INFORMATION THEORY FROM A
FUNCTIONAL VIEWPOINT

JINGBO LIU

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
ELECTRICAL ENGINEERING
ADVISER: PAUL CUFF AND SERGIO VERDÚ

JANUARY 2018
Abstract

A perennial theme of information theory is to find new methods to determine the fundamental limits of various communication systems, which potentially helps the engineers to find better designs by eliminating the deficient ones. Traditional methods have focused on the notion of “sets”: the method of types concerns the cardinality of subsets of the typical sets; the blowing-up lemma bounds the probability of the neighborhood of decoding sets; the single-shot (information-spectrum) approach uses the likelihood threshold to define sets. This thesis promotes the idea of deriving the fundamental limits using functional inequalities, where the central notion is “functions” instead of “sets”. A functional inequality follows from the entropic definition of an information measure by convex duality. For example, the Gibbs variational formula follows from the Legendre transform of the relative entropy.

As a first example, we propose a new methodology of deriving converse (i.e. impossibility) bounds based on convex duality and the reverse hypercontractivity of Markov semigroups. This methodology is broadly applicable to network information theory, and in particular resolves the optimal scaling of the second-order rate for the previously open “side-information problems”. As a second example, we use the functional inequality for the so-called $E_{\gamma}$ metric to prove non-asymptotic achievability (i.e. existence) bounds for several problems including source coding, wiretap channels and mutual covering.

Along the way, we derive general convex duality results leading to a unified treatment to many inequalities and information measures such as the Brascamp-Lieb inequality and its reverse, strong data processing inequality, hypercontractivity and its reverse, transportation-cost inequalities, and Rényi divergences. Capitalizing on such dualities, we demonstrate information-theoretic approaches to certain properties of functional inequalities, such as the Gaussian optimality. This is the antithesis of the main thesis (functional approaches to information theory).
Acknowledgements

First, I thank my advisors Professor Paul Cuff and Professor Sergio Verdú for their unceasing guidance and support over the past five years. I was new to information theory at the start of my PhD, and Professor Cuff was generous with his time in helping getting on the road of this field. The puzzle meetings with Professor Cuff and other students was a great source of diversion and inspiration for our graduate life. Professor Verdú demonstrated to us the beauty of research and the charisma of teaching. He encouraged us to delve deeply into serious research topics, revise papers carefully, and work hard even on holidays, by doing so himself.

Thanks, Prof. Abbe, Braverman, Cuff and Verdú, for serving as the committee; Prof. Chen, Poor, and Verdú, for being the readers; the administrative staff from ELE, for their assistance over the years. Thanks to all faculty members of Princeton who have taught me classes; what I learnt from them is reflected in every aspect of this thesis. I am indebted to Princeton University for providing a wonderful environment for studying and research.

I owe sincere gratitude to my mentors and collaborators: Sanket Satpathy, Sudeep Kamath, Wei Yang for numerous illuminating discussions; Professor Emmanuel Abbe for letting me catch a glimpse of working on coding theory; Professor Thomas Courtade for showing me how to prove Gaussian optimality; Professor Ramon van Handel for advising me to look into semigroup operators, as opposed to the maximal operators which I was previously stuck with.

I am thankful to all my friends at Princeton, my colleagues from the information sciences and systems group, especially my roommates, Sen Tao, Lanqing Yu and Amir Reza Asadi, for helping me out with many things in life.

Last but not the least, I thank my parents for everything, and in particular for providing me a place where a large part of the thesis was written.
To my parents.
## Contents

Abstract ................................................................. iii
Acknowledgements ...................................................... iv
List of Tables .......................................................... xi
List of Figures .......................................................... xii

1 Introduction 1

1.1 Information Theory ........................................ 1

1.1.1 Information Measures and Operational Problems ........ 1

1.1.2 Channel Coding Revisited ................................ 3

1.1.3 More on Conveses in Information Theory ................. 7

1.1.4 The Side Information Problems ........................... 10

1.1.5 Common Randomness Generation ......................... 11

1.2 Concentration of Measure .................................... 13

1.2.1 The Gromov-Milman Formulation .......................... 13

1.2.2 Is It Natural for Information Theorists? .................. 15

1.3 Functional Analysis ........................................... 16

1.3.1 In Large Deviation Theory ................................. 16

1.3.2 Hypercontractivity and Its Reverse ....................... 17

1.3.3 Brascamp-Lieb Inequality .................................. 19

1.4 Organization of the Thesis ................................. 21
2 A Tale of Two Optimizations

2.1 The Duality between Functions and Measures .................................. 24

2.2 Dual Formulations of the Forward Brascamp-Lieb Inequality .......... 26

2.3 Application to Common Randomness Generation .............................. 33

2.3.1 The One Communicator Problem: the Setup ............................... 33

2.3.2 A Zero-Rate One-Shot Converse for the Omniscient Helper Problem ................................................................. 35

2.4 Dual Formulation of a Forward-Reverse Brascamp-Lieb Inequality.... 42

2.4.1 Some Mathematical Preliminaries ............................................. 43

2.4.2 Compact $\mathcal{X}$ ................................................................... 49

2.4.3 Noncompact $\mathcal{X}$ ............................................................... 56

2.4.4 Extension to General Convex Functionals ................................. 59

2.5 Some Special Cases of the Forward-Reverse Brascamp-Lieb Inequality 61

2.5.1 Variational Formula of Rényi Divergence .................................. 62

2.5.2 Strong Data Processing Constant ............................................. 64

2.5.3 Loomis-Whitney Inequality and Shearer’s Lemma ................. 66

2.5.4 Hypercontractivity .................................................................. 68

2.5.5 Reverse Hypercontractivity (Positive Parameters) ............... 69

2.5.6 Reverse Hypercontractivity (One Negative Parameter) ........ 70

2.5.7 Transportation-Cost Inequalities ............................................. 71

3 Smoothing ....................................................................................... 74

3.1 The BL Divergence and Its Smooth Version .................................. 74

3.2 Upper Bounds on $d_\delta$ .............................................................. 80

3.2.1 Discrete Case ........................................................................ 80

3.2.2 Gaussian Case ................................................................. 86

3.2.3 Continuous Density Case ................................................... 90

3.3 Application to Common Randomness Generation ....................... 91
3.3.1 Second-order Converse for the Omniscient Helper Problem

4 The Pump

4.1 Prelude: Binary Hypothesis Testing

4.1.1 Review of the Blowing-Up Method, and Its Second-order Sub-optimality

4.1.2 The Fundamental Sub-optimality of the Set-Inequality Approaches

4.1.3 The New Pumping-up Approach

4.2 Optimal Second-order Fano’s Inequality

4.2.1 Bounded Probability Density Case

4.2.2 Gaussian Case

4.3 Optimal Second-order Change-of-Measure

4.4 Applications

4.4.1 Output Distribution of Good Channel Codes

4.4.2 Broadcast Channels

4.4.3 Source Coding with Compressed Side Information

4.4.4 One-communicator CR Generation

4.5 About Extensions to Processes with Weak Correlation

5 More Dualities

5.1 Air on $g^*$

5.1.1 Review of the Image-size Problem

5.1.2 $g^*$ and the Optimal Second-order Image Size Lemmas

5.1.3 Relationship between $g$ and $g^*$

5.2 Variations on $E_\gamma$

5.2.1 $E_\gamma$: Definition and Dual Formulation
List of Tables

1.1 Comparison of previous converse proof techniques . . . . . . . . . . . 10

4.1 Comparison of the basic ideas in BUL and the new methodology . . 107
List of Figures

1.1 CR generation with one-communicator ........................................ 12

2.1 Diagrams for Theorem 2.4.5 ....................................................... 51

2.2 The “partial relations” between the inequalities discussed in this chapter with respect to implication ........................................ 61

2.3 Diagram for hypercontractivity (HC) ............................................ 68

2.4 Diagram for reverse HC ............................................................ 69

2.5 Diagram for reverse HC with one negative parameter ....................... 70

4.1 Schematic comparison of $1_A$, $1_{A_{\ell t}}$, and $T_{\ell t} \circ n 1_A$, where $A$ is the indicator function of a Hamming ball ...................................................... 106

4.2 Illustration of the action of $T_{x,a,t}$. The original function (an indicator function) is convolved with a Gaussian measure and then dilated (with center $x^n$) ...................................................... 111

4.3 A set $A \subseteq Y$ and its “preimage” under $Q_{Y|X}$ in $X'$ .................... 114

4.4 Source coding with compressed side information .............................. 126

5.1 The tradeoff between the sizes of the two images ............................. 139

5.2 $E_\gamma(P \| Q)$ as a function of $\gamma$ where $P = Ber(0.1)$ and $Q = Ber(0.5)$ 157

5.3 Setup for channel resolvability ................................................... 173

5.4 The wiretap channel ................................................................ 207

6.1 Key generation with an omniscient helper .................................... 267
Chapter 1

Introduction

1.1 Information Theory

1.1.1 Information Measures and Operational Problems

The scope of a research field is not only defined by its very name, but also shaped by the people who left their mark on it. Shannon’s ground-breaking 1948 paper [1] characterized the capacity of transmitting messages through a channel in terms of the maximum mutual information. Since then, information theory has been closely associated with communication engineering, even though the ideas and the analytical tools from information theory have found applications in many other research fields, such as physics, statistics, computer science, networking, control theory and financial engineering [2][3][4].

Roughly speaking, the research by Shannon and his successors may be split into the following two categories:

- Information measures. Just as physicists are intrigued by the notion of a space, information theorists are perpetually obsessed with finding a good notion of an information measure, which is a quantity defined for a probability distribution that characterizes, say, the complexity of a random object, or the correlation
between random objects. Examples of information measures include the relative entropy, mutual information, common information \cite{5} \cite{6}, the strong data processing coefficient \cite{7} \cite{8} \cite{9}, and the maximal correlation coefficient \cite{10} \cite{11} \cite{12}. A good definition of an information measure entails certain mathematical conveniences while being the answers to interesting problems.

- **Operational problems.** Once a practical system, say from communication engineering, is formulated as an abstract model, it then becomes a well-posed mathematical question to determine whether an operational task (e.g. data transmission) can be performed for a given parameter range of the model. The critical value of the parameter is called the **fundamental limit**, which the information theorists seek to express or bound in terms of the information measures. Examples of operational problems in communication engineering include channel coding \cite{1}, source coding \cite{13}, randomness generation \cite{14} \cite{15} \cite{16}, their generalizations to network settings \cite{17} \cite{18}, or their variants \cite{19} \cite{20} \cite{21} \cite{22} \cite{23} \cite{24}.

Research in the above two categories are not independent. Firstly, knowing the properties of information measures certainly helps computing the fundamental limits of operational problems. Secondly, the study of operational problems often serves as a source of motivations and ideas for proposing new information measures, or the proofs of information theoretic inequalities.

This thesis investigates the functional representations\(^1\) (sometimes called the variational formulae) of various information measures, ubiquitous in information theory, and their implications on the fundamental limits of operational problems. The most basic example of such a functional representation is the **Gibbs variational formula**

\(^1\)In the mathematical literature a “functional” usually refers to a mapping from a vector space to the reals. In the scope of this thesis, this vector space is the set of all test functions (e.g. bounded continuous real-valued functions) on an alphabet.
(also known as the Donsker-Varadhan formula) of the relative entropy,

\[ D(P\|Q) = \sup_f \left\{ \int f \, dP - \log \int \exp(f) \, dQ \right\} \quad (1.1) \]

where \( P \) and \( Q \) are probability distributions and \( f \) is a real-valued measurable function, all on the same measurable space. As simple and well-known as (1.1) is, it may be surprising that a new, yet powerful and canonical methodology of attacking a wide range of operational problems in communication engineering lies beneath these functional representations. A main objective of this thesis is to introduce this new methodology, and apply it to certain problems regarded as challenging or open by some researchers in the community. As another example, the \( E_\gamma \) metric to be introduced in Section 5.2 has the following functional representation:

\[ E_\gamma(\pi\|Q) = \sup_{f \colon 0 \leq f \leq 1} \left\{ \int f \, d\pi - \gamma \int f \, dQ \right\}. \quad (1.2) \]

While there are other ways of deriving (1.2), a very illuminating viewpoint, which also matches the philosophy of this thesis, is to view \( E_\gamma(\cdot\|Q) \) as the Legendre-Fenchel dual of the convex functional

\[ f \mapsto \begin{cases} \gamma \int f \, dQ, & 0 \leq f \leq 1 \text{ a.e.}; \\ +\infty, & \text{otherwise.} \end{cases} \quad (1.3) \]

### 1.1.2 Channel Coding Revisited

To get a glimpse of the above ideas through a concrete example, let us recall the data transmission (or channel coding) problem considered by Shannon [1], which consists of the following (in the single-shot formulation):

- The input and the output alphabets \( \mathcal{X} \) and \( \mathcal{Y} \), which represent the sets of the possible values of the input and output signals for a given communication chan-
nel. The communication channel is modeled as a random transformation (or transition probability) $P_{Y|X}$.

- A set of messages $\mathcal{M}$ to be transmitted by the channel. The semantic meaning of the messages is irrelevant for the solution of the operational problem, so it is without loss of generality to consider $\mathcal{M} = \{1, \ldots, M\}$ for some integer $M$.

- The encoder is a mapping from $\mathcal{M}$ to $\mathcal{X}$, which can be represented by a codebook consisting of codewords $c_1, \ldots, c_M \in \mathcal{X}$.

- The decoder is a rule of reconstructing a message $\hat{m} \in \mathcal{M}$ based on the observation $y \in \mathcal{Y}$. A deterministic decoder can be specified by disjoint sets $\mathcal{D}_1, \ldots, \mathcal{D}_M \subseteq \mathcal{Y}$. A stochastic decoder can be represented by nonnegative functions $f_1, \ldots, f_M$ where $f_m: \mathcal{Y} \to [0, 1]$ is the probability of declaring message $m$ upon observing $y$. Hence $\sum_{m=1}^{M} f_m = 1$ (i.e. $(f_m)_{m=1}^{M}$ is a partition of the unity).

The goal is to find the tradeoff between $M$ and the maximum error probability

$$\epsilon := 1 - \min_{1 \leq m \leq M} P_{Y|X=c_m}[\mathcal{D}_m].$$

(1.4)

Of course, in general such a problem cannot have a simple yet explicit answer. Shannon considered a special case called the discrete stationary memoryless channel, where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets, and the channel can be used repeatedly for $n$ times; in other

\footnote{To streamline the presentation we consider the maximum error probability here rather than the more popular average error probability. By a simple expurgation argument (see e.g. \cite{Shannon}) it is known that the choice of the two error probability criteria is irrelevant for the fundamental limit in an asymptotic sense.}
words, one considers

\[ X \leftarrow X^n, \quad (1.5) \]

\[ Y \leftarrow Y^n, \quad (1.6) \]

\[ P_{Y|X} \leftarrow P_{Y|X}^{\otimes n}, \quad (1.7) \]

Then it is known \[1\] that \( \epsilon \in (0,1) \), one can send at most

\[ M = \exp(n(C + o_\epsilon(1))) \quad (1.8) \]

messages as \( n \to \infty \), where

\[ C := \max_{P_X} I(X; Y) \quad (1.9) \]

is the maximum mutual information over all input distributions. To establish such a result, one needs to prove lower and upper bounds on \( M \), which are called the \textit{achievability} and the \textit{converse}, respectively. For the converse part, the classical \textit{Fano's inequality} claims that

\[ nI(X^n; Y^n) \geq (1 - \epsilon) \log M - h(\epsilon). \quad (1.10) \]

where \( X^n \) equiprobably takes the values \( c_1, \ldots, c_M \), and \( Y^n \) is the corresponding output. Information theorists call \((1.10)\) a \textit{weak converse} because it only shows that \( \lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log M \geq C \) for the optimal codebook. In contrast, the classical \textit{blowing-up} lemma argument \[27\] (see Section \[4.1.1\] for an introduction) strengthens

\(^3\)Conventionally, \( X^n := \times \cdots \times X \) denotes the Cartesian product. The product channel (tensor product) \( P_{Y^X}^{\otimes n} \) is defined by \( P_{Y|X}^{\otimes n}(y^n|x^n) := \prod_{i=1}^n P_{Y|X}(y_i|x_i) \) for any \( x^n \in X^n \) and \( y^n \in Y^n \).

\(^4\)The notation \( \phi(\epsilon, n) = o_\epsilon(f(n)) \) means that \( \lim_{n \to \infty} \frac{\phi(\epsilon, n)}{f(n)} = 0 \) for any \( \epsilon > 0 \).
Note that (1.11) implies

\[ M \geq \exp(n(C + o_{1}(1))) \]  

(1.12)

which is called a strong converse. However, the \( o_{1}(1) \) term in (1.12) has not yet captured the correct order of the second-order term. In this thesis we propose a new converse method, further strengthening (1.12) to

\[ nI(X^n;Y^n) \geq \log M - O\left(\sqrt{n \log \frac{1}{1-\epsilon}}\right) \]  

(1.13)

which, as we will see, is sharp in both \( \epsilon \uparrow 1 \) and \( n \rightarrow \infty \)! Although there are other methods for obtaining sharp converse bounds for channel coding, to our knowledge there was no previous method for proving the sharp Fano’s inequality in (1.13), which is a versatile tool immediately applicable to a number of other operational problems in network information theory.

The fundamental limits for most operational problems in network information theory can be expressed in terms of the linear combinations of information measures, which admit functional representations (analogous to the Gibbs variational formula) via convex duality, so we can play a similar game. For certain network information theory problems, we thus obtain by far the only existing method for a converse bound which is sharp as \( n \rightarrow \infty \).
1.1.3 More on Converses in Information Theory

While the early work in information theory centered around first-order asymptotic results, such as (1.8), information theorists have always been striving for improvements:

- **Non-asymptotic bounds** where the blocklength $n$ is assumed to be fixed. For example, Fano’s inequality [28] provides a simple, though not very strong, non-asymptotic bound. While traditionally information theory has focused on the stationary memoryless settings, it is certainly of practical importance to consider processes with memory, or sources and channels with only one realization (as opposed to repeated independent realizations). In fact the latter setting is often adopted in the computer science literature, for problems in the intersection of theoretical computer science and information theory (see e.g. [29]). Moreover, even in the stationary memoryless setting, it is sometimes more desirable to consider the scenario where the alphabet size or the number of users is large compared to the number of channel/source realizations [30] [31] [32], perhaps even more so in the so-called “big data” and “IoT” era. In those cases, the traditional $n \to \infty$ assumption appears inadequate, while a versatile non-asymptotic bound may be applicable to a more relevant regime of asymptotics.

- **More precise asymptotics.** For example, in channel coding, it is possible to improve (1.8) to the precise second-order asymptotic result:

$$ M = \exp \left( nC - \sqrt{nVQ^{-1}(\epsilon)} + o(\sqrt{n}) \right) $$

(1.14)

where $Q(\cdot)$ denotes the Gaussian tail probability. The study of the second-order asymptotics in information theory was initiated by Strassen [33], and recently flourished thanks to the work of Polyanskiy et al. [34], Hayashi [35], Kostina et al. [36], and others. When $n$ is not too small, a good approximation of $M$ can
be obtained from (1.14) (sometimes significantly better than the error exponent approximation [34]). This is relevant to the practical setting where a short delay, hence a short blocklength of the code, is demanded for data transmission.

The key to achieving the above goals is to find new and better ways of proving achievability and converse bounds. For example, suppose that the goal is to determine the second-order asymptotics of a certain operational problem in network information theory. The achievability (i.e. existence) part is rather well understood: for almost all problems from network information theory with known first-order asymptotics, an achievability bound with an $O(\sqrt{n})$ second-order term can be derived [37][38]. Indeed, since the achievability part is constructive, one can usually use random coding to produce a joint distribution which is close to the ideal distribution appearing in the formula of the first-order fundamental limit, so the second-order analysis is generally reduced to Berry-Esseen central limit arguments. The problem of producing a desired joint distribution is sometimes called coordination [39].

On the other hand, the converse (i.e. impossibility) part appears subtler since the possibility of success by any code must be excluded. It is recognized within the community that the converse part is usually the bottleneck for nailing down the exact second-order asymptotics (see e.g. [40, Section V]). Below, we review three most influential ideas of proving converse results in information theory:

1. **Information spectrum method**: given probability measures $P$ and $Q$ on the same alphabet, the **information spectrum** refers to the cumulative distribution function of the **relative information** $\log \frac{dP}{dQ}(X)$ where $X \sim P$. It is possible to bound the error probability (maximum or average) of channel coding in terms of the information spectrum [26][41]. The meta-converse approach (based on binary hypothesis testing) and the Rényi divergence approach [42] are both closely related to the information spectrum approach: it is easy to see that the knowl-

---

[5]Here we do not discuss the variants such as universal coding or zero-error coding.
edge of the information spectrum is equivalent to the knowledge of the binary hypothesis testing tradeoff region for \( P \) and \( Q \) (see e.g. [33]), and the Rényi divergence between \( P \) and \( Q \) is recoverable from the information spectrum. Besides being very general and not limited to the stationary memoryless systems, the information spectrum method shows that the \( o(n) \) term in (1.12) can be improved to \( O(\sqrt{n}) \), which is in fact the optimal central limit theorem rate (see e.g. [34]). Other information-theoretic problems\(^6\) with known information spectrum converses include lossy source coding [36], Slepian-Wolf coding, multiple access channels [44], and broadcast channels [45].

2. **Type class analysis** [8]: this method is restricted to, but very powerful for, the discrete memoryless systems. The idea is to first look at, instead of a stationary memoryless distribution, the equiprobable distribution on sequences with the same empirical distribution (a.k.a *type*). In other words, imagine that a genie reveals to the encoder and the decoder the information of the type, and a different code can be constructed for each type. The evaluation of the probability then becomes a combinatorial (counting) problem, which has been referred to as the “skeleton” or the “combinatorial kernel” of the information theoretic problem [46]. It appears that those \( O(\sqrt{n}) \) second-order converses for those problems proved by the information spectrum methods can also be proved by the method of types (see [47] for the example of lossy source coding); this is not surprising since the type determines the value of the relative information, and hence may be viewed as a finer resolution than the information spectrum. For those problems, it is generally easy to see (by a combinatorial covering or packing argument) that the error probability given the type of the source or channel realization is either very close to 0 or 1, hence the task of estimating the

\(^6\)By an “information-theoretic problem” we mean an operational task such as channel coding or source coding.
error probability in the $O(\sqrt{n})$ second-order rate regime is reduced to calculating the probability of those “bad types”, which is an exercise of the central limit theorem and the Taylor expansion of the information quantities. The type class analysis has also been used extensively in [8] for deriving error exponents.

3. **Blowing-up lemma (BUL):** this method was introduced in [27] and exploited extensively in the classical book [8]; see also the recent survey [48]. We review the BUL method in Section 4.1.1. BUL draws on the nonlinear measure concentration mechanism, and has been successful in establishing the strong converse property in a wide range of multiuser information theoretic problems (including all settings in [8] with known single-letter[7] rate regions; see [8, Ch. 16/Discussion].)

<table>
<thead>
<tr>
<th></th>
<th>Information spectrum (meta-converse)</th>
<th>Type class analysis</th>
<th>Blowing-up (image size characterization)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beyond $</td>
<td>\mathcal{Y}</td>
<td>\leq \infty$?</td>
<td>✓</td>
</tr>
<tr>
<td>Second-order term</td>
<td>optimal</td>
<td>optimal</td>
<td>$O(\sqrt{n} \log^{3/2} n)$</td>
</tr>
<tr>
<td>Extension to multiuser</td>
<td>more selected</td>
<td>selected</td>
<td>all source channel networks with known first order region</td>
</tr>
</tbody>
</table>

Table 1.1: Comparison of previous converse proof techniques

1.1.4 **The Side Information Problems**

A class of problems involving side information (also called “helper” in [8]) are known to be challenging in the converse part. An archetypical example is source coding

---

7A single-letter expression (as opposed to a multi-letter expression) for a stationary memoryless system is a formula involving only the per-letter distribution.
with compressed side information. Originally studied by Wyner [49], Ahlswede, Gács and Körner [27] in the 1970’s, this problem is the first instance in network information theory where the single letter expression of the rate region requires an auxiliary random variable. Other examples of the side information problems include common randomness generation under the one-way communication protocol [15] and some source and channel networks considered in [25]. For the side information problems, BUL remained hitherto the only method for establishing the strong converse, hence the precise second-order asymptotics has been recognized as a tough problem [50, Section 9.2].

On reason such problems require BUL might be that the “combinatorial kernels” (see the paragraph explaining the type class analysis above) are not easily solved, hence it is not clear that the error probability can be estimated by simply computing the error probability of certain “bad types”. We remark that the Gray-Wyner network [51], although also requiring an auxiliary in the expression of the single-letter region, does not assume this degree of difficulty as the side information problems. In fact the exact second-order asymptotics of the Gray-Wyner network can be derived using the type class analysis described above [52].

1.1.5 Common Randomness Generation

Common randomness (CR) generation refers to the operational problem of generating shared random bits among terminals. Although the converse methodology proposed in this thesis is broadly applicable, we will use examples from the common randomness generation problem, simply as a consequence of the author’s personal research background.

The scenario depicted in Figure 1.1, which we call CR generation with one communicator, is a typical example of the side information problem. The terminals \( T_0, \ldots, T_m \) observe correlated sources. Terminal \( T_0 \) is allowed to communicate to the
other terminals, after which all terminals aim to produce a common random number \( K \). The goal is to characterize the tradeoff between the communication rates and the rate at which the random numbers are produced, under a lower bound on the probability that the random numbers produced by the terminals agree. The precise formulation of the problem is given in Section 2.3.1.

The first three chapters of the thesis form a trilogy that offers the three key ingredients for proving a \( O(\sqrt{n}) \) second-order converse for this problem in the stationary memoryless setting, which, to the best of the author’s knowledge, was beyond the reach of previous techniques. More precisely, the three ingredients allow one to accomplish the following, respectively:

- Chapter 2 presents equivalences between entropic optimizations and functional optimizations, which allow one to derive a sharp converse bound for the case where \( Y_1, \ldots, Y_m \) are functions of \( X \) and when the communication rates are close to zero.

- Chapter 3 shows that the behavior of an entropic optimization under a perturbation on the underlying measure is approximated by a mutual information optimization. This forms the basis for deriving an \( O(\sqrt{n}) \) second-order converse bound when the vanishing communication rates assumption is dropped.
To show an $O(\sqrt{n})$ second-order converse bound when further dropping the assumption that $Y_1, \ldots, Y_m$ are functions of $X$, we need the pumping-up technology in Chapter 4.

On the other hand, previous works have obtained one-shot converses using smooth Rényi entropy [53] or the meta-converse idea [54][55], for which the asymptotic tightness are achieved in the other extreme of limited correlated sources but unlimited communications. Moreover, Mossel [56] used reverse hypercontractivity to bound the probability of agreement when there is no communication at all for the binary symmetric sources, and derived the tight asymptotics when the number of terminals is growing and the probability of agreement is vanishing.

