+1))| \\
&\overset{(b)}{\leq} \sum_{i=1}^{J_j} 2^{-s} + (2^{-r} + P_j^n(K_n(J_j+1))) \\
&\overset{(c)}{\leq} J_j 2^{-s} + 2^{-r} + 2^{r-s} + 2^{-r} - 2^{-s} \\
&= 2\left(2^{r-s} + 2^{-r} - 2^{-s}\right) < 2\left(2^{r-s} + 2^{-r}\right).
\end{aligned}$$

where (a) follows from (3.25), (b) is due to the triangle inequality and (3.23) and (c) follows from (3.31). Setting $r = nH_\infty^*(Y|X) - n\delta$ and $s = nH_\infty^*(Y|X) - n\frac{\delta}{2}$ for some $0 < \delta \leq \frac{2}{3}H_\infty^*(Y|X)$, we conclude that

$$\mathsf{TV}\left(\tilde{P}_j^n, U^r\right) \leq 2(2^{r-s} + 2^{-r}) \leq 2^{-\frac{\delta}{2}n+2},$$

and hence for sufficiently large n, there exists $\delta > 0$ such that $\mathsf{TV}\left(\tilde{P}_j^n, U^r\right) \leq \frac{\varepsilon}{2}$.

Now consider $\tilde{P}_j^n$ and $\tilde{P}_k^n$ for $j \neq k$. We can write

$$\mathsf{TV}(\tilde{P}_j^n, \tilde{P}_k^n) \leq \mathsf{TV}(\tilde{P}_j^n, U^r) + \mathsf{TV}(\tilde{P}_k^n, U^r) \leq \varepsilon.$$

Let $Z_n \coloneqq f_n(Y^n)$. Then by Jensen's inequality, we can write

$$\mathsf{TV}(P_{Z_n X}, P_{Z_n} P_X) \leq \sum_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} P_X(x) P_X(x') \mathsf{TV}\left(\tilde{P}_j^n, \tilde{P}_k^n\right) \leq \varepsilon. \tag{3.32}$$

Invoking Lemma 3.17, we conclude from (3.32) that $I(X; Z_n) \leq \varepsilon$ for sufficient large n. Notice that Z_n is a random variable which is almost uniformly distributed over a set of cardinality $2^r = 2^{nH_\infty^*(Y|X) - n\delta}$ and hence $\frac{1}{n} H(Z_n) = H_\infty^*(Y|X) - \delta$. □

This theorem implies that, even if Y is binary, one can have information transfer at a positive rate while allowing perfect privacy only in the limit instead of requiring absolutely zero privacy leakage. In fact, Theorem 3.16 implies that for any joint distribution P which satisfies $P_{Y|X}(y|x) < 1$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have $\tilde{g}_n(\varepsilon) > 0$ and consequently $g(P_{XY^n}, \varepsilon) > 0$. This result seems similar in essence to the main result of [31] which states that for n i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$ from $(X, Y) \sim \mathsf{P}$, we have $g(P_{X^n Y^n}, \varepsilon) > 0$ for n sufficiently large, unless X is a deterministic function of Y.

3.7 Operational Interpretation of Rate-Privacy Function

In this section, we propose a coding-theoretic setting, the so-called *dependence dilution* model, and show that the dual of the rate-privacy function is a boundary point of its achievable rate region, thereby giving an information-theoretic operational interpretation for the rate-privacy function. It must be noted that another operational interpretation of $g(\varepsilon)$ was recently shown in [96].

Inspired by the problems of information amplification [88] and state masking [110], Courtade [36] proposed the *information-masking tradeoff* problem as follows. The tuple $(R_u, R_v, \Delta_A, \Delta_M) \in \mathbb{R}^4$ is said to be achievable if for two given separated sources $U \in \mathcal{U}$ and $V \in \mathcal{V}$ and any $\varepsilon > 0$ there exist mappings $f : \mathcal{U}^n \to [2^{nR_u}]$ and $g : \mathcal{V}^n \to [2^{nR_v}]$ such that $I(U^n; f(U^n), g(V^n)) \leq n(\Delta_M + \varepsilon)$ and $I(V^n; f(U^n), g(V^n)) \geq n(\Delta_A - \varepsilon)$. That is, $(R_u, R_v, \Delta_A, \Delta_M)$ is achievable if there exist indices K and J of rates R_u and R_v given U^n and V^n, respectively, such that the receiver in possession of (K, J) can recover at most $n\Delta_M$ bits about U^n and at least $n\Delta_A$ about V^n. The closure of the set of all achievable tuple $(R_u, R_v, \Delta_A, \Delta_M)$ is characterized in [36]. Here, we look at a similar problem but for a joint encoder. In fact, we want to examine the achievable rate of an encoder observing both X^n and Y^n which masks X^n and amplifies Y^n at the same time, by rates Δ_M and Δ_A, respectively.

We define a $(2^{nR}, n)$ *dependence dilution* code by an encoder

$$f_n : \mathcal{X}^n \times \mathcal{Y}^n \to [2^{nR}],$$

and a list decoder

$$g_n : [2^{nR}] \to 2^{\mathcal{Y}^n},$$

having a fixed list size

$$|g_n(J)| = 2^{n(H(Y) - \Delta_A)}, \qquad \forall J \in [2^{nR}], \tag{3.33}$$

where $J := f_n(X^n, Y^n)$ is the encoder's output and $2^{\mathcal{Y}^n}$ denotes the power set of $\mathcal{Y}^n$. Let the error probability be defined as $p_e^{(n)} := \Pr\left(Y^n \notin g_n(J)\right)$. A *dependence dilution* *triple* $(R, \Delta_A, \Delta_M) \in \mathbb{R}_+^3$ is said to be achievable if, for any $\delta > 0$, there exists a $(2^{nR}, n)$ dependence dilution code that satisfies the utility constraint:

$$p_e^{(n)} \to 0, \tag{3.34}$$

as $n \to \infty$ and the privacy constraint:

$$\frac{1}{n} I(X^n; J) \leq \Delta_M + \delta. \tag{3.35}$$

Intuitively speaking, upon receiving J, the decoder is required to construct list $g_n(J) \subset \mathcal{Y}^n$ of fixed size which contains likely candidates of the actual sequence Y^n. Without any observation, the decoder can only construct a list of size $2^{nH(Y)}$ which contains Y^n with probability close to one. However, after J is observed and the list $g_n(J)$ is formed, the decoder's list size can be reduced to $2^{n(H(Y)-\Delta_A)}$ and thus reducing the uncertainty about Y^n by $0 \leq n\Delta_A \leq nH(Y)$. This observation led Kim et al. [88] to show that the utility constraint (3.34) is equivalent to the amplification requirement

$$\frac{1}{n} I(Y^n; J) \geq \Delta_A - \delta, \tag{3.36}$$

which lower bounds the amount of information that J carries about Y^n. The following lemma gives an outer bound for the achievable dependence dilution region.

Theorem 3.18. *Any achievable dependence dilution triple* (R, Δ_A, Δ_M) *satisfies*

$$\begin{cases} R \geq \Delta_A \\ \Delta_A \leq I(Y;U) \\ \Delta_M \geq I(X;U) - I(Y;U) + \Delta_A, \end{cases}$$

for some auxiliary random variable $U \in \mathcal{U}$ *with a finite alphabet and jointly distributed with* X *and* Y.

Before we prove this theorem, we need two preliminary lemmas. The first lemma is an extension of Fano's inequality for list decoders and the second one makes use of a single-letterization technique to express $I(X^n;J) - I(Y^n;J)$ in a single-letter form in the sense of Csiszár and Körner [40].

Lemma 3.19 ([88, 5])**.** *Given a pair of random variables* (U, V) *defined over* $\mathcal{U} \times \mathcal{V}$ *for finite* $\mathcal{V}$ *and arbitrary* $\mathcal{U}$, *any list decoder* $g : \mathcal{U} \to 2^{\mathcal{V}}$ *of fixed list size* m *(i.e.,* $|g(u)| = m, \ \forall u \in \mathcal{U}$*), satisfies*

$$H(V|U) \leq h_{\mathsf{b}}(p_e) + p_e \log |\mathcal{V}| + (1 - p_e) \log m,$$

where $p_e := \Pr(V \notin g(U))$.

This lemma, applied to J and Y^n in place of U and V, respectively, implies that for any list decoder with the property (3.34), we have

$$H(Y^n|J) \leq \log |g_n(J)| + n\varepsilon_n, \tag{3.37}$$

where $\varepsilon_n := \frac{1}{n} + (\log |\mathcal{Y}| - \frac{1}{n} \log |g_n(J)|) p_e^{(n)}$ and $p_e^{(n)} = \Pr(Y^n \notin g_n(J))$ and hence, according to (3.34), $\varepsilon_n \to 0$ as $n \to \infty$.

Lemma 3.20. *Let (X^n, Y^n) be n i.i.d. copies of a pair of random variables (X, Y). Then for a random variable J jointly distributed with (X^n, Y^n), we have*

$$I(X^n; J) - I(Y^n; J) = \sum_{i=1}^{n} [I(X_i; U_i) - I(Y_i; U_i)],$$

where $U_i := (J, X_{i+1}^n, Y^{i-1})$.

Proof. Using the chain rule for the mutual information, we can express $I(X^n; J)$ as follows

$$\begin{aligned} I(X^n; J) &= \sum_{i=1}^{n} I(X_i; J | X_{i+1}^n) = \sum_{i=1}^{n} I(X_i; J, X_{i+1}^n) \\ &= \sum_{i=1}^{n} [I(X_i; J, X_{i+1}^n, Y^{i-1}) - I(X_i; Y^{i-1} | J, X_{i+1}^n)] \\ &= \sum_{i=1}^{n} I(X_i; U_i) - \sum_{i=1}^{n} I(X_i; Y^{i-1} | J, X_{i+1}^n). \end{aligned} \tag{3.38}$$

Similarly, we can expand $I(Y^n; J)$ as

$$\begin{aligned} I(Y^n; J) &= \sum_{i=1}^{n} I(Y_i; J | Y^{i-1}) = \sum_{i=1}^{n} I(Y_i; J, Y^{i-1}) \\ &= \sum_{i=1}^{n} [I(Y_i; J, X_{i+1}^n, Y^{i-1}) - I(Y_i; X_{i+1}^n | J, Y^{i-1})] \\ &= \sum_{i=1}^{n} I(Y_i; U_i) - \sum_{i=1}^{n} I(Y_i; X_{i+1}^n | J, Y^{i-1}). \end{aligned} \tag{3.39}$$

Subtracting (3.39) from (3.38), we get

$$I(X^n; J) - I(Y^n; J) = \sum_{i=1}^{n} [I(X_i; U_i) - I(Y_i; U_i)]$$

$$
\begin{aligned}
& -\sum_{i=1}^{n}[I(X_i;Y^{i-1}|J,X_{i+1}^n) - I(X_{i+1}^n;Y_i|J,Y^{i-1})] \\
\stackrel{(a)}{=}\ & \sum_{i=1}^{n}[I(X_i;U_i) - I(Y_i;U_i)],
\end{aligned}
$$

where (a) follows from the Csiszár sum identity [87]. □

Proof of Theorem 3.18. The rate R can be bounded as

$$
\begin{aligned}
nR &\geq H(J) \geq I(Y^n;J) = nH(Y) - H(Y^n|J) \\
&\stackrel{(a)}{\geq} nH(Y) - \log|g_n(J)| - n\varepsilon_n \stackrel{(b)}{=} n\mathit{\Delta}_A - n\varepsilon_n,
\end{aligned}
$$

where (a) follows from Fano's inequality (3.37) with $\varepsilon_n \to 0$ as $n \to \infty$ and (b) is due to (3.33). We can also upper bound $\mathit{\Delta}_A$ as

$$
\begin{aligned}
\mathit{\Delta}_A &\stackrel{(a)}{=} H(Y^n) - \log|g_n(J)| \\
&\stackrel{(b)}{\leq} H(Y^n) - H(Y^n|J) + n\varepsilon_n = \sum_{i=1}^{n} H(Y_i) - H(Y_i|Y^{i-1},J) + n\varepsilon_n \\
&\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|Y^{i-1},X_{i+1}^n,J) + n\varepsilon_n = \sum_{i=1}^{n} I(Y_i;U_i) + n\varepsilon_n, \qquad (3.40)
\end{aligned}
$$

where (a) follows from (3.33), (b) follows from (3.37), and in the last equality the auxiliary random variable $U_i := (Y^{i-1}, X_{i+1}^n, J)$ is introduced.

We shall now lower bound $I(X^n;J)$:

$$
\begin{aligned}
n(\mathit{\Delta}_M + \delta) &\geq I(X^n;J) \\
&\stackrel{(a)}{=} I(Y^n;J) + \sum_{i=1}^{n}[I(X_i;U_i) - I(Y_i;U_i)]
\end{aligned}
$$

$$\stackrel{(b)}{\geq} \quad n\Delta_A + \sum_{i=1}^{n}[I(X_i;U_i) - I(Y_i;U_i)] - n\varepsilon_n. \tag{3.41}$$

where (a) follows from Lemma 3.20 and (b) is due to Fano's inequality and (3.33) (or equivalently from (3.36)).

Combining (3.40), (3.40) and (3.41), we can write

$$\begin{aligned}
R &\geq \Delta_A - \varepsilon_n \\
\Delta_A &\leq I(Y_Q;U_Q|Q) + \varepsilon_n = I(Y_Q;U_Q,Q) + \varepsilon_n \\
\Delta_M &\geq \Delta_A + I(X_Q;U_Q|Q) - I(Y_Q;U_Q|Q) - \varepsilon'_n \\
&= \Delta_A + I(X_Q;U_Q,Q) - I(Y_Q;U_Q,Q) - \varepsilon'_n
\end{aligned}$$

where $\varepsilon'_n := \varepsilon_n + \delta$ and Q is a random variable distributed uniformly over $\{1,2,\ldots,n\}$ which is independent of (X,Y) and hence $I(Y_Q;U_Q|Q) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i;U_i)$. The results follow by denoting $U := (U_Q,Q)$ and noting that Y_Q and X_Q have the same distributions as Y and X, respectively. □

If the encoder does not have direct access to the private source X^n, then we can define the encoder mapping as $f_n : \mathcal{Y}^n \to [2^{nR}]$. The following corollary is an immediate consequence of Theorem 3.18.

Corollary 3.21. *If the encoder does not see the private source, then for all achievable* *dependence dilution triple (R, Δ_A, Δ_M), we have*

$$\begin{cases} R \geq \Delta_A \\ \Delta_A \leq I(Y;U) \\ \Delta_M \geq I(X;U) - I(Y;U) + \Delta_A, \end{cases}$$

for some joint distribution $P_{XYU} = P_{XY}P_{U|Y}$ where the auxiliary random variable $U \in \mathcal{U}$ satisfies $|\mathcal{U}| \leq |\mathcal{Y}| + 1$.

Remark 3.22. If source Y is required to be amplified (according to (3.36)) at maximum rate, that is, $\Delta_A = I(Y;U)$ for an auxiliary random variable U which satisfies $X \multimap Y \multimap U$, then by Corollary 3.21, the best privacy performance one can expect from the dependence dilution setting is

$$\Delta_M^* = \min_{\substack{U: X \multimap Y \multimap U \\ I(Y;U) \geq \Delta_A}} I(X;U), \tag{3.42}$$

which is equal to the dual of the rate-privacy function evaluated at Δ_A, i.e., $t(\mathsf{P}, \Delta_A)$, as defined in (3.4).

The dependence dilution problem is closely related to the discriminatory lossy source coding problem studied in [138]. In this problem, an encoder f observes (X^n, Y^n) and wants to describe this source to a decoder whose task is to recover Y^n within distortion level D and $I(f(X^n, Y^n); X^n) \leq n\Delta_M$. If the distortion level is Hamming measure, then the distortion constraint and the amplification constraint are closely related via Fano's inequality.

3.8 Observation Channels for Minimal and Maximal $\boldsymbol{g(\varepsilon)}$

In this section, we characterize the observation channels which achieve the lower or upper bounds on the rate-privacy function in (3.5). We first derive general conditions for achieving the lower bound and then present a large family of observation channels $P_{Y|X}$ which achieve the lower bound. We also give a family of $P_{Y|X}$ for which $g(\varepsilon)$ attains the upper bound in (3.5).

3.8.1 Conditions for Minimal $\boldsymbol{g(\varepsilon)}$

Assuming that $g(0) = 0$, we seek a set of conditions on P under which $g(\varepsilon)$ is linear in ε, or equivalently, $g(\varepsilon) = \varepsilon\frac{H(Y)}{I(X;Y)}$. In order to do this, we shall examine the slope of $g(\varepsilon)$ at zero. Recall that by concavity of g, it is clear that $g'(0) \geq \frac{H(Y)}{I(X;Y)}$. We strengthen this bound in the following lemmas.

Lemma 3.23. *For a given joint distribution* P *with marginals* p_X *and* q_Y*, if* $g(0) = 0$*, then we have*

$$g'(0) \geq \max_{y\in\mathcal{Y}} \frac{-\log \mathsf{q}(y)}{D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))}.$$

Proof. Given a joint distribution P defined over $\mathcal{X}\times\mathcal{Y}$ where $\mathcal{X} = [M]$ and $\mathcal{Y} = [N]$ with[7] $N \leq M$, we consider the following privacy filter: for $\delta > 0$ and $\mathcal{Z} = \{k, \mathsf{e}\}$ with a fixed integer $k \in \mathcal{Y}$

$$P_{Z|Y}(k|y) = \delta 1_{\{y=k\}} \tag{3.43}$$

$$P_{Z|Y}(\mathsf{e}|y) = 1 - \delta 1_{\{y=k\}}, \tag{3.44}$$

where $1_{\{\cdot\}}$ denotes the indicator function. The system of $X \multimap Y \multimap Z$ in this case is

[7] Recall that, according to Theorem 3.10, if $N > M$ then $g(0) > 0$.

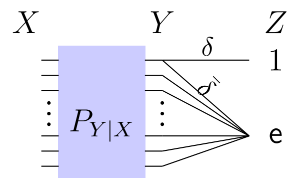


Figure 3.3: The privacy filter associated with (3.43) and (3.44) with $k = 1$.

depicted in Fig. 3.3 when $k = 1$. We clearly have $P_Z(k) = \delta \mathsf{q}_Y(k)$ and $P_Z(\mathsf{e}) = 1 - \delta \mathsf{q}_Y(k)$, and hence

$$P_{X|Z}(x|k) = \frac{P_{XZ}(x,k)}{\delta \mathsf{q}_Y(k)} = \frac{P_{XYZ}(x,k,k)}{\delta \mathsf{q}_Y(k)} = \frac{\delta \mathsf{P}(x,k)}{\delta \mathsf{q}_Y(k)} = P_{X|Y}(x|k),$$

and also,

$$\begin{aligned} P_{X|Z}(x|\mathsf{e}) &= \frac{P_{XZ}(x,\mathsf{e})}{1-\delta \mathsf{q}_Y(k)} = \frac{\sum_y P_{XYZ}(x,y,\mathsf{e})}{1-\delta \mathsf{q}_Y(k)} \\ &= \frac{\sum_{y\neq k} P_{XYZ}(x,y,\mathsf{e}) + \bar{\delta}\mathsf{P}(x,k)}{1-\delta \mathsf{q}_Y(k)} = \frac{\mathsf{p}_X(x) - \delta \mathsf{P}(x,k)}{1-\delta \mathsf{q}_Y(k)}. \end{aligned}$$

Therefore, we obtain $H(X|Z=k) = H(X|Y=k)$ for $k \in \mathcal{Y}$ and

$$H(X|Z=\mathsf{e}) = H\left(\frac{\mathsf{p}_X(1) - \delta \mathsf{P}(1,k)}{1-\delta \mathsf{q}_Y(k)}, \dots, \frac{\mathsf{p}_X(M) - \delta \mathsf{P}(M,k)}{1-\delta \mathsf{q}_Y(k)}\right) =: \hbar_X(\delta).$$

We then write

$$I(X;Z) = H(X) - H(X|Z) = H(X) - \delta \mathsf{q}_Y(k) H(X|Y=k) - (1-\delta \mathsf{q}_Y(k))\hbar_X(\delta),$$

and hence,

$$\frac{\mathrm{d}}{\mathrm{d}\delta} I(X;Z) = -\mathsf{q}_Y(k) H(X|Y=k) + \mathsf{q}_Y(k)\hbar_X(\delta) - (1-\delta \mathsf{q}_Y(k))\hbar'_X(\delta),$$

where

$$\hbar'_X(\delta) \coloneqq \frac{\mathrm{d}}{\mathrm{d}\delta}\hbar_X(\delta) = -\sum_{x=1}^{M} \frac{\mathsf{p}_X(x)\mathsf{q}_Y(k) - \mathsf{P}(x,k)}{[1-\delta\mathsf{q}_Y(k)]^2} \log\left(\frac{\mathsf{p}_X(x) - \delta\mathsf{P}(x,y)}{1-\delta\mathsf{q}_Y(k)}\right).$$

Using the first-order approximation of mutual information about $\delta = 0$, we can write

$$\begin{aligned} I(X;Z) &= \frac{\mathrm{d}}{\mathrm{d}\delta} I(X;Z)|_{\delta=0}\delta + o(\delta) = \delta\left[\sum_{x=1}^{M} \mathsf{P}(x,k)\log\left(\frac{\mathsf{P}(x,k)}{\mathsf{p}_X(x)\mathsf{q}_Y(k)}\right)\right] + o(\delta) \\ &= \delta\mathsf{q}_Y(k) D(P_{X|Y}(\cdot|k)\|\mathsf{p}_X(\cdot)) + o(\delta). \end{aligned} \tag{3.45}$$

Similarly, we can write

$$\begin{aligned} I(Y;Z) &= h(Z) - \sum_{y=1}^{N} \mathsf{q}_Y(y) h(Z|Y=y) = h(Z) - \mathsf{q}_Y(k)h(\delta) = h(\delta\mathsf{q}_Y(k)) - \mathsf{q}_Y(k)h(\delta) \\ &= -\delta\mathsf{q}_Y(k)\log(\mathsf{q}_Y(k)) - \Psi(1-\delta\mathsf{q}_Y(k)) + \mathsf{q}_Y(k)\Psi(\bar{\delta}), \end{aligned}$$

where $\Psi(x) \coloneqq x\log x$ which yields

$$\frac{\mathrm{d}}{\mathrm{d}\delta} I(Y;Z) = -\Psi(\mathsf{q}_Y(k)) + \mathsf{q}_Y(k)\log\left(\frac{1-\delta\mathsf{q}_Y(k)}{\bar{\delta}}\right).$$

From the above, we obtain

$$\begin{aligned} I(Y;Z) &= \frac{\mathrm{d}}{\mathrm{d}\delta} I(Y;Z)|_{\delta=0}\delta + o(\delta) \\ &= -\delta\Psi(\mathsf{q}_Y(k)) + o(\delta). \end{aligned} \tag{3.46}$$

Expression (3.45) implies that the filter $P_{Z|Y}$, specified in (3.43) and (3.44), satisfies the privacy constraint $I(X;Z) \leq \varepsilon$ (and thus belongs to $\mathcal{D}_\varepsilon(\mathsf{P})$) if

$$\frac{\varepsilon}{\delta} = \mathsf{q}_Y(k) D(P_{X|Y}(\cdot|k) \| \mathsf{p}_X(\cdot)) + \frac{o(\delta)}{\delta},$$

and hence from (3.46), we have

$$I(Y;Z) = \frac{-\Psi(\mathsf{q}_Y(k))}{\mathsf{q}_Y(k) D(P_{X|Y}(\cdot|k) \| \mathsf{p}_X(\cdot))} \varepsilon + o(\delta).$$

This immediately implies that

$$g'(0) = \lim_{\varepsilon \downarrow 0} \frac{g(\varepsilon)}{\varepsilon} \geq \frac{-\Psi(\mathsf{q}_Y(k))}{\mathsf{q}_Y(k) D(P_{X|Y}(\cdot|k) \| \mathsf{p}_X(\cdot))} = \frac{-\log(\mathsf{q}_Y(k))}{D\left(P_{X|Y}(\cdot|k) \| \mathsf{p}_X(\cdot)\right)}, \tag{3.47}$$

where we have used the assumption $g(0) = 0$ in the first equality. □

Remark 3.24. Note that with the assumption $g(0) = 0$, the right-hand side of inequality in Lemma 3.23 can not be infinity. We prove this fact by contradiction. To do this, suppose that there exists $y_0 \in \mathcal{Y}$ such that $D(P_{X|Y}(\cdot|y_0) \| \mathsf{p}_X(\cdot)) = 0$, and consequently $P_{X|Y}(\cdot|y_0) = \mathsf{p}_X(\cdot)$. Consider the binary random variable $Z \in \{1, \mathsf{e}\}$ constructed according to the distribution $P_{Z|Y}(1|y_0) = 1$ and $P_{Z|Y}(\mathsf{e}|y) = 1$ for all $y \in \mathcal{Y} \backslash \{y_0\}$. We can now claim that Z is independent of X, because $P_{X|Z}(\cdot|1) = P_{X|Y}(\cdot|y_0) = \mathsf{p}_X(\cdot)$, and for all $x \in \mathcal{X}$ we have

$$\begin{aligned} P_{X|Z}(x|\mathsf{e}) &= \sum_{y \neq y_0} P_{X|Y}(x|y) P_{Y|Z}(y|\mathsf{e}) = \sum_{y \neq y_0} P_{X|Y}(x|y) \frac{\mathsf{q}_Y(y)}{1 - \mathsf{q}_Y(y_0)} \\ &= \frac{1}{1 - \mathsf{q}_Y(y_0)} \sum_{y \neq y_0} \mathsf{P}(x, y) = \mathsf{p}_X(x). \end{aligned}$$

On the other hand, Z and Y are clearly not independent. Therefore, we have $g(0) > 0$

which contradicts our assumption $g(0) = 0$.

In order to prove the main result, we need the following simple lemma.

Lemma 3.25. *For any joint distribution* P *with marginals* p_X *and* q_Y*, we have*

$$\frac{H(Y)}{I(X;Y)} \le \max_{y \in \mathcal{Y}} \frac{-\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y) \| \mathsf{p}_X(\cdot))},$$

where equality holds if and only if there exists a constant $c > 0$ *such that* $-\log \mathsf{q}_Y(y) = cD(P_{X|Y}(\cdot|y) \| \mathsf{p}_X(\cdot))$ *for all* $y \in \mathcal{Y}$*.*

Proof. It is clear that

$$\frac{H(Y)}{I(X;Y)} = \frac{-\sum_{y \in \mathcal{Y}} \mathsf{q}_Y(y) \log \mathsf{q}_Y(y)}{\sum_{y \in \mathcal{Y}} \mathsf{q}_Y(y) D(P_{X|Y}(\cdot|y) \| \mathsf{p}_X(\cdot))} \le \max_{y \in \mathcal{Y}} \frac{-\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y) \| \mathsf{p}_X(\cdot))},$$

where the inequality follows from the fact that for any three sequences of positive numbers $\{a_i\}_{i=1}^n$, $\{b_i\}_{i=1}^n$, and $\{\lambda_i\}_{i=1}^n$ we have $\frac{\sum_{i=1}^n \lambda_i a_i}{\sum_{i=1}^n \lambda_i b_i} \le \max_{1 \le i \le n} \frac{a_i}{b_i}$, where equality occurs if and only if $\frac{a_i}{b_i} = c$ for all $i \in [n]$. $\square$

Now we are ready to state the main result of this section.

Theorem 3.26. *For a given joint distribution* P *with marginals* p_X *and* q_Y*, if* $g(0) = 0$ *and* $g(\cdot)$ *is linear on* $[0, I(X;Y)]$*, then for any* $y \in \mathcal{Y}$

$$\frac{H(Y)}{I(X;Y)} = \frac{-\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y) \| \mathsf{p}_X(\cdot))}.$$

Proof. Note that the facts that $g(0) = 0$ and $g(\cdot)$ is linear are equivalent to $g(\varepsilon) = \varepsilon \frac{H(Y)}{I(X;Y)}$.

It is, therefore, immediate from Lemmas 3.23 and 3.25 that we have

$$g'(0) \stackrel{(a)}{=} \frac{H(Y)}{I(X;Y)} \stackrel{(b)}{\leq} \max_{y\in\mathcal{Y}} \frac{-\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))} \stackrel{(c)}{\leq} g'(0), \tag{3.48}$$

where (a) follows from the fact that $g(\varepsilon) = \varepsilon\frac{H(Y)}{I(X;Y)}$ and (b) and (c) are due to Lemmas 3.25 and 3.23, respectively. Hence, we obtain

$$\frac{H(Y)}{I(X;Y)} = \max_{y\in\mathcal{Y}} \frac{-\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))}. \tag{3.49}$$

According to Lemma 3.25, (3.49) implies that the ratio $\frac{-\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))}$ does not depend on $y \in \mathcal{Y}$ and hence the result follows. □

This theorem implies that if there exists $y = y_1$ and $y = y_2$ such that $\frac{\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))}$ results in two different values, then the lower bound in (3.5) is not achievable, that is, we have

$$g(\varepsilon) > \varepsilon\frac{H(Y)}{I(X;Y)}.$$

This, therefore, gives a necessary condition for the lower bound to be achievable. The following corollary simplifies this necessary condition.

Corollary 3.27. *If $g(0) = 0$ and $g(\cdot)$ is linear, then the following are equivalent:*

(i) Y *is uniformly distributed,*

(ii) $D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))$ *is constant for all* $y \in \mathcal{Y}$.

Proof. $(ii) \Rightarrow (i)$:

From Theorem 3.26, we have for all $y \in \mathcal{Y}$

$$\frac{H(Y)}{I(X;Y)} = \frac{-\log \mathsf{q}_Y(y))}{D\left(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot)\right)}. \tag{3.50}$$

Letting $D := D\left(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot)\right)$ for any $y \in \mathcal{Y}$, we have $\sum_y \mathsf{q}_Y(y)D = I(X;Y)$ and hence $D = I(X;Y)$, which together with (3.50) implies that $H(Y) = -\log \mathsf{q}_Y(y)$ for all $y \in \mathcal{Y}$ and hence Y is uniformly distributed.

$(i) \Rightarrow (ii)$:

When Y is uniformly distributed, we have from (3.50) that $I(X;Y) = D\left(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot)\right)$ which implies that $D\left(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot)\right)$ is constant for all $y \in \mathcal{Y}$. □

To illustrate this corollary, consider the following examples.

Example 3.28. Suppose $P_{Y|X} = \mathsf{BSC}(\alpha)$ with $\alpha \in (0,1)$ and $\mathsf{p}_X = \mathsf{Bernoulli}(\frac{1}{2})$. In this case, we have $P_{X|Y} = \mathsf{BSC}(\alpha)$ with $\mathsf{q}_Y = \mathsf{Bernoulli}(\frac{1}{2})$. Note that Corollary 3.13 implies that $g(0) = 0$. It was shown in Theorem 3.8 that $g(\cdot)$ is linear and hence according to Corollary 3.27, $D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))$ must be constant for $y \in \{0,1\}$. It is simple to verify that $D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot)) = 1 - h_{\mathsf{b}}(\alpha)$ for $y \in \{0,1\}$.

Example 3.29. It was shown in Theorem 3.7 that if $P_{X|Y} = \mathsf{BEC}(\delta)$ and $\mathsf{q}_Y = \mathsf{Bernoulli}(q)$ with $0 \le q \le \frac{1}{2}$, then $g(\cdot)$ is linear. In this case, we have $D(P_{X|Y}(\cdot|0)\|\mathsf{P}_X(\cdot)) = -\bar{\delta}\log\bar{q}$ and $D(P_{X|Y}(\cdot|1)\|\mathsf{P}_X(\cdot)) = -\bar{\delta}\log q$ which show that $D(P_{X|Y}(\cdot|y)\|\mathsf{P}_X(\cdot))$ is constant if and only if $q = \frac{1}{2}$.

Example 3.30. Now suppose $P_{X|Y}$ is a binary asymmetric channel such that $P_{X|Y}(\cdot|0) = \mathsf{Bernoulli}(\alpha)$, and $P_{X|Y}(\cdot|1) = \mathsf{Bernoulli}(\beta)$ for some $0 < \alpha, \beta < 1$ and input distribution $\mathsf{q}_Y = \mathsf{Bernoulli}(q)$ with $0 < q \le \frac{1}{2}$. It is easy to see that if $\alpha + \beta = 1$ then

$D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))$ does not depend on y and hence we can conclude from Corollary 3.27 that $g(\cdot)$ is not linear for any $q < \frac{1}{2}$ and thus we have $g(\varepsilon) > \varepsilon\frac{h_{\mathsf{b}}(q)}{I(X;Y)}$.

In Theorem 3.26, we showed that when $g(\varepsilon)$ achieves its lower bound, given in (3.5), the slope of the mapping $\varepsilon \mapsto g(\varepsilon)$ at zero is equal to $\frac{-\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))}$ for any $y \in \mathcal{Y}$. We will show in the next section that the reverse direction is also true at least for a large family of binary input symmetric output channels, thereby showing that in this case,

$$g'(0) = \frac{-\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))}, \quad \forall y \in \mathcal{Y} \Longleftrightarrow g(\varepsilon) = \varepsilon\frac{H(Y)}{I(X;Y)}, \quad 0 \le \varepsilon \le I(X;Y).$$

3.8.2 Binary Input Symmetric Output Channels

In this section, we apply the results of the previous section to a particular joint distribution. Specifically, we look at the case where Y is binary and the reverse channel $P_{X|Y}$ respects a certain notion of symmetry.

Suppose $\mathcal{Y} = \{0, 1\}$, $\mathcal{X} = \{0, \pm 1, \pm 2, \ldots, \pm k\}$ for some integer $k \ge 1$, and $P_{X|Y}(x|1) = P_{X|Y}(-x|0)$ for any $x \in \mathcal{X}$. This channel is called *binary input symmetric output* (BISO) [63, 136]. For $x = 0$, we have $p_0 \coloneqq P_{X|Y}(0|0) = P_{X|Y}(0|1)$. We notice that with this definition of symmetry, we can always assume that the output alphabet $\mathcal{X}$ has even number of elements because we can split $X = 0$ into two outputs, $X = 0^+$ and $X = 0^-$, with $P_{X|Y}(0^-|0) = P_{X|Y}(0^+|0) = \frac{p_0}{2}$ and $P_{X|Y}(0^-|1) = P_{X|Y}(0^+|1) = \frac{p_0}{2}$. The new channel is clearly essentially equivalent to the original one. This family of channels can also be characterized using the definition of *quasi-symmetric* channels [7, Definition 4.17]. A channel W is BISO if (after making $|\mathcal{X}|$ even) the transition matrix $P_{X|Y}$ can be partitioned along its columns into binary input binary output sub-arrays in which rows are permutations of each other and the column sums are equal. It is clear that BSC and BEC are both examples of BISO. The following lemma gives an upper bound for $g(\varepsilon)$ when $P_{X|Y}$ belongs to such a family of channels.

Lemma 3.31. *If $P_{X|Y}$ is BISO, then we have for $\varepsilon \in [0, I(X;Y)]$*

$$\varepsilon \frac{H(Y)}{I(X;Y)} \leq g(\varepsilon) \leq H(Y) - \frac{I(X;Y)-\varepsilon}{C(P_{X|Y})},$$

where $C(P_{X|Y})$ denotes the capacity of $P_{X|Y}$.

Proof. The lower bound was already shown in Lemma 3.1. To prove the upper bound note that by Markov condition $X \multimap Y \multimap Z$, we have for any $x \in \mathcal{X}$ and $z \in \mathcal{Z}$

$$P_{X|Z}(x|z) = P_{X|Y}(x|0)P_{Y|Z}(0|z) + P_{X|Y}(x|1)P_{Y|Z}(1|z). \tag{3.51}$$

Now suppose $\mathcal{Z}_0 := \{z : P_{Y|Z}(0|z) \leq P_{Y|Z}(1|z)\}$ and similarly $\mathcal{Z}_1 := \{z : P_{Y|Z}(1|z) \leq P_{Y|Z}(0|z)\}$. Then (3.51) allows us to write for $z \in \mathcal{Z}_0$

$$P_{X|Z}(x|z) = P_{X|Y}(x|0)h_b^{-1}(H(Y|Z=z)) + P_{X|Y}(x|1)(1 - h_b^{-1}(H(Y|Z=z))), \tag{3.52}$$

where $h_b^{-1} : [0,1] \to [0, \frac{1}{2}]$ is the inverse of binary entropy function, and for $z \in \mathcal{Z}_1$,

$$P_{X|Z}(x|z) = P_{X|Y}(x|0)(1 - h_b^{-1}(H(Y|Z=z))) + P_{X|Y}(x|1)h_b^{-1}(H(Y|Z=z)). \tag{3.53}$$

Letting $P \otimes h_b^{-1}(H(Y|z))$ and $\tilde{P} \otimes h_b^{-1}(H(Y|z))$ denote the right-hand sides of (3.52) and (3.53), respectively, we can write

$$H(X|Z) = \sum_{z \in \mathcal{Z}} P_Z(z) H(X|Z=z)$$

$$
\begin{aligned}
&\stackrel{(a)}{=} \sum_{z\in\mathcal{Z}_0} P_Z(z)H(P\otimes h_b^{-1}(H(Y|Z=z))) + \sum_{z\in\mathcal{Z}_1} P_Z(z)H(\tilde{P}\otimes h_b^{-1}(H(Y|Z=z)))\\
&\stackrel{(b)}{\leq} \sum_{z\in\mathcal{Z}_0} P_Z(z)\left[(1-H(Y|Z=z))H(P\otimes h_b^{-1}(0)) + H(Y|Z=z)H(P\otimes h_b^{-1}(1))\right]\\
&\quad + \sum_{z\in\mathcal{Z}_1} P_Z(z)\left[(1-H(Y|Z=z))H(\tilde{P}\otimes h_b^{-1}(0)) + H(Y|Z=z)H(\tilde{P}\otimes h_b^{-1}(1))\right]\\
&\stackrel{(c)}{=} \sum_{z\in\mathcal{Z}_0} P_Z(z)\left[(1-H(Y|Z=z))H(X|Y) + H(Y|Z=z)H(X_{\text{unif}})\right]\\
&\quad + \sum_{z\in\mathcal{Z}_1} P_Z(z)\left[(1-H(Y|Z=z))H(X|Y) + H(Y|Z=z)H(X_{\text{unif}})\right]\\
&= H(X|Y)[1-H(Y|Z)] + H(Y|Z)H(X_{\text{unif}}),
\end{aligned}
$$

where $H(X_{\text{unif}})$ denotes the entropy of X when Y is uniformly distributed. Here, (a) is due to (3.52) and (3.53), (b) follows form convexity of $u \mapsto H(P\otimes h_b^{-1}(u))$ for all $u\in[0,1]$ [34] and Jensen's inequality. In (c), we used the symmetry of channel $P_{X|Y}$ to show that $H(X|Y=0) = H(X|Y=1) = H(X|Y)$. Hence, we obtain

$$
H(Y|Z) \geq \frac{H(X|Z)-H(X|Y)}{H(X_{\text{unif}})-H(X|Y)} = \frac{I(X;Y)-I(X;Z)}{C(P_{X|Y})},
$$

where the equality follows from the fact that for BISO channels (and in general for any quasi-symmetric channels) the uniform input distribution is the capacity-achieving distribution [7, Lemma 4.18]. Since $g(\varepsilon)$ is attained when $I(X;Z)=\varepsilon$, the conclusion immediately follows. $\square$

This lemma demonstrates that the larger the gap between $I(X;Y)$ and $I(X;Y')$ is for $Y' \sim \mathsf{Bernoulli}(\frac{1}{2})$, the more $g(\cdot)$ deviates from its lower bound. When $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$, then $C(P_{Y|X}) = I(X;Y)$ and $H(Y) = 1$ and hence Lemma 3.31 implies that

$$\frac{\varepsilon}{I(X;Y)} \le g(\varepsilon) \le 1 - \frac{I(X;Y) - \varepsilon}{I(X;Y)} = \frac{\varepsilon}{I(X;Y)},$$

and hence we have the following corollary.

Corollary 3.32. *If $P_{X|Y}$ is BISO and $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$, then we have for any $\varepsilon \le I(X;Y)$*

$$g(\varepsilon) = \frac{\varepsilon}{I(X;Y)}.$$

This corollary now enables us to prove the reverse direction of Theorem 3.26 for the family of BISO channels.

Theorem 3.33. *If $P_{X|Y}$ is a BISO channel, then the following statements are equivalent:*

(i) $g(\varepsilon) = \varepsilon \frac{H(Y)}{I(X;Y)}$ *for* $0 \le \varepsilon \le I(X;Y)$.

(ii) The initial efficiency of the privacy-constrained information extraction is

$$g'(0) = \frac{-\log \mathsf{q}_Y(y)}{D(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))}, \quad \forall y \in \mathcal{Y}.$$

Proof. The fact that (i) implies (ii) follows directly from Theorem 3.26. To show that (ii) implies (i), let $\mathsf{q}_Y = \mathsf{Bernoulli}(q)$ and as before $\mathcal{X} = \{\pm 1, \pm 2, \ldots, \pm k\}$. We then have

$$\frac{-\log \mathsf{q}_Y(0)}{D(P_{X|Y}(\cdot|0)\|\mathsf{p}_X(\cdot))} = \frac{\log \bar{q}}{H(X|Y) + \sum_{x=-k}^{k} P_{X|Y}(x|0) \log \mathsf{p}_X(x)}, \tag{3.54}$$

and

$$\frac{-\log \mathsf{q}_Y(1)}{D(P_{X|Y}(\cdot|1)\|\mathsf{p}_X(\cdot))} = \frac{\log q}{H(X|Y) + \sum_{x=-k}^{k} P_{X|Y}(x|1) \log \mathsf{p}_X(x)}. \tag{3.55}$$

By assumption, we can write

$$\frac{\log \bar{q}}{H(X|Y) + \sum_{x=-k}^{k} P_{X|Y}(x|0) \log \mathsf{p}_X(x)} = \frac{\log q}{H(X|Y) + \sum_{x=-k}^{k} P_{X|Y}(x|1) \log \mathsf{p}_X(x)}. \tag{3.56}$$

It is shown in Appendix A that (3.56) holds if and only if $q = \frac{1}{2}$. Now we can invoke Corollary 3.32 to conclude that $g(\varepsilon) = \varepsilon \frac{H(Y)}{I(X;Y)}$. $\square$

Remark 3.34. Theorem 3.33 states that if $P_{X|Y}$ is BISO and $\mathsf{q}_Y = \mathsf{Bernoulli}(\frac{1}{2})$, then $g(\varepsilon) = \frac{\varepsilon}{I(X;Y)}$ which is a generalization of Theorems 3.8 and 3.7 in the uniform case. Furthermore, since in this case $g(\varepsilon)$ coincides with its lower bound, the erasure filter, illustrated in Fig. 3.2, is an optimal filter for any ε.

Note that if $P_{X|Y} = \mathsf{BSC}(\alpha)$ and $\mathsf{q}_Y = \mathsf{Bernoulli}(\frac{1}{2})$, then $P_{Y|X} = \mathsf{BSC}(\alpha)$ with $\mathsf{p}_X = \mathsf{Bernoulli}(\frac{1}{2})$. The following corollary specializes Corollary 3.32 for this case.

Corollary 3.35. *If* $\mathsf{p}_X = \mathsf{Bernoulli}(\frac{1}{2})$ *and* $P_{Y|X} = \mathsf{BSC}(\alpha)$ *with* $0 < \alpha < \frac{1}{2}$*, then* $g(\varepsilon) = \frac{\varepsilon}{I(X;Y)}$ *for* $0 \leq \varepsilon \leq I(X;Y)$*. Furthermore,* $\mathsf{BEC}(\delta(\varepsilon, \alpha))$ *is an optimal filter, where*

$$\delta(\varepsilon, \alpha) := 1 - \frac{\varepsilon}{I(X;Y)}. \tag{3.57}$$

3.8.3 Erasure Observation Channel

In this section, we obtain a sufficient condition for the joint distribution P under which $g(\varepsilon)$ attains its upper bound given in (3.5). Before that, recall from (3.9) and Lemma 3.1 that for $0 \leq \varepsilon \leq I(X;Y)$

$$\varepsilon \frac{H(Y)}{I(X;Y)} + g(0) \left[1 - \frac{\varepsilon}{I(X;Y)}\right] \leq g(\varepsilon) \leq H(Y|X) + \varepsilon, \tag{3.58}$$

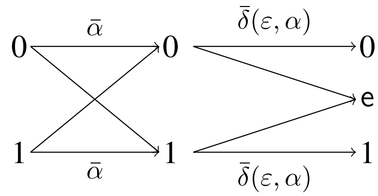


Figure 3.4: Optimal privacy filter for $P_{Y|X} = \mathsf{BSC}(\alpha)$ and uniform X, where $\delta(\varepsilon, \alpha)$ is specified in (3.57).

In the following, we show that the above upper and lower bounds coincide when $P_{Y|X}$ is an erasure channel, i.e., $\mathcal{Y} = \mathcal{X} \cup \{\mathsf{e}\}$ and there exists $0 \leq \delta \leq 1$ such that $P_{Y|X}(x|x) = 1 - \delta$ and $P_{Y|X}(\mathsf{e}|x) = \delta$ for all $x \in \mathcal{X}$.

Lemma 3.36. *If $P_{Y|X}$ is an erasure channel (as defined above), then for any $0 \leq \varepsilon \leq I(X;Y)$*

$$g(\varepsilon) = H(Y|X) + \varepsilon.$$

Proof. Notice that if $g(0) = H(Y|X)$, then the lower bound in (3.58) becomes $H(Y|X)+\varepsilon$ and thus $g(\varepsilon) = H(Y|X) + \varepsilon$. Therefore, it suffices to show that if $P_{Y|X}$ is an erasure channel, then $g(0) = H(Y|X)$.

Recall that $|\mathcal{X}| = M$ and $\mathcal{Y} = \mathcal{X} \cup \{\mathsf{e}\}$. Consider the following privacy filter that generates $Z \in \mathcal{Y}$:

$$P_{Z|Y}(z|y) = \begin{cases} \frac{1}{M}, & \text{if } y \neq \mathsf{e}, z \neq \mathsf{e}, \\ 1, & \text{if } y = z = \mathsf{e}. \end{cases}$$

For any $x \in \mathcal{X}$, we have

$$P_{Z|X}(z|x) = P_{Z|Y}(z|x)P_{Y|X}(x|x) + P_{Z|Y}(z|\mathsf{e})P_{Y|X}(\mathsf{e}|x) = \left[\frac{\bar{\delta}}{M}\right] 1_{\{z \neq \mathsf{e}\}} + \delta 1_{\{z = \mathsf{e}\}},$$

which implies $Z \perp\!\!\!\perp X$ and thus $I(X;Z) = 0$. On the other hand, $P_Z(z) = \left(\frac{\bar{\delta}}{M}\right) 1_{\{z \neq \mathsf{e}\}} +$

$\delta 1_{\{z=\mathsf{e}\}}$, and therefore we have

$$\begin{aligned} g(0) &\geq I(Y;Z) = H(Z) - H(Z|Y) = H\left(\frac{\bar{\delta}}{M}, \dots, \frac{\bar{\delta}}{M}, \delta\right) - \bar{\delta}\log M \\ &= h_{\mathsf{b}}(\delta) = H(Y|X). \end{aligned}$$

It follows from Lemma 3.1 that $g(0) = H(Y|X)$, and thus the proof is complete. $\square$

Although Lemma 3.36 does not specify the optimal filter, we demonstrate in the following example that if $P_{Y|X} = \mathsf{BEC}(\delta)$, then a ternary-valued Z is sufficient to achieve $g(\varepsilon)$.

Example 3.37. Suppose $\mathsf{p}_X = \mathsf{Bernoulli}(p)$ and $P_{Y|X} = \mathsf{BEC}(\delta)$. Consider the privacy filter described as: $P_{Z|Y}(\mathsf{e}|\mathsf{e}) = 1$, $P_{Z|Y}(0|y) = \bar{\alpha}$, and $P_{Z|Y}(1|y) = \alpha$ for $y \neq \mathsf{e}$, with a fixed $0 \leq \alpha \leq \frac{1}{2}$. Easy calculation reveals that $I(X;Z) = \bar{\delta}[h_{\mathsf{b}}(\alpha * p) - h_{\mathsf{b}}(\alpha)]$ and $I(Y;Z) = h_{\mathsf{b}}(\delta) + \bar{\delta}[h_{\mathsf{b}}(\alpha * p) - h_{\mathsf{b}}(\alpha)]$. Setting $I(X;Z) = \varepsilon$ therefore implies $I(Y;Z) = h_{\mathsf{b}}(\delta) + \varepsilon$, which is the upper bound given in (3.58). Thus, the optimal privacy filter is a combination of an identity channel and a $\mathsf{BSC}(\alpha(\varepsilon, \delta))$, as shown in Fig. 3.5, where $0 \leq \alpha(\varepsilon, \delta) \leq \frac{1}{2}$ is the unique solution of

$$\bar{\delta}[h_{\mathsf{b}}(\alpha * p) - h_{\mathsf{b}}(\alpha)] = \varepsilon. \tag{3.59}$$

We note that for fixed $0 < \delta < 1$ and $0 \leq p \leq 1$, the map $\alpha \mapsto \bar{\delta}[h_{\mathsf{b}}(\alpha * p) - h_{\mathsf{b}}(\alpha)]$ is monotonically decreasing on $[0, \frac{1}{2}]$ ranging over $[0, \bar{\delta}h_{\mathsf{b}}(p)]$ and since $\varepsilon \leq I(X;Y) = \bar{\delta}h_{\mathsf{b}}(p)$, the solution of equation 3.59 is unique.

Combining Corollary 3.35 with Lemma 3.36, we obtain the following *extremal property* of the BEC and BSC. For $X \sim \mathsf{p}_X = \mathsf{Bernoulli}(\frac{1}{2})$, we have for any channel $P_{Y|X}$,

$$g(\varepsilon) \geq \frac{\varepsilon H(Y)}{I(X;Y)} = g(\mathsf{p}_X \times \mathsf{BSC}(\hat{\alpha}), \varepsilon),$$

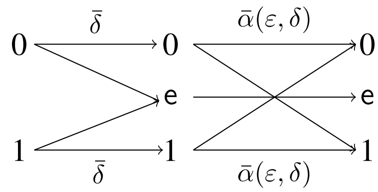


Figure 3.5: Optimal privacy filter for $P_{Y|X} = \mathsf{BEC}(\delta)$, where $\delta(\varepsilon, \alpha)$ is specified in (3.59).

where $\hat{\alpha} \coloneqq h_b^{-1}\left(\frac{H(Y|X)}{H(Y)}\right)$. Similarly, if $\mathsf{P}_X = \mathsf{Bernoulli}(p)$, we have for any channel $P_{Y|X}$ with $H(Y|X) \leq 1$

$$g(\varepsilon) \leq H(Y|X) + \varepsilon = g(\mathsf{p}_X \times \mathsf{BEC}(\hat{\delta}), \varepsilon),$$

where $\hat{\delta} \coloneqq h_{\mathsf{b}}^{-1}(H(Y|X))$.

Chapter 4

Information Extraction Under an Information-Theoretic Privacy Constraint: Absolutely Continuous Case

4.1 Overview

In this section, we extend the rate-privacy function g to the continuous case. Specifically, we assume that the private and observable data are continuous random variables and that the filter is composed of two stages: first Gaussian noise is added to the observable data and then the resulting random variable is quantized using an M-bit accuracy uniform scalar quantizer (for some positive integer $M \in \mathbb{N}$). These filters are of practical interest as they can be easily implemented. This section is divided in two parts, in the first we discuss general properties of the rate-privacy function and in the second we study approximating the rate-privacy function for sufficiently small privacy level ε.

4.1.1 Main Contributions

The main contributions of this chapter are as follows:

- We formulate the rate-privacy function for the continuous random variables X and Y by assuming that the privacy filter belongs to a family of additive noise channels followed by an M-level uniform scalar quantizer.

- We obtain asymptotic bounds as $M \to \infty$ for the rate-privacy function and show that some of the properties of g in the discrete case do not hold in the continuous case.
- We further show that $g(0) = 0$ for any joint distribution P, thus perfect privacy implies trivial utility. We then express the initial efficiency of privacy-constrained information extraction, $g'(0)$, in terms of the so-called *one-sided maximal correlation*.
- Finally, we obtain a second-order approximation for $g(\varepsilon)$ when ε is in the almost perfect privacy regime and show the accuracy of this approximation in the Gaussian case.
- As by-products, we derive two strong data processing inequalities for mutual information as well as MMSE in the special case of AWGN channel.