1.2 Concentration of Measure

1.2.1 The Gromov-Milman Formulation

Concentration of measure is a collection of tools and results from analysis and probability theory that have proved successful in many areas of pure and applied mathematics [57][48]. Concentration of a real valued random variable simply means that the random variable is close to its “center” (e.g. its mean or its median) with high probability.

The relevance of concentration of measure to non-asymptotic information theory is easy to see: consider the example of channel coding. By the information spectrum bounds [41], the error probability is essentially

$$\epsilon \approx \mathbb{P}[\iota_{X,Y}(X;Y) < \log M]$$

(1.15)

where $M$ denotes the number of messages, $\iota_{X,Y}(\cdot)$ is a quantity called information density, and $X$ and $Y$ are the input and the output, respectively. Note that the
expectation of $\iota_{X,Y}(X;Y)$ equals the mutual information $I(X;Y)$, hence bounding $\epsilon$ can be reduced to the concentration of $\iota_{X,Y}(X;Y)$.

Many standard tail bounds in statistics, such as Hoeffding’s inequality, Chernoff’s inequality, and Bernstein’s inequality, capture the concentration of measure of the i.i.d. sum of real valued random variables. They are sometimes called “linear concentration inequalities” since the summation is linear.

In the more general case, in which these random variables are not real valued, there are no notions of “linear functions”, “median” or “mean” of these random variables. However, intuitively we should still be able to describe the concentration phenomenon of, say, a random variable taking values in a metric space. Indeed, there are two widely accepted formulations of the concentration of measure phenomenon, equivalent in an exact way, for general metric measure spaces (See for example [57][58] for a proof of the equivalence):

1. Any 1-Lipschitz function of the random variable deviates from its median with small probability.

2. The blow-up of any set with probability at least 1/2 has a high probability.

Here, the blow-up of a set means the collection of all points whose distance to the set is not exceeding a given threshold. Setting the threshold close to zero, we see that the second formulation above implies an isoperimetric property (i.e. relationship between the surface area and the volume of the set). Conversely, isoperimetry implies concentration of measure by integration. While isoperimetry had long been considered as merely a curiosity in geometry, Milman showed in the early 1970s [59] that sphere isoperimetry can be used to give a simpler proof of the Dvoretzky theorem, a fundamental result in convex geometry originally conjectured by Grothendieck. Since then, Milman vigorously promoted the concentration of measure as a useful tool for attacking serious mathematical problems (see the discussions in [58][60]). Moreover,
Gromov studied concentration of measure on the manifolds. Thus the above formulation for the concentration of measure has been called the “Gromov-Milman” formulation (see e.g. [58]). Margulis investigated isoperimetry on the hypercube with application to a phase transition problem on the random graph in the 1970s [61]. In information theory, concentration of measure has also been called the blowing-up lemma [27] [8].

Peter Gács recalled the following (see [62]) about how the ideas of concentration of measure in the early 1970s, and in particular the result of Margulis, influenced his work (with Ahlswede and Körner) on the strong converse of the side information problem [27]:

Rudi Ahlswede, at a visit in Budapest, posed a question that I found solvable. I remembered a Dobrushin seminar presentation of Margulis about a certain phase transition in random graphs, in particular a certain isoperimetric theorem used in the proof. This is how the Blowup Lemma was born. We had a pleasant collaboration with János Körner in the course of working it into a paper, along with a couple of other results.

In the 1980s, Marton [63] invented a transportation method for proving concentration of measure, yielding simpler proofs and stronger claims for some previous results [64]. The heart of this methodology is the equivalence between the transportation-cost inequality and the concentration inequality, which can also be viewed as an instance of convex duality (see Section 2.5.7).

1.2.2 Is It Natural for Information Theorists?

Although the blowing-up lemma (the Gromov-Milman formulation of concentration for metric-measure spaces) has been successfully applied to the strong converses of many network information theory problems [8], the drawbacks are also evident: as
we alluded before, the BUL approach does not yield the optimal second-order term, and is limited to the finite alphabet case. Perhaps more importantly, assuming a metric structure on the alphabet which plays an active role is not logically satisfying. Whether it be channel coding or the side information problem, it is clear that the fundamental limit does not change if a measure-preserving map (which may very well distort the metric structure) is applied to the alphabet.

The new converse approach we introduce in this thesis addresses these issues. In particular, the Gromov-Milman formulation of concentration of measure is sidestepped, and a metric structure need not be imposed on the alphabet. We now proceed to review some elements in functional analysis that form the basis of this new methodology.

1.3 Functional Analysis

1.3.1 In Large Deviation Theory

While most textbooks on information theory first define the relative entropy by

\[ D(P\|Q) := \mathbb{E} \left[ \log \frac{dP}{dQ}(X) \right], \tag{1.16} \]

\( X \sim P \), and then prove the variational formula (1.1), Varadhan, the founder of the general theory of large deviations, chose the other way around: he defined the relative entropy via (1.1) as the convex dual (Legendre-Fenchel dual) of the cumulant

\footnote{Although BUL can be extended beyond finite alphabets, the BUL approach to strong converses involves a union bound step which requires finite alphabets.}
generating function and then proved its equivalence to \((1.16)\) Section 10. The “functional approach” of the present thesis follows the same spirit.

In fact, the idea of defining an object “dually” is very common in mathematics. For example, in probability theory, some authors prefer to define a probability measure as a functional on the linear space of continuous functions (i.e. using the claim of the Riesz-Kakutani theorem as the definition), and then recover the usual definition (assigning real numbers to measurable sets) and other properties \([67, 68]\). Other examples include cohomology in topology (dual of homology) and vector fields in differential geometry (dual of “local functions” on a manifold). Although not the most intuitive way of defining those objects, such “dual” definitions often turn out to be more technically convenient in deeper studies of the subject.

1.3.2 Hypercontractivity and Its Reverse

Let \(L^p(Q_Y)\) denote the linear space of all measurable function \(f\) equipped with the \(L^p\) norm:

\[
\|f\|_{L^p(Q_Y)} := \mathbb{E}^{\frac{1}{p}}[f^p(Y)]
\]  

(1.17)

where \(Y \sim Q_Y\). Let \(T\) be a linear operator from \(L^p(Q_Y)\) to \(L^q(Q_X)\); if

\[
\|T\|_{L^p(Q_Y) \rightarrow L^q(Q_X)} \leq 1
\]

(1.18)

for some \(q \geq p \geq 0\) we say \(T\) is a hypercontraction (the case of \(p = q\) is called a contraction).

\(^9\)While the equivalence of \((1.16)\) and \((1.1)\) follows immediately from the nonnegativity of the relative entropy when the \(f\) in \((1.1)\) is assumed to be measurable, in the large deviation theory the underlying space is usually assumed to be a metric-measure space (e.g. a Polish space), and the function \(f\) is assumed to be bounded and continuous. In that case, the proof of the equivalence of \((1.16)\) and \((1.1)\) (due to Donsker and Varadhan) is more involved \([66]\).

\(^{10}\)Recall that the norm of an operator \(T\) is defined as the supremum norm of \(Tf\) over \(f\) whose norm is bounded by 1.
For information theorists, such an operator $T$ arises naturally when a random transformation $Q_{Y|X}$ is considered (in which case $T$ is called a conditional expectation operator): Given $Q_{Y|X}$, the corresponding $T$ maps a function $f$ on $\mathcal{Y}$ to

$$x \mapsto \mathbb{E}[f(Y)|X = x]$$

(1.19)

which is a function on $\mathcal{X}$, where $Y \sim Q_{Y|X=x}$ conditioned on $X = x$. For some delicate problems (e.g. the proof of the dual formulation of certain functional inequalities in Chapter 2), it is more technically convenient (and perhaps also more elegant) to define those objects “dually”: as alluded before, a measure can be thought of as a linear functional for a suitable space of test functions. Thus, a random transformation $Q_{Y|X}$ (yielding a linear mapping taking a measure on $\mathcal{X}$ to a measure on $\mathcal{Y}$) can be thought of as the dual (or conjugate) of a conditional expectation operator $T$ mapping functions on $\mathcal{Y}$ to functions on $\mathcal{X}$.

Hypercontractivity was initially considered in theoretical physics [69] and later in functional analysis [70], theoretical computer science [71], and statistics [72] [73]. In information theory, it was introduced by Ahlswede-Gács [7], and more recently received revived interest through the work of Anantharam et al. [9]. It is known [74] [75] [76] that the functional definition (1.18) is equivalent to an information theoretic characterization, which we discuss in more details in Section 2.5.4. The strong data processing constant [25] can be recovered from the region of hypercontractive parameters $p$ and $q$ via a limiting argument [7].

The operator $T$ is said to satisfy a reverse hypercontractivity if

$$\|Tf\|_{L^q(Q_X)} \geq \|f\|_{L^p(Q_Y)}, \quad \forall f \geq 0$$

(1.20)

for some $0 \leq q \leq p \leq 1$. Initially studied by Borell in the Gaussian and the Bernoulli cases in the 1970s, reverse hypercontractivity has not found many applications until
the recent work of Mossel et al. [77]. Information theoretic formulations of the reverse hypercontractivity were studied in [78] [79] [80].

In this thesis, we show that the reverse hypercontractivity mechanism gives rise to a general approach of proving strong converses which uniformly improves the traditional blowing-up methodology. In hindsight, this makes good sense, since literally, the reverse of “contraction” is nothing but “expansion” (i.e. blowing-up). However, this cute observation was not exploited before, even though Ahlswede, Gác and Körner have worked on both hypercontractivity [7] and the strong converse problem [27] around 1976.

We remark that various bounds between hypercontractivity (and its reverse) and measure concentration have been discussed in certain special cases such as Gaussian [81, P116] [82] or Bernoulli [56]. However, our present application of reverse hypercontractivity relies on different arguments, and is also more canonical and convincing than those discussions in terms of the generality and the sharpness of our results.

### 1.3.3 Brascamp-Lieb Inequality

The Brascamp-Lieb inequality is a functional inequality that assumes as special cases, or closely relates to, many other familiar inequalities, including Hölder’s inequality, the sharp Young inequality, the Loomis-Whitney inequality, hypercontractivity, the entropy power inequality, and the logarithmic Sobolev inequality. Originally studied by Brascamp and Lieb [83] in the 1970s motivated by problems in particle physics, the Brascamp-Lieb inequality has now found applications in numerous other fields such as convex geometry [84], statistics [85], and theoretical computer science [86]. The connection to information theory was already observed in the paper of Cover and Dembo [87] in the context of the entropy power inequality. However, the present thesis will investigate a more general connection to information via convex duality and its implications, as we explain next.
By convex duality, the (functional form of the) Brascamp-Lieb inequality has an equivalent formulation in terms of a bound on the linear combination of relative entropies. Such an equivalent formulation was observed in the work of Carlen and Cordero-Erausquin [76]. Alternative proofs (without using the Gibbs variational formula) and extensions are recently proposed by Nair [74] and Liu et al. [80]. Similarly, the reverse Brascamp-Lieb inequality [88] also has equivalent information theoretic formulations; see [79] [80].

In high dimensions, we show that a small perturbation of the underlying measure in the Brascamp-Lieb inequality changes the first-order asymptotics to a certain linear combination of mutual informations (instead of the linear combinations of the relative entropy). Such a perturbation idea is a generalization of that of the smooth entropy arising in quantum information theory [53]. Thus, we obtain a correspondence between these mutual informations (which appear in the expression of single-letter regions) and a functional inequality, which can be used in the strong converse proofs if we take the functions to be the indicator functions of certain encoding or decoding sets. In the case of finite alphabets, similar results were derived under the moniker “image-size characterization” [8], using the monotonicity of relative entropy instead of convex duality.

The Brascamp-Lieb inequality can be restated as an optimization problem over some measurable functions. The original work of Brascamp and Lieb [83] proved that Gaussian functions achieve the optimal value. Since then, various other proofs of the Gaussian optimality have been found [89] [90] [75] [76] [91]. Among them, Lieb’s proof [89] used the fact that two independent random variables whose sum and difference are also independent must be Gaussian. Using the same property of the Gaussian distribution but working with information theoretic inequalities, Geng-Nair [92] proved the Gaussian optimality in the single-letter region of several network information theory problems. Liu et al. [93] gave information-theoretic proofs of the Brascamp-Lieb
inequality and its extensions with several improvements over the methodology of Lieb [89] and Geng-Nair [92].

1.4 Organization of the Thesis

The trilogy of the proposal of the functional approach to information theoretic converses is comprised of the first three chapters of the thesis. Each of these chapters describes one of the three main ingredients: the convex duality of the functional and entropic optimizations; the behavior of the optimal value under a perturbation of the underlying measure; and the reverse hypercontractivity of Markov semigroups. In each of these chapters we will return to the one-communicator common randomness generation problem introduced in Section 1.1.5 and show how the introduction of the new ingredients gets us closer to the goal of obtaining an optimal $O(\sqrt{n})$ second-order converse. Towards the end of Chapter 4 we also discuss how to apply the proposed methodology to other problems in network information theory, such as source networks, broadcast channels, and the empirical distribution of good channel codes. Chapter 2 contains additional results on convex duality which are of independent interest, but not needed for the applications to common randomness generation.

Chapter 5 presents additional examples of information measures, their dual formulations, and the applications. $E_\gamma$ is a generalization of the total variational distance. We get a glimpse of the power of $E_\gamma$ through channel resolvability in the large deviation regime, and applications of the result to source coding, broadcast channels, and wiretap channels. $g^*$ is a functional variant of the quantity $g$ in the chapter on “image-size characterizations” in the book of Csiszár-Körner [8]. The proposed $g^*$ quantity can substitute the $g$ in the converse proofs of network information theory problems such as Gelfand-Pinsker coding, but the analysis of the former is cleaner and also yields the optimal scaling of the second-order term.
While Chapters 2-5 are concerned with the applications of functional inequalities to problems in information theory, Chapter 6 offers the counterpoint that entropic inequalities and standard techniques in information theory provide simpler alternative approaches to certain functional inequalities, such as establishing the Gaussian optimality in the Brascamp-Lieb inequality.

Given the successes of functional inequalities in common randomness generation, source coding and channel coding, it is tempting to extend this paradigm to the secret key generation problem, which imposes security constraints on common randomness generation, as well as the interactive common randomness generation problem where terminals may communicate back-and-forth under a rate constraint. We discuss these operational problems and the challenges of extending the functional approach to these settings in Chapter 7.

With the understanding that readers from various backgrounds may be interested in different aspects of the thesis, we provide the following recommended itineraries for the first reading:

- Readers interested in strong converse and non-asymptotic information theory may prefer to look closely at the first four chapters, excluding Sections 2.4-2.5.
- Readers interested in information security, especially common generation or key agreement, may prefer to read Sections 2.3, 3.3, 4.4.4 for the converse proof for the one-communicator common randomness generation model, and then visit Section 5.2.6 for the application of $E_\gamma$ to the wiretap channels. Moreover, a summary of some other results on secret key generation by the author can be found in Chapter 7.
- Readers interested in information measures or the proof of Gaussian optimality in the rate regions in network information theory may prefer to focus on Chapters 2 and 6.
Chapter 2

A Tale of Two Optimizations

It is a functional optimization; it is an entropic optimization. We investigate which viewpoint is better, for various problems under consideration.

This chapter investigates the fundamental convex duality machinery that explains the connection between the functional optimization and the entropic optimization. To be concrete, we look at the example of the Brascamp-Lieb inequality from functional analysis. More precisely, we consider a slight extension of the Brascamp-Lieb inequality tailored to our information-theoretic applications. The entropic formulation of such a functional inequality is proved in Sections 2.1-2.2. The result is then applied to the common randomness generation problem in Section 2.3 where a zero-rate single-shot converse bound is derived. The reverse Brascamp-Lieb inequality, although formally symmetrical to the Brascamp-Lieb inequality, requires more sophistication in the proof of the equivalence to the entropic formulation (Section 2.4), and may be skipped by readers who are only interested in the applications to network information theory. Section 2.5 reviews several special cases of the Brascamp-Lieb inequality and its reverse.
2.1 The Duality between Functions and Measures

We first introduce some general notations and definitions used in this thesis. An alphabet (denoted by $\mathcal{X}$, $\mathcal{Y}$ etc.) is a measurable space, that is, a set equipped with a $\sigma$-algebra. The set of all real-valued measurable functions (denoted by $f$, $g$ etc.) on a given alphabet forms a linear space. We use Greek letters (e.g. $\mu$) for unnormalized measures, capital Roman letters (e.g. $P$ and $Q$) for probability measures. A measure $\mu$ defines a linear functional on the space of functions via integration:

$$ f \mapsto \int f \, d\mu, \quad (2.1) $$

hence the space of measures may be viewed as the dual space (see e.g. [67]) of the space of “test functions”. Note that this is a general and informal statement since we haven’t specified what are the “test functions” yet. When we deal with the reverse Brascamp-Lieb inequality in later sections of this chapter, it is technically necessary to choose a “nice” space of functions (e.g. bounded continuous functions) so that a certain space of measures indeed forms the set of all bounded linear functionals, meaning that the space of measures is the dual of the space of functions in the precise mathematical sense [67]. However, in other places in this thesis, it is generally more convenient to consider the ensembles of non-negative functions and non-negative measures instead. In this case, the duality between the two linear spaces no longer holds in the conventional precise sense of functional analysis (since $\mu$ may no longer be bounded as a linear functional), but the integration (2.1) still always makes sense (possibly equal to $+\infty$) thanks to the nonnegativity assumption. Hence, except in sections related to the reverse-type inequalities, no additional assumptions (e.g. a topological structure) need to be imposed on the alphabet other than being a measurable space.

A fundamental quantity in information theory is the relative entropy. First, given two nonnegative $\sigma$-finite measures $\theta \ll \mu$ on $\mathcal{X}$, let us define the relative information
as the logarithm of the Radon-Nikodym derivative:

$$\theta_{\mu}(x) := \log \frac{d\theta}{d\mu}(x)$$

(2.2)

where $x \in \mathcal{X}$. Note that there is no assumption about $|\mathcal{X}|$, and recall that the right side of (2.2) is unique up to values on a set of $\mu$-measure zero. Then, the relative entropy between a probability measure $P$ and a $\sigma$-finite measure $\mu$ on the same measurable space is defined as

$$D(P \| \mu) := \mathbb{E}[\theta_{P\|\mu}(X)]$$

(2.3)

where $X \sim P$, if $P \ll \mu$, and infinity otherwise.

A fundamental quantity in functional analysis is the norm: for $p \in (0, \infty)$ and a real-valued measurable function $f$ on $\mathcal{X}$, define the $p$-th norm

$$\|f\|_{L^p(\mu)} := \left( \int |f|^p d\mu \right)^{\frac{1}{p}}.$$ 

(2.4)

Sometimes we abbreviate $\|f\|_{L^p(\mu)}$ as $\|f\|_p$, or as $\|f(X)\|_p$ if $\mu$ is a probability measure and $X \sim \mu$. Note that if $\mu$ is a probability measure, and $g$ is a real-valued function on $\mathcal{X}$, then

$$\log \|\exp(g)\|_p = \frac{1}{p} \log \mathbb{E}[\exp(pg(X))]$$

(2.5)

$$= \frac{1}{p} \Lambda_{g(\bar{X})}(p)$$

(2.6)

where $\Lambda_{g(\bar{X})}(p)$ denotes the cumulant-generating function of the real-valued random variable $g(X)$.
Using a bit of convex analysis (see e.g. [94]) one can verify that the convex functional

\[ P \mapsto \frac{1}{p} D(P\|\mu) \quad (2.7) \]

is the Legendre dual of the convex functional

\[ g \mapsto \log \| \exp(g) \|_p. \quad (2.8) \]

This basic fact hints that functional inequalities in functional analysis (typically involving integrals and norms) have dual correspondences with inequalities in information theory (typically involving the relative entropy). In the next section, we explore a very canonical instance of such a correspondence, in the context of the Brascamp-Lieb inequality.

### 2.2 Dual Formulations of the Forward Brascamp-Lieb Inequality

The Brascamp-Lieb inequality\(^1\), originally motivated by problems from particle physics [83], can be viewed as a property of the “correlation” among random variables, and turns out to be naturally suited for the analysis of the common randomness generation problem. In the literature, the Brascamp-Lieb inequality usually refers to the fact that Gaussian functions saturate a certain functional inequality. To be concrete, let us look at a modern formulation of the result from Barthe’s paper [88]\(^2\).

\(^1\)Not to be confused with the “variance Brascamp-Lieb inequality” (cf. [95] [96] [97]), a different type of inequality that generalizes the Poincaré inequality.

\(^2\) [88, Theorem 1] actually contains additional assumptions, which make the best constants \(D\) and \(F\) positive and finite, but not really necessary for the conclusion to hold ([88, Remark 1]).
Theorem 2.2.1 ([88, Theorem 1]) Let $E$, $E_1$, ..., $E_m$ be Euclidean spaces, equipped with the Lebesgue measure, and $B_i : E \rightarrow E_i$ be linear maps. Let $(c_i)_{i=1}^m$ and $D$ be positive real numbers. Then the Brascamp-Lieb inequality

$$\int \prod_{i=1}^m f_i(B_i x) \, dx \leq D \prod_{i=1}^m \|f_i\|_{\frac{1}{c_i}},$$

(2.9)

where

$$\|f_j\|_{\frac{1}{c_j}} := \mathbb{E}^{c_j} \left[ \frac{1}{f_j^{\frac{1}{c_j}}}(Y_j) \right],$$

(2.10)

for all nonnegative measurable functions $f_i$ on $E_i$, $i = 1, \ldots, m$, holds if and only if it holds whenever $f_i$, $i = 1, \ldots, m$ are centered Gaussian functions.

Various proofs of the Gaussian optimality in the Brascamp-Lieb inequality have been proposed since the original paper of Brascamp and Lieb [83]; we further discuss these matters in Section 6.2. In this chapter, however, we are only interested in the form of the inequality (2.9) rather than the Gaussian optimality result. In particular, we do not restrict our attention to the setting of Euclidean spaces and linear maps: the underlying spaces and the mappings between them can be arbitrary.

The form of the inequality (2.9) can be seen as a generalization of several other familiar inequalities, including Hölder’s inequality, the sharp Young inequality, the Loomis-Whitney inequality, the entropy power inequality, hypercontractivity and the logarithmic Sobolev inequality; we further discuss these special cases in Section 2.5 (see also [98] [99] [70]).

The Brascamp-Lieb inequality is known to imply several information-theoretic inequalities, such as the entropy power inequality [87, Theorem 12] and the entropic uncertainty principle [100]; see also the summary in [93]. In this thesis, however, we

---

3 A centered Gaussian function is of the form $x \mapsto \exp(-x^\top A x)$ where $A$ is a positive semidefinite matrix.
are mainly interested in the dual formulations of the inequality (2.9). Such duality was noted by Carlen and Cordero-Erausquin [76, Theorem 2.1] (see also [75] for a special case). We reproduce their result here:

**Theorem 2.2.2 ([76, Theorem 2.1])** Let $X, Y_1, \ldots, Y_m$ be measurable spaces. Consider a probability measure $Q_X$, measurable maps $\phi_j : X \to Y_j$, $j = 1, \ldots, m$, probability measures $Q_{Y_1}, \ldots, Q_{Y_m}$ (not necessarily induced by $Q_X$ and $(\phi_j)_{j=1}^m$), and $c_1, \ldots, c_m \in (0, \infty)$. For any $d \in \mathbb{R}$, the following statements are equivalent:

1. For any non-negative measurable functions $f_j : Y_j \to \mathbb{R}$, $j = 1, \ldots, m$, it holds that

\[
\mathbb{E} \left[ \prod_{j=1}^{m} f_j(\phi_j(X)) \right] \leq \exp(d) \prod_{j=1}^{m} \| f_j \|_{\phi_j}^{\frac{1}{c_j}} \tag{2.11}
\]

where $X \sim Q_X$ and $Y_j \sim Q_{Y_j}$.

2. For any probability measure $P_X \ll Q_X$,

\[
D(P_X \| Q_X) + d \geq \sum_{j=1}^{m} c_j D(P_{Y_j} \| Q_{Y_j}) \tag{2.12}
\]

where $P_{Y_j}$ is induced by $P_X$ and $\phi_j$, $j = 1, \ldots, m$.

In this section we prove an extension of the duality of Carlen and Cordero-Erausquin, with a functional inequality that generalizes (2.11) by allowing a cost function and random transformations in lieu of $\{\phi_j\}$. Both generalizations are essential for certain information-theoretic applications. The “forward-reverse Brascamp-Lieb inequality” alluded before will be introduced in later sections since, although the forward-reverse inequality essentially generalizes the forward inequality (2.11),
the proof of its equivalent formulation is more involved and applies only to certain “regular” (though fairly general) spaces.\footnote{More precisely, the “entropic⇒functional” direction is not more difficult than the forward inequality, but the “functional⇒entropic” direction requires sophisticated min-max theorems and is only proved in for Polish spaces. In the finite alphabet case, the latter difficulty can be circumvented by using the KKT conditions \cite{80}.}

**Theorem 2.2.3** \textit{Fix a probability measure} $Q_X$, \textit{an integer} $m \in \{1, 2, \ldots \}$, \textit{c}$_j$ \textit{∈ (0,} $\infty$), \textit{and random transformation} $Q_{Y_j\lvert X}$ \textit{for each} $j \in \{1, \ldots, m\}$. \textit{Let} $(X, Y_j) \sim Q_X Q_{Y_j\lvert X}$. \textit{Assume that} $d: X \to (-\infty, \infty]$ \textit{is a measurable function satisfying}

$$0 < \mathbb{E}[\exp(-d(X))] < \infty. \quad (2.13)$$

The following statements are equivalent:

1. \textit{For any non-negative measurable functions} $f_j: Y_j \to \mathbb{R}$, \textit{j} \textit{∈} \{1, \ldots, m\}, \textit{it holds that}

$$\mathbb{E} \left[ \exp \left( \sum_{j=1}^{m} \mathbb{E}[\log f_j(Y_j\lvert X)] - d(X) \right) \right] \leq \prod_{j=1}^{m} \|f_j\|_{1/c_j} \quad (2.14)$$

where the norm $\|f_j\|_{1/c_j}$ is with respect to $Q_{Y_j}$.

2. \textit{For any distribution} $P_X \ll Q_X$, \textit{it holds that}

$$D(P_X\lvert\rvert Q_X) + \mathbb{E}[d(\hat{X})] \geq \sum_{j=1}^{m} c_j D(P_{Y_j}\lvert\rvert Q_{Y_j}) \quad (2.15)$$

where $\hat{X} \sim P_X$, and $P_X \to Q_{Y_j\lvert X} \to P_{Y_j}$ \textit{for} $j \in \{1, \ldots, m\}$.