4.2 General properties of the rate-privacy function

We assume throughout this chapter that the random vector (X, Y) is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$. Additionally, we assume that its joint density $f_{X,Y}$ satisfies the following:

(a) there exist constants $C_1 > 0$, $p > 1$ and bounded function $C_2 : \mathbb{R} \to \mathbb{R}$ such that

$$f_Y(y) \leq C_1 |y|^{-p},$$

and also for $x \in \mathbb{R}$

$$f_{Y|X}(y|x) \leq C_2(x) |y|^{-p},$$

(b) $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^2]$ are both finite,

(c) the differential entropy of (X,Y) satisfies $h(X,Y) > -\infty$,

Note that assumptions (b) and (c) together imply that $h(X)$, $h(Y)$, and $h(X,Y)$ are finite, i.e., the maps $x \mapsto f_X(x)\log f_X(x)$, $y \mapsto f_Y(y)\log f_Y(y)$, and $(x,y) \mapsto f_{X,Y}(x,y)\log(f_{X,Y}(x,y))$ are integrable. Note also that assumption (b) implies that $H(\lfloor Y \rfloor) < \infty$, where $\lfloor a \rfloor$ denotes the largest integer ℓ such that $\ell \le a$ [149]. We also assume that X and Y are not independent, since otherwise the problem of characterizing $g(\varepsilon)$ becomes trivial by assuming that the displayed data Z can equal the observable data Y.

We are interested in filters of the form $\mathcal{Q}_M(Y + \lambda N_{\mathsf{G}})$, where $\lambda \ge 0$, $N_{\mathsf{G}} \sim \mathcal{N}(0,1)$ is independent of (X,Y), and for any positive integer M, $\mathcal{Q}_M$ denotes the M-bit accuracy uniform scalar quantizer, i.e., for all $x \in \mathbb{R}$

$$\mathcal{Q}_M(x) = \frac{1}{2^M}\left\lfloor 2^M x \right\rfloor.$$

Let $U_\lambda := Y + \lambda N_{\mathsf{G}}$ and $U_\lambda^M := \mathcal{Q}_M(U_\lambda) = \mathcal{Q}_M(Y + \lambda N_{\mathsf{G}})$. We define, for any $M \in \mathbb{N}$,

$$g_M(\varepsilon) := \sup_{\substack{\lambda \ge 0, \\ I(X;U_\lambda^M) \le \varepsilon}} I(Y; U_\lambda^M), \tag{4.1}$$

and similarly

$$g(\varepsilon) := \sup_{\substack{\lambda \ge 0, \\ I(X;U_\lambda) \le \varepsilon}} I(Y; U_\lambda). \tag{4.2}$$

The main result of this section proves that $g(\varepsilon)$ is indeed the limit of $g_M(\varepsilon)$ as $M \to \infty$. In order to prove this result, we need the following lemmas whose proofs are given in

Appendices B.1, B.2, and B.3.

Lemma 4.1. *The function $\lambda \mapsto I(Y;U_\lambda)$ is strictly decreasing and continuous. Additionally, it satisfies*

$$I(Y;U_\lambda) \le \frac{1}{2}\log\left(1 + \frac{\mathsf{var}(Y)}{\lambda^2}\right).$$

with equality if and only if Y is Gaussian. In particular, $I(Y;U_\lambda) \to 0$ as $\lambda \to \infty$.

Lemma 4.2. *The function $\lambda \mapsto I(X;U_\lambda)$ is strictly decreasing and continuous. Moreover, $I(X;U_\lambda) \to 0$ when $\lambda \to \infty$.*

In light of Lemmas 4.2 and 4.1, there exists a unique $\lambda_\varepsilon \in (0,\infty)$ for every $0 < \varepsilon < I(X;Y)$ such that $I(X;U_{\lambda_\varepsilon}) = \varepsilon$ and $g(\varepsilon) = I(Y;U_{\lambda_\varepsilon})$, and thus $g(\varepsilon)$ corresponds to the smallest variance of Gaussian noise which results in $I(X;U_\lambda) = \varepsilon$.

Lemma 4.3. *The functions $\lambda \mapsto I(X;U_\lambda^M)$ and $\lambda \mapsto I(Y;U_\lambda^M)$ are continuous for each $M \in \mathbb{N}$ and satisfy for any $\lambda \ge 0$*

$$\lim_{M\to\infty} I(X;U_\lambda^M) = I(X;U_\lambda) \qquad \text{and} \qquad \lim_{M\to\infty} I(Y;U_\lambda^M) = I(Y;U_\lambda). \tag{4.3}$$

We are now in position to state the main result of this section.

Theorem 4.4. *Let $\varepsilon > 0$ be fixed. Then $\lim_{M\to\infty} g_M(\varepsilon) = g(\varepsilon)$.*

Proof. For every $M \in \mathbb{N}$, let $\Lambda_\varepsilon^M := \{\lambda \ge 0 : \ I(X;U_\lambda^M) \le \varepsilon\}$. The Markov chain $X \multimap Y \multimap U_\lambda \multimap U_\lambda^{M+1} \multimap U_\lambda^M$ and the data processing inequality imply that

$$I(X;U_\lambda) \ge I(X;U_\lambda^{M+1}) \ge I(X;U_\lambda^M),$$

and, in particular,

$$\varepsilon = I(X;U_{\lambda_\varepsilon}) \ge I(X;U_{\lambda_\varepsilon}^{M+1}) \ge I(X;U_{\lambda_\varepsilon}^M),$$

This in turn implies that

$$\lambda_\epsilon \in \Lambda_\varepsilon^{M+1} \subset \Lambda_\varepsilon^M, \tag{4.4}$$

and thus

$$I(Y; U_{\lambda_\varepsilon}^M) \leq g_M(\varepsilon).$$

Taking limits in both sides, we conclude from (4.3) that

$$g(\varepsilon) = I(Y; U_{\lambda_\varepsilon}) \leq \liminf_{M\to\infty} g_M(\varepsilon). \tag{4.5}$$

Observe, on the other hand, that

$$g_M(\varepsilon) = \sup_{\lambda\in\Lambda_\varepsilon^M} I(Y; U_\lambda^M) \leq \sup_{\lambda\in\Lambda_\varepsilon^M} I(Y; U_\lambda) = I(Y; U_{\lambda_{\varepsilon,\mathsf{min}}^M}), \tag{4.6}$$

where inequality follows from Markovity and $\lambda_{\varepsilon,\mathsf{min}}^M := \inf_{\Lambda_\varepsilon^M} \lambda$. Since $\lambda_\varepsilon \in \Lambda_\varepsilon^{M+1} \subset \Lambda_\varepsilon^M$, we have $\lambda_{\varepsilon,\mathsf{min}}^M \leq \lambda_{\varepsilon,\mathsf{min}}^{M+1} \leq \lambda_\varepsilon$. Thus, $(\lambda_{\varepsilon,\mathsf{min}}^M)$ is an increasing sequence in M and bounded from above and hence has a limit. Let $\lambda_{\varepsilon,\mathsf{min}} = \lim\limits_{M\to\infty} \lambda_{\varepsilon,\mathsf{min}}^M$. Clearly, we have

$$\lambda_{\varepsilon,\mathsf{min}} \leq \lambda_\varepsilon. \tag{4.7}$$

By Lemma 4.3, we know that $I(X; U_\lambda^M)$ is continuous in λ, so Λ_ε^M is closed for all $M \in \mathbb{N}$. Thus, we have $\lambda_{\varepsilon,\mathsf{min}}^M = \min_{\Lambda_\varepsilon^M} \lambda$ and in particular $\lambda_{\varepsilon,\mathsf{min}}^M \in \Lambda_\varepsilon^M$. By the inclusion $\Lambda_\varepsilon^{M+1} \subset \Lambda_\varepsilon^M$, we obtain $\lambda_{\varepsilon,\mathsf{min}}^{M+n} \in \Lambda_\varepsilon^M$ for all $n \in \mathbb{N}$. By closedness of Λ_ε^M, we have that $\lambda_{\varepsilon,\mathsf{min}} \in \Lambda_\varepsilon^M$ for all $M \in \mathbb{N}$. In particular, $I(X; U_{\lambda_{\varepsilon,\mathsf{min}}}^M) \leq \varepsilon$, for all $M \in \mathbb{N}$. We obtain from (4.3)

$$I(X; U_{\lambda_{\varepsilon,\mathsf{min}}}) \leq \varepsilon = I(X; U_{\lambda_\varepsilon}),$$

and by monotonicity of $\lambda \mapsto I(X; U_\lambda)$, we conclude that

$$\lambda_\varepsilon \leq \lambda_{\varepsilon,\mathsf{min}}. \tag{4.8}$$

Combining (4.8) with (4.7), we conclude that $\lambda_{\varepsilon,\mathsf{min}} = \lambda_\varepsilon$. Taking limits in the inequality (4.6), we have

$$\limsup_{M\to\infty} g_M(\varepsilon) \leq \limsup_{M\to\infty} I(Y; U_{\lambda^M_{\varepsilon,\mathsf{min}}}) = I(Y; U_{\lambda_{\varepsilon,\mathsf{min}}}).$$

Plugging $\lambda_{\varepsilon,\mathsf{min}} = \lambda_\varepsilon$ in above, we conclude that

$$\limsup_{M\to\infty} g_M(\varepsilon) \leq I(Y; U_{\lambda_\varepsilon}) = g(\varepsilon)$$

and therefore $\lim\limits_{M\to\infty} g_M(\varepsilon) = g(\varepsilon)$. □

As shown in this lemma, in the limit of large M, $g(\varepsilon)$ approximates $g_M(\varepsilon)$. This motivates us to focus on $g(\varepsilon)$. The following theorem summarizes some general properties of $g(\varepsilon)$.

Theorem 4.5. *The function $\varepsilon \mapsto g(\varepsilon)$ is non-negative, strictly increasing, and satisfies*

$$\lim_{\varepsilon\to 0} g(\varepsilon) = 0 \qquad \text{and} \qquad g(I(X;Y)) = \infty.$$

Proof. The nonnegativity of $g(\varepsilon)$ follows directly from the definition. According to Lemma 4.2, it is easy to verify that $\varepsilon \mapsto \lambda_\varepsilon$ is strictly decreasing. Since $\lambda \mapsto I(Y; U_\lambda)$ is strictly decreasing, we conclude that $\varepsilon \mapsto g(\varepsilon)$ is strictly increasing.

The fact that $\varepsilon \mapsto \lambda_\varepsilon$ is strictly decreasing also implies that $\lambda_\varepsilon \to \infty$ as $\varepsilon \to 0$. In particular,

$$\lim_{\varepsilon\to 0} g(\varepsilon) = \lim_{\varepsilon\to 0} I(Y;U_{\lambda_\varepsilon}) = \lim_{\lambda\to\infty} I(Y;U_\lambda) = 0.$$

By the data processing inequality, we have that $I(X;U_\lambda) \le I(X;Y)$ for all $\lambda \ge 0$, i.e., any filter satisfies the privacy constraint for $\varepsilon = I(X;Y)$. Thus, $g(I(X;Y)) \ge I(Y;Y) = \infty$. □

In Section 4.4, we will make use of the I-MMSE relationship [70] to compute g' the derivative of g. To do this, it is easier to equivalently describe the privacy filter as $Z_\gamma := \sqrt{\gamma}Y + N_{\mathsf{G}}$, instead of U_λ. Note that assuming $\sqrt{\gamma} = \frac{1}{\lambda}$, we have $I(X;Z_\gamma) = I(X;U_\lambda)$ and $I(Y;Z_\gamma) = I(Y;U_\lambda)$. With this representation, the rate-privacy function corresponds to the largest signal-to-noise ration (SNR) of the privacy-preserving additive Gaussian channel. Lemmas 4.1 and 4.2 imply together that $\gamma \mapsto I(X;Z_\gamma)$ and $\gamma \mapsto I(Y;Z_\gamma)$ are both strictly increasing and continuous and there exists a unique γ_ε (corresponding to the largest SNR which provides privacy level of ε) such that $I(X;Z_\gamma) = \varepsilon$ and $g(\varepsilon) = I(Y;Z_\gamma)$. Also, Lemma 4.2 implies that the map $\varepsilon \mapsto \gamma_\varepsilon$ is strictly increasing, and it satisfies $\gamma_0 = 0$ and $\gamma_{I(X;Y)} = \infty$. The following proposition provides upper and lower bounds for $g(\varepsilon)$ in terms of γ_ε.

Proposition 4.6. *For a pair of absolutely continuous random variables (X,Y), we have*

$$\frac{1}{2}\log\left(1+\gamma_\varepsilon 2^{-2D(Y)}\mathsf{var}(Y)\right) \le g(\varepsilon) \le \frac{1}{2}\log(1+\gamma_\varepsilon \mathsf{var}(Y)),$$

where $D(Y)$ denote the "non-Gaussianness" of Y, defined as

$$D(Y) := D(P_Y \| P_{Y_{\mathsf{G}}}), \tag{4.9}$$

with Y_{G} being the Gaussian random variable having the same mean and variance as Y.

Proof. The upper bound is a direct consequence of Lemma 4.1. The lower bound follows from the entropy power inequality [37, Theorem 17.7.3] which states that $2^{2h(Z_\gamma)} \geq \gamma 2^{2h(Y)} + 2\pi e$ and hence

$$g(\varepsilon) = I(Y; Z_{\gamma_\varepsilon}) \geq \frac{1}{2}\log\left(\gamma_\varepsilon 2^{2h(Y)} + 2\pi e\right) - \frac{1}{2}\log(2\pi e),$$

from which and the fact that $D(Y) = h(Y_{\mathsf{G}}) - h(Y)$, the lower bound immediately follows. $\square$

As opposed to the discrete case, in the continuous case g is no longer bounded and concave. A counterexample is given in the next section.

4.3 Gaussian Information

In the study of additive white Gaussian noise (AWGN) channel in information theory literatures, there exist several extremal properties of Gaussian distribution. For instance, (i) $I(Y; Y + N_{\mathsf{G}}) \leq I(Y_{\mathsf{G}}; Y_{\mathsf{G}} + N_{\mathsf{G}})$ which establishes the optimality of Gaussian input distribution for AWGN channels, (ii) $\mathsf{mmse}(Y|Y+N_{\mathsf{G}}) \leq \mathsf{mmse}(Y_{\mathsf{G}}|Y_{\mathsf{G}}+N_{\mathsf{G}})$ which establishes the fact that the Gaussian source is the hardest to estimate given its Gaussian perturbation, and (iii) $\mathsf{mmse}(Y_{\mathsf{G}}|Y_{\mathsf{G}} + N) \leq \mathsf{mmse}(Y_{\mathsf{G}}|Y_{\mathsf{G}} + N_{\mathsf{G}})$ which characterizes the worst additive noise for a Gaussian input. Here in this section, we provide another extremal property of Gaussian distribution. Before that, we first derive the rate-privacy function for Gaussian $(X_{\mathsf{G}}, Y_{\mathsf{G}})$.

Theorem 4.7. *Let $(X_{\mathsf{G}}, Y_{\mathsf{G}})$ be a pair of Gaussian random variables with zero mean and* *correlation coefficient ρ. Then, for any $\varepsilon \in [0, I(X;Y))$ we have*

$$g(\varepsilon) = \frac{1}{2}\log\left(\frac{\rho^2}{2^{-2\varepsilon} + \rho^2 - 1}\right).$$

Proof. One can always write $Y_{\mathsf{G}} = aX_{\mathsf{G}} + M_{\mathsf{G}}$ where $a^2 = \rho^2 \frac{\mathsf{var}(Y_{\mathsf{G}})}{\mathsf{var}(X_{\mathsf{G}})}$ and M_{G} is a Gaussian random variable with mean 0, variance $\sigma^2 = (1-\rho^2)\mathsf{var}(Y_{\mathsf{G}})$, and independent of X_{G}. Therefore, $Z_\gamma = a\sqrt{\gamma}X_{\mathsf{G}} + \sqrt{\gamma}M_{\mathsf{G}} + N_{\mathsf{G}}$ is also a Gaussian random variable. Then

$$I(X_{\mathsf{G}}; Z_\gamma) = \frac{1}{2}\log\left(\frac{1+\gamma\mathsf{var}(Y_{\mathsf{G}})}{1+\gamma\sigma^2}\right),$$

and hence for any $\varepsilon \in [0, I(X_{\mathsf{G}}; Y_{\mathsf{G}}))$ the equation $I(X_{\mathsf{G}}; Z_\gamma) = \varepsilon$ has the unique solution

$$\gamma_\varepsilon = \frac{1-2^{-2\varepsilon}}{\mathsf{var}(Y_{\mathsf{G}})(2^{-2\varepsilon}+\rho^2-1)}, \tag{4.10}$$

from which and the increasing property of $\gamma \mapsto I(Y; Z_\gamma)$, the result immediately follows. □

The graph of $g(\varepsilon)$ is depicted in Fig. 4.1 for jointly Gaussian X_{G} and Y_{G} with $\rho = 0.45$ and $\rho = 0.85$. It is worth noting that $g(\varepsilon)$ is related to the Gaussian rate-distortion function $R_{\mathsf{G}}(D)$ [37]. In fact, $g(\varepsilon) = R_{\mathsf{G}}(D_\varepsilon)$ for $\varepsilon \le I(X_{\mathsf{G}}; Y_{\mathsf{G}})$, where

$$D_\varepsilon = \frac{2^{-2\varepsilon} - 2^{-2I(X_{\mathsf{G}};Y_{\mathsf{G}})}}{\rho^2},$$

is the mean squared distortion incurred in reconstructing Y given the displayed data Z_γ.

According to Theorem 4.7, we conclude that the optimal privacy filter for jointly Gaussian $(X_{\mathsf{G}}, Y_{\mathsf{G}})$ is an additive Gaussian channel with SNR equal to $\frac{1-2^{-2\varepsilon}}{2^{-2\varepsilon}+\rho^2-1}$, which

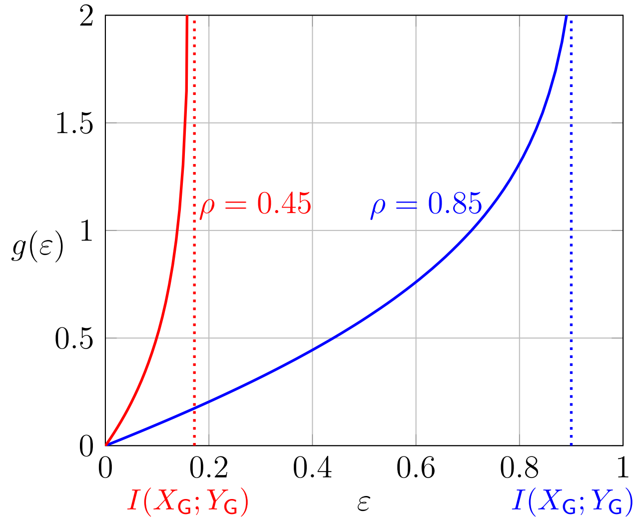


Figure 4.1: The map $\varepsilon \mapsto g_{\varepsilon}(P_{X_{\mathsf{G}}Y_{\mathsf{G}}}, \varepsilon)$ for two cases where $\rho^2 = 0.45$ and $\rho^2 = 0.85$.

shows that if perfect privacy is required, then the displayed data is independent of the observable data Y, i.e., $g(0) = 0$ as expected. Fig. 4.1 reveals that unlike to the discrete case (cf. Lemma 3.2), the mapping $\varepsilon \mapsto g(\varepsilon)$ is not necessarily concave.

Remark 4.8. We assumed that the privacy filter is a Gaussian additive channel. More generally, we could instead assume that the privacy filter adds non-Gaussian noise to the observable data and define the rate-privacy function as

$$g^{\mathsf{f}}(\varepsilon) \coloneqq \sup_{\substack{\gamma \geq 0, \\ I(X; Z_{\gamma}^{\mathsf{f}})}} I(Y; Z_{\gamma}^{\mathsf{f}}),$$

where $Z_{\gamma}^{\mathsf{f}} \coloneqq \sqrt{\gamma} Y + N_{\mathsf{f}}$ and N_{f} is a real-valued random variable having density f with $\operatorname{supp} \mathsf{f} = \mathbb{R}$ and independent of (X, Y). In this case, we use a technique similar to Oohama [114] to lower bound $g^{\mathsf{f}}(\varepsilon)$ for jointly Gaussian X_{G} and Y_{G} with correlation coefficient ρ. Since X_{G} and Y_{G} are jointly Gaussian, we can write $X_{\mathsf{G}} = aY_{\mathsf{G}} + bN_{\mathsf{G}}$ where $a^2 = \rho^2 \frac{\mathsf{var}(X_{\mathsf{G}})}{\mathsf{var}(Y_{\mathsf{G}})}$, $b^2 = (1 - \rho^2)\mathsf{var}(X_{\mathsf{G}})$, and N_{G} is standard Gaussian random variable independent of Y_{G}. Applying the conditional entropy power inequality (cf., [87, Page 22]) for a random variable Z independent of N_{G}, we obtain

$$2^{2h(X_{\mathsf{G}}|Z)} \geq 2^{2h(aY_{\mathsf{G}}|Z)} + 2^{2h(bN_{\mathsf{G}})} = a^2 2^{2h(Y_{\mathsf{G}}|Z)} + 2\pi e(1-\rho^2)\mathsf{var}(X_{\mathsf{G}}),$$

and hence

$$2^{-2I(X_{\mathsf{G}};Z)} 2^{2h(X_{\mathsf{G}})} \geq a^2 2^{2h(Y_{\mathsf{G}})} 2^{-2I(Y_{\mathsf{G}};Z)} + 2\pi e(1-\rho^2)\mathsf{var}(X_{\mathsf{G}}). \tag{4.11}$$

Assuming $Z = Z_\gamma^{\mathsf{f}}$, we obtain

$$g^{\mathsf{f}}(\varepsilon) \geq \frac{1}{2}\log\left(\frac{\rho^2}{2^{-2\varepsilon}+\rho^2-1}\right) = g(\varepsilon),$$

where the equality comes from Theorem 4.7. Therefore, for jointly Gaussian X_{G} and Y_{G}, Gaussian noise is the *worst* additive noise in the sense of the privacy-constrained information extraction.

The rate-privacy function for Gaussian Y_{G} has an interesting interpretation from an estimation-theoretic point of view. Given the private and observable data (X, Y_{G}), suppose an agent is required to *estimate* Y_{G} based on the output of the privacy filter Z_γ^{f}. We wish to know the effect of imposing the privacy constraint $I(X; Z_\gamma^{\mathsf{f}}) \leq \varepsilon$ on the estimation efficiency. The following lemma shows that $g^{\mathsf{f}}(\varepsilon)$ bounds the best performance of the predictability of Y_{G} given Z_γ^{f}.

Proposition 4.9. *For a given (X, Y_{G}), we have for any $\varepsilon \geq 0$*

$$\inf_{\substack{\gamma\geq 0,\\ I(X;Z_\gamma^{\mathsf{f}})\leq\varepsilon}} \mathsf{mmse}(Y_{\mathsf{G}}|Z_\gamma^{\mathsf{f}}) \geq \mathsf{var}(Y_{\mathsf{G}}) 2^{-2g^{\mathsf{f}}(\varepsilon)}.$$

Proof. It is well-known from the rate-distortion theory that

$$I(Y_{\mathsf{G}};\hat{Y_{\mathsf{G}}}) \geq \frac{1}{2}\log\frac{\mathsf{var}(Y_{\mathsf{G}})}{\mathbb{E}[(Y_{\mathsf{G}}-\hat{Y_{\mathsf{G}}})^2]},$$

where $\hat{Y_{\mathsf{G}}}$ is an estimation of Y_{G}. Hence, by setting $\hat{Y_{\mathsf{G}}} = \mathbb{E}[Y_{\mathsf{G}}|Z_\gamma^{\mathsf{f}}]$ and noting that $I(Y_{\mathsf{G}};\hat{Y_{\mathsf{G}}}) \leq I(Y_{\mathsf{G}};Z_\gamma^{\mathsf{f}})$, we obtain

$$\mathsf{mmse}(Y_{\mathsf{G}}|Z_\gamma^{\mathsf{f}}) \geq \mathsf{var}(Y_{\mathsf{G}})2^{-2I(Y_{\mathsf{G}};Z_\gamma^{\mathsf{f}})}, \tag{4.12}$$

from which the result follows immediately. □

Motivated by Lemma 4.9, the quantity $\eta_\varepsilon := 2^{-2g^{\mathsf{f}}(\varepsilon)}$ can be viewed as a parameter that bounds the difficulty of estimating Y_{G} when observing an additive perturbation Z_γ^{f} with privacy constraint $I(X;Z_\gamma^{\mathsf{f}}) \leq \varepsilon$. Note that $0 < \eta_\varepsilon \leq 1$, and therefore, provided that the privacy threshold is not trivial (i.e, $\varepsilon < I(X;Y)$), $\mathsf{mmse}(Y_{\mathsf{G}}|Z_\gamma^{\mathsf{f}})$ is bounded away from zero, however the bound decays exponentially at rate $g^{\mathsf{f}}(\varepsilon)$.

4.4 Approximation of $g(\varepsilon)$ in Almost Perfect Privacy Regime

We observed in the last section that perfect privacy results in a trivial utility, i.e., $g(0) = 0$. In this section, we derive a second-order approximation for $g(\varepsilon)$ for the "almost" perfect privacy regime, i.e., for sufficiently small ε . We also obtain the first and second derivatives of the mapping g.

To state the main result, we need the so-called I-MMSE relationship [70]:

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}I(Y;Z_\gamma) = \frac{1}{2}\mathsf{mmse}(Y|Z_\gamma). \tag{4.13}$$

Since X, Y and Z_γ form the Markov chain $X \multimap\!\!- Y \multimap\!\!- Z_\gamma$, it follows that $I(X; Z_\gamma) = I(Y; Z_\gamma) - I(Y; Z_\gamma|X)$ and hence two applications of (4.13) yields [70, Theorem 10]

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} I(X; Z_\gamma) = \frac{1}{2}\left[\mathsf{mmse}(Y|Z_\gamma) - \mathsf{mmse}(Y|Z_\gamma, X)\right]. \tag{4.14}$$

The next result provides the first derivative $g'(\varepsilon)$ of the function $\varepsilon \mapsto g(\varepsilon)$.

Theorem 4.10. *We have for any* $\varepsilon \in [0, I(X;Y))$

$$g'(\varepsilon) = \frac{\mathsf{mmse}(Y|Z_{\gamma_\varepsilon})}{\mathsf{mmse}(Y|Z_{\gamma_\varepsilon}) - \mathsf{mmse}(Y|Z_{\gamma_\varepsilon}, X)}.$$

Proof. Since $g(\varepsilon) = I(Y; Z_{\gamma_\varepsilon})$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} g(\varepsilon) = \left[\frac{\mathrm{d}}{\mathrm{d}\gamma} I(Y; Z_\gamma)\right]_{\gamma=\gamma_\varepsilon} \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\gamma_\varepsilon \overset{(a)}{=} \frac{1}{2}\mathsf{mmse}(Y|Z_{\gamma_\varepsilon})\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\gamma_\varepsilon, \tag{4.15}$$

where (a) follows from (4.13). In order to calculate $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\gamma_\varepsilon$, notice that $\varepsilon = I(X; Z_{\gamma_\varepsilon})$ and hence taking the derivative of both sides of this equation with respect to ε yields

$$1 = \left[\frac{\mathrm{d}}{\mathrm{d}\gamma} I(X; Z_\gamma)\right]_{\gamma=\gamma_\varepsilon} \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\gamma_\varepsilon,$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\gamma_\varepsilon = \frac{1}{\left[\frac{\mathrm{d}}{\mathrm{d}\gamma} I(X; Z_\gamma)\right]_{\gamma=\gamma_\varepsilon}} \overset{(a)}{=} \frac{2}{\mathsf{mmse}(Y|Z_{\gamma_\varepsilon}) - \mathsf{mmse}(Y|Z_{\gamma_\varepsilon}, X)}, \tag{4.16}$$

where (a) follows from (4.14). The result then follows by plugging (4.16) into (4.15). $\square$

As a simple illustration of Theorem 4.10, consider $(X_{\mathsf{G}}, Y_{\mathsf{G}})$ whose rate-privacy function was computed in Theorem 4.7. In particular, we have

$$g'(\varepsilon) = \frac{2^{-2\varepsilon}}{2^{-2\varepsilon} + \rho^2 - 1}. \tag{4.17}$$

On the other hand, since $X_{\mathsf{G}} = aY_{\mathsf{G}} + bN_{\mathsf{G}}$, where $a^2 = \rho^2 \frac{\mathsf{var}(X_{\mathsf{G}})}{\mathsf{var}(Y_{\mathsf{G}})}$ and $b^2 = (1-\rho^2)\mathsf{var}(X_{\mathsf{G}})$, one can conclude from [71, Proposition 3] that

$$\mathsf{mmse}(Y_{\mathsf{G}}|Z_\gamma, X_{\mathsf{G}}) = \mathsf{mmse}(Y_{\mathsf{G}}|Z_{\gamma+c}),$$

where $c = \frac{\rho^2}{1-\rho^2}$. Recalling that $\mathsf{mmse}(Y_{\mathsf{G}}|Z_\gamma) = \frac{\mathsf{var}(Y_{\mathsf{G}})}{1+\gamma\mathsf{var}(Y_{\mathsf{G}})}$, we obtain from (4.10) that

$$\frac{\mathsf{mmse}(Y_{\mathsf{G}}|Z_{\gamma_\varepsilon})}{\mathsf{mmse}(Y_{\mathsf{G}}|Z_{\gamma_\varepsilon}) - \mathsf{mmse}(Y_{\mathsf{G}}|Z_{\gamma_\varepsilon+c})} = \frac{1+(1-\rho^2)\gamma_\varepsilon\mathsf{var}(Y_{\mathsf{G}})}{\rho^2} = \frac{2^{-2\varepsilon}}{2^{-2\varepsilon}+\rho^2-1},$$

which equals (4.17).

In light of Theorem 4.10, we can now show that g is in fact infinitely differentiable on $(0, I(X;Y))$. This conclusion clears the way towards calculating the second derivative of g.

Corollary 4.11. *The map $\varepsilon \mapsto g(\varepsilon)$ is infinitely differentiable at any $\varepsilon \in (0, I(X;Y))$. Moreover, if $\mathbb{E}[Y^{2k+2}] < \infty$, then $\varepsilon \mapsto g(\varepsilon)$ is $(k+1)$ right-differentiable at $\varepsilon = 0$.*

Proof. It is shown in [71, Proposition 7] that $\gamma \mapsto \mathsf{mmse}(Y|Z_\gamma)$ is infinitely differentiable at any $\gamma > 0$ and k right-differentiable at $\gamma = 0$ if $\mathbb{E}[Y^{2k+2}] < \infty$. Thus the corollary follows from Theorem 4.10 noting that since $\mathbb{E}[Y^{2k+2}] < \infty$, we also have $\mathbb{E}[Y^{2k+2}|X = x] < \infty$ for almost all x (except for x in a set of zero p_X-measure). It therefore follows that $\gamma \mapsto \mathsf{mmse}(Y|Z_\gamma, X)$ is k right-differentiable at $\gamma = 0$. $\square$

It is shown in [71, Proposition 9] that for every $\gamma > 0$

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathsf{mmse}(Y|Z_\gamma, X) = -\mathbb{E}[\mathsf{var}^2(Y|Z_\gamma, X)], \tag{4.18}$$

which, together with Theorem 4.10, implies

$$\begin{aligned} g''(\varepsilon) &= \frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2} g(\varepsilon) \qquad (4.19) \\ &= \frac{2\left(\mathsf{mmse}(Y|Z_{\gamma_\varepsilon}, X)\mathbb{E}[\mathsf{var}^2(Y|Z_{\gamma_\varepsilon})] - \mathsf{mmse}(Y|Z_{\gamma_\varepsilon})\mathbb{E}[\mathsf{var}^2(Y|Z_{\gamma_\varepsilon}, X)]\right)}{\left[\mathsf{mmse}(Y|Z_{\gamma_\varepsilon}) - \mathsf{mmse}(Y|Z_{\gamma_\varepsilon}, X)\right]^3}, \end{aligned}$$

for any $\varepsilon > 0$. We notice that $g''(0)$ is guaranteed to exist (due to Corollary 4.11) if $\mathbb{E}[Y^4] < \infty$. The following corollary, which is an immediate consequence of Theorem 4.10, provides a second-order approximation for $g(\varepsilon)$ as $\varepsilon \downarrow 0$. Before we get to the corollary, we need to make a definition. Rényi [122] defined the *one-sided maximal correlation*[1] between U and V as

$$\eta_V^2(U) := \sup_g \rho^2(U, g(V)) = \frac{\mathsf{var}(\mathbb{E}[U|V])}{\mathsf{var}(U)}, \tag{4.20}$$

where ρ is the (Pearson) correlation coefficient, the supremum is taken over all measurable functions g, and the equality follows from the Cauchy-Schwarz inequality. The law of total variance implies that

$$\mathsf{mmse}(U|V) = \mathsf{var}(U)(1 - \eta_V^2(U)). \tag{4.21}$$

Corollary 4.12. *If* $\mathbb{E}[Y^4] < \infty$*, then we have as* $\varepsilon \downarrow 0$*,*

$$g(\varepsilon) = \frac{\varepsilon}{\eta_X^2(Y)} + \Delta(X, Y)\varepsilon^2 + o(\varepsilon^2),$$

[1]This name is taken from [28, Definition 7.4]. Originally, Rényi named this quantity ”correlation ratio”.

where

$$\Delta(X,Y) = \frac{1}{\eta_X^4(Y)}\left(\frac{\mathsf{var}^2(Y) - \mathbb{E}[\mathsf{var}^2(Y|X)]}{\mathsf{var}^2(Y)\eta_X^2(Y)} - 1\right). \tag{4.22}$$

Proof. According to Corollary 4.11, we can use the second-order Taylor expansion to approximate $g(\varepsilon)$ around $\varepsilon = 0$, resulting in

$$g(\varepsilon) = \varepsilon g'(0) + \frac{\varepsilon^2}{2} g''(0) + o(\varepsilon^2).$$

From Theorem 4.10 and (4.19) we have $g'(0) = \frac{1}{\eta_X^2(Y)}$ and $g''(0) = 2\Delta(X,Y)$, respectively, from which the corollary follows. $\square$

It can be shown that for jointly Gaussian X_{G} and Y_{G} with correlation coefficient ρ, $\eta_{X_{\mathsf{G}}}^2(Y_{\mathsf{G}}) = \rho^2$ and $\Delta(X_{\mathsf{G}}, Y_{\mathsf{G}}) = \frac{1-\rho^2}{\rho^4}$, and therefore Corollary 4.12 implies that for small $\varepsilon > 0$,

$$g(\varepsilon) = \frac{1}{\rho^2}\varepsilon + \frac{1-\rho^2}{\rho^4}\varepsilon^2 + o(\varepsilon^2).$$

This second-order approximation is illustrated in Fig. 4.2 for $\rho^2 = 0.45$ and $\rho^2 = 0.85$.

Polyanskiy and Wu [118] have recently generalized the strong data processing inequality of Anantharam et al. [11] for the case of continuous random variables X and Y with joint distribution P. Their result states that

$$\sup_{\substack{X -\!\circ\!- Y -\!\circ\!- U,\\ 0<I(U;Y)<\infty}} \frac{I(X;U)}{I(Y;U)} = S^*(Y,X), \tag{4.23}$$

where

$$S^*(Y,X) := \sup_{\substack{\mathsf{q},\\ 0<D(\mathsf{q}\|\mathsf{q}_Y)<\infty}} \frac{D(\mathsf{p}\|\mathsf{p}_X)}{D(\mathsf{q}\|\mathsf{q}_Y)},$$

where p_X and q_Y are the marginals of P and $\mathsf{p}(\cdot) = \int P_{X|Y}(\cdot|y)\mathsf{q}(\mathrm{d}y)$. In addition, it is

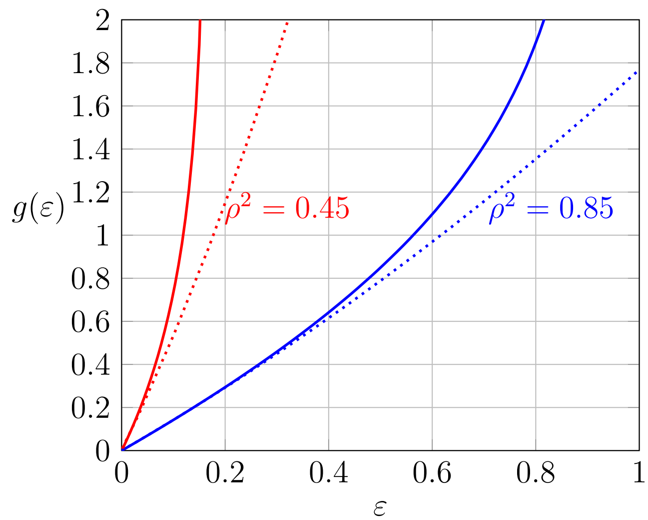


Figure 4.2: The second-order approximation for $g(\varepsilon)$ when X_{G} and Y_{G} are jointly Gaussian random variables with correlation coefficient $\rho^2 = 0.45$ or $\rho^2 = 0.85$.

shown in [118] that the supremum in (4.23) is achieved by a binary U. Replacing U with Z_γ, we can conclude from (4.23) that $\frac{I(X;Z_\gamma)}{I(Y;Z_\gamma)} \leq S^*(Y,X)$, for any $\gamma > 0$. Letting $\gamma = \gamma_\varepsilon$, the above yields

$$g(\varepsilon) \geq \frac{\varepsilon}{S^*(Y,X)}. \tag{4.24}$$

Clearly, this bound may be expected to be tight only for small $\varepsilon > 0$ since $g(\varepsilon) \to \infty$ as $\varepsilon \to I(X;Y)$, as shown in Proposition 4.6. Note that Theorem 4.10 implies that $\lim_{\varepsilon \downarrow 0} \frac{g(\varepsilon)}{\varepsilon} = \frac{1}{\eta_X^2(Y)}$. On the other hand, it can be easily shown that $\eta_X^2(Y) \leq S^*(Y,X)$, with equality when X and Y are jointly Gaussian and hence the inequality (4.24) becomes tight for small ε and jointly Gaussian X and Y.

The bound in (4.24) would be significantly improved if we could show that $g(P_{XY},\varepsilon) \geq g(P_{X_{\mathsf{G}}Y_{\mathsf{G}}},\varepsilon)$, where X_{G} and Y_{G} are jointly Gaussian having the same means, variances, and correlation coefficient as (X, Y). This is because in that case we could write

$$g(P_{XY}, \varepsilon) \geq g(P_{X_\mathsf{G}Y_\mathsf{G}}, \varepsilon) \geq \frac{\varepsilon}{\eta^2_{X_\mathsf{G}}(Y_\mathsf{G})} = \frac{\varepsilon}{\rho^2(X_\mathsf{G}, Y_\mathsf{G})} = \frac{\varepsilon}{\rho^2(X, Y)} \geq \frac{\varepsilon}{\eta^2_X(Y)}. \tag{4.25}$$

However, as shown in the following theorem, the inequality $g(P_{XY}, \varepsilon) \geq g(P_{X_\mathsf{G}Y_\mathsf{G}}, \varepsilon)$ does not in general hold.

Theorem 4.13. *For any continuous random variable X correlated with Gaussian Y_G, we have*

$$g(P_{X_\mathsf{G}Y_\mathsf{G}}, \varepsilon) \geq g(P_{XY_\mathsf{G}}, \varepsilon),$$

where $(X_\mathsf{G}, Y_\mathsf{G})$ is a pair of Gaussian random variables having the same mean, variance and correlation coefficient as (X, Y_G).

Proof. For any pair of random variables (U, V) with $I(U; V) < \infty$, let $P_{V|U}(\cdot|u)$ be the conditional density of V given $U = u$. Let $(U_\mathsf{G}, V_\mathsf{G})$ be a pair of Gaussian random variables having the same means, variances and correlation coefficient as (U, V), and $P_{V_\mathsf{G}|U_\mathsf{G}}(\cdot|u)$ the conditional density of V_G given $U_\mathsf{G} = u$. Similar to $D(V)$ the non-Gaussianness of V, defined in (4.9), we can define $D(V|U)$ the conditional non-Gaussianness of V given U as

$$D(V|U) := \int D\Big(P_{V|U}(\cdot|u) \| P_{V_\mathsf{G}|U_\mathsf{G}}(\cdot|u)\Big) \mathrm{d}P_U(u) = \mathbb{E}_{UV}\left[\log \frac{P_{V|U}(V|U)}{P_{V_\mathsf{G}|U_\mathsf{G}}(V|U)}\right].$$

It is straightforward to show that

$$I(U; V) = I(U_\mathsf{G}; V_\mathsf{G}) + D(V|U) - D(V). \tag{4.26}$$

Replacing U and V with X and Z_γ, respectively, in the decomposition (4.26) and noticing that $D(Z_\gamma) = 0$, we obtain

$$I(X; Z_\gamma) = I(X_{\mathsf{G}}; Z_\gamma) + D(Z_\gamma|X).$$

Since $D(Z_\gamma|X) \geq 0$, we have $I(X; Z_\gamma) \geq I(X_{\mathsf{G}}; Z_\gamma)$. Therefore, the condition $I(X; Z_\gamma) \leq \varepsilon$ implies that $I(X_{\mathsf{G}}; Z_\gamma) \leq \varepsilon$, from which the result follows. $\square$

In light of this theorem, it is therefore possible to have $g(\varepsilon) < \frac{\varepsilon}{\eta_X^2(Y)}$ for some $0 < \varepsilon < I(X;Y)$. To construct an example, it suffices to construct P for which $\varepsilon \mapsto g(\varepsilon)$ has negative second-derivative at zero and hence its graph lies below the tangent line $\frac{\varepsilon}{\eta_X^2(Y)}$ for some $\varepsilon > 0$.

Example 4.14. Let $Y_{\mathsf{G}} \sim \mathcal{N}(0, 1)$ and $X = Y_{\mathsf{G}} \cdot 1_{\{Y_{\mathsf{G}} \in [-1,1]\}}$. Then it can be readily shown that $\mathbb{E}[\mathsf{var}(Y_{\mathsf{G}}|X)] < \mathbb{E}[\mathsf{var}^2(Y_{\mathsf{G}}|X)]$, which implies that $\mathit{\Delta}(X, Y_{\mathsf{G}}) < 0$. Hence, since $g''(0) = 2\mathit{\Delta}(X, Y)$, we have that $g''(0) < 0$. This observation is illustrated in Fig. 4.3.

The above example also shows that $\varepsilon \mapsto \frac{g(\varepsilon)}{\varepsilon}$ cannot be increasing, because if it were, it would have implied $\frac{g(\varepsilon)}{\varepsilon} \geq \lim_{\varepsilon \to 0} \frac{g(\varepsilon)}{\varepsilon} = \frac{1}{\eta_X^2(Y)}$. However, it can be shown that $g(\varepsilon)$ lies always above the line $\frac{\varepsilon}{\eta_X^2(Y)}$ if P has certain structures. In the next theorem, we assume that Y is a noisy version of X through an AWGN channel.

Theorem 4.15. *For a given absolutely continuous X with variance $\mathsf{var}(X)$, and $Y = aX + \sigma M_{\mathsf{G}}$, where M_{G} is a standard Gaussian random variable independent of X, we have:*

1. *The map $\varepsilon \mapsto g(\varepsilon)$ has positive second-derivative at $\varepsilon = 0$.*

2. *For any $a > 0$ and $\varepsilon \in [0, I(X;Y))$, we have*

$$g(\varepsilon) \geq \frac{\varepsilon}{\eta_X^2(Y)}. \tag{4.27}$$

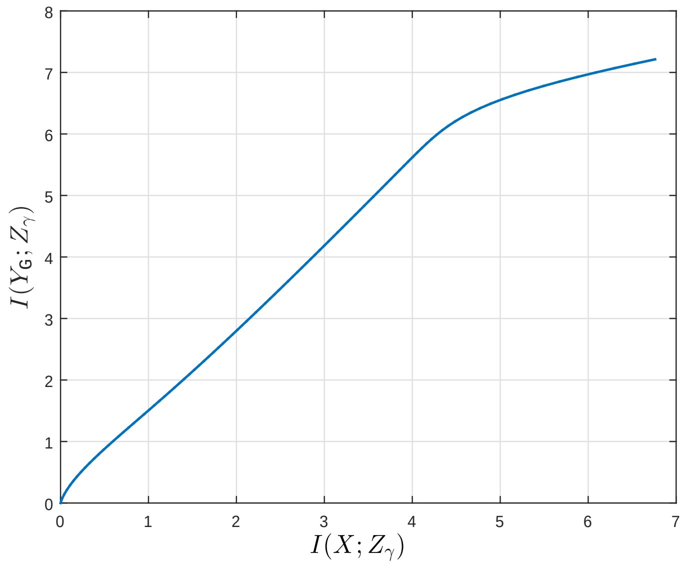


Figure 4.3: The rate-privacy function for $Y_{\mathsf{G}} \sim \mathcal{N}(0,1)$ and $X = Y_{\mathsf{G}} \cdot 1_{\{Y_{\mathsf{G}} \in [-1,1]\}}$. The map $\varepsilon \mapsto g(\varepsilon)$ has negative second-derivative at zero. Note that here $I(X; Y_{\mathsf{G}}) = \infty$ and hence ε is unbounded.

Furthermore, we have

$$\inf_{\gamma \geq 0} \frac{\mathsf{mmse}(Y|Z_\gamma, X)}{\mathsf{mmse}(Y|Z_\gamma)} = 1 - \eta_X^2(Y), \tag{4.28}$$

and

$$\sup_{\gamma > 0} \frac{I(X; Z_\gamma)}{I(Y; Z_\gamma)} = \eta_X^2(Y). \tag{4.29}$$

Proof. To see the first part, notice that $\mathsf{var}(Y) = a^2\mathsf{var}(X) + \sigma^2$, $\mathbb{E}[\mathsf{var}^2(Y|X)] = \sigma^4$, and $\eta_X^2(Y) = \frac{a^2\mathsf{var}(X)}{a^2\mathsf{var}(X)+\sigma^2}$, from which we can show that $\mathsf{var}^2(Y) - \mathbb{E}[\mathsf{var}^2(Y|X)] \geq \mathsf{var}^2(Y)\eta_X^2(Y)$, and consequently $\Delta(X, Y) \geq 0$.

To prove the second part, note that for any $\gamma > 0$, we have

$$\begin{aligned}
\mathsf{mmse}(Y|Z_\gamma) &= \mathsf{mmse}(aX + \sigma M_{\mathsf{G}}|a\sqrt{\gamma}X + \sqrt{\gamma}\sigma M_{\mathsf{G}} + N_{\mathsf{G}}) \\
&\overset{(a)}{=} \frac{1}{\gamma}\mathsf{mmse}\left(N_{\mathsf{G}}|a\sqrt{\gamma}X + \sqrt{\gamma}\sigma M_{\mathsf{G}} + N_{\mathsf{G}}\right) \\
&\overset{(b)}{\leq} \frac{a^2\mathsf{var}(X) + \sigma^2}{1 + \gamma(a^2\mathsf{var}(X) + \sigma^2)} < \frac{a^2\mathsf{var}(X) + \sigma^2}{1 + \gamma\sigma^2} \\
&\overset{(c)}{=} \frac{1}{\gamma}\left(\frac{a^2\mathsf{var}(X) + \sigma^2}{\sigma^2}\right)\mathsf{mmse}\left(N_{\mathsf{G}}|\sqrt{\gamma}\sigma M_{\mathsf{G}} + N_{\mathsf{G}}\right) \\
&\overset{(d)}{=} \left(\frac{a^2\mathsf{var}(X) + \sigma^2}{\sigma^2}\right)\mathsf{mmse}(Y|Z_\gamma, X), \qquad (4.30)
\end{aligned}$$

where (a) follows from the fact that $\mathsf{mmse}(U|\alpha U + V) = \frac{1}{\alpha^2}\mathsf{mmse}(V|\alpha U + V)$ for $\alpha \neq 0$, and (b) and (c) follows from [150, Theorem 12] which states that $\mathsf{mmse}(U|U + V_{\mathsf{G}}) \leq \mathsf{mmse}(U_{\mathsf{G}}|U_{\mathsf{G}} + V_{\mathsf{G}}) = \frac{\mathsf{var}(U)\mathsf{var}(V_{\mathsf{G}})}{\mathsf{var}(U)+\mathsf{var}(V_{\mathsf{G}})}$. Finally, (d) follows from the following chain of equalities

$$\begin{aligned}
\mathsf{mmse}(Y|Z_\gamma, X) &= \mathsf{mmse}(aX + \sigma M_{\mathsf{G}}|a\sqrt{\gamma}X + \sqrt{\gamma}\sigma M_{\mathsf{G}} + N_{\mathsf{G}}, X) \\
&= \mathsf{mmse}(\sigma M_{\mathsf{G}}|\sqrt{\gamma}\sigma M_{\mathsf{G}} + N_{\mathsf{G}}, X) \overset{(e)}{=} \mathsf{mmse}(\sigma M_{\mathsf{G}}|\sqrt{\gamma}\sigma M_{\mathsf{G}} + N_{\mathsf{G}}) \\
&= \frac{1}{\gamma}\mathsf{mmse}(N_{\mathsf{G}}|\sqrt{\gamma}\sigma M_{\mathsf{G}} + N_{\mathsf{G}}),
\end{aligned}$$

where (e) holds since X and M_{G} are independent. We can therefore write

$$g'(\varepsilon) = \frac{\mathsf{mmse}(Y|Z_{\gamma_\varepsilon})}{\mathsf{mmse}(Y|Z_{\gamma_\varepsilon}) - \mathsf{mmse}(Y|Z_{\gamma_\varepsilon}, X)} \overset{(a)}{\geq} \frac{a^2\mathsf{var}(X) + \sigma^2}{a^2\mathsf{var}(X)} \overset{(b)}{=} \frac{1}{\eta_X^2(Y)} = g'(0), \tag{4.31}$$

where (a) is due to (4.30) and (b) holds since $\mathsf{var}(Y) = a^2\mathsf{var}(X) + \sigma^2$ and $\mathsf{var}(\mathbb{E}[Y|X]) = a^2\mathsf{var}(X)$. The identity $g(\varepsilon) = \int_0^\varepsilon g'(t)\mathrm{d}t$ and inequality (4.31) together imply that $g(\varepsilon) \geq \frac{\varepsilon}{\eta_X^2(Y)}$.

It is straightforward to show that (4.31) yields (4.28). Using the integral representation of mutual information in (4.13) and (4.14), we can write for any $\gamma \geq 0$

$$\begin{aligned} I(X; Z_\gamma) &= \frac{1}{2}\int_0^\gamma [\mathsf{mmse}(Y|Z_t) - \mathsf{mmse}(Y|Z_t, X)]\, \mathrm{d}t \\ &\leq \frac{\eta_X^2(Y)}{2}\int_0^\gamma \mathsf{mmse}(Y|Z_t)\mathrm{d}t = \eta_X^2(Y) I(Y; Z_\gamma), \end{aligned} \tag{4.32}$$

where the inequality is due to (4.28). The equality (4.29) then follows from (4.32) by noticing that $\frac{I(X;Z_\gamma)}{I(Y;Z_\gamma)} \to \eta_X^2(Y)$ as $\gamma \to 0$. $\square$

It should be noted that both MMSE and mutual information satisfy the data processing inequality, see, [150] and [11], that is, $\mathsf{mmse}(U|V) \leq \mathsf{mmse}(U|W)$, and $I(U; W) \leq I(U; V)$ for $U \multimap V \multimap W$. Therefore, (4.28) can be viewed as a strong version of the data processing inequality for MMSE for the trivial Markov chain $Y \multimap (Z_\gamma, X) \multimap Z_\gamma$. Also, (4.29) can be viewed as a strong data processing inequality for the mutual information for the Markov chain $X \multimap Y \multimap Z_\gamma$.

Remark 4.16. As mentioned earlier, it is immediate from [4, Theorem 3] that $\eta_X^2(Y) \leq S^*(Y, X)$ where the equality occurs if X and Y are jointly Gaussian (see [113, Theorem 3]). Equality (4.29) provides an interesting implication of the equality $\eta_X^2(Y) = S^*(Y, X)$. Combining (4.23) with (4.29), we conclude that for $Y = aX + \sigma M_{\mathsf{G}}$ with $a, \sigma > 0$, AWGN channel is an optimal channel $P_{U|Y}$, in the sense of (4.23), if and only if $S^*(Y, X) = \eta_X^2(Y)$. For instance, if X is Gaussian, then the ratio $\frac{I(X;U)}{I(Y;U)}$ is maximized over $X \multimap Y \multimap U$ when the channel from Y to U is an AWGN channel with SNR approaching zero.