**Proof** \hspace{1em} 1)⇒2) Define

$$d_0 := \log \mathbb{E} \left[ \exp(-d(X)) \prod_{j=1}^{m} \exp(c_j \mathbb{E}[t_{P_{Y_j\lvert Q_{Y_j}}}(Y_j\lvert X)]) \right]. \quad (2.16)$$

\footnote{Throughout the thesis with the exception of the discussions on the reverse Brascamp-Lieb inequality (mostly, Section \ref{3.4}), the definition of a random transformation is the same as the conventional definition of a regular conditional probability (see e.g. \cite{43}).}

29
Invoking statement 1) with

\[ f_j \leftarrow \left( \frac{dP_{Y_j}}{dQ_{Y_j}} \right)^{c_j} \]  \hspace{1cm} (2.17)

we obtain

\[ \exp(d_0) \leq \prod_{j=1}^{m} \left( \mathbb{E} \left[ \frac{dP_{Y_j}}{dQ_{Y_j}}(Y_j) \right] \right)^{c_j} = 1. \]  \hspace{1cm} (2.18)

Now if \( 0 < \exp(d_0) \leq 1 \) then

\[ dS_X(x) := \exp(-d(x) - d_0) \prod_{j=1}^{m} \exp(c_j \mathbb{E} [\mathbb{I}_{P_{Y_j}||Q_{Y_j}}(Y_j)|X = x])dQ_X(x) \]  \hspace{1cm} (2.19)

is a probability measure. Then (2.18) combined with the nonnegativity of relative entropy shows that

\[ \sum_{j=1}^{m} c_j D(P_{Y_j}||Q_{Y_j}) \leq D(P_X||S_X) - d_0 + \sum_{j=1}^{m} c_j D(P_{Y_j}||Q_{Y_j}) \]  \hspace{1cm} (2.20)

\[ = D(P_X||Q_X) + \mathbb{E}[d(\hat{X})] \]  \hspace{1cm} (2.21)

and statement 2) holds. On the other hand, if \( \exp(d_0) = 0 \), then for \( Q_X \)-almost all \( x \),

\[ \exp(-d(x)) \prod_{j=1}^{m} \exp(c_j \mathbb{E} [\mathbb{I}_{P_{Y_j}||Q_{Y_j}}(Y_j)|X = x]) = 0. \]  \hspace{1cm} (2.22)

Taking logarithms on both sides and taking the expectation with respect to \( P_X \), we have

\[ -\mathbb{E}[d(\hat{X})] + \sum_{j=1}^{m} c_j D(P_{Y_j}||Q_{Y_j}) = -\infty \]  \hspace{1cm} (2.23)
and statement 2) also follows.

- 2)⇒1) It suffices to prove for \( f_j \)'s such that \( 0 < a < f_j < b < \infty \) for some \( a \) and \( b \), since the general case will then follow by taking limits (e.g. using monotone convergence theorem). By this assumption and (2.13), we can always define \( P_X \) through

\[
\log \left( \sum_{j=1}^{m} \log f_j(Y_j) \right) = \sum_{j=1}^{m} \log f_j(Y_j) - d_0 + \mathbb{E} \left[ \sum_{j=1}^{m} \log f_j(Y_j) \middle| X = x \right] \tag{2.24}
\]

and \( S_{Y_j} \) through

\[
\log \left( \sum_{j=1}^{m} \log f_j(Y_j) \right) = \sum_{j=1}^{m} \log f_j(Y_j) - d_j , \tag{2.25}
\]

for each \( j \in \{1, \ldots, m\} \), where \( d_j \in \mathbb{R}, j \in \{0, \ldots, m\} \) are normalization constants, therefore

\[
\exp(d_0) = \mathbb{E} \left[ \exp \left( -d(X) + \mathbb{E} \left[ \sum_{j=1}^{m} \log f_j(Y_j) \middle| X \right] \right) \right] ; \tag{2.26}
\]

\[
\exp(d_j) = \mathbb{E} \left[ \exp \left( \frac{1}{c_j} \log f_j(Y_j) \right) \right] , \quad j \in \{1, \ldots, m\} \tag{2.27}
\]

But direct computation gives

\[
D(P_X||Q_X) = -\mathbb{E}[d(\hat{X})] - d_0 + \mathbb{E} \left[ \sum_{j=1}^{m} \log f_j(\hat{Y}_j) \right] \tag{2.28}
\]

\[
D(P_{Y_j}||Q_{Y_j}) = D(P_{Y_j}||S_{Y_j}) + \mathbb{E} \left[ \log \left( \frac{1}{c_j} \log f_j(\hat{Y}_j) \right) \right] \tag{2.29}
\]

\[
= D(P_{Y_j}||S_{Y_j}) - d_j + \mathbb{E} \left[ \frac{1}{c_j} \log f_j(\hat{Y}_j) \right] \tag{2.30}
\]
where \( \hat{Y}_j \sim P_{Y_j} \). Therefore statement 2) yields

\[
-d_0 + \mathbb{E} \left[ \sum_{j=1}^{m} \log f_j(\hat{Y}_j) \right] \geq \sum_{j=1}^{m} c_j D(P_{Y_j} \| S_{Y_j}) - \sum_{j=1}^{m} c_j d_j + \mathbb{E} \left[ \sum_{j=1}^{m} \log f_j(\hat{Y}_j) \right].
\]

(2.31)

Since \( f_j \)'s are assumed to be bounded, \(-\infty < \mathbb{E} \left[ \sum_{j=1}^{m} \log f_j(\hat{Y}_j) \right] < \infty\) so we can cancel it from the two sides of the inequality. It then follows from the non-negativity of relative entropy that

\[
d_0 \leq \sum_{j=1}^{m} c_j d_j
\]

(2.32)

which is equivalent to statement 1) in view of (2.26) and (2.27).

Remark 2.2.1 We prove Theorem 2.2.3 by defining certain auxiliary measures and then reducing (2.14) or (2.15) to the nonnegativity of relative entropy. Alternatively, Theorem 2.2.3 may be proved with a variational formula for the relative entropy in a similar manner as the proof of [76, Theorem 2.1].

Remark 2.2.2 As a convention, if the left side of (2.15) is \( \infty - \infty \) it should be understood as \( \mathbb{E} \left[ t_{P_X|Q_X}(\hat{X}) + d(\hat{X}) \right] \) with \( \hat{X} \sim P_X \). Note that the assumption (2.13) ensures that

\[
dT_X(x) := \frac{\exp(-d(x))dQ_X(x)}{\mathbb{E}[\exp(-d(X))]}
\]

(2.33)

defines a probability measure \( T_X \) so that

\[
\mathbb{E} \left[ t_{P_X|Q_X}(\hat{X}) + d(\hat{X}) \right] = D(P_X \| T_X) - \log \mathbb{E}[\exp(-d(X))] \]

(2.34)
is always well-defined (finite or $+\infty$).

**Remark 2.2.3** In the statements of Theorem 2.2.3, $Q_X$ is a probability measure and $Q_X$ and $Q_{Y_j}$ are connected through the random transformation $Q_{Y_j|X}$.

These help to keep our notations simple and suffice for most of our applications. However, from the proof it is clear that these restrictions are not really necessary. In other words, we have the extension of Theorem 2.2.3 that the following two statements are equivalent:

\[
\int \exp \left( \sum_{j=1}^{m} \mathbb{E}[\log f_j(Y_j)|X = x] - d(x) \right) d\nu(x) \leq \prod_{j=1}^{m} \|f_j\|_{\frac{1}{c_j}}, \quad \forall f_1, \ldots, f_m; \tag{2.35}
\]

\[
D(P_X||\nu) + \int d(x)d\nu(x) \geq \sum_{j=1}^{m} c_j D(P_{Y_j||\mu_j}), \quad \forall P_X, \tag{2.36}
\]

where again $f_j: \mathcal{Y}_j \to \mathbb{R}, j \in \{1, \ldots, m\}$ are nonnegative measurable functions, $P_X \ll \nu$, and $P_X \to Q_{Y_j|X} \to P_{Y_j}$ for $j \in \{1, \ldots, m\}$. Here the measures $\nu$ and $\mu_j$ need not be normalized and need not be connected by $Q_{Y_j|X}$, and $\| \cdot \|_{\frac{1}{c_j}}$ is with respect to $\mu_j$.

## 2.3 Application to Common Randomness Generation

We now use Theorem 2.2.3 to prove a preliminary converse bound for common randomness generation.

### 2.3.1 The One Communicator Problem: the Setup

We first formulate the problem in the one-shot setting. Let $Q_{XY^m}$ be the joint distribution of sources $X, Y_1, \ldots, Y_m$, observed by terminals $T_0, \ldots, T_m$ as shown in Figure 1.1. The communicator $T_0$ computes/encodes (possibly stochastically) the
integers $W_1(X), \ldots, W_m(X)$ and sends them to $T_1, \ldots, T_m$, respectively. Then, terminals $T_0, \ldots, T_m$ compute/decode (either deterministically or stochastically, which will be specified in the statements of our results) integers $K(X), K_1(Y_1, W_1), \ldots, K_m(Y_m, W_m)$. The goal is to produce $K = K_1 = \cdots = K_m$ with high probability with $K$ almost equiprobable.

In the stationary memoryless case where the sources have the per-letter distribution $Q_{XY}$, take $X \leftarrow X^n$, $Y_j \leftarrow Y_j^n$ in the above formulation.

**Definition 2.3.1** The $(m+1)$-tuple $(R, R_1, \ldots, R_m)$ is said to be achievable if a sequence of key generation schemes can be designed to fulfill the following conditions:

\[
\lim_{n \to \infty} \frac{1}{n} H(K) \geq R; \tag{2.37}
\]

\[
\lim_{n \to \infty} \frac{1}{n} \log |W_l| \leq R_l, \quad l = 1, \ldots, m; \tag{2.38}
\]

\[
\lim_{n \to \infty} \max_{1 \leq l \leq m} \mathbb{P}[K \neq K_l] = 0. \tag{2.39}
\]

Irrespective of the stochasticity of the decoders (cf. [16][101]), the achievable region is the set of $(R, R_1, \ldots, R_m) \in [0, \infty)^{m+1}$ such that

\[
\sup_{R_{U|X}} \left\{ \sum_{j=1}^{m} c_j I(U; Y_j) - I(U; X) \right\} + \sum_{j} c_j R_j \geq \left( \sum_{j} c_j - 1 \right) R \tag{2.40}
\]

for all $c^n \in (0, \infty)^m$ (equivalently, for all $c^n \in (0, \infty)^m$ such that $\sum c_j > 1$), where $(U, X, Y_j) \sim P_{U|X}Q_{XY}^j$.

**Remark 2.3.1** It is widely known in the information-theoretic secrecy literature (see e.g. [102] for a discussion) that the rate region (2.40) is rather “robust” under various definitions. For example, (2.40) does not change if one replaces (2.37) and (2.39) in
the conventional definition with

\[
\liminf_{n \to \infty} \frac{1}{n} \log |\mathcal{K}| \geq R; \tag{2.41}
\]

\[
\lim_{n \to \infty} |P_{KK^m} - T_{KK^m}| = 0, \tag{2.42}
\]

where \(T_{KK^m}\) denotes the ideal distribution:

\[
T_{KK^m}(k, k^m) = \frac{1}{|\mathcal{K}|} 1\{k = k_1 = \cdots = k_m\}, \quad \forall k, k_1, \ldots, k_m. \tag{2.43}
\]

In fact, such a definition appears to be more natural for the second-order analysis in the nonvanishing error regime.

### 2.3.2 A Zero-Rate One-Shot Converse for the Omniscient Helper Problem

Although the single-letter rate region is known for the one-communicator problem (2.40), the previous proof based on Fano’s inequality only establishes a weak converse result, that is, the error probability is assumed to be vanishing. As mentioned before, a main goal of the first three chapters is to supply general converse tools to solve, in particular, the optimal scaling of the second-order rate for the one-communicator problem with nonvanishing error probability. In this section, however, we achieve an intermediate goal, which is more modest than the final goal in the following senses:

- The first order asymptotics of the bound is tight only when the supremum in (2.40) is zero. In other words, we bound the maximum ratio of the log alphabet sizes of the key and the messages such that the key can be successfully generated in the omniscient helper problem. Since this ratio is supremized as the key rate and the communication rates tend to zero, such a converse bound may also be
called a zero-rate converse. The bound is only asymptotically tight in the case of abundant correlated sources but limited communication rates.

- We consider only the omniscient helper special case of the one-communicator problem, where \(X = Y^m\). Actually, the analysis should work as long as each \(Y_j\) is a (deterministic) function of \(X\), but let us consider \(X = Y^m\) for simplicity.

We remark that the above two limitations will be remedied by techniques introduced in Chapters 3, 4 respectively.

Now, the rate region (2.40) suggests that we should find a correspondence between the linear combination of mutual informations to a certain functional inequality. Suppose that random variables \(Y_1, \ldots, Y_m\) and \(c_1, \ldots, c_m \in (0, \infty)\) satisfy

\[
\mathbb{E} \left[ \prod_{l=1}^{m} f_l(Y_l) \right] \leq \prod_{l=1}^{m} \|f_l(Y_l)\|_{\frac{1}{\alpha_l}}
\]

(2.44)

for all bounded real-valued measurable functions \(f_l\) defined on \(\mathcal{Y}_l, l = 1, \ldots, m\). Then Theorem 2.2.3 shows that (2.44) is equivalent to

\[
D(P_{Y^m} \| Q_{Y^m}) \geq \sum_{l=1}^{m} c_l D(P_{Y_l} \| Q_{Y_l})
\]

(2.45)

for all \(P_{Y^m} \ll Q_{Y^m}\). To further relate (2.45) to the rate region formula (2.40), we need the following result:

**Proposition 2.3.1** For given \(Q_{Y^m}\) and \(c_1, \ldots, c_m \in [1, +\infty)\), (2.45) holds for all \(P_{Y^m} \ll Q_{Y^m}\) if and only if

\[
I(U; Y^m) \geq \sum_{l=1}^{m} c_l I(U; Y_l)
\]

(2.46)

holds for all \(P_{U|Y^m}\).
Proof  The \((2.45) \Rightarrow (2.46)\) part follows immediately from the definition of the mutual information. The less trivial direction \((2.46) \Rightarrow (2.45)\) can be shown by constructing a random variable \(U\) that takes two values; see for example [74].

In view of the rate region formula \((2.40)\), a rate tuple \((R, R_1, \ldots, R_m)\) is not achievable if

\[
R < \sum_{l=1}^{m} c_l(R - R_l) \tag{2.47}
\]

holds for some \(c_1, \ldots, c_m\) satisfying the assumption in Proposition 2.3.1. Conversely, if \(r_1, \ldots, r_m \in (0, \infty)\) satisfies the property that \(1 \geq \sum_{l=1}^{m} c_l(1 - r_l)\) for all \(c_1, \ldots, c_m\) such that the assumption in Proposition 2.3.1 is satisfied, then for any \(\delta > 0\) there exists \((R, R_1, R_2, \ldots, R_m)\) achievable such that \(\frac{R_l}{R} < r_l + \delta\) for each \(l = 1, \ldots, m\). In other words, the set of admissible \((c_1, \ldots, c_m)\) completely characterizes the set of the achievable ratios of \(\frac{R_1}{R}, \ldots, \frac{R_m}{R}\).

We are now ready to prove a zero-rate one-shot converse for the omniscient helper problem. Suppose the (possibly stochastic) encoder for the public messages is specified by \(P_{W^m|Y^m}\) and the (possibly stochastic) decoder for the key is given by \(\prod_{l=1}^{m} P_{K_l|Y_lW_l}\). Let \(T_{K^m}\) be the correct distribution under which \(K_1 = K_2 = \cdots = K_m\) is equiprobably distributed on \(\mathcal{K}\). Clearly, a small total variation \(|P_{K^m} - T_{K^m}|\) implies both uniformity of the key distribution and a small probability of key disagreement; in fact in the stationary memoryless case it can be shown that this is asymptotically equivalent to imposing both the entropy constraint \((2.37)\) and the error constraint \((2.39)\).
Theorem 2.3.2 In the omniscient helper problem, if \( Y^m \) and \( c_1, \ldots, c_m \) satisfies the condition in Proposition 2.3.1, then

\[
\frac{1}{2} |P_{K^m} - T_{K^m}| \geq 1 - \frac{1}{|\mathcal{K}|} - \left[ |\mathcal{K}| \prod_{l=1}^{m} \left( \frac{|W_l|}{|\mathcal{K}|} \right)^{c_l} \right]^{1/\sum c_l}.
\] (2.48)

Proof For any \( k \in \mathcal{K} \),

\[
P \left[ \bigcap_{l=1}^{m} \{ K_l = k \} \right] = \int_{Y^m} \sum_{w^m} \prod_{l=1}^{m} P_{K_l=k|Y_l=W_l=w_l} P_{W^m|Y^m} dP_{Y^m} \quad (2.49)
\]

\[
\leq \int_{Y^m} \max_{w^m} \prod_{l=1}^{m} P_{K_l=k|Y_l=W_l=w_l} dP_{Y^m} \quad (2.50)
\]

\[
\leq \int_{Y^m} \prod_{l=1}^{m} \max_{w_l} P_{K_l=k|Y_l=W_l=w_l} dP_{Y^m} \quad (2.51)
\]

\[
\leq \prod_{l=1}^{m} \left[ \int_{Y_l} \left( \max_{w_l} P_{K_l=k|Y_l=W_l=w_l} \right)^{1/c_l} dP_{Y_l} \right]^{c_l} \quad (2.52)
\]

\[
\leq \prod_{l=1}^{m} \left[ \int_{Y_l} \max_{w_l} P_{K_l=k|Y_l=W_l=w_l} dP_{Y_l} \right]^{c_l} \quad (2.53)
\]

\[
\leq \prod_{l=1}^{m} \left[ \sum_{w_l} \int_{Y_l} P_{K_l=k|Y_l=W_l=w_l} dP_{Y_l} \right]^{c_l} \quad (2.54)
\]

\[
\leq \prod_{l=1}^{m} \left[ \sum_{w_l} \int_{Y_l} P_{K_l=k|Y_l=W_l=w_l} dP_{Y_l} \right] \quad (2.55)
\]

where

- (2.53) uses the definition of hypercontractivity;

- (2.54) uses \( 0 < c_l < 1 \) and \( \max_{w_l} P_{K_l=k|Y_l=W_l=w_l} \leq 1 \).

Raising both sides of (2.55) to the power of \( \frac{1}{\sum_{l=1}^{m} c_l} \), we obtain

\[
P \left[ \bigcap_{l=1}^{m} \{ K_l = k \} \right]^{1/\sum c_l} \leq \prod_{l=1}^{m} \left[ \sum_{w_l} \int_{Y_l} P_{K_l=k|Y_l=W_l=w_l} dP_{Y_l} \right]^{c_l/\sum c_l} \quad (2.56)
\]
But the function \( t^m \rightarrow \prod_{l=1}^m t_i^{\frac{ci}{\sum\mathcal{c}_i}} \) is a concave function on \([0, \infty)^m\), so by Jensen’s inequality,

\[
\frac{1}{|\mathcal{K}|} \sum_{k=1}^{|\mathcal{K}|} \prod_{l=1}^m \left[ \sum_{y_l} \int_{Y_l} \frac{1}{|\mathcal{K}|} P_{K_l=k|Y_l=W_l=w_l} dP_{Y_l} \right]^{\frac{ci}{\sum\mathcal{c}_i}} \quad (2.57)
\]

\[
\leq \prod_{l=1}^m \left[ \sum_{w_l} \int_{Y_l} \frac{1}{|\mathcal{K}|} \sum_{k=1}^{|\mathcal{K}|} P_{K_l=k|Y_l=W_l=w_l} dP_{Y_l} \right]^{\frac{ci}{\sum\mathcal{c}_i}} \quad (2.58)
\]

\[
= \prod_{l=1}^m \left[ \sum_{w_l} \int_{Y_l} \frac{1}{|\mathcal{K}|} dP_{Y_l} \right]^{\frac{ci}{\sum\mathcal{c}_i}} \quad (2.59)
\]

\[
= \prod_{l=1}^m \left( \frac{\mathcal{W}_l}{|\mathcal{K}|} \right)^{\frac{ci}{\sum\mathcal{c}_i}} \quad (2.60)
\]

Combining (2.56) and (2.60) we obtain

\[
\frac{1}{|\mathcal{K}|} \sum_{k=1}^{|\mathcal{K}|} \mathbb{P} \left[ \bigcap_{l=1}^m \{ K_l = k \} \right]^{\frac{1}{\sum\mathcal{c}_i}} \leq \prod_{l=1}^m \left( \frac{\mathcal{W}_l}{|\mathcal{K}|} \right)^{\frac{ci}{\sum\mathcal{c}_i}}. \quad (2.61)
\]

Now, if we can show the following bound

\[
\frac{1}{2} |P_{K^m} - T_{K^m}| \\
= \sum_{k=1}^{|\mathcal{K}|} \left| \mathbb{P} \left[ \bigcap_{l=1}^m \{ K_l = k \} \right] - \frac{1}{|\mathcal{K}|} \right| + \sum_{k=1}^{|\mathcal{K}|} \mathbb{P} \left[ \bigcap_{l=1}^m \{ K_l = k \} \right] \\
\geq 1 - \frac{1}{|\mathcal{K}|} - |\mathcal{K}|^{\frac{1}{\sum\mathcal{c}_i}} \sum_{k=1}^{|\mathcal{K}|} \mathbb{P} \left[ \bigcap_{l=1}^m \{ K_l = k \} \right]^{\frac{ci}{\sum\mathcal{c}_i}} \quad (2.62)
\]

then the proof is finished by combining (2.61) and (2.63).
To finish the proof we need to find an lower bound on (2.62) in terms of the left side of (2.61). Consider the optimization problem over substochastic vector $a^{\mathcal{K}}$:

\[
\begin{align*}
\text{minimize} & \quad f(a^{\mathcal{K}}) := \sum_{k=1}^{\vert \mathcal{K} \vert} \left| a_k - \frac{1}{\vert \mathcal{K} \vert} \right| + 1 - \sum_{k=1}^{\vert \mathcal{K} \vert} a_k \\
\text{subject to} & \quad g(a^{\mathcal{K}}) := \frac{1}{\vert \mathcal{K} \vert} \sum_{k=1}^{\vert \mathcal{K} \vert} a_k \frac{1}{c_l} \leq \lambda
\end{align*}
\]

where $\lambda > 0$ is some constant. Notice that if $a^{\mathcal{K}}$ and $b^{\mathcal{K}}$ satisfies

\[
\frac{1}{\vert \mathcal{K} \vert} \sum_{k=1}^{\vert \mathcal{K} \vert} a_k = \frac{1}{\vert \mathcal{K} \vert} \sum_{k=1}^{\vert \mathcal{K} \vert} b_k
\]

and that $a_k - \frac{1}{\vert \mathcal{K} \vert}$ and $b_k - \frac{1}{\vert \mathcal{K} \vert}$ have the same sign, then (because $\frac{1}{\sum c_l} \leq 1$)

\[
\begin{align*}
&f\left(\frac{1}{2}(a^{\mathcal{K}} + b^{\mathcal{K}})\right) = \frac{1}{2}(f(a^{\mathcal{K}}) + f(b^{\mathcal{K}})) \\
g\left(\frac{1}{2}(a^{\mathcal{K}} + b^{\mathcal{K}})\right) \geq \frac{1}{2}(g(a^{\mathcal{K}}) + g(b^{\mathcal{K}})).
\end{align*}
\]

This implies that if $c^{\mathcal{K}}$ is an optimizer of (2.64), then

\[
\vert \{k : c_k \in (0, \frac{1}{\vert \mathcal{K} \vert})\} \vert \leq 1.
\]

Moreover for any $a^{\mathcal{K}}$, define $b_k := a_k 1\{a_k < \frac{1}{\vert \mathcal{K} \vert}\} + \frac{1}{\vert \mathcal{K} \vert} 1\{a_k \geq \frac{1}{\vert \mathcal{K} \vert}\}$, then

\[
\begin{align*}
f(b^{\mathcal{K}}) &= f(a^{\mathcal{K}}) \\
g(b^{\mathcal{K}}) &\leq (g(a^{\mathcal{K}}))
\end{align*}
\]

Thus one optimizer $c^{\mathcal{K}}$ of (2.64) is of the following form:

\[
c^{\mathcal{K}} = \left[ \frac{1}{\vert \mathcal{K} \vert}, \frac{1}{\vert \mathcal{K} \vert}, \ldots, \frac{1}{\vert \mathcal{K} \vert}, \eta, 0, 0, \ldots, 0 \right]
\]
for some $0 < \eta \leq \frac{1}{|\mathcal{K}|}$, and then it is elementary to show that the optimal value for

\[(2.64)\]

satisfies

\[f \geq 1 - \lambda |\mathcal{K}| \sum_{i=1}^{|\mathcal{K}|} \frac{1}{|\mathcal{K}|}. \quad (2.73)\]

Thus we have shown that

\[\frac{1}{2} |P_{K^m} - T_{K^m}| \geq 1 - \frac{1}{|\mathcal{K}|} - |\mathcal{K}| \sum_{i=1}^{|\mathcal{K}|} \sum_{k=1}^{K_m} \mathbb{P}(K_1 = \cdots = K_m = k) \sum_{l=1}^{q_{K,l}} \log |\mathcal{K}| \quad (2.74)\]

as desired.

\[\blacklozenge\]

\textbf{Remark 2.3.2} Theorem 2.3.2 is stronger than a conventional converse by Fano’s inequality (which proves exactly the converse part of (2.40)) in the following senses:

\begin{itemize}
  \item The conventional converse fails to bound the ratio of the number of common randomness bits to the communication bits if the communication rate is zero. In contrast, Theorem 2.3.2 is still gives a tight bound on this ratio in the zero-rate case where the log size of the key alphabet grows sub-linearly in the blocklength. In fact, as long as

\[\log |\mathcal{K}| - \sum_{l=1}^{m} c_l (\log |\mathcal{K}| - \log |\mathcal{W}_l|) \rightarrow -\infty \quad (2.75)\]

which is weaker than (2.47), Theorem 2.3.2 implies that $\frac{1}{2} |P_{K^m} - T_{K^m}|$ converges to 1.

  \item Even if (2.47) holds, the conventional Fano-based converse does not guarantee that the error probability tends to 1.
\end{itemize}
2.4 Dual Formulation of a Forward-Reverse Brascamp-Lieb Inequality

In this section we introduce a new type of functional inequality which may be called the “forward-reverse Brascamp-Lieb inequality”, and prove its equivalent entropic formulation. The forward-reverse Brascamp-Lieb inequality is a generalization of both the forward inequality (which we have discussed in Section 2.2) and the reverse Brascamp-Lieb inequality. The proof of the equivalent formulation of the forward-reverse inequality is more technically demanding because of the “reverse” component. Readers who are only interested in applications to network information theory may skip this section. However, as shown in Section 2.5 the forward-reverse inequality is useful in providing a unified treatment to almost all duality results in the literature, previously proved by various other methods.