Chapter 5

Information Extraction Under an Estimation-Theoretic Privacy Constraint

5.1 Overview

In the last two chapters, we proposed to measure the privacy leakage in terms of the mutual information. By imposing constraint on the mutual information between private data and the displayed data, we make sure that only limited bits of private information are revealed during the process of transferring Y. Despite the dependence dilution setting studied in Section 3.7, mutual information does not lead to an arguably operational privacy interpretation and thus cannot serve as an appropriate privacy leakage function [54, 47]. For the discrete case, one may invoke Fano's inequality to interpret the requirement $I(X;Z) \leq \varepsilon$. That is, for any estimator $\hat{X} : \mathcal{Z} \to \mathcal{X}$ we have $\Pr(\hat{X}(Z) \neq X) \geq \frac{H(X)-1-\varepsilon}{\log|\mathcal{X}|}$, and consequently the probability that an adversary, observing Z, can correctly guess X is lower-bounded. Unfortunately, Fano's inequality proves to be loose in most practical cases. For example, if $|\mathcal{X}| = 2$, then the above lower bound is negative for any $\varepsilon \geq 0$. In this chapter, we provide a better motivated measure of privacy for discrete random variables, study the corresponding privacy-constrained information extraction $\hat{g}(\varepsilon)$ and obtain tight bounds on

$\hat{g}(\varepsilon)$ in terms of $g(\varepsilon)$.

5.1.1 Main Contributions

The main contributions of this chapter are as follows:

- After justifying the use of the maximal correlation ρ_m as an operational privacy measure in the discrete case, we introduce a variant rate-privacy function as an operationally better-justified privacy-utility tradeoff. Specifically, we define the function $\hat{g}(\varepsilon)$ as the maximum $I(Y;Z)$ over all $P_{Z|Y}$ satisfying $X \multimap Y \multimap Z$ and $\rho_m^2(X,Z) \leq \varepsilon$. We show that if Y is binary, then it suffices to consider ternary-valued Z.

- Some of the functional properties of $\hat{g}$ are derived. Specifically, we show that $\hat{g}$ shares many properties with g: it is strictly increasing, concave and lower-bounded by the erasure mechanism. We also derive bounds on $\hat{g}$ in terms of g. These bounds, in particular, show that $\hat{g}(\varepsilon) = g(\varepsilon)$ for any ε in the domain when $P_{Y|X}$ is BEC and $X \sim \mathsf{Bernoulli}(\frac{1}{2})$.

- Finally, we study in detail the characterization of linear behavior of $\hat{g}$ when $P_{X|Y}$ is BISO and show that $\hat{g}$ is linear only if Y is uniform.

5.2 Maximal Correlation: Definition and Properties

Given the collection $\mathcal{C}$ of all pairs of random variables $(U,V) \in \mathcal{U} \times \mathcal{V}$ where $\mathcal{U}$ and $\mathcal{V}$ are general alphabets, a mapping $T : \mathcal{C} \to [0,1]$ defines a *measure of correlation* [61] if $T(U,V) = 0$ if and only if U and V are independent (in short, $U \perp\!\!\!\perp V$) and $T(U,V)$ attains its maximum value if $g(U) = f(V)$ almost surely for some measurable real-valued functions f and g. There are many different examples of measures of correlation including the Hirschfeld-Gebelein-Rényi maximal correlation [77, 61, 122], the information measure [102], mutual information and f-divergence [38], MMSE [70], and χ^2-divergence. In his seminal paper, Rényi postulated seven properties of an "appropriate" measure of correlation and showed that the maximal correlation satisfies all the properties.

Definition 5.1 ([77, 61, 122])**.** *Given random variables U and V defined over general (discrete or continuous) alphabets $\mathcal{U}$ and $\mathcal{V}$, respectively, the maximal correlation $\rho_m(U,V)$ is defined as*

$$\rho_m(U,V) \coloneqq \sup_{(f,g)\in\mathcal{S}} \mathbb{E}[f(U)g(V)],$$

where $\mathcal{S} \coloneqq \{(f,g) :\ \mathbb{E}[f(U)] = \mathbb{E}[g(V)] = 0, \mathsf{var}(f(U)) = \mathsf{var}(g(V)) = 1\}$. If $\mathcal{S}$ is empty (which happens precisely when at least one of U and V is constant almost surely) then one defines $\rho_m(U,V)$ to be 0.

Applying the Cauchy-Schwarz inequality, Rényi [122] derived an equivalent "one-function" characterization of the maximal correlation as follows:

$$\rho_m^2(U,V) = \sup_{f\in\mathcal{S}_U} \mathbb{E}\left[\mathbb{E}^2[f(U)|V]\right], \tag{5.1}$$

where $\mathcal{S}_U$ is the set of all measurable real-valued functions f on $\mathcal{U}$ such that $\mathbb{E}f(U) = 0$ and $\mathsf{var}(f(U)) = 1$.

It is worth mentioning that maximal correlation has a discontinuous property. To see this discontinuous property, let $\delta \in (0, \frac{1}{2}]$ and (X^δ, Y^δ) be defined as follows: with probability δ, one samples X^δ and Y^δ independently according to uniform distribution over $[0,\delta]$ and with probability $\bar{\delta}$, one samples X^δ and Y^δ independently according to uniform distribution over $[\delta, 1]$. It is straightforward to show that $\rho_m^2(X^\delta, Y^\delta) = 1$ for all $\delta > 0$ while

(X^δ, Y^δ) converge in distribution to (X, Y) where $X \perp\!\!\!\perp Y$. Kimeldorf and Sampson [89] constructed another example to show the discontinuity of maximal correlation. In what follows, we provide some functional properties of maximal correlation including the lower semi-continuity.

Proposition 5.2. *Let random variables U and V be defined over general alphabets $\mathcal{U}$ and $\mathcal{V}$ with joint distribution P_{UV} and marginals P_U and P_V. Then*

1. $\rho_m(U, V) \geq 0$ *with equality if and only if* $U \perp\!\!\!\perp V$.

2. $\rho_m(U, V) \leq 1$ *with equality if and only if there exists a pair of measurable functions* $(f, g) \in \mathcal{S}$ *such that* $\Pr(f(U) = g(V)) = 1$.

3. *If U and V are jointly Gaussian with correlation coefficient ρ, then $\rho_m^2(U, V) = \rho^2$.*

4. *$\rho_m(U, V)$ is equal to the second largest singular value of the operator*[1] $T : \mathcal{L}^2(P_U) \to \mathcal{L}^2(P_V)$ *given by* $(Tf)(v) = \mathbb{E}[f(U)|V = v]$. *In particular, if $\mathcal{U}$ and $\mathcal{V}$ are finite alphabets, then $\rho_m(U, V)$ is equal to the second largest singular value of matrix*
$$B = \left[\frac{P_{UV}(u, v)}{\sqrt{P_U(u) P_V(v)}}\right]_{u \in \mathsf{supp}(P_U), v \in \mathsf{supp}(P_V)}.$$

5. *ρ_m satisfies the data processing inequality, i,e., given random variables R and S which form Markov chain $R \multimap U \multimap V \multimap S$, we have $\rho_m(R, S) \leq \rho_m(U, V)$.*

6. *(Tensorization property) Let (U_i, V_i) for $i \in [n]$ be n independent pairs of random variables with joint distribution $P_{U_iV_i}$, $i \in [n]$. Then* $\rho_m(U^n, V^n) = \max_{1 \leq i \leq n} \rho_m(U_i, V_i)$.

[1] That is, the the square root of the second largest number in the point spectrum of the operator TT^* where T^* is the adjoint operator of T.

7. *If N is an infinitely divisible random variable independent of X and Y, then $\lambda \mapsto \rho_m(X; Y + \lambda N)$ is non-increasing right continuous function on $[0, \infty)$.*

8. *We have*

$$\max\{\eta_U^2(V), \eta_V^2(U)\} \leq \rho_m^2(U, V) \leq \chi^2(P_{UV} \| P_U P_V),$$

where the one-sided maximal correlation $\eta_U(V)$ was defined in (4.20) *and the χ^2-divergence between two probability distributions P and Q is defined as*

$$\chi^2(P\|Q) := \int \left(\frac{dP}{dQ}\right)^2 dQ - 1. \tag{5.2}$$

9. *The map $P_{UV} \mapsto \rho_m(U, V)$ is weakly lower semi-continuous.*

Proof. Parts 1, 2 and 4 were proved by Rényi [122]. Two complicated proofs for part 3 were given in [61] and [93] using Hermite-Chebyshev polynomial decomposition. More recently, a rather easier proof was given in [115]. An interesting (yet indirect) proof can also be obtained by combining [4, Theorem 3.b] and [113, Theorem 3]. Different proofs for part 5 were provided in [84], [32] and [117]. A lengthy proof for part 6 was constructed in [148] and an easier proof was given in [91]. Part 7 was proved in [27]. Part 8 can be proved by noticing that $\chi^2(P_{UV} \| P_U P_V)$ is equal to the sum of squares of the singular values of operator T minus 1 (the largest one) [148] (see also the proof of Lemma 5.5) while ρ_m is equal to the second largest one. To prove part 9, we define $\delta(U, V)$ as

$$\begin{aligned}
\delta(U; V) &:= \inf\{\mathbb{E}[(f(U) - g(V))^2] : \ (f(U), g(V)) \in \mathcal{S}\} \\
&= \inf\{\mathbb{E}[(f(U) - g(V))^2] : \ f \in \mathcal{C}_b(\mathcal{U}), g \in \mathcal{C}_b(\mathcal{V}), \mathbb{E}[f^2(X)] = \mathbb{E}[g^2(Y)] = 1\},
\end{aligned}$$

where $\mathcal{C}_b(\mathcal{U})$ and $\mathcal{C}_b(\mathcal{V})$ denote the collection of all real-valued continuous bounded functions over $\mathcal{U}$ and $\mathcal{V}$, respectively, and the last equality is due to the denseness of $\mathcal{C}_b$ in $\mathcal{L}^2$. It is clear that $\delta(U,V) = 2(1-\rho_m(U,V))$. It therefore suffices to prove that $\delta(P_{UV})$ is weakly upper semi-continuous. For fixed $f \in \mathcal{C}_b(\mathcal{U})$ and $g \in \mathcal{C}_b(\mathcal{V})$,

$$\mathbb{E}[(f(U)-g(V))^2] = \int (f(u)-g(v))^2 P_{UV}(\mathrm{d}u, \mathrm{d}v),$$

is weakly continuous in P_{UV}. The result then follows from the fact that pointwise infimum of a family of weakly continuous functions is weakly upper semi-continuous. $\square$

We next show that the data processing inequality shown in part 5 can be strengthened. The following lemma proves the *strong* data processing inequality for the maximal correlation from which the typical data processing inequality immediately follows.

Lemma 5.3. *For random variables U and V with a joint distribution P_{UV}, we have*

$$\sup_{\substack{U \multimap\!- V \multimap\!- S \\ \rho_m(V,S)\neq 0}} \frac{\rho_m(U,S)}{\rho_m(V,S)} = \rho_m(U,V).$$

Proof. Fix a joint distribution P_{UVS} satisfying $U \multimap\!- V \multimap\!- S$. For measurable functions $f \in \mathcal{S}_U$ and $g \in \mathcal{S}_S$, we have

$$\begin{aligned} \mathbb{E}^2[f(U)g(S)] &= \mathbb{E}^2[\mathbb{E}[f(U)g(S)|V]] = \mathbb{E}^2\left[\mathbb{E}[f(U)|V]\mathbb{E}[g(S)|V]\right] \\ &\leq \mathbb{E}[\mathbb{E}^2[f(U)|V]]\mathbb{E}[\mathbb{E}^2[g(S)|V]], \end{aligned} \tag{5.3}$$

where the inequality follows from the Cauchy-Schwarz inequality. Taking supremum from both sides of (5.3) over (f,g) and recalling (5.1), we obtain $\rho_m^2(U,S) \leq \rho_m^2(U,V)\rho_m^2(V,S)$.

In the following we show that this bound holds with equality for the special case of $U \multimap V \multimap \hat{U}$, where $P_{\hat{U}|V}$ is the backward channel associated with $P_{V|U}$. To this end, first note that the above implies that $\rho_m(U, \hat{U}) \leq \rho_m(U, V)\rho_m(\hat{U}, V)$. Since $P_{UV} = P_{\hat{U}V}$, it follows that $\rho_m(U, V) = \rho_m(\hat{U}, V)$ and hence in this case $\rho_m(U, \hat{U}) \leq \rho_m^2(U, V)$. On the other hand, we have

$$\mathbb{E}[[\mathbb{E}[f(U)|V]]^2] = \mathbb{E}[\mathbb{E}[f(U)|V]\mathbb{E}[f(\hat{U})|V]] = \mathbb{E}[\mathbb{E}[f(U)f(\hat{U})|V]] = \mathbb{E}[f(U)f(\hat{U})],$$

which together with (5.1) implies that

$$\rho_m^2(U, V) = \sup_{f \in \mathcal{S}_U} \mathbb{E}[f(U)f(\hat{U})] \leq \rho_m(U; \hat{U}).$$

Thus, $\rho_m^2(U, V) = \rho_m(U, \hat{U})$ which completes the proof. □

5.3 Maximal Correlation as a Privacy Measure

It is proposed in [104] and [95] to consider $\rho_m^2(X, Z) \leq \varepsilon$ as a privacy guarantee without giving an operational justification. However, an interesting interpretation for this constraint was given in [32]. Before giving this result, we need the following definition.

Definition 5.4. *Given discrete random variables U and V taking values respectively in $\mathcal{U}$ and $\mathcal{V}$ with joint distribution $P_{UV} = P_V \times P_{U|V}$, the Bayes map $\Phi : \mathcal{V} \to \mathcal{U}$ is given by $\Phi(v) := \arg\max_{u \in \mathcal{U}} P_{U|V}(u|v)$. Furthermore, $\mathsf{P_c}(U|V)$ the probability of correctly guessing U given V (also known as the Bayes risk [37]) is defined as*

$$\mathsf{P_c}(U|V) \;:=\; \sup_{P_{\hat{U}|V}:\, U \multimap V \multimap \hat{U}} \Pr(U = \hat{U}) = \sum_{v \in \mathcal{V}} P_V(v) \Pr(U = \Phi(v)|V = v)$$

$$= \sum_{v\in\mathcal{V}} P_V(v) \max_{u\in\mathcal{U}} P_{U|V}(u|v) = \sum_{v\in\mathcal{V}} \max_{u\in\mathcal{U}} P_{UV}(u,v).$$

When side information is not available, i.e., $V=\emptyset$, then $\mathsf{P_c}(U)$, the probability of correctly guessing U, is given by

$$\mathsf{P_c}(U) \coloneqq \max_{u\in\mathcal{U}} P_U(u).$$

Given a pair of discrete random variables (X,Z) with joint distribution P_{XZ}, it is clearly easier to correctly guess X with side information Z than without it, i.e., $\mathsf{P_c}(X|Z) \geq \mathsf{P_c}(X)$. However, it has been recently shown [28, Theorem 5.6] (see [32, Corollary 3] for a weaker result) that $\mathsf{P_c}(f(X)|Z)$ cannot be much larger than $\mathsf{P_c}(f(X))$, for any deterministic function f, if the maximal correlation between X and Z is small:

$$\mathsf{P_c}(f(X)) \leq \mathsf{P_c}(f(X)|Z) \leq \mathsf{P_c}(f(X)) + \rho_m(X,Z)\sqrt{S_2(\mathsf{p}_f)},$$

where $S_2(\mathsf{p}_f) \coloneqq 1-\sum_i P^2_{f(X)}(i)$ and $P_{f(X)}$ is the distribution of $f(X)$ induced by p_X, the distribution of X. Consequently, $\rho_m(X,Z)\leq\varepsilon$ for small $\varepsilon\geq 0$ implies that $\mathsf{P_c}(f(X)|Z)$ is close to $\mathsf{P_c}(f(X))$ and hence the observation Z cannot be used to efficiently guess any deterministic function of X. This justifies to use $\rho_m(X,Z)$ as a privacy measure.

Similar to Chapter 3, we define the rate-privacy function $\hat{g}(\varepsilon)$ for a pair of given discrete random variables (X,Y) with joint distribution P and marginals p_X and q_Y over *finite* alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, as

$$\hat{g}(\varepsilon) \coloneqq \sup_{P_{Z|Y}\in\hat{\mathcal{D}}_\varepsilon(\mathsf{P})} I(Y;Z),$$

where $\varepsilon \geq 0$ and

$$\hat{\mathcal{D}}_\varepsilon(\mathsf{P}) \coloneqq \{P_{Z|Y} : X \multimap Y \multimap Z, \rho_m^2(X,Z) \leq \varepsilon\}.$$

In words, $\hat{g}(\varepsilon)$ quantifies the maximum number of bits of information one can extract from Y such that X cannot be efficiently guessed from the extracted information. Again, we refer to $\hat{g}(\varepsilon)$ as the privacy-constrained information extraction function, where here the privacy is guaranteed by $\rho_m^2(X,Z) \leq \varepsilon$.

Setting $\varepsilon = 0$ corresponds to the case where X and Z are required to be statistically independent, i.e., no information leakage about X is allowed. This case is called *perfect privacy.* Since the independence of X and Z is equivalent to $I(X;Z) = \rho_m(X;Z) = 0$, we have $\hat{g}(0) = g(0)$. This in turn implies that weak independence, defined in Definition 3.9, is still a necessary and sufficient condition for $\hat{g}(0) > 0$. However, for $\varepsilon > 0$, both $g(\varepsilon) \leq \hat{g}(\varepsilon)$ and $g(\varepsilon) \geq \hat{g}(\varepsilon)$ may occur in general. We also note that it is not clear how to bound the cardinality of Z in the definition of $\hat{g}(\varepsilon)$. However, we will show that if $|\mathcal{Y}| = 2$, then Z with $|\mathcal{Z}| = 3$ is sufficient to achieve $\hat{g}(\varepsilon)$.

According to Lemma 5.3, maximal correlation satisfies the data processing inequality and thus $\rho_m^2(X,Z) \leq \rho_m^2(X,Y)$. Therefore, for any $\varepsilon \geq \rho_m^2(X,Y)$, setting $Z = Y$ results in $\hat{g}(\varepsilon) = H(Y)$. We can hence restrict ε to the interval $\varepsilon \in [0, \rho_m^2(X,Y))$.

The following proposition provides a bound for $\hat{g}(\varepsilon)$ in terms of $g(\varepsilon)$ for any $\varepsilon \geq 0$.

Lemma 5.5. *If $|\mathcal{X}| = M$, then*

$$\hat{g}(\varepsilon) \leq g((M-1)\varepsilon).$$

Proof. First we notice that[2]

$$
\begin{aligned}
I(X;Z) &= \mathbb{E}\left[\log \frac{P_{XZ}(X,Z)}{\mathsf{p}_X(X)P_Z(Z)}\right] \overset{(a)}{\leq} \log \mathbb{E}\left[\frac{P_{XZ}(X,Z)}{\mathsf{p}_X(X)P_Z(Z)}\right] \\
&\overset{(b)}{=} \log\left(1+\chi^2(P_{XZ}\|\mathsf{p}_X P_Z)\right) \leq \chi^2(P_{XZ}\|\mathsf{p}_X P_Z),
\end{aligned}
\tag{5.4}
$$

where (a) follows from Jensen's inequality and (b) holds due to the definition of the χ^2-divergence given in (5.2). On the other hand, in light of Proposition 5.2 we know that $\rho_m(X,Z)$ is equal to the second largest singular value of matrix B of size $M \times |\mathcal{Z}|$ with entries $\frac{P_{XZ}(x,z)}{\sqrt{\mathsf{p}_X(x)P_Z(z)}}$. Let $\sigma_0 \geq \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$ be the singular values of B where $k \coloneqq \min\{M-1, |\mathcal{Z}|-1\}$. It is easy to verify that $\sigma_0 = 1$. Letting $\mathsf{Tr}(\cdot)$ denote the trace of a matrix and B^* denote the conjugate of matrix B, we can then write

$$
1+\sum_{i=1}^{k}\sigma_i^2 = \mathsf{Tr}(BB^*) = \sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} \frac{P_{XZ}^2(x,z)}{\mathsf{p}_X(x)P_Z(z)},
\tag{5.5}
$$

and hence $\sum_{i=1}^{k}\sigma_i^2 = \chi^2(P_{XZ}\|\mathsf{p}_X P_Z)$, which implies that

$$
\chi^2(P_{XZ}\|\mathsf{p}_X P_Z) \leq k\sigma_1^2 \leq (M-1)\rho_m^2(X;Z).
\tag{5.6}
$$

Combining (5.4) and (5.6), we obtain that $I(X;Z) \leq (M-1)\rho_m^2(X,Z)$. Therefore, we conclude that the requirement $\rho_m^2(X,Z) \leq \varepsilon$ implies $I(X;Z) \leq (M-1)\varepsilon$, which completes the proof. □

[2]Note that all the logarithms in this chapter are natural.

5.4 Properties of $\hat{g}(\varepsilon)$

Similar to $g(\varepsilon)$, we clearly have $\hat{\mathcal{D}}_{\varepsilon_1}(\mathsf{P}) \subset \hat{\mathcal{D}}_{\varepsilon_2}(\mathsf{P})$ for $\varepsilon_1 \leq \varepsilon_2$, and hence $\varepsilon \mapsto \hat{g}(\varepsilon)$ is non-decreasing. The following lemma, which is a counterpart of Lemma 3.2, establishes the concavity of $\hat{g}(\varepsilon)$.

Lemma 5.6. *The mapping $\varepsilon \mapsto \hat{g}(\varepsilon)$ is concave for $\varepsilon \geq 0$.*

Proof. Let the privacy filters $\hat{\mathcal{D}}_{\varepsilon_1}(\mathsf{P}) \ni P_{Z_1|Y} : Y \to Z_1$ and $\hat{\mathcal{D}}_{\varepsilon_3}(\mathsf{P}) \ni P_{Z_3|Y} : Y \to Z_3$ be optimal, i.e., $g(\varepsilon_1) = I(Y; Z_1)$ and $g(\varepsilon_3) = I(Y; Z_3)$. Let also the channel $P_{Z_\lambda|Y} : Y \to Z_\lambda$ with output alphabet $\mathcal{Z}_1 \cup \mathcal{Z}_3$ be the random filter constructed in the proof of Lemma 3.2. Then the proof is similar to the proof of Lemma 3.2 except that here we need to show that $P_{Z_\lambda|Y} \in \hat{\mathcal{D}}_{\varepsilon_2}(\mathsf{P})$, where $0 \leq \varepsilon_1 < \varepsilon_2 < \varepsilon_3 \leq \rho_m^2(X, Y)$. To show this, consider $f \in \mathcal{S}_X$ and let U be a binary random variable as in the proof of Lemma 3.2. We then have

$$\begin{aligned}
\mathbb{E}[\mathbb{E}^2[f(X)|Z_\lambda]] &= \mathbb{E}\left[\mathbb{E}[\mathbb{E}^2[f(X)|Z_\lambda]|U]\right] \\
&= \lambda\mathbb{E}[\mathbb{E}^2[f(X)|Z_3]] + \bar{\lambda}\mathbb{E}[\mathbb{E}^2[f(X)|Z_1]], \qquad (5.7)
\end{aligned}$$

We obtain from (5.7) and (5.1) that

$$\begin{aligned}
\rho_m^2(X, Z_\lambda) &= \sup_{f \in \mathcal{S}_X} \mathbb{E}[\mathbb{E}^2[f(X)|Z_\lambda]] \\
&= \sup_{f \in \mathcal{S}_X} \left[\lambda\mathbb{E}[\mathbb{E}^2[f(X)|Z_3]] + \bar{\lambda}\mathbb{E}[\mathbb{E}^2[f(X)|Z_1]]\right] \\
&\leq \lambda\rho_m^2(X; Z_3) + \bar{\lambda}\rho_m^2(X; Z_1) \leq \lambda\varepsilon_3 + \bar{\lambda}\varepsilon_1,
\end{aligned}$$

from which we conclude that $P_{Z_\lambda|Y} \in \hat{\mathcal{D}}_{\varepsilon_2}(\mathsf{P})$ and hence the proof is complete. $\square$

In light of this result, the following corollaries are immediate.

Corollary 5.7. *The mapping* $\varepsilon \mapsto \frac{\hat{g}(\varepsilon)}{\varepsilon}$ *is non-increasing on* $(0, \infty)$.

Proof. We note that since $\varepsilon \mapsto \hat{g}(\varepsilon)$ is concave, the chordal slope $\frac{\hat{g}(\varepsilon)-\hat{g}(0)}{\varepsilon}$ is non-increasing in ε. The corollary then follows by noticing that $\frac{\hat{g}(\varepsilon)}{\varepsilon} = \frac{\hat{g}(\varepsilon)-\hat{g}(0)}{\varepsilon} + \frac{\hat{g}(0)}{\varepsilon}$. □

Corollary 5.8. *For any* $\varepsilon \in [0, \rho_m^2(X,Y))$, *we have*

$$\hat{g}(\varepsilon) \geq \varepsilon \frac{H(Y)}{\rho_m^2(X,Y)} + \hat{g}(0)\left(1 - \frac{\varepsilon}{\rho_m^2(X,Y)}\right).$$

Proof. Due to the concavity, $\hat{g}(\varepsilon)$ must lie above the chord connecting $(0, \hat{g}(0))$ and $(\rho_m^2(X,Y), H(Y))$. □

Remark 5.9. When X is weakly independent of Y (and thus $\hat{g}(0) = 0$), this corollary then implies that $\hat{g}(\varepsilon) \geq \varepsilon \frac{H(Y)}{\rho_m^2(X,Y)}$. This lower bound can be achieved by the simple erasure privacy filter shown in Fig. 3.2 with erasure probability $1 - \frac{\varepsilon}{\rho_m^2(X,Y)}$. This is because for $X \multimap Y \multimap Z_\delta$, where $P_{Z_\delta|Y}$ is an erasure channel with erasure probability δ, we have $\rho_m^2(X, Z_\delta) = \bar{\delta}\rho_m^2(X,Y)$ [157, Page 8] and $I(Y; Z_\delta) = \bar{\delta}H(Y)$.

The lower bound for $\hat{g}(\varepsilon)$ given in Corollary 5.8 is similar to the lower bound for $g(\varepsilon)$ given in (3.9) with $I(X;Y)$ replaced by $\rho_m^2(X,Y)$. Hence, these two bounds coincide if for the given P we have $I(X;Y) = \rho_m^2(X,Y)$. For example if $P_{Y|X} = \mathsf{BEC}(\delta)$ and $X \sim \mathsf{Bernoulli}(\frac{1}{2})$, then $\rho_m^2(X,Y) = I(X;Y) = \bar{\delta}$ and then according to Lemmas 5.5 and 3.36, we have $\hat{g}(\varepsilon) = g(\varepsilon)$. The following lemma generalizes this observation to the case where $P_{Y|X}$ is an erasure channel (see Section 3.8.3 for definition).

Lemma 5.10. *If* $P_{Y|X}$ *is an erasure channel (defined in Section 3.8.3) with erasure probability* δ *with* $0 \leq \delta \leq 1$ *and* $\mathcal{X} = [M]$, *then for any* $0 \leq \varepsilon \leq \bar{\delta}$, *we have*

$$h_{\mathsf{b}}(\delta) + \varepsilon H(X) \leq \hat{g}(\varepsilon) \leq h_{\mathsf{b}}(\delta) + (M-1)\varepsilon.$$

In particular, if $X \sim \mathsf{Bernoulli}(\frac{1}{2})$*, then*

$$\hat{g}(\varepsilon) = g(\varepsilon) = h_{\mathsf{b}}(\delta) + \varepsilon.$$

Proof. The upper bound follows immediately from Lemmas 5.5 and 3.36. For the lower bound we use the observations that $\rho_m^2(X, Y) = \bar{\delta}$ [157, Page 8], $\hat{g}(0) = g(0) = h_{\mathsf{b}}(\delta)$ (Lemma 3.36), and $H(Y) = h_{\mathsf{b}}(\delta) + \bar{\delta} H(X)$, and then apply Corollary 5.8. □

5.5 Binary Observable Data

In this section, we assume that Y is binary and show that it is sufficient to consider a ternary random variable Z in the definition of $\hat{g}(\varepsilon)$. We also derive bounds for $\hat{g}(\varepsilon)$ in the special case of $P_{X|Y}$ being BISO.

5.5.1 Cardinality Bound

We start by the following lemma.

Lemma 5.11. *For a given* P *with marginals* p_X *and* $\mathsf{q}_Y = \mathsf{Bernoulli}(q)$*,* $\hat{g}(\varepsilon)$ *is attained by a privacy filter with a ternary output alphabet, i.e.,* $|\mathcal{Z}| = 3$*.*

Proof. We first recall that $\rho_m^2(X, Z) = \sigma_1$, where σ_1 is the second largest singular value of the matrix B with entries $\frac{P_{XZ}(x,z)}{\sqrt{\mathsf{p}_X(x) P_Z(z)}}$. We remark that B can also be written as $B = AC$, where A and C have entries $\frac{\mathsf{P}(x,y)}{\sqrt{\mathsf{p}_X(x)\mathsf{q}_Y(y)}}$ and $\frac{P_{YZ}(y,z)}{\sqrt{\mathsf{q}_Y(y) P_Z(z)}}$, respectively. This implies that $\mathsf{rank}(B) \le \min\{\mathsf{rank}(A), \mathsf{rank}(C)\}$. It follows that for binary Y, we have $\mathsf{rank}(B) = 2$ (excluding the trivial cases where $\mathsf{rank}(A) = 1$ or $\mathsf{rank}(C) = 1$) and hence in light of (5.5), $\rho_m^2(X, Z) = \chi^2(P_{XZ} \| \mathsf{p}_X P_Z)$.

Consider the mapping $L:[0,1]\to[0,1]^3$ given by $r\mapsto(r,\chi^2(\mathsf{r}_X\|\mathsf{p}_X),h_\mathsf{b}(r))$, where $\mathsf{r}_X(\cdot)\coloneqq P_{X|Y}(\cdot|0)\bar{r}+P_{X|Y}(\cdot|1)r$ is the X-marginal when $Y\sim\mathsf{Bernoulli}(r)$. Let $\mathcal{S}$ be the image of $[0,1]$ under this mapping and $\mathcal{C}$ its convex hull, i.e.,

$$\mathcal{C}=\left\{\sum_{i=1}^k\omega_iL(r_i):\ r_i\in[0,1],k>0,\omega_i\geq 0,\sum_{i=1}^k\omega_i=1\right\}.$$

Since $\mathcal{S}$ is a compact and connected set in $[0,1]^3$, so is $\mathcal{C}$, and hence according to the Carathéodory-Fenchel theorem every points in C can be written as a convex combination of no more than $k=3$ points of $\mathcal{S}$.

Now let $P_{Z|Y}$ be an optimal privacy filter which generates Z taking values in $\mathcal{Z}$. Note that

$$(q,\chi^2(P_{XZ}\|\mathsf{p}_XP_Z),H(Y|Z))=\sum_{z=1}^{|\mathcal{Z}|}P_Z(z)L(P_{Y|Z}(1|z)),$$

and hence $(q,\chi^2(P_{XZ}\|\mathsf{p}_XP_Z),H(Y|Z))\in\mathcal{C}$. Thus there exists a ternary $\tilde{Z}$ such that $\chi^2(P_{XZ}\|\mathsf{p}_XP_Z)=\chi^2(P_{X\tilde{Z}}\|\mathsf{p}_XP_{\tilde{Z}})$ and $H(Y|Z)=H(Y|\tilde{Z})$. $\square$

Remark 5.12. Since Z with $|\mathcal{Z}|=3$ is sufficient to achieve $\hat{g}(\varepsilon)$ when Y is binary, we can improve Lemma 5.5 for binary Y as follows:

$$\hat{g}(\varepsilon)\leq g(\hat{\kappa}\varepsilon),$$

where $\hat{\kappa}\coloneqq\min\{|\mathcal{X}|-1,2\}$. In particular, we have

- If $X\sim\mathsf{Bernoulli}(\frac{1}{2})$ and $P_{Y|X}=\mathsf{BSC}(\alpha)$ with $0\leq\alpha\leq\frac{1}{2}$, then according to Corollaries 3.35 and 5.8 we have for $0\leq\varepsilon\leq(1-2\alpha)^2$

$$\frac{\varepsilon}{(1-2\alpha)^2}\leq\hat{g}(\varepsilon)\leq\frac{\varepsilon}{1-h_\mathsf{b}(\alpha)}.$$

- If $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$ and $P_{X|Y}$ is BISO, then according to Corollaries 3.32 and 5.8 we have for $0 \leq \varepsilon \leq \rho_m^2(X,Y)$

$$\frac{\varepsilon}{\rho_m^2(X,Y)} \leq \hat{g}(\varepsilon) \leq \frac{2\varepsilon}{I(X;Y)}.$$

5.5.2 Binary Input Symmetric Output Channels

In this section, we first derive an necessary condition for the linearity of $\hat{g}$ when $P_{X|Y}$ is BISO. We then derive bounds for two special cases: (i) $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$ and (ii) $P_{X|Y} = \mathsf{BSC}(\alpha)$ and $Y \sim \mathsf{Bernoulli}(q)$ for any $0 \leq q \leq \frac{1}{2}$.

Similar to Section 3.8, we can define the initial efficiency of privacy-constrained information extraction as the derivative $\hat{g}'(0)$ of $\hat{g}(\varepsilon)$ at $\varepsilon = 0$. Analogous to Lemma 3.23, the following lemma provides a lower bound for the initial efficiency.

Lemma 5.13. *For a given joint distribution* P *with marginals* p_X *and* q_Y*, if* X *is weakly independent of* Y *(i.e.,* $\hat{g}(0) = 0$*), then*

$$\hat{g}'(0) \geq \max_{y \in \mathcal{Y}} \frac{-\log \mathsf{q}_Y(y)}{\chi^2(P_{X|Y}(\cdot|y) \| \mathsf{p}_X(\cdot))},$$

where the χ^2*-divergence is defined in* (5.2)*.*

Proof. First, note that it can be verified using (5.5) that if either X or Z is binary, then

$$\rho_m^2(X,Z) = \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \left[\frac{P_{XZ}^2(x,z)}{\mathsf{p}_X(x) P_Z(z)} \right] - 1. \tag{5.8}$$

We use the same privacy filter as in the proof of Lemma 3.23, illustrated in Fig. 3.3. Specifically, let $\mathcal{Z} = \{k, \mathsf{e}\}$ for some fixed $k \in \mathcal{Y}$ and the erasure symbol e, and define the privacy filter $P_{Z|Y}$ by $P_{Z|Y}(k|y) = \delta 1_{\{y=k\}}$, and $P_{Z|Y}(\mathsf{e}|y) = 1 - \delta 1_{\{y=k\}}$, which imply $P_Z(k) = \delta \mathsf{q}(k)$ and $P_Z(\mathsf{e}) = 1 - \delta \mathsf{q}(k)$. Since Z is binary, from (5.8) we can write

$$\begin{aligned}\rho_m^2(X,Z) &= -1 + \sum_{x\in\mathcal{X}} \frac{P_{XZ}^2(x,k)}{\mathsf{p}_X(x)P_Z(k)} + \sum_{x\in\mathcal{X}} \frac{P_{XZ}^2(x,\mathsf{e})}{\mathsf{p}_X(x)P_Z(\mathsf{e})} \\ &\overset{(a)}{=} -1 + \delta \mathsf{q}_Y(k) \sum_{x\in\mathcal{X}} \frac{P_{X|Y}^2(x|k)}{\mathsf{p}_X(x)} + \sum_{x\in\mathcal{X}} \frac{(\mathsf{p}_X(x) - \delta \mathsf{P}(x,k))^2}{\mathsf{p}_X(x)(1-\delta \mathsf{q}_Y(y))}\end{aligned}$$

where (a) follows from the fact that $P_{X|Z}(x|k) = P_{X|Y}(x|k)$ for $k \in \mathcal{Y}$ and $P_{X|Z}(x|\mathsf{e}) = \frac{\mathsf{p}_X(x) - \delta \mathsf{P}(x,k)}{1-\delta \mathsf{q}_Y(k)}$. We can therefore write

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\delta}\rho_m^2(X,Z) &= \mathsf{q}_Y(k) \sum_{x\in\mathcal{X}} \frac{P_{X|Y}^2(x|k)}{\mathsf{p}_X(x)} \\ &\quad + \sum_{x\in\mathcal{X}} \frac{\mathsf{p}_X(x)(1-\delta P_{Y|X}(k|x))}{(1-\delta \mathsf{q}_Y(k))^2}(\delta P_{Y|X}(k|x)\mathsf{q}_Y(k) + \mathsf{q}_Y(k) - 2P_{Y|X}(k|x)),\end{aligned}$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}\delta}\rho_m^2(X,Z)|_{\delta=0} = \mathsf{q}_Y(k)\left[\sum_{x\in\mathcal{X}} \frac{P_{X|Y}^2(x|k)}{\mathsf{p}_X(x)} - 1\right] = \mathsf{q}_Y(k)\chi^2(P_{X|Y}(\cdot|k)\|\mathsf{p}_X(\cdot)). \quad (5.9)$$

The rest follows similarly as in the proof of Lemma 3.23. □

The following result establishes a similar result as Lemma 3.25.

Lemma 5.14. *For any joint distribution* P *with marginals* q_Y *and* p_X*, we have*

$$\frac{H(Y)}{\rho_m^2(X,Y)} \le \frac{dH(Y)}{\chi^2(\mathsf{P}\|\mathsf{p}_X\mathsf{q}_Y)} \le \max_{y\in\mathcal{Y}} \frac{-d\log \mathsf{q}_Y(y)}{\chi^2(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))},$$

where $d \coloneqq \min\{|\mathcal{X}|,|\mathcal{Y}|\} - 1$ *and the second inequality becomes equality if and only if there exists a constant* $c > 0$ *such that* $-\log \mathsf{q}_Y(y) = c\chi^2(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(x))$ *for all* $y \in \mathcal{Y}$.

Proof. First note that from Proposition 4.6 and (5.6), we have

$$\frac{1}{d}\chi^2(\mathsf{P}\|\mathsf{p}_X\mathsf{q}_Y) \le \rho_m^2(X,Y) \le \chi^2(\mathsf{P}\|\mathsf{p}_X\mathsf{q}_Y). \tag{5.10}$$

We therefore obtain

$$\begin{aligned}\frac{H(Y)}{\rho_m^2(X,Y)} &\le \frac{dH(Y)}{\chi^2(\mathsf{P}\|\mathsf{p}_X\mathsf{q}_Y)} = \frac{-d\sum_{y\in\mathcal{Y}}\mathsf{q}_Y(y)\log\mathsf{q}_Y(y)}{\sum_{y\in\mathcal{Y}}\mathsf{q}_Y(y)\chi^2(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))}\\ &\le \max_{y\in\mathcal{Y}}\frac{-d\log\mathsf{q}_Y(y)}{\chi^2(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))},\end{aligned}$$

where the second inequality is analogous to Lemma 3.25. □

Combining Lemmas 5.13 and 5.14, we obtain a necessary condition for the linearity of $\hat{g}$ when $P_{X|Y}$ is BISO.

Theorem 5.15. *If $P_{X|Y}$ is BISO, then $\hat{g}$ is linear only if $Y\sim \mathsf{Bernoulli}(\frac{1}{2})$.*

Proof. First, we notice that since Y is binary (i.e., $d=1$), we have from (5.10) that $\rho_m^2(X,Y)=\chi^2(\mathsf{P}\|\mathsf{p}_X\mathsf{q}_Y)$. Since for binary Y we have $\hat{g}(0)=0$, linearity and concavity of $\hat{g}$ imply $\hat{g}(\varepsilon)=\varepsilon\frac{H(Y)}{\rho_m^2(X,Y)}$. Thus we can write

$$\max_{y\in\{0,1\}}\frac{-\log\mathsf{q}_Y(y)}{\chi^2(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))} \overset{(a)}{\le} \hat{g}'(0) = \frac{H(Y)}{\rho_m^2(X,Y)} \overset{(b)}{\le} \max_{y\in\{0,1\}}\frac{-\log\mathsf{q}_Y(y)}{\chi^2(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))},$$

where (a) and (b) follow from Lemmas 5.13 and 5.14, respectively. We then conclude that $\hat{g}'(0)=\frac{H(Y)}{\rho_m^2(X,Y)}=\max_{y\in\{0,1\}}\frac{-\log\mathsf{q}_Y(y)}{\chi^2(P_{X|Y}(\cdot|y)\|\mathsf{p}_X(\cdot))}$ and hence

$$\frac{\chi^2(P_{X|Y}(\cdot|0)\|\mathsf{p}_X(\cdot))}{\log\bar{q}} = \frac{\chi^2(P_{X|Y}(\cdot|1)\|\mathsf{p}_X(\cdot))}{\log q}, \tag{5.11}$$

where $q := \mathsf{q}_Y(1)$. It is straightforward to modify Lemma A.1 to show that equation (5.11) has only one solution $q = \frac{1}{2}$. $\square$

In light of this theorem, if $P_{X|Y}$ is BISO, then the lower bound in Corollary 5.8 is attained only if Y is uniform. In an attempt to prove the converse, i.e., if $P_{X|Y}$ is BISO and Y is uniform, then $\hat{g}(\varepsilon) = \frac{\varepsilon}{\rho_m^2(X,Y)}$, we obtain the following result. It is still not clear that the converse of Theorem 5.15 holds.

Theorem 5.16. *If $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$ and $P_{X|Y}$ is BISO, then*

$$\hat{g}(\varepsilon) \leq \log\left(\frac{2+2\varepsilon}{1+\rho_m^2(X,Y)}\right).$$

Furthermore, the bound is tight if there exists a function f such that $f(X) = Y$ with probability one.

Proof. Let $\mathcal{X} = \{\pm k, \ldots, \pm 2, \pm 1\}$. As shown earlier, for binary Y we have $\rho_m^2(X,Z) = \chi^2(P_{XZ}\|\mathsf{p}_X P_Z)$. Thus we can write

$$\begin{aligned}
\rho_m^2(X,Z) &= -1 + \sum_{z\in\mathcal{Z}}\sum_{x=-k}^{k}\frac{P_{XZ}^2(x,z)}{\mathsf{p}_X(x)P_Z(z)} = -1 + \sum_{z\in\mathcal{Z}}\sum_{x=-k}^{k}\frac{\mathsf{p}_X(x)P_{Z|X}^2(z|x)}{P_Z(z)} \\
&= -1 + \sum_{z\in\mathcal{Z}}\sum_{x=-k}^{k}\frac{\mathsf{p}_X(x)(P_{Z|Y}(z|0)P_{Y|X}(0|x) + P_{Z|Y}(z|1)P_{Y|X}(1|x))^2}{P_Z(z)} \\
&= -1 + \sum_{z\in\mathcal{Z}}\frac{P_{Z|Y}^2(z|0)}{P_Z(z)}\sum_{x=-k}^{k}\mathsf{p}_X(x)P_{Y|X}^2(0|x) + \sum_{z\in\mathcal{Z}}\frac{P_{Z|Y}^2(z|1)}{P_Z(z)}\sum_{x=-k}^{k}\mathsf{p}_X(x)P_{Y|X}^2(1|x) \\
&\quad +2\sum_{z\in\mathcal{Z}}\frac{P_{Z|Y}(z|0)P_{Z|Y}(z|1)}{P_Z(z)}\sum_{x=-k}^{k}\mathsf{p}_X(x)P_{Y|X}(0|x)P_{Y|X}(1|x) \\
&= -1 + \frac{1}{4}\sum_{z\in\mathcal{Z}}\frac{P_{Z|Y}^2(z|0)}{P_Z(z)}\sum_{x=-k}^{k}\frac{P_{X|Y}^2(x|0)}{\mathsf{p}_X(x)} + \frac{1}{4}\sum_{z\in\mathcal{Z}}\frac{P_{Z|Y}^2(z|1)}{P_Z(z)}\sum_{x=-k}^{k}\frac{P_{X|Y}^2(x|1)}{\mathsf{p}_X(x)}
\end{aligned}$$

$$+\frac{1}{2}\sum_{z\in\mathcal{Z}}\frac{P_{Z|Y}(z|0)P_{Z|Y}(z|1)}{P_Z(z)}\sum_{x=-k}^{k}\frac{P_{X|Y}(x|0)P_{X|Y}(x|1)}{\mathsf{p}_X(x)}. \tag{5.12}$$

Note that

$$\begin{aligned}\frac{1}{2}\sum_{x=-k}^{k}\frac{P^2_{X|Y}(x|0)}{\mathsf{p}_X(x)} &= \sum_{x=-k}^{-1}\frac{P^2_{X|Y}(-x|1)}{P_{X|Y}(-x|1)+P_{X|Y}(x|1)}+\sum_{x=1}^{k}\frac{P^2_{X|Y}(-x|1)}{P_{X|Y}(-x|1)+P_{X|Y}(x|1)}\\ &= \sum_{x=1}^{k}\frac{P^2_{X|Y}(x|1)}{P_{X|Y}(-x|1)+P_{X|Y}(x|1)}+\sum_{x=1}^{k}\frac{P^2_{X|Y}(-x|1)}{P_{X|Y}(-x|1)+P_{X|Y}(x|1)}\\ &= \frac{1}{2}\sum_{x=-k}^{k}\frac{P^2_{X|Y}(x|1)}{\mathsf{p}_X(x)}.\end{aligned}$$

On the other hand, we can write

$$\rho_m^2(X,Y)=-1+\sum_{x\in\mathcal{X}}\sum_{y\in\{0,1\}}\frac{\mathsf{P}^2(x,y)}{\mathsf{p}_X(x)\mathsf{q}_Y(y)}=-1+\frac{1}{2}\sum_{x=-k}^{k}\frac{P^2_{X|Y}(x|0)}{\mathsf{p}_X(x)}+\frac{1}{2}\sum_{x=-k}^{k}\frac{P^2_{X|Y}(x|1)}{\mathsf{p}_X(x)}.$$

Thus we have

$$\frac{1}{2}\sum_{x=-k}^{k}\frac{P^2_{X|Y}(x|0)}{\mathsf{p}_X(x)}=\frac{1}{2}\sum_{x=-k}^{k}\frac{P^2_{X|Y}(x|1)}{\mathsf{p}_X(x)}=\frac{1}{2}(1+\rho_m^2(X,Y)). \tag{5.13}$$

Plugging (5.13) into (5.12), we obtain

$$\begin{aligned}\rho_m^2(X,Z) &= -1+\frac{1}{2}(1+\rho_m^2(X,Y))\left[\frac{1}{2}\sum_{z\in\mathcal{Z}}\frac{P^2_{Z|Y}(z|0)}{P_Z(z)}+\frac{1}{2}\sum_{z\in\mathcal{Z}}\frac{P^2_{Z|Y}(z|1)}{P_Z(z)}\right]\\ &\quad+\frac{1}{2}\mathsf{K}(P_{Z|Y}(\cdot|0)\|P_{Z|Y}(\cdot|1))\mathsf{K}(P_{X|Y}(\cdot|0)\|P_{X|Y}(\cdot|1))\\ &= -1+\frac{1}{2}(1+\rho_m^2(X,Y))(1+\rho_m^2(Y,Z))\\ &\quad+\frac{1}{2}\mathsf{K}(P_{Z|Y}(\cdot|0)\|P_{Z|Y}(\cdot|1))\mathsf{K}(P_{X|Y}(\cdot|0)\|P_{X|Y}(\cdot|1))\end{aligned}$$

$$\overset{(a)}{\geq} \quad -1+\frac{1}{2}(1+\rho_m^2(X,Y))(1+\rho_m^2(Y,Z)), \tag{5.14}$$

where

$$\mathsf{K}(P\|Q) \coloneqq \sum_{i\in\mathcal{I}} \frac{P(i)Q(i)}{0.5P(i)+0.5Q(i)},$$

for any pairs of probability distributions P and Q supported over a set $\mathcal{I}$. Inequality (a) becomes equality if $\mathsf{K}(P_{X|Y}(\cdot|0)\|P_{X|Y}(\cdot|1))=0$ or equivalently $P_{X|Y}(x|0)P_{X|Y}(x|1)=0$ for all $x\in\mathcal{X}$. After relabeling if necessary, we can assume that there exist disjoints sets $\mathcal{X}_0$ and $\mathcal{X}_1$ such that $\mathcal{X}=\mathcal{X}_0\cup\mathcal{X}_1$ and $P_{X|Y}(x|1)=0$ for all $x\in\mathcal{X}_0$, and $P_{X|Y}(x|0)=0$ for all $x\in\mathcal{X}_1$. We then define a boolean function f as $f(x)=0$ if $x\in\mathcal{X}_0$ and $f(x)=1$ if $x\in\mathcal{X}_1$.

The inequality in (5.14) implies

$$1+\rho_m^2(Y,Z)\leq\frac{2+2\rho_m^2(X,Z)}{1+\rho_m^2(X,Y)},$$

from which, and the fact that $I(Y;Z)\leq\log(1+\rho_m^2(Y,Z))$ proved in (5.4), the result follows. □

Remark 5.17. If there exists a function f such that $f(X)=Y$ with probability one, or equivalently $\mathsf{K}(P_{X|Y}(\cdot|0)\|P_{X|Y}(\cdot|1))=0$, then we obtain two Markov chains $X \multimap Y \multimap Z$ and $Y \multimap X \multimap Z$. Due to the data processing inequality for maximal correlation (5.3), we conclude that $\rho_m^2(X,Z)=\rho_m^2(Y,Z)$, and hence, according to the inequality $I(Y;Z)\leq\log(1+\rho_m^2(Y,Z))$, we obtain that $\hat{g}(\varepsilon)\leq\log(1+\varepsilon)$. In fact, Theorem 5.16 proves that in this case this bound holds with equality because the existence of such a function f implies that $\rho_m^2(X,Y)=1$.

We close this chapter by modifying the geometric approach that Witsenhausen and

Wyner [147] proposed to generalize Mrs. Gerber's Lemma [153] (see Section 3.5). To this end, let $\varrho^2_{\mathsf{T}}(q, \Delta)$ be defined as

$$\varrho^2_{\mathsf{T}}(q, \Delta) \coloneqq \min_{\substack{P_{Z|Y}: \rho^2_m(Y,Z) \geq \Delta \\ X -\!\circ\!- Y -\!\circ\!- Z}} \rho^2_m(X, Z),$$

for any $0 \leq \Delta \leq 1$, where $Y \sim \mathsf{q}_Y = \mathsf{Bernoulli}(q)$ and T is the channel from Y to X. Recall that $X \sim \mathsf{p}_X$. Consider the map $\tau : [0,1] \to [0,1]^3$ given by $r \mapsto (r, \chi^2(\mathsf{r}_Y \| \mathsf{q}_Y), \chi^2(\mathsf{r}_X \| \mathsf{p}_X))$, where $\mathsf{r}_Y = \mathsf{Bernoulli}(r)$ and $\mathsf{r}_X(\cdot) = P_{X|Y}(\cdot|0)\bar{r} + P_{X|Y}(\cdot|1)r$ is the X-marginal when $Y \sim \mathsf{r}_Y$. Let $\mathcal{S}$ be the image of $[0,1]$ under τ and $\mathcal{C}$ be its convex hull. By definition, $\mathcal{C}$ can be written as the collection of triplets $(s, \chi^2(P_{Y'|Z}P_Z \| \mathsf{q}_Y P_Z), \chi^2(P_{X'|Z}P_Z \| \mathsf{p}_X P_Z))$, where $s = \sum_{i=1}^k \omega_i r_i$, $P_Z(i) = \omega_i$, $P_{Y'|Z}(\cdot|i) = \mathsf{Bernoulli}(r_i)$, and $P_{X'|Z}(\cdot|i) = P_{X|Y}(\cdot|0)\bar{r}_i + P_{X|Y}(\cdot|1)r_i$ for $i \in [k]$ and some integer k. Note that if and only if $s = q$, then the pair (X', Y') has the same distribution as the given pair (X, Y). Since Y is binary $\rho^2_m(X, Z) = \chi^2(P_{XZ} \| \mathsf{p}_X P_Z)$ and $\rho^2_m(Y, Z) = \chi^2(P_{YZ} \| \mathsf{q}_Y P_Z)$. Consequently, the graph of $\varrho^2_{\mathsf{T}}(q, \cdot)$ is the lower boundary of the convex set $\mathcal{C}_q \coloneqq \mathcal{C} \cap \{s = q\}$.