To begin with, let us review a common form of the reverse Brascamp-Lieb inequality, again from Barthe’s paper [88]:

**Theorem 2.4.1** Let $E$, $E_1$, $\ldots$, $E_m$ be Euclidean spaces, and $B_i: E \to E_i$ be linear maps. Let $(c_i)_{i=1}^m$ and $F$ be positive real numbers. The reverse Brascamp-Lieb inequality

$$\int \sup_{(y_i)} \prod_{i=1}^m f_i(y_i) \, dx \geq F \prod_{i=1}^m \|f_i\|_{\frac{1}{c_i}},$$

(2.76)

for all nonnegative measurable functions $f_i$ on $E_i$, $i = 1, \ldots, m$, holds if and only if it holds for all centered Gaussian functions.

As in the forward case, in this chapter we are only be interested in the form of the reverse Brascamp-Lieb inequality rather than the Gaussian optimality result.

$^6$Denotes the adjoint of $B_i$. In other words, the matrix of $B_i^*$ is the transpose of the matrix of $B_i$. 

42
In particular, we consider general (possibly nonlinear) mappings between general (possibly non-Euclidean) spaces.

Lehec [103, Theorem 18] essentially proved that (2.76) is implied by an entropic inequality, but did not prove the converse implication (which, as we shall see, is the more challenging direction).

### 2.4.1 Some Mathematical Preliminaries

Before we introduce the new forward-reverse inequality, some mathematical notations and preliminaries are necessary: this is because we will prove the “functional implies entropic” part of the equivalent formulation result, which is much more sophisticated than the case of the forward inequality. Some regularity assumptions on the alphabets and the measures appear to be necessary, and some definitions (e.g. that of a random transformation) are slightly different than the other sections of the thesis in order to make the proof work (although of course these definitions are equivalent for finite alphabets).

Throughout this thesis, when the forward-reverse inequality is considered, we always assume that the alphabets are Polish spaces, and the measures are Borel measures\[7\]. Of course, this covers the cases where the alphabet is Euclidean or discrete (endowed with the Hamming metric, which induces the discrete topology, making every function on the discrete set continuous), among others. Readers interested in finite-alphabets only may refer to the (much simpler) argument in [80] based on the KKT condition. However, the proof here for the general Polish space case better manifests the “function-measure duality” viewpoint emphasized in this thesis.

**Notation 1** Let $\mathcal{X}$ be a topological space.

\[7\]A Polish space is a complete separable metric space. It enjoys several nice properties that we use heavily in this section, including Prokhorov theorem and Riesz-Kakutani theorem (the latter is related to the fact that every Borel probability measure on a Polish space is inner regular, hence a Radon measure). Short introductions to the Polish space can be found in e.g. [104][105].
• \( C_c(X) \) denotes the space of continuous functions on \( X \) with a compact support;

• \( C_0(X) \) denotes the space of all continuous function \( f \) on \( X \) that vanishes at infinity (i.e. for any \( \epsilon > 0 \) there exists a compact set \( K \subseteq X \) such that \( |f(x)| < \epsilon \) for \( x \in X \setminus K \));

• \( C_b(X) \) denotes the space of bounded continuous functions on \( X \);

• \( \mathcal{M}(X) \) denotes the space of finite signed Borel measures on \( X \);

• \( \mathcal{P}(X) \) denotes the space of probability measures on \( X \).

We consider \( C_c, C_0 \) and \( C_b \) as topological vector spaces, with the topology induced from the sup norm. The following theorem, usually attributed to Riesz, Markov and Kakutani, is well-known in functional analysis and can be found in, e.g. [67][106].

**Theorem 2.4.2 (Riesz-Markov-Kakutani)** If \( X \) is a locally compact, \( \sigma \)-compact Polish space, the dual
\( 8 \) of both \( C_c(X) \) and \( C_0(X) \) is \( \mathcal{M}(X) \).

**Remark 2.4.1** The dual space of \( C_b(X) \) can be strictly larger than \( \mathcal{M}(X) \), since it also contains those linear functionals that depend on the “limit at infinity” of a function \( f \in C_b(X) \) (originally defined for those \( f \) that do have a limit at the infinity, and then extended to the whole \( C_b(X) \) by Hahn-Banach theorem; see e.g. [67]).

Of course, any \( \mu \in \mathcal{M}(X) \) is a continuous linear functional on \( C_0(X) \) or \( C_c(X) \), given by

\[
f \mapsto \int f \, d\mu \tag{2.77}
\]

where \( f \) is a function in \( C_0(X) \) or \( C_c(X) \). Remarkably, Theorem 2.4.2 states that the converse is also true under mild regularity assumptions on the space. Thus, we

---

The dual of a topological vector space consists of all continuous linear functionals on that space, which is, naturally, also a topological vector space (with the weak* topology).
can view measures as continuous linear functionals on a certain function space; this justifies the shorthand notation

$$\mu(f) := \int f \, d\mu$$

which we use frequently in the rest of the thesis. This viewpoint is the most natural for our setting since in the proof of the equivalent formulation of the forward-reverse Brascamp-Lieb inequality we shall use the Hahn-Banach theorem to show the existence of certain linear functionals.

**Definition 2.4.1** Let $\Lambda: C_b(\mathcal{X}) \to (-\infty, +\infty]$ be a lower semicontinuous, proper convex function. Its Legendre-Fenchel transform $\Lambda^*: C^*_b(\mathcal{X}) \to (-\infty, +\infty]$ is given by

$$\Lambda^*(\ell) := \sup_{u \in C_b(\mathcal{X})} [\ell(u) - \Lambda(u)].$$

Let $\nu$ be a nonnegative finite Borel measure on a Polish space $\mathcal{X}$. The relative entropy has the following definition: for any $\mu \in \mathcal{M}(\mathcal{X})$,

$$D(\mu \| \nu) := \sup_{f \in C_b(\mathcal{X})} [\mu(f) - \Lambda(f)]$$

where we have defined the convex functional on $C_b(\mathcal{X})$:

$$\Lambda(f) := \log \nu(\exp(f))$$

$$= \log \int \exp(f) \, d\nu.$$
The definition (2.80) agrees with the definition (2.3) when \( \nu \) is a probability measure, by the Donsker-Varadhan formula (c.f. [105, Lemma 6.2.13]). If \( \mu \) is not a probability measure, then \( D(\mu \| \nu) \) as defined in (2.80) is \(+\infty\).

Given a bounded linear operator \( T: C_b(\mathcal{Y}) \rightarrow C_b(\mathcal{X}) \), the dual operator \( T^*: C_b(\mathcal{X})^* \rightarrow C_b(\mathcal{Y})^* \) is defined in terms of

\[
T^* \mu_X: f \in C_b(\mathcal{Y}) \mapsto \mu_X(Tf), \tag{2.83}
\]

for any \( \mu_X \in C_b(\mathcal{X})^* \). Since \( \mathcal{P}(\mathcal{X}) \subseteq \mathcal{M}(\mathcal{X}) \subseteq C_b(\mathcal{X})^* \), we can define a conditional expectation operator as any \( T \) such that \( T^* P \in \mathcal{P}(\mathcal{Y}) \) for any \( P \in \mathcal{P}(\mathcal{X}) \). A random transformation \( T^* \) is defined as the dual of some conditional expectation operator.

**Remark 2.4.2** From the viewpoint of category theory (see for example [107,108]), \( C_b \) is a functor from the category of topological spaces to the category of topological vector spaces, which is contra-variant because for any continuous, \( \phi: \mathcal{X} \rightarrow \mathcal{Y} \) (morphism between topological spaces), we have \( C_b(\phi): C_b(\mathcal{Y}) \rightarrow C_b(\mathcal{X}) \), \( u \mapsto u \circ f \) where \( u \circ \phi \) denotes the composition of two continuous functions, reversing the arrows in the maps (i.e. the morphisms). On the other hand, \( \mathcal{M} \) is a covariant functor and \( \mathcal{M}(\phi): \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{M}(\mathcal{Y}) \), \( \mu \mapsto \mu \circ \phi^{-1} \), where \( \mu \circ \phi^{-1}(\mathcal{B}) := \mu(\phi^{-1}(\mathcal{B})) \) for any Borel measurable \( \mathcal{B} \subseteq \mathcal{Y} \). “Duality” itself is a contra-variant functor between the category of topological spaces (note the reversal of arrows in Fig. 2.1). Moreover, \( C_b(\mathcal{X})^* = \mathcal{M}(\mathcal{X}) \) and \( C_b(\phi)^* = \mathcal{M}(\phi) \) if \( \mathcal{X} \) and \( \mathcal{Y} \) are compact metric spaces and \( \phi: \mathcal{X} \rightarrow \mathcal{Y} \) is continuous. Definition 2.4.2 below can therefore be viewed as the special case where \( \phi \) is the projection map:

**Definition 2.4.2** Suppose \( \phi: \mathcal{Z}_1 \times \mathcal{Z}_2 \rightarrow \mathcal{Z}_1 \), \((z_1, z_2) \mapsto z_1\) is the projection to the first coordinate.

- \( C_b(\phi): C_b(\mathcal{Z}_1) \rightarrow C_b(\mathcal{Z}_1 \times \mathcal{Z}_2) \) is called a canonical map, whose action is almost trivial: it sends a function of \( z_i \) to itself, but viewed as a function of \((z_1, z_2)\).
• $\mathcal{M}(\phi) : \mathcal{M}(\mathcal{Z}_1 \times \mathcal{Z}_2) \to \mathcal{M}(\mathcal{Z}_1)$ is called marginalization, which simply takes a joint distribution to a marginal distribution.

Next, recall that the Fenchel-Rockafellar duality (see [104, Theorem 1.9], or [94] in the case of finite dimensional vector spaces) from convex analysis usually refers to the $k = 1$ special case of the following result, for which we provide a proof here for completeness:

**Theorem 2.4.3** Assume that $A$ is a topological vector space whose dual is $A^*$. Let $\Theta_j : A \to \mathbb{R} \cup \{+\infty\}, j = 0, 1, \ldots, k$, for some positive integer $k$. Suppose there exist some $(u_j)_{j=1}^k$ and $u_0 := -(u_1 + \cdots + u_k)$ such that

$$\Theta_j(u_j) < \infty, \quad j = 0, \ldots, k$$

and $\Theta_0$ is upper semicontinuous at $u_0$. Then

$$-\inf_{\ell \in A^*} \left[ \sum_{j=0}^k \Theta_j^*(\ell) \right] = \inf_{u_1, \ldots, u_k \in A} \left[ \Theta_0 \left( -\sum_{j=1}^k u_j \right) + \sum_{j=1}^k \Theta_j(u_j) \right].$$

For completeness, we provide a proof of this result, which is based on the Hahn-Banach theorem (Theorem 2.4.4) and is similar to the proof of [104, Theorem 1.9].

**Proof** Let $m_0$ be the right side of (2.85). The $\leq$ part of (2.85) follows trivially from the (weak) min-max inequality since

$$m_0 = \inf_{u_0, \ldots, u_k \in A} \sup_{\ell \in A^*} \left\{ \sum_{j=0}^k \Theta_j(u_j) - \ell \left( \sum_{j=0}^k u_j \right) \right\}.$$  

$$\geq \sup_{\ell \in A^*} \inf_{u_0, \ldots, u_k \in A} \left\{ \sum_{j=0}^k \Theta_j(u_j) - \ell \left( \sum_{j=0}^k u_j \right) \right\}$$

$$= -\inf_{\ell \in A^*} \left[ \sum_{j=0}^k \Theta_j^*(\ell) \right].$$  

47
It remains to prove the $\geq$ part, and it suffices to assume without loss of generality that $m_0 > -\infty$. Note that (2.84) also implies that $m_0 < +\infty$. Define convex sets

$$C_j := \{(u, r) \in A \times \mathbb{R} : r > \Theta_j(u)\}, \quad j = 0, \ldots, k;$$

(2.89)

$$D := \{(0, m) \in A \times \mathbb{R} : m \leq m_0\}.$$  

(2.90)

Observe that these are nonempty sets by the assumption (2.84). Also $C_0$ has nonempty interior by the assumption that $\Theta_0$ is upper semicontinuous at $u_0$. Thus, the Minkowski sum

$$C := C_0 + \cdots + C_k$$

(2.91)

is a convex set with a nonempty interior. Moreover, $C \cup D = \emptyset$. By the Hahn-Banach theorem (Theorem 2.4.4), there exists $(\ell, s) \in A^* \times \mathbb{R}$ such that

$$sm \leq \ell(\sum_{j=0}^{k} u_j) + s \sum_{j=0}^{k} r_j.$$  

(2.92)

For any $m \leq m_0$ and $(u_j, r_j) \in C_j$, $j = 0, \ldots, k$. From (2.90) we see (2.92) can only hold when $s \geq 0$. Moreover, from (2.84) and the upper semicontinuity of $\Theta_0$ at $u_0$ we see the $\sum_{j=0}^{k} u_j$ in (2.92) can take value in a neighbourhood of $0 \in A$, hence $s \neq 0$. Thus, by dividing $s$ on both sides of (2.92) and setting $\ell \leftarrow -\ell/s$, we see that

$$m_0 \leq \inf_{u_0, \ldots, u_k \in A} \left[ -\ell(\sum_{j=0}^{k} u_j) + \sum_{j=0}^{k} \Theta_j(u_j) \right]$$

(2.93)

$$= - \left[ \sum_{j=0}^{k} \Theta_j^*(\ell) \right]$$

(2.94)

which establishes the $\geq$ part in (2.85).
Theorem 2.4.4 (Hahn-Banach) Let $C$ and $D$ be convex, nonempty disjoint subsets of a topological vector space $A$. If the interior of $C$ is non-empty, then there exists $\ell \in A^*$, $\ell \neq 0$ such that

$$\sup_{u \in D} \ell(u) \leq \inf_{u \in C} \ell(u).$$

(2.95)

Remark 2.4.3 The assumption in Theorem 2.4.4 that $C$ has nonempty interior is only necessary in the infinite dimensional case. However, even if $A$ in Theorem 2.4.3 is finite dimensional, the assumption in Theorem 2.4.3 that $\Theta_0$ is upper semicontinuous at $u_0$ is still necessary, because this assumption was not only used in applying Hahn-Banach, but also in concluding that $s \neq 0$ in (2.92).

2.4.2 Compact $\mathcal{X}$

We first state a duality theorem for the case of compact alphabets to streamline the proof. Later we show that the argument can be extended to a particular non-compact case.\footnote{Theorem 2.4.5 is not included in the conference paper [80], but was announced in the conference presentation.} Our proof based on the Legendre-Fenchel duality (Theorem 2.4.3) was inspired by the proof of the Kantorovich duality in the theory of optimal transportation (see [104], Chapter 1], where the idea is credited to Brenier).

Theorem 2.4.5 (Dual formulation of forward-reverse Brascamp-Lieb inequality) Assume that

- $m$ and $l$ are positive integers, $d \in \mathbb{R}$, $\mathcal{X}$ is a compact metric space (hence also a Polish space);

- For each $i = 1, \ldots, l$, $b_i \in (0, \infty)$; $\nu_i$ is a finite Borel measure on a Polish space $Z_i$, and $Q_{Z_i|X} = S_i^*$ is a random transformation;
• For each $j = 1, \ldots, m$, $c_j \in (0, \infty)$, $\mu_j$ is a finite Borel measure on a Polish space $Y_j$, and $Q_{Y_j|X} = T_j^*$ is a random transformation.

• For any $P_{Z_i}$ such that $D(P_{Z_i}||\nu_i) < \infty$, $i = 1, \ldots, l$, there exists $P_X \in \cap_i (S_i^*)^{-1} P_{Z_i}$ such that $\sum_{j=1}^m c_j D(P_{Y_j}||\mu_j) < \infty$, where $P_{Y_j} := T_j^* P_X$.

Then the following two statements are equivalent:

1. If nonnegative continuous functions $(g_i)$, $(f_j)$ are bounded away from 0 and such that

\[ \sum_{i=1}^l b_i S_i \log g_i \leq \sum_{j=1}^m c_j T_j \log f_j \]  

(2.96)

then (see (2.78) for the notation of the integral)

\[ \prod_{i=1}^l \nu_i^{b_i}(g_i) \leq \exp(d) \prod_{j=1}^m \mu_j^{c_j}(f_j). \]  

(2.97)

2. For any $(P_{Z_i})$ such that $D(P_{Z_i}||\nu_i) < \infty$ $i = 1, \ldots, l$,

\[ \sum_{i=1}^l b_i D(P_{Z_i}||\nu_i) + d \geq \inf_{P_X} \sum_{j=1}^m c_j D(P_{Y_j}||\mu_j) \]  

(2.98)

where the infimum is over $P_X$ such that $S_i^* P_X = P_{Z_i}$, $i = 1, \ldots, l$, and $P_{Y_j} = T_j^* P_X$, $j = 1, \ldots, m$.

**Proof** We can safely assume $d = 0$ below without loss of generality (since otherwise we can always substitute $\mu_1 \leftarrow \exp\left(\frac{d}{c_1}\right) \mu_1$).

\[ \text{Of course, this assumption is not essential (once we adopt the convention that the infimum in (2.98) is } +\infty \text{ when it runs over an empty set).} \]
1) ⇒ 2) This is the nontrivial direction which relies on certain (strong) min-max type results. In Theorem 2.4.3 put

\[ Θ_0 : u ∈ C_b(\mathcal{X}) → \begin{cases} 
0 & u \leq 0; \\
+∞ & \text{otherwise.}
\end{cases} \quad (2.99) \]

Then,

\[ Θ_0^* : π ∈ M(\mathcal{X}) → \begin{cases} 
0 & π \geq 0; \\
+∞ & \text{otherwise.}
\end{cases} \quad (2.100) \]

For each \( j = 1, \ldots, m \), set

\[ Θ_j(u) = c_j \inf \log \mu_j \left( \exp \left( \frac{1}{c_j} v \right) \right) \quad (2.101) \]

where the infimum is over \( v ∈ C_b(\mathcal{Y}) \) such that \( u = T_j v \); if there is no such \( v \) then \( Θ_j(u) := +∞ \) as a convention. Observe that

- \( Θ_j \) is convex;
\[ \Theta_j(u) > -\infty \text{ for any } u \in C_b(\mathcal{X}). \] If otherwise, for any \( P_X \) and \( P_{Y_j} := T_j^* P_X \) we have

\[
D(P_{Y_j} \| \mu_j) = \sup_{v} \{ P_{Y_j}(v) - \log \mu_j(\exp(v)) \} = \sup_{v} \{ P_X(T_jv) - \log \mu_j(\exp(v)) \} = \sup_{u \in C_b(\mathcal{X})} \left\{ P_X(u) - \frac{1}{c_j} \Theta_j(c_j u) \right\} = +\infty
\]

which contradicts the assumption in the theorem;

- From the steps (2.102)-(2.104), we see \( \Theta_j^*(\pi) = c_j D(T_j^* \pi \| \mu_j) \) for any \( \pi \in \mathcal{M}(\mathcal{X}) \), where the definition of \( D(\cdot \| \mu_j) \) is extended using the Donsker-Varadhan formula (that is, it is infinite when the argument is not a probability measure).

Finally, for the given \( (P_{Z_i})_{i=1}^l \), choose

\[
\Theta_{m+1} : u \in C_b(\mathcal{X}) \mapsto \begin{cases} 
\sum_{i=1}^l P_{Z_i}(w_i) & \text{if } u = \sum_{i=1}^l S_i w_i \text{ for some } w_i \in C_b(\mathcal{Z}_i); \\
+\infty & \text{otherwise.}
\end{cases}
\]

(2.106)

Notice that

- \( \Theta_{m+1} \) is convex;
• $\Theta_{m+1}$ is well-defined (that is, the choice of $(w_i)$ in (2.106) is inconsequential). Indeed if $(w_i)_{i=1}^{l}$ is such that $\sum_{i=1}^{l} S_i w_i = 0$, then

$$
\sum_{i=1}^{l} P_{Z_i}(w_i) = \sum_{i=1}^{l} S_i^* P_X(w_i) 
= \sum_{i=1}^{l} P_X(S_i w_i) 
= 0,
$$

(2.107)

(2.108)

(2.109)

where $P_X$ is such that $S_i^* P_X = P_{Z_i}$, $i = 1, \ldots, l$, whose existence is guaranteed by the assumption of the theorem. This also shows that $\Theta_{m+1} > -\infty$.

•

$$
\Theta_{m+1}^*(\pi) := \sup_u \{\pi(u) - \Theta_{m+1}(u)\}
$$

(2.110)

$$
= \sup_{w_1, \ldots, w_l} \left\{ \pi \left( \sum_{i=1}^{l} S_i w_i \right) - \sum_{i=1}^{l} P_{Z_i}(w_i) \right\}
= \sup_{w_1, \ldots, w_l} \left\{ \sum_{i=1}^{l} S_i^* \pi(w_i) - \sum_{i=1}^{l} P_{Z_i}(w_i) \right\}
= \left\{ \begin{array}{ll}
0 & \text{if } S_i^* \pi = P_{Z_i}, \ i = 1, \ldots, l; \\
+\infty & \text{otherwise.} 
\end{array} \right.
$$

(2.111)

(2.112)

(2.113)

Invoking Theorem 2.4.3 (where the $u_j$ in Theorem 2.4.3 can be chosen as the constant function $u_j \equiv 1$, $j = 1, \ldots, m + 1$):

$$
\inf_{\pi: \pi \geq 0, S_i^* \pi = P_{Z_i}} \sum_{j=1}^{m} c_j D(T_j^* \pi \| \mu_j)
$$

(2.114)

$$
= -\inf_{v^m, w^l: \sum_{j=1}^{m} T_j v_j + \sum_{i=1}^{l} S_i w_i \geq 0} \left[ \sum_{j=1}^{m} c_j \log \mu_j \left( \exp \left( \frac{1}{c_j} v_j \right) \right) + \sum_{i=1}^{l} P_{Z_i}(w_i) \right]
$$

(2.115)
Note that the left side of (2.115) is exactly the right side of (2.98). For any \( \epsilon > 0 \), choose \( v_j \in C_b(Y_j), j = 1, \ldots, m \) and \( w_i \in C_b(Z_i), i = 1, \ldots, l \) such that \( \sum_{j=1}^{m} T_j v_j + \sum_{i=1}^{l} S_i w_i \geq 0 \) and

\[
\epsilon - \sum_{j=1}^{m} c_j \log \mu_j \left( \exp \left( \frac{1}{c_j} v_j \right) \right) - \sum_{i=1}^{l} P_{Z_i} (w_i) > \inf_{\pi : \pi \geq 0, S^\pi = P_{Z_i}} \sum_{j=1}^{m} c_j D(T_j^\pi \| \mu_j)
\]

(2.116)

Now invoking (2.97) with \( f_j := \exp \left( \frac{1}{c_j} v_j \right), j = 1, \ldots, m \) and \( g_i := \exp \left( -\frac{1}{b_i} w_i \right), i = 1, \ldots, l \), we upper bound the left side of (2.116) by

\[
\epsilon - \sum_{i=1}^{l} b_i \log \nu_i (g_i) + \sum_{i=1}^{l} b_i P_{Z_i} (\log g_i) \leq \epsilon + \sum_{i=1}^{l} b_i D(P_{Z_i} \| \nu_i)
\]

(2.117)

where the last step follows by the Donsker-Varadhan formula. Therefore (2.98) is established since \( \epsilon > 0 \) is arbitrary.

2) \( \Rightarrow \) 1) Since \( \nu_i \) is finite and \( g_i \) is bounded by assumption, we have \( \nu_i (g_i) < \infty, i = 1, \ldots, l \). Moreover (2.97) is trivially true when \( \nu_i (g_i) = 0 \) for some \( i \), so we will assume below that \( \nu_i (g_i) \in (0, \infty) \) for each \( i \). Define \( P_{Z_i} \) by

\[
\frac{dP_{Z_i}}{d\nu_i} = \frac{g_i}{\nu_i (g_i)}, \quad i = 1, \ldots, l.
\]

(2.118)

Then for any \( \epsilon > 0 \),

\[
\sum_{i=1}^{l} b_i \log \nu_i (g_i) = \sum_{i=1}^{l} b_i [P_{Z_i} (\log g_i) - D(P_{Z_i} \| \nu_i)]
\]

(2.119)

\[
< \sum_{j=1}^{m} c_j P_{Y_j} (\log f_j) + \epsilon - \sum_{j=1}^{m} c_j D(P_{Y_j} \| \mu_j)
\]

(2.120)

\[
\leq \epsilon + \sum_{j=1}^{m} c_j \log \mu_j (f_j)
\]

(2.121)
where

- (2.120) uses the Donsker-Varadhan formula, and we have chosen $P_X, P_{Y_j} := T_j^* P_X, j = 1, \ldots, m$ such that

$$
\sum_{i=1}^l b_i D(P_{Z_i} \| \nu_i) > \sum_{j=1}^m c_j D(P_{Y_j} \| \mu_j) - \epsilon
$$

(2.122)

- (2.121) also follows from the Donsker-Varadhan formula.

The result follows since $\epsilon > 0$ can be arbitrary.

Remark 2.4.4 The infimum in (2.98) is in fact achievable: For any $(P_{Z_i})$, there exists a $P_X$ that minimizes $\sum_{j=1}^m c_j D(P_{Y_j} \| \mu_j)$ subject to the constraints $S_i^* P_X = P_{Z_i}, i = 1, \ldots, m$, where $P_{Y_j} := T_j^* P_X, j = 1, \ldots, m$. Indeed, since the singleton $\{P_{Z_i}\}$ is weak*-closed and $S_i^*$ is weak*-continuous\(^{12}\) the set $\bigcap_{i=1}^l (S_i^*)^{-1} P_{Z_i}$ is weak*-closed in $\mathcal{M}(X)$; hence its intersection with $\mathcal{P}(\mathcal{X})$ is weak*-compact in $\mathcal{P}(\mathcal{X})$, because $\mathcal{P}(\mathcal{X})$ is weak*-compact by (a simple version for the setting of a compact underlying space $\mathcal{X}$ of) the Prokhorov theorem \([109]\). Moreover, by the weak*-lower semicontinuity of $D(\cdot \| \mu_j)$ (easily seen from the variational formula/Donsker-Varadhan formula of the relative entropy, cf. \([43]\)) and the weak*-continuity of $T_j^*; j = 1, \ldots, m$, we see $\sum_{j=1}^m c_j D(T_j^* P_X \| \mu_j)$ is weak*-lower semicontinuous in $P_X$, and hence the existence of a minimizing $P_X$ is established.

Remark 2.4.5 Abusing the terminology from the min-max theory, Theorem 2.4.5 may be interpreted as a “strong duality” result which establishes the equivalence of

---

\(^{12}\) Generally, if $T: A \to B$ is a continuous map between two topologically vector spaces, then $T^*: B^* \to A^*$ is a weak* continuous map between the dual spaces. Indeed, if $y_n \to y$ is a weak*-convergent subsequence in $B^*$, meaning $y_n(b) \to y(b)$ for any $b \in B$, then we must have $T^* y_n(a) = y_n(Ta) \to y(Ta) = T^* y(a)$ for any $a \in A$, meaning that $T^* y_n$ converges to $T^* y$ in the weak* topology.
two optimization problems. The 1)⇒2) part is the non-trivial direction which requires regularity on the spaces. In contrast, the 2)⇒1) direction can be thought of as a “weak duality” which establishes only a partial relation but holds for more general spaces.

**Remark 2.4.6** The equivalent formulations of the forward Brascamp-Lieb inequality (Theorem 2.2.3) can be recovered from Theorem 2.4.5 by taking \( l = 1, \ Z_1 = \mathcal{X} \), and letting \( S_1 \) be the identity map/isomorphism, except that Theorem 2.2.3 is established for completely general alphabets. In other words, the forward Brascamp-Lieb inequality is the special case of the forward-reverse Brascamp-Lieb inequality when there is only one reverse channel which is the identity.

### 2.4.3 Noncompact \( \mathcal{X} \)

Our proof of 1)⇒2) in Theorem 2.4.5 makes use of the Hahn-Banach theorem, and hence relies crucially on the fact that the measure space is the dual of the function space. Naively, one might want to extend the the proof to the case of *locally compact* \( \mathcal{X} \) by considering \( C_0(\mathcal{X}) \) instead of \( C_b(\mathcal{X}) \), so that the dual space is still \( M(\mathcal{X}) \).

However, this would not work: consider the case when \( \mathcal{X} = Z_1 \times \ldots \times Z_l \) and each \( S_i \) is the canonical map. Then \( \Theta_m(\mathcal{X}) \) as defined in (2.106) is \( +\infty \) unless \( u \equiv 0 \), thus \( \Theta_m^* \equiv 0 \). Luckily, we can still work with \( C_b(\mathcal{X}) \); in this case \( \ell \in C_b(\mathcal{X})^* \) may not be a measure, but we can decompose it into \( \ell = \pi + R \) where \( \pi \in M(\mathcal{X}) \) and \( R \) is a linear functional “supported at the infinity”. Below we use the techniques in [104, Chapter 1.3] to prove a particular extension of Theorem 2.4.5 to a non-compact case.

**Theorem 2.4.6** Theorem 2.4.5 still holds if

- The assumption that \( \mathcal{X} \) is a compact metric space is relaxed to the assumption that it is a locally compact and \( \sigma \)-compact Polish space;
\[ X = \prod_{i=1}^{l} Z_i \text{ and } S_i: C_b(Z_i) \to C_b(X), \ i = 1, \ldots, l \text{ are canonical maps (see Definition 2.4.2).} \]

**Proof**  The proof of the “weak duality” part 2) \( \implies 1) \) still works in the noncompact case, so we only need to explain what changes need to be made in the proof of 1) \( \implies 2) \) part. Let \( \Theta_0 \) be defined as before, in (2.99). Then for any \( \ell \in C_b(X)^* \),

\[
\Theta_0^*(\ell) = \sup_{u \leq 0} \ell(u) \tag{2.123}
\]

which is 0 if \( \ell \) is nonnegative (in the sense that \( \ell(u) \geq 0 \) for every \( u \geq 0 \)), and \(+\infty\) otherwise. This means that when computing the infimum on the left side of (2.85), we only need to take into account of those nonnegative \( \ell \).

Next, let \( \Theta_{m+1} \) be also defined as before. Then directly from the definition we have

\[
\Theta_{m+1}^*(\ell) = \begin{cases} 
0 & \text{if } \ell(\sum_i S_i w_i) = \sum_i P_{Z_i}(w_i), \ \forall w_i \in C_b(Z_i), \ i = 1, \ldots l; \\
+\infty & \text{otherwise.}
\end{cases} \tag{2.124}
\]

For any \( \ell \in C_b^*(X) \). Generally, the condition in the first line of (2.124) does not imply that \( \ell \) is a measure. However, if \( \ell \) is also nonnegative, then using a technical result in [104, Lemma 1.25] we can further simplify:

\[
\Theta_{m+1}^*(\ell) = \begin{cases} 
0 & \text{if } \ell \in M(X) \text{ and } S_i^* \ell = P_{Z_i}, \ i = 1, \ldots, l; \\
+\infty & \text{otherwise.}
\end{cases} \tag{2.125}
\]

This further shows that when we compute the left side of (2.85) the infimum can be taken over \( \ell \) which is a coupling of \( (P_{Z_i}) \). In particular, if \( \ell \) is a probability measure, then \( \Theta_j^*(\ell) = c_j D(T_j^* \ell \| \mu_j) \) still holds with the \( \Theta_j \) defined in (2.101), \( j = 1, \ldots, m \). Thus the rest of the proof can proceed as before. \[\qed\]
**Remark 2.4.7** The second assumption is made in order to achieve (2.125) in the proof.

**Remark 2.4.8** In [80] we studied a version of “reverse Brascamp-Lieb inequality” which is a special case of Theorem 2.4.6 when there is only one forward channel: in the setting of Theorem 2.4.5 consider $m = 1$, $c_1 = 1$, $X = Z_1 \times \ldots \times Z_1$. Let $S_i$ be the canonical map, $g_i \leftarrow g_{i_1}^{Z_i}$, $i = 1, \ldots, l$. Then (2.97) becomes

$$
\prod_{i=1}^{l} \| g_i \|_{\frac{1}{Z_i}} \leq \exp(d) \mu_1(f_1)
$$

(2.126)

for any nonnegative continuous $(g_i)_{i=1}^{l}$ and $f_1$ bounded away from 0 and $+\infty$ such that

$$
\sum_{i=1}^{l} \log g_i(z_i) \leq \mathbb{E}[\log f_1(Y_1)|Z^l = z^l], \quad \forall z^l.
$$

(2.127)

Note that (2.127) can be simplified in the deterministic special case: let $\phi: X \to Y_1$ be any continuous function, and $T_1 \leftarrow C_b(\phi)$ (that is, $T_1$ sends a function $f$ on $Y_1$ to the function $f \circ \phi$ on $X$). Then (2.127) becomes

$$
\prod_{i=1}^{l} g_i(z_i) \leq f_1(\phi(z_1, \ldots, z_l)), \quad \forall z_1, \ldots, z_l.
$$

(2.128)

Then the optimal choice of $f_1$ admits an explicit formula, since for any given $(g_i)$, to verify (2.126) we only need to consider

$$
f_1(y) := \sup_{\phi(z_1, \ldots, z_l) = y} \left\{ \prod_{i} g_i(z_i) \right\}, \quad \forall y.
$$

(2.129)

Thus when $\phi$ is a linear function, (2.126) is essentially Barthe’s formulation of reverse BL (2.76) (the exception being that Theorem 2.4.6, in contrast to (2.76), restricts attention to finite $\nu_i$, $i = 1, \ldots, l$ and $\mu_1$). The more straightforward part of the duality
(entropic inequality⇒functional inequality) has essentially been proved by Lehec [103, Theorem 18] in a special setting.

2.4.4 Extension to General Convex Functionals

For certain applications (e.g. the transportation-cost inequalities, see Section 2.5.7 ahead), we may be interested in convex functionals beyond the relative entropy. Recall the definition of the Legendre-Fenchel transform in Definition 2.4.1. Then we have, from the theory of convex analysis (see for example [105, Lemma 4.5.8]),

\[
\Lambda(u) = \sup_{\ell \in C_b(X)^*} [\ell(u) - \Lambda^*(\ell)],
\]

for any \( u \in C_b(X) \). Moreover, if \( \Lambda^*(\ell) = +\infty \) for any \( \ell \notin \mathcal{P}(X) \), then from (2.130) we must also have

\[
\Lambda(u) = \sup_{\ell \in \mathcal{P}(X)} [\ell(u) - \Lambda^*(\ell)].
\]

For example, the function \( \Lambda \) defined in (2.82) satisfies the property in (2.131). We need this property in the proof of Theorem 2.4.5 because of step (2.119). From the proof of Theorem 2.4.5 we see that we can obtain the following generalization to convex functionals with no additional cost. An application of this generalization to transportation-cost inequalities is given in Section 2.5.7.

**Theorem 2.4.7** Assume that

- \( m \) and \( l \) are positive integers, \( d \in \mathbb{R} \), \( X \) is a compact metric space (hence also a Polish space);

- For each \( i = 1, \ldots, l \), \( Z_i \) is a Polish space, \( \Lambda_i: C_b(Z_i) \to \mathbb{R} \cup \{+\infty\} \) is proper convex such that \( \Lambda^*_i(\ell) = +\infty \) for \( \ell \notin \mathcal{P}(Z_i) \), and \( S_i: C_b(Z_i) \to C_b(X) \) is a conditional expectation operator;
• For each $j = 1, \ldots, m$, $\mathcal{Y}_j$ is a Polish space, $\Theta_j : C_b(\mathcal{Y}_j) \to \mathbb{R} \cup \{+\infty\}$ is proper convex such that $\Theta_j(u) < \infty$ for some $u \in C_b(\mathcal{Y}_j)$ which is bounded below, and $T_j : C_b(\mathcal{Y}_j) \to C_b(\mathcal{X})$ is a conditional expectation operator;

• For any $\ell_{Z_i} \in \mathcal{M}(\mathcal{Z}_i)$ such that $\Lambda^*_i(\ell_{Z_i}) < \infty$, $i = 1, \ldots, l$, there exists $\ell_X \in \bigcap_i (S^*_i)^{-1} \ell_{Z_i}$ such that $\sum_{j=1}^m \Theta^*_j(\ell_{Y_j}) < \infty$, where $\ell_{Y_j} := T^*_j \ell_X$.

Then the following two statements are equivalent:

1. If $g_i \in C_b(\mathcal{Z}_i)$, $f_j \in C_b(\mathcal{Y}_j)$, $i = 1, \ldots, l$, $j = 1, \ldots, m$ satisfy
   \[
   \sum_{i=1}^l S_i g_i \leq \sum_{j=1}^m T_j f_j \tag{2.132}
   \]
   then
   \[
   \sum_{i=1}^l \Lambda_i(g_i) \leq \sum_{j=1}^m \Theta_j(f_j). \tag{2.133}
   \]

2. For any\footnote{Since by assumption $\Lambda^*_i(\ell_{Z_i}) = +\infty$ when $\ell \not\in \mathcal{P}(\mathcal{Z}_i)$, in which case (2.134) is trivially true, it is equivalent to assuming here that $\ell \in \mathcal{P}(\mathcal{Z}_i)$.} $\ell_{Z_i} \in \mathcal{M}(\mathcal{Z}_i)$, $i = 1, \ldots, l$,
   \[
   \sum_{i=1}^l \Lambda^*_i(\ell_{Z_i}) \geq \inf_{\ell_X} \sum_{j=1}^m \Theta^*_j(\ell_{Y_j}) \tag{2.134}
   \]
   where the infimum is over $\ell_X$ such that $S^*_i \ell_X = P_{Z_i}$, $i = 1, \ldots, l$, and $\ell_{Y_j} := T^*_j \ell_X$, $j = 1, \ldots, m$.

**Remark 2.4.9** Just like Theorem 2.4.6, it is possible to extend Theorem 2.4.7 to the case of noncompact $\mathcal{X}$, provided that $\mathcal{X} = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_l$ and $S_i$, $i = 1, \ldots, l$ are canonical maps.
2.5 Some Special Cases of the Forward-Reverse Brascamp-Lieb Inequality

In this section we discuss some notable special cases of the duality results for the forward-reverse Brascamp-Lieb Inequality (Theorems 2.2.3-2.4.7). The relationship among the involved inequalities is illustrated in Figure 2.2. The equivalent formulations of Rényi divergence, the strong data processing inequality, hypercontractivity and its reverse (with positive or negative parameters), Loomis-Whitney inequality/Shearer’s lemma, and the transportation-cost inequalities have been proved by different methods (see for example [110][7][111][112][74][79][113]). In some of these examples (e.g. strong data processing [7]) the previous approaches rely heavily on the finiteness of the alphabet, whereas the present approach (essentially based on the nonnegativity of the relative entropy) is simpler and holds for general alphabets.

Due in part to their utility in establishing impossibility bounds, these functional inequalities have attracted a lot of attention in information theory [114][115][116][117][118][101][73][119], theoretical computer science [71][120][121][122][86], and statistics [123][124][125][72][126][127], to name only a small subset of the literature.
2.5.1 Variational Formula of Rényi Divergence

As the first example, we show how (2.14) recovers the variational formula of Rényi divergence \[128\] \[129\] in a special case. A prototype of the variational formula of Rényi divergence appeared in the context of control theory \[128\] as a technical lemma. Its utility in information theory was then noticed by \[129\] \[110\], which further developed the result and elaborated on its applications in other areas of probability theory.

Suppose \( R \) and \( Q \) are nonnegative measures on \( (\mathcal{X}, \mathcal{F}) \), \( \alpha \in (0, 1) \cup (1, \infty) \), and \( g: \mathcal{X} \to \mathbb{R} \) is a bounded measurable function. Also, let \( T \) be a probability measure such that \( R, Q \ll T \). Define the Rényi divergence

\[
D_\alpha(Q \| R) := \frac{1}{\alpha - 1} \log \left( \mathbb{E} \left[ \exp \left( \alpha \mathbbm{1}_{Q \ll T}(\hat{X}) + (1 - \alpha) \mathbbm{1}_{R \ll T}(\hat{X}) \right) \right] \right)
\]

(2.135)

where \( \hat{X} \sim T \), which is independent of the particular choice of the reference measure \( T \) \[130\]. Then the variational formula of Rényi divergence \[129\] Remark 2.2] can be equivalently stated as the functional inequality\[14\]

\[
\frac{1}{\alpha - 1} \log \mathbb{E}[\exp((\alpha - 1)g(\hat{X}))] - \frac{1}{\alpha} \log \mathbb{E}[\exp(\alpha g(X))] \leq \frac{1}{\alpha} D_\alpha(Q \| R)
\]

(2.136)

where \( X \sim R \) and \( \hat{X} \sim Q \), with equality achieved when

\[
\frac{dQ}{dR}(x) = \frac{\exp(g(x))}{\mathbb{E}[\exp(g(X))]}.
\]

(2.137)

The well-known variational formula of the relative entropy (see e.g. \[43\]) can be recovered by taking \( \alpha \to 1 \). In the \( \alpha \in (1, \infty) \) case, we can choose \( \exp(g) \) to be the indicator function of an arbitrary measurable set \( \mathcal{A} \subseteq \mathcal{X} \), to obtain the logarithmic

\[\text{Note that our definition of Rényi divergence is different from \[129\] by a factor of } \alpha.\]
probability comparison bound (LPCB)\cite{129}\footnote{15}{129} focuses on the case of probability measure, but (2.138) continues to hold if $Q$ and $R$ are replaced by any unnormalized nonnegative measures.

\[
\frac{1}{\alpha - 1} \log Q(A) - \frac{1}{\alpha} \log R(A) \leq \frac{1}{\alpha} D_{\alpha}(Q\|R). \tag{2.138}
\]

Now we give a new proof of the functional inequality (2.136) using a well-known entropic inequality in information theory. First consider $\alpha \in (1, \infty)$. In Theorem 2.2.3 set $m \leftarrow 1$, $Y_1 = X$, $c \leftarrow \frac{a-1}{\alpha}$, $P_{Y_1|X} = \text{id}$ (the identity mapping). We may assume without loss of generality that $Q \ll R$, since otherwise $D_{\alpha}(Q\|R) = \infty$ and (2.136) always holds. Thus, setting the cost function as

\[
d(x) \leftarrow -\nu_{Q|R}(x) + \frac{\alpha - 1}{\alpha} D_{\alpha}(Q\|R) \tag{2.139}
\]

we see that (2.15) is reduced to

\[
D(P\|R) + \mathbb{E} \left[ -\nu_{Q|R}(\hat{X}) \right] + \frac{\alpha - 1}{\alpha} D_{\alpha}(Q\|R) \geq \frac{\alpha - 1}{\alpha} D(P\|R) \tag{2.140}
\]

which, by our convention in Remark 2.2.2, can be simplified to

\[
D(P\|Q) + \frac{\alpha - 1}{\alpha} D_{\alpha}(Q\|R) \geq \frac{\alpha - 1}{\alpha} D(P\|R). \tag{2.141}
\]

It is a well-known result that (2.141) holds for all $P$ absolutely continuous with respect to $Q$ and $R$ (see for example \cite{130} Theorem 30, \cite{131} Theorem 1, \cite{132} Corollary 2), due to its relation to the fundamental problem of characterizing the error exponents in binary hypothesis testing. By Theorem 2.2.3 with $m = 1$ we translate (2.140) into the functional inequality:

\[
\mathbb{E} \left[ \exp \left( \log f(\hat{X}) + \nu_{Q|R}(\hat{X}) - \frac{\alpha - 1}{\alpha} D_{\alpha}(Q\|R) \right) \right] \leq \left( \mathbb{E} \left[ f^{\frac{\alpha}{\alpha - 1}}(X) \right] \right)^{\frac{\alpha - 1}{\alpha}} \tag{2.142}
\]
for all nonnegative measurable $f$. Finally, by taking the logarithms and dividing by $\alpha - 1$ on both sides and setting $f \leftarrow \exp((\alpha - 1)g)$, the functional inequality (2.136) is recovered.

Note that the choice of $d(\cdot)$ in (2.139) is simply for the purpose of change-of-measure, since the two relative entropy terms in (2.141) have different reference measures $Q$ and $R$. Thus, an alternative proof is to take $d(\cdot) = 0$ but invoke the extension of Theorem 2.2.3 in Remark 2.2.3 with $\mu \leftarrow Q$ and $\nu \leftarrow R$.

The case of $\alpha \in (0, 1)$ can be proved in a similar fashion using Theorem 2.4.5 with $m \leftarrow 2$, $l \leftarrow 1$, $X = \mathcal{Y}_1 = \mathcal{Y}_2 \leftarrow \mathcal{X}$, $Q_{\mathcal{Y}_1|\mathcal{X}} = Q_{\mathcal{Y}_2|\mathcal{X}} = \text{id}$, $Q_{\mathcal{Y}_1} \leftarrow Q$, $Q_{\mathcal{Y}_2} \leftarrow R$, and $|\mathcal{Z}| = 1$; we omit the details here.

Note that the original proofs of the functional inequality (2.136) in [128][129] were based on Hölder’s inequality, whereas the present proof relies on the duality between functional inequalities and entropic inequalities, and the property (2.141) (which amounts to the nonnegativity of relative entropy). Moreover, the weaker probability version (2.138) can be easily proved by a data processing argument; see for example [12][Section II.B][133].

### 2.5.2 Strong Data Processing Constant

The strong data processing inequality (SDPI) [7][8][9] has received considerable interest recently. It has proved fruitful in providing impossibility bounds in various problems; see [117] for a recent list of its applications. It generally refers to an inequality of the form

$$D(P_X\|Q_X) \geq cD(P_Y\|Q_Y), \text{ for all } P_X \ll Q_X$$

(2.143)

where $P_X \rightarrow Q_{\mathcal{Y}|\mathcal{X}} \rightarrow P_Y$, and we have fixed $Q_{\mathcal{X}\mathcal{Y}} = Q_{\mathcal{X}}Q_{\mathcal{Y}|\mathcal{X}}$. The conventional data processing inequality corresponds to the case of $c = 1$. The study of the best
(largest) constant $c$ for (2.143) to hold can be traced to Ahlswede and Gács [7], who showed, among other things, its equivalence to the functional inequality

$$\mathbb{E}[\exp(\mathbb{E}[\log f(Y) | X])] \leq \|f\|_c^c \text{ for all nonnegative } f.$$  (2.144)

The strong data processing inequality can be viewed as a special case of the forward-reverse Brascamp-Lieb inequality where there is only one forward and one identity reverse channel. In other words, the equivalence between (2.143) and (2.144) can be readily seen from either Theorem 2.2.3 or Remark 2.4.8. Its original proof of such an equivalence [7, Theorem 5], on the other hand, relies on a limiting property of hypercontractivity, which, in turn, relies heavily on the finiteness of the alphabet and the proof is quite technical even in that case.

As we saw in Section 2.5.1, a functional inequality often implies an inequality of the probabilities of sets when specialized to the indicator functions. In the case of (2.144), however, a more rational choice is

$$f(y) = (1_A(y) + Q_Y[A]1_A(y))^c$$  (2.145)

where $A$ is an arbitrary measurable subset of $Y$ and $\bar{A} := Y \backslash A$. Then using (2.144),

$$Q_Y^c[A]Q_X[x: Q_{Y|X=x}[A] \geq 1 - \epsilon] = Q_Y^c[A]Q_X[x: Q_{Y|X=x}[\bar{A}] \leq \epsilon]$$

$$\leq \int \exp \left( \log Q_Y[A]^c \cdot Q_{Y|X=x}[\bar{A}] \right) dQ_X(x)$$

$$= \mathbb{E}[\exp(\mathbb{E}[\log f(Y) | X])]]$$

$$\leq \mathbb{E}^c[f^\frac{1}{c}(Y)]$$

$$\leq 2^c Q_Y^c[A].$$

65
Rearranging, we obtain the following bound on conditional probabilities:

\[
Q_X \left[ x : Q_{Y|x=x}[A] \geq 1 - \epsilon \right] \leq 2^\epsilon Q_Y^{(1-\epsilon)}[A]
\] (2.151)

which, by a blowing-lemma argument (cf. [27]), would imply the asymptotic result of [27] Theorem 1, a useful tool in establishing strong converses in source coding problems. Note that [27, Section 2] proved a result essentially the same as (2.151) by working on (2.143) rather than (2.144). \(^{16}\)

2.5.3 Loomis-Whitney Inequality and Shearer’s Lemma

The duality between Loomis-Whitney Inequality and Shearer’s Lemma is yet another special case of Theorem 2.2.3. This is already contained in the duality theorem of Carlen and Cordero-Erausquin [76], but we briefly discuss it here.

The combinatorial Loomis-Whitney inequality [134, Theorem 2] says that if \( A \) is a subset of \( A^m \), where \( A \) is a finite or countably infinite set, then

\[
|A| \leq \prod_{j=1}^m |\pi_j(A)|^{\frac{1}{m-1}}
\] (2.152)

where we defined the projection

\[
\pi_j : A^m \to A^{m-1},
\]

\[
(x_1, \ldots, x_m) \mapsto (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m).
\] (2.153)

\(^{16}\) Another difference is that [27, Theorem 1] involves an auxiliary r.v. \( U \) with \( |U| \leq 3 \), where the cardinality bound comes from convexifying a subset in \( \mathbb{R}^2 \). Here (2.151) holds if (2.143), which is slightly simpler involving only relative entropy terms, because we are essentially working with the supporting lines of the convex hull, and the supporting line of a set is the same as the supporting line of its convex hull.

66
for each $j \in \{1, \ldots, m\}$. The combinatorial inequality (2.152) can be recovered from the following integral inequality: let $\mu$ be the counting measure on $A^m$, then

$$
\int_{A^m} \prod_{j=1}^m f_j(\pi_j(x)) d\mu(x) \leq \prod_{j=1}^m \|f_j\|_{m-1}
$$

(2.155)

for all nonnegative $f_j$'s, where the norm on the right side is with respect to the counting measure on $A^{m-1}$. We claim that (2.155) is an inequality of the form (2.14). To see how (2.155) recovers (2.152), let $f_j$ be the indicator function of $\pi_j(A)$ for each $j$. Then the left side of (2.155) upper-bounds the left side of (2.152), while the right side of (2.155) is equal to the right side of (2.152). Now we invoke Remark 2.2.3 with $X \leftarrow A^m$, $\nu$ and $\mu_j$ being the counting measure on $A^m$ and $A^{m-1}$, respectively, $Q_{Y_j|X}$ being the projection mappings in (2.153)-(2.154), and let $(X_1, \ldots, X_m)$ be distributed according to a given $P_X$, to obtain

$$
-H(X_1, \ldots, X_m) = D(P_X \| \nu)
$$

(2.156)

$$
\geq \sum_{j=1}^m \frac{1}{m-1} D(P_{X_j} \| \mu_j)
$$

(2.157)

$$
\geq -\sum_{j=1}^m \frac{1}{m-1} H(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_m)
$$

(2.158)

where $H(\cdot)$ is the Shannon entropy. This is a special case of Shearer’s Lemma [135] [112].