Similar to (3.12), we can define the conjugate function

$$\boldsymbol{r}^2_{\mathsf{T}}(q, \lambda) \coloneqq \min_{0 \leq \Delta \leq 1} \varrho^2_{\mathsf{T}}(q, \Delta) - \lambda\Delta = \min\{\eta - \lambda\Delta : \ (\Delta, \eta) \in \mathcal{C}_q\}, \tag{5.15}$$

for every $\lambda \geq 0$. It can be verified that

$$\varrho^2_{\mathsf{T}}(q, \Delta) = \max_{\lambda \geq 0} \boldsymbol{r}^2_{\mathsf{T}}(q, \lambda) + \lambda\Delta. \tag{5.16}$$

Using a technique similar to Section 3.5, we can show that for a fixed λ the graph of $\boldsymbol{r}^2_{\mathsf{T}}(\cdot, \lambda)$

is the lower convex envelope of the map $\phi(\,\cdot\,,\lambda):[0,1]\to\mathbb{R}$, given by

$$\phi(r,\lambda)=\chi^2(\mathsf{r}_X\|\mathsf{p}_X)-\lambda\chi^2(\mathsf{r}_Y\|\mathsf{q}_Y).$$

Hence, similar to Section 3.5, we only need to focus on the domain of $\phi(\cdot,\lambda)$ on which it differs from $\mathscr{r}_{\mathsf{T}}^2(\cdot,\lambda)$ for a given λ. Computing $\mathscr{r}_{\mathsf{T}}^2(\cdot,\lambda)$ from $\phi(\cdot,\lambda)$ follows the same procedure as given in Section 3.5. The following lemma clarifies this approach in the simple binary symmetric case.

Lemma 5.18. *Let* $P_{X|Y}=\mathsf{BSC}(\alpha)$ *and* $Y\sim\mathsf{q}_Y=\mathsf{Bernoulli}(q)$ *with* $0\le\alpha,q\le\frac{1}{2}$. *Then for any* $\varepsilon\le\rho_m^2(X,Y)$

$$\hat{g}(\varepsilon)\le\log\left(1+\frac{\varepsilon}{\rho_m^2(X,Y)}\right).$$

Proof. For notational simplicity let $\chi_b^2(a\|b):=\frac{a^2}{b}+\frac{\bar{a}^2}{\bar{b}}-1$ for $0<a,b<1$. Then, we can write $\phi(r,\lambda)=\chi_b^2(r*\alpha\|q*\alpha)-\lambda\chi_b^2(r\|q)$. It is straightforward to show that $\phi(\cdot,\lambda)$ is convex for $0<\lambda<\rho_m^2(X,Y)$ and concave for $\lambda\ge\rho_m^2(X,Y)$. Therefore we need to focus on $\lambda\ge\rho_m^2(X,Y)$. Note that $(q,\mathscr{r}_{\mathsf{T}}^2(q,\lambda))$ can be written as a convex combination of the points $(0,\phi(0,\lambda))$ and $(1,\phi(1,\lambda))$ with weights $\bar{q}$ and q, respectively. Thus, for any $\lambda\ge\rho_m^2(X,Y)$

$$\begin{aligned}\mathscr{r}_{\mathsf{T}}^2(q,\lambda) &= \frac{\alpha^2\bar{q}+\bar{\alpha}^2q}{\alpha*q}+\frac{\bar{\alpha}^2\bar{q}+\alpha^2q}{1-\alpha*q}-1-\lambda\\ &= \rho_m^2(X,Y)-\lambda.\end{aligned}$$

Now that we obtain $\mathscr{r}_{\mathsf{T}}^2(q,\lambda)$, we invoke (5.16) to write

$$\varrho_{\mathsf{T}}^2(q,\mathit{\Delta}) = \max_{\lambda\ge\rho_m^2(X,Y)}\rho_m^2(X,Y)-\lambda+\lambda\mathit{\Delta}$$

$$= \quad \varDelta\rho_m^2(X,Y) \tag{5.17}$$

Since $\varrho_{\mathsf{T}}^2(q,\varDelta) = \varDelta\rho_m^2(X,Y)$, we conclude that

$$\max_{\substack{P_{Z|Y}:\rho_m^2(X,Z)\leq\varepsilon \\ X \multimap Y \multimap Z}} \rho_m^2(Y,Z) = \frac{\varepsilon}{\rho_m^2(X,Y)}. \tag{5.18}$$

Since we have $I(Y;Z) \leq \log(1+\rho_m^2(Y,Z))$, the results immediately follows. $\square$

It is interesting to note that, on the one hand, the strong data processing inequality for maximal correlation (5.3) implies that $\rho_m(X,Z) \leq \rho_m(Y,Z)\rho_m(X,Y)$ and, on the other hand, (5.18) implies that $\rho_m(X,Z) \geq \rho_m(Y,Z)\rho_m(X,Y)$. Hence, if the channel from Y to X is BSC, then we have

$$\rho_m(X,Z) = \rho_m(Y,Z)\rho_m(X,Y), \tag{5.19}$$

for any *arbitrary* channel $P_{Z|Y}$ which forms the Markov chain $X \multimap Y \multimap Z$. A mutual information counterpart of (5.19) can be obtained from Corollary 3.32. If $Y \sim \mathsf{Bernoulli}(\frac{1}{2})$ and $P_{X|Y}$ is BISO, then

$$I(X;Z) \geq I(X;Y)I(Y;Z),$$

for any *arbitrary* channel $P_{Z|Y}$ which forms the Markov chain $X \multimap Y \multimap Z$.

Chapter 6

Privacy-Aware Guessing Efficiency

6.1 Overview

As seen in the previous chapter, the privacy measure based on maximal correlation results in an operational interpretation; that is the requirement $\rho_m^2(X,Z) \leq \varepsilon$ implies that $\mathsf{P_c}(f(X)|Z) - \mathsf{P_c}(f(X)) \leq O(\sqrt{\varepsilon})$ for any non-constant function f. However, the corresponding rate-privacy function $\hat{g}(\varepsilon)$ is difficult to calculate in closed form. In order to overcome this difficulty and at the same time to enjoy the operational interpretation of privacy, we propose to measure privacy when both X and Y are discrete in terms of Arimoto's mutual information . In fact, we utilize Arimoto's mutual information to measure both utility and privacy and then define a *parametric family* $g^{(\nu,\mu)}(\varepsilon)$ of utility-privacy tradeoffs. In the uniform case, the parameters $\nu \in [1,\infty]$ and $\mu \in [1,\infty]$ correspond to the sensitivity of privacy and utility, respectively, i.e., for $\nu_1 \leq \nu_2$ and $\mu_1 \leq \mu_2$ we have $g^{(\nu_2,\mu_1)}(\varepsilon) \leq g^{(\nu_1,\mu_1)}(\varepsilon) \leq g^{(\nu_1,\mu_2)}(\varepsilon)$. Of this family of utility-privacy tradeoffs, two extreme cases are particularly interesting: $g^{(1,1)}(\varepsilon)$, which equals $g(\varepsilon)$, and $g^{(\infty,\infty)}(\varepsilon)$, that is the limit of $g^{(\nu,\mu)}(\varepsilon)$ as both ν and μ tend to ∞. As seen in Chapter 3, the former provides an information-theoretic formulation for the utility-privacy tradeoff. In this chapter, we see that the latter provides an estimation-theoretic formulation for the utility-privacy tradeoff. In fact, $g^{(\infty,\infty)}(\varepsilon)$ provides a quantitative answer to the following question: Among all discrete random variables Z satisfying $X \multimap Y \multimap Z$, what is the largest $\log \frac{\mathsf{P_c}(Y|Z)}{\mathsf{P_c}(Y)}$ such that $\log \frac{\mathsf{P_c}(X|Z)}{\mathsf{P_c}(X)} \leq \varepsilon$ for a given ε?

6.1.1 Main Contribution

The main contributions of this chapter are as follows:

- We first describe a decision-theoretic setting where given two discrete random variables X and Z we define the so-called *information leakage* as the amount of information leaking from X to Z. We then show that, in some particular cases, the information leakage is in a one-to-one correspondence with Arimoto's mutual information and thus provide an operational interpretation for the privacy measure based on Arimoto's mutual information.

- Given $1 \leq \nu, \mu \leq \infty$, we then define a parametric family $g^{(\nu,\mu)}(\varepsilon)$ of utility-privacy tradeoffs which is shown to include $g(\varepsilon)$. It is argued that $g^{(1,1)}(\varepsilon) = g(\varepsilon)$ and $g^{(\infty,\infty)}(\varepsilon)$ can lower and upper bound each $g^{(\nu,\mu)}(\varepsilon)$ for $\nu, \mu > 1$. This observation motivates us to concentrate on evaluating $g^{(\infty,\infty)}(\varepsilon)$.

- In evaluating $g^{(\infty,\infty)}(\varepsilon)$, we define the so-called *privacy-constrained guessing probability* $\hbar(\varepsilon)$ as the maximum of $\mathsf{P_c}(Y|Z)$, where the maximization is taken over $P_{Z|Y}$ such that $X \multimap Y \multimap Z$ and $\mathsf{P_c}(X|Z) \leq \varepsilon$. We observe that $\hbar(\varepsilon)$ has a one-to-one relationship with $g^{(\infty,\infty)}(\varepsilon)$ and that it is easier to deal with, and thus we turn our attention to $\hbar(\varepsilon)$.

- Using geometric properties of the set of the privacy filters, we prove some functional properties of $\hbar$. In particular, we show that it is strictly increasing, concave and piecewise linear on $[\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$. These properties allow us to derive a closed form expression for $\hbar$ in the binary case, i.e., $|\mathcal{X}| = |\mathcal{Y}| = 2$. The optimal privacy filter in this case is a simple Z-channel which establishes an optimal privacy-preserving mechanism to avoid survey response bias [145]; see Example 1.3.

- To study the non-binary case, we define $\underline{\hbar}$ similar to $\hbar$ with a further assumption that $\mathcal{Z} = \mathcal{Y}$. Using a novel technique, we compute, $\underline{\hbar}'(\mathsf{P_c}(X|Y))$, the derivate of $\underline{\hbar}(\cdot)$ at $\varepsilon = \mathsf{P_c}(X|Y)$ in closed form and show that there exists a constant $0 \leq \varepsilon_{\mathsf{L}} < \mathsf{P_c}(X|Y)$ such that $\underline{\hbar}(\varepsilon) = 1 - (\mathsf{P_c}(X|Y))\underline{\hbar}'(\mathsf{P_c}(X|Y))$ for $\varepsilon \in [\varepsilon_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$. Thus, this technique yields a closed form expression for $\underline{\hbar}(\varepsilon)$ for general pair of discrete random variables (X, Y) for sufficiently large, but nontrivial, values of ε. By assuming $\mathcal{X} = \mathcal{Y} = \{0, 1\}^n$, this result enables us to derive expression for $\underline{\hbar}$ corresponding to n-tuples (X^n, Y^n), where X^n consists of the first n samples of either a memoryless or a first-order Markov process with a symmetric transition matrix and Y^n is the output of a memoryless $\mathsf{BSC}(\alpha)$ fed with X^n.

In this chapter, we need the following definitions. The Rényi entropy $H_\nu(X)$ of order $\nu \in [1, \infty]$ is defined as

$$H_\nu(X) := \begin{cases} H(X), & \nu = 1, \\ \frac{1}{1-\nu} \log\left(\sum_{x\in\mathcal{X}} \mathsf{p}_X^\nu(x)\right), & 1 < \nu < \infty, \\ -\log\left(\mathsf{P_c}(X)\right), & \nu = \infty, \end{cases} \tag{6.1}$$

where $\mathsf{P_c}(X)$ was defined in Definition 5.4. Arimoto's conditional entropy of order $\nu \in$

$[1, \infty]$ is defined as

$$H_\nu(X|Z) \coloneqq \begin{cases} H(X|Z), & \nu = 1, \\ \frac{\nu}{1-\nu} \log \left(\sum_{z \in \mathcal{Z}} P_Z(z) \left[\sum_{x \in \mathcal{X}} P^\nu_{X|Z}(x|z) \right]^{1/\nu} \right), & 1 < \nu < \infty, \\ -\log \left(\mathsf{P_c}(X|Z) \right), & \nu = \infty, \end{cases} \tag{6.2}$$

where $\mathsf{P_c}(X|Z)$ was defined in Definition 5.4. *Arimoto's mutual information of order* $\nu \in [1, \infty]$ is then defined as $I_\nu(X; Z) \coloneqq H_\nu(X) - H_\nu(X|Z)$ (see, e.g. [142]). Note that $I_1(X; Z) = I(X; Z)$.

6.2 Loss-Based Information Leakage: A General Framework

Consider a pair of random variables $(X, Z) \in \mathcal{X} \times \mathcal{Z}$. Using a decision rule $\psi : \mathcal{Z} \to \hat{\mathcal{X}}$, a decision maker, say Bob, takes $\hat{x} = \psi(z)$ as a prediction of the target variable X whenever $Z = z$. In this context, a loss function $\ell : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ quantifies through $\ell(x, \hat{x})$ the loss suffered by Bob when the true value of X is x but he used $\hat{x}$ as an estimate of X.

The Bayes map for loss function ℓ is the optimal decision rule, i.e., the map ψ that minimizes the expected loss. When no side information is available, the loss corresponding to the Bayes map is $\inf_{\hat{x} \in \hat{\mathcal{X}}} \mathbb{E}[\ell(X, \hat{x})]$ and when side information Z is available, the corresponding loss is

$$\inf_{\psi : \mathcal{Z} \to \hat{\mathcal{X}}} \mathbb{E}[\ell(X, \psi(Z))] = \inf_{P_{\hat{X}|Z} : X -\circ- Z -\circ- \hat{X}} \mathbb{E}[\ell(X, \hat{X})].$$

In this context, we propose the difference between the log-losses with and without side information Z as a measure of the information leakage from X to Z.

Definition 6.1. *For a given loss function $\ell : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, the information leakage from X to Z is defined as*

$$\mathcal{L}_\ell(X \to Z) \coloneqq \sup_{P_{\hat{X}|Z}: X -\!\!\circ\!\!- Z -\!\!\circ\!\!- \hat{X}} \log \frac{\inf_{\hat{x} \in \hat{\mathcal{X}}} \mathbb{E}[\ell(X, \hat{x})]}{\mathbb{E}[\ell(X, \hat{X})]}.$$

Since side information can only improve the performance, $\mathcal{L}_\ell(X \to Z) \geq 0$.

Example 6.2 (Hamming loss function). Let $\hat{\mathcal{X}} = \mathcal{X} = [M]$ and $\ell_{\mathsf{H}}(x, \hat{x}) = 1_{\{x \neq \hat{x}\}}$, where $1_{\{\cdot\}}$ is the indicator function. Then, the associated information leakage $\mathcal{L}_{\mathsf{H}}(X \to Z)$ is given by

$$\mathcal{L}_{\mathsf{H}}(X \to Z) = \log \frac{1 - \mathsf{P}_{\mathsf{c}}(X)}{1 - \mathsf{P}_{\mathsf{c}}(X|Z)} = \log \frac{1 - 2^{-H_\infty(X)}}{1 - 2^{-H_\infty(X|Z)}},$$

where the second equality follows from (6.1) and (6.2).

Note that if $X \sim \mathsf{Bernoulli}(p)$ and $P_{Z|X} = \mathsf{BSC}(\alpha)$ with $p \in [\frac{1}{2}, 1]$ and $\alpha \in [0, \frac{1}{2}]$, then $\mathsf{P}_{\mathsf{c}}(X) = p$ and $\mathsf{P}_{\mathsf{c}}(X|Z) = p\bar{\alpha} + \max\{\bar{p}\bar{\alpha}, \alpha p\}$. In this case, it is straightforward to verify that $\mathcal{L}_{\mathsf{H}}(X \to Z) = 0$ if and only if $p \geq \bar{\alpha}$. Therefore, if $\frac{1}{2} < \bar{\alpha} \leq p < 1$ then $\mathcal{L}_{\mathsf{H}}(X \to Z) = 0$ even though Z is not independent of X. This example shows that, in general, $\mathcal{L}_\ell(X \to Z) = 0$ does not imply $X \perp\!\!\!\perp Z$. This contrasts with other notions of information leakage: mutual information in Chapter 3, maximal correlation in Chapter 5, and Sibson's mutual information in [79], where zero leakage is equivalent to independence.

Example 6.3 (Generalized Hamming loss function). Let $\mathcal{X} = [M]$ and $\hat{\mathcal{X}}$ be the set of all probability distributions over $\mathcal{X}$. For $\nu \in (1, \infty)$, the generalized Hamming loss function of order ν is defined as [141, eq. 10.27]

$$\ell_\nu(x, Q) \coloneqq \frac{\nu}{\nu - 1}\left(1 - Q(x)^{\frac{\nu-1}{\nu}}\right).$$

For this loss function, the associated information leakage $\mathcal{L}_\nu^{\mathsf{GH}}(X \to Z)$ satisfies

$$\mathcal{L}_\nu^{\mathsf{GH}}(X \to Z) = \log \frac{1 - 2^{-\frac{\nu-1}{\nu} H_\nu(X)}}{1 - 2^{-\frac{\nu-1}{\nu} H_\nu(X|Z)}}. \tag{6.3}$$

It is straightforward to show that $\lim_{\nu \to 1} \mathcal{L}_\nu^{\mathsf{GH}}(X \to Z) = \dfrac{H(X)}{H(X|Z)}$ and $\lim_{\nu \to \infty} \mathcal{L}_\nu^{\mathsf{GH}}(X \to Z) = \mathcal{L}_{\mathsf{H}}(X \to Z)$.

Example 6.4 (Squared-error loss function). Let $\mathcal{X} = \hat{\mathcal{X}} = \mathbb{R}$ and $\ell(x, \hat{x}) = (x - \hat{x})^2$. The corresponding information leakage $\mathcal{L}_{\mathsf{MS}}(X \to Z)$ is given by

$$\mathcal{L}_{\mathsf{MS}}(X \to Z) = \log \frac{\mathsf{var}(X)}{\mathsf{mmse}(X|Z)},$$

where $\mathsf{mmse}(X|Z) := \mathbb{E}[(X - \mathbb{E}[X|Z])^2] = \mathbb{E}[\mathsf{var}(X|Z)]$.

We end this section by considering the following decision problem. Suppose that Alice observes Y and, in order to receive a utility, she has to disclose it to Bob. In general, Y might be correlated to her private information, represented by X. To maintain her privacy, Alice would like to disclose another random variable Z which, on the one hand, maximizes the utility, and on the other hand, preserves the privacy of X. From an estimation theoretic point of view, it is reasonable to measure the utility in terms of the efficiency of Bob in estimating Y. A way to measure this estimation efficiency is by the expectation of the reward function associated to a loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}^+$. In this case, for a given Z, Bob seeks to maximize $-\mathbb{E}[\ell(Y, \hat{Y}(Z))]$ over all estimators $\hat{Y} : \mathcal{Z} \to \hat{\mathcal{Y}}$. Since such a maximum is in a one-to-one correspondence with $\mathcal{L}_\ell(Y \to Z)$, we will take the latter as a utility measure. Similarly, we will take $\mathcal{L}_{\ell'}(X \to Z)$ as a measure of privacy where $\ell' : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ is a loss function. In order to quantify the tradeoff between $\mathcal{L}_\ell(Y \to Z)$ and $\mathcal{L}_{\ell'}(X \to Z)$, we introduce the *utility-privacy function* for discrete X and Y in Section 6.3

using generalized Hamming loss functions $\ell = \ell_\mu$ and $\ell' = \ell_\nu$ and the *estimation-noise-to-signal ratio* for continuous X and Y in Chapter 7 using the squared-error loss function for both ℓ and ℓ'.

6.3 Discrete Scalar Case

In this section, we assume that X and Y are discrete random variables taking values in $\mathcal{X} = [M]$ and $\mathcal{Y} = [N]$, respectively. Let $\mathsf{P}(x, y) \coloneqq P_{XY}(x, y)$ with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be their joint distribution and p_X and q_Y the marginal distributions. For every $\nu, \mu \in [1, \infty]$, we define $\mathcal{G}^{(\nu,\mu)}(\mathsf{P}, \cdot\,) : [0, \infty) \to \mathbb{R}$ by

$$\mathcal{G}^{(\nu,\mu)}(\mathsf{P}, \varepsilon) \coloneqq \sup_{\substack{X -\circ- Y -\circ- Z \\ \mathcal{L}^{\mathsf{GH}}_{\nu}(X \to Z) \leq \varepsilon}} \mathcal{L}^{\mathsf{GH}}_{\mu}(Y \to Z),$$

which is a measure of the tradeoff between estimation efficiency and privacy discussed in the previous section. In connection with the rate-privacy function in Chapter 3, we define the following family of utility-privacy functions.

Definition 6.5. *For every $\nu, \mu \in [1, \infty]$, we define the utility-privacy function $g^{(\nu,\mu)}(\mathsf{P}, \cdot\,) : [0, \infty) \to \mathbb{R}$ by*

$$g^{(\nu,\mu)}(\mathsf{P}, \varepsilon) \coloneqq \sup_{\substack{X -\circ- Y -\circ- Z \\ I_{\nu}(X;Z) \leq \varepsilon}} I_\mu(Y; Z).$$

Notice that the function $g(\varepsilon)$ introduced in Chapter 3 equals $g^{(1,1)}(\mathsf{P}, \varepsilon)$. If $Z \perp\!\!\!\perp X$, then both $\mathcal{L}^{\mathsf{GH}}_{\nu}(X \to Z)$ and $I_\nu(X; Z)$ equal zero. Therefore, $\mathcal{G}^{(\nu,\mu)}(\mathsf{P}, \varepsilon)$ and $g^{(\nu,\mu)}(\mathsf{P}, \varepsilon)$ are well defined. When there is no risk of confusion, we will omit P in $\mathcal{G}^{(\nu,\mu)}(\mathsf{P}, \varepsilon)$ and $g^{(\nu,\mu)}(\mathsf{P}, \varepsilon)$.

It is easy to see from (6.3) that

$$\mathcal{G}^{(\nu,\mu)}(\varepsilon) = \log\left(\frac{1-2^{-\frac{\mu-1}{\mu}H_\mu(Y)}}{1-2^{-\frac{\mu-1}{\mu}[H_\mu(Y)-g^{(\nu,\mu)}(\vartheta(\varepsilon))]}}\right),$$

where $\vartheta(\varepsilon) = \dfrac{\nu}{\nu-1}\log\left(2^{\frac{\nu-1}{\nu}H_\nu(X)}\left(1-2^{-\varepsilon}\right)+2^{-\varepsilon}\right)$. This shows that $\mathcal{G}^{(\nu,\mu)}$ can be obtained from the utility-privacy function $g^{(\nu,\mu)}$. To characterize $\mathcal{G}^{(\nu,\mu)}$, we can therefore focus on $g^{(\nu,\mu)}$.

Computing $g^{(\nu,\mu)}$ for every $\nu,\mu>1$ seems to be complicated even for the simple binary case. However, the following lemma provides lower and upper bounds for $g^{(\nu,\mu)}$ in terms of $g^{(\infty,\infty)}$. For notational simplicity, we let $g^{\nu}(\varepsilon)$ denote $g^{(\nu,\nu)}(\varepsilon)$.

Lemma 6.6. *Let (X,Y) be a pair of random variables and $\nu,\mu\in(1,\infty)$. Then*

$$g^{(\nu,\mu)}(\varepsilon) \le g^{\infty}\left(\frac{\nu-1}{\nu}\varepsilon+\frac{1}{\nu}H_\infty(X)\right)+H_\mu(Y)-H_\infty(Y),$$

for any $\varepsilon\ge 0$, and

$$g^{(\nu,\mu)}(\varepsilon) \ge \frac{\mu}{\mu-1}g^{\infty}\left(\varepsilon-H_\nu(X)+H_\infty(X)\right)-\frac{1}{\mu-1}H_\infty(Y),$$

for any $\varepsilon\ge H_\nu(X)-H_\infty(X)$.

Proof. The facts that $\nu\mapsto H_\nu(X|Z)$ is non-increasing on $[1,\infty]$ [56, Proposition 5] and $(\sum_i |x_i|^p)^{1/p}\ge \max_i |x_i|$ for all $p\ge 0$ imply

$$\frac{\nu-1}{\nu}H_\nu(X|Z)\le H_\infty(X|Z)\le H_\nu(X|Z). \tag{6.4}$$

Since $I_\infty(X;Z) = H_\infty(X) - H_\infty(X|Z)$, the above lower bound implies

$$\begin{aligned} I_\infty(X;Z) &\le H_\infty(X) - \frac{\nu-1}{\nu}H_\nu(X) + \frac{\nu-1}{\nu}I_\nu(X;Z) \\ &\le \frac{1}{\nu}H_\infty(X) + \frac{\nu-1}{\nu}I_\nu(X;Z), \end{aligned} \tag{6.5}$$

where the second inequality follows from the fact that $\nu \mapsto H_\nu(X)$ is non-increasing. Since $I_\mu(Y;Z) = H_\mu(Y) - H_\mu(Y|Z)$, the upper bound in (6.4) implies

$$I_\mu(Y;Z) \le I_\infty(Y;Z) + H_\mu(Y) - H_\infty(Y). \tag{6.6}$$

Therefore,

$$g^{(\nu,\mu)}(\varepsilon) \le g^\infty\left(\frac{1}{\nu}H_\infty(X) + \frac{\nu-1}{\nu}\varepsilon\right) + H_\mu(Y) - H_\infty(Y).$$

Similarly, interchanging X, ν and Y, μ in (6.5) and (6.6), we obtain

$$g^{(\nu,\mu)}(\varepsilon) \ge \frac{\mu}{\mu-1}g^\infty\Big(\varepsilon - H_\nu(X) + H_\infty(X)\Big) - \frac{1}{\mu-1}H_\infty(Y)$$

whenever $\varepsilon \ge H_\nu(X) - H_\infty(X)$. $\square$

In light of this lemma, we can focus on g^∞ as it provides upper and lower bounds for $g^{(\nu,\mu)}$, $1 < \nu, \mu \le \infty$. In order to study g^∞, we need the following definition.

Definition 6.7. *Given a pair of discrete random variables $(X,Y) \sim \mathsf{P}$ and $\varepsilon > 0$, the privacy-constrained guessing probability is defined as*

$$\hbar(\mathsf{P},\varepsilon) := \sup_{\substack{P_{Z|Y}: X -\!\!\circ\!\!- Y -\!\!\circ\!\!- Z, \\ \mathsf{P_c}(X|Z) \le \varepsilon}} \mathsf{P_c}(Y|Z). \tag{6.7}$$

For brevity, we write $\hslash(\varepsilon)$ for $\hslash(\mathsf{P},\varepsilon)$. The fact that $I_\infty(X;Z) = \log\left(\frac{\mathsf{P_c}(X|Z)}{\mathsf{P_c}(X)}\right)$ implies

$$g^\infty(\varepsilon) = \log\frac{\hslash(2^\varepsilon\mathsf{P_c}(X))}{\mathsf{P_c}(Y)}. \tag{6.8}$$

The above functional relation allows us to translate results for $\hslash$ into results for g^∞.

In summary, $\mathcal{G}^{(\nu,\mu)}$ and $g^{(\nu,\mu)}$ are intrinsically related and both quantify the tradeoff between utility and privacy. The latter family of functions can be bounded using the function g^∞ which, by the functional relation (6.8), can be obtained from $\hslash$. The quantity $\hslash$ is thus not only operational by itself, but it provides bounds for the family of utility-privacy functions $g^{(\nu,\mu)}$ for any $\nu,\mu > 1$. It is therefore natural to focus on $\hslash$ in the remainder of this chapter. However, before delving into $\hslash$, we generalize the geometric approach given in Section 3.5 to compute g^ν in the binary symmetric case.

6.3.1 Computation of $g^{(\nu,\nu)}$

In this section, we generalize the geometric approach given in Section 3.5 to derive an expression for $g^\nu(\varepsilon)$ in the special case $P_{X|Y} = \mathsf{BSC}(\alpha)$ and $Y \sim \mathsf{q}_Y = \mathsf{Bernoulli}(q)$ with $q \in [0, \frac{1}{2}]$.

Recall that the function ϕ was defined in (3.14) and it was shown that its upper concave envelope (resp. lower convex envelop) equals G^*_{T} (resp. F^*_{T}), the conjugate of G_{T} (resp. F_{T}), defined in (3.15) (resp. (3.11)). This argument can also be used to characterize $g^\nu(\varepsilon)$ for any $\nu \geq 2$ when $P_{X|Y}$ is a BSC, as shown below.

Define

$$K_\nu(X|Z) := \mathbb{E}\left[\|P_{X|Z}(\cdot|Z)\|_\nu\right] = \sum_{z\in\mathcal{Z}} P_Z(z)\left[\sum_{x\in\mathcal{X}} P^\nu_{X|Z}(x|z)\right]^{1/\nu},$$

and also $K_\nu(X) \coloneqq \|\mathsf{p}_X\|_\nu$ for $\nu > 1$. We may also use $K_\nu(\mathsf{p}_X)$ to denote $K_\nu(X)$. Note that

$$K_\nu(X|Z) = \exp\left\{\frac{1-\nu}{\nu} H_\nu(X|Z)\right\},$$

and also $K_\nu(X) = \exp\left\{\frac{1-\nu}{\nu} H_\nu(X)\right\}$. Note that conditioning reduces entropy, i.e., $H_\nu(X|Z) \le H_\nu(X)$ for $\nu \in [1,\infty]$ [56, Theorem 2]. Note also that the map $x \mapsto \exp\{\frac{1-\nu}{\nu} x\}$ is strictly decreasing, thus we have $K_\nu(X|Z) \ge K_\nu(X)$. For a given $\nu > 1$, $\mathsf{T} = P_{X|Y}$, and $Y \sim \mathsf{q}_Y$, let $F_{\mathsf{T}}^{(\nu)}(\mathsf{q}_Y, \cdot\,) : [K_\nu(Y), 1] \to [K_\nu(X), 1]$ be defined as

$$F_{\mathsf{T}}^{(\nu)}(\mathsf{q}_Y, \mathit{\Delta}) \coloneqq \min_{K_\nu(Y|Z) \ge \mathit{\Delta}} K_\nu(X|Z). \tag{6.9}$$

Having defined $F_{\mathsf{T}}^{(\nu)}$ as above, we can write

$$\max_{H_\nu(Y|Z) \le \kappa} H_\nu(X|Z) = \frac{\nu}{1-\nu} \log F_{\mathsf{T}}^{(\nu)}(\mathsf{q}_Y, \mathit{\Delta}), \tag{6.10}$$

where $\kappa = \frac{\nu}{1-\nu}\log \mathit{\Delta}$. The above expression yields a relationship between $F_{\mathsf{T}}^{(\nu)}$ and the dual of g^ν:

$$\min_{I_\nu(Y;Z) \ge R} I_\nu(X;Z) = H_\nu(X) - \frac{\nu}{1-\nu} \log F_{\mathsf{T}}^{(\nu)}(\mathsf{q}_Y, \mathit{\Delta}), \tag{6.11}$$

where $R = H_\nu(Y) - \frac{\nu}{1-\nu}\log \mathit{\Delta}$. Consequently, analogous to Section 3.5, characterizing g^ν is equivalent to characterizing $F_{\mathsf{T}}^{(\nu)}$. In what follows, we generalize the approach given in Section 3.5 to characterize $F_{\mathsf{T}}^{(\nu)}$.

Recall that $\mathsf{T} = P_{X|Y}$ and $Y \sim \mathsf{q}_Y$ are given. Now consider the $(|\mathcal{Y}|+1)$-dimensional set $\mathcal{S}^{(\nu)} \coloneqq \{(\mathsf{q}, K_\nu(\mathsf{q}), K_\nu((\mathsf{T}\mathsf{q})_{\mathcal{X}})) : \mathsf{q} \in \mathcal{P}_{\mathcal{Y}}\}$, where $(\mathsf{T}\mathsf{q})_{\mathcal{X}} \in \mathcal{P}_{\mathcal{X}}$ is the marginal distribution of X when $Y \sim \mathsf{q}$. Let $\mathcal{C}^{(\nu)}$ be the convex hull of $\mathcal{S}$. It can be shown that $\mathcal{C}^{(\nu)}$

can be characterized as

$$\mathcal{C}^{(\nu)} = \{(\mathsf{q}, K_\nu(Y'|Z), K_\nu(X'|Z)) : Y' \sim \mathsf{q}, X' \sim (\mathsf{T}\mathsf{q})_{\mathcal{X}}, X' \multimap Y' \multimap Z\}$$

Note that if and only if $\mathsf{q} = \mathsf{q}_Y$, then the pair (X', Y') has the same distribution as the given pair (X, Y). It is clear that we can write

$$F_{\mathsf{T}}^{(\nu)}(\mathsf{q}_Y, \mathit{\Delta}) = \min\{\eta : (\mathsf{q}_Y, \mathit{\Delta}, \eta) \in \mathcal{C}^{(\nu)}\} = \min\{\eta : (\mathit{\Delta}, \eta) \in \mathcal{C}_Y^{(\nu)}\},$$

where $\mathcal{C}_Y^{(\nu)} := \mathcal{C}^{(\nu)} \cap \{\mathsf{q} = \mathsf{q}_\mathsf{Y}\}$. Hence, the graph of $F_{\mathsf{T}}^{(\nu)}(\mathsf{q}_Y, \cdot)$ is the lower boundary of the convex set $\mathcal{C}_Y^{(\nu)}$. Although this observation establishes the convexity of $F_{\mathsf{T}}^{(\nu)}(\mathsf{q}_Y, \cdot)$, the function g^ν need not be convex nor concave. Let now $F_{\mathsf{T}}^{(\nu)*}(\mathsf{q}, \cdot) : \mathbb{R} \to \mathbb{R}$ be the conjugate of $F_{\mathsf{T}}^{(\nu)}(\mathsf{q}, \cdot)$, i.e.,

$$F_{\mathsf{T}}^{(\nu)*}(\mathsf{q}, \lambda) := \min\{F_{\mathsf{T}}^{(\nu)}(\mathsf{q}, \mathit{\Delta}) - \lambda\mathit{\Delta} : K_\nu(\mathsf{q}) \leq \mathit{\Delta} \leq 1\}.$$

In fact, the line $\lambda x + F_{\mathsf{T}}^{(\nu)*}(\mathsf{q}, \lambda)$, of slope λ, supports $\mathcal{C}_Y^{(\nu)}$ from below and thus supports the graph of $F_{\mathsf{T}}^{(\nu)}(\mathsf{q}, \cdot)$. We can recover $F_{\mathsf{T}}^{(\nu)}$ from $F_{\mathsf{T}}^{(\nu)*}$ according the following relationship

$$F_{\mathsf{T}}^{(\nu)}(\mathsf{q}, \mathit{\Delta}) = \max\{F_{\mathsf{T}}^{(\nu)*}(\mathsf{q}, \lambda) + \lambda\mathit{\Delta} : \lambda \in \mathbb{R}\}.$$

The above setting coincides exactly with the Witsenhausen and Wyner's setting [147], described in Section 3.5. Consequently, we can invoke the procedure given in Section 3.5 to compute $F_{\mathsf{T}}^{(\nu)}$. For a given λ, define the mapping $\phi^{(\nu)}(\,\cdot\,, \lambda) : \mathcal{P}_{\mathcal{Y}} \to \mathbb{R}$, given by $\phi^{(\nu)}(\mathsf{q}, \lambda) = K_\nu((\mathsf{T}\mathsf{q})_{\mathcal{X}}) - \lambda K_\nu(\mathsf{q})$. The procedure is as follows:

- Fix $\lambda \in \mathbb{R}$ and compute the lower convex envelope of $\phi^{(\nu)}(\cdot, \lambda)$ (i.e., $F_{\mathsf{T}}^{(\nu)*}(\cdot, \lambda)$),

- If a point of the graph of $F_{\mathsf{T}}^{(\nu)*}(\cdot,\lambda)$ can be written as a convex combination of $\phi^{(\nu)}(\mathsf{q}_i,\lambda)$ with weights ω_i, $i\in[k]$ for some $k\geq 2$, then

$$F_{\mathsf{T}}^{(\nu)}\left(\sum_{i=1}^{k}\omega_i\mathsf{q}_i,\sum_{i=1}^{k}\omega_iK_\nu(\mathsf{q}_i)\right)=\sum_{i=1}^{k}\omega_iK_\nu((\mathsf{T}\mathsf{q}_i)_{\mathcal{X}}).$$

- If, for some λ, the function $F_{\mathsf{T}}^{(\nu)*}(\mathsf{q}_Y,\lambda)$ coincides with $\phi(\mathsf{q}_Y,\lambda)$, then this corresponds to a line of slope λ supporting the graph of $F_{\mathsf{T}}^{(\nu)}$ at point $\varDelta=K_\nu(\mathsf{q}_Y)$.

We next apply this procedure for the special case $\mathsf{T}=\mathsf{BSC}(\alpha)$ and $\mathsf{q}_Y=\mathsf{Bernoulli}(q)$ with $\alpha,q\in[0,\frac{1}{2}]$. Let $K_\nu(q)$ be defined as $K_\nu(\mathsf{q})$, where $\mathsf{q}=\mathsf{Bernoulli}(q)$ and also $\phi_{\mathsf{b}}^{(\nu)}(q,\lambda):=K_\nu(\alpha*q)-\lambda K_\nu(q)$. Note that $K_\nu(q)$ is strictly decreasing in q for $q\in[0,\frac{1}{2}]$. Let also $K_\nu^{-1}:[0,1]\to[0,\frac{1}{2}]$ be the functional inverse of K_ν. The following result characterizes g^ν for any $\nu\geq 2$ by characterizing its functional dual in the same spirit that Theorem 3.8 characterized the dual of $g(\varepsilon)$.

Lemma 6.8. *Let $\mathsf{T}=\mathsf{BSC}(\alpha)$ and $\mathsf{q}_Y=\mathsf{Bernoulli}(q)$ with $\alpha,q\in[0,\frac{1}{2}]$. Then, we have for $\nu\geq 2$ and $K_\nu(q)\leq\varDelta\leq 1$*

$$F_{\mathsf{T}}(\mathsf{q}_Y,\varDelta)=K_\nu(K_\nu^{-1}(\varDelta)*\alpha). \tag{6.12}$$

Consequently,

$$\min_{I_\nu(Y;Z)\geq R}I_\nu(X;Z)=\frac{\nu}{1-\nu}\log\frac{K_\nu(q*\alpha)}{K_\nu(K_\nu^{-1}(\varDelta)*\alpha)}, \tag{6.13}$$

where $R=\frac{\nu}{1-\nu}\log\frac{K_\nu(q)}{\varDelta}$.

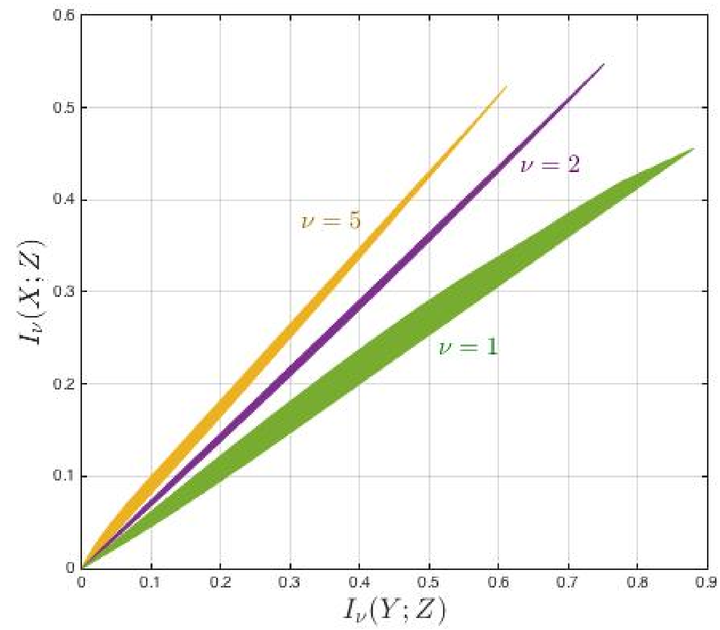


Figure 6.1: The set of achievable pairs $\{(I_\nu(Y;Z), I_\nu(X;Z))\}$, for $P_{X|Y} = \mathsf{BSC}(0.1)$ and $\mathsf{q}_Y = \mathsf{Bernoulli}(0.3)$ in different cases $\nu = 1$, $\nu = 2$ and $\nu = 5$. The lower boundaries of these sets correspond to Theorem 3.8 for $\nu = 1$ and to Lemma 6.8 for $\nu = 2$ and 5.

Proof. First note that (6.13) follows directly from (6.11) and (6.12). To prove (6.12), observe that the second derivative of $\phi_{\mathsf{b}}^{(\nu)}(\cdot, \lambda)$ is given by

$$\frac{\mathrm{d}^2}{\mathrm{d}p^2}\phi_{\mathsf{b}}^{(\nu)}(p,\lambda) = (\nu-1)(1-2\alpha)^2 A(r)(C(r)-B(r)) - \lambda(\nu-1)A(p)(C(p)-B(p)), \quad (6.14)$$

where $r := \alpha * p$ and $A(r) := (r^\nu + \bar{r}^\nu)^{1/\nu-2}$, $B(r) := (r^{\nu-1} - \bar{r}^{\nu-1})^2$, and $C(r) := (r^\nu + \bar{r}^\nu)(r^{\nu-2} + \bar{r}^{\nu-2})$. This expression can be shown to be positive if $\lambda \le (1-2\alpha)^2$ and $\nu \ge 2$. For $\lambda \ge (1-2\alpha)^2$ and $\nu \ge 2$, the right-hand side of (6.14) is negative on an interval $[p_{\lambda,\nu}, \bar{p}_{\lambda,\nu}]$ symmetric about $p = \frac{1}{2}$ and is positive elsewhere with the local maximum at $p = \frac{1}{2}$ (Fig. 6.2). Therefore, we only need to focus on the interval $[p_{\lambda,\nu}, \bar{p}_{\lambda,\nu}]$. It can be verified that we have $\frac{\mathrm{d}}{\mathrm{d}p}\phi_{\mathsf{b}}^{(\nu)}(p_{\lambda,\nu}, \lambda) = \frac{\mathrm{d}}{\mathrm{d}p}\phi_{\mathsf{b}}^{(\nu)}(\bar{p}_{\lambda,\nu}, \lambda) = 0$. By symmetry, the lower convex

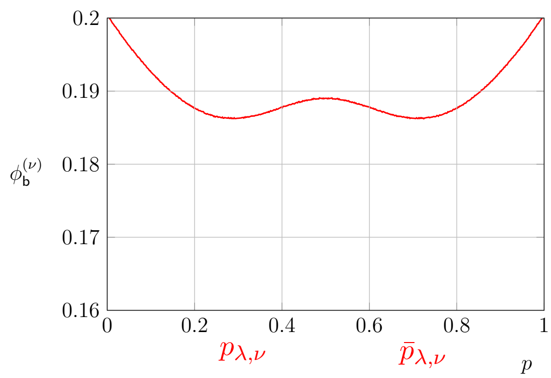


Figure 6.2: The function $\phi_{\mathsf{b}}^{(3)}(p, 0.7)$, where $\mathsf{T} = \mathsf{BSC}(0.1)$ and $\mathsf{q}_Y = \mathsf{Bernoulli}(0.3)$.

envelope of the graph $\phi_{\mathsf{b}}^{(\nu)}(\cdot, \lambda)$ is obtained by replacing p in $\phi_{\mathsf{b}}^{(\nu)}(p, \lambda)$ on the interval $[p_{\lambda,\nu}, \bar{p}_{\lambda,\nu}]$ by $p_{\lambda,\nu}$. Therefore, for the given $q \leq \frac{1}{2}$, if $p_{\lambda,\nu} \leq q$, then $(q, F_{\mathsf{T}}^{(\nu)*}(\mathsf{q}_Y, \lambda))$ is a convex combination of $(p_{\lambda,\nu}, \phi_{\mathsf{b}}^{(\nu)}(p_{\lambda,\nu}, \lambda)$ and $(\bar{p}_{\lambda,\nu}, \phi_{\mathsf{b}}^{(\nu)}(\bar{p}_{\lambda,\nu}, \lambda)$ with weights ω and $\bar{\omega}$. Hence, we can write

$$F_{\mathsf{T}}^{(\nu)}(\mathsf{q}_Y, \omega K_\nu(p_{\lambda,\nu}) + \bar{\omega} K_\nu(\bar{p}_{\lambda,\nu})) = \omega K_\nu(p_{\lambda,\nu} * \alpha) + \bar{\omega} K_\nu(\bar{p}_{\lambda,\nu} * \alpha),$$

where ω satisfies $q = \omega p_{\lambda,\nu} + \bar{\omega}\bar{p}_{\lambda,\nu}$. Note that $K_\nu(p_{\lambda,\nu}) = K_\nu(\bar{p}_{\lambda,\nu})$ and also $K_\nu(p_{\lambda,\nu} * \alpha) = K_\nu(\bar{p}_{\lambda,\nu} * \alpha)$. Thus, denoting $p_{\lambda,\nu}$ by p, we conclude that $F_{\mathsf{T}}^{(\nu)}(\mathsf{q}_Y, K_\nu(p)) = K_\nu(p * \alpha)$ for $0 \leq p \leq q$. $\square$

6.3.2 Geometric Properties of $\hbar$

First, note that $\mathsf{P_c}(X|YZ) \geq \mathsf{P_c}(X|Z) \geq \mathsf{P_c}(X)$ for random variables X, Y and Z. Therefore from (6.7) we have $\mathsf{P_c}(Y) \leq \hbar(\varepsilon) \leq 1$, and $\hbar(\varepsilon) = 1$ if and only if $\varepsilon \geq \mathsf{P_c}(X|Y)$. Thus it is enough to study $\hbar(\cdot)$ over the interval $[\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$. An application of the Support Lemma [40, Lemma 15.4] shows that it is enough to consider random variables Z supported on $\mathcal{Z} = [N+1]$. Thus, the privacy filter $P_{Z|Y}$ can be realized by an $N \times (N+1)$

stochastic matrix $F \in \mathcal{M}_{N\times(N+1)}$, where $\mathcal{M}_{N\times M}$ denotes the set of all real-valued $N \times M$ matrices. Let $\mathcal{F}$ be the set of all such matrices. Then both utility $\mathcal{U}(P,F) = \mathsf{P_c}(Y|Z)$ and privacy $\mathcal{P}(P,F) = \mathsf{P_c}(X|Z)$ are functions of $F \in \mathcal{F}$ and can be written as

$$\mathcal{P}(P,F) := \sum_{z=1}^{N+1} \max_{1\leq x\leq M} \sum_{y=1}^{N} P(x,y)F(y,z), \qquad \mathcal{U}(P,F) := \sum_{z=1}^{N+1} \max_{1\leq y\leq N} q(y)F(y,z). \tag{6.15}$$

In particular, we can express $\hbar(\varepsilon)$ as

$$\hbar(\varepsilon) = \sup_{\substack{F\in\mathcal{F},\\ \mathcal{P}(P,F)\leq\varepsilon}} \mathcal{U}(P,F). \tag{6.16}$$

As before, consider P fixed and omit it in $\mathcal{U}(P,F)$ and $\mathcal{P}(P,F)$ when there is no risk of confusion. It is straightforward to verify that $\mathcal{P}$ and $\mathcal{U}$ are continuous and convex on $\mathcal{F}$. On the other hand, we show in the following theorem that $\hbar$ is concave and continuous on $[\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$ and consequently, for every $\varepsilon \in [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$ there exists $G \in \mathcal{F}$ such that $\mathcal{P}(G) = \varepsilon$ and $\mathcal{U}(G) = \hbar(\varepsilon)$.

Theorem 6.9. *The mapping $\varepsilon \mapsto \hbar$ is concave on* $[\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$.

Proof. This result can be proved using a proof technique similar to [147, Theorem 2.3] (see Section 3.5). However, we provide an easier proof based on the random filter argument presented in the proof of Theorem3.2. Let $P_{Z_1|Y} : Y \to Z_1$ and $P_{Z_2|Y} : Y \to Z_2$ be two optimal privacy filters with disjoint output alphabets $\mathcal{Z}_1$ and $\mathcal{Z}_2$, and corresponding privacy levels ε_1 and ε_2, respectively. We introduce an auxiliary binary random variable $U \sim \mathsf{Bernoulli}(\lambda)$, independent of (X,Y), for some $\lambda \in [0,1]$ and define the following random privacy filter $P_{Z_\lambda|Y}$: We pick $P_{Z_2|Y}$ if $U = 1$ and $P_{Z_1|Y}$ if $U = 0$. Note that U is a

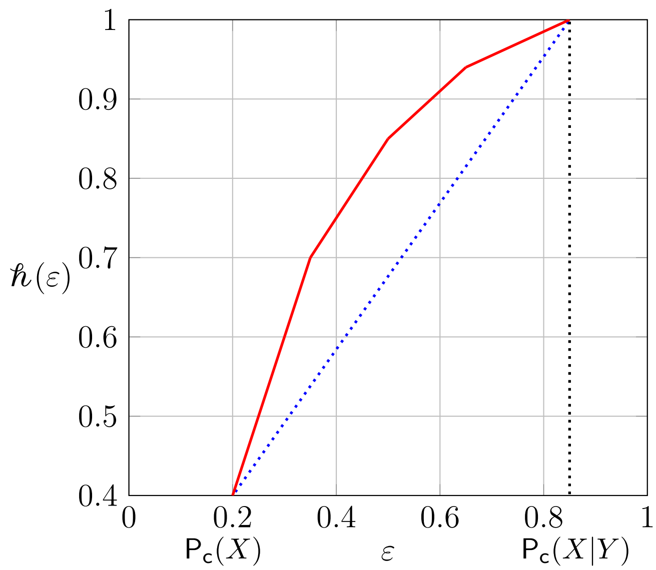


Figure 6.3: Typical $\hbar$ and its trivial lower bound, the chord connecting $(\mathsf{P_c}(X), \hbar(\mathsf{P_c}(X)))$ and $(\mathsf{P_c}(X|Y), 1)$.

deterministic function of Z_λ. Then, we have

$$\mathsf{P_c}(X|Z_\lambda) = \mathsf{P_c}(X|Z_\lambda, U) = \bar{\lambda}\mathsf{P_c}(X|Z_1) + \lambda\mathsf{P_c}(X|Z_2) = \bar{\lambda}\varepsilon_1 + \lambda\varepsilon_2.$$

Analogously, we obtain $\mathsf{P_c}(Y|Z_\lambda) = \bar{\lambda}\hbar(\varepsilon_1) + \lambda\hbar(\varepsilon_2)$. Since $\hbar(\bar{\lambda}\varepsilon_1 + \lambda\varepsilon_2) \geq \mathsf{P_c}(Y|Z_\lambda)$, the result immediately follows. $\square$

The following theorem states that $\hbar$ is a piecewise linear function, as illustrated in Fig. 6.3.

Theorem 6.10. *The function $\hbar : [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)] \to \mathbb{R}^+$ is piecewise linear, i.e., there exist $K \geq 1$ and thresholds $\mathsf{P_c}(X) = \varepsilon_0 \leq \varepsilon_1 \leq \ldots \leq \varepsilon_K = \mathsf{P_c}(X|Y)$ such that $\hbar$ is linear on $[\varepsilon_{i-1}, \varepsilon_i]$ for all $i \in [K]$.*

The proof of this theorem, which is given in Appendix C.1, relies on the geometric formulation of $\hbar$. In particular, it is proved that $\mathcal{P}$ and $\mathcal{U}$, are piecewise linear functions in

$\mathcal{F}$. Using this fact, we establish the existence of a piecewise linear path of *optimal filters* in $\mathcal{F}$. The proof technique allows us to derive the slope of $\hbar$ on $[\varepsilon_{i-1}, \varepsilon_i]$, given the family of optimal filters at a single point $\varepsilon \in [\varepsilon_{i-1}, \varepsilon_i]$. For example, since the family of optimal filters at $\varepsilon = \mathsf{P_c}(X|Y)$ is easily obtainable in the binary case, it is possible to compute $\hbar$ on the last interval. We utilize this observation in Section 6.3.4 to prove that in the binary case $\hbar$ is indeed linear.

6.3.3 Perfect Privacy

When $\varepsilon = \mathsf{P_c}(X)$, observing Z does not increase the probability of guessing X. In this case we say that perfect privacy holds. An interesting problem is to characterize when non-trivial utility can be obtained under perfect privacy, that is, to characterize when $\hbar(\mathsf{P_c}(X)) > \mathsf{P_c}(Y)$ holds. To the best of our knowledge, a general necessary and sufficient condition for this requirement is unknown.

Notice that $\hbar(\mathsf{P_c}(X)) > \mathsf{P_c}(Y)$ is equivalent to $g^{\infty}(0) > 0$. As opposed to $I_\nu(X;Z)$ with $1 \leq \nu < \infty$, $I_\infty(X;Z) = 0$ does not necessarily imply that $X \perp\!\!\!\perp Z$. In particular, the *weak independence* arguments from Chapter 3 cannot be applied for g^{∞}. However, we have the following.

Proposition 6.11. *Let (X, Z) be a pair of random variables with X uniformly distributed. If $I_\infty(X;Z) = 0$, then $X \perp\!\!\!\perp Z$.*

This proposition follows easily from the following lemma.

Lemma 6.12. *If X is uniformly distributed, then the mapping $\nu \mapsto I_\nu(X;Z)$ is non-decreasing on $[1, \infty]$. In particular, $I(X;Z) \leq I_\infty(X;Z)$.*

Proof. From the definition of Arimoto's mutual information, we can write

$$I_\nu(X;Z) = -\frac{\nu}{1-\nu}\log\sum_{z\in\mathcal{Z}}\left[\sum_{x\in\mathcal{X}} r_\nu(x)P^{\nu}_{Z|X}(z|x)\right]^{\frac{1}{\nu}},$$

where $r_\nu(x) := \frac{\mathsf{p}^{\nu}_X(x)}{\sum_{x'\in\mathcal{X}}\mathsf{p}^{\nu}_X(x')}$. Since X is uniformly distributed, $r_\nu(x) = \mathsf{p}_X(x) = \frac{1}{M}$. Thus we obtain

$$I_\nu(X;Z) = \frac{E_0(\rho, p_X, P_{Z|X})}{\rho},$$

where $\rho := \frac{1-\nu}{\nu}$, and for any channel W with input distribution Q, $E_0(\rho, Q, \mathsf{W})$ is *Gallager's error exponent function* [60], defined as

$$E_0(\rho, Q, \mathsf{W}) := -\log\sum_{z\in\mathcal{Z}}\left[\sum_{x\in\mathcal{X}} Q(x)\mathsf{W}^{\frac{1}{1+\rho}}(z|x)\right]^{1+\rho}.$$

Arimoto [12] showed that for any Q and W fixed, the mapping $\rho\mapsto\frac{E_0(\rho,Q,\mathsf{W})}{\rho}$ is decreasing. Since ρ is decreasing in ν, the result follows. $\square$

As a consequence of Proposition 6.11, when X and Y are uniformly distributed, one can apply the weak independence arguments from Chapter 3 to obtain the following.

Corollary 6.13. *If X and Y are uniformly distributed, then $g^{\infty}(0) > 0$ if and only if X is weakly independent of Y.*

When X is uniform, the privacy requirement $I_\infty(X;Z)\le\varepsilon$ guarantees that an adversary observing Z cannot efficiently estimate any arbitrary *randomized function* of X. To see this, consider a random variable U which satisfies $U \multimap X \multimap Z$. Then we have

$$\mathsf{P_c}(U|Z) = \sum_{z\in\mathcal{Z}}\max_{u\in\mathcal{U}}\sum_{x\in\mathcal{X}} P_{UX}(u,x)P_{Z|X}(z|x)$$

$$
\begin{aligned}
&\leq \sum_{z\in\mathcal{Z}}\left(\max_{x\in\mathcal{X}} P_{Z|X}(z|x)\right)\left(\max_{u\in\mathcal{U}}\sum_{x\in\mathcal{X}} P_{UX}(u,x)\right)\\
&= \frac{\mathsf{P_c}(X|Z)\mathsf{P_c}(U)}{\mathsf{P_c}(X)},
\end{aligned}
$$

which can be rearranged to yield $I_\infty(U;Z) \leq I_\infty(X;Z)$. It is worth mentioning that the data processing inequality for I_∞ [56] states that $I_\infty(Z;U) \leq I_\infty(Z;X)$. However, $I_\infty(Z;U)$ is not necessarily equal to $I_\infty(U;Z)$.

6.3.4 Binary Case

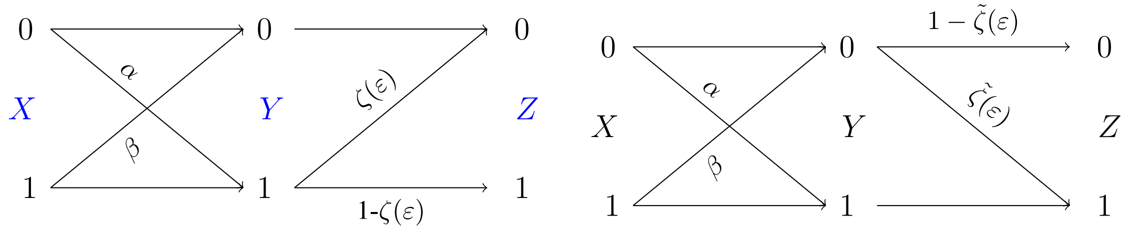
A channel W is called a binary input binary output channel with crossover probabilities α and β, denoted by $\mathsf{BIBO}(\alpha,\beta)$, if $\mathsf{W}(\cdot|0) = (\bar{\alpha},\alpha)$ and $\mathsf{W}(\cdot|1) = (\beta,\bar{\beta})$. Notice that if $X \sim \mathsf{Bernoulli}(p)$ with $p \in [\frac{1}{2},1)$ and $P_{Y|X} = \mathsf{BIBO}(\alpha,\beta)$ with $\alpha,\beta \in [0,\frac{1}{2})$, then $\mathsf{P_c}(X) = p$ and $\mathsf{P_c}(X|Y) = \max\{\bar{\alpha}\bar{p},\beta p\} + \bar{\beta}p$. In this case, if $\bar{\alpha}\bar{p} \leq \beta p$ then $\mathsf{P_c}(X|Y) = p = \mathsf{P_c}(X)$ and hence $\hbar(p) = 1$. The following theorem, whose proof is given in Appendix C.2, establishes the linear behavior of $\hbar$ in the non-trivial case $\bar{\alpha}\bar{p} > \beta p$.

Theorem 6.14. *Let $X \sim \mathsf{Bernoulli}(p)$ with $p \in [\frac{1}{2},1)$ and $P_{Y|X} = \mathsf{BIBO}(\alpha,\beta)$ with $\alpha,\beta \in [0,\frac{1}{2})$ such that $\bar{\alpha}\bar{p} > \beta p$. Then, for any $\varepsilon \in [p,\bar{\alpha}\bar{p}+\bar{\beta}p] = [\mathsf{P_c}(X),\mathsf{P_c}(X|Y)]$,*

$$
\hbar(\varepsilon) = \begin{cases} 1-\zeta(\varepsilon)q, & \alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2, \\ 1-\tilde{\zeta}(\varepsilon)\bar{q}, & \alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2, \end{cases}
$$

where $q := q_Y(1) = \alpha\bar{p}+\bar{\beta}p$,

$$
\zeta(\varepsilon) := \frac{\bar{\alpha}\bar{p}+\bar{\beta}p-\varepsilon}{\bar{\beta}p-\alpha\bar{p}}, \quad \text{and} \quad \tilde{\zeta}(\varepsilon) := \frac{\bar{\alpha}\bar{p}+\bar{\beta}p-\varepsilon}{\bar{\alpha}\bar{p}-\beta p}. \tag{6.17}
$$



(a) $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$.　　(b) $\alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2$.

Figure 6.4: Optimal privacy mechanisms in Theorem 6.14.

Furthermore, the Z-channel $\mathsf{Z}(\zeta(\varepsilon))$ *and the reverse Z-channel* $\tilde{\mathsf{Z}}(\tilde{\zeta}(\varepsilon))$ *achieve* $\hbar(\varepsilon)$ *when* $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$ *and* $\alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2$*, respectively. The optimal privacy filters are depicted in Fig. 6.4.*

Note that the condition $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$ can be equivalently written as $P_{X|Y}(0|0) < P_{X|Y}(1|1)$. Consequently, when $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$, the event $Y = 1$ reveals more useful information about X and hence it needs to be distorted to maintain the privacy of X.

Under the hypotheses of Theorem 6.14, there exists a Z-channel for every $\varepsilon \in [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$ that achieves $\hbar(\varepsilon)$. A minor modification to the proof of Theorem 6.14 shows that the Z-channel is the only binary channel with this property. It must be noted that even in the symmetric case (i.e., $\alpha = \beta$), the optimal filter cannot be a symmetric channel for $p \in (\frac{1}{2}, 1)$. However, when $\alpha = \beta$ and $p = \frac{1}{2}$, the channel $\mathsf{BSC}(0.5\zeta(\varepsilon))$ can be easily shown to be an optimal privacy filter for every $\varepsilon \in [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$.

It is straightforward to show that $1-\zeta(p)q > \bar{q}$ if and only if $p \in (\frac{1}{2}, 1)$, and $1-\zeta(p)q > q$ if and only if $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$. In particular, we have the following necessary and sufficient condition for the non-trivial utility under perfect privacy.

Corollary 6.15. *Let* $X \sim \mathsf{Bernoulli}(p)$ *with* $p \in [\frac{1}{2}, 1)$ *and* $P_{Y|X} = \mathsf{BIBO}(\alpha, \beta)$ *with* $\alpha, \beta \in [0, \frac{1}{2})$ *such that* $\bar{\alpha}\bar{p} > \beta p$*. Then* $g^{\infty}(0) > 0$ *if and only if* $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$ *and*

$p \in (\frac{1}{2}, 1)$.

6.3.5 A variant of $\hbar$

Thus far, we studied the privacy-constrained guessing probability $\hbar$ where no constraint on the cardinality of the displayed data Z is imposed (other than being finite). Nevertheless, it was shown that it is sufficient to consider $\mathcal{Z}$ with cardinality $|\mathcal{Y}|+1 = N+1$. However, this condition may be practically inconvenient. Moreover, for the scalar binary case examined in the last section we showed that a binary Z was sufficient to achieve $\hbar(\varepsilon)$ for any $\varepsilon \in [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$. Hence to simplify the constrained optimization problem involved, it is natural to require that $|\mathcal{Z}| = |\mathcal{Y}| = N$, which leads to the following variant quantity of $\hbar$, denoted by $\underline{\hbar}$.

Definition 6.16. *For arbitrary discrete random variables X and Y supported on $\mathcal{X}$ and $\mathcal{Y}$ respectively, let $\underline{\hbar} : [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)] \to \mathbb{R}^+$ be defined by*

$$\underline{\hbar}(\varepsilon) := \sup_{P_{Z|Y} \in \underline{\mathfrak{D}}_\varepsilon} \mathsf{P_c}(Y|Z),$$

where $\underline{\mathfrak{D}}_\varepsilon := \left\{P_{Z|Y} : \mathcal{Z} = \mathcal{Y}, X -\!\!\circ\!\!- Y -\!\!\circ\!\!- Z, \mathsf{P_c}(X|Z) \le \varepsilon\right\}$.

Unlike $\hbar$, the definition of $\underline{\hbar}$ requires $\mathcal{Z} = \mathcal{Y}$. This difference makes the tools from [147] unavailable. In particular, the concavity and hence the piecewise linearity of $\hbar$ do not carry over to $\underline{\hbar}$. However, we have the following theorem for $\underline{\hbar}$ whose proof is given in Appendix C.3. For $(y_0, z_0) \in \mathcal{Y} \times \mathcal{Y}$, a channel W is said to be an N-ary Z-channel with crossover probability γ from y_0 to z_0, denoted by $\mathsf{Z}^{y_0,z_0}(\gamma)$, if the input and output alphabets are $\mathcal{Y}$ and $\mathsf{W}(y|y) = 1$ for $y \neq y_0$, $\mathsf{W}(z_0|y_0) = \gamma$, and $\mathsf{W}(y_0|y_0) = \bar{\gamma}$. We also let $\underline{\hbar}'(\mathsf{P_c}(X|Y))$ denote the left derivative of $\underline{\hbar}(\cdot)$ evaluated at $\varepsilon = \mathsf{P_c}(X|Y)$. For notational convenience, we adopt the convention $\frac{x}{0} = +\infty$ for $x > 0$.

Theorem 6.17. *Let X and Y be discrete random variables. If $\mathsf{P_c}(X) < \mathsf{P_c}(X|Y)$, then there exists $\varepsilon_{\mathsf{L}} < \mathsf{P_c}(X|Y)$ such that $\underline{\hbar}$ is linear on $[\varepsilon_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$. In particular, for every $\varepsilon \in [\varepsilon_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$,*

$$\underline{\hbar}(\varepsilon) = 1 - (\mathsf{P_c}(X|Y) - \varepsilon)\underline{\hbar}'(\mathsf{P_c}(X|Y)). \tag{6.18}$$

Moreover, if $q_Y(y) > 0$ for all $y \in \mathcal{Y}$ and if there exists (a unique) $x_y \in \mathcal{X}$ for each $y \in \mathcal{Y}$ such that $\Pr(X = x_y|Y = y) > \Pr(X = x|Y = y)$ for all $x \neq x_y$, then

$$\underline{\hbar}'(\mathsf{P_c}(X|Y)) = \min_{(y,z)\in\mathcal{Y}\times\mathcal{Y}} \frac{q_Y(y)}{\Pr(X = x_y, Y = y) - \Pr(X = x_z, Y = y)}. \tag{6.19}$$

In addition, if $(y_0, z_0) \in \mathcal{Y} \times \mathcal{Y}$ attains the minimum in (6.19), then there exists $\varepsilon_{\mathsf{L}}^{y_0,z_0} < \mathsf{P_c}(X|Y)$ such that $\mathsf{Z}^{y_0,z_0}(\zeta^{y_0,z_0}(\varepsilon))$ achieves $\underline{\hbar}(\varepsilon)$ for every $\varepsilon \in [\varepsilon_{\mathsf{L}}^{y_0,z_0}, \mathsf{P_c}(X|Y)]$, where

$$\zeta^{y_0,z_0}(\varepsilon) = \frac{\mathsf{P_c}(X|Y) - \varepsilon}{\Pr(X = x_{y_0}, Y = y_0) - \Pr(X = x_{z_0}, Y = y_0)}.$$

Although (6.18) establishes the linear behavior of $\underline{\hbar}$ over $[\varepsilon_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$ for general X and Y, a priori it is not clear how to obtain $\underline{\hbar}'(\mathsf{P_c}(X|Y))$. Under the assumptions of Theorem 6.17, (6.19) expresses $\underline{\hbar}'(\mathsf{P_c}(X|Y))$ as the minimum of *finitely* many numbers, and a suitable Z-channel achieves $\underline{\hbar}$ for ε close to $\mathsf{P_c}(X|Y)$. As we will see in the following section, these assumptions are rather general and allow us to derive a closed form expression for $\underline{\hbar}$ for a pair of binary random *vectors* (X^n, Y^n) with $X^n, Y^n \in \{0, 1\}^n$.

6.4 Binary Vector Case

We next study privacy aware guessing for a pair of binary random vectors (X^n, Y^n). First note that since having more side information only improves the probability of correct guessing, one can write $\mathsf{P_c}(X^n) \leq \mathsf{P_c}(X^n|Z^n) \leq \mathsf{P_c}(X^n|Y^n, Z^n) = \mathsf{P_c}(X^n|Y^n)$ for $X^n \multimap Y^n \multimap Z^n$ and thus, we can restrict ε^n in the following definition to $[\mathsf{P_c}(X^n), \mathsf{P_c}(X^n|Y^n))]$.

Definition 6.18. *For a given pair of binary random vectors (X^n, Y^n), the function $\underline{\hbar}_n$ is defined, for $\varepsilon \in [\mathsf{P_c}^{1/n}(X^n), \mathsf{P_c}^{1/n}(X^n|Y^n)]$, as*

$$\underline{\hbar}_n(\varepsilon) \coloneqq \sup_{P_{Z^n|Y^n} \in \underline{\mathfrak{D}}_{n,\varepsilon}} \mathsf{P_c}^{1/n}(Y^n|Z^n) \tag{6.20}$$

where $\underline{\mathfrak{D}}_{n,\varepsilon} = \{P_{Z^n|Y^n} : \mathcal{Z}^n = \{0,1\}^n, X^n \multimap Y^n \multimap Z^n, \mathsf{P_c}^{1/n}(X^n|Z^n) \leq \varepsilon\}$.

Notice that this definition does not make any assumption about the privacy filters $P_{Z^n|Y^n}$ apart from $\mathcal{Z}^n = \{0,1\}^n$. Nonetheless, this restriction makes the functional properties of $\underline{\hbar}_n$ different from those of $\hbar$.

In order to study $\underline{\hbar}_n$, we consider the following two scenarios for (X^n, Y^n):

(a$_1$) $X_1, \ldots, X_n$ are i.i.d. samples drawn from $\mathsf{Bernoulli}(p)$,

(a$_2$) $X_1 \sim \mathsf{Bernoulli}(p)$ and $X_k = X_{k-1} \oplus U_k$ for $k = 2, \ldots, n$, where $U_2, \ldots, U_n$ are i.i.d. samples drawn from $\mathsf{Bernoulli}(r)$ and independent of X_1, where $\oplus$ denotes mod 2 addition,

and in both cases, we assume that

(b) $Y_k = X_k \oplus V_k$ for $k \in [n]$, where $V_1, \ldots, V_n$ are i.i.d. samples drawn from $\mathsf{Bernoulli}(\alpha)$ and independent of X^n.

We study $\underline{\hbar}_n$ for (X^n, Y^n) satisfying the assumptions (a$_1$) and (b) in Section 6.4.1 and for (X^n, Y^n) satisfying the assumptions (a$_2$) and (b) in Section 6.4.2. We study $\underline{\hbar}_n$ in the special case $r = 0$ in more detail.

6.4.1 I.I.D. Case

Here, we assume that (X^n, Y^n) satisfy (a$_1$) and (b) and apply Theorem 6.17 to derive a closed form expression for $\underline{\hbar}_n(\varepsilon)$ for ε close to $\mathsf{P_c}(X^n|Y^n)$. Additionally, we determine an optimal filter in the same regime.

We begin by identifying the domain $[\mathsf{P_c}(X^n), \mathsf{P_c}(X^n|Y^n)]$ of $\underline{\hbar}$ in the following lemma, whose proof follows directly from the definition of $\mathsf{P_c}$.

Lemma 6.19. *Assume that $(X_1, Z_1), \ldots, (X_n, Z_n)$ are independent pairs of random variables. Then*

$$\mathsf{P_c}(X^n|Z^n) = \prod_{k=1}^{n} \mathsf{P_c}(X_k|Z_k).$$

Thus, according to this lemma, if $p \in [\frac{1}{2}, 1)$ and $\alpha \in [0, \frac{1}{2})$ then $\mathsf{P_c}(X^n) = p^n$ and $\mathsf{P_c}(X^n|Y^n) = \bar{\alpha}^n$. The following proposition, whose proof is given in Appendix C.4, is a straightforward consequence of Theorem 6.17. A channel W is said to be a 2^n-ary Z-channel with crossover probability γ, denoted by $\mathsf{Z}_n(\gamma)$, if its input and output alphabets are $\{0,1\}^n$ and it is $\mathsf{Z}^{\mathbf{1},\mathbf{0}}(\gamma)$, where $\mathbf{0} = (0, 0, \ldots, 0)$ and $\mathbf{1} = (1, 1, \ldots, 1)$.

Theorem 6.20. *Assume that (X^n, Y^n) satisfy (a$_1$) and (b) with $p \in [\frac{1}{2}, 1)$ and $\alpha \in [0, \frac{1}{2})$ such that $\bar{\alpha} > p$. Then there exists $\varepsilon_{\mathsf{L}} < \bar{\alpha}$ such that, for all $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$,*

$$\underline{\hbar}_n^n(\varepsilon) = 1 - \zeta_n(\varepsilon) q^n$$

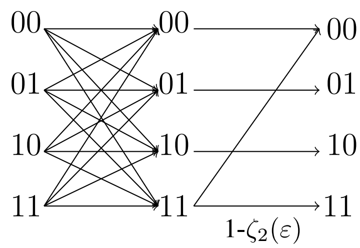


Figure 6.5: The optimal filter for $\underline{\hbar}_2(\varepsilon)$ for $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$.

where $q = \alpha\bar{p} + \bar{\alpha}p$ *and*

$$\zeta_n(\varepsilon) := \frac{\bar{\alpha}^n - \varepsilon^n}{(\bar{\alpha}p)^n - (\alpha\bar{p})^n}.$$

Moreover, the 2^n*-ary Z-channel* $\mathsf{Z}_n(\zeta_n(\varepsilon))$ *achieves* $\underline{\hbar}_n(\varepsilon)$ *in this interval.*

The optimal privacy filter achieving $\underline{\hbar}_2(\varepsilon)$ is depicted in Fig. 6.5. From an implementation point of view, the simplest privacy filter is a memoryless filter such that Z_k is a noisy version of Y_k for each $k \in [n]$. This privacy mechanism generates Z_k, given Y_k, using a *single* BIBO channel W, and thus

$$P_{Z^n|Y^n}(z^n|y^n) = \prod_{k=1}^{n} \mathsf{W}(z_k|y_k). \tag{6.21}$$

Now, let $\hbar_n^{\mathsf{i}}(\varepsilon) = \sup \mathsf{P_c}^{1/n}(Y^n|Z^n)$, where the supremum is taken over all $P_{Z^n|Y^n}$ satisfying (6.21) and $\mathsf{P_c}^{1/n}(X^n|Z^n) \leq \varepsilon$. Clearly, $\hbar_n^{\mathsf{i}}(\varepsilon) \leq \underline{\hbar}_n(\varepsilon)$ for all $\varepsilon \in [\mathsf{P_c}^{1/n}(X^n), \mathsf{P_c}^{1/n}(X^n|Y^n)]$. The following proposition shows that if we restrict the privacy filter $P_{Z^n|Y^n}$ to be memoryless, then the optimal filter coincides with the optimal filter in the scalar case, which in this case is $\mathsf{Z}(\zeta(\varepsilon))$, defined in Theorem 6.14.

Proposition 6.21. *Assume that* (X^n, Y^n) *satisfy* (a_1) *and* (b) *with* $p \in [\frac{1}{2}, 1)$ *and* $\alpha \in [0, \frac{1}{2})$

such that $\bar{\alpha} > p$. Then, for all $\varepsilon \in [p, \bar{\alpha}]$,

$$\hbar_n^{\mathsf{i}}(\varepsilon) = \hbar(\varepsilon) = 1 - \zeta(\varepsilon)q,$$

where $q = \alpha\bar{p} + \bar{\alpha}p$ and $\zeta(\varepsilon) = \dfrac{\bar{\alpha}\bar{p} + \bar{\alpha}p - \varepsilon}{\bar{\alpha}p - \alpha\bar{p}}$.

Proof. For any privacy filter satisfying (6.21), (X^n, Z^n) and (Y^n, Z^n) are i.i.d. By Lemma 6.19, we have $\mathsf{P_c}(X^n|Z^n) = (\mathsf{P_c}(X|Z))^n$ and $\mathsf{P_c}(Y^n|Z^n) = (\mathsf{P_c}(Y|Z))^n$ where (X, Y, Z) has the common distribution of $\{(X_k, Y_k, Z_k)\}_{k=1}^n$. In particular,

$$\hbar_n^{\mathsf{i}}(\varepsilon) = \sup_{\mathsf{P_c}^{1/n}(X^n|Z^n) \leq \varepsilon} \mathsf{P_c}^{1/n}(Y^n|Z^n) = \sup_{\mathsf{P_c}(X|Z) \leq \varepsilon} \mathsf{P_c}(Y|Z),$$

where the first supremum assumes (6.21) and the second supremum is implicitly constrained to $\mathcal{Z} = \{0, 1\}$. The result then follows from Theorem 6.14. □

It must be noted that, despite the fact that (X^n, Y^n) is i.i.d., the memoryless privacy filter associated to $\hbar_n^{\mathsf{i}}(\varepsilon)$ is not optimal, as $\underline{\hbar}_n(\varepsilon)$ is a function of n while $\hbar_n^{\mathsf{i}}(\varepsilon)$ is not. The following corollary, whose proof is given in Appendix C.5, bounds the loss resulting from using a memoryless filter instead of an optimal one for $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$. Clearly, for $n = 1$, there is no gap as $\underline{\hbar}_1(\varepsilon) = \hbar(\varepsilon) = \hbar_1^{\mathsf{i}}(\varepsilon)$.

Corollary 6.22. *Let (X^n, Y^n) satisfy (a_1) and (b) with $p \in [\frac{1}{2}, 1)$ and $\alpha \in [0, \frac{1}{2})$ such that $\bar{\alpha} > p$. Let ε_{L} be as in Theorem 6.20. If $p > \frac{1}{2}$ and $\alpha > 0$, then for $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$ and sufficiently large n*

$$\underline{\hbar}_n(\varepsilon) - \hbar_n^{\mathsf{i}}(\varepsilon) \geq (\bar{\alpha} - \varepsilon)[\Phi(1) - \Phi(n)], \tag{6.22}$$

where $q = \alpha\bar{p} + \bar{\alpha}p$ and

$$\Phi(n) := \frac{q^n \bar{\alpha}^{n-1}}{(\bar{\alpha}p)^n - (\alpha\bar{p})^n}.$$

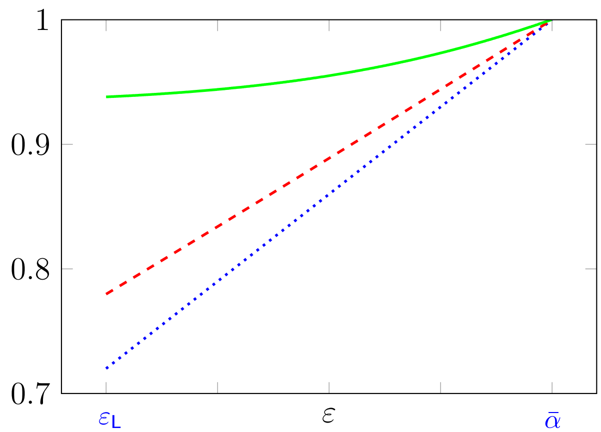


Figure 6.6: The graphs of $\underline{\hbar}_{10}(\varepsilon)$ (green solid curve), $\underline{\hbar}_{2}(\varepsilon)$ (red dashed curve), and $\hbar^{\mathsf{i}}_{2}(\varepsilon) = \hbar^{\mathsf{i}}_{10}(\varepsilon)$ (blue dotted line) given in Proposition 6.21 and Theorem 6.20 for i.i.d. (X^n, Y^n) with $X \sim \mathsf{Bernoulli}(0.6)$ and $P_{Y|X} = \mathsf{BSC}(0.2)$.

If $p = \frac{1}{2}$, *then*

$$\hbar^{\mathsf{i}}_{n}(\varepsilon) \leq \underline{\hbar}_{n}(\varepsilon) \leq \hbar^{\mathsf{i}}_{n}(\varepsilon) + \frac{\alpha}{2\bar{\alpha}}, \tag{6.23}$$

for every $n \geq 1$ *and* $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$.

Note that $\Phi(n) \downarrow 0$ as $n \to \infty$. Thus (6.22) implies that, as expected, the gap between the performance of the optimal privacy filter and that of the optimal memoryless privacy filter increases as n increases. This observation is numerically illustrated in Fig. 6.6, where $\underline{\hbar}_{n}(\varepsilon)$ is plotted as a function of ε for $n = 2$ and $n = 10$.

Moreover, (6.23) implies that when $p = \frac{1}{2}$ and α is small, $\underline{\hbar}_{n}(\varepsilon)$ can be approximated by $\hbar^{\mathsf{i}}_{n}(\varepsilon)$. Thus, we can approximate the optimal filter $\mathsf{Z}_n(\zeta_n(\varepsilon))$ with a simple memoryless filter given by $Z_k = Y_k \oplus W_k$, where $W_1, \ldots, W_n$ are i.i.d. $\mathsf{Bernoulli}(0.5\zeta(\varepsilon))$ random variables that are independent of (X^n, Y^n).

6.4.2 Markov Private Data

In this section, we assume that X^n comprises the first n samples of a homogeneous first-order Markov process having a symmetric transition matrix; i.e., (X^n, Y^n) satisfy (a$_2$) and (b). In practice, this may account for data that follows a pattern, such as a password.

It is easy to see that under assumptions (a$_2$) and (b),

$$\Pr(X^n = x^n) = \bar{p}\bar{r}^{n-1} \left(\frac{p}{\bar{p}}\right)^{x_1} \prod_{k=2}^{n} \left(\frac{r}{\bar{r}}\right)^{x_k \oplus x_{k-1}}.$$

In particular, if $r < \frac{1}{2} \leq p$, then a direct computation shows that $\mathsf{P_c}(X^n) = p\bar{r}^{n-1}$. The values of $\mathsf{P_c}(X^n|Y^n)$ for odd and even n are slightly different. For simplicity, in what follows we assume that n is odd. In this case, as shown in equation (C.42) in Appendix C.6,

$$\mathsf{P_c}(X^n|Y^n) = \bar{\alpha}^n \bar{r}^{n-1} \sum_{k=0}^{(n-1)/2} \binom{n}{k} \left(\frac{\alpha}{\bar{\alpha}}\right)^k. \tag{6.24}$$

Theorem 6.17 established the optimality of a Z-channel Z^{y_0,z_0} for some $y_0, z_0 \in \{0,1\}^n$. In order to find a closed form expression for $\underline{\hbar}_n$, it is necessary to find (y_0, z_0) which in principle depends on the parameters (p, α, r). The following theorem, whose proof is given in Appendix C.6, bounds $\underline{\hbar}_n$ for different values of (p, α, r).

Theorem 6.23. *Assume that $n \in \mathbb{N}$ is odd and (X^n, Y^n) satisfy (*a$_2$*) and (*b*) with $p \in [\frac{1}{2}, 1)$, $\alpha \in (0, \frac{1}{2})$, $\bar{\alpha} > p$ and $\mathsf{P_c}(X^n) < \mathsf{P_c}(X^n|Y^n)$. If $\frac{r}{\bar{r}} < \left(\frac{\alpha}{\bar{\alpha}}\right)^{n-1}$, then there exists $\varepsilon_{\mathsf{L}} < \mathsf{P_c}(X^n|Y^n)$ such that*

$$1 - \zeta_n(\varepsilon) \Pr(Y^n = \mathbf{1}) \leq \underline{\hbar}_n^n(\varepsilon) \leq 1 - \zeta_n(\varepsilon)\alpha^n,$$

for every $\varepsilon \in [\varepsilon_{\mathsf{L}}, \mathsf{P_c}(X^n|Y^n)]$*, where*

$$\zeta_n(\varepsilon) := \bar{r}\frac{\mathsf{P_c}(X^n|Y^n) - \varepsilon^n}{p(\bar{\alpha}\bar{r})^n - \bar{p}(\alpha\bar{r})^n}.$$

Furthermore, the 2^n*-ary Z-channel* $\mathsf{Z}_n(\zeta_n(\varepsilon))$ *achieves the lower bound in this interval.*

The special case of $r = 0$ is of particular interest. Note that when $r = 0$, then (a$_2$) corresponds to $X_1 = \cdots = X_n = \theta \in \{0, 1\}$. Here, $Y^n \in \{0,1\}^n$ are i.i.d. copies drawn from $P_{Y|\theta} = \mathsf{Bernoulli}(\bar{\alpha}^\theta \alpha^{\bar{\theta}})$. The prior distribution of the parameter θ is $\mathsf{Bernoulli}(p)$. The parameter θ is considered to be private and Y^n must be guessed as accurately as possible. This problem can be viewed as a reverse version of *privacy-aware learning* studied in [47]. The following proposition, whose proof is given in Appendix C.7, provides a closed form expression for $\underline{\hbar}_n$ in the low privacy regime. Note that in this case, $\mathsf{P_c}(\theta) = p$ and the value of $\mathsf{P_c}(\theta|Y^n)$ is obtained from (6.24) by setting $r = 0$.

Proposition 6.24. *Assume that* n *is odd. Let* $\theta \sim \mathsf{Bernoulli}(p)$ *with* $p \in [\frac{1}{2}, 1)$ *and* Y^n *be* n *i.i.d.* $\mathsf{Bernoulli}(\bar{\alpha}^\theta \alpha^{\bar{\theta}})$ *samples with* $\alpha \in (0, \frac{1}{2})$, $\bar{\alpha} > p$ *and* $p < \mathsf{P_c}(\theta|Y^n)$*. Then, there exists* $\varepsilon_{\mathsf{L}} < \mathsf{P_c}(\theta|Y^n)$ *such that*

$$\max_{\substack{P_{Z^n|Y^n}: \mathcal{Z}^n = \{0,1\}^n, \\ \mathsf{P_c}(\theta|Z^n) \le \varepsilon^n}} \mathsf{P_c}(Y^n|Z^n) = 1 - \zeta_n(\varepsilon)(p\bar{\alpha}^n + \bar{p}\alpha^n),$$

for every $\varepsilon \in [\varepsilon_{\mathsf{L}}, \mathsf{P_c}(\theta|Y^n)]$ *where*

$$\zeta_n(\varepsilon) = \frac{\mathsf{P_c}(\theta|Y^n) - \varepsilon^n}{p\bar{\alpha}^n - \bar{p}\alpha^n}.$$

Moreover, the 2^n*-ary Z-channel* $\mathsf{Z}_n(\zeta_n(\varepsilon))$ *achieves* $\underline{\hbar}_n(\varepsilon)$ *in this interval.*

Chapter 7

Privacy-Aware MMSE Estimation Efficiency

7.1 Overview

In Chapter 4, we studied the problem of information extraction under an information-theoretic privacy constraint for absolutely continuous X and Y and Gaussian additive privacy filters. In the previous chapter, we replaced the information-theoretic privacy requirement $I(X;Z) \leq \varepsilon$ by an estimation-theoretic requirement $I_\infty(X;Z) \leq \varepsilon$ in the discrete case. It is thus natural to follow the same spirit for the continuous case as well. Specifically, we focus on the additive Gaussian channels as the privacy filters and replace the privacy constraint $I(X;Z_\gamma) \leq \varepsilon$ as in (4.2) by a better justified estimation-theoretic constraint. The new constraint ensures that the minimum mean-squared error (MMSE) in estimating any arbitrary real-valued non-constant function f of X given the observation Z_γ is lower bounded. As such, an adversary observing Z cannot estimate efficiently any arbitrary function f of X, thus maintaining a very strong privacy guarantee.

Furthermore, we define the utility between Y and Z_γ by the efficiency of Z_γ in estimating Y. The estimation efficiency is defined as $\frac{1}{\mathsf{mmse}(Y|Z_\gamma)}$ which is equal to infinity if Y can be perfectly estimated from Z_γ. Therefore, we seek $\gamma \geq 0$ which minimizes $\mathsf{mmse}(Y|Z_\gamma)$

among all privacy-preserving Z_γ.

7.1.1 Main Contributions

The main contributions of this chapter are as follows:

- We first present an operational "definition" for an ε-private mechanism (i.e., channel $P_{Z|Y}$). This definition is motivated by the notion of information leakage in the previous chapter for the squared-error loss function and corresponds to the semantic privacy: all non-constant functions of the private data X need to remain private. We then show that this definition is equivalent to a certain constraint about maximal correlation, and thus we provide an operational interpretation for maximal correlation as a privacy measure.

- We then concentrate on the additive Gaussian filters and introduce the so-called estimation noise-to-signal ratio function sENSR as the corresponding utility-privacy tradeoff. We obtain tight bounds for sENSR by assuming Y is Gaussian and derive some extremal property for jointly Gaussian X and Y.

- Finally, we derive a tight bound for sENSR for arbitrary (X, Y) and show that this bound leads to a connection between sENSR and g in the special case $Y = aX + M$, for M being a noise random variable having density and independent of X.

7.2 Estimation Noise-to-Signal Ratio

We assume that X and Y are both real-valued absolutely continuous random variables (so that $\mathcal{X} = \mathcal{Y} = \mathbb{R}$) and the filter $P_{Z|Y}$ is realized by an independent additive Gaussian noise random variable $N_{\mathsf{G}} \sim \mathcal{N}(0,1)$ which is independent of (X,Y). Similar to Chapter 4, we denote the mechanism's output by $Z_\gamma = \sqrt{\gamma}Y + N_{\mathsf{G}}$, for some $\gamma \geq 0$.

As shown in Example 6.4 in the last chapter, if $\hat{\mathcal{X}} = \mathbb{R}$ and the loss function is given by $\ell(x,\hat{x}) = (x-\hat{x})^2$, the corresponding information leakage $\mathcal{L}_{\mathsf{MS}}(X \to Z)$ (see Definition 6.1) is

$$\mathcal{L}_{\mathsf{MS}}(X \to Z_\gamma) = \log \frac{\mathsf{var}(X)}{\mathsf{mmse}(X|Z_\gamma)}.$$

Hence, in order to have small information leakage from X to Z_γ, the filter must be such that $\mathsf{mmse}(X|Z_\gamma)$ is close to $\mathsf{var}(X)$. Since $\mathsf{mmse}(X|Z_\gamma) \leq \mathsf{var}(X)$, it is natural to consider filters which satisfy

$$1-\varepsilon \leq \frac{\mathsf{mmse}(X|Z_\gamma)}{\mathsf{var}(X)} \leq 1, \tag{7.1}$$

for a given $0 \leq \varepsilon \leq 1$, which clearly implies $\mathcal{L}_{\mathsf{MS}}(X \to Z_\gamma) \leq -\log(1-\varepsilon)$; thus $\mathcal{L}_{\mathsf{MS}}(X \to Z)$ is close to zero for $\varepsilon \ll 1$. Analogous to the discrete case, studied in Chapter 6, we formulate a stronger version of privacy where the information leakage from $f(X)$, any arbitrary non-constant deterministic function of X, to Z_γ is limited. In other words, we require that $\mathcal{L}_{\mathsf{MS}}(f(X) \to Z_\gamma)$ be small for *any* measurable real-valued function of X, or equivalently, as above

$$1-\varepsilon \leq \frac{\mathsf{mmse}(f(X)|Z_\gamma)}{\mathsf{var}(f(X))} \leq 1, \tag{7.2}$$

for a given $0 \leq \varepsilon \leq 1$ and all non-constant f. It is worth mentioning that the strong privacy guarantee introduced in (7.2) is related to *semantic security* [65] in the cryptographic literature. An encryption mechanism is said to be semantically secure if the adversary's advantage for correctly guessing any function of the private data given an observation of the mechanism's output (i.e., the ciphertext) is required to be negligible.

The operational privacy requirement (7.2) motivates the following definition.

Definition 7.1. *Given a pair of absolutely continuous random variables* (X, Y) *with distribution* P *and* $\varepsilon \geq 0$*, we say that* Z_γ *satisfies* ε-strong estimation privacy*, denoted as* $Z_\gamma \in \mathit{\Gamma}(\mathsf{P}, \varepsilon)$*, if* (7.2) *holds for any measurable real-valued non-constant function* f*. Similarly,* Z_γ *is said to satisfy* ε-weak estimation privacy*, denoted by* $Z_\gamma \in \partial\mathit{\Gamma}(\mathsf{P}, \varepsilon)$*, if* (7.2) *holds only for the identity function* $f(x) = x$*, as in* (7.1).

Similar to privacy, the utility between Y and Z_γ will be measured in terms of $\mathcal{L}_{\mathsf{MS}}(Y \to Z_\gamma)$. Since maximizing $\mathcal{L}_{\mathsf{MS}}(Y \to Z_\gamma)$ amounts to minimizing $\mathsf{mmse}(Y|Z_\gamma)$, to quantify the tradeoff between utility and information leakage, we define the strong and weak *estimation noise to signal ratio* (ENSR), respectively, as

$$\mathsf{sENSR}(\mathsf{P}, \varepsilon) := \inf_{\gamma: Z_\gamma \in \mathit{\Gamma}(\mathsf{P}, \varepsilon)} \frac{\mathsf{mmse}(Y|Z_\gamma)}{\mathsf{var}(Y)},$$

and

$$\mathsf{wENSR}(\mathsf{P}, \varepsilon) := \inf_{\gamma: Z_\gamma \in \partial\mathit{\Gamma}(\mathsf{P}, \varepsilon)} \frac{\mathsf{mmse}(Y|Z_\gamma)}{\mathsf{var}(Y)}.$$

Note that both $\mathsf{sENSR}(\mathsf{P}, \varepsilon)$ and $\mathsf{wENSR}(\mathsf{P}, \varepsilon)$ are inversely proportional for the respective utilities in these problems. For the sake of brevity, we omit P in $\mathit{\Gamma}(\mathsf{P}, \varepsilon)$, $\partial\mathit{\Gamma}(\mathsf{P}, \varepsilon)$, $\mathsf{sENSR}(\mathsf{P}, \varepsilon)$, and $\mathsf{wENSR}(\mathsf{P}, \varepsilon)$ when there is no confusion.

In what follows we derive an equivalent characterization of the random mapping $P_{Z|X}$ that generates $Z \in \mathit{\Gamma}(\varepsilon)$.

Theorem 7.2. *Let* U *and* V *be non-degenerate random variables and* $\varepsilon \in [0, 1]$*. Then*

$$\mathsf{mmse}(f(U)|V) \geq (1 - \varepsilon)\mathsf{var}(f(U)),$$

for all $f \in \mathcal{S}_U$ *if and only if* $\rho_m^2(U,V) \le \varepsilon$. *In particular,* $Z_\gamma \in \Gamma(\varepsilon)$ *if and only if* $\rho_m^2(X, Z_\gamma) \le \varepsilon$.

Proof. Fix f and define $\tilde{f}(U) := f(U) - \mathbb{E}[f(U)]$. Since $\mathsf{mmse}(\tilde{f}(U)|V) = \mathsf{mmse}(f(U)|V)$ and $\mathsf{var}(\tilde{f}(U)) = \mathsf{var}(f(U))$, without loss of generality, we can assume that $\mathbb{E}[f(U)] = 0$. Recalling the alternative characterization of the maximal correlation (5.1), we can write

$$\begin{aligned}
\inf_f \frac{\mathsf{mmse}(f(U)|V)}{\mathsf{var}(f(U))} &= \inf_{f\in\mathcal{S}_U} \mathsf{mmse}(f(U)|V) = 1 - \sup_{f\in\mathcal{S}_U} \mathsf{var}(\mathbb{E}[f(U)|V]) \\
&= 1 - \sup_{f\in\mathcal{S}_U} \mathbb{E}[\mathbb{E}^2[f(U)|V]] \qquad (7.3) \\
&= 1 - \rho_m^2(U,V). \qquad (7.4)
\end{aligned}$$

If $\rho_m^2(U,V) \le \varepsilon$, then it is clear from (7.4) that $\mathsf{mmse}(f(U)|V) \ge (1-\varepsilon)\mathsf{var}(f(U))$. Conversely, let P_{UV} satisfy $\mathsf{mmse}(f(U)|V) \ge (1-\varepsilon)\mathsf{var}(f(U))$ for any measurable f. In view of (7.3) and (7.4), there exists real-valued measurable f for arbitrary $\delta > 0$ such that

$$1 - \varepsilon \le \frac{\mathsf{mmse}(f(U)|V)}{\mathsf{var}(f(U))} \le 1 - \rho_m^2(U,V) + \delta,$$

which implies $\rho_m^2(U,V) \le \varepsilon$. □

From this theorem and (4.21), we can equivalently express $\mathsf{sENSR}(\varepsilon)$ and $\mathsf{wENSR}(\varepsilon)$ as

$$\mathsf{sENSR}(\varepsilon) = 1 - \sup_{\gamma\ge 0:\ \rho_m^2(X,Z_\gamma)\le\varepsilon} \eta_{Z_\gamma}^2(Y),$$

$$\mathsf{wENSR}(\varepsilon) = 1 - \sup_{\gamma\ge 0:\ \eta_{Z_\gamma}^2(X)\le\varepsilon} \eta_{Z_\gamma}^2(Y).$$

As observed in previous chapters, η and ρ_m satisfy the data processing inequality and hence $\eta_{Z_\gamma}(X) \leq \eta_Y(X)$ and $\rho_m(X, Z_\gamma) \leq \rho_m(X, Y)$. Therefore, we can restrict ε in the definition of $\mathsf{wENSR}(\varepsilon)$ and $\mathsf{sENSR}(\varepsilon)$ to the intervals $[0, \eta_Y^2(X)]$ and $[0, \rho_m^2(X, Y)]$, respectively. Unlike the discrete case, it is clear that perfect privacy $\varepsilon = 0$ implies $\gamma = 0$. Thus perfect privacy yields trivial utility; i.e., $\mathsf{sENSR}(0) = 1$ and $\mathsf{wENSR}(0) = 1$.

Note that $\gamma \mapsto \mathsf{mmse}(Y|Z_\gamma)$ is continuous and decreasing on $(0, \infty)$ [70] and $\gamma \mapsto \rho_m^2(X, Z_\gamma)$ is left-continuous and increasing on $(0, \infty)$ (see Proposition 5.2). Thus we can define $\gamma_\varepsilon^* := \max\{\gamma \geq 0 : \rho_m^2(X, Z_\gamma) \leq \varepsilon\}$ for which we have $\mathsf{sENSR}(\varepsilon) = \dfrac{\mathsf{mmse}(Y|Z_{\gamma_\varepsilon^*})}{\mathsf{var}(Y)}$. The left-continuity of $\gamma \mapsto \rho_m^2(X, Z_\gamma)$ implies that $\varepsilon \mapsto \gamma_\varepsilon^*$ is right-continuous, and thus $\varepsilon \mapsto \mathsf{sENSR}(\varepsilon)$ is right-continuous on $(0, \rho_m^2(X, Y))$.

Example 7.3. Let $(X_{\mathsf{G}}, Y_{\mathsf{G}})$ be jointly Gaussian random variables with mean zero and correlation coefficient ρ. Since $\rho_m^2(X_{\mathsf{G}}, Z_\gamma) = \rho^2(X_{\mathsf{G}}, Z_\gamma)$, we have that

$$\rho_m^2(X_{\mathsf{G}}, Z_\gamma) = \rho^2 \frac{\gamma \mathsf{var}(Y_{\mathsf{G}})}{1 + \gamma \mathsf{var}(Y_{\mathsf{G}})},$$

and hence the mapping $\gamma \mapsto \rho_m^2(X_{\mathsf{G}}, Z_\gamma)$ is strictly increasing. As a consequence, for $0 \leq \varepsilon \leq \rho^2$, the equation $\rho_m^2(X_{\mathsf{G}}, Z_\gamma) = \varepsilon$ has a unique solution

$$\gamma_\varepsilon := \frac{\varepsilon}{\mathsf{var}(Y_{\mathsf{G}})(\rho^2 - \varepsilon)},$$

and $\rho_m^2(X_{\mathsf{G}}, Z_\gamma) \leq \varepsilon$ if and only if $\gamma \leq \gamma_\varepsilon$. On the other hand,

$$\mathsf{mmse}(Y_{\mathsf{G}}|Z_\gamma) = \frac{\mathsf{var}(Y_{\mathsf{G}})}{1 + \gamma \mathsf{var}(Y_{\mathsf{G}})},$$

which shows that the map $\gamma \mapsto \mathsf{mmse}(Y_{\mathsf{G}}|Z_\gamma)$ is strictly decreasing. Therefore,

$$\mathsf{sENSR}(\varepsilon) = \frac{\mathsf{mmse}(Y_{\mathsf{G}}|Z_{\gamma_\varepsilon})}{\mathsf{var}(Y_{\mathsf{G}})} = 1 - \frac{\varepsilon}{\rho^2}. \tag{7.5}$$

Clearly, for jointly Gaussian X_{G} and Y_{G}, we have $\eta^2_{Z_\gamma}(X_{\mathsf{G}}) = \rho_m^2(X_{\mathsf{G}}, Z_\gamma)$ for any $\gamma \geq 0$. Consequently, $\Gamma(\varepsilon) = \partial\Gamma(\varepsilon)$ and, for $0 \leq \varepsilon \leq \rho^2$,

$$\mathsf{sENSR}(\varepsilon) = \mathsf{wENSR}(\varepsilon) = 1 - \frac{\varepsilon}{\rho^2}. \tag{7.6}$$

Next, we obtain bounds on $\mathsf{sENSR}(\varepsilon)$ for the special case of Gaussian non-private data Y_{G}.

Theorem 7.4. *Let X be jointly distributed with Gaussian Y_{G}. Then,*

$$1 - \frac{\varepsilon}{\rho^2(X, Y_{\mathsf{G}})} \leq \mathsf{sENSR}(\mathsf{P}_{XY_{\mathsf{G}}}, \varepsilon) \leq 1 - \frac{\varepsilon}{\rho_m^2(X, Y_{\mathsf{G}})},$$

Proof. Without loss of generality, assume $\mathbb{E}(X) = \mathbb{E}(Y_{\mathsf{G}}) = 0$. Since Y_{G} is Gaussian, we have $\rho_m^2(Y_{\mathsf{G}}, Z_\gamma) = \eta^2_{Z_\gamma}$ and thus (4.21) implies that

$$\mathsf{sENSR}(\varepsilon) = \inf_{\gamma:\rho_m^2(X,Z_\gamma)\leq\varepsilon} \frac{\mathsf{mmse}(Y_{\mathsf{G}}|Z_\gamma)}{\mathsf{var}(Y_{\mathsf{G}})} = 1 - \sup_{\gamma:\rho_m^2(X,Z_\gamma)\leq\varepsilon} \rho_m^2(Y_{\mathsf{G}}; Z_\gamma). \tag{7.7}$$

A straightforward computation leads to

$$\rho_m^2(Y_{\mathsf{G}}, Z_\gamma) = \rho^2(Y_{\mathsf{G}}, Z_\gamma) = \frac{\gamma\mathsf{var}(Y_{\mathsf{G}})}{1 + \gamma\mathsf{var}(Y_{\mathsf{G}})}, \tag{7.8}$$

$$\rho_m^2(X, Z_\gamma) \geq \rho^2(X, Z_\gamma) = \rho^2(X, Y_{\mathsf{G}})\rho_m^2(Y_{\mathsf{G}}, Z_\gamma).$$

The preceding inequality and (7.7) imply

$$\mathsf{sENSR}(\varepsilon) \geq 1 - \sup_{\gamma:\rho_m^2(X,Z_\gamma)\leq\varepsilon} \frac{\rho_m^2(X,Z_\gamma)}{\rho^2(X,Y_{\mathsf{G}})} = 1 - \frac{\varepsilon}{\rho^2(X,Y_{\mathsf{G}})},$$

which proves the lower bound.

The strong data processing inequality for maximal correlation proved in Lemma5.3 implies that if $\rho_m^2(Y_{\mathsf{G}},Z_\gamma) \leq \dfrac{\varepsilon}{\rho_m^2(X,Y)}$, then $\rho_m^2(X,Z_\gamma)\leq\varepsilon$. Therefore, (7.7) implies

$$\mathsf{sENSR}(\varepsilon) \leq 1 - \sup_{\gamma:\rho_m^2(Y_{\mathsf{G}},Z_\gamma)\leq\frac{\varepsilon}{\rho_m^2(X,Y_{\mathsf{G}})}} \rho_m^2(Y_{\mathsf{G}};Z_\gamma) = 1 - \frac{\varepsilon}{\rho_m^2(X,Y_{\mathsf{G}})},$$

where the last equality follows from the continuity of $\gamma\mapsto\rho_m^2(Y_{\mathsf{G}},Z_\gamma)$, established in (7.8), finishing the proof of the upper bound. □

Combined with (7.6), this theorem shows that for a Gaussian Y, a Gaussian X_{G} minimizes $\mathsf{sENSR}(\varepsilon)$ among all continuous random variables X having identical $\rho(X,Y_{\mathsf{G}})$ and maximizes $\mathsf{sENSR}(\varepsilon)$ among all continuous random variables X having identical $\rho_m(X,Y_{\mathsf{G}})$. These observations establish another extremal property of Gaussian distribution over AWGN channels, see e.g., [150, Theorem 12] for another example. This theorem also implies that

$$\mathsf{sENSR}(\mathsf{P}_{X_{\mathsf{G}}Y_{\mathsf{G}}},\varepsilon) - \mathsf{sENSR}(\mathsf{P}_{XY_{\mathsf{G}}},\varepsilon) \leq \varepsilon\left[\frac{1}{\rho^2(X,Y_{\mathsf{G}})} - \frac{1}{\rho_m^2(X,Y_{\mathsf{G}})}\right]$$

for Gaussian X_{G} which satisfies $\rho_m^2(X_{\mathsf{G}},Y_{\mathsf{G}}) = \rho_m^2(X,Y_{\mathsf{G}})$. This demonstrates that if the difference $\rho_m^2(X,Y_{\mathsf{G}}) - \rho^2(X,Y_{\mathsf{G}})$ is small, then $\mathsf{sENSR}(\mathsf{P}_{XY_{\mathsf{G}}},\varepsilon)$ is very close to $\mathsf{sENSR}(\mathsf{P}_{X_{\mathsf{G}}Y_{\mathsf{G}}},\varepsilon)$.

As stated before, for any given joint density P, perfect privacy results in trivial utility, i.e., $\mathsf{sENSR}(0) = 1$. Therefore, it is interesting to study the approximation of $\mathsf{sENSR}(\varepsilon)$ for sufficiently small ε, i.e., in the almost perfect privacy regime. The next result provides such an approximation and also shows that the lower bound in Theorem 7.4 holds for general Y for ε in the almost perfect privacy regime.