Similarly, the continuous Loomis-Whitney inequality for Lebesgue measure, that is,

$$
\int_{\mathbb{R}^m} \prod_{j=1}^m f_j(\pi_j(x^n)) dx^n \leq \prod_{j=1}^m \|f_j\|_{m-1}
$$

(2.159)
is the dual of a continuous version of Shearer’s lemma involving differential entropies:

\[ h(X^m) \leq \sum_{j=1}^{m} \frac{1}{m-1} h(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_m). \quad (2.160) \]

### 2.5.4 Hypercontractivity

\[ \mathcal{P}(Z_1) \overset{\cong}{\leftarrow} \mathcal{P}(Y_1 \times Y_2) \xrightarrow{T_1^*} \mathcal{P}(Y_1) \]
\[ \xrightarrow{T_2^*} \mathcal{P}(Y_2) \]

Figure 2.3: Diagram for hypercontractivity (HC)

In Theorem 2.4.5 take \( l \leftarrow 1, m \leftarrow 2, b_1 \leftarrow 1, d \leftarrow 0, f_1 \leftarrow f_1^{\frac{1}{c_1}}, f_2 \leftarrow f_2^{\frac{1}{c_2}}, \nu_1 \leftarrow Q_{Y_1Y_2}, \mu_1 \leftarrow Q_{Y_1}, \mu_2 \leftarrow Q_{Y_2}. \) Also, put \( Z_1 = X = (Y_1, Y_2), \) and let \( T_1 \) and \( T_2 \) be the canonical maps (Definition 2.4.2). We obtain the equivalence between\(^\dagger\)

\[ \| f_1 \|_{c_1} \| f_2 \|_{c_2} \geq \mathbb{E}[f_1(Y_1)f_2(Y_2)], \quad \forall f_1 \in L^{\frac{1}{c_1}}(Q_{Y_1}), f_2 \in L^{\frac{1}{c_2}}(Q_{Y_2}) \quad (2.161) \]

and

\[ \forall P_{Y_1Y_2}, \quad D(P_{Y_1Y_2} \| Q_{Y_1Y_2}) \geq c_1D(P_{Y_1} \| Q_{Y_1}) + c_2D(P_{Y_2} \| Q_{Y_2}). \quad (2.162) \]

This equivalence can also be obtained from Theorem 2.2.3. By Hölder’s inequality, \((2.161)\) is equivalent to saying that the norm of the linear operator sending \( f_1 \in L^{\frac{1}{c_1}}(Q_{Y_1}) \) to \( \mathbb{E}[f_1(Y_1)|Y_2 = \cdot] \in L^{\frac{1}{1-c_2}}(Q_{Y_2}) \) does not exceed 1. The interesting case is \( \frac{1}{1-c_2} > \frac{1}{c_1} \), hence the name hypercontractivity. The equivalent formulation of hypercontractivity was shown in [74] using a different proof via the method of types/typicality, which relies on the finite nature of the alphabet. In contrast, the proof based on the nonnegativity of relative entropy removes this constraint, allow-

\(^\dagger\)By a standard dense-subspace argument, we see that it is inconsequential that \( f_1 \) and \( f_2 \) in \((2.161)\) are not assumed to be continuous.
2.5.5 Reverse Hypercontractivity (Positive Parameters)

\[ \mathcal{P}(Z_1) \xrightarrow{S_1^*} \mathcal{P}(Z_1 \times Z_2) \xrightarrow{\sim} \mathcal{P}(Y) \xleftarrow{S_2^*} \mathcal{P}(Z_2) \]

Figure 2.4: Diagram for reverse HC

In Theorem 2.4.5, take \( l \leftarrow 2, m \leftarrow 1, c_1 \leftarrow 1, d \leftarrow 0, g_1 \leftarrow \frac{1}{\mu_1}, g_2 \leftarrow \frac{1}{\mu_2}, \mu_1 \leftarrow Q_{Z_1 Z_2}, \nu_1 \leftarrow Q_{Z_1}, \nu_2 \leftarrow Q_{Z_2} \). Also, put \( Y = X = (Z_1, Z_2) \), and let \( S_1 \) and \( S_2 \) be the canonical maps (Definition 2.4.2). Note that in the title, by “positive parameters” we mean the \( b_1 \) and \( b_2 \) in (2.164) are positive. We obtain the equivalence between

\[ \|g_1\|_{\frac{1}{\mu_1}} \|g_2\|_{\frac{1}{\mu_2}} \leq \mathbb{E}[g_1(Z_1)g_2(Z_2)], \quad \forall g_1 \in L^{\frac{1}{\mu_1}}(Q_{Z_1}), g_2 \in L^{\frac{1}{\mu_2}}(Q_{Z_2}) \quad (2.163) \]

and

\[ \forall P_{Z_1}, P_{Z_2}, \exists P_{Z_1 Z_2}, \quad D(P_{Z_1 Z_2} \| Q_{Z_1 Z_2}) \leq b_1 D(P_{Z_1} \| Q_{Z_1}) + b_2 D(P_{Z_2} \| Q_{Z_2}). \quad (2.164) \]

Note that the function \( f_1 \) in (2.97) disappears since its “optimal value” is easily computed in terms of \( g_1 \) and \( g_2 \) from (2.96) in this special setting. Moreover, in this set-up, if \( Z_1 \) and \( Z_2 \) are finite, then the condition in the last bullet in Theorem 2.4.5 is equivalent to \( Q_{Z_1 Z_2} \prec Q_{Z_1} \times Q_{Z_2} \). The equivalent formulations of reverse hypercontractivity were observed in [78], where the proof is based on the method of types argument.
Figure 2.5: Diagram for reverse HC with one negative parameter

2.5.6 Reverse Hypercontractivity (One Negative Parameter)

In Theorem 2.4.5, take \( l \leftarrow 1, m \leftarrow 2, c_1 \leftarrow 1, d \leftarrow 0, g_1 \leftarrow g^{\frac{1}{c_1}}, f_2 \leftarrow f^{\frac{1}{-c_2}}, \mu_1 \leftarrow Q_{Z_1Y_2}, \nu_1 \leftarrow Q_{Z_1}, \mu_2 \leftarrow Q_{Y_2}. \) Also, put \( Y_1 = X = (Z_1,Y_2), \) and let \( S_1 \) and \( T_2 \) be the canonical maps (Definition 2.4.2). Note that in the title, by “one negative parameter” we mean the \( b_1 \) is positive and \(-c_2\) is negative in (2.166). We obtain the equivalence between

\[
\|f\|_{\frac{1}{c_2}} \|g\|_{\frac{1}{c_1}} \leq \mathbb{E}[f(Y_2)g(Z_1)], \quad \forall f \in L^{\frac{1}{c_2}}(Q_{Y_2}), \ g \in L^{\frac{1}{c_1}}(Q_{Z_1}) \tag{2.165}
\]

and

\[
\forall P_{Z_1}, \exists P_{Z_1Y_2}, \ D(P_{Z_1Y_2} \| Q_{Z_1Y_2}) \leq b_1 D(P_{Z_1} \| Q_{Z_1}) + (-c_2) D(P_{Y_2} \| Q_{Y_2}). \tag{2.166}
\]

Again, the function \( f_1 \) in (2.97) does not exist in (2.165) since its “optimal choice”

\[
f_1(z_1,y_2) = \frac{g_1^{b_1}(z_1)}{f_2^{c_2}(y_2)} = f(y_2)g(z_1), \quad \forall y_2, z_1 \tag{2.167}
\]

is easily computed in terms of \( g_1 \) and \( f_2 \) from (2.96) in this special setting. Inequality (2.165) is called reverse hypercontractivity with a negative parameter in [79], where an equivalent formulation is established for finite alphabets using the method of types. Multiterminal extensions of (2.165) and (2.166) (called reverse Brascamp-Lieb type inequality with negative parameters in [79]) can also be recovered from Theorem 2.4.5.
in the same fashion, i.e., we move all negative parameters to the other side of the inequality so that all parameters become positive.

In summary, from the viewpoint of Theorem 2.4.5, the results in Sections 2.5.4, 2.5.5 and 2.5.6 are degenerate special cases, in the sense that in any of the three cases the optimal choice of one of the functions in (2.97) can be explicitly expressed in terms of the other functions, hence this “hidden function” disappears in (2.161), (2.163) or (2.165).

2.5.7 Transportation-Cost Inequalities

Definition 2.5.1 (see for example [57]) We say that a probability measure $Q$ on a metric space $(Z, d)$ satisfies the $T_p(\lambda)$ inequality, $p \in [1, \infty)$, $\lambda \in (0, \infty)$, if

$$\inf_{\pi} \mathbb{E}_{\pi}^P \left[ d^p(X, Y) \right] \leq \sqrt{2\lambda D(P \| Q)}$$

(2.168)

for every $P \ll Q$, where the infimum is over all coupling $\pi$ of $P$ and $Q$, and $(X, Y) \sim \pi$.

It suffices to focus on the case of $\lambda = 1$, since results for general $\lambda \in (0, \infty)$ can usually be obtained by a scaling argument.

As a consequence of Theorem 2.4.7 and Remark 2.4.9, we have

Corollary 2.5.1 Let $(Z, d)$ be a locally compact, $\sigma$-compact Polish space.

(a) A probability measure $Q$ on $Z$ satisfies $T_2(1)$ inequality if and only if for any $f \in C_b(Z)$,

$$\log Q \left( \exp \left( \inf_{z \in Z} \left[ f(z) + \frac{d^2(\cdot, z)}{2} \right] \right) \right) \leq Q(f).$$

(2.169)
(b) A probability measure \( Q \) on \( Z \) satisfies \( T_p(1) \) inequality, \( p \in [1,2) \), if and only if

\[
\log Q \left( t \inf_{z \in Z} \left[ f(z) + \frac{d^p(\cdot, z)}{p} \right] \right) \leq \left( \frac{1}{p} - \frac{1}{2} \right) t^{2/p} + tQ(f), \quad \forall t \in [0, \infty), \, f \in C_b(Z). 
\]

(2.170)

**Proof** (a) In Theorem 2.4.7 put \( l = 2, \, m = 1, \, Z_1 = Z_2 \leftarrow Z, \, \mathcal{X} = \mathcal{Y}_1 \leftarrow Z \times Z, \) and

\[
\Lambda_1(u) := 2 \log Q \left( \exp \left( \frac{u}{2} \right) \right); \\
\Lambda_2(u) := Q(u); \\
\Theta_1(u) := \begin{cases} 0 & u \leq d^2; \\ +\infty & \text{otherwise}. \end{cases}
\]

(2.171) \hspace{1cm} (2.172) \hspace{1cm} (2.173)

For \( \ell \in \mathcal{M}(\mathcal{X}) \), we can compute

\[
\Lambda_1^*(\ell) = 2D(\ell\| Q); \\
\Lambda_2^*(\ell) := \begin{cases} 0 & \ell = Q; \\ +\infty & \text{otherwise}; \end{cases} \\
\Theta_1^*(\ell) = \begin{cases} \ell(d^2) & \ell \geq 0; \\ +\infty & \text{otherwise}. \end{cases}
\]

(2.174) \hspace{1cm} (2.175) \hspace{1cm} (2.176)

We also have \( \Lambda_i^*(\ell) = +\infty \) for any \( \ell \notin \mathcal{P}(\mathcal{X}), \, i = 1,2. \) Thus by Theorem 2.4.7 \[(2.169)\] is equivalent to the following: for any \( f_1 \in C_b(Z_1 \times Z_2), \, g_1 \in C_b(Z_1), \, g_2 \in C_b(Z_2) \) such that \( g_1 + g_2 \leq f_1 \), it holds that

\[
\Lambda_1(g_1) + \Lambda_2(g_2) \leq \Theta_1(f_1). 
\]

(2.177)
By the monotonicity of $\Lambda_1$, this is equivalent to

$$\Lambda_1 \left( \inf_z [d^2(\cdot, z) - g_2(z)] \right) + \Lambda_2(g_2) \leq 0$$

(2.178)

for any $g_2 \in C_b(Z)$, which is the same as (2.169).

(b) The proof is similar to Part (a), except that we now pick

$$\Theta_1(f) := \begin{cases} 
2^{-\frac{2}{p} - \frac{p}{2p}} (2 - p) \sup_{x} (\frac{f(x)}{f(x)^{p-1}}) & \text{if } \sup f \geq 0; \\
0 & \text{otherwise,}
\end{cases}$$

(2.179)

so that for any $\ell \geq 0$,

$$\Theta_1^*(\ell) = [\ell(d^p)]^{\frac{2}{p}}.$$ 

(2.180)

Remark 2.5.1 Actually, the proof of Corollary 2.5.4 does not use the assumption that $d$ is a metric (other than that it is a continuous function which is bounded below). The equivalent formulation of the $T_1$ inequality (special case of (2.170)) was known to Rachev [136] and Bobkov and Götze [113] (who actually slightly simplified the formula using the fact that $d$ is a metric). The equivalent formulation of the $T_2$ inequality in (2.169) also appeared in [113], and was employed in [96, 82] to show a connection to the logarithmic Sobolev inequality. The equivalent formulation of the $T_p$ inequality, $p \in [1, 2)$ in (2.170) appeared in [57, Proposition 22.3].

Transportation-cost inequalities have been fruitful in obtaining measure concentration results (since [137, 138]). Section 6.4.3 contains further discussions on $T_2$ inequalities in the Gaussian case.
Chapter 3
Smoothing

In Chapter 2 we studied the fundamentals of the functional-entropic duality and used the result to derive a strong converse bound for the omniscient helper common randomness generation problem. The bound is only (first-order) asymptotically tight for vanishing communication rates, because the mutual information optimization arising from the single-letter rate region coincides with the corresponding relative entropy optimization only in that case. In this chapter we introduce a machinery called smoothing to overcome this issue. By perturbing a product distribution by a tiny bit in the total variation distance, we regain the (first-order asymptotic) equivalence between the relative entropy optimization and the mutual information optimization in rather general settings. Analysis of the second-order asymptotics is carried out in the discrete and the Gaussian cases, implying the optimal $O(\sqrt{n})$ second-order converse of the whole rate region for the omniscient-helper problem in those cases.

3.1 The BL Divergence and Its Smooth Version

This section introduces some necessary definitions and notations, and summarizes the main goal of this chapter.

**Definition 3.1.1** Given a nonnegative finite measure $\mu$ on $X$, nonnegative $\sigma$-finite measures $\nu_1, \ldots, \nu_m$ on $Y_1, \ldots, Y_m$, random transformations $Q_{Y_1|X}, \ldots, Q_{Y_m|X}$, and
Define the Brascamp-Lieb divergence

\[ d(\mu, (Q_{Y_j}|X), (\nu_j), c^m) := \sup_{P_X} \left\{ \sum_{j=1}^{m} c_j D(P_{Y_j||\nu_j}) - D(P_X\|\mu) \right\} \]  

(3.1)

where the supremum is over \( P_X : P_X \ll \mu \) such that \( D(P_X\|\mu) \) and \( D(P_{Y_j||\nu_j}) \), \( j = 1, \ldots, m \) are finite, and \( P_X \rightarrow Q_{Y_j|X} \rightarrow P_{Y_j} \).

We remark that the optimization in (3.1) is generally not convex. Moreover, as a convention of this thesis, if the feasible set for a supremum is empty, then we set the supremum equal to \(-\infty\).

By Theorem 2.2.3, we have the following alternative expression for the Brascamp-Lieb (BL) divergence:

**Proposition 3.1.1** Under the assumptions in Definition 3.1.1,

\[ d(\mu, (Q_{Y_j}|X), (\nu_j), c^m) = \sup_{f_1, \ldots, f_m} \left\{ \log \int \exp \left( \sum_{j=1}^{m} \mathbb{E}[\log f_j(Y_j)|X = x] \right) \, d\mu(x) - \sum_{j=1}^{m} \log \| f_j \|_{L^1[\mathcal{Y}_j]} \right\} \]  

(3.2)

where the supremum is over nonnegative functions on \( \mathcal{Y}_1, \ldots, \mathcal{Y}_m \), and \( Y_j \sim Q_{Y_j|X=x} \) conditioned on \( X = x \).

The entropic definition (3.1) is more close to the answer (the single-letter formula of network information theory problems, usually expressed using the linear combination of mutual informations), whereas the functional (3.2) is more “operational” in the derivation of bounds.

When other parameters are fixed, the BL divergence \( d(\mu, (Q_{Y_j}|X), (Q_{Y_j}), c^m) \) can be viewed as a function of \( \mu \). Loosely speaking, the “smoothing” operation in the title of the chapter can be thought of as the “semicontinuous regularization” of a function on the measure space in the total variation distance. That is, we infimize (or supremize) a function within a neighbourhood of its argument. The rationale for
considering such a regularization is that a small perturbation of the distribution in the
total variation distance implies only a small change of the observed error probability,
but the asymptotic behavior of the value of the function can be vastly different. For
example, in cryptography, it has long been known that the collision entropy (Rényi
entropy of order 2) bounds the performance of the privacy amplification (see e.g.
[139][140][141][142]). However, Renner and Wolf [53] showed that the smooth Rényi
entropy has the asymptotic behavior of the Shannon entropy (Rényi entropy of order
1). Hence, the correct rate formula for privacy amplification is still expressed in terms
of the Shannon information quantities.

Now, let us try to apply the same smoothing operation to the BL divergence. Instead of using the total variation distance, however, we shall consider a very similar,
but mathematically more convenient, quantity:

**Definition 3.1.2** For nonnegative measures \( \nu \) and \( \mu \) on the same measurable space
\((\mathcal{X}, \mathcal{F})\) where \( \nu(\mathcal{X}) < \infty \), and \( \gamma \in [1, \infty) \), the \( E_\gamma \) divergence is defined as

\[
E_\gamma(\nu \parallel \mu) := \sup_{A \in \mathcal{F}} \{\nu(A) - \gamma \mu(A)\}.
\] (3.3)

Note that under this definition \( E_1(P \parallel \mu) = \frac{1}{2}|P - \mu| \) if and only if \( \mu \) is a probability
measure. We will return to some other applications of \( E_\gamma \) in chapter [5]

Now let us proceed to define the smooth BL divergence. Since in most of our
usage, the random transformations \( Q_{Y_1|X}, \ldots, Q_{Y_m|X} \), We sometimes omit these ran-
don transformations from the argument of the BL divergences, to keep the notations
simple.

**Definition 3.1.3** For \( \delta \in [0, 1) \), \( Q_X \), \( (Q_{Y_j|X}) \), \( (\nu_j) \) and \( c^m \in [0, \infty)^m \), define

\[
d_\delta(Q_X, (\nu_j), c^m) := \inf_{\mu : E_1(Q_X \parallel \mu) \leq \delta} d(\mu, (\nu_j), c^m)
\] (3.4)
with $d(\cdot)$ is defined in (3.1.1).

**Remark 3.1.1** According to discussions in the previous chapter, the BL divergence is a generalization of several information measures, including the strong data processing constant, hypercontractivity, and Rényi divergence. For example, for $\alpha \in (1, \infty)$, the Rényi divergence between two probability measures $P$ and $Q$ on the same alphabet can be expressed in terms of the BL divergence:

$$D_\alpha(P\|Q) = \frac{\alpha}{\alpha - 1} d\left( P, Q, \frac{\alpha - 1}{\alpha} \right).$$

(3.5)

where the random transformation in the definition of the BL divergence is the identity. Consequently, the smooth Rényi divergence [53] can be expressed in terms of a smooth BL divergence:

$$D^\delta_\alpha(P\|Q) = \frac{\alpha}{\alpha - 1} d^\delta\left( P, Q, \frac{\alpha - 1}{\alpha} \right)$$

(3.6)

for $\delta \in (0, 1)$.

**Remark 3.1.2** If we restrict $\mu$ in (3.4) to be a probability measure, then all asymptotic results remain the same. However, allowing unnormalized measures avoids the unnecessary step of normalization in the proof, and is in accordance with the literature on smooth Rényi entropy, where such a relaxation generally gives rise to nicer properties and tighter non-asymptotic bounds, cf. [53][143].

We now introduce a quantity which resembles the BL divergence and plays a central role in the characterization of the asymptotic performance of the smooth BL divergence.
Definition 3.1.4 Given $Q_X$, $(Q_{Y_j|X})$, $(\nu_j)$ and $c^m \in (0, \infty)^m$ such that $|D(Q_{Y_j}\|\nu_j)| < \infty$, $j = 1, \ldots, m$, define

\[
d^*(Q_X, (\nu_j), c^m) := \sup_{P_{UX} : P_X = Q_X} \left[ \sum_{j=1}^{m} c_j D(P_{Y_j|U}\|\nu_j|P_U) - I(U; X) \right],
\]

where $(U, X) \sim P_{UX}$, $P_{Y_j|U}$ is the concatenation of $P_{X|U}$ and $Q_{Y_j|X}$, and in the supremum we also assume that $P_{U|X}$ is such that each term in (3.7) is well-defined and finite.

Remark 3.1.3 In Definition [3.1.4] it is without loss of generality to impose that $|U| < \infty$, which can be shown using the fact that the mutual information equals its supremum over finite partitions (see [114] for more details).

In later applications to the common randomness generation problem, we take the reference measures $\nu_j = Q_{Y_j}$, $j = 1, \ldots, m$, whereas in other applications we may take $\nu_j$ equal to the equiprobable distribution. Note that in the case of $\nu_j = Q_{Y_j}$, we have

\[
d^*(Q_X, (Q_{Y_j}), c^m) = \sup_{P_{U|X}} \left\{ \sum_{i=1}^{m} c_i I(U; Y_i) - I(U; X) \right\}
\]

which is the quantity appearing in the single-letter region of the common randomness generation problem (Section 2.3.1)! Moreover, as Proposition 2.3.1 shows, we must have $d^* = d$ if one of them is known to be 0. However, in general such an equivalence may not hold, which is why our converse bound in the previous chapter is not (first-order asymptotically) tight – the issue that we are now trying to fix in this chapter.
Notice that from these definitions and the tensorization property of $d(\cdot)$ (Section 6.1), we have

$$
    d(Q_X, (Q_{Y_j}), c^m) = \frac{1}{n} d(Q_X^\otimes n, (Q_{Y_j}^\otimes n), c^m) \\
    \geq \frac{1}{n} d_\delta(Q_X^\otimes n, (Q_{Y_j}^\otimes n), c^m).
$$

(3.9) (3.10)

The goal of this chapter is to study the asymptotics of $d_\delta(Q_X^\otimes n, (Q_{Y_j}^\otimes n), c^m)$ and discuss its implication for some coding theorems.

The converse of the omniscient-helper common randomness (CR) generation problem that we will derive using $d_\delta$ implies the general lower bound

$$
    d_\delta(Q_X^\otimes n, (Q_{Y_j}^\otimes n), c^m) \geq n d^*(Q_X, (Q_{Y_j}), c^m) + O(\sqrt{n})
$$

(3.11)

for any $\delta \in (0, 1)$, since otherwise the converse bound would contradict the known achievability bound for the CR generation problem, which is impossible.

Upper bounding $d_\delta(Q_X^\otimes n, (Q_{Y_j}^\otimes n), c^m)$ is the more nontrivial part. We would still expect that $d^*(Q_X, (Q_{Y_j}), c^m)$ will emerge as the asymptotic limit in rather general settings. The results are summarized as follows:

- In the cases of 1) finite $X$, 2) $Q_{XY}^m$ is jointly Gaussian, the second-order term is order $O(\sqrt{n})$:

  $$
  d_\delta(Q_X^\otimes n, (Q_{Y_j}^\otimes n), c^m) = n d^*(Q_X, (Q_{Y_j}), c^m) + O(\sqrt{n}).
  $$

  (3.12)

- We have conclusive results on the first-order asymptotics for a rather large class of continuous distributions. More precisely, if $X = Y^m$ and $\nu_j$ is an atomless

\footnote{That is, for any measurable $\mathcal{A}: \mathcal{A} \subseteq \mathcal{Y}_j$ with $\nu_{Y_j}(\mathcal{A}) > 0$, there exists a measurable $\mathcal{B}: \mathcal{B} \subseteq \mathcal{A}$ such that $0 < \nu_{Y_j}(\mathcal{B}) < \nu_{Y_j}(\mathcal{A})$.}
nonnegative σ-finite measures on a Polish space for each \( j \), and a certain relative information quantity is continuous and absolutely integrable, then

\[
\lim_{n \to \infty} \frac{1}{n} d_\delta(Q_X^n, (Q_{Yj}^n), c^m) = d^*(Q_X, (Q_{Yj}), c^m).
\] (3.13)

The second-order asymptotic result of (3.12) will lead to second-order converses for coding theorems, whereas the first-order result of (3.13) will lead to a strong converse result. The proposed approach to strong converses has several advantages compared with existing approaches such as the method of types approach in [8], which are nicely illustrated by our example of CR generation:

• The approach is applicable to continuous sources satisfying certain regularity conditions for which the method of types is futile.

• In the omniscient helper CR generation problem, our approach covers possibly stochastic encoders and decoders. Stochastic encoders and decoders are tricky to handle with the method of types or change-of-measure techniques since they can not be described by encoding and decoding sets.

Strong converses for a continuous source when the rate region involves auxiliaries are rare [50]. The smooth BL divergence offers a canonical method for this situation.

### 3.2 Upper Bounds on \( d_\delta \)

#### 3.2.1 Discrete Case

The main result for the discrete case can be summarized as follows:

\[\text{A Polish space is a complete separable metric space. It enjoys nice properties that allow us to perform some technical steps smoothly; see e.g. [105, 104] for brief introductions.}\]

\[\text{The (asymptotic) rate region with stochastic encoders can be strictly larger than with deterministic encoders, since in the former case the CR rate is unbounded whereas in the latter case it is bounded by the entropy of the sources. Regarding the decoders, we argue in Remark 3.3.2 that allowing stochasticity can strictly decrease the (one-shot) error, but within a constant factor.}\]
Theorem 3.2.1  Fix $Q_X$, $(Q_{Y|X})$, $(\nu_j)$ and $c^m \in (0, \infty)^m$. Assume that $X$ is finite, and

\begin{align*}
D(P_{Y_j} \| \nu_j) &< +\infty, \quad \forall P_X \ll Q_X, \ j = 1, \ldots m; \quad (3.14) \\
d^*(Q_X; (\nu_j), c^m) &> -\infty. \quad (3.15)
\end{align*}

Then

\begin{align*}
d_d(Q_X^\otimes n, (\nu_j^\otimes n), c^m) &\leq n d^*(Q_X, (\nu_j), c^m) + O(\sqrt{n}). \quad (3.16)
\end{align*}

We prove a non-asymptotic version of Theorem 3.2.1 here. For simplicity we only consider the $m = 1$ case here, but the analysis completely goes through the general case as well.

Lemma 3.2.2  Consider $Q_X$, $Q_{Y|X}$, $\nu$, $\delta \in (0, 1)$, $c \in (0, \infty)$. Define $\beta_X$ and $\alpha_Y$ as in Lemma 3.2.3. Then there exists some $C_n$, $Q_X^\otimes n [C_n] \geq 1 - \delta$, such that $\mu_n := Q_X^\otimes n |_{C_n}$ satisfies

\begin{align*}
d(\mu_n, Q_{Y|X}^\otimes n, \nu^\otimes n, c) \\
\leq nd^*(Q_X, Q_{Y|X}, \nu, c) \\
+ \ln(\alpha_Y^c \beta_X^{c+1}) \sqrt{3n \beta_X \ln \frac{|X|}{\delta}} \quad (3.17)
\end{align*}

for all $n > 3 \beta_X \ln \frac{|X|}{\delta}$.

The intuition behind the second-order estimate in Lemma 3.2.2 is not hard to comprehend: we can choose

\[ C_n = \{ x^n : |\hat{P}_{x^n} - Q_X| = O(n^{-1/2}) \} \quad (3.18) \]
where $\hat{P}_{xn}$ denotes the empirical measure of $x^n$, to guarantee that $Q_X^{\otimes n}(C_n) \geq 1 - \delta$ (in Lemma 3.2.2 a slightly more sophisticated $C_n$ is chosen). A simple single-letterization argument shows that $\frac{1}{n}d(\mu_n, Q_X^{\otimes n}, \nu^{\otimes n}, c)$ can be bounded using only the empirical distributions of the sequences on the support of $\mu_n$, so that the $O(n^{-1/2})$ term in (3.18) contributes to a second-order correction term in (3.61).