Lemma 7.5. *For any given joint density* P*, we have as* $\varepsilon \to 0$

$$\mathsf{sENSR}(\varepsilon) \geq 1 - \frac{\varepsilon}{\rho^2(X,Y)} + o(\varepsilon).$$

Proof. Let

$$\gamma_\varepsilon^* \coloneqq \sup\{\gamma \geq 0 : \rho_m^2(X, Z_\gamma) \leq \varepsilon\}. \tag{7.9}$$

Recall that

$$\rho_m^2(X, Z_\gamma) \geq \rho^2(X, Z_\gamma) = \frac{\gamma \rho^2(X,Y)\mathsf{var}(Y)}{1 + \gamma \mathsf{var}(Y)}. \tag{7.10}$$

Since $\varepsilon \to 0$, we can assume that $\varepsilon < \rho^2(X,Y)$. Thus, from (7.10) we obtain

$$\gamma_\varepsilon^* \leq \frac{\varepsilon}{\mathsf{var}(Y)(\rho^2(X,Y) - \varepsilon)}. \tag{7.11}$$

In particular, $\gamma_\varepsilon^* \to 0$ as $\varepsilon \to 0$. Since $\gamma \mapsto \mathsf{mmse}(Y|Z_\gamma)$ is decreasing, we have that $\mathsf{sENSR}(\varepsilon) = \mathsf{mmse}(Y|Z_{\gamma_\varepsilon^*})$. Therefore, the first-order approximation of $\mathsf{sENSR}(\cdot)$ around zero yields

$$\begin{aligned}
\mathsf{sENSR}(\varepsilon) &= 1 + \frac{\gamma_\varepsilon^*}{\mathsf{var}(Y)} \frac{\mathrm{d}}{\mathrm{d}\gamma_\varepsilon^*} \mathsf{mmse}(Y|Z_{\gamma_\varepsilon^*})\Big|_{\varepsilon=0} + o(\gamma_\varepsilon^*) \\
&\overset{(a)}{=} 1 - \mathsf{var}(Y)\gamma_\varepsilon^* + o(\gamma_\varepsilon^*) \\
&\overset{(b)}{\geq} 1 - \frac{\varepsilon}{\rho^2(X,Y)} + o(\varepsilon)
\end{aligned}$$

where (a) follows from the fact that $\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathsf{mmse}(Y|Z_\gamma) = -\mathbb{E}[\mathsf{var}^2(Y|Z_\gamma)]$ [70, Prop. 9] and (b) follows from (7.11). $\square$

We close this chapter by providing an interpretation for the rate-privacy function for continuous random variables introduced in Chapter 4. We showed in Corollary 4.12 that as $\varepsilon \to 0$

$$g(\varepsilon) = \frac{\varepsilon}{\eta_X^2(Y)} + o(\varepsilon). \tag{7.12}$$

Now assume that $P_{Y|X}$ is additive, i.e., $Y = aX + M$ for $a \in \mathbb{R}$ and an independent noise random variable M with a density having zero mean and variance σ_M^2. Then it is easy to verify that $\eta_X^2(Y) = \rho^2(X, Y)$ and hence in light of (7.12) and Lemma 7.5, we have for $\varepsilon \to 0$

$$g(\varepsilon) \geq 1 - \mathsf{sENSR}(\varepsilon) + o(\varepsilon),$$

which shows that in the almost perfect privacy regime the gap between $\mathsf{sENSR}(\varepsilon)$ and 1 is bounded by $g(\varepsilon)$, thereby providing an interpretation for $g(\varepsilon)$.

Chapter 8

Summary and Concluding Remarks

In this thesis, we mathematically formulated a more general local privacy setting. This setting takes into consideration the existence of two sets of correlated data: private data X and non-private (or observable) data Y, which is correlated with X via a fixed joint distribution P_{XY}. The ultimate goal is to generate the so-called displayed data Z based on Y such that Z maximizes the "utility" with respect to Y while limiting the "information leakage" about X. We proposed information-theoretic and estimation-theoretic metrics for utility and information leakage measures and quantified the corresponding privacy-utility tradeoff. We presented converse bounds on the achievable maximal utility under different metrics of information leakage. In particular, these bounds provide provably unconditional privacy guarantees: regardless of the computational resources available to the recipient of Z, he will not be able to guess/estimate X with the estimation error smaller than the proposed converse bounds. We then used these bounds to both evaluate and design optimal privacy-preserving mechanisms.

We claimed that this setting is more general than the setting studied in the differential privacy literature [50]. Indeed, the standard differential privacy analysis used in the centralized statistical databases can be mapped to this general framework: Y can represent a query response over a database, and X a binary variable that indicates whether or not a particular user is present in the database. The goal then is to distort the query response Y (in differential privacy this is often done by adding noise) in order to produce Z .

As a first step, we used mutual information as a metric for both utility and information leakage and defined the so-called rate-privacy function g as the corresponding privacy-utility tradeoff. If $g(\varepsilon) = R$, then one can maximally extract R bits of information about Y in a single shot such that the extracted information does not carry more than ε bits of private information about X. Apart from its interpretation in the context of privacy, g has an interesting geometric interpretation: it is closely related to the upper boundary of the convex set $\{(I(Y;Z), I(X;Z)) : X \multimap Y \multimap Z\}$. We mentioned that the mutual information does not arguably lead to an operational interpretation of privacy. Despite this fact, we showed that if both X and Y are discrete and the channel from Y to X enjoys a notion of symmetry, then g admits a simple expression. We also studied the properties of g when X and Y are continuous.

Second, we took an estimation-theoretic viewpoint on privacy while keeping the utility in terms of mutual information. We introduced $\hat{g}$ as the corresponding privacy-utility tradeoff, which quantifies the maximum number of bits one can extract from Y such that no deterministic function of X can be efficiently estimated from the extracted information. Specifically, we showed that this strong semantic privacy requirement is equivalent to a certain condition on the Hirschfeld-Gebelein-Rényi maximal correlation between X and the displayed data Z. Although $\hat{g}$ seems to be more complicated to deal with than g, we showed that g can serve as a tight bound for $\hat{g}$.

Third, we took a fully inferential point of view by bringing both utility and privacy in contact with statistical efficiency. In the discrete case, we used the Arimoto's mutual information of order infinity for both utility and privacy and defined g^{∞} as the corresponding privacy-utility tradeoff. In fact, g^{∞} quantifies the highest probability of correctly guessing Y from Z such that the probability of correctly guessing X from Z does not exceed a threshold. We derived simple closed-form expressions for g^{∞} in the binary case and also for a (practically-motivated) variant of g^{∞} in the non-binary case. In the continuous case, the corresponding privacy-utility tradeoff concerns a balance between the minimum mean-squared error (MMSE) of estimating Y from Z and MMSE of estimating X from Z.

We believe that the results and approaches presented here can be applied to develop theory and methods for distributed processing of statistical data. The fundamental limits of guessing and estimation under privacy constraints 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. In addition, the information and estimation-theoretic measures presented here can also be used to quantify the security threat posed if one of the processing nodes is attacked.

A possible direction for future work is extending our information and estimation-theoretic approaches to an asymptotic theory for information processing under a privacy constraint in distributed systems, which leads to a better understanding of the tradeoffs involved when acquiring, processing, securing and storing data.

Our approaches have also applications in centralized systems. An agency (e.g., a bank, hospital, or government) in possession of a large database of private and non-private data of individuals is often requested to respond to a query. Our approach can, at least in theory, help guide the design of a privacy-preserving query response mechanism.

Bibliography

[1] HIV and diabetes. https://aidsinfo.nih.gov/education-materials/fact-sheets/22/59/hiv-and-diabetes. Accessed: 2016-11-21.

[2] N. R. Adam and J. C. Worthmann. Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv.*, 21(4):515–556, December 1989.

[3] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proc. 20th ACM Symposium on Principles of Database Systems (PODS)*, pages 247–255, 2001.

[4] R. Ahlswede and P. Gács. Spreading of sets in product spaces and hypercontraction of the markov operator. *The annals of probability*, 4(6):925–939, 1976.

[5] R. Ahlswede and J. Körner. Source coding with side information and a converse for degraded broadcast channels. *IEEE Trans. Inf. Theory*, 21(6):629–637, Nov. 1975.

[6] E. Akyol, C. Langbort, and T. Başar. Privacy constrained information processing. In *Proc. 54th IEEE Conference on Decision and Control (CDC)*, pages 4511–4516, Dec. 2015.

[7] F. Alajaji and P. N. Chen. *Information Theory for Single User Systems, Part I.* Course Notes, Queen's University, http://www.mast.queensu.ca/~math474/it-lecture-notes.pdf, 2015.

[8] M. Alvim, M. E. Andrés, K. Chatzikokolakis, and C. Palamidessi. Quantitative information flow and applications to differential privacy. In *Foundations of Security Analysis and Design VI - FOSAD Tutorial Lectures*, pages 211–230, 2011.

[9] M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi. *Differential Privacy: On the Trade-Off between Utility and Information Leakage*, pages 39–54. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[10] V. Anantharam, A. Gohari, S. Kamath, and C. Nair. On hypercontractivity and a data processing inequality. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 3022–3026, June 2014.

[11] V. Anantharam, A. Gohari, S. Kamath, and C. Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. *[Online],* `arXiv:1304.6133v1`, 2014.

[12] S. Arimoto. Information measures and capacity of order α for discrete memoryless channels. *Colloq. on Inf. Theory*, 16:41–52, 1977.

[13] S. Asoodeh, F. Alajaji, and T. Linder. Notes on information-theoretic privacy. In *Proc. 52nd Annual Allerton Conf. Comm., Cont., Comput.*, pages 1272–1278, Sep. 2014.

[14] S. Asoodeh, F. Alajaji, and T. Linder. Lossless secure source coding: Yamamoto's setting. In *Proc. 53rd Annual Allerton Conf. Comm., Cont., and Comput.*, pages 1032–1037, Sep. 2015.

[15] S. Asoodeh, F. Alajaji, and T. Linder. On maximal correlation, mutual information and data privacy. In *Proc. IEEE Canadian Workshop on Inf. Theory (CWIT)*, pages 27–31, June 2015.

[16] S. Asoodeh, F. Alajaji, and T. Linder. Almost perfect privacy for additive Gaussian privacy filters. In *Information-Theoretic Security, [Online]* `arXiv:1608.04001v1`, pages 259–278. Springer-Verlag Lecture Notes in Computer Science, 2016.

[17] S. Asoodeh, F. Alajaji, and T. Linder. Privacy-aware MMSE estimation. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 1989–1993, July 2016.

[18] S. Asoodeh, F. P. Calmon, and S. Salamatian. On Mr. Gerber's Lemma. *To be submitted*, 2017.

[19] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder. Information extraction under privacy constraints. *Information*, 7, 2016.

[20] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder. Estimation efficiency under privacy constraints. *To be submitted*, 2017.

[21] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder. Privacy-aware guessing efficiency. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, June 2017.

[22] M. Benammar and A. Zaidi. Secure lossy helper and gray-wyner problems. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 2454–2458, July 2016.

[23] C. H. Bennett, G. Brassard, C. Crepeau, and U. M. Maurer. ”Generalized privacy amplification”. *IEEE Trans. Inf. Theory*, 41(6):1915–1923, Nov. 1995.

[24] T. Berger and R.W. Yeung. Multiterminal source encoding with encoder breakdown. *IEEE Trans. Inf. Theory*, 35(2):237–244, March 1989.

[25] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proc. 40th Annual ACM Symposium on the Theory of Computing*, pages 1123–1127, 2008.

[26] C. Braun, K. Chatzikokolakis, and C. Palamidessi. Quantitative notions of leakage for one-try attacks. *Electronic Notes in Theoretical Computer Science*, 249:75 – 91, 2009.

[27] W. Bryc, A. Dembo, and A. Kagan. On the maximum correlation coefficient. *Theory Probab. Appl.*, 49(1):132–138, Mar. 2005.

[28] F. P. Calmon. *Information-Theoretic Metrics for Security and Privacy*. PhD thesis, MIT, Sep. 2015.

[29] F. P. Calmon. private communication, 2016.

[30] F. P. Calmon and N. Fawaz. Privacy against statistical inference. In *Proc. 50th Annual Allerton Conf. Comm., Cont., and Comput.*, pages 1401–1408, Oct. 2012.

[31] F. P. Calmon, A. Makhdoumi, and M. Médard. Fundamental limits of perfect privacy. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 1796–1800, 2015.

[32] F. P. Calmon, M. Varia, M. Médard, M. M. Christiansen, K. R. Duffy, and S. Tessaro. Bounds on inference. In *Proc. 51st Annual Allerton Conf. Comm., Cont., and Comput.*, pages 567–574, Oct. 2013.

[33] K. Chatzikokolakis, C. Palamidessi, and P. Panangaden. On the Bayes Risk in Information-Hiding Protocols. *Journal of Computer Security*, 16(5):531–571, 2008.

[34] N. Chayat and S. Shamai. Extension of an entropy property for binary input memoryless symmetric channels. *IEEE Trans.Inf. Theory*, 35(5):1077–1079, March 1989.

[35] T. A. Courtade and T. Weissman. Multiterminal source coding under logarithmic loss. *IEEE Trans. Inf. Theory*, 60(1):740–761, Jan. 2014.

[36] T.A. Courtade. Information masking and amplification: The source coding setting. In *IEEE Int. Symp. Inf. Theory (ISIT)*, pages 189–193, 2012.

[37] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[38] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, (2):229–318, 1967.

[39] I. Csiszár and J. Körner. ”Broadcast channels with confidential messages”. *IEEE Trans. Inf. Theory*, 24(3):339–348, May 1978.

[40] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[41] P. Cuff. A framework for partial secrecy. In *Proc. IEEE Global Telecom. Conference GLOBE-COM*, pages 1–5, Dec. 2010.

[42] P. Cuff and L. Yu. Differential privacy as a mutual information constraint. In *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS)*, pages 43–54, 2016.

[43] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:2–1, 1977.

[44] A. De. *Lower Bounds in Differential Privacy*, pages 321–338. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[45] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proc. 22nd Symposium on Principles of Database Systems*, pages 202–210, 2003.

[46] L. E. Dubins. On extreme points of convex sets. *Journal of Mathematical Analysis and Applications*, 5(2):237 – 244, 1962.

[47] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. *Journal of the Association for Computing Machinery (ACM)*, 61(6), Dec. 2014.

[48] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81:10–18, 1986.

[49] G. T. Duncan and D. Lambert. The risk of disclosure for microdata. *Journal of Business and Economic Statistics,*, 7(2):207217, 1989.

[50] C. Dwork. Differential privacy. *Automata, Languages and Programming*, pages 1–12, 2006.

[51] C. Dwork. Differential privacy: a survey of results. *In Theory and Applications of Models of Computation, Lecture Notes in Computer Science*, (4978):1–19, 2008.

[52] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *3rd Conference on Theory of Cryptography (TCC'06)*, pages 265–284, 2006.

[53] E. Ekrem and S. Ulukus. Secure lossy source coding with side information. In *Proc. Annual Allerton Conf. Comm., Cont., and Comput.*, pages 1098–1105, Sep. 2011.

[54] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proc. 22nd Symposium on Principles of Database Systems*, pages 211–222, 2003.

[55] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Proc. 8th ACM Conf. Knowledge Discovery and Data Mining (KDD)*, pages 217–228, 2002.

[56] S. Fehr and S. Berens. On the conditional Rényi entropy. *IEEE Trans. Inf. Theory*, 60(11):6801–6810, Nov. 2014.

[57] I. P. Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67:7–18, 1972.

[58] S. E. Fienberg, U. E. Makov, and R. J. Steele. Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14:485–502, 1998.

[59] P. Gács and J. Körner. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2(2):149–162, 1973.

[60] R. G. Gallager. A simple derivation of the coding theorem and some applications. *IEEE Trans. Inf. Theory,*, 11(1):3–18, Jan.

[61] H. Gebelein. Das statistische problem der korrelation als variations- und eigenwert-problem und sein zusammenhang mit der ausgleichungsrechnung. *Zeitschrift f ur angew. Math. und Mech.*, (21):364–379, 1941.

[62] Q. Geng and P. Viswanath. The optimal noise-adding mechanism in differential privacy. *IEEE Trans. Inf. Theory*, 62(2):925–951, Feb. 2016.

[63] Y. Geng, C. Nair, S. Shamai, and Z. V. Wang. On broadcast channels with binary inputs and symmetric outputs. *IEEE Trans. Inf. Theory*, 59(11):6980–6989, March 2013.

[64] N. Gilbert. Researchers criticize genetic data restrictions. *Nature News*, 10.38(10), Sep. 2008.

[65] S. Goldwasser and S. Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 28(2):270 – 299, 1984.

[66] P. K. Gopala, L. Lai, and H. El Gamal. On the secrecy capacity of fading channels. *IEEE Trans. Inf. Theory*, 54(10):4687–4698, Oct. 2008.

[67] R. M. Gray and A. D. Wyner. Source coding for a simple network. *Bell System Technical Journal*, 53(9):1681–1721, 1974.

[68] D. Gündüz, E. Erkip, and H.V. Poor. Lossless compression with security constraints. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 111–115, July 2008.

[69] D. Gündüz, E. Erkip, and H.V. Poor. Secure lossless compression with side information. In *Proc. IEEE Inf. Theory Workshop (ITW)*, pages 169–173, May 2008.

[70] D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Inf. Theory*, 51(4):1261–1282, April 2005.

[71] D. Guo, Y. Wu, S. Shamai, and S. Verdú. Estimation in Gaussian noise: properties of the minimum mean-square error. *IEEE Trans. Inf. Theory*, 57(4):2371–2385, April 2011.

[72] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, March 2009.

[73] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, March 2009.

[74] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. *Advances in Neural Information Processing Systems*, pages 2348–2356, 2012.

[75] M. Hardt and K. Talwar. On the geometry of differential privacy. In *Proc. 42nd ACM Symposium on Theory of Computing, (STOC)*, pages 705–714, New York, NY, USA, 2010. ACM.

[76] M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In *Proc. 9th IEEE International Conference on Data Mining, ICDM*, pages 169–178, 2009.

[77] H. O. Hirschfeld. A connection between correlation and contingency. *Cambridge Philosophical Soc.*, 31:520–524, 1935.

[78] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genetics*, 4(8):1–9, Aug. 2008.

[79] I. Issa, S. Kamath, and A. B. Wagner. An operational measure of information leakage. In *Proc. Annual Conf. Inf. Science and Systems (CISS)*, pages 234–239, March 2016.

[80] I. Issa and A. B. Wagner. Measuring secrecy by the probability of a successful guess. *Submitted to IEEE Trans. Inf. Theory*, `http://arxiv.org/abs/1507.02342`, 2015.

[81] P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. In *Proc. 32nd Int. Conf. Machine Learning, ICML*, pages 1376–1385, 2015.

[82] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Journal of Machine Learning Research*, 17(1):492–542, January 2016.

[83] K. Kalantari, L. Sankar, and A. D. Sarwate. Optimal differential privacy mechanisms under hamming distortion for structured source classes. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 2069–2073, July 2016.

[84] W. Kang and S. Ulukus. A new data processing inequality and its applications in distributed source and channel coding. *IEEE Trans. Inf. Theory*, 57(1):56–69, Jan. 2011.

[85] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? In *Proc. IEEE Annual Symp. Foundations of Computer Science*, pages 531–540, 2008.

[86] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *Proc. ACM SIGMOD Int. Conf. Management of Data, SIGMOD*, pages 193–204, 2011.

[87] Y. H. Kim and A. El Gamal. *Network Information Theory*. Cambridge University Press, Cambridge.

[88] Y. H. Kim, A. Sutivong, and T.M. Cover. State amplification. *IEEE Trans. Inf. Theory*, 54(5):1850–1859, April 2008.

[89] G. Kimeldorf and A. R. Sampson. Monotone dependence. *The Annals of Statistics*, 6(4):895–903, July, 1978.

[90] K. Kittichokechai and G. Caire. Privacy-constrained remote source coding. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 1078–1082, July 2016.

[91] G. Kumar. On sequences of pairs of dependent random variables: A simpler proof of the main result using SVD. `http://www.gowthamiitm.com/research/Witsenhausen_simpleproof.pdf`, 2010.

[92] L. Lai and H. El Gamal. The relay-eavesdropper channel: cooperation for secrecy. *Proc. IEEE Trans. Inf. Theory*, 54(9):4005–4019, Sep. 2008.

[93] H.O. Lancaster. Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, pages 289–292, 1957.

[94] J. Lee and C. Clifton. Differential identifiability. In *Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, KDD*, pages 1041–1049, 2012.

[95] C. T. Li and A. El Gamal. Maximal correlation secrecy. `arXiv:1412.5374`, 2015.

[96] C. T. Li and A. El Gamal. Extended Gray-Wyner system with complementary causal side information. `http://arxiv.org/abs/1701.03207`, 2017.

[97] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010.

[98] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang. Membership privacy: a unifying framework for privacy definitions. In *Proc. ACM SIGSAC conf. on Comput., commun. security (CCS)*, pages 889–900, 2013.

[99] J. Liao, L. Sankar, F. P. Calmon, and V. Y. F. Tan. Hypothesis testing under maximal leakage privacy constraints. `http://arxiv.org/abs/1701.07099`, 2017.

[100] J. Liao, L. Sankar, V. Y. F. Tan, and F. P. Calmon. Hypothesis testing in the high privacy limit. `http://arxiv.org/abs/1607.00533`, 2016.

[101] T. Linder and R. Zamir. On the asymptotic tightness of the Shannon lower bound. *IEEE Trans. Inf. Theory*, 40(6):2026–2031, Nov. 1994.

[102] E. H. Linfoot. An informational measure of correlation. *Information and Control*, 1(1):85–89, 1957.

[103] J. Liu, P. Cuff, and S. Verdú. Resolvability in E_γ with applications to lossy compression and wiretap channels. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 755–759, June 2015.

[104] A. Makhdoumi and N. Fawaz. Privacy-utility tradeoff under statistical uncertainty. In *Proc. 51st Annual Allerton Conf. Comm., Cont., and Comput.*, pages 1627–1634, Oct. 2013.

[105] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard. From the information bottleneck to the privacy funnel. In *Proc. IEEE Inf. Theory Workshop (ITW)*, pages 501–505, 2014.

[106] P. Malacaria and H. Chen. Lagrange multipliers and maximum information leakage in different observational models. In *Proc. 3rd ACM SIGPLAN Workshop on Programming Languages and Analysis for Security*, pages 135–146, 2008.

[107] J. L. Massey. Guessing and entropy. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 204–, June 1994.

[108] N. Merhav. Shannon's secrecy system with informed receivers and its application to systematic coding for wiretapped channels. *IEEE Trans. Inf. Theory*, 54(6):2723–2734, June 2008.

[109] N. Merhav and E. Arikan. The shannon cipher system with a guessing wiretapper. *IEEE Trans. Inf. Theory*, 45(6):1860–1866, Sep. 1999.

[110] N. Merhav and S. Shamai. Information rates subject to state masking. *IEEE Trans. Inf. Theory*, 53(6):2254–2261, June 2007.

[111] D. J. Mir. Information-theoretic foundations of differential privacy. *Foundations and Practice of Security*, pages 374–381, 2013.

[112] F. Mokhtarinezhad, J. Kliewer, and O. Simeone. Lossy compression with privacy constraints: Optimality of polar codes. In *IEEE Inf. Theory Workshop (ITW)*, pages 182–186, Oct. 2015.

[113] C. Nair. An extremal inequality related to hypercontractivity of Gaussian random variables. In *Proc. Inf. Theory and Applications Workshop (ITA)*, pages 1–6, Feb. 2014.

[114] Y. Oohama. Gaussian multiterminal source coding. *IEEE Trans. Inf. Theory*, 43(6):2254–2261, July 1997.

[115] N. Papadatos and T. Xifara. A simple method for obtaining the maximal correlation coefficient and related characterizations. *Journal of Multivariate Analysis*, 118:102 – 114, 2013.

[116] A. Pastore and M. Gastpar. Locally differentially-private distribution estimation. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 2694–2698, July 2016.

[117] Y. Polyanskiy. Hypothesis testing via a comparator. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pages 2206–2210, July 2012.

[118] Y. Polyanskiy and Y. Wu. Dissipation of information in channels with input constraints. *IEEE Trans. Inf. Theory*, 62(1):35–55, Jan. 2016.

[119] V. Prabhakaran and K. Ramchandran. On secure distributed source coding. In *Proc. IEEE Inf. Theory Workshop (ITW)*, pages 442–447, Sep. 2007.

[120] S. R. Rajagopalan, L. Sankar, S. Mohajer, and H. V. Poor. Smart meter privacy: A utility-privacy framework. In *Proc. IEEE Int. Conf. Smart Grid Communications*, pages 190–195, Oct. 2011.

[121] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *IEEE Trans. Knowl. Data Eng.*, 22(11):1623–1636, Nov. 2010.

[122] A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3):441–451, 1959.

[123] A. Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(1):193–215, 1959.

[124] A. Rényi. On measures of entropy and information. In *4th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561, Berkeley, Calif., 1961. University of California Press.

[125] P. B. Rubinstein, L. Bartlett, J. Huang, and N. Taft. Learning in a large function space: privacy-preserving mechanisms for svm learning. *Journal of Privacy and Confidentiality*, 4(1):65–100, 2012.

[126] L. Sankar, S. R. Rajagopalan, and H. V. Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Trans. Inform. Forensics and Security*, 8(6):838–852, June 2013.

[127] L. Sankar, S.R. Rajagopalan, and H.V. Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Trans. Inf. Forensics Security*, 8(6):838–852, 2013.

[128] A. D. Sarwate and L. Sankar. A rate-disortion perspective on local differential privacy. In *Proc. 52nd Annual Allerton Conf. Comm., Cont., and Comput.*, pages 903–908, Sept 2014.

[129] F. B. Schneider, A. C. Myers, and M. R. Clarkson. Belief in information flow. *Proc. IEEE Computer Security Foundations Workshop*, pages 31–45, 2005.

[130] C. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28:656–715, Oct. 1949.

[131] N. Sharma and N. A. Warsi. Fundamental bound on the reliability of quantum information transmission. *Phys. Rev. Lett.*, 110:080501, Feb. 2013.

[132] N. Shulman and M. Feder. The uniform distribution as a universal prior. *IEEE Trans. Inf. Theory*, 50(6):1356–1362, June 2004.

[133] R. Sibson. Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14(2):149–160, 1969.

[134] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proc. of the Forty-third Annual ACM Symposium on Theory of Computing (STOC)*, pages 813–822, 2011.

[135] G. Smith. On the foundations of quantitative information flow. In *Proc. 12th International Conference on Foundations of Software Science and Computational Structures (FOSSACS)*, pages 288–302, 2009.

[136] I. Sutskover, S. Shamai, and J. Ziv. Extremes of information combining. *IEEE Trans. Inf. Theory*, 51(4):1313–1325, April 2005.

[137] L. Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, Oct. 2002.

[138] R. Tandon, L. Sankar, and H.V. Poor. Discriminatory lossy source coding: side information privacy. *IEEE Trans. Inf. Theory*, 59(9):5665–5677, April 2013.

[139] R. Tandon, S. Ulukus, and K. Ramchandran. Secure source coding with a helper. *IEEE Trans. Inf. Theory*, 59(4):2178–2187, April 2013.

[140] N. Tishby, F. C. Pereira, and W.Bialek. The information bottleneck method. *arXiv:physics/0004057*, April 2000.

[141] I. Vajda. *Theory of Statistical Inference and Information*. Kluwer Academic Publishers, 1989.

[142] S. Verdú. α-mutual information. In *Proc. Inf. Theory and Applications Workshop (ITA)*, pages 1–6, Feb. 2015.

[143] J. Villard and P. Piantanida. Secure multiterminal source coding with side information at the eavesdropper. *IEEE Trans. Inf. Theory,*, 59(6):3668–3692, June 2013.

[144] W. Wang, L. Ying, and J. Zhang. On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Trans. Inf. Theory*, 62(9):5018–5029, Sep. 2016.

[145] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(39):63–69, March 1965.

[146] H. Witsenhausen. Some aspects of convexity useful in information theory. *IEEE Trans. Inf. Theory*, 26(3):265–271, May 1980.

[147] H. Witsenhausen and A. Wyner. A conditional entropy bound for a pair of discrete random variables. *IEEE Trans. Inf. Theory*, 21(5):493–501, Sep. 1975.

[148] H. S. Witsenhausen. On sequence of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(2):100–113, 1975.

[149] Y. Wu and S. Verdú. Rényi information dimension: fundamental limits of almost lossless analog compression. *IEEE Trans. Inf. Theory*, 56(8):3721–3748, Aug. 2010.

[150] Y. Wu and S. Verdú. Functional properties of minimum mean-square error and mutual information. *IEEE Trans. Inf. Theory,*, 58(3):1289–1301, March 2012.

[151] A. Wyner. The common information of two dependent random variables. *IEEE Trans. Inf. Theory*, 21(2):163–179, March 1975.

[152] A. Wyner. On source coding with side information at the decoder. *IEEE Trans. Inf. Theory*, 21(3):294–300, May 1975.

[153] A. Wyner and J. Ziv. A theorem on the entropy of certain binary sequences and applications-part I. *IEEE Trans. Inf. Theory*, 19(6):769–772, Nov. 1973.

[154] A. D. Wyner. The wire-tap channel. *Bell System Technical Journal*, 54:1355–1387, 1975.

[155] H. Yamamoto. A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers. *IEEE Trans. Inf. Theory*, 29(6):918–923, Nov. 1983.

[156] L. Yu, H. Li, and C. W. Chen. Generalized common information: Common information extraction and private sources synthesis. `arXiv:1610.09289`, 2016.

[157] L. Zhao. *Common randomness, efficiency, and actions*. PhD thesis, Stanford University, 2011.

[158] E. Zheleva, E. Terzi, and L. Getoor. *Privacy in Social Networks*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, 2012.

Appendix A

Completion of Proof of Theorem 3.33

To prove that the equality (3.56) has only one solution $p = \frac{1}{2}$, we first show the following lemma.

Lemma A.1. *Let P and Q be two distributions over $\mathcal{X} = \{\pm 1, \pm 2, \dots, \pm k\}$ which satisfy $P(x) = Q(-x)$. Let $R_\lambda := \lambda P + (1-\lambda)Q$ for $\lambda \in (0,1)$. Then*

$$\frac{D(P||R_{1-\lambda})}{D(P||R_\lambda)} < \frac{\log(1-\lambda)}{\log(\lambda)}, \tag{A.1}$$

for $\lambda \in (0, \frac{1}{2})$ and

$$\frac{D(P||R_{1-\lambda})}{D(P||R_\lambda)} > \frac{\log(1-\lambda)}{\log(\lambda)}, \tag{A.2}$$

for $\lambda \in (\frac{1}{2}, 1)$.

Note that it is easy to see that the map $\lambda \mapsto D(P||R_\lambda)$ is convex and strictly decreasing and hence $D(P||R_\lambda) > D(P||R_{1-\lambda})$ when $\lambda \in (0, \frac{1}{2})$ and $D(P||R_\lambda) < D(P||R_{1-\lambda})$ when $\lambda \in (\frac{1}{2}, 1)$. Inequality (A.1) and (A.2) strengthen these monotonic behavior and show that $D(P||R_\lambda) > \frac{\log(\lambda)}{\log(1-\lambda)} D(P||R_{1-\lambda})$ and $D(P||R_\lambda) < \frac{\log(\lambda)}{\log(1-\lambda)} D(P||R_{1-\lambda})$ for $\lambda \in (0, \frac{1}{2})$ and $\lambda \in (\frac{1}{2}, 1)$, respectively.

Proof. Without loss of generality, we can assume that $P(x) > 0$ for all $x \in \mathcal{X}$. Let $\mathcal{X}_+ := \{x \in \mathcal{X} | P(X) > P(-x)\}$, $\mathcal{X}_- := \{x \in \mathcal{X} | P(X) < P(-x)\}$ and $\mathcal{X}_0 := \{x \in \mathcal{X} | P(X) =$

$P(-x)\}$. We notice that when $x \in \mathcal{X}_+$, then $-x \in \mathcal{X}_-$, and hence $|\mathcal{X}_+| = |\mathcal{X}_-| = m$ for a $0 < m \leq k$. After relabelling if needed, we can therefore assume that $\mathcal{X}_+ = \{1, 2, \ldots, m\}$ and $\mathcal{X}_- = \{-m, \ldots, -2, -1\}$. We can write

$$
\begin{aligned}
D(P||R_\lambda) &= \sum_{x=-k}^{k} P(x) \log\left(\frac{P(x)}{\lambda P(x) + (1-\lambda)Q(x)}\right) \\
&= \sum_{x=-k}^{k} P(x) \log\left(\frac{P(x)}{\lambda P(x) + (1-\lambda)P(-x)}\right) \\
&\overset{(a)}{=} \sum_{x=1}^{m} \left[P(x) \log\left(\frac{P(x)}{\lambda P(x) + (1-\lambda)P(-x)}\right)\right. \\
&\qquad\qquad \left. + P(-x) \log\left(\frac{P(-x)}{\lambda P(-x) + (1-\lambda)P(x)}\right)\right] \\
&\overset{(b)}{=} \sum_{x=1}^{m} \left[P(x) \log\left(\frac{1}{\lambda + (1-\lambda)\zeta_x}\right) + P(x)\zeta_x \log\left(\frac{1}{\lambda + \frac{(1-\lambda)}{\zeta_x}}\right)\right] \\
&\overset{(c)}{=} \sum_{x=1}^{m} P(x)\Upsilon(\lambda, \zeta_x) \, \log\left(\frac{1}{\lambda}\right),
\end{aligned}
$$

where (a) follows from the fact that for $x \in \mathcal{X}_0$, $\log\left(\frac{P(x)}{R_\lambda(x)}\right) = 0$ for any $\lambda \in (0,1)$, and in (b) and (c) we introduced $\zeta_x := \frac{P(-x)}{P(x)}$ and

$$
\Upsilon(\lambda, \zeta) := \frac{1}{\log\left(\frac{1}{\lambda}\right)} \left(\log\left(\frac{1}{\lambda + (1-\lambda)\zeta}\right) + \zeta \log\left(\frac{1}{\lambda + \frac{(1-\lambda)}{\zeta}}\right)\right).
$$

Similarly, we can write

$$
\begin{aligned}
D(P||R_{1-\lambda}) &= \sum_{x=-k}^{k} \log\left(\frac{P(x)}{(1-\lambda)P(x) + \lambda Q(x)}\right) \\
&= \sum_{x=-k}^{k} \log\left(\frac{P(x)}{(1-\lambda)P(x) + \lambda P(-x)}\right) \\
&= \sum_{x=1}^{m} \left[P(x) \log\left(\frac{P(x)}{(1-\lambda)P(x) + \lambda P(-x)}\right)\right.
\end{aligned}
$$

$$+P(-x)\log\left(\frac{P(-x)}{(1-\lambda)P(-x)+\lambda P(x)}\right)\Bigg]$$

$$= \sum_{x=1}^{m}\left[P(x)\log\left(\frac{1}{1-\lambda+\lambda\zeta_x}\right)+P(x)\zeta_x\log\left(\frac{1}{1-\lambda+\frac{\lambda}{\zeta_x}}\right)\right]$$

$$= \sum_{x=1}^{m}P(x)\Upsilon(1-\lambda,\zeta_x)\,\log\left(\frac{1}{1-\lambda}\right),$$

which implies that

$$\frac{D(P||R_\lambda)}{-\log(\lambda)}-\frac{D(P||R_{1-\lambda})}{-\log(1-\lambda)}=\sum_{x=1}^{m}P(x)\left[\Upsilon(\lambda,\zeta_x)-\Upsilon(1-\lambda,\zeta_x)\right].$$

Hence, in order to show (A.1), it suffices to verify that

$$\Phi(\lambda,\zeta):=\Upsilon(\lambda,\zeta)-\Upsilon(1-\lambda,\zeta)>0, \tag{A.3}$$

for any $\lambda\in(0,\frac{1}{2})$ and $\zeta\in(1,\infty)$. Since $\log(\lambda)\log(1-\lambda)$ is always positive for $\lambda\in(0,\frac{1}{2})$, it suffices to show that

$$h(\zeta):=\Phi(\lambda,\zeta)\log(1-\lambda)\log(\lambda)>0, \tag{A.4}$$

for $\lambda\in(0,\frac{1}{2})$ and $\zeta\in(1,\infty)$. We have

$$h''(\zeta)=A(\lambda,\zeta)B(\lambda,\zeta), \tag{A.5}$$

where

$$A(\lambda,\zeta):=\frac{1+\zeta}{(1-\lambda+\lambda\zeta)^2(\lambda+(1-\lambda)\zeta)^2\zeta},$$

and

$$B(\lambda,\zeta):=\lambda^2(1+\lambda(\lambda-2)(\zeta-1)^2+\zeta(\zeta-1))\log(\lambda)-(1-\lambda)^2(\lambda^2(\zeta-1)^2+\zeta)\log(1-\lambda).$$

We have

$$\frac{\partial^2}{\partial \zeta^2} B(\lambda, \zeta) = 2\lambda^2(1-\lambda)^2 \log\left(\frac{\lambda}{1-\lambda}\right) < 0,$$

because $\lambda \in (0, \frac{1}{2})$ and hence $\lambda < 1-\lambda$. This implies that the map $\zeta \mapsto B(\lambda, \zeta)$ is concave for any $\lambda \in (0, \frac{1}{2})$ and $\zeta \in (1, \infty)$. Moreover, since $\zeta \mapsto B(\lambda, \zeta)$ is a quadratic polynomial with negative leading coefficient, it is clear that $\lim_{\zeta\to\infty} B(\lambda, \zeta) = -\infty$. Consider now $g(\lambda) := B(\lambda, 1) = \lambda^2 \log(\lambda) - (1-\lambda)^2 \log(1-\lambda)$. We have $\lim_{\lambda\to 0} g(\lambda) = g(\frac{1}{2}) = 0$ and $g''(\lambda) = 2\log\left(\frac{\lambda}{1-\lambda}\right) < 0$ for $\lambda \in (0, \frac{1}{2})$. It implies that $\lambda \mapsto g(\lambda)$ is concave on $(0, \frac{1}{2})$ and hence $g(\lambda) > 0$ over $(0, \frac{1}{2})$ which implies that $B(\lambda, 1) > 0$. This together with the fact that $\zeta \mapsto B(\lambda, \zeta)$ is concave and it approaches to $-\infty$ as $\zeta \to \infty$ imply that there exists a real number $c = c(\lambda) > 1$ such that $B(\lambda, \zeta) > 0$ for all $\zeta \in (1, c)$ and $B(\lambda, \zeta) < 0$ for all $\zeta \in (c, \infty)$. Since $A(\lambda, \zeta) > 0$, it follows from (A.5) that $\zeta \mapsto h(\zeta)$ is convex on $(1, c)$ and concave on (c, ∞). Since $h(1) = h'(1) = 0$ and $\lim_{\zeta\to\infty} h(\zeta) = \infty$, we can conclude that $h(\zeta) > 0$ over $(1, \infty)$. That is, $\Phi(\lambda, \zeta) > 0$ and thus $\Upsilon(\lambda, \zeta) - \Upsilon(1-\lambda, \zeta) > 0$, for $\lambda \in (0, \frac{1}{2})$ and $\zeta \in (1, \infty)$.

The inequality (A.2) can be proved by (A.1) and switching λ to $1-\lambda$. $\square$

Letting $P(\cdot) = P_{X|Y}(\cdot|1)$ and $Q(\cdot) = P_{X|Y}(\cdot|0)$ and $\lambda = \Pr(Y = 1) = p$, we have $R_p(x) = P_X(x) = pP(x) + (1-p)Q(x)$ and $R_{1-p} = P_X(-x) = (1-p)P(x) + pQ(x)$. Since $D(P_{X|Y}(\cdot|0)||P_X(\cdot)) = D(P||R_{1-p})$, we can conclude from Lemma A.1 that

$$\frac{D(P_{X|Y}(\cdot|0)||P_X(\cdot))}{-\log(1-p)} < \frac{D(P_{X|Y}(\cdot|1)||P_X(\cdot))}{-\log(p)},$$

over $p \in (0, \frac{1}{2})$ and

$$\frac{D(P_{X|Y}(\cdot|0)||P_X(\cdot))}{-\log(1-p)} > \frac{D(P_{X|Y}(\cdot|1)||P_X(\cdot))}{-\log(p)},$$

on $p \in (\frac{1}{2}, 1)$, and hence equation (3.56) has only solution $p = \frac{1}{2}$.

Appendix B

Proofs of Chapter 4

B.1 Proof of Lemma 4.1

We first prove the following version of the data processing inequality which will be required in the proofs.

Lemma B.1. *Let X and Y be absolutely continuous random variables such that X, Y and (X,Y) have finite differential entropies. If V is an absolutely continuous random variable independent of X and Y, then*

$$I(X;Y+V) \leq I(X;Y)$$

with equality if and only if X and Y are independent.

Proof. Since $X \multimap Y \multimap (Y+V)$, the data processing inequality implies that $I(X;Y+V) \leq I(X;Y)$. It therefore suffices to show that this inequality is tight if and only X and Y are independent. It is known that data processing inequality is tight if and only if $X \multimap (Y+V) \multimap Y$. This is equivalent to saying that for any measurable set $A \subset \mathbb{R}$ and for P_{Y+V}-almost all z, $\Pr(X \in A|Y+V=z, Y=y) = \Pr(X \in A|Y+V=z)$. On the other hand, due to the independence of V and (X,Y), we have $\Pr(X \in A|Y+V=z, Y=y) = \Pr(X \in A|Y=y)$. Hence, the equality in data processing inequality holds if and only if $\Pr(X \in A|Y+V=z) =$

$\Pr(X \in A|Y = y)$ which implies that X and Y must be independent. $\square$

Proof of Lemma 4.1. Recall that, by assumption (b), $\mathsf{var}(Y)$ is finite. This implies that the entropy of Y is also finite. The finiteness of the entropy of U_λ then follows directly from the entropy power inequality [37, Theorem 17.7.3] and the fact that $\mathsf{var}(U_\lambda) = \mathsf{var}(Y) + \lambda^2 < \infty$. The data processing inequality, as stated in Lemma B.1, implies that for any $\delta > 0$, we have $I(Y; U_{\lambda+\delta}) \geq I(Y; U_\lambda)$. Clearly, Y and U_λ are not independent for any $\lambda < \infty$, therefore the inequality is strict and thus $\lambda \mapsto I(Y, U_\lambda)$ is strictly increasing.

Continuity is proved for $\lambda = 0$ and $\lambda > 0$ separately. Let first $\lambda = 0$. Recall that $h(\lambda N_{\mathsf{G}}) = \frac{1}{2}\log(2\pi e\lambda^2)$. In particular, $\lim_{\lambda\to 0} h(\lambda N_{\mathsf{G}}) = -\infty$, which together with the entropy power inequality implies that $\lim_{\lambda\to 0} I(Y; U_\lambda) = \infty$. This coincides with the convention $I(Y; Z_0) = I(Y; Y) = \infty$. For $\lambda > 0$, let $(\lambda_n)_{n\geq 1}$ be a sequence of positive numbers such that $\lambda_n \to \lambda$. Observe that

$$I(Y; U_{\lambda_n}) = h(Y + \lambda_n N_{\mathsf{G}}) - h(\lambda_n N_{\mathsf{G}}) = h(Y + \lambda_n N_{\mathsf{G}}) - \frac{1}{2}\log(2\pi e\lambda_n^2).$$

Since $\lim_{n\to\infty} \frac{1}{2}\log(2\pi e\lambda_n^2) = \frac{1}{2}\log(2\pi e\lambda^2)$, we only have to show that $h(Y + \lambda_n N) \to h(Y + \lambda N)$ as $n \to \infty$ to establish the continuity at λ. This, in fact, follows from de Bruijn's identity (cf., [37, Theorem 17.7.2]).

Since the channel from Y to U_λ is an additive Gaussian noise channel, we have $I(Y; U_\lambda) \leq \frac{1}{2}\log\left(1 + \lambda^{-2}\mathsf{var}(Y)\right)$ with equality if and only if Y is Gaussian. The claimed limit as $\lambda \to 0$ is clear. $\square$

B.2 Proof of Lemma 4.2

The proof of the strictly decreasing behavior of $\lambda \mapsto I(X; U_\lambda)$ is analogous to the proof of Lemma 4.1. To prove continuity, let $\lambda \geq 0$ be fixed. Let $(\lambda_n)_{n\geq 1}$ be any sequence of positive numbers converging to λ. First suppose that $\lambda > 0$. Recall that $I(X; U_{\lambda_n}) = h(U_{\lambda_n}) - h(U_{\lambda_n}|X)$, for all $n \geq 1$. As shown in Lemma 4.1, $h(U_{\lambda_n}) \to h(U_\lambda)$ as $n \to \infty$. Therefore, it is enough to show that $h(U_{\lambda_n}|X) \to h(U_\lambda|X)$ as $n \to \infty$. Note that by de Bruijn's identity, we have $h(Z_{\lambda_n}|X = x) \to h(Z_\lambda|X = x)$ as $n \to \infty$ for all $x \in \mathbb{R}$. Note also that since

$$h(U_{\lambda_n}|X = x) \le \frac{1}{2}\log\left(2\pi e \mathsf{var}(U_{\lambda_n}|x)\right),$$

we can write

$$h(U_{\lambda_n}|X) \le \mathbb{E}\left[\frac{1}{2}\log(2\pi e \mathsf{var}(U_{\lambda_n}|X))\right] \le \frac{1}{2}\log\left(2\pi e \mathbb{E}[\mathsf{var}(U_{\lambda_n}|X)]\right),$$

and hence we can apply dominated convergence theorem to show that $h(U_{\lambda_n}|X) \to h(U_\lambda|X)$ as $n \to \infty$. To prove the continuity at $\lambda = 0$, we first note that Linder and Zamir [101, Page 2028] showed that $h(U_{\lambda_n}|X = x) \to h(Y|X = x)$ as $n \to \infty$, and hence as before by dominated convergence theorem we can show that $h(U_{\lambda_n}|X) \to h(Y|X)$. Similarly [101] implies that $h(U_{\lambda_n}) \to h(Y)$. This concludes the proof of the continuity of $\lambda \mapsto I(X; U_\lambda)$. To prove the last claim, note that the data processing inequality and Lemma 4.1 imply

$$0 \le I(X; U_\lambda) \le I(Y; U_\lambda) \le \frac{1}{2}\log\left(1 + \frac{\mathsf{var}(Y)}{\lambda^2}\right),$$

and hence $\lim\limits_{\lambda\to\infty} I(X; U_\lambda) = 0$.

B.3 Proof of Lemma 4.3

In order to prove Lemma 4.3, we first prove some preliminary results.

Theorem B.2 ([123]). *If U is an absolutely continuous random variable with density f_U and if $H(\lfloor U \rfloor) < \infty$, then*

$$\lim_{n\to\infty} H(n^{-1}\lfloor nU \rfloor) - \log(n) = -\int_{\mathbb{R}} f_U(x)\log f_U(x)dx,$$

provided that the integral on the right hand side exists.

We will need the following consequence of the previous theorem. Rényi [123] proved that $H(\mathcal{Q}_M(U)) \le H(\lfloor U \rfloor) + M$; however, one can improve this inequality using Jensen's inequality as $H(\mathcal{Q}_{M-1}(U)) \le H(\mathcal{Q}_M(U)) \le H(\mathcal{Q}_{M-1}(U)) + 1$.

Lemma B.3. *If U is an absolutely continuous random variable with density f_U and if $H(\lfloor U \rfloor) < \infty$, then $H(\mathcal{Q}_M(U)) - M \ge H(\mathcal{Q}_{M+1}(U)) - (M+1)$ for all $M \ge 1$ and*

$$\lim_{M\to\infty} H(\mathcal{Q}_M(U)) - M = -\int_{\mathbb{R}} f_U(x) \log f_U(x) dx,$$

provided that the integral on the right hand side exists.

Lemma B.4. *Fix $M \in \mathbb{N}$. Assume that $f_Y(y) \le C|y|^{-p}$ for some positive constant C and $p > 1$. For integer k and $\lambda \ge 0$, let*

$$p_{k,\lambda} \coloneqq \Pr\left(U_\lambda^M = \frac{k}{2^M}\right).$$

Then

$$p_{k,\lambda} \le \frac{C 2^{(p-1)M+p}}{k^p} + 1_{\{\lambda>0\}} \frac{\lambda 2^{M+1}}{k\sqrt{2\pi}} e^{-k^2/2^{2M+3}\lambda^2}.$$

Proof. The case $\lambda = 0$ is trivial, so we assume that $\lambda > 0$. For notational simplicity, let $r_a \coloneqq \frac{a}{2^M}$ for all $a \in \mathbb{Z}$. Assume that $k \ge 0$. Observe that

$$p_{k,\lambda} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\lambda N_{\mathsf{G}}}(n) f_Y(y) 1_{[r_k, r_{k+1})}(y+n) \mathrm{d}y \mathrm{d}n = \int_{-\infty}^{\infty} \frac{e^{-n^2/2\lambda^2}}{\sqrt{2\pi\lambda^2}} \Pr\left(Y \in [r_k, r_{k+1}) - n\right) \mathrm{d}n.$$

We will estimate the above integral by breaking it up into two pieces. First, we consider

$$\int_{-\infty}^{\frac{r_k}{2}} \frac{e^{-n^2/2\lambda^2}}{\sqrt{2\pi\lambda^2}} \Pr\left(Y \in [r_k, r_{k+1}) - n\right) \mathrm{d}n.$$

When $n \leq \frac{r_k}{2}$, then $r_k - n \geq r_k/2$. By the assumption on the density of Y,

$$\Pr\left(Y \in [r_k, r_{k+1}) - n\right) \leq \frac{C}{2^M}\left(\frac{r_k}{2}\right)^{-p}.$$

(The previous estimate is the only contribution when $\lambda = 0$.) Therefore,

$$\int_{-\infty}^{\frac{r_k}{2}} \frac{e^{-n^2/2\lambda^2}}{\sqrt{2\pi\lambda^2}} \Pr\left(Y \in [r_k, r_{k+1}) - n\right) \mathrm{d}n \leq \frac{C}{2^M}\left(\frac{r_k}{2}\right)^{-p} \int_{-\infty}^{\frac{r_k}{2}} \frac{e^{-n^2/2\lambda^2}}{\sqrt{2\pi\lambda^2}} \mathrm{d}n \leq \frac{C 2^{(p-1)M+p}}{k^p}.$$

Using the trivial bound $\Pr\left(Y \in [r_k, r_{k+1}) - n\right) \leq 1$ and well known estimates for the error function, we obtain that

$$\int_{\frac{r_k}{2}}^{\infty} \frac{e^{-n^2/2\lambda^2}}{\sqrt{2\pi\lambda^2}} \Pr\left(Y \in [r_k, r_{k+1}) - n\right) \mathrm{d}n < \frac{1}{\sqrt{2\pi}} \frac{2\lambda}{r_k} e^{-r_k^2/8\lambda^2} = \frac{\lambda 2^{M+1}}{k\sqrt{2\pi}} e^{-k^2/2^{2M+3}\lambda^2}.$$

Therefore, we have

$$p_{k,\lambda} \leq \frac{C 2^{(p-1)M+p}}{k^p} + \frac{\lambda 2^{M+1}}{k\sqrt{2\pi}} e^{-k^2/2^{2M+3}\lambda^2}.$$

The proof for $k < 0$ is completely analogous. □

Lemma B.5. *Fix $M \in \mathbb{N}$. Assume that $f_Y(y) \leq C|y|^{-p}$ for some positive constant C and $p > 1$. The mapping $\lambda \mapsto H(U_\lambda^M)$ is continuous.*

Proof. Let $(\lambda_n)_{n\geq 1}$ be a sequence of non-negative real numbers converging to λ_0. First, we prove continuity at $\lambda_0 > 0$. Without loss of generality, assume that $\lambda_n > 0$ for all $n \in \mathbb{N}$. Define $\lambda_* \coloneqq \inf\{\lambda_n :\ n \geq 1\}$ and $\lambda^* \coloneqq \sup\{\lambda_n :\ n \geq 1\}$. Clearly $0 < \lambda_* \leq \lambda^* < \infty$. Recall that

$$p_{k,\lambda} = \int_{\mathbb{R}} \frac{e^{-z^2/2\lambda^2}}{\sqrt{2\pi\lambda^2}} \Pr\left(Y \in \left[\frac{k}{2^M}, \frac{k+1}{2^M}\right) - z\right) \mathrm{d}z.$$

Since, for all $n \in \mathbb{N}$ and $z \in \mathbb{R}$,

$$\frac{e^{-z^2/2\lambda_n^2}}{\sqrt{2\pi\lambda_n^2}} \Pr\left(Y \in \left[\frac{k}{2^M}, \frac{k+1}{2^M}\right) - z\right) \leq \frac{e^{-z^2/2(\lambda^*)^2}}{\sqrt{2\pi\lambda_*^2}},$$

the dominated convergence theorem implies that

$$\lim_{n\to\infty} p_{k,\lambda_n} = p_{k,\lambda_0}. \tag{B.1}$$

Lemma B.4 implies that for all $n \geq 0$ and $|k| > 0$,

$$p_{k,\lambda_n} \leq \frac{C2^{(p-1)M+p}}{k^p} + \frac{\lambda_n 2^{M+1}}{k\sqrt{2\pi}} e^{-k^2/2^{2M+3}\lambda_n^2}.$$

Thus, for k large enough, $p_{k,\lambda_n} \leq \dfrac{A}{k^p}$ for a suitable positive constant A that does not depend on n. Since the function $x \mapsto -x\log(x)$ is increasing in $[0, 1/2]$, there exists $K' > 0$ such that for $|k| > K'$

$$-p_{k,\lambda_n} \log(p_{k,\lambda_n}) \leq \frac{A}{k^p} \log(A^{-1}k^p).$$

Since $\displaystyle\sum_{|k|>K'} \frac{A}{k^p} \log(A^{-1}k^p) < \infty$, for any $\varepsilon > 0$ there exists K_ε such that

$$\sum_{|k|>K_\varepsilon} \frac{A}{k^p} \log(A^{-1}k^p) < \varepsilon.$$

In particular, for all $n \geq 0$,

$$H(U_{\lambda_n}^M) - \sum_{|k|\leq K_\varepsilon} -p_{k,\lambda_n} \log(p_{k,\lambda_n}) = \sum_{|k|>K_\varepsilon} -p_{k,\lambda_n} \log(p_{k,\lambda_n}) < \varepsilon.$$

Therefore, for all $n \geq 1$,

$$\left|H(U_{\lambda_n}^M) - H(U_{\lambda_0}^M)\right|$$

$$\leq \sum_{|k|>K_\varepsilon} -p_{k,\lambda_n}\log(p_{k,\lambda_n}) + \left|\sum_{|k|\leq K_\varepsilon} p_{k,\lambda_0}\log(p_{k,\lambda_0}) - p_{k,\lambda_n}\log(p_{k,\lambda_n})\right| + \sum_{|k|>K_\varepsilon} -p_{k,\lambda_0}\log(p_{k,\lambda_0})$$
$$\leq \varepsilon + \left|\sum_{|k|\leq K_\varepsilon} p_{k,\lambda_0}\log(p_{k,\lambda_0}) - p_{k,\lambda_n}\log(p_{k,\lambda_n})\right| + \varepsilon.$$

From (B.1) and the continuity of the function $x \mapsto -x\log(x)$ on $[0,1]$, we conclude that

$$\limsup_{n\to\infty} \left|H(U_{\lambda_n}^M) - H(U_{\lambda_0}^M)\right| \leq 3\varepsilon.$$

Since ε is arbitrary, we have $\lim_{n\to\infty} H(U_{\lambda_n}^M) = H(U_{\lambda_0}^M)$, which completes the proof.

To prove continuity at $\lambda_0 = 0$, observe that equation (B.1) holds in this case as well. The rest is analogous to the case $\lambda_0 > 0$. □

Proof of Lemma 4.3. Since $\lambda \to H(U_\lambda^M)$ is continuous, it suffices to show that $\lambda \mapsto H(U_\lambda^M|Y)$ and $\lambda \mapsto H(U_\lambda^M|X)$ are continuous. To show these, we will first prove that $H(U_\lambda^M|Y)$ and $H(U_\lambda^M|X)$ are bounded and then apply the dominated convergence theorem.

We note that $H(U_\lambda^M|Y) \leq M + H(\lfloor U_\lambda \rfloor|Y)$, and hence we can write

$$\begin{aligned} H(U_\lambda^M|Y) &\leq M + H(\lfloor U_\lambda \rfloor|Y) \leq M + \sup_{t\in[0,1]} H(\lfloor t + \lambda N_{\mathsf{G}} \rfloor) \\ &\overset{(a)}{\leq} M + \frac{1}{2}\sup_{t\in[0,1]} \log\left(2\pi e\mathbb{E}[(\lfloor t+\lambda N_{\mathsf{G}}\rfloor)^2] + \frac{2\pi e}{12}\right), \end{aligned}$$

where (a) follows from [37, Problem 8.7]. Since $|\lfloor t+\lambda N_{\mathsf{G}}\rfloor| \leq |t+\lambda N_{\mathsf{G}}| + 1$ almost surely, we have $\mathbb{E}[(\lfloor t+\lambda N_{\mathsf{G}}\rfloor)^2] \leq 2t^2 + \lambda^2 c_1$, where c_1 is a positive constant. Consequently,

$$H(U_\lambda^M|Y) \leq M + \frac{1}{2}\log\left(4\pi e + \lambda^2 c_1 + \frac{2\pi e}{12}\right),$$

and hence the continuity of $\lambda \mapsto H(U_\lambda^M|Y)$ follows from the dominated convergence theorem.

Analogously, we can write for any $x \in \mathbb{R}$

$$\begin{aligned} H(U_\lambda^M|X=x) &\leq M + H(\lfloor U_\lambda \rfloor | X = x) \\ &\stackrel{(b)}{\leq} M + \frac{1}{2}\log\left(2\pi e \mathbb{E}[(\lfloor U_\lambda \rfloor)^2|X=x] + \frac{2\pi e}{12}\right), \end{aligned}$$

where (b) follows from [37, Problem 8.7]. As before, we can easily show that $\mathbb{E}[(\lfloor U_\lambda \rfloor)^2|X=x] \leq 2\mathbb{E}[Y^2|X=x] + \lambda^2 c_2$, where c_2 is a constant. Consequently, we can write

$$\begin{aligned} H(U_\lambda^M|X) &= \int H(U_\lambda^M|X=x) f_X(x)\mathrm{d}x \\ &\leq M + \int H(\lfloor U_\lambda \rfloor|X=x) f_X(x)\mathrm{d}x \\ &\leq M + \frac{1}{2}\int \log\left(2\pi e\mathbb{E}[(\lfloor U_\lambda \rfloor)^2|X=x] + \frac{2\pi e}{12}\right) f_X(x)\mathrm{d}x \\ &\leq M + \frac{1}{2}\int \log\left(4\pi e\mathbb{E}[Y^2|X=x] + \lambda^2 c_2 + \frac{2\pi e}{12}\right) f_X(x)\mathrm{d}x \\ &\stackrel{(c)}{\leq} M + \frac{1}{2}\log\left(4\pi e\mathbb{E}[Y^2] + \lambda^2 c_2 + \frac{2\pi e}{12}\right), \end{aligned}$$

where (c) follows from Jensen's inequality. The continuity of $\lambda \mapsto H(U_\lambda^M|X)$ then follows immediately from the dominated convergence theorem.

Observe that

$$\begin{aligned} I(X;U_\lambda^M) &= I(X;\mathcal{Q}_M(U_\lambda)) = H(\mathcal{Q}_M(U_\lambda)) - H(\mathcal{Q}_M(U_\lambda)|X) \\ &= [H(\mathcal{Q}_M(U_\lambda)) - M] - \int_{\mathbb{R}} f_X(x)[H(\mathcal{Q}_M(U_\lambda)|X=x) - M]\mathrm{d}x. \end{aligned}$$

By Lemma B.3, the integrand is decreasing in M, and thus we can take the limit with respect to M inside the integral. Thus,

$$\lim_{M\to\infty} I(X;U_\lambda^M) = h(U_\lambda) - h(U_\lambda|X) = I(X;U_\lambda).$$

The proof for $I(Y; Z_\lambda^M)$ is analogous. $\square$

Appendix C

Proofs of Chapter 6

C.1 Proof of Theorem 6.10

Before proving Theorem 6.10, we need to establish some technical facts. Recall that $\mathcal{X} = [M]$, $\mathcal{Y} = [N]$, and $\mathcal{Z} = [N+1]$.

Consider the map $\mathcal{H} : \mathcal{F} \to [0,1] \times [0,1]$ given by

$$\mathcal{H}(F) = (\mathcal{P}(F), \mathcal{U}(F)),$$

with $\mathcal{P}(F)$ and $\mathcal{U}(F)$ defined in (6.15). For ease of notation, let $\mathcal{D} = \left\{ D \in \mathcal{M}_{N\times(N+1)} : \|D\| = 1 \right\}$ where $|| \cdot ||$ denotes the Euclidean norm in $\mathcal{M}_{N\times(N+1)} \equiv \mathbb{R}^{N\times(N+1)}$. For $G \in \mathcal{F}$, let

$$\mathcal{D}(G) = \{ D \in \mathcal{D} : G + tD \in \mathcal{F} \text{ for some } t > 0 \} .$$

In graphical terms, $\mathcal{D}$ is the set of all possible directions in $\mathcal{M}_{N\times(N+1)}$ and $\mathcal{D}(G)$ is the set of directions that make $t \mapsto G + tD$ ($t \geq 0$) stay locally in $\mathcal{F}$.

Lemma C.1. *For every $G \in \mathcal{F}$, the set $\mathcal{D}(G)$ is compact.*

Proof. Let $A = \{(y,z) \in \mathcal{Y} \times \mathcal{Z} : G(y,z) = 0\}$ and $B = \{(y,z) \in \mathcal{Y} \times \mathcal{Z} : G(y,z) = 1\}$. It is straightforward to verify that

$$\mathcal{D}(G) = \mathcal{A} \cap \mathcal{B} \cap \mathcal{C} \cap \mathcal{D},$$

where

$$\mathcal{A} = \bigcap_{(y,z)\in A} \left\{ D \in \mathcal{M}_{N\times(N+1)} : D(y,z) \geq 0 \right\},$$
$$\mathcal{B} = \bigcap_{(y,z)\in B} \left\{ D \in \mathcal{M}_{N\times(N+1)} : D(y,z) \leq 0 \right\},$$
$$\mathcal{C} = \left\{ D \in \mathcal{M}_{N\times(N+1)} : \sum_{z=1}^{N+1} D(y,z) = 0,\ y \in \mathcal{Y} \right\}.$$

Observe that since sets $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ are closed, so is $\mathcal{D}(G)$. Since $\mathcal{D}$ is bounded, we have that $\mathcal{D}(G)$ is bounded as well. In particular, $\mathcal{D}(G)$ is closed and bounded and thus compact. □

The following lemma shows the local linear nature of the mapping $\mathcal{H}$. Let $[G_1, G_2] = \{\lambda G_1 + (1-\lambda)G_2 : \lambda \in [0,1]\}$.

Lemma C.2. *For every $G \in \mathcal{F}$, there exists $\delta > 0$ such that $F \mapsto \mathcal{H}(F)$ is linear on $[G, G + \delta D]$ for every $D \in \mathcal{D}(G)$.*

Proof. Let $P = [P(x,y)]_{x\in\mathcal{X},y\in\mathcal{Y}}$ be the joint probability matrix of X and Y, and Q the diagonal matrix with $q_1, \ldots, q_N$ as diagonal entries, where $q_y = \Pr(Y = y)$ for $y \in \mathcal{Y}$. For $G \in \mathcal{F}$ fixed, consider the function $\tau : \mathcal{D}(G) \to \mathbb{R}$ given by

$$\tau(D) = \sup\{t \geq 0 \mid G + tD \in \mathcal{F}\}.$$

Using the fact that $\mathcal{F}$ is a convex polytope, it can be shown that τ is continuous. The definition of $\mathcal{D}(G)$ clearly implies that $\tau(D) > 0$ for all $D \in \mathcal{D}(G)$. For $x \in \mathcal{X}$, $z \in \mathcal{Z}$, and $D \in \mathcal{D}(G)$, consider the function $f_{x,z}^{(D)} : \mathbb{R} \to \mathbb{R}$ given by

$$f_{x,z}^{(D)}(t) := [PG](x,z) + t[PD](x,z), \tag{C.1}$$

where PG (resp., PD) is the product of matrices P and G (resp., P and D). Note that $\mathcal{P}(G+tD) = \sum_{z\in\mathcal{Z}} \max_{x\in\mathcal{X}} f_{x,z}^{(D)}(t)$ for all $t \in [0, \tau(D)]$ (see (6.15)). Let

$$a_z = \max_{x\in\mathcal{X}}[PG](x,z), \quad \mathcal{M}_z = \{x \in \mathcal{X} : [PG](x,z) = a_z\}, \quad \text{and} \quad b_z^{(D)} = \max_{x\in\mathcal{M}_z}[PD](x,z). \tag{C.2}$$

Let $t_{x,z}^{(D)} := -\dfrac{a_z - [PG](x,z)}{b_z^{(D)} - [PD](x,z)}$ whenever $[PD](x,z) \neq b_z^{(D)}$, and $t_{x,z}^{(D)} = \infty$ otherwise. Notice that $f_{x,z}^{(D)}(t_{x,z}^{(D)}) = a_z + t_{x,z}^{(D)} b_z^{(D)}$. Since $t_{x,z}^{(D)} \neq 0$ for all $x \notin \mathcal{M}_z$,

$$t^{(D)} := \min_{z\in\mathcal{Z}} \min_{x\notin\mathcal{M}_z} \min\{|t_{x,z}^{(D)}|, \tau(D)\} > 0.$$

It is easy to see that $a_z + tb_z^{(D)} = \max_{x\in\mathcal{X}} f_{x,z}^{(D)}(t)$ for all $t \in [0, t^{(D)}]$. In particular,

$$\mathcal{P}(G+tD) = \sum_{z=1}^{N+1} \max_{x\in\mathcal{X}} f_{x,z}^{(D)}(t) = \sum_{z=1}^{N+1} a_z + t\sum_{z=1}^{N+1} b_z^{(D)} = \mathcal{P}(G) + tb^{(D)}, \tag{C.3}$$

for every $D \in \mathcal{D}(G)$ and $t \in [0, t^{(D)}]$, where $b^{(D)} := \sum_{z=1}^{N+1} b_z^{(D)}$. Consequently, $\mathcal{P}$ is linear on $[G, G + t^{(D)}D]$. Since $\tau : \mathcal{D}(G) \to \mathbb{R}$ is continuous and bounded, it follows that the map $D \mapsto \min\{|t_{x,z}^{(D)}|, \tau(D)\}$ ($x \notin \mathcal{M}_z$) is also continuous. In particular, the map $D \mapsto t^{(D)}$ is continuous. By compactness of $\mathcal{D}(G)$ established in Lemma C.1, we conclude that $\delta_{\mathcal{P}} := \min_{D\in\mathcal{D}(G)} t^{(D)} > 0$. Thus, $\mathcal{P}$ is linear on $[G, G + \delta_{\mathcal{P}}D]$ for every $D \in \mathcal{D}(G)$.

For $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, and $D \in \mathcal{D}(G)$, consider the function $g_{y,z}^{(D)} : \mathbb{R} \to \mathbb{R}$ given by

$$g_{y,z}^{(D)}(t) = [QG](y,z) + t[QD](y,z).$$

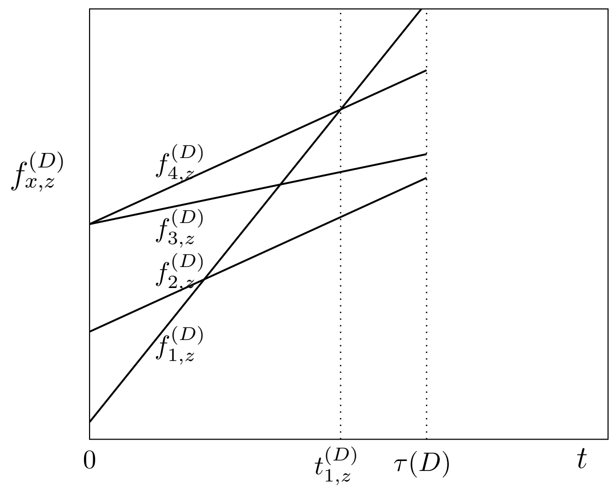


Figure C.1: Typical functions $f_{x,z}^{(D)}$ ($x \in \{1,2,3,4\}$) for a given $z \in \mathcal{Z}$ and $D \in \mathcal{D}(G)$. In this example, we have $\mathcal{M}_z = \{3,4\}$ and $a_z + tb_z^{(D)} = f_{4,z}^{(D)}(t)$. Notice that $t_{2,z}^{(D)} = \infty$ and $t_{3,z}^{(D)} = t_{4,z}^{(D)} = 0$.

Observe that $\mathcal{U}(G+tD) = \sum_{z\in\mathcal{Z}} \max_{y\in\mathcal{Y}} g_{y,z}^{(D)}(t)$ for all $t \in [0, \tau(D)]$ (see (6.15)). Similarly to (C.2), let

$$\alpha_z = \max_{y\in\mathcal{Y}}[QG](y,z), \quad \mathcal{N}_z = \{y \in \mathcal{Y} : [QG](y,z) = \alpha_z\}, \quad \text{and} \quad \beta_z^{(D)} = \max_{y\in\mathcal{N}_z}[QD](y,z).$$

Using a similar argument that resulted in (C.3), it can be shown that there exists $\delta_U > 0$ such that

$$\mathcal{U}(G+tD) = \sum_{z=1}^{N+1} g_{y_z,z}^{(D)}(t) = \sum_{z=1}^{N+1} \alpha_z + t\sum_{z=1}^{N+1} \beta_z^{(D)} = \mathcal{U}(G) + t\beta^{(D)}, \tag{C.4}$$

for every $D \in \mathcal{D}(G)$ and $t \in [0, \delta_{\mathcal{U}}]$, where $\beta^{(D)} := \sum_{z=1}^{N+1} \beta_z^{(D)}$. Consequently, $\mathcal{U}$ is linear on $[G, G+\delta_{\mathcal{U}}D]$ for every $D \in \mathcal{D}(G)$. Therefore, $F \mapsto \mathcal{H}(F) = (\mathcal{P}(F), \mathcal{U}(F))$ is linear on $[G, G+\delta D]$ for every $D \in \mathcal{D}(G)$, where $\delta = \min(\delta_{\mathcal{P}}, \delta_{\mathcal{U}})$. $\square$

We say that a filter $F \in \mathcal{F}$ is optimal if $\mathcal{U}(F) = \hslash(\mathcal{P}(F))$. If F is an optimal filter and $\mathcal{P}(F) = \varepsilon$, we say that F is optimal at ε. The following result is a straightforward application of the concavity of $\hslash$, and thus its proof is omitted.

Lemma C.3. *For $G \in \mathcal{F}$, let $\delta > 0$ be as in Lemma C.2. If there exist $D \in \mathcal{D}(G)$ and $0 < t_1 < t_2 \leq \delta$ such that G, $G + t_1 D$ and $G + t_2 D$ are optimal filters, then $G + tD$ is an optimal filter for each $t \in [0, \delta]$.*

A function $[\mathsf{P_c}(X), \mathsf{P_c}(X|Y)] \ni \varepsilon \mapsto F_\varepsilon \in \mathcal{F}$ is called a *path* of optimal filters if $\mathcal{P}(F_\varepsilon) = \varepsilon$ and $\mathcal{U}(F_\varepsilon) = \hslash(\varepsilon)$ for every $\varepsilon \in [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$. As mentioned in Section 6.3.2, for every ε there exists F_ε such that $\mathcal{P}(F_\varepsilon) = \varepsilon$ and $\mathcal{U}(F_\varepsilon) = \hslash(\varepsilon)$, i.e., a path of optimal filters always exists. In the rest of this section we establish the existence of a *piecewise* linear path of optimal filters.

Lemma C.4. *For every $\varepsilon \in [\mathsf{P_c}(X), \mathsf{P_c}(X|Y))$, there exists $F_\varepsilon \in \mathcal{F}$ and $D \in \mathcal{D}(F_\varepsilon)$ such that F_ε is an optimal filter at ε, $\mathcal{P}(F_\varepsilon + \delta D) > \varepsilon$, and $F_\varepsilon + tD$ is an optimal filter for each $t \in [0, \delta]$ with $\delta > 0$ as in Lemma C.2 for F_ε.*

Proof. Let $K = 2(\mathsf{P_c}(X|Y) - \varepsilon)^{-1}$. For every $n, m > K$, let $G_{n,m}$ be an optimal filter at $\varepsilon + \frac{1}{n} + \frac{1}{m}$. For every $n > K$, the set $\{G_{n,m} : m > K\}$ is an infinite set. Since $\mathcal{F}$ is compact, $\{G_{n,m} : m > K\}$ has at least one accumulation point, say G_n. Let $(G_{n,m_k})_{k\geq 1} \subset \{G_{n,m} : m > K\}$ be a subsequence with $\lim_k G_{n,m_k} = G_n$. By continuity of $\mathcal{P}, \mathcal{U}$, and $\hslash$, we have that

$$\mathcal{P}(G_n) = \lim_{k\to\infty} \mathcal{P}(G_{n,m_k}) = \varepsilon + \frac{1}{n},$$
$$\mathcal{U}(G_n) = \lim_{k\to\infty} \mathcal{U}(G_{n,m_k}) = \lim_{k\to\infty} \hslash(\mathcal{P}(G_{n,m_k})) = \hslash(\mathcal{P}(G_n)),$$

i.e., G_n is an optimal filter at $\varepsilon + \frac{1}{n}$. By the same arguments as before, the set $\{G_n : n > K\}$ has at least one accumulation point, say F_ε, and this accumulation point is an optimal filter at ε. Let $\delta > 0$ be as in Lemma C.2 for F_ε. By construction of F_ε, there exists $n_1 > K$ such that $\|G_{n_1} - F_\varepsilon\| < \frac{\delta}{2}$. The filter G_{n_1} can be written as $G_{n_1} = F_\varepsilon + t_1 D_1$ with $t_1 \in (0, \frac{\delta}{2})$ and $D_1 \in \mathcal{D}(F_\varepsilon)$. Recall that, by (C.3) and (C.4), for every $D \in \mathcal{D}(F_\varepsilon)$ and $t \in [0, \delta]$,

$$\mathcal{P}(F_\varepsilon + tD) = \varepsilon + tb^{(D)} \quad \text{and} \quad \mathcal{U}(F_\varepsilon + tD) = \hslash(\varepsilon) + t\beta^{(D)}.$$

Notice that the maps $D \mapsto b^{(D)}$ and $D \mapsto \beta^{(D)}$ are continuous. Since $\mathcal{P}(G_{n_1}) = \varepsilon + \frac{1}{n_1} > \varepsilon$, we conclude that $b^{(D_1)} > 0$ and, in particular, $\mathcal{P}(F_\varepsilon + \delta D_1) > \varepsilon$.

Let $(G_{n_1,m_k})_{k\geq 1} \subset \{G_{n_1,m} : m > K\}$ be such that $\lim_k G_{n_1,m_k} = G_{n_1}$. For k large enough, we can write $G_{n_1,m_k} = F_\varepsilon + \theta_k E_k$ with $\theta_k \in [0,\delta]$ and $E_k \in \mathcal{D}(F_\varepsilon)$. Since $\theta_k \to t_1$ and $E_k \to D_1$ as $k \to \infty$, there exists $n_2 > K$ such that $\theta_{n_2} < \frac{\delta}{2}$ and $|b^{(E_{n_2})} - b^{(D_1)}| < \frac{b^{(D_1)}}{2}$. Let $t_2 := \theta_{n_2}$ and $D_2 := E_{n_2}$. Clearly, $t_2 < \frac{\delta}{2}$ and $\frac{1}{2}b^{(D_1)} < b^{(D_2)} < 2b^{(D_1)}$. These inequalities yield $\mathcal{P}(F_\varepsilon + \delta D_1) > \mathcal{P}(F_\varepsilon + t_2 D_2)$ and $\mathcal{P}(F_\varepsilon + \delta D_2) > \mathcal{P}(F_\varepsilon + t_1 D_1)$. Thus, there exist $s_1, s_2 \in [0,\delta]$ such that $\mathcal{P}(F_\varepsilon + t_2 D_2) = \mathcal{P}(F_\varepsilon + s_1 D_1)$ and $\mathcal{P}(F_\varepsilon + t_1 D_1) = \mathcal{P}(F_\varepsilon + s_2 D_2)$. In particular,

$$\varepsilon + t_2 b^{(D_2)} = \varepsilon + s_1 b^{(D_1)} \quad \text{and} \quad \varepsilon + t_1 b^{(D_1)} = \varepsilon + s_2 b^{(D_2)}. \tag{C.5}$$

By the optimality of $G_{n_1} = F_\varepsilon + t_1 D_1$ and $G_{n_1,m_{n_2}} = F_\varepsilon + t_2 D_2$,

$$\begin{aligned}
\mathcal{U}(F_\varepsilon + t_2 D_2) &= \hbar(\varepsilon) + t_2 \beta^{(D_2)} \geq \hbar(\varepsilon) + s_1 \beta^{(D_1)} = \mathcal{U}(F_\varepsilon + s_1 D_1), \\
\mathcal{U}(F_\varepsilon + t_1 D_1) &= \hbar(\varepsilon) + t_1 \beta^{(D_1)} \geq \hbar(\varepsilon) + s_2 \beta^{(D_2)} = \mathcal{U}(F_\varepsilon + s_2 D_2).
\end{aligned}$$

By the equations in (C.5), the above inequalities are in fact equalities. In particular, F_ε, $F_\varepsilon + t_1 D_1$ and $F_\varepsilon + s_1 D_1$ are optimal filters. Invoking Lemma C.3, we conclude that $F_\varepsilon + tD_1$ is an optimal filter for all $t \in [0,\delta]$. $\square$

Using an analogous proof, we can also prove the following lemma.

Lemma C.5. *For every $\varepsilon \in (\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$, there exists $F_\varepsilon \in \mathcal{F}$ and $D \in \mathcal{D}(F_\varepsilon)$ such that F_ε is an optimal filter at ε, $\mathcal{P}(F_\varepsilon + \delta D) < \varepsilon$, and $F_\varepsilon + tD$ is an optimal filter for each $t \in [0,\delta]$ with $\delta > 0$ as in Lemma C.2 for F_ε.*

We are in position to prove Theorem 6.10.

Proof of Theorem 6.10. For notational simplicity, we define $S := \mathsf{P_c}(X)$ and $T := \mathsf{P_c}(X|Y)$. In light of Lemmas C.4 and C.5, for every $\varepsilon \in (S,T)$ there exist optimal filters F_ε and G_ε at ε, $\delta_\varepsilon > 0$,

$D_\varepsilon \in \mathcal{D}(F_\varepsilon)$, and $E_\varepsilon \in \mathcal{D}(G_\varepsilon)$ such that $F_\varepsilon + tD_\varepsilon$ and $G_\varepsilon + tE_\varepsilon$ are optimal filters for each $t \in [0, \delta_\varepsilon]$, and $\mathcal{P}(G_\varepsilon + \delta_\varepsilon E_\varepsilon) < \varepsilon < \mathcal{P}(F_\varepsilon + \delta_\varepsilon D_\varepsilon)$. Note that $\delta_\varepsilon = \min\{\delta_{F_\varepsilon}, \delta_{G_\varepsilon}\}$, where δ_{F_ε} and δ_{G_ε} are the constants obtained in Lemma C.2 for filters F_ε and G_ε, respectively. For every $\varepsilon \in (S, T)$, let $V_\varepsilon = (\mathcal{P}(F_\varepsilon + \delta_\varepsilon E_\varepsilon), \mathcal{P}(G_\varepsilon + \delta_\varepsilon D_\varepsilon))$. Similarly, there exist

a) an optimal filter F_S at S, $\delta_S > 0$, and $D_S \in \mathcal{D}(F_S)$ such that $F_S + tD_S$ is an optimal filter for each $t \in [0, \delta_S]$ and $\mathcal{P}(F_S + \delta_S D_S) > S$;

b) an optimal filter G_T at T, $\delta_T > 0$, and $E_T \in \mathcal{D}(G_T)$ such that $G_T + tE_T$ is an optimal filter for each $t \in [0, \delta_T]$ and $\mathcal{P}(G_T + \delta_T E_T) < T$.

Let $V_S = [S, \mathcal{P}(F_S + \delta_S D_S))$ and $V_T = (\mathcal{P}(G_T + \delta_T E_T), T]$. The family $\{V_\varepsilon : \varepsilon \in [S, T]\}$ forms an open cover of $[S, T]$ (in the subspace topology). By compactness, there exist $S = \varepsilon_0 < \cdots < \varepsilon_l = T$ such that $\{V_{\varepsilon_0}, \ldots, V_{\varepsilon_l}\}$ forms an open cover for $[S, T]$. For each $i \in \{0, \ldots, l-1\}$, the mapping

$$[\varepsilon_i, \mathcal{P}(F_{\varepsilon_i} + \delta_{\varepsilon_i} D_{\varepsilon_i})) \ni \varepsilon \mapsto F_{\varepsilon_i} + \frac{\varepsilon - \varepsilon_i}{b^{(D_{\varepsilon_i})}} D_{\varepsilon_i} \in \mathcal{F}, \tag{C.6}$$

is clearly linear. Similarly, for each $i \in \{0, \ldots, l-1\}$, the mapping

$$(\mathcal{P}(G_{\varepsilon_i} + \delta_{\varepsilon_i} E_{\varepsilon_i}), \varepsilon_i] \ni \varepsilon \mapsto G_{\varepsilon_i} + \frac{\varepsilon - \varepsilon_i}{b^{(E_{\varepsilon_i})}} E_{\varepsilon_i} \in \mathcal{F}, \tag{C.7}$$

is also linear. Notice that $\mathcal{P}\left(F_{\varepsilon_i} + \frac{\varepsilon - \varepsilon_i}{b^{(D_{\varepsilon_i})}} D_{\varepsilon_i}\right) = \varepsilon = \mathcal{P}\left(G_{\varepsilon_i} + \frac{\varepsilon - \varepsilon_i}{b^{(E_{\varepsilon_i})}} E_{\varepsilon_i}\right)$. Since $\{V_{\varepsilon_0}, \ldots, V_{\varepsilon_l}\}$ forms an open cover for $[S, T]$, the mappings in (C.6) and (C.7) implement a piecewise linear path of optimal filters. □

The proof provided in this appendix establishes the existence of $\delta_* > 0$, an optimal filter F_* at $T := \mathsf{P}_\mathsf{c}(X|Y)$, and $D_* \in \mathcal{D}(F_*)$ such that $\mathcal{P}(F_* + \delta_* D_*) < T$ (or equivalently $b^{(D_*)} < 0$) and

$$\hbar(\varepsilon) = 1 + (\varepsilon - T)\frac{\beta^{(D_*)}}{b^{(D_*)}},$$

for every $\varepsilon \in [T + \delta_* b^{(D_*)}, T]$. The previous equation and the maximality of $\hbar(\varepsilon)$ imply that

$$\hbar'(T) = \min_{\substack{F\in\mathcal{F}\\ \mathcal{P}(F)=T}} \min_{\substack{D\in\mathcal{D}(F)\\ b^{(D)}<0}} \frac{\beta^{(D)}}{b^{(D)}}. \tag{C.8}$$

C.2 Proof of Theorem 6.14

We first note that since $\hbar$ is concave on $[\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$, its right derivative exists at $\mathsf{P_c}(X|Y)$. Therefore, we have by concavity

$$\hbar(\varepsilon) \leq 1 - (\mathsf{P_c}(X|Y) - \varepsilon)\hbar'(\mathsf{P_c}(X|Y)), \tag{C.9}$$

for all $\varepsilon \in [p, \mathsf{P_c}(X|Y)]$. In Lemma C.6 below, we show that

$$\hbar'(\mathsf{P_c}(X|Y)) = \frac{q}{\bar{\beta}p - \alpha\bar{p}} 1_{\{\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2\}} + \frac{\bar{q}}{\bar{\alpha}\bar{p} - \beta p} 1_{\{\alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2\}}.$$

Thus, (C.9) becomes

$$\hbar(\varepsilon) \leq \begin{cases} 1 - \zeta(\varepsilon)q, & \alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2, \\ 1 - \tilde{\zeta}(\varepsilon)\bar{q}, & \alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2. \end{cases} \tag{C.10}$$

To finish the proof of Theorem 6.14 we show that the Z-channel $\mathsf{Z}(\zeta(\varepsilon))$ and the reverse Z-channel $\tilde{\mathsf{Z}}(\tilde{\zeta}(\varepsilon))$ achieve (C.9) and (C.10), when $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$ and $\alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2$, respectively.

For $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$, consider the filter $P_{Z|Y} = \begin{bmatrix} 1 & 0 \\ \zeta(\varepsilon) & 1-\zeta(\varepsilon) \end{bmatrix}$. Notice that

$$P_{XZ} = \begin{bmatrix} \bar{p}(\bar{\alpha} + \alpha\zeta(\varepsilon)) & \bar{p}\alpha(1-\zeta(\varepsilon)) \\ p(\beta + \bar{\beta}\zeta(\varepsilon)) & p\bar{\beta}(1-\zeta(\varepsilon)) \end{bmatrix} \quad \text{and} \quad P_{YZ} = \begin{bmatrix} \bar{q} & 0 \\ q\zeta(\varepsilon) & q(1-\zeta(\varepsilon)) \end{bmatrix}. \tag{C.11}$$

It is straightforward to verify that $\bar{p}(\bar{\alpha} + \alpha\zeta(\varepsilon)) \geq p(\beta + \bar{\beta}\zeta(\varepsilon))$. As a consequence, $\mathsf{P_c}(X|Z) = \varepsilon$. Since $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$, we have that $\frac{\bar{q}}{q} > \zeta(\varepsilon)$. Thus, $\mathsf{P_c}(Y|Z) = 1 - \zeta(\varepsilon)q$.

For $\alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2$, consider the filter $P_{Z|Y} = \begin{bmatrix} 1-\tilde{\zeta}(\varepsilon) & \tilde{\zeta}(\varepsilon) \\ 0 & 1 \end{bmatrix}$. Notice that

$$P_{XZ} = \begin{bmatrix} \bar{p}\bar{\alpha}(1-\tilde{\zeta}(\varepsilon)) & \bar{p}(\alpha+\bar{\alpha}\tilde{\zeta}(\varepsilon)) \\ p\beta(1-\tilde{\zeta}(\varepsilon)) & p(\bar{\beta}+\beta\tilde{\zeta}(\varepsilon)) \end{bmatrix} \quad \text{and} \quad P_{YZ} = \begin{bmatrix} \bar{q}(1-\tilde{\zeta}(\varepsilon)) & \bar{q}\tilde{\zeta}(\varepsilon) \\ 0 & q \end{bmatrix}. \tag{C.12}$$

Recall that $\bar{\alpha}\bar{p} > \beta p$ and also observe that $p(\bar{\beta}+\beta\tilde{\zeta}(\varepsilon)) \geq \bar{p}(\alpha+\bar{\alpha}\tilde{\zeta}(\varepsilon))$. As a consequence, $\mathsf{P_c}(X|Z) = \varepsilon$. The fact that $\alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2$ implies $q \geq \bar{q}\tilde{\zeta}(\varepsilon)$. Therefore, $\mathsf{P_c}(Y|Z) = 1-\tilde{\zeta}(\varepsilon)\bar{q}$.

Lemma C.6. *Let $X \sim \mathsf{Bernoulli}(p)$ with $p \in [\frac{1}{2}, 1)$ and $P_{Y|X} = \mathsf{BIBO}(\alpha,\beta)$ with $\alpha,\beta \in [0, \frac{1}{2})$ such that $\bar{\alpha}\bar{p} > \beta p$. Then $\hbar'(\mathsf{P_c}(X|Y)) = \dfrac{q}{\bar{\beta}p-\alpha\bar{p}} 1_{\{\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2\}} + \dfrac{\bar{q}}{\bar{\alpha}\bar{p}-\beta p} 1_{\{\alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2\}}$.*

Proof. As before, let $T := \mathsf{P_c}(X|Y)$. We begin the proof by noticing that the Z-channels defined in (C.11) and (C.12) provide a lower bound on $\hbar(\varepsilon)$ as follows:

$$\hbar(\varepsilon) \geq 1 - \zeta(\varepsilon) q 1_{\{\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2\}} - \tilde{\zeta}(\varepsilon)\bar{q} 1_{\{\alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2\}}. \tag{C.13}$$

By concavity of $\hbar$, this inequality implies

$$\hbar'(T) \leq \frac{q}{\bar{\beta}p-\alpha\bar{p}} 1_{\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2} + \frac{\bar{q}}{\bar{\alpha}\bar{p}-\beta p} 1_{\alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2}.$$

The rest of the proof is devoted to establishing the reverse inequality. To this end, we use the variational formula for $\hbar'(T)$ given in (C.8). Let $P = [P(x,y)]_{x,y\in\{0,1\}}$ be the joint probability matrix of X and Y. Without loss of generality we can assume $\mathcal{Z} = \{z_1, z_2, z_3\}$. It follows from (C.3) and (C.4) that for every $F \in \mathcal{F} \subset \mathcal{M}_{2\times 3}$ there exists $\delta > 0$ such that

$$\mathcal{P}(F+tD) = \mathcal{P}(F) + tb^{(D)} \quad \text{and} \quad \mathcal{U}(F+tD) = \mathcal{U}(F) + t\beta^{(D)}, \tag{C.14}$$

for every $t \in [0,\delta]$ and $D \in \mathcal{D}(F)$, where $b^{(D)} = \displaystyle\sum_{i=1}^{3} \max_{x\in\mathcal{M}_{z_i}} [PD](x,z_i)$ and $\beta^{(D)} =$

$\sum_{i=1}^{3} \max_{y \in \mathcal{N}_{z_i}} q(y)D(y, z_i)$ with

$$\mathcal{M}_{z_i} = \Big\{x \in \{0,1\} : (PF)(x, z_i) = \max_{x' \in \{0,1\}} (PF)(x', z_i)\Big\},$$
$$\mathcal{N}_{z_i} = \Big\{y \in \{0,1\} : q(y)F(y, z_i) = \max_{y' \in \{0,1\}} q(y')F(y', z_i)\Big\}.$$

Up to permutation of columns, which corresponds to permuting the elements of $\mathcal{Z}$, the set of filters $F \in \mathcal{F}$ such that $\mathcal{P}(F) = T$ equals

$$\left\{\begin{bmatrix} 1 & 0 & 0 \\ 0 & u & v \end{bmatrix} : \begin{matrix} 0 < v \le u \\ u + v = 1 \end{matrix}\right\} \bigcup \left\{\begin{bmatrix} 0 & u & v \\ 1 & 0 & 0 \end{bmatrix} : \begin{matrix} 0 < v \le u \\ u + v = 1 \end{matrix}\right\} \bigcup \left\{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}\right\}. \tag{C.15}$$

To compute $\hbar'(T)$ using formula (C.8) we need to compute $\beta^{(D)}$ and $b^{(D)}$ for each $D \in \mathcal{D}(F)$ with F of the form described in (C.15).

Let $F = \begin{bmatrix} 1 & 0 & 0 \\ 0 & u & v \end{bmatrix}$ for some $0 < v \le u$ and $u + v = 1$. A direct computation shows that

$$PF = \begin{bmatrix} \bar{\alpha}\bar{p} & u\alpha\bar{p} & v\alpha\bar{p} \\ \beta p & u\bar{\beta}p & v\bar{\beta}p \end{bmatrix}. \tag{C.16}$$

In particular, $\mathcal{M}_{z_1} = \{0\}$, $\mathcal{M}_{z_2} = \{1\}$, and $\mathcal{M}_{z_3} = \{1\}$. For every $D \in \mathcal{D}(F)$,

$$PD = \begin{bmatrix} \bar{\alpha}\bar{p}D_{11} + \alpha\bar{p}D_{21} & \bar{\alpha}\bar{p}D_{12} + \alpha\bar{p}D_{22} & \bar{\alpha}\bar{p}D_{13} + \alpha\bar{p}D_{23} \\ \beta pD_{11} + \bar{\beta}pD_{21} & \beta pD_{12} + \bar{\beta}pD_{22} & \beta pD_{13} + \bar{\beta}pD_{23} \end{bmatrix},$$

and hence $b^{(D)} = \bar{\alpha}\bar{p}D_{11} + \alpha\bar{p}D_{21} + \beta pD_{12} + \bar{\beta}pD_{22} + \beta pD_{13} + \bar{\beta}pD_{23}$. Notice that, for $1 \le i \le 3$, we have that $D_{i1} + D_{i2} + D_{i3} = 0$. In particular, $b^{(D)} = (\bar{\alpha}\bar{p} - \beta p)D_{11} + (\alpha\bar{p} - \bar{\beta}p)D_{21}$. Consider the matrices,

$$\begin{bmatrix} \bar{q} & 0 \\ 0 & q \end{bmatrix} F = \begin{bmatrix} \bar{q} & 0 & 0 \\ 0 & qu & qv \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \bar{q} & 0 \\ 0 & q \end{bmatrix} D = \begin{bmatrix} \bar{q}D_{11} & \bar{q}D_{12} & \bar{q}D_{13} \\ qD_{21} & qD_{22} & qD_{33} \end{bmatrix},$$

from which we obtain $\mathcal{N}_{z_1} = \{0\}, \mathcal{N}_{z_2} = \{1\}, \mathcal{N}_{z_3} = \{1\}$, and therefore, $\beta^{(D)} = \bar{q}D_{11} + qD_{22} + qD_{23} = \bar{q}D_{11} - qD_{21}$. In what follows we use the simple fact that $\dfrac{ax+y}{bx+y} \geq \min\left\{\dfrac{a}{b}, 1\right\}$ for $a, b > 0$ and $x, y \geq 0$ with $x + y > 0$. For notational simplicity, let $\eta \coloneqq \frac{\bar{q}}{q}$ and $\zeta \coloneqq \zeta(p)$, where $\zeta(\cdot)$ is defined in (6.17).

From the form of F, it is clear that $-D_{11} \geq 0$ and $D_{21} \geq 0$. If $b^{(D)} < 0$, then D_{11} and D_{21} cannot be simultaneously zero, and hence

$$\frac{\beta^{(D)}}{b^{(D)}} = \frac{q}{\bar{\beta}p - \alpha\bar{p}} \frac{\eta(-D_{11}) + D_{21}}{\zeta(-D_{11}) + D_{21}} \geq \frac{q}{\bar{\beta}p - \alpha\bar{p}} \min\left\{\frac{\eta}{\zeta}, 1\right\} = \begin{cases} \frac{q}{\bar{\beta}p - \alpha\bar{p}}, & \alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2, \\ \frac{\bar{q}}{\bar{\alpha}\bar{p} - \beta p}, & \alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2. \end{cases}$$

In particular, we obtain that

$$\min_{\substack{D \in \mathcal{D}(F) \\ b^{(D)} < 0}} \frac{\beta^{(D)}}{b^{(D)}} \geq \begin{cases} \frac{q}{\bar{\beta}p - \alpha\bar{p}}, & \alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2, \\ \frac{\bar{q}}{\bar{\alpha}\bar{p} - \beta p}, & \alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2. \end{cases} \tag{C.17}$$

The case $F = \begin{bmatrix} 0 & u & v \\ 1 & 0 & 0 \end{bmatrix}$ for $0 < v \leq u$ and $u + v = 1$ is analogous.

Now, let $F = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$. By (C.16) with $u = 1$ and $v = 0$, we obtain that $\mathcal{M}_{z_1} = \{0\}$, $\mathcal{M}_{z_2} = \{1\}$, and $\mathcal{M}_{z_3} = \{0, 1\}$. In a similar way, $\mathcal{N}_{z_1} = \{0\}$, $\mathcal{N}_{z_2} = \{1\}$, and $\mathcal{N}_{z_3} = \{0, 1\}$. Hence

$$b^{(D)} = \bar{\alpha}\bar{p}D_{11} + \alpha\bar{p}D_{21} + \beta pD_{12} + \bar{\beta}pD_{22} + \max\{\bar{\alpha}\bar{p}D_{13} + \alpha\bar{p}D_{23}, \beta pD_{13} + \bar{\beta}pD_{23}\},$$

$$\beta^{(D)} = \bar{q}D_{11} + qD_{22} + \max\{\bar{q}D_{13}, qD_{23}\}.$$

We therefore need to consider the following cases:

Case I: $\bar{\alpha}\bar{p}D_{13} + \alpha\bar{p}D_{23} \le \beta pD_{13} + \bar{\beta}pD_{23}$ and $\bar{q}D_{13} \le qD_{23}$. The computation in this case reduces to the computation for $F = \begin{bmatrix} 1 & 0 & 0 \\ 0 & u & v \end{bmatrix}$.

Case II: $\bar{\alpha}\bar{p}D_{13} + \alpha\bar{p}D_{23} \le \beta pD_{13} + \bar{\beta}pD_{23}$ and $\bar{q}D_{13} > qD_{23}$. Notice that these conditions imply that $\zeta D_{13} \le D_{23} < \eta D_{13}$, and therefore this case requires $\zeta < \eta$ (or equivalently, $\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2$). This yields

$$b^{(D)} = (\bar{\alpha}\bar{p} - \beta p)D_{11} + (\alpha\bar{p} - \bar{\beta}p)D_{21} \quad \text{and} \quad \beta^{(D)} = qD_{22} - \bar{q}D_{12}.$$

Hence, we have

$$\frac{\beta^{(D)}}{b^{(D)}} = \frac{q}{\bar{\beta}p - \alpha\bar{p}} \frac{D_{22} - \eta D_{12}}{\zeta D_{11} - D_{21}}.$$

By the form of F, we have that $-D_{11}, D_{12}, D_{21} \ge 0$. The inequalities $\zeta < \eta$ and $\zeta D_{13} \le D_{23}$ imply that $\dfrac{D_{22} - \eta D_{12}}{\zeta D_{11} - D_{21}} \ge 1$, and hence

$$\frac{\beta^{(D)}}{b^{(D)}} \ge \frac{q}{\bar{\beta}p - \alpha\bar{p}} 1_{\{\alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2\}}. \tag{C.18}$$

Case III: $\bar{\alpha}\bar{p}D_{13} + \alpha\bar{p}D_{23} > \beta pD_{13} + \bar{\beta}pD_{23}$ and $\bar{q}D_{13} \le qD_{23}$. Notice that these conditions imply that $\eta D_{13} \le D_{23} < \zeta D_{13}$, and hence this case requires $\zeta > \eta$ (or equivalently, $\alpha\bar{\alpha}\bar{p}^2 > \beta\bar{\beta}p^2$). In this case, we have

$$b^{(D)} = (\beta p - \bar{\alpha}\bar{p})D_{12} + (\bar{\beta}p - \alpha\bar{p})D_{22} \quad \text{and} \quad \beta^{(D)} = \bar{q}D_{11} - qD_{21}.$$

Therefore,

$$\frac{\beta^{(D)}}{b^{(D)}} = \frac{\bar{q}}{\bar{\alpha}\bar{p} - \beta p} \frac{D_{11} - \eta^{-1} D_{21}}{-D_{12} + \zeta^{-1} D_{22}}.$$

By the form of F, we have that $-D_{22}, D_{12}, D_{21} \geq 0$. The inequalities $\zeta^{-1} < \eta^{-1}$ and $\zeta D_{13} > D_{23}$ imply that $\dfrac{D_{11} - \eta^{-1} D_{21}}{-D_{12} + \zeta^{-1} D_{22}} > 1$, and hence

$$\frac{\beta^{(D)}}{b^{(D)}} > \frac{\bar{q}}{\bar{\alpha}\bar{p} - \beta p} 1_{\{\alpha\bar{\alpha}\bar{p}^2 > \beta\bar{\beta}p^2\}}. \tag{C.19}$$

Case IV: $\bar{\alpha}\bar{p}D_{13} + \alpha\bar{p}D_{23} > \beta p D_{13} + \bar{\beta} p D_{23}$ and $\bar{q}D_{13} > qD_{23}$. Notice that these two inequalities imply that $D_{23} < \min\{\zeta, \eta\} D_{13}$. For this case we have that

$$b^{(D)} = (\beta p - \bar{\alpha}\bar{p}) D_{12} + (\bar{\beta} p - \alpha\bar{p}) D_{22} \quad \text{and} \quad \beta^{(D)} = qD_{22} - \bar{q}D_{12}.$$

Hence, we have

$$\frac{\beta^{(D)}}{b^{(D)}} = \frac{q}{\bar{\beta}p - \alpha\bar{p}} \frac{\eta D_{12} - D_{22}}{\zeta D_{12} - D_{22}}.$$

By the form of F, we have that $-D_{22}, D_{12} \geq 0$. As before, we conclude that

$$\frac{\beta^{(D)}}{b^{(D)}} \geq \frac{q}{\bar{\beta}p - \alpha\bar{p}} \min\left\{\frac{\eta}{\zeta}, 1\right\} = \begin{cases} \frac{q}{\bar{\beta}p - \alpha\bar{p}}, & \alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2, \\ \frac{\bar{q}}{\bar{\alpha}\bar{p} - \beta p}, & \alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2. \end{cases} \tag{C.20}$$

Combining (C.17), (C.18), (C.19), and (C.20), we obtain

$$\min_{\substack{F \in \mathcal{F} \\ \mathcal{P}(F) = T}} \min_{\substack{D \in \mathcal{D}(F) \\ b^{(D)} < 0}} \frac{\beta^{(D)}}{b^{(D)}} \geq \begin{cases} \frac{q}{\bar{\beta}p - \alpha\bar{p}}, & \alpha\bar{\alpha}\bar{p}^2 < \beta\bar{\beta}p^2, \\ \frac{\bar{q}}{\bar{\alpha}\bar{p} - \beta p}, & \alpha\bar{\alpha}\bar{p}^2 \geq \beta\bar{\beta}p^2, \end{cases}$$

as desired. $\square$

C.3 Proof of Theorem 6.17

Recall that $\mathcal{X} = [M]$ and $\mathcal{Y} = \mathcal{Z} = [N]$, and $P = [P(x,y)]_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$ is the joint probability matrix of X and Y, and the marginals are $p_X(x) = \Pr(X = x)$ and $q_Y(y) = \Pr(Y = y)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Similar to $\hbar$, the function $\underline{\hbar}$ admits the alternative formulation

$$\underline{\hbar}(\varepsilon) = \sup_{F\in\underline{\mathcal{F}}\colon\, \underline{\mathcal{P}}(F)\leq\varepsilon} \underline{\mathcal{U}}(F),$$

where $\underline{\mathcal{F}}$ is the set of all stochastic matrices $F \in \mathcal{M}_{N\times N}$,

$$\underline{\mathcal{P}}(F) = \sum_{z\in\mathcal{Z}} \max_{x\in\mathcal{X}} (PF)(x,z), \quad \text{ and } \quad \underline{\mathcal{U}}(F) = \sum_{z\in\mathcal{Z}} \max_{y\in\mathcal{Y}} q_Y(y)F(y,z).$$

We let $\underline{\mathcal{D}} = \{D \in \mathcal{M}_{N\times N} : \|D\| = 1\}$ and, for each $F \in \underline{\mathcal{F}}$, we define

$$\underline{\mathcal{D}}(F) \coloneqq \left\{D \in \underline{\mathcal{D}} : F + tD \in \underline{\mathcal{F}} \text{ for some } t > 0\right\}.$$

Before proving Theorem 6.17, we need to establish some technical lemmas. Notice that the proofs of Lemmas C.1 and C.2 do not depend on the alphabets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$. Therefore, $\underline{\mathcal{D}}(F)$ is compact for any $F \in \underline{\mathcal{F}}$ and also we obtain the following lemma.

Lemma C.7. *Let $\underline{\mathcal{H}} : \underline{\mathcal{F}} \to [0,1] \times [0,1]$ be the mapping given by $\underline{\mathcal{H}}(F) = (\underline{\mathcal{P}}(F), \underline{\mathcal{U}}(F))$. For every $F \in \underline{\mathcal{F}}$, there exists $\delta > 0$ such that $\underline{\mathcal{H}}$ is linear on $[F, F + \delta D]$ for every $D \in \underline{\mathcal{D}}(F)$.*

The convex analysis tools used to study $\hbar$ heavily rely on the fact that $|\mathcal{Z}| = |\mathcal{Y}| + 1$. Hence, they are unavailable in this case, and thus we need an alternative approach to establish the desired functional properties of $\underline{\hbar}$.