**Proof** Let $\hat{P}_{X^n}$ denote the empirical distribution of $X^n \sim Q_X^{\otimes n}$. For $n > 3\beta_X \ln \frac{|X|}{\delta}$, define

$$\epsilon_n := \sqrt{\frac{3\beta_X}{n} \ln \frac{|X|}{\delta}} \in (0, 1). \quad (3.19)$$

For each $x$, we have, from the Chernoff bound for Bernoulli random variables (see below),

$$\mathbb{P}[\hat{P}_{X^n}(x) > (1 + \epsilon_n)Q_X(x)] \leq e^{-\frac{n}{2}Q_X(x)\epsilon_n^2}, \quad (3.20)$$

therefore by the union bound,

$$\mathbb{P}[\hat{P}_{X^n} \leq (1 + \epsilon_n)Q_X] \geq 1 - \delta. \quad (3.21)$$

Finally, we note that for any $P_{X^n}$ supported on

$$C_n := \{x^n : \hat{P}_{x^n} \leq (1 + \epsilon_n)Q_X\}, \quad (3.22)$$
we have
\[ cD(P_{Y^n} \| \nu^{\otimes n}) - D(P_{X^n} \| \mu_n) \]
\[ = cD(P_{Y^n} \| \nu^{\otimes n}) - D(P_{X^n} \| Q_X^{\otimes n}) \]  
(3.23)
\[ = c \sum_{i=1}^{n} D(P_{Y_i \mid X^{i-1}} \| \nu | P_{Y^{i-1}}) - \sum_{i=1}^{n} D(P_{X_i \mid X^{i-1}} \| Q_X | P_{X^{i-1}}) \]  
(3.24)
\[ \leq c \sum_{i=1}^{n} D(P_{Y_i \mid X^{i-1}} \| \nu | P_{X^{i-1}}) - \sum_{i=1}^{n} D(P_{X_i \mid X^{i-1}} \| Q_X | P_{X^{i-1}}) \]  
(3.25)
\[ = n[c D(P_{Y_i \mid I^{X-1}} \| \nu | P_{I^{X-1}}) - D(P_{X_i \mid I^{X-1}} \| Q_X | P_{I^{X-1}})] \]  
(3.26)
\[ \leq n \sup_{P_X: P_X \leq (1 + \epsilon_n)Q_X} \phi(P_X) \]  
(3.27)
where (3.25) follows since $Y_i - X^{i-1} - Y^{i-1}$ under $P$; in (3.26) we defined $I$ to be a random variable equiprobably distributed on $\{1, \ldots, n\}$ and independent of everything else; and in (3.27) $\phi$ is defined in Lemma 3.2.3. Note that (3.23)-(3.27) is essentially the same as the single-letterization property [145, Lemma 9].

Since the probability of the set in (3.22) is lower bounded by (3.21), the result follows by taking the supremum over $P_{X^n} \ll \mu_n$ in (3.23)-(3.27) and using Lemma 3.2.3 below.

The following technical result was used in the proof of the main result. For notational simplicity, the base of log in this lemma is natural.

**Lemma 3.2.3** Fix $(Q_X, Q_{Y \mid X}, \nu)$ and $c \in (0, \infty)$. Define $\beta_X := (\min_x Q_X(x))^{-1}$ and
\[ \alpha_Y = \| \nu \| \frac{dQ_Y}{d\nu} \|_\infty . \]  
(3.28)

Let
\[ \phi(P_X) := \sup_{P_{U \mid X}} \left\{ cD(P_{Y \mid U} \| \nu \mid P_{U}) - D(P_{X \mid U} \| Q_X \mid P_U) \right\} , \]  
(3.29)
for any $P_X \ll Q_X$. Then

$$\phi(P_X) \leq \phi(Q_X) + \ln(\beta_X^c \alpha_X^c) \epsilon$$  \hspace{1cm} (3.30)

for $P_X : P_X \leq (1 + \epsilon)Q_X$, $\epsilon \in [0, 1)$.

**Proof**  Given an arbitrary $P_X : P_X \leq (1 + \epsilon)Q_X$, suppose that $P_{U|X}$ is a maximizer for (3.29) (if a maximizer does not exist, the argument can still carry through by approximation). Then consider $P_{\tilde{U}, \tilde{X}}$ where $\tilde{U} = \mathcal{U} \cup \{\star\}$, $\tilde{X} = \mathcal{X}$, defined as the follows

$$P_{\tilde{U}} := \frac{1}{1 + \epsilon} P_U + \frac{\epsilon}{1 + \epsilon} \delta_*;$$  \hspace{1cm} (3.31)

$$P_{\tilde{X}|\tilde{U} = u} := P_{X|U = u}, \quad \forall u \in \mathcal{U};$$  \hspace{1cm} (3.32)

$$P_{\tilde{X}|\tilde{U} = \star} := \frac{1 + \epsilon}{\epsilon} \left( Q_X - \frac{1}{1 + \epsilon} P_X \right)$$  \hspace{1cm} (3.33)

where $\delta_*$ is the one-point distribution on $\star$. Then observe that $P_{\tilde{X}} = Q_X$, and

$$D(P_{\tilde{X}|\tilde{U}} \| Q_X | P_{\tilde{U}}) = \frac{1}{1 + \epsilon} D(P_{X|U} \| Q_X | P_U)$$

$$+ \frac{\epsilon}{1 + \epsilon} D \left( \frac{1 + \epsilon}{\epsilon} Q_X - \frac{1}{\epsilon} P_X \| Q_X \right)$$  \hspace{1cm} (3.34)

and

$$D(P_{Y|\tilde{U}} \| \nu | P_{\tilde{U}}) = \frac{1}{1 + \epsilon} D(P_{Y|U} \| \nu | P_U)$$

$$+ \frac{\epsilon}{1 + \epsilon} D \left( \frac{1 + \epsilon}{\epsilon} Q_Y - \frac{1}{\epsilon} P_Y \| \nu \right)$$  \hspace{1cm} (3.35)
which imply that

\[
D(P_X|U \| Q_X|P_U) \geq D(P_{X|\tilde{U}}|Q_X|P_{\tilde{U}})
- \epsilon D\left(\frac{1 + \epsilon}{\epsilon} Q_X - \frac{1}{\epsilon} P_X \| Q_X \right)
\geq D(P_{X|\tilde{U}}|Q_X|P_{\tilde{U}})
- \epsilon \ln \beta_X;
\]

\[
(3.36)
\]

\[
D(P_Y|U \| \nu|P_U) = (1 + \epsilon)D(P_{Y|\tilde{U}}|\nu|P_{\tilde{U}}) - \epsilon D\left(\frac{1 + \epsilon}{\epsilon} Q_Y - \frac{1}{\epsilon} P_Y \| \nu \right)
\leq D(P_{Y|\tilde{U}}|\nu|P_{\tilde{U}}) + \epsilon D(P_{Y|\tilde{U}}|\nu|P_{\tilde{U}}) + \epsilon \ln |\nu|
\]

\[
(3.39)
\]

\[
= D(P_{Y|\tilde{U}}|\nu|P_{\tilde{U}}) + \epsilon D(P_{Y|\tilde{U}}|Q_Y|P_{\tilde{U}}) + \epsilon \mathbb{E}[t_{Q_Y|\nu}(\tilde{Y})] + \epsilon \ln |\nu|
\leq D(P_{Y|\tilde{U}}|\nu|P_{\tilde{U}}) + \epsilon \ln(\beta_X \alpha_Y)
\]

\[
(3.40)
\]

where \( t_{Q_Y|\nu} := \ln \frac{dQ_Y}{d\nu} \) and the last step used the data processing inequality

\[
D(P_{Y|\tilde{U}}|Q_Y|P_{\tilde{U}}) \leq D(P_{X|\tilde{U}}|Q_X|P_{\tilde{U}}) \leq \ln \beta_X.
\]

\[
(3.42)
\]

The proof of the main result used the following standard large deviation bound

\[64\]:

**Theorem 3.2.4 (Chernoff Bound for Bernoulli random variables)** Assume that \( X_1, \ldots, X_n \) are i.i.d. Ber\((p)\). Then for any \( \epsilon \in (0, \infty) \),

\[
\mathbb{P}\left[ \sum_{i=1}^{n} X_i \geq (1 + \epsilon)np \right] \leq e^{-nD((1+\epsilon)p|p)}
\]

\[
(3.43)
\]

\[
\leq e^{-\min(\epsilon^2, \epsilon)np}. \quad (3.44)
\]
3.2.2 Gaussian Case

The main result for the Gaussian case can be summarized as follows:

**Theorem 3.2.5** Fix a Gaussian measure $Q_X$, Gaussian random transformations $(Q_{Y|X})$, Lebesgue or Gaussian measures $(\nu_j)$, and $c^m \in (0, \infty)^m$. Then

$$d_\delta(Q_X^{\otimes n}, (\nu_j^{\otimes n}), c^m) \leq n d^*(Q_X, (\nu_j), c^m) + O(\sqrt{n}). \quad (3.45)$$

In the following we prove a non-asymptotic version of the theorem. For simplicity we consider the setting where $m = 1$ and $\nu_j$ is the Lebesgue measure, but the analysis completely goes through the general case as well. Such a setting is useful in proving the second-order converse for certain distributed lossy source compression problems with quadratic cost functions. The base of log in this lemma is natural.

**Lemma 3.2.6** Let $Q_X = \mathcal{N}(0, \sigma^2)$, $Q_{Y|X=x} = \mathcal{N}(x, 1)$, and suppose that $\nu = \lambda$ is Lebesgue. Then for any $c \in (0, \infty)$, $\delta \in (0, 1)$ and $n \geq 24 \ln \frac{2}{\delta}$,

$$d_\delta(Q_X^{\otimes n}, Q_{Y|X,x}^{\otimes n}, c^m) \leq n d^*(Q_X, Q_{Y|X}, \nu, c) + \sqrt{6n \ln \frac{2}{\delta}}. \quad (3.46)$$

**Proof** Define the two kinds of typical sets:

$$S_{\epsilon_1}^n = \left\{ x^n: \frac{1}{n} \|x^n\|^2 \leq (1 + \epsilon_1)\sigma^2 \right\}, \quad (3.47)$$

and

$$T_{\epsilon_2}^n = \left\{ x^n: \frac{1}{n} \sum_{i=1}^n \nu_{Q_X}(x_i) \leq -h(Q_X) + \frac{\epsilon_2}{2} \right\} \quad (3.48)$$

$$= \left\{ x^n: \frac{1}{n} \|x^n\|^2 \geq (1 - \epsilon_2)\sigma^2 \right\}. \quad (3.49)$$

By a “Gaussian random transformation” we mean the output is a linear transformation of the input plus an independent Gaussian noise.
If we choose $\epsilon_i = \frac{A_i}{\sqrt{n}}$, $i = 1, 2$, since

$$\ln E \left[ e^{\lambda |X_n|^2/\sigma^2} \right] = \frac{n}{2} \ln \frac{1}{1 - 2\lambda}, \quad \forall \lambda \in (-\infty, 1/2),$$

we have

$$P \left[ \sum_{i=1}^{n} X_i^2 > n(1 + \epsilon_1) \sigma^2 \right] \leq e^{-n \sup_{\lambda \in (0, \frac{1}{2})} \{ \frac{1}{2} \ln(1 - 2\lambda) + \lambda (1 + \epsilon_1) \}}$$

$$= e^{-n(-\frac{1}{2} \ln(1 + \epsilon_1) + \frac{\epsilon_1}{2})}$$

$$\leq e^{-n(\epsilon_1^2/4 - \epsilon_1^2/6)}.$$  

$$P \left[ \sum_{i=1}^{n} X_i^2 < n(1 - \epsilon_2) \sigma^2 \right] \leq e^{-n \sup_{\lambda \in (-\infty, 0)} \{ \frac{1}{2} \ln(1 - 2\lambda) + \lambda (1 - \epsilon_2) \}}$$

$$= e^{-n(-\frac{1}{2} \ln(1 - \epsilon_2) - \frac{\epsilon_2}{2})}$$

$$\leq e^{-n\epsilon_2^2/4}.$$  

Therefore by the union bound,

$$1 - Q_X^\otimes \left[ S^n_{\epsilon_1} \cap T^n_{\epsilon_2} \right] \leq e^{-\frac{A_1^2}{4} + \frac{A_3^3}{6\sqrt{n}}} + e^{-\frac{A_2^2}{4}}.$$  

Now, suppose that we can show the following result: if $\mu_n$ is the restriction of $Q_X^\otimes$ on $S^n_{\epsilon_1} \cap T^n_{\epsilon_2}$, then

$$d(\mu_n, Q_{Y|X}^\otimes, \nu^\otimes, c) \leq nd^\ast(Q_X, Q_{Y|X}, \nu, c) + \sqrt{n} \left( \frac{1 - c}{2} A_1 + \frac{A_2}{2} \right).$$

Then, taking

$$A_1 = A_2 = \sqrt{6 \ln \frac{2}{\delta}}$$

87
we see that for \( n \geq 24 \ln \frac{2}{\delta} \), the right side of (3.57) is bounded above by \( \delta \) and (3.46) holds.

It remain to prove (3.58). Not complicating the notations too much, let us shift to the case of general \( m \) for the convenience of future use. Given \( (Q_{Y_j|X})_{j=1}^m \) and \( c^m \), let \( \lambda \) and \( (\lambda_j)_{j=1}^m \) denote the Lebesgue measures on the Euclidean spaces \( X \) and \( (Y_j)_{j=1}^m \), respectively. Define

\[
F(M) := \sup \left\{ - \sum c_j h(Y_j|U) + h(X|U) \right\}
\]

\[
= \sup \left\{ \sum c_j D(P_{Y_j|U} || \lambda_j | P_U) - D(P_{X|U} || \lambda | P_U) \right\}
\]

where the suprema are over \( P_{U|X} \) such that \( \Sigma_X \leq M \) for the given positive semidefinite matrix \( M \). In Chapter 6 we will discuss results about the Gaussian optimality in this type of optimizations. In particular, we will be able to show that \( F(M) \) equals the sup in (3.61) restricted to constant \( U \) and Gaussian \( P_X \), which implies that

\[
F(\Sigma) + C = d^*(Q_X, \nu_j, c^m)
\]

for

\[
C := - \sum_j c_j E[U_{\nu_j||Y_j}(Y_j)] - h(X)
\]
where \((Y_j, X) \sim Q_{Y_j|X}Q_X\). For \(\mu_n\) defined as the restriction of \(Q_X^{\otimes n}\) on \(S_{\epsilon_1}^n \cap T_{\epsilon_2}^n\) and \(P_{X^n} \ll \mu_n\), we have:

\[
\frac{1}{n} \left[ \sum_j c_j D(P_{Y_j^n}^{\otimes n} \| \lambda_j^{\otimes n}) - D(P_{X^n} \| \lambda^{\otimes n}) \right] \\
= \frac{1}{n} \sum_{i=1}^n \left[ \sum_j c_j D(P_{Y_{i,j}} \| \lambda_{Y_{i,j-1}}) - D(P_{X_{i-1}} \| \lambda_{X_{i-1}}) \right] \tag{3.64}
\]

\[
\leq \frac{1}{n} \sum_{i=1}^n \left[ \sum_j c_j D(P_{Y_{i,j}} \| \lambda_{X_{i-1}}) - D(P_{X_{i-1}} \| \lambda_{X_{i-1}}) \right] \tag{3.65}
\]

\[
\leq F((1 + \epsilon_1)\Sigma). \tag{3.66}
\]

Since \(P_{X^n}\) is supported on \(T_{\epsilon_2}^n\), we also have

\[
\frac{1}{n} \left[ \sum_j c_j D(P_{Y_j^n}^{\otimes n} \| \nu_j^{\otimes n}) - D(P_{X^n} \| \lambda^{\otimes n}) \right] + C \\
\geq \frac{1}{n} \left[ \sum_j c_j D(P_{Y_j^n}^{\otimes n} \| Q_X^{\otimes n}) - D(P_{X^n} \| Q_X^{\otimes n}) \right] - \epsilon_2 \tag{3.67}
\]

Hence from (3.66)-(3.67) we conclude

\[
\frac{1}{n} \left[ \sum_j c_j D(P_{Y_j^n}^{\otimes n} \| \mu_j^{\otimes n}) - D(P_{X^n} \| \mu_n) \right] \\
\leq F((1 + \epsilon_1)\Sigma) + C + \epsilon_2 \tag{3.68}
\]

where we used \(D(P_{X^n} \| Q_X^{\otimes n}) = D(P_{X^n} \| \mu_n)\). Also, a simple scaling property of the differential entropy shows that

\[
F((1 + \epsilon_1)\Sigma) = F(\Sigma) + \frac{\ln(1 + \epsilon_1)}{2} \left( m - \sum c_j \right) \tag{3.69}
\]

\[
\leq F(\Sigma) + \frac{1}{2\sqrt{n}} \left( m - \sum c_j \right) A_1. \tag{3.70}
\]

\(^5\)As is clear from the context, in \(Y_j^n\) the \(n\) indexes the blocklength and the \(j\) indexes the terminal, so \(Y_j^n\) is not an abbreviation for \((Y_j, \ldots, Y_n)\).
Continuing (3.68), we see

\[
\frac{1}{n} \left[ \sum_j c_j D(P_{Y_j^n} \| \nu_j^{\otimes n}) - D(P_{X^n} \| \mu_n) \right]
\]

\[
\leq F(\Sigma) + \frac{1}{2 \sqrt{n}} \left( m - \sum c_j \right) A_1 + C + \frac{A_2}{\sqrt{n}}
\]

\[
\leq d^* (Q_X, \nu_j, c^m) + \frac{1}{2 \sqrt{n}} \left( m - \sum c_j \right) A_1 + \frac{A_2}{\sqrt{n}}
\] (3.71)

(3.72)

where the last step used (3.62). Thus (3.58) is established, as desired.

3.2.3 Continuous Density Case

As alluded before, the first-order asymptotics of the BL divergence is known when

\[ X = Y^m \] and \( Y^m \) has a continuous distribution. This result is sufficient for es-
tablishing the strong converse of omniscient-helper CR generation with continuous
distributions, but the second-order analysis is still elusive at this point.

The main result of this subsection is the following:

**Theorem 3.2.7** Assume that \( Y_1, \ldots, Y_m \) are Polish spaces, \( \nu_{Y_1}, \ldots, \nu_{Y_m} \) and
\( Q_{Y^m} \) are Borel measures on \( Y_1, \ldots, Y_m \) and \( Y^m \) respectively, and \( \nu_{Q_{Y^m} \| Y^m} \) is a
(finite-valued) continuous function on \( Y^m \), where \( \nu_{Y^m} := \nu_{Y_1} \times \cdots \times \nu_{Y_m} \), such
that \( \mathbb{E}[\nu_{Q_{Y^m} \| Y^m}(Y^m)] < \infty \), where \( Y^m \sim Q_{Y^m} \). Then for any \( \delta \in (0, 1) \) and
\( c_1, \ldots, c_m \in (0, \infty) \),

\[
\liminf_{n \to \infty} d_\delta (Q_{Y^m}^{\otimes n}, (\nu_{Y_j}^{\otimes n}), c^m) \leq d^* (Q_{Y^m}, (\nu_{Y_j}), c^m).
\] (3.73)

The proof of this result can be found in [144]. It is technically involved, so we
omit it here. The proof idea is outlined as follows: first, we show that if \( c_j > 1 \) for
any \( j \in \{1, \ldots, m\} \), then the result is trivial since both sides of the desired inequality
equal \( +\infty \). Next we prove the theorem in the special case where \( Y_1, \ldots, Y_m \), using
quantization and the result for discrete distributions (Lemma 3.2.3). Finally, we extend the result to general Polish alphabets using the fact that a probability measure on a Polish space can be approximated by its restriction on a compact set.

The regularity condition \( \mathbb{E}[|t_{Q_{Y}^{m}}(\nu_{Y}^{m}(Y^{m}))|] < \infty \) is rather mild. For example, if \( \nu_{Y} \) is the Lebesgue measure, then the condition is equivalent to assuming that the positive and negative parts of the integral for computing the differential entropy of \( Y^{m} \) are both finite. When \( \nu_{Y} = Q_{Y} \) and \( m = 2 \), the condition translates into the assumption that \( I(Y_1; Y_2) < \infty \).

### 3.3 Application to Common Randomness Generation

The smooth BL divergence can be used to give single-shot converse bounds for problems in multiuser information theory. In this section we will focus on the example of omniscient-helper common randomness generation. The applications to the discrete and Gaussian Gray-Wyner source coding problem, which we do not discuss here, is similar and can be found in our journal version [144].

#### 3.3.1 Second-order Converse for the Omniscient Helper Problem

The following result can be proved based on the Theorem 2.3.2 and the smoothing idea.

**Theorem 3.3.1 (one-shot converse for omniscient helper CR generation)**

Fix \( Q_{Y}^{m} \), \( \delta \in [0, 1) \), and \( \epsilon^{m} \in (0, \infty)^{m} \) such that \( \sum_{j=1}^{m} c_{j} > 1 \). Let \( Q_{K}^{m} \) be the actual CR distribution in a coding scheme for omniscient helper CR generation, using stochastic encoders and deterministic decoders (or stochastic decoders, if \( c_{j} \leq 1 \),
\( j = 1, \ldots, m \). Then

\[
\frac{1}{2} |Q_{K^m} - T_{K^m}| \geq 1 - \frac{1}{|K|} - \frac{\prod_{j=1}^{m} |\mathcal{W}_j|^{c_j}}{|K|^{1 - \frac{1}{\Sigma_{c_i}}} \Sigma_{c_i}} \exp \left( \frac{\delta(Q_{Y^m}, (Q_{Y_j})^m, c^m)}{\sum c_i} \right) - \delta, \tag{3.74}
\]

where

\[
T_{K^m}(k^m) := \frac{1}{|K|} 1 \{k_1 = \cdots = k_m\} \tag{3.75}
\]

is the target CR distribution and \( \mathcal{K} \) and \( (\mathcal{W}_j)^m_{j=1} \) denote the CR and message alphabets.

**Proof** The \( \delta = 0 \) special case of the result was established in Theorem 2.3.2. Note the argument can be extended to the case where the source probability distribution is replaced by unnormalized measures (see Proposition 4.4.5 which holds for unnormalized measures). In other words, suppose \( \mu_{Y^m} \) is an unnormalized measure for the source variables, and let \( \mu_{K^m} \) denote the corresponding measure for the CR when the given encoders and decoders are applied. Then we have

\[
E_1(T_{K^m} \| \mu_{K^m}) \geq 1 - \frac{1}{|K|} - \frac{\prod_{j=1}^{m} |\mathcal{W}_j|^{c_j}}{|K|^{1 - \frac{1}{\Sigma_{c_i}}} \Sigma_{c_i}} \exp \left( \frac{\delta(\mu_{Y^m}, (Q_{Y_j})^m, c^m)}{\sum c_i} \right). \tag{3.76}
\]

Now Theorem 3.3.1 follows using

\[
\frac{1}{2} |Q_{K^m} - T_{K^m}| = E_1(T_{K^m} \| Q_{K^m}) \tag{3.77}
\]

\[
\geq E_1(T_{K^m} \| \mu_{K^m}) - E_1(Q_{K^m} \| \mu_{K^m}) \tag{3.78}
\]

\[
\geq E_1(T_{K^m} \| \mu_{K^m}) - E_1(Q_{Y^m} \| \mu_{Y^m}). \tag{3.79}
\]
Remark 3.3.1 Let \( Q_{K^m} \) and \( T_{K^m} \)

de note the actual and the target distributions of the CR generated by \( T_0, T_1, \ldots, T_m \), respectively. Since

\[
|Q_{K^m} - T_{K^m}| \geq |Q_{K^m} - T_{K^m}|,
\]

Theorem 3.3.1 also provides a lower bound on \( |Q_{K^m} - T_{K^m}| \). Actually, if the decoders are deterministic, \( T_0 \) can always produce \( K \) such that the two total variations are equal, because \( T_0 \) is aware of the CR produced by the other terminals.

Remark 3.3.2 Allowing stochastic decoders can strictly lower \( \frac{1}{2}|Q_{K^m} - T_{K^m}| \): consider the special case where \( m = 2 \), \( X \) and \( Y \) are constant, and there are no messages sent. Then the minimum \( \frac{1}{2}|Q_{K^m} - T_{K^m}| \) achieved by deterministic decoders is \( 1 - \frac{1}{|K|} \). On the other hand, \( T_1 \) and \( T_2 \) can each independently output an integer in \( \{1, \ldots, \sqrt{|K|}\} \) equiprobably, achieving \( \frac{1}{2}|Q_{K^m} - T_{K^m}| = 1 - \frac{1}{\sqrt{|K|}} \). On the other hand, we can argue that allowing stochastic decoders can at most reduce the error by a factor of 4, using the following argument: suppose \( \frac{1}{2}|Q_{K^m} - T_{K^m}| \leq \delta \) for some stochastic decoders, then \( \frac{1}{2}|Q_{KK_1} - T_{KK_1}| \leq \delta \) and \( Q(K_1 = K_2 = \ldots, = K_m) \leq \delta \). We can then remove the stochasticity of decoders at \( T_2 \ldots T_m \) but retain the last two inequalities. This is possible since each of \( K_j, j = 2, \ldots, m \) is independent of all other CR conditioned on the observation of \( T_j \), and we used the fact that average is no greater than the maximum. Thus \( \frac{1}{2}|Q_{KK^m} - T_{KK^m}| \leq 2\delta \) is achievable with deterministic decoders at \( T_2, \ldots, T_m \). Applying a similar argument again we can further remove the stochasticity of the decoder at \( T_1 \), at the cost of another factor of 2.

Using Theorem 3.3.1 and the definition of the BL divergence, we see that
Corollary 3.3.2 (Strong converse for omniscient helper CR generation)

Suppose that $Q_{Y^m}$, $(Q_{Y_j^m})$ and $c^m \in (0, \infty)^m$ satisfy

$$\liminf_{n \to \infty} d_\delta(Q_{Y^m}^n, (Q_{Y_j}^n), c^m) \leq d^*(Q_{Y^m}, (Q_{Y_j}), c^m) \tag{3.82}$$

and that $(R, R_1, \ldots, R_m) \in (0, \infty)^{m+1}$ satisfy

$$d^*(Q_{Y^m}, (Q_{Y_j}), c^m) + \sum_j c_j R_j \geq \left( \sum_j c_j - 1 \right) R. \tag{3.83}$$

Then, any omniscient helper CR generation scheme for the stationary memoryless source with per-letter distribution $Q_{Y^m}$ allowing stochastic encoders and decoders at the rates $(R, R_1, \ldots, R_m)$ must satisfy

$$\lim_{n \to \infty} \frac{1}{2} |Q_{K^m} - T_{K^m}| = 1, \tag{3.84}$$

where $Q_{K^m}$ denotes the actual CR distribution and $T_{K^m}$ denotes the ideal distribution under which $K_1 = \cdots = K_m$ is equiprobable.