Lemma C.8. *If $\mathsf{P_c}(X) < \mathsf{P_c}(X|Y)$, then $\underline{\hbar}$ is continuous at $\mathsf{P_c}(X|Y)$.*

Proof. Without loss of generality, we will assume that $q_Y(1) > 0$. Let $D_* \in \underline{\mathcal{D}}(\mathrm{I}_N)$ be given by

$$D_* = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \lambda & -\lambda & 0 & \cdots & 0 \\ \lambda & 0 & -\lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda & 0 & 0 & \cdots & -\lambda \end{bmatrix},$$

where $\lambda = (2(N-1))^{-1/2}$. As in the proof of Lemma C.2, one can show that there exist $\delta_1 > 0$ and $(x_z)_{z\in\mathcal{Z}} \subset \mathcal{X}$ such that for every $t \in [0, \delta_1]$,

$$\underline{\mathcal{P}}(\mathrm{I}_N + tD_*) = \sum_{z\in\mathcal{Z}} \max_{x\in\mathcal{X}} (P(\mathrm{I}_N + tD_*))(x,z) = \sum_{z\in\mathcal{Z}} (P(\mathrm{I}_N + tD_*))(x_z, z). \tag{C.21}$$

In this case, we have that

$$\begin{aligned} \underline{\mathcal{P}}(\mathrm{I}_N + tD_*) &= P(x_1, 1) + t\lambda \sum_{z=2}^{N} P(x_1, z) + (1 - t\lambda) \sum_{z=2}^{N} P(x_z, z) \\ &= \sum_{z\in\mathcal{Z}} P(x_z, z) - t\lambda \left(\sum_{z\in\mathcal{Z}} P(x_z, z) - P(x_1, z) \right). \end{aligned}$$

Note that $\mathsf{P_c}(X|Y) = \underline{\mathcal{P}}(\mathrm{I}_N) = \sum_{z\in\mathcal{Z}} P(x_z, z)$. Hence,

$$\underline{\mathcal{P}}(\mathrm{I}_N + tD_*) = \mathsf{P_c}(X|Y) - t\lambda\sigma, \tag{C.22}$$

where $\sigma = \sum_{z\in\mathcal{Z}} (P(x_z, z) - P(x_1, z))$. Setting $t = 0$ in (C.21), we have that $P(x_z, z) \geq P(x, z)$ for all $(x, z) \in \mathcal{X} \times \mathcal{Z}$. If $P(x_z, z) = P(x_1, z)$ for all $z \geq 1$, then

$$\mathsf{P_c}(X|Y) = \sum_{z\in\mathcal{Z}} P(x_1, z) = p_X(x_1) \leq \mathsf{P_c}(X),$$

which contradicts the hypothesis of the lemma. Therefore, there exists $z \in \mathcal{Z}$ such that $P(x_z, z) >$

$P(x_1, z)$ and hence $\sigma > 0$. Similarly, there exists $\delta_2 > 0$ such that for every $t \in [0, \delta_2]$,

$$\underline{\mathcal{U}}(\mathrm{I}_N + tD_*) = q_Y(1) + (1 - t\lambda) \sum_{z=2}^{N} q_Y(z) = 1 - t\lambda(1 - q_Y(1)). \tag{C.23}$$

Let $\delta = \min(\delta_1, \delta_2)$. From (C.22) and (C.23), we have for every $t \in [0, \delta]$

$$1 - t\lambda(1 - q_Y(1)) \leq \underline{\hbar}(\mathsf{P_c}(X|Y) - t\lambda\sigma) \leq 1. \tag{C.24}$$

In particular,

$$\lim_{\varepsilon \to \mathsf{P_c}(X|Y)} \underline{\hbar}(\varepsilon) = \lim_{t \to 0} \underline{\hbar}(\mathsf{P_c}(X|Y) - t\lambda\sigma) = 1 = \underline{\hbar}(\mathsf{P_c}(X|Y)),$$

i.e., $\underline{\hbar}$ is continuous at $\mathsf{P_c}(X|Y)$. $\square$

We say that $F \in \underline{\mathcal{F}}$ is an optimal filter at ε if $\underline{\mathcal{U}}(F) = \underline{\hbar}(\varepsilon)$ and $\underline{\mathcal{P}}(F) \leq \varepsilon$. As opposed to $\hbar$, the concavity of $\underline{\hbar}$ is unknown and hence the existence of an optimal filter at ε with $\underline{\mathcal{P}}(F) = \varepsilon$ is not immediate. Nonetheless, since $\underline{\mathcal{P}}$ and $\underline{\mathcal{U}}$ are continuous functions, there exists an optimal filter F at ε (with $\underline{\mathcal{P}}(F) \leq \varepsilon$) for every $\varepsilon \in [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$. For any $F \in \underline{\mathcal{F}}$ and $\delta > 0$, let $B(F, \delta) = \{G \in \underline{\mathcal{F}} : \|G - F\| < \delta\}$.

Lemma C.9. *Let $\delta > 0$ be as in Lemma C.7 for I_N, i.e., $\underline{\mathcal{U}}$ and $\underline{\mathcal{P}}$ are linear on $[\mathrm{I}_N, \mathrm{I}_N + \delta D]$ for every $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$. If $\mathsf{P_c}(X) < \mathsf{P_c}(X|Y)$ and $q_Y(y) > 0$ for all $y \in \mathcal{Y}$, then there exists $\varepsilon_{\mathsf{L}} < \mathsf{P_c}(X|Y)$ such that for every $\varepsilon \in [\varepsilon_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$ there exists an optimal filter F_ε at ε with $F_\varepsilon \in B(\mathrm{I}_N, \delta)$.*

Proof. Let $\underline{\mathcal{F}}^1 = \{F \in \underline{\mathcal{F}} : \underline{\mathcal{U}}(F) = 1\}$ and let $\mathcal{B} = \displaystyle\bigcup_{F \in \underline{\mathcal{F}}^1} B(F, \delta)$. The proof is based on the following claim.

Claim. There exists $\varepsilon_{\mathsf{L}} < \mathsf{P_c}(X|Y)$ such that if F is an optimal filter at ε with $\varepsilon \geq \varepsilon_{\mathsf{L}}$, then $F \in \mathcal{B}$.

Proof of the claim. The proof is by contradiction. Assume that for every $\varepsilon < \mathsf{P_c}(X|Y)$ there exists an optimal filter $G_{\varepsilon'}$ at $\varepsilon' \in [\varepsilon, \mathsf{P_c}(X|Y))$ with $G_{\varepsilon'} \notin \mathcal{B}$. Since $\underline{\hbar}$ is a non-decreasing function, we have that $\underline{\mathcal{U}}(G_{\varepsilon'}) = \underline{\hbar}(\varepsilon') \geq \underline{\hbar}(\varepsilon)$. Let $K := (\mathsf{P_c}(X|Y) - \mathsf{P_c}(X))^{-1}$. For each $n > K$, let $F_n = G_{\mathsf{P_c}(X|Y)-1/n}$. Since $\mathcal{F}\backslash\mathcal{B}$ is compact, there exist $\{n_1 < n_2 < \cdots\}$ and $F \in \mathcal{F}\backslash\mathcal{B}$ such that $F_{n_k} \to F$ as $k \to \infty$. By continuity of $\underline{\mathcal{U}}$ and $\underline{\hbar}$ at $\mathsf{P_c}(X|Y)$, established in the Lemma C.8, we have

$$1 \geq \underline{\mathcal{U}}(F) = \lim_{k\to\infty} \underline{\mathcal{U}}(F_{n_k}) \geq \lim_{k\to\infty} \underline{\hbar}(\mathsf{P_c}(X|Y) - n_k^{-1}) = \underline{\hbar}(\mathsf{P_c}(X|Y)) = 1.$$

In particular, we have that $F \in \underline{\mathcal{F}}^1 \subset \mathcal{B}$, which contradicts the fact that $F \in \mathcal{F}\backslash\mathcal{B}$. $\square$

The assumption $q_Y(y) > 0$ for every $y \in \mathcal{Y}$ implies that $F \in \underline{\mathcal{F}}^1$ if and only if F is a permutation matrix, i.e., F can be obtained by permuting the columns of I_N. In particular, the mapping $G \mapsto GF^{-1}$ is a bijection between $B(F, \delta)$ and $B(\mathrm{I}_N, \delta)$ which preserves $\underline{\mathcal{P}}$ and $\underline{\mathcal{U}}$, i.e., $\underline{\mathcal{P}}(G) = \underline{\mathcal{P}}(GF^{-1})$ and $\underline{\mathcal{U}}(G) = \underline{\mathcal{U}}(GF^{-1})$ for every $G \in B(F, \delta)$. As mentioned earlier, there exists an optimal filter F_ε at ε for every $\varepsilon \in [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$. By the claim, F_ε, for $\varepsilon \geq \varepsilon_{\mathsf{L}}$, belongs to $\mathcal{B}$ and, in particular, $F_\varepsilon \in B(F, \delta)$ for some $F \in \underline{\mathcal{F}}^1$. By the aforementioned properties of the bijection $G \mapsto GF^{-1}$, the filter $F_\varepsilon F^{-1}$ is an optimal filter at ε with $F_\varepsilon F^{-1} \in B(\mathrm{I}_N, \delta)$. $\square$

Now we are in position to prove Theorem 6.17.

Proof of Theorem 6.17. If $q_Y(y) = 0$ for some $y \in \mathcal{Y}$, the effective cardinality of the alphabet of Y is $|\mathcal{Y}| - 1$ and thus $\underline{\hbar}(\varepsilon)$ equals $\hbar(\varepsilon)$ for every $\varepsilon \in [\mathsf{P_c}(X), \mathsf{P_c}(X|Y)]$. In this case, $\underline{\hbar}$ is piecewise linear and (6.18) follows trivially by Theorem 6.10. In what follows, we assume that $q_Y(y) > 0$ for all $y \in \mathcal{Y}$.

Let $\delta > 0$ and $\varepsilon'_{\mathsf{L}} < \mathsf{P_c}(X|Y)$ be as in Lemma C.9. For each $\varepsilon \in [\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y))$, let G_ε be an optimal filter at ε with $G_\varepsilon \in B(\mathrm{I}_N, \delta)$ whose existence was established in Lemma C.9. Let $t_\varepsilon \in [0, \delta]$ and $D_\varepsilon \in \underline{\mathcal{D}}(\mathrm{I}_N)$ be such that $G_\varepsilon = \mathrm{I}_N + t_\varepsilon D_\varepsilon$ for every $\varepsilon \in [\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y))$. As in (C.3) and (C.4) in the proof of Lemma C.2, for every $t \in [0, \delta]$ and $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$,

$$\underline{\mathcal{P}}(\mathrm{I}_N + tD) = \mathsf{P_c}(X|Y) + tb^{(D)} \quad \text{and} \quad \underline{\mathcal{U}}(\mathrm{I}_N + tD) = 1 + t\beta^{(D)}, \tag{C.25}$$

where

$$b^{(D)} = \sum_{z \in \mathcal{Z}} \max_{x \in \mathcal{M}_z} (PD)(x, z) \quad \text{and} \quad \beta^{(D)} = \sum_{z \in \mathcal{Z}} q(z) D(z, z), \tag{C.26}$$

where $\mathcal{M}_z = \{x \in \mathcal{X} : P(x, z) \geq P(x', z) \text{ for all } x' \in \mathcal{X}\}$. Since $\underline{\mathcal{P}}(F) \leq \mathsf{P_c}(X|Y)$ for all $F \in \underline{\mathcal{F}}$, it is immediate that $b^{(D)} \leq 0$ for every $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$. Moreover, since $\underline{\mathcal{P}}(G_\varepsilon) \leq \varepsilon$, we have that $b^{(D_\varepsilon)} < 0$ for all $\varepsilon \in [\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y))$. By definition of $\underline{\mathcal{D}}(\mathrm{I}_N)$, it is clear that if $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$, then we have $D(y, y) \leq 0$ for all $y \in \mathcal{Y}$, which together with the fact that $\|D\| = 1$ for all $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$, implies that $\beta^{(D)} < 0$ for all $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$. We first establish the following intuitive claim.

Claim. Let $\varepsilon'_{\mathsf{L}} < \mathsf{P_c}(X|Y)$ be as defined in Lemma C.9. Then, there exists an optimal filter G_ε at ε for each $\varepsilon \in [\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$ such that $\underline{\mathcal{P}}(G_\varepsilon) = \varepsilon$ and $\underline{\mathcal{U}}(G_\varepsilon) = \underline{\hbar}(\varepsilon)$.

Proof of Claim. The filter $G_\varepsilon = \mathrm{I}_N + t_\varepsilon D_\varepsilon$ is optimal at ε for every $\varepsilon \in [\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y))$. To reach contradiction, assume that there exists $\varepsilon_0 < \varepsilon$ such that $\underline{\mathcal{P}}(G_\varepsilon) = \varepsilon_0$. According to (C.25), we obtain $\mathsf{P_c}(X|Y) + t_\varepsilon b^{(D_\varepsilon)} = \varepsilon_0 < \varepsilon$ and hence

$$t_\varepsilon > \frac{\mathsf{P_c}(X|Y) - \varepsilon}{-b^{(D_\varepsilon)}} =: t'.$$

Now consider the filter $\mathrm{I}_N + t' D_\varepsilon$. Since $t' \leq \delta$, we have from (C.25) that $\underline{\mathcal{P}}(\mathrm{I}_N + t' D_\varepsilon) = \varepsilon$ and

$$\underline{\hbar}(\varepsilon) \stackrel{(a)}{=} 1 + t_\varepsilon \beta^{(D_\varepsilon)} \stackrel{(b)}{<} \underline{\mathcal{U}}(\mathrm{I}_N + t' D_\varepsilon) = 1 + t' \beta^{(D_\varepsilon)},$$

where (a) is due to the optimality of G_ε and (b) follows from the negativity of $\beta^{(D_\varepsilon)}$. The above inequality contradicts the maximality of $\underline{\hbar}(\varepsilon)$. This implies that $\underline{\mathcal{P}}(G_\varepsilon) = \varepsilon$ which, according to (C.25), yields

$$\underline{\hbar}(\varepsilon) = 1 - (\mathsf{P_c}(X|Y) - \varepsilon)\frac{\beta^{(D_\varepsilon)}}{b^{(D_\varepsilon)}}, \tag{C.27}$$

for all $\varepsilon \in [\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y))$. $\square$

Now fix $\varepsilon' \in [\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$ with $\varepsilon \leq \varepsilon'$. On the one hand, according to (C.27), we know that

$$\underline{\hbar}(\varepsilon') = 1 - (\mathsf{P_c}(X|Y) - \varepsilon')\frac{\beta^{(D_{\varepsilon'})}}{b^{(D_{\varepsilon'})}}. \tag{C.28}$$

On the other hand, we obtain from (C.25) that $0 \leq \frac{\mathsf{P_c}(X|Y)-\varepsilon'}{-b^{(D_\varepsilon)}} \leq t_\varepsilon$ and hence

$$\underline{\mathcal{P}}\left(\mathrm{I}_N + \frac{\mathsf{P_c}(X|Y) - \varepsilon'}{-b^{(D_\varepsilon)}}D_\varepsilon\right) = \varepsilon', \tag{C.29}$$

$$\underline{\mathcal{U}}\left(\mathrm{I}_N + \frac{\mathsf{P_c}(X|Y) - \varepsilon'}{-b^{(D_\varepsilon)}}D_\varepsilon\right) = 1 - (\mathsf{P_c}(X|Y) - \varepsilon')\frac{\beta^{(D_\varepsilon)}}{b^{(D_\varepsilon)}}. \tag{C.30}$$

Comparing (C.28) and (C.30), we conclude that

$$1 - (\mathsf{P_c}(X|Y) - \varepsilon')\frac{\beta^{(D_{\varepsilon'})}}{b^{(D_{\varepsilon'})}} = \underline{\hbar}(\varepsilon') \geq 1 - (\mathsf{P_c}(X|Y) - \varepsilon')\frac{\beta^{(D_\varepsilon)}}{b^{(D_\varepsilon)}},$$

and hence the function $\varepsilon \mapsto \dfrac{\beta^{(D_\varepsilon)}}{b^{(D_\varepsilon)}}$ is non-increasing over $[\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y))$. Therefore, since $\dfrac{\beta^{(D_\varepsilon)}}{b^{(D_\varepsilon)}} > 0$, the limit $\lim\limits_{\varepsilon \to \mathsf{P_c}(X|Y)^-} \dfrac{\beta^{(D_\varepsilon)}}{b^{(D_\varepsilon)}} =: A$ exists.

Let $K = (\mathsf{P_c}(X|Y) - \varepsilon'_{\mathsf{L}})^{-1}$. For each $n > K$, let $F_n = G_{\mathsf{P_c}(X|Y) - \frac{1}{n}}$. Write $F_n = \mathrm{I}_N + t_n D_n$ with $t_n \in [0, \delta]$ and $D_n \in \underline{\mathcal{D}}(\mathrm{I}_N)$. Since $\underline{\mathcal{D}}(\mathrm{I}_N)$ is compact, there exist $\{n_1 < n_2 < \cdots\}$ and $D^* \in \underline{\mathcal{D}}(\mathrm{I}_N)$ such that $D_{n_k} \to D^*$ as $k \to \infty$. By continuity of the mappings $D \mapsto b^{(D)}$ and $D \mapsto \beta^{(D)}$, we have that $b^{(D_{n_k})} \to b^{(D^*)}$ and $\beta^{(D_{n_k})} \to \beta^{(D^*)}$ as $k \to \infty$.

Claim. We have that $b^{(D^*)} < 0$ and, in particular, $A = \dfrac{\beta^{(D^*)}}{b^{(D^*)}}$.

Proof of Claim. Recall that $F \in \underline{\mathcal{F}}^1$ if and only if F is a permutation matrix. In particular, $\underline{\mathcal{F}}^1$ is finite with $|\underline{\mathcal{F}}^1| = N!$. Recall that $b^{(D^*)} \leq 0$. Assume that $b^{(D^*)} = 0$. Since $\dfrac{\beta^{(D_{n_k})}}{b^{(D_{n_k})}} \to A \in [0, \infty)$ and $b^{(D_{n_k})} \to b^{(D^*)} = 0$ as $k \to \infty$, we have that $\beta^{(D_{n_k})} \to 0$ and hence $\beta^{(D^*)} = 0$. This implies that $\underline{\mathcal{U}}(\mathrm{I}_N + tD^*) = 1$ for all $t \in [0, \delta]$, i.e., $\mathrm{I}_N + tD^* \in \underline{\mathcal{F}}^1$ for all $t \in [0, \delta]$. This contradicts the fact that $\underline{\mathcal{F}}^1$ is finite. $\square$

The claim implies that for $\varepsilon \in [\mathsf{P_c}(X|Y) + \delta b^{(D^*)}, \mathsf{P_c}(X|Y)]$,

$$\underline{\mathcal{P}}\left(\mathrm{I}_N + \frac{\mathsf{P_c}(X|Y) - \varepsilon}{-b^{(D^*)}} D^*\right) = \varepsilon,$$
$$\underline{\mathcal{U}}\left(\mathrm{I}_N + \frac{\mathsf{P_c}(X|Y) - \varepsilon}{-b^{(D^*)}} D^*\right) = 1 - (\mathsf{P_c}(X|Y) - \varepsilon)A.$$

Recall that $\dfrac{\beta^{(D^*)}}{b^{(D^*)}} = A \leq \dfrac{\beta^{(D_\varepsilon)}}{b^{(D_\varepsilon)}}$ for all $\varepsilon \in [\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y))$. Let $\varepsilon_{\mathsf{L}} := \max\{\varepsilon'_{\mathsf{L}}, \mathsf{P_c}(X|Y) + \delta b^{(D^*)}\}$. Then

$$\underline{\hbar}(\varepsilon) \geq 1 - (\mathsf{P_c}(X|Y) - \varepsilon)\frac{\beta^{(D^*)}}{b^{(D^*)}} \geq 1 - (\mathsf{P_c}(X|Y) - \varepsilon)\frac{\beta^{(D_\varepsilon)}}{b^{(D_\varepsilon)}} \overset{(a)}{=} \underline{\hbar}(\varepsilon), \tag{C.31}$$

for all $\varepsilon \in [\varepsilon_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$, where the equality in (a) follows from (C.27). This proves that $\underline{\hbar}$ is linear on $\varepsilon \in [\varepsilon_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$.

Recall that $\beta^{(D)} < 0$ for all $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$. The maximality of $\underline{\hbar}$ and (C.31) imply then

$$\underline{\hbar}'(\mathsf{P_c}(X|Y)) = \min_{D \in \underline{\mathcal{D}}(\mathrm{I}_N)} \frac{\beta^{(D)}}{b^{(D)}}. \tag{C.32}$$

If $b^{(D)} = 0$ for some $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$, the term $\dfrac{\beta^{(D)}}{b^{(D)}}$ is defined to be $+\infty$. Notice that this convention agrees with the fact that if $b^{(D)} = 0$ then D cannot be an *optimal direction*. Furthermore, for every $D' \in \underline{\mathcal{D}}(\mathrm{I}_N)$ such that $\underline{\hbar}'(\mathsf{P_c}(X|Y)) = \frac{\beta^{(D')}}{b^{(D')}}$, there exists $\varepsilon_{\mathsf{L}} < \mathsf{P_c}(X|Y)$ (depending on D') such that

$$\mathrm{I}_N + \frac{\mathsf{P_c}(X|Y) - \varepsilon}{-b^{(D')}} D' \tag{C.33}$$

achieves $\underline{\hbar}(\varepsilon)$ for every $\varepsilon \in [\varepsilon_{\mathsf{L}}, \mathsf{P_c}(X|Y)]$. In addition, assume that for each $y \in \mathcal{Y}$ there exists (a unique) $x_y \in \mathcal{X}$ such that, for all $x \neq x_y$,

$$\Pr(X = x_y|Y = y) > \Pr(X = x|Y = y).$$

In particular, $\mathcal{M}_z = \{x_z\}$ for every $z \in \mathcal{Z}$ and hence (C.26) becomes

$$b^{(D)} = \sum_{z\in\mathcal{Z}} (PD)(x_z, z) \quad \text{and} \quad \beta^{(D)} = \sum_{z\in\mathcal{Z}} q_Y(z)D(z, z),$$

for every $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$. Using the fact that $\sum\limits_{z\in\mathcal{Z}} D(y, z) = 0$ for all $y \in \mathcal{Y}$, we obtain

$$b^{(D)} = -\sum_{y\in\mathcal{Y}}\sum_{z\neq y}(P(x_y, y) - P(x_z, y))D(y, z) \quad \text{and} \quad \beta^{(D)} = -\sum_{y\in\mathcal{Y}}\sum_{z\neq y} q_Y(y)D(y, z).$$

Therefore, for every $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$,

$$\frac{\beta^{(D)}}{b^{(D)}} = \frac{\sum_{y\in\mathcal{Y}}\sum_{z\neq y} q_Y(y)D(y, z)}{\sum_{y\in\mathcal{Y}}\sum_{z\neq y}(P(x_y, y) - P(x_z, y))D(y, z)}. \tag{C.34}$$

Since $\dfrac{\sum_k a_k x_k}{\sum_k b_k x_k} \geq \min\limits_k \dfrac{a_k}{b_k}$ for $a_k > 0$ and $b_k, x_k \geq 0$ with $\sum_k x_k > 0$, we obtain from (C.34) that for every $D \in \underline{\mathcal{D}}(\mathrm{I}_N)$

$$\frac{\beta^{(D)}}{b^{(D)}} \geq \min_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} \frac{q_Y(y)}{P(x_y, y) - P(x_z, y)}.$$

Equation (C.32) implies that

$$\underline{\hbar}'(\mathsf{P_c}(X|Y)) \geq \min_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} \frac{q_Y(y)}{P(x_y, y) - P(x_z, y)}.$$

Assume that (y_0, z_0) attains the above minimum. We note that one can easily show from (C.24) that $0 \leq \underline{\hbar}'(\varepsilon) \leq \frac{1-q_Y(1)}{\sigma} < \infty$, where $\sigma := \sum_{z\in\mathcal{Z}} (P(x_z, z) - P(x_1, z)) > 0$. Hence, we have

$y_0 \neq z_0$. Now, consider the direction D_* such that

$$D_*(y,z) = \begin{cases} \lambda, & y = y_0, z = z_0 \\ -\lambda, & y = z = y_0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda = 2^{-1/2}$. Equation (C.34) implies then that

$$\frac{\beta^{(D_*)}}{b^{(D_*)}} = \frac{q_Y(y_0)}{P(x_{y_0}, y_0) - P(x_{z_0}, y_0)},$$

and hence

$$\underline{\hbar}'(\mathsf{P_c}(X|Y)) \leq \frac{q_Y(y_0)}{P(x_{y_0}, y_0) - P(x_{z_0}, y_0)} = \min_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} \frac{q_Y(y)}{P(x_y, y) - P(x_z, y)}.$$

As a consequence,

$$\underline{\hbar}'(\mathsf{P_c}(X|Y)) = \min_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} \frac{q_Y(y)}{P(x_y, y) - P(x_z, y)}.$$

Moreover, (C.33) implies that there exists $\varepsilon_{\mathsf{L}}^{y_0,z_0} < \mathsf{P_c}(X|Y)$ such that $\mathrm{I}_N + \dfrac{\mathsf{P_c}(X|Y) - \varepsilon}{-b^{(D_*)}} D_*$ achieves $\underline{\hbar}(\varepsilon)$ for every $\varepsilon \in [\varepsilon_{\mathsf{L}}^{y_0,z_0}, \mathsf{P_c}(X|Y)]$. Note that

$$\mathrm{I}_N + \frac{\mathsf{P_c}(X|Y) - \varepsilon}{-b^{(D_*)}} D_* = \mathsf{Z}^{y_0,z_0}(\zeta^{y_0,z_0}(\varepsilon)),$$

where $\zeta^{y_0,z_0}(\varepsilon) = \dfrac{\mathsf{P_c}(X|Y) - \varepsilon}{P(x_{y_0}, y_0) - P(x_{z_0}, y_0)}$. $\square$

C.4 Proof of Theorem 6.20

Let $P = [P(x^n, y^n)]_{x^n, y^n \in \{0,1\}^n}$ denotes the joint probability matrix of X^n and Y^n and $q(y^n) = \Pr(Y^n = y^n)$ for $y^n \in \{0,1\}^n$. Let $\mathbf{0} = (0, 0, \ldots, 0)$ and $\mathbf{1} = (1, 1, \ldots, 1)$. We will show that

(X^n, Y^n) satisfies the hypotheses of Theorem 6.17 with $y_0 = \mathbf{1}$ and $z_0 = \mathbf{0}$.

Under the assumptions (a_1) and (b), it is straightforward to verify that

$$P(x^n, y^n) = (\bar{\alpha}\bar{p})^n \prod_{k=1}^{n} \left(\frac{p}{\bar{p}}\right)^{x_k} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_k \oplus y_k}, \tag{C.35}$$

for every $x^n, y^n \in \{0,1\}^n$. By assumption, $\mathsf{P_c}(X^n) = p^n < \bar{\alpha}^n = \mathsf{P_c}(X^n|Y^n)$. It is also straightforward to verify that $q(y^n) > 0$ for all $y \in \{0,1\}^n$. Since $\bar{\alpha}\bar{p} > \alpha p$, we have from (C.35) that

$$\Pr(X^n = z^n, Y^n = z^n) > \Pr(X^n = x^n, Y^n = z^n),$$

for all $x^n \neq z^n$. In the notation of Theorem 6.17, $x^n_{z^n} = z^n$ for all $z^n \in \{0,1\}^n$. Note that

$$\min_{y^n, z^n \in \{0,1\}^n} \frac{q(y^n)}{P(x^n_{y^n}, y^n) - P(x^n_{z^n}, y^n)} = \min_{y^n \in \{0,1\}^n} \frac{q(y^n)}{P(y^n, y^n) - \min\limits_{z^n \neq y^n} P(z^n, y^n)}.$$

It is easy to show that $\min\limits_{z^n \neq y^n} P(z^n, y^n) = (\alpha p)^n \prod\limits_{k=1}^{n} \left(\frac{p}{\bar{p}}\right)^{-y_k}$ and that the minimum is attained by $z^n = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n)$. As a consequence,

$$\begin{aligned}
\min_{y^n, z^n \in \{0,1\}^n} \frac{q(y^n)}{P(x^n_{y^n}, y^n) - P(x^n_{z^n}, y^n)} &= \min_{y^n \in \{0,1\}^n} \frac{\sum\limits_{x^n \in \{0,1\}^n} \prod\limits_{k=1}^{n} \left(\frac{p}{\bar{p}}\right)^{x_k - y_k} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_k \oplus y_k}}{1 - \left(\frac{p\alpha}{\bar{p}\bar{\alpha}}\right)^n \Pi_{y^n}^{-2}} \\
&= \min_{y^n \in \{0,1\}^n} \frac{\prod\limits_{k=1}^{n} \left[(\frac{p}{\bar{p}})^{-y_k} (\frac{\alpha}{\bar{\alpha}})^{y_k} + (\frac{p}{\bar{p}})^{1-y_k} (\frac{\alpha}{\bar{\alpha}})^{1-y_k} \right]}{1 - \left(\frac{p\alpha}{\bar{p}\bar{\alpha}}\right)^n \Pi_{y^n}^{-2}},
\end{aligned}$$

where $\Pi_{y^n} = \prod\limits_{k=1}^{n} \left(\frac{p}{\bar{p}}\right)^{y_k}$. Observe that the denominator is maximized when $y^n = \mathbf{1}$. Using the fact that $p \geq \frac{1}{2} \geq \bar{p}$, one can show that the numerator is minimized when $y^n = \mathbf{1}$. In particular,

$$\min_{y^n, z^n \in \{0,1\}^n} \frac{q(y^n)}{P(x^n_{y^n}, y^n) - P(x^n_{z^n}, y^n)} = \frac{(\alpha\bar{p} + \bar{\alpha}p)^n}{(\bar{\alpha}p)^n - (\alpha\bar{p})^n},$$

and the minimum is attained by $(y_0^n, z_0^n) = (\mathbf{1}, \mathbf{0})$.

Therefore (X^n, Y^n) satisfies the hypotheses of Theorem 6.17 with $(y_0^n, z_0^n) = (\mathbf{1}, \mathbf{0})$. Thus, there exists $\varepsilon'_{\mathsf{L}} < \bar{\alpha}^n$ such that for every $\varepsilon \in [\varepsilon'_{\mathsf{L}}, \bar{\alpha}^n]$

$$\underline{\hbar}(\varepsilon) = 1 - \frac{\bar{\alpha}^n - \varepsilon}{(\bar{\alpha}p)^n - (\alpha\bar{p})^n} q^n.$$

Moreover, $\mathsf{Z}^{\mathbf{1},\mathbf{0}}(\zeta^{y_0,z_0}(\varepsilon))$ achieves $\underline{\hbar}(\varepsilon)$ for every $\varepsilon \in [\varepsilon'_{\mathsf{L}}, \bar{\alpha}^n]$, where

$$\zeta^{y_0,z_0}(\varepsilon) = \frac{\bar{\alpha}^n - \varepsilon}{(\bar{\alpha}p)^n - (\alpha\bar{p})^n}.$$

Recall that $\underline{\hbar}(\varepsilon) = \underline{\hbar}_n^n(\varepsilon^{1/n})$ and let $\varepsilon_{\mathsf{L}} = (\varepsilon'_{\mathsf{L}})^{1/n}$. Therefore, $\underline{\hbar}_n^n(\varepsilon) = 1 - \zeta_n(\varepsilon)q^n$ for all $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$ which is attained by the Z-channel $\mathsf{Z}_n(\zeta_n(\varepsilon))$, where $\zeta_n(\varepsilon) := \zeta^{y_0,z_0}(\varepsilon^n)$.

C.5 Proof of Corollary 6.22

Assume that $p > \frac{1}{2}$. By Theorem 6.20, for every $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$ we have $\underline{\hbar}_n(\varepsilon) = [A_n\varepsilon^n + B_n]^{1/n}$, where $A_n = \dfrac{q^n}{(\bar{\alpha}p)^n - (\alpha\bar{p})^n}$ and $B_n = 1 - \dfrac{\bar{\alpha}^n q^n}{(\bar{\alpha}p)^n - (\alpha\bar{p})^n}$. In particular,

$$\begin{aligned}
\underline{\hbar}'_n(\varepsilon) &= A_n \left(\frac{\varepsilon}{\underline{\hbar}_n(\varepsilon)}\right)^{n-1}, \\
\underline{\hbar}''_n(\varepsilon) &= (n-1)\frac{A_n B_n}{\underline{\hbar}_n^{n+1}(\varepsilon)} \left(\frac{\varepsilon}{\underline{\hbar}_n(\varepsilon)}\right)^{n-2}.
\end{aligned} \tag{C.36}$$

Since $p > \frac{1}{2}$ and $\alpha > 0$, we have $B_n \to 1$ as $n \to \infty$. Let $N_0 \geq 1$ be such that $B_n \geq 0$ for all $n \geq N_0$. In this case, we have that $\underline{\hbar}''_n(\varepsilon) \geq 0$ for all $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$ and $n \geq N_0$. In particular, $\underline{\hbar}_n$ is convex on $[\varepsilon_{\mathsf{L}}, \bar{\alpha}]$. As a consequence, for all $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$ and $n \geq N_0$

$$\underline{\hbar}_n(\varepsilon) \geq 1 - (\bar{\alpha} - \varepsilon)\underline{\hbar}'_n(\bar{\alpha}).$$

Since $\hbar^{\mathsf{i}}_n(\varepsilon) = \underline{\hbar}_1(\varepsilon) = 1 - (\bar{\alpha} - \varepsilon)\underline{\hbar}_1'(\bar{\alpha})$ for all $\varepsilon \in [p, \bar{\alpha}]$, the above inequality implies that

$$\underline{\hbar}_n(\varepsilon) - \hbar^{\mathsf{i}}_n(\varepsilon) \geq (\bar{\alpha} - \varepsilon)(\underline{\hbar}_1'(\bar{\alpha}) - \underline{\hbar}_n'(\bar{\alpha}))$$

for all $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$ and $n \geq N_0$. The result follows from (C.36).

Now, assume that $p = \frac{1}{2}$. In this case, we have for all $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$

$$\underline{\hbar}_n(\varepsilon) = \left(\frac{\varepsilon^n - \alpha^n}{\bar{\alpha}^n - \alpha^n}\right)^{1/n} \quad \text{and} \quad \hbar^{\mathsf{i}}_n(\varepsilon) = \frac{\varepsilon - \alpha}{\bar{\alpha} - \alpha}.$$

Let $\Xi_n : [\frac{1}{2}, \bar{\alpha}] \to \mathbb{R}$ be given by $\Xi_n(\varepsilon) = \underline{\hbar}_n(\varepsilon) - \hbar^{\mathsf{i}}_n(\varepsilon)$.

Claim. The function Ξ_n is decreasing on $[\frac{1}{2}, \bar{\alpha}]$.

Proof of Claim. We shall show that $\Xi_n'(\varepsilon) \leq 0$ for all $\varepsilon \in [\frac{1}{2}, \bar{\alpha}]$. A straightforward computation shows that

$$\Xi_n'(\varepsilon) = \frac{1}{\left[1 - \left(\frac{\alpha}{\varepsilon}\right)^n\right]^{(n-1)/n}} \frac{1}{[\bar{\alpha}^n - \alpha^n]^{1/n}} - \frac{1}{\bar{\alpha} - \alpha}.$$

This function is clearly decreasing, and so it is enough to show that $\Xi_n'(\frac{1}{2}) \leq 0$. Note that $\Xi_n'(\frac{1}{2}) \leq 0$ if and only if

$$\frac{\left(1 - \frac{\alpha}{\bar{\alpha}}\right)^n}{1 - \left(\frac{\alpha}{\bar{\alpha}}\right)^n} \leq [1 - (2\alpha)^n]^{n-1}. \tag{C.37}$$

Observe that $\dfrac{\left(1 - \frac{\alpha}{\bar{\alpha}}\right)^n}{1 - \left(\frac{\alpha}{\bar{\alpha}}\right)^n} \leq \left(1 - \dfrac{\alpha}{\bar{\alpha}}\right)^{n-1}$. Using the fact that $4\alpha\bar{\alpha} \leq 1$, it is straightforward to verify that (C.37) holds. □

Since Ξ_n is decreasing over $[\frac{1}{2}, \bar{\alpha}]$, we obtain for all $\varepsilon \in [\varepsilon_{\mathsf{L}}, \bar{\alpha}]$

$$0 \leq \underline{\hbar}_n(\varepsilon) - \hbar^{\mathsf{i}}_n(\varepsilon) \leq \Xi_n\left(\frac{1}{2}\right) = \frac{1}{2}\left[\left(\frac{1 - (2\alpha)^n}{\bar{\alpha}^n - \alpha^n}\right)^{1/n} - 1\right].$$

Since $1 - (2\alpha)^n \leq 1 - \left(\frac{\alpha}{\bar{\alpha}}\right)^n$, it is straightforward to show that $\Xi_n\left(\frac{1}{2}\right) \leq \frac{\alpha}{2\bar{\alpha}}$, which completes the proof.

C.6 Proof of Theorem 6.23

As before, let $P = [P(x^n, y^n)]_{x^n, y^n \in \{0,1\}^n}$ denote the joint probability matrix of X^n and Y^n and let $q(y^n) = \Pr(Y^n = y^n)$ for $y^n \in \{0,1\}^n$. We first show that (X^n, Y^n) satisfies the hypotheses of Theorem 6.17, and thus we can use (6.19) to obtain bounds on $\underline{\hbar}'(\mathsf{P_c}(X^n|Y^n))$.

Assumptions (a$_2$) and (b) imply that, for all $x^n, y^n \in \{0,1\}^n$

$$P(x^n, y^n) = (\bar{\alpha}\bar{r})^n \frac{\bar{p}}{\bar{r}} \left(\frac{p}{\bar{p}}\right)^{x_1} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_1 \oplus y_1} \prod_{k=2}^{n} \left(\frac{r}{\bar{r}}\right)^{x_k \oplus x_{k-1}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_k \oplus y_k}, \tag{C.38}$$

where the product equals one if $n = 1$. Since $\alpha > 0$, it is clear that $q(y^n) > 0$ for all $y^n \in \{0,1\}^n$. Let $N_0(z^n) = |\{1 \leq k \leq n : z_k = 0\}|$ and $N_1(z^n) = |\{1 \leq k \leq n : z_k = 1\}|$ for any binary vector $z^n \in \{0,1\}^n$. Recall that n is odd, so either $N_0(z^n) < N_1(z^n)$ or $N_0(z^n) > N_1(z^n)$. The following lemma shows that for every $y^n \in \{0,1\}^n$ there exists (a unique) $x^n_{y^n} \in \{0,1\}^n$ such that $P(x^n_{y^n}, y^n) > P(x^n, y^n)$ for all $x^n \neq x^n_{y^n}$.

Lemma C.10. *Let (X^n, Y^n) be as in the hypothesis of Theorem 6.23. Then, we have for any $y^n \in \{0,1\}^n$*

$$P(x^n, y^n) \leq \begin{cases} P(\mathbf{0}, y^n) = (\bar{\alpha}\bar{r})^n \frac{\bar{p}}{\bar{r}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}, & \textit{if } N_0(y^n) > N_1(y^n), \\ P(\mathbf{1}, y^n) = (\bar{\alpha}\bar{r})^n \frac{p}{\bar{r}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)}, & \textit{if } N_0(y^n) < N_1(y^n), \end{cases}$$

for all $x^n \in \{0,1\}^n$ with equality if and only if $x^n = \mathbf{0}$ or $x^n = \mathbf{1}$, respectively.

To prove this lemma, we will make use of the following fact.

Claim. Let $y^n \in \{0,1\}^n$ be given. If $x^n \in \{0,1\}^n$ maximizes $P(x^n, y^n)$, then $x_1 = x_2 = \cdots = x_n$.

Proof of Claim. We prove the result using backward induction. To do so, we assume that the maximizer x^n satisfies $x_n = x_{n-1} = \cdots = x_l$ for $2 \leq l \leq n$. It is sufficient to show that

$x_n = \cdots = x_l = x_{l-1}$. In light of (C.38), we have

$$P(x^n, y^n) = A_{l-1} \left(\frac{r}{\bar{r}}\right)^{x_l \oplus x_{l-1}} \prod_{k=l}^{n} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_l \oplus y_k}, \tag{C.39}$$

where[1]

$$A_{l-1} := (\bar{\alpha}\bar{r})^n \frac{\bar{p}}{\bar{r}} \left(\frac{p}{\bar{p}}\right)^{x_1} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_1 \oplus y_1} \prod_{k=2}^{l-1} \left(\frac{r}{\bar{r}}\right)^{x_k \oplus x_{k-1}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_k \oplus y_k}.$$

Notice that A_{l-1} depends only on $x_1, \ldots, x_{l-1}$. By the induction hypothesis, we have $x_l = \cdots = x_n$. In particular, x^n equals either

$$\tilde{x}^n := \{x_1, \ldots, x_{l-1}, \underbrace{\bar{x}_{l-1}, \ldots, \bar{x}_{l-1}}_{n-l+1}\} \quad \text{or} \quad \hat{x}^n := \{x_1, \ldots, x_{l-1}, \underbrace{x_{l-1}, \ldots, x_{l-1}}_{n-l+1}\}.$$

By (C.39), we have that

$$P(\tilde{x}^n, y^n) = A_{l-1} \frac{r}{\bar{r}} \prod_{k=l}^{n} \left(\frac{\alpha}{\bar{\alpha}}\right)^{1 - x_{l-1} \oplus y_k} \quad \text{and} \quad P(\hat{x}^n, y^n) = A_{l-1} \prod_{k=l}^{n} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_{l-1} \oplus y_k}.$$

By the assumptions on r and α, we have

$$\frac{r}{\bar{r}} \prod_{k=l}^{n} \left(\frac{\alpha}{\bar{\alpha}}\right)^{1 - x_{l-1} \oplus y_k} \leq \frac{r}{\bar{r}} < \left(\frac{\alpha}{\bar{\alpha}}\right)^{n-1} \leq \left(\frac{\alpha}{\bar{\alpha}}\right)^{n-l+1} \leq \prod_{k=l}^{n} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_{l-1} \oplus y_k},$$

which shows that $P(\tilde{x}^n, y^n) < P(\hat{x}^n, y^n)$ and hence $x^n = \hat{x}^n$. In other words, $x_{l-1} = x_l = \cdots = x_n$. This completes the induction step. $\square$

Proof of Lemma C.10. By the above claim, for any given $y^n \in \{0,1\}^n$, the maximizer $x^n \in \{0,1\}^n$ of $P(x^n, y^n)$ is either $x^n = \mathbf{0}$ or $x^n = \mathbf{1}$, for which we have

$$P(\mathbf{0}, y^n) = (\bar{\alpha}\bar{r})^n \frac{\bar{p}}{\bar{r}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}, \tag{C.40}$$

[1] When $l \leq 3$, we use the convention that $\prod_{k=2}^{l-1} \left(\frac{r}{\bar{r}}\right)^{x_k \oplus x_{k-1}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_k \oplus y_k} = 1$.

$$P(\mathbf{1}, y^n) = (\bar{\alpha}\bar{r})^n \frac{p}{\bar{r}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)}. \tag{C.41}$$

Assume $N_0(y^n) > N_1(y^n)$ and recall that $\alpha p < \bar{\alpha}\bar{p}$. In this case,

$$p\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} \leq \frac{\alpha p}{\bar{\alpha}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)} < \bar{p}\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)},$$

which implies $P(\mathbf{0}, y) > P(\mathbf{1}, y)$, and hence $x^n = \mathbf{0}$ is the only maximizer. If $N_0(y^n) < N_1(y^n)$, then $\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} > \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}$. Since $p \geq \bar{p}$, we conclude that

$$p\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} > \bar{p}\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}.$$

Consequently, $P(\mathbf{1}, y) > P(\mathbf{0}, y)$ and hence $x^n = \mathbf{1}$ is the only maximizer. $\square$

Note that

$$\begin{aligned}
\mathsf{P_c}(X^n|Y^n) &= \sum_{y^n \in \{0,1\}^n} \max_{x^n \in \{0,1\}^n} P(x^n, y^n) \\
&\overset{(a)}{=} \sum_{y^n : N_0(y^n) > N_1(y^n)} P(\mathbf{0}, y^n) + \sum_{y^n : N_0(y^n) < N_1(y^n)} P(\mathbf{1}, y^n) \\
&\overset{(b)}{=} \bar{\alpha}^n \bar{r}^{n-1} \sum_{k=0}^{(n-1)/2} \binom{n}{k} \left(\frac{\alpha}{\bar{\alpha}}\right)^k,
\end{aligned} \tag{C.42}$$

where (a) is due to Lemma C.10 and (b) comes from (C.40) and (C.41).

Now that all the hypotheses of Theorem 6.17 are shown to be satisfied, we can use (6.19) to study $\underline{\hbar}'(\mathsf{P_c}(X^n|Y^n))$. The following lemma is important in bounding $\underline{\hbar}'(\mathsf{P_c}(X^n|Y^n))$.

Lemma C.11. *Let (X^n, Y^n) be as in the hypothesis of Theorem 6.23. Then, for all $y^n \in \{0,1\}^n$,*

$$q(y^n) \geq \alpha^n.$$

Proof. From (C.38), we have

$$\begin{aligned} P(x^n, y^n) &= (\bar{\alpha}\bar{r})^n \frac{\bar{p}}{\bar{r}} \left(\frac{p}{\bar{p}}\right)^{x_1} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_1 \oplus y_1} \prod_{k=2}^{n} \left(\frac{r}{\bar{r}}\right)^{x_k \oplus x_{k-1}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{x_k \oplus y_k} \\ &\geq \left(\frac{\alpha}{\bar{\alpha}}\right)^n (\bar{\alpha}\bar{r})^n \frac{\bar{p}}{\bar{r}} \left(\frac{p}{\bar{p}}\right)^{x_1} \prod_{k=2}^{n} \left(\frac{r}{\bar{r}}\right)^{x_k \oplus x_{k-1}} \\ &= \alpha^n \bar{r}^n \frac{\bar{p}}{\bar{r}} \left(\frac{p}{\bar{p}}\right)^{x_1} \prod_{k=2}^{n} \left(\frac{r}{\bar{r}}\right)^{x_k \oplus x_{k-1}}. \end{aligned}$$

Summing over all $x^n \in \{0,1\}^n$, we obtain

$$q(y^n) \geq \alpha^n \bar{r}^{n-1} \bar{p} \sum_{x^n \in \{0,1\}^n} \left(\frac{p}{\bar{p}}\right)^{x_1} \prod_{k=2}^{n} \left(\frac{r}{\bar{r}}\right)^{x_k \oplus x_{k-1}}. \tag{C.43}$$

On the other hand, it is straightforward to verify that

$$1 = \sum_{x \in \{0,1\}^n} \Pr(X^n = x^n) = \bar{r}^{n-1} \bar{p} \sum_{x^n \in \{0,1\}^n} \left(\frac{p}{\bar{p}}\right)^{x_1} \prod_{k=2}^{n} \left(\frac{r}{\bar{r}}\right)^{x_k \oplus x_{k-1}}. \tag{C.44}$$

Plugging (C.44) into (C.43), the result follows. □

By (6.19) and the previous lemma,

$$\underline{\hbar}'(\mathsf{P_c}(X^n|Y^n)) \geq \min_{y^n \in \{0,1\}^n} \min_{\substack{z^n \in \{0,1\}^n \\ z^n \neq y^n}} \frac{\alpha^n}{P(x^n_{y^n}, y^n) - P(x^n_{z^n}, y^n)}.$$

Since both $x^n_{y^n}$ and $x^n_{z^n}$ are either $\mathbf{0}$ or $\mathbf{1}$, we have to maximize

$$\vartheta := \begin{cases} (\bar{\alpha}\bar{r})^n \frac{\bar{p}}{\bar{r}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)} - (\bar{\alpha}\bar{r})^n \frac{p}{\bar{r}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)}, & \text{if } N_0(y^n) > N_1(y^n), \\ (\bar{\alpha}\bar{r})^n \frac{p}{\bar{r}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} - (\bar{\alpha}\bar{r})^n \frac{\bar{p}}{\bar{r}} \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}, & \text{if } N_0(y^n) < N_1(y^n). \end{cases}$$

Clearly, ϑ is maximized when $y^n = \mathbf{1}$ and thus

$$\underline{\hbar}'(\mathsf{P_c}(X^n|Y^n)) \geq \frac{\bar{r}\alpha^n}{p(\bar{\alpha}\bar{r})^n - \bar{p}(\alpha\bar{r})^n}.$$

By (6.18) and the fact that $\underline{\hbar}_n^n(\varepsilon) = \underline{\hbar}(\varepsilon^n)$,

$$\underline{\hbar}_n^n(\varepsilon) \leq 1 - \bar{r}\frac{\mathsf{P_c}(X^n|Y^n) - \varepsilon^n}{p(\bar{\alpha}\bar{r})^n - \bar{p}(\alpha\bar{r})^n}\alpha^n,$$

where $\mathsf{P_c}(X^n|Y^n)$ is computed in (C.42).

The lower bound follows from considering the direction $\tilde{D} \in \underline{\mathcal{D}}(\mathrm{I}_{2^n})$, whose entries are all zero except $\tilde{D}(\mathbf{1}, \mathbf{0}) = \lambda$ and $\tilde{D}(\mathbf{1}, \mathbf{1}) = -\lambda$ for $\lambda = 2^{-1/2}$. In particular, plugging $\tilde{D}$ into (C.34), we obtain an upper bound for $\underline{\hbar}'(\mathsf{P_c}(X^n|Y^n))$ and thus a lower bound for $\underline{\hbar}(\varepsilon)$ for the desired range of ε. Note that the filter $\mathrm{I}_{2^n}+\zeta_n(\varepsilon)\tilde{D}$ corresponds to the 2^n-ary Z-channel $\mathsf{Z}_n(\zeta_n(\varepsilon))$.

C.7 Proof of Proposition 6.24

Since $r = 0$, the joint distribution $P_{\theta Y^n}$ can be equivalently written as the joint probability matrix $P = [P(x^n, y^n)]_{x^n, y^n \in \{0,1\}^n}$ with $x_1 = x_2 = \cdots = x_n = \theta$. As in the proof of Theorem 6.23, the hypotheses of Theorem 6.17 are fulfilled. In particular,

$$\underline{\hbar}'(\mathsf{P_c}(\theta|Y^n)) = \min_{y^n \in \{0,1\}^n} \min_{\substack{z^n \in \{0,1\}^n \\ z^n \neq y^n}} \frac{q(y^n)}{P(x^n_{y^n}, y^n) - P(x^n_{z^n}, y^n)}. \tag{C.45}$$

In this case, (C.38) becomes

$$P(\mathbf{0}, y^n) = \bar{p}\bar{\alpha}^n \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)} \quad \text{and} \quad P(\mathbf{1}, y^n) = p\bar{\alpha}^n \left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)}.$$

In particular,

$$\underline{\hbar}'(\mathsf{P_c}(\theta|Y^n)) = \min_{y^n\in\{0,1\}^n} \min_{\substack{z^n\in\{0,1\}^n\\ z^n\neq y^n}} \frac{p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} + \bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}}{P(x^n_{y^n},y^n)-P(x^n_{z^n},y^n)}.$$

Lemma C.10 implies that both $x^n_{y^n}$ and $x^n_{z^n}$ are either $\mathbf{0}$ or $\mathbf{1}$. If $N_0(y^n) > N_1(y^n)$, then

$$\frac{p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} + \bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}}{P(x^n_{y^n},y^n)-P(x^n_{z^n},y^n)} \geq \frac{p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} + \bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}}{\bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)} - p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)}},$$

with equality if and only if $N_1(z^n) > N_0(z^n)$. It is not hard to show that

$$\frac{p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} + \bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}}{\bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)} - p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)}} \geq \frac{\bar{p}+p\left(\frac{\alpha}{\bar{\alpha}}\right)^n}{\bar{p}-p\left(\frac{\alpha}{\bar{\alpha}}\right)^n}, \tag{C.46}$$

with equality if and only if $y^n = \mathbf{0}$. Similarly, if $N_1(y^n) > N_0(y^n)$, then

$$\frac{p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} + \bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}}{P(x^n_{y^n},y^n)-P(x^n_{z^n},y^n)} \geq \frac{p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} + \bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}}{p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} - \bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}},$$

with equality if and only if $N_0(z^n) > N_1(z^n)$. As before,

$$\frac{p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)} + \bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)}}{\bar{p}\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_1(y^n)} - p\bar{\alpha}^n\left(\frac{\alpha}{\bar{\alpha}}\right)^{N_0(y^n)}} \geq \frac{p+\bar{p}\left(\frac{\alpha}{\bar{\alpha}}\right)^n}{p-\bar{p}\left(\frac{\alpha}{\bar{\alpha}}\right)^n}, \tag{C.47}$$

with equality if and only if $y^n = \mathbf{1}$. From (C.46) and (C.47), we conclude that

$$\underline{\hbar}'(\mathsf{P_c}(\theta|Y^n)) = \frac{p+\bar{p}\left(\frac{\alpha}{\bar{\alpha}}\right)^n}{p-\bar{p}\left(\frac{\alpha}{\bar{\alpha}}\right)^n} = \frac{p\bar{\alpha}^n+\bar{p}\alpha^n}{p\bar{\alpha}^n-\bar{p}\alpha^n},$$

and $y_0 = \mathbf{1}$ and $z_0 = \mathbf{0}$ achieve the minimum in (C.45). From the last part of Theorem 6.17 the optimality of the 2^n-ary Z-channel $\mathsf{Z}_n(\zeta_n(\varepsilon))$ is evident.