Note that (3.82) is satisfied in the settings discussed in the previous section. Moreover, in the case of finite alphabets or Gaussian distributions, the second-order upper bounds on the smooth BL divergence in the previous section imply second-order converses for the omniscient helper CR generation problem:

Corollary 3.3.3 (Second-order converse) Assume that the per-letter source distribution $Q_{Y^m}$ is either

1. $m$-dimensional Gaussian with a non-degenerate covariance matrix, or

2. a probability distribution on a finite alphabet.
Suppose that there is a sequence of CR generation schemes such that
\[
\liminf_{n \to \infty} \frac{1}{2} |Q_{K_n^m} - T_{K_n^m}| < 1 
\] (3.85)

Then,
\[
\liminf_{n \to \infty} \sqrt{n} \left[ \left( \sum c_j - 1 \right) R_n - \sum c_j R_{jn} - d^*(Q_{\gamma m}, (Q_{\gamma j}), c^m) \right] 
\geq O(\sqrt{n}) 
\] (3.86)

where
\[
R_n := \frac{1}{n} \log |\mathcal{K}|; 
\] (3.87)
\[
R_{jn} := \frac{1}{n} \log |\mathcal{W}_j|, \quad j = 1, \ldots, m 
\] (3.88)

are the rates at blocklength \( n \).
Chapter 4

The Pump

This chapter is the third part of the trilogy on the functional approach to information-theoretic converses (Chapters 2-4), and also one of the highlights of the thesis. We introduce the “pumping-up” argument, where we design an operator $T$ with the property that for any function $f$ taking values in $[0, 1]$ whose 1-norm is not too small, the 0-norm of $Tf$ is not too small either. The Markov semigroups provide an excellent source of such operators. This method is reminiscent of the philosophy of the classical blowing-up method (the concentration of measure in the Hamming space), and can indeed substitute the parts of the strong converse proofs which used the blowing-up argument. However, in contrast to the blowing-up lemma which yields a sub-optimal $O(\sqrt{n \log^3 n})$ second-order term for discrete memoryless systems in the regime of non-vanishing error probability, the proposed method yields an $O(\sqrt{n \log \frac{1}{1-\epsilon}})$ bound on the second-order term which is sharp in both the blocklength $n$ and the error probability $\epsilon \uparrow 1$. Moreover, the new approach extends beyond stationary memoryless settings: we discuss how it works under the assumptions of bounded probability density, Gaussian distribution, and processes with memory and weak correlation. The optimality and the wide applicability of the new method suggests that it may be the “right” approach to multiuser second-order converses, and perhaps, that the classical blowing-up method works for strong converse purposes simply because the blowing-up operation approximates the optimal semigroup operation.
While Chapters 2 and 3 proved the $O(\sqrt{n})$ second-order converse for the omniscient helper CR generation problem, the “pumping-up” machinery allows us to extend the result to the general one-communicator problem, which is a canonical example of the side information problems.

4.1 Prelude: Binary Hypothesis Testing

This section we review the classical blowing-up approach and then introduce the proposed pumping-up approach through the basic example of binary hypothesis testing. Although this is a rather simple setting where many alternative approaches are applicable, it is rich enough to demonstrate many features of the new approach.

A few words about the notations of the chapter: $\mathcal{H}(\mathcal{Y})$ denotes the set of nonnegative measurable functions on $\mathcal{Y}$ and $\mathcal{H}_{[0,1]}(\mathcal{Y})$ is the subset of $\mathcal{H}(\mathcal{Y})$ with range in $[0,1]$. For a measure $\nu$ and $f \in \mathcal{H}(\mathcal{Y})$, we write $\nu(f) := \int f d\nu$ and $\|f\|_p^p = \|f\|_{L_p(\nu)}^p = \int |f|^p d\nu$, while the measure of a set is denoted as $\nu[A]$. A random transformation $Q_{Y|X}$, mapping measures on $\mathcal{X}$ to measures on $\mathcal{Y}$, is viewed as an operator mapping $\mathcal{H}(\mathcal{Y})$ to $\mathcal{H}(\mathcal{X})$ according to $Q_{Y|X}(f) := \mathbb{E}[f(Y)|X = \cdot]$ where $(X,Y) \sim Q_X Q_{Y|X}$.

4.1.1 Review of the Blowing-Up Method, and Its Second-order Sub-optimality

Many converses in information theory (most notably, the meta-converse [34]) rely on the analysis of certain binary hypothesis tests (BHT). Consider probability distributions $P$ and $Q$ on $\mathcal{Y}$. Let $f \in \mathcal{H}_{[0,1]}(\mathcal{Y})$ be the probability of deciding the hypothesis $P$ upon observing $y$. Denote by $\pi_{P|Q} = Q(f)$ the probability that $P$ is decided when $Q$ is true and vice versa by $\pi_{Q|P}$. By the data processing property of the relative
entropy

\[ D(P\|Q) \geq d(\pi_{Q\|P}1 - \pi_{P\|Q}) \]  
\[ \geq (1 - \pi_{Q\|P}) \log \frac{1}{\pi_{P\|Q}} - h(\pi_{Q\|P}), \]  
(4.1)

where \(d(\cdot\|\cdot)\) is the binary relative entropy function on \([0, 1]^2\). In the special case of product measures \(P \leftarrow P^\otimes n, Q \leftarrow Q^\otimes n\), and \(\pi_{Q^\otimes n|P^\otimes n} \leq \epsilon \in (0, 1),\) (4.2) yields

\[ \pi_{P^\otimes n|Q^\otimes n} \geq \exp \left( -n \frac{D(P\|Q)}{1 - \epsilon} - O(1) \right), \]  
(4.3)

which is not sufficiently powerful to yield the converse part of the Chernoff-Stein Lemma \([3]\) (because of the factor \(\frac{1}{1-\epsilon}\)).

In the case of deterministic tests \((f = 1_A \text{ for some } A \subseteq \mathcal{Y}^n)\) we show how to improve (4.3) by means of a remarkable property enjoyed by product measures: a small blowup of a set of nonvanishing probability suffices to increase its probability to nearly 1 \([27]\). The following “modern version” was due to Marton \([63]\); see also \([48, \text{Lemma 3.6.2}]\).

**Lemma 4.1.1 (Blowing-up)** Denote the \(r\)-blowup of \(A \subseteq \mathcal{Y}^n\) by

\[ A_r := \{v^n \in \mathcal{Y}^n : d_n(v^n, A) \leq r\}, \]  
(4.4)

where \(d_n\) is the Hamming distance on \(\mathcal{Y}^n\). Then, for any \(c > 0\),

\[ P^\otimes n[A_r] \geq 1 - e^{-c^2} \]  
(4.5)

where

\[ r = \sqrt{\frac{n}{2}} \left( \sqrt{\ln \frac{1}{P^\otimes n[A]} + c} \right). \]  
(4.6)
Moreover, as every element of $A_r$ is obtained from an element of $A$ by changing at most $r$ coordinates, a simple counting argument \cite[Lemma 5]{27} shows that

$$Q^\otimes_n[A_r] \leq C^r (r + 1) \binom{n}{r} Q^\otimes_n[A]$$

$$= \exp(nh(r/n) + O(r))Q^\otimes_n[A]$$

for $r \in \Omega(\sqrt{n}) \cap o(n)$, where $C = |\mathcal{Y}|/\min_y Q(y)$.

**Proposition 4.1.2** Assuming $|\mathcal{Y}| < \infty$ and $\pi_{Q^\otimes_n|P^\otimes_n} \leq \epsilon \in (0, 1)$. Any deterministic test between $P^\otimes_n$ and $Q^\otimes_n$ on $\mathcal{Y}^n$ satisfies

$$\pi_{P^\otimes_n|Q^\otimes_n} \geq \exp \left( -nD(P\|Q) - O(\sqrt{n} \log \frac{3}{2} n) \right).$$

**Proof** Fix $A \subseteq \mathcal{Y}$. If, in lieu of (4.1), we use

$$nD(P\|Q) \geq d \left( P^\otimes_n[A_r]\|Q^\otimes_n[A_r] \right),$$

then similarly as (4.3) we obtain that

$$Q^\otimes_n[A_r] \geq \exp \left( -n \frac{D(P\|Q)}{P^\otimes_n[A_r]} - O(1) \right).$$

Now, applying (4.5) and (4.8), and taking the (optimal) $r = \sqrt{\alpha n \log n}$ for any fixed $\alpha \in (\frac{1}{4}, \infty)$, we have

$$P^\otimes_n[A_r] \geq 1 - o_r(n^{-\frac{1}{2}});$$

$$Q^\otimes_n[A_r] \leq \exp(O(\sqrt{n} \log \frac{3}{2} n))Q^\otimes_n[A].$$

$$= \exp(O(\sqrt{n} \log \frac{3}{2} n))\pi_{P^\otimes_n|Q^\otimes_n}. $$

\footnote{The subscript in $O_\epsilon$ indicates that the implicit constant in the big $O$ notation depends on $\epsilon$. Later we may omit this subscript when the discussion focuses on the asymptotics in $n$.}
Then (4.9) follows from (4.11), (4.12) and (4.13).

Although Proposition 4.1.2 is sufficient to obtain the converse part of the Chernoff-Stein Lemma, there are far easier and more general methods to accomplish that goal. Moreover, the sublinear term in the exponent of (4.9) is not optimal. Indeed, our aim is to lower bound $Q_{b^r A_s \geq 1 - \epsilon}$ by the Neyman-Pearson lemma, the extremal $A$ is a sublevel set of $\log d_{P_{b^n A}}$, which is a sum of i.i.d. random variables under $P_{b^n}$. The optimal rate of the second-order term in (4.9) is thus the central limit theorem rate $O(\sqrt{n})$ (see for example the analysis in Lemma 2.15) via the Chebyshev inequality). Such an optimal second-order rate is fundamentally beyond the blowing-up method: simple examples show that neither (4.5) nor (4.8) can be essentially improved and the choice $r = \Theta(\sqrt{n \log n})$ is optimal.

In other classical applications of BUL such as the settings in Section 4.2 and 4.3, the suboptimal $O(p^\log n)$ second term emerges for a similar reason.

It can be seen that the present setting of BHT is a special case of the change-of-measure problem in Section 4.3 by taking $X$ to be a singleton. However, in contrast to BHT, the extremal set for the problem in Section 4.3 cannot be so simply characterized, and BUL appears to be the only previous method to achieve the second-order strengthening (to a suboptimal $O(\sqrt{n \log \frac{3}{2} n})$ order).

4.1.2 The Fundamental Sub-optimality of the Set-Inequality Approaches

Previously we have argued that the blowing-up lemma only achieves the $\sqrt{n \log \frac{3}{2} n}$ second-order rate. Here we remark that the optimal $\sqrt{n}$ rate is, in fact, fundamentally beyond the reach of the idea of transforming sets and applying the data processing argument (e.g. (4.1) and (4.10)): suppose, generalizing the blowing-up operation, we apply a certain transformation of the set $A \mapsto \tilde{A}$, and then proceed with the same
argument as Proposition 4.1.2. It is not hard to see that in order to achieve the $\sqrt{n}$ rate, such a set transformation must have the property that for any $A$ such that $P^{\otimes n}[A]$ is bounded away from 0 and 1, say $P^{\otimes n}[A] \in [1/3, 2/3]$ for all $n$, we have

- Lower bound

$$P^{\otimes n}[\tilde{A}] \geq 1 - O(n^{-\frac{1}{2}}). \quad (4.15)$$

- Upper bound

$$Q^{\otimes n}[\tilde{A}] \leq \exp(O(\sqrt{n}))Q^{\otimes n}[A]. \quad (4.16)$$

We remark that for proving a strong converse (i.e. $o(n)$ second-order rate), it suffices to relax (4.15) and (4.16) to

$$P^{\otimes n}[\tilde{A}] \geq 1 - o(1); \quad (4.17)$$

$$Q^{\otimes n}[\tilde{A}] \leq \exp(o(n))Q^{\otimes n}[A], \quad (4.18)$$

which are of course satisfied by the blowing-up operation, see (4.12) and (4.13).

Now we will assume that (4.15) and (4.16) indeed hold, and we will reach a contradiction. Upon choosing $A$ such that $P^{\otimes n}[A] \in [1/3, 2/3]$ and $Q[A] = \exp(-nD(P\|Q) + O(\sqrt{n}))$ we get

$$P^{\otimes n}[\tilde{A}] \geq 1 - O(n^{-\frac{1}{2}}), \quad (4.19)$$

$$Q^{\otimes n}[\tilde{A}] \leq \exp(-nD(P\|Q) + O(\sqrt{n})). \quad (4.20)$$

However, under the constraint (4.19), $Q^{\otimes n}[\tilde{A}]$ is minimized by $\tilde{A}$ of the form:

$$\tilde{A} := \{x^n: i_{P^{\otimes n}\|Q^{\otimes n}}(x^n) > nD(P\|Q) - C\sqrt{n\ln n}\} \quad (4.21)$$
for some $C > 0$, which can be seen from the Neyman-Pearson lemma and the Gaussian approximation of a normalized iid sum. However, for any $C' \in (0, C)$,

$$Q_n^\otimes[\tilde{A}] \geq \mathbb{P} \left[ C' \sqrt{n \ln n} < n D(P\|Q) - \mathbb{I}_{P\otimes^n}|Q\otimes^n}(Y^n) < C' \sqrt{n \ln n} \right]$$

$$\geq \exp \left( -n D(P\|Q) + C' \sqrt{n \ln n} \right)$$

$$\cdot \mathbb{P} \left[ C' \sqrt{n \ln n} < n D(P\|Q) - \mathbb{I}_{P\otimes^n}|Q\otimes^n}(X^n) < C' \sqrt{n \ln n} \right]$$

(4.23)

where $X^n \sim P\otimes^n$ and $Y^n \sim Q\otimes^n$. We can show [147, Lemma 8.1] that the second factor in (4.23) is of the order of $\exp(-O(\log n))$. Then (4.23) contradicts (4.20). In fact, by refining the above analysis we see that the best possible second-order rate that can be obtained from the data processing argument is $\sqrt{n \log n}$.

### 4.1.3 The New Pumping-up Approach

In this subsection we propose the pumping-up approach, which is a gentler alternative to BUL that achieves the optimal $O(\sqrt{n})$ second-order rate, while retaining the applicability of BUL to multiuser information theory problems. Instead of applying a set transformation $A \mapsto \tilde{A}$, we will apply a linear transformation $T$ to a function $f \in \mathcal{H}_{[0, 1]}$ (which can be thought of as the indicator function of $A$, but not necessarily). In contrast to (4.15) and (4.16), which cannot be achieved simultaneously as we have shown, we will find $T$ with the following properties:

- Lower bound: for any $f \in \mathcal{H}_{[0, 1]}(\mathcal{Y})$, with, say\footnote{We define $\|f\|_{L^2(P)} := \lim_{q \downarrow 0} \|f\|_{L^q(P)} = \exp(P(\ln f))$.} $P\otimes^n(f) \geq \frac{1}{2}$,

$$\|Tf\|_{L^2(P\otimes^n)} \geq \exp(-O(\sqrt{n})).$$

(4.24)
• Upper bound: for any $f \in \mathcal{H}_{[0,1]}(\mathcal{Y})$,

$$Q^{\otimes n}(Tf) \leq \exp(O(\sqrt{n}))Q^{\otimes n}(f).$$

(4.25)

We now explain how (4.24) and (4.25) can lead to an $O(\sqrt{n})$ second-order converse. Instead of applying the data processing argument as in (4.1) and (4.10), we note that, by the variational formula for the relative entropy (e.g. [48, (2.4.67)])

$$D(P\|Q) \geq P(\ln g) - \ln Q(g), \quad \forall g \in \mathcal{H}_+(\mathcal{Y}).$$

(4.26)

Taking $\mathcal{Y} \leftarrow \mathcal{Y}^n$, $P \leftarrow P^{\otimes n}$ and $Q \leftarrow Q^{\otimes n}$ in (4.26), we see that (4.24) and (4.25) lead to

$$Q^{\otimes n}(f) \geq e^{-nD(P\|Q) - O(\sqrt{n})}.$$  

(4.27)

Next, we need to find the operator $T$ satisfying (4.24) and (4.25). Markov semigroups appear to be tailor-made for this purpose. We say $(T_t)_{t \geq 0}$ is a simple semigroup$^4$ with stationary measure $P$ if

$$T_t:\mathcal{H}_+(\mathcal{Y}) \rightarrow \mathcal{H}_+(\mathcal{Y}), \quad f \mapsto e^{-t}f + (1 - e^{-t})P(f).$$

(4.28)

In the i.i.d. case $P \leftarrow P^{\otimes n}$ we consider their tensor product

$$T_t := [e^{-t} + (1 - e^{-t})P]^{\otimes n}$$

(4.29)

---

$^3$While the bases of the information theoretic quantities, log and exp were arbitrary (but consistent) up to here (see the convention in [43]), henceforth they are natural.

$^4$ Readers who are unfamiliar with semigroups may safely ignore this terminology; while the semigroup property plays an important role in the proof of Theorem 4.1.3, it is not used directly in this thesis.
Note that (4.24) implies that $T$ is an *positivity improving* operator, that is, it maps a nonnegative, not identically zero function to a strictly positive function. The positivity-improving property of $T_t$ is precisely quantified by the *reverse hypercontractivity* phenomenon discovered by Borell [148] and recently generalized by Mossel et al. [77].

**Theorem 4.1.3** [77] Let $(T_t)_{t \geq 0}$ be a simple semigroup (defined in (4.28)) or an arbitrary tensor product of simple semigroups. Then for all $0 < q < p < 1$, $f \in \mathcal{H}_+$ and $t \geq \ln \frac{1-q}{1-p}$,

$$\|T_t f\|_q \geq \|f\|_p. \quad (4.30)$$

When $T_t$ is defined by (4.29), for any $f \in \mathcal{H}_{[0,1]}(\mathcal{Y})$,

$$P^{\otimes n}(\ln T_t f) = \ln \|T_t f\|_{L^0(P^{\otimes n})} \quad (4.31)$$

$$\geq \ln \|f\|_{L^{1-e^{-t}}(P^{\otimes n})} \quad (4.32)$$

$$\geq \frac{1}{1-e^{-t}} \ln P^{\otimes n}(f) \quad (4.33)$$

$$\geq \left(\frac{1}{t} + 1\right) \ln P^{\otimes n}(f) \quad (4.34)$$

where (4.34) follows from $e^t \geq 1 + t$. On the other hand, instead of the counting argument (4.8), in the present setting we use

$$Q^{\otimes n}(T_t f) = Q^{\otimes n}((e^{-t} + (1 - e^{-t})P)^{\otimes n} f) \quad (4.35)$$

$$\leq Q^{\otimes n}((e^{-t} + \alpha(1 - e^{-t})Q)^{\otimes n} f) \quad (4.36)$$

$$= (e^{-t} + \alpha(1 - e^{-t}))^n Q^{\otimes n}(f) \quad (4.37)$$

$$\leq e^{(\alpha - 1)nt} Q^{\otimes n}(f), \quad (4.38)$$

for all $f \in \mathcal{H}_+(\mathcal{Y})$, where $\alpha := \frac{dP}{dQ}_{\infty} \geq 1$.  

104
Theorem 4.1.4  If \( \pi_{Q^\otimes n | P^\otimes n} \leq \epsilon \in (0,1) \), any (possibly stochastic) test between \( P^\otimes n \) and \( Q^\otimes n \) satisfies

\[
\pi_{P^\otimes n | Q^\otimes n} \geq (1 - \epsilon) \exp \left( -n D(P \| Q) - 2 \sqrt{n \left\| \frac{dP}{dQ} \right\|_\infty} \ln \frac{1}{1 - \epsilon} \right). \tag{4.39}
\]

Proof  Take \( g \leftarrow T_t f \) in (4.26) with \( f(y) \) the probability for which the test decides \( P \) observing \( y \). Apply (4.34) and (4.38). The result follows by optimizing over \( t > 0 \).

\( \blacksquare \)

Remark 4.1.1  In the setting of Theorem 4.1.4, the likelihood ratio test is optimal by the Neyman-Pearson lemma, and using the central limit theorem it is easy to show that (see e.g. [50, Proposition 2.3])

\[
\pi_{P^\otimes n | Q^\otimes n} \geq \exp \left( -n D(P \| Q) + \sqrt{n \text{Var}(\pi_{P|Q}(X))Q^{-1}(\epsilon)} - o(\sqrt{n}) \right) \tag{4.40}
\]

where \( X \sim P \), which is tight up to the \( \sqrt{n} \) second-order term. Since \( Q^{-1}(\epsilon) \sim -2\sqrt{\ln \frac{1}{1 - \epsilon}} \) as \( \epsilon \uparrow 1 \), we see that the second-order term in (4.39) is tight up to a constant factor depending only on \( P \) and \( Q \) for \( \epsilon \) close to 1.

Remark 4.1.2  It is instructive to compare the blowing-up and semigroup operations. Note that

\[
1_{\mathcal{A}^r}(x) = \sup_{|S| \leq r} \sup \left( 1_{\mathcal{A}}((z_i)_{i \in S}, (x_i)_{i \in S^c}) \right), \tag{4.41}
\]

while expanding (4.29) gives (cf. (4.51))

\[
T_t 1_{\mathcal{A}}(x) = \mathbb{E}[1_{\mathcal{A}}((X_i)_{i \in S}, (x_i)_{i \in S^c})] \tag{4.42}
\]

where the set \( S \) is uniformly distributed over sets of size \( |S| \sim \text{Binom}(n, 1 - e^{-t}) \) and \( X_i \sim P \) are i.i.d. Thus the semigroup operation can be viewed as an “average” coun-
part of blowing-up (with $r \approx n(1 - e^{-1})$). In contrast to the maximum, averaging increases the small values of $f$ (positivity-improving) while preserving the total mass $P^\otimes n(T_t f) = P^\otimes n(f)$, so that $Q^\otimes n(T_t f)$ does not increase too much.

Figure 4.1: Schematic comparison of $1_A$, $1_{A_{nt}}$ and $T_t^\otimes n 1_A$, where $A$ is the indicator function of a Hamming ball.

We note that the new method draws on a philosophy similar to BUL, but enjoys the following advantages:

- Achieves the optimal $O(\sqrt{n \ln \frac{1}{1-\epsilon}})$ second-order rate, which is sharp in both $n$ (as $n \to \infty$) and $\epsilon$ (as $\epsilon \to 1$);

- Purely measure-theoretic in nature: finiteness of the alphabet is sufficient, but not necessary, for Theorem 4.1.4. In fact, the boundedness of the Radon-Nikodym derivative is sometimes not necessary if we apply semigroups other than (4.29): see Section 4.2.2 for a Gaussian example. In contrast, while analogues of the blowing-up property (4.5) exist for many other measures, no analogue of the counting estimate (4.8) can hold in continuous settings as the blow-up of a set of measure zero may have positive measure.

A comparison between the BUL method and the pumping-up method is given in Table 4.1.
Table 4.1: Comparison of the basic ideas in BUL and the new methodology

<table>
<thead>
<tr>
<th></th>
<th>BUL approach</th>
<th>Pumping-up approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connecting information measures and observables</td>
<td>Data processing property</td>
<td>Convex duality</td>
</tr>
<tr>
<td>Lower bound with respect to a given measure</td>
<td>Blowing-Up lemma</td>
<td>Reverse hypercontractivity</td>
</tr>
<tr>
<td>Upper bound with respect to the reference measure</td>
<td>A counting argument (4.8)</td>
<td>Change of measure (4.36) (bounded density case); scaling argument (4.61) (Gaussian case)</td>
</tr>
</tbody>
</table>

4.2 Optimal Second-order Fano’s Inequality

4.2.1 Bounded Probability Density Case

Consider a random transformation $P_{Y|X}$. For any $x \in \mathcal{X}$, denote by $(T_{x,t})_{t \geq 0}$ the simple Markov semigroup:

$$T_{x,t} := e^{-t} + (1 - e^{-t})P_{Y|X=x}.$$  

(4.43)

Motivated by the steps (4.35)-(4.38), for $\alpha \in [1, \infty)$, $t \in [0, \infty)$ and a probability measure $\nu$ on $\mathcal{Y}$, define a linear operator $\Lambda^\nu_{\alpha,t} : \mathcal{H}_+(\mathcal{Y}^n) \to \mathcal{H}_+(\mathcal{Y}^n)$ by

$$\Lambda^\nu_{\alpha,t} := \prod_{i=1}^n (e^{-t} + \alpha(1 - e^{-t})\nu(i))$$  

(4.44)
where \( \nu^{(i)}: \mathcal{H}(\mathcal{Y}^n) \to \mathcal{H}(\mathcal{Y}^n) \) is the linear operator that integrates the \( i \)-th coordinate with respect to \( \nu \). Since \( \nu^{(i)} 1 = 1 \), we see from the binomial formula that

\[
\Lambda_{\alpha,t}^\nu 1 = \sum_{k=0}^{n} e^{-(n-k)t} \alpha^k (1 - e^{-t})^k \binom{n}{k} \tag{4.45}
\]

\[
= (e^{-t} + \alpha(1 - e^{-t}))^n \tag{4.46}
\]

\[
\leq e^{(\alpha-1)nt}. \tag{4.47}
\]

**Lemma 4.2.1** Fix \((P_{Y|X}, \nu, t)\). Suppose that

\[
\alpha := \sup_x \left\| \frac{dP_{Y|X=x}}{d\nu} \right\|_\infty \in [1, \infty); \tag{4.48}
\]

\[
T_{x^n,t} := \bigotimes_{i=1}^{n} T_{x_i,t}. \tag{4.49}
\]

Then for \( n \geq 1 \) and \( f \in \mathcal{H}_+(\mathcal{Y}^n) \),

\[
\sup_{x^n} T_{x^n,t} f \leq \Lambda_{\alpha,t}^\nu f. \tag{4.50}
\]

**Proof** For any \( x^n \in \mathcal{X}^n \), observe that

\[
T_{x^n,t} f = \prod_{i=1}^{n} [e^{-t} + (1 - e^{-t})P_{Y|X=x_i}^{(i)}] f
\]

\[
= \sum_{S \subseteq \{1, \ldots, n\}} e^{-|S|t} (1 - e^{-t})^{|S|} \left( \prod_{i \in S} P_{Y|X=x_i}^{(i)} \right) (f), \tag{4.51}
\]

The result then follows from (4.44) and (4.48). \( \square \)

**Theorem 4.2.2** Fix \( P_{Y|X} \) and positive integers \( n \) and \( M \). Suppose (4.48) holds for some probability measure \( \nu \) on \( \mathcal{Y} \). If there exists \( c_1, \ldots, c_M \in \mathcal{X}^n \) and disjoint
\( \mathcal{D}_1, \ldots, \mathcal{D}_M \subseteq \mathcal{Y}^n \) such that

\[
\prod_{m=1}^{M} P_{Y^n|X^n=c_m}^{\frac{1}{M}}[\mathcal{D}_m] \geq 1 - \epsilon, \tag{4.52}
\]

then

\[
I(X^n; Y^n) \geq \ln M - 2 \sqrt{(\alpha - 1)n \ln \frac{1}{1 - \epsilon} - \ln \frac{1}{1 - \epsilon}}, \tag{4.53}
\]

where \( X^n \) is equiprobable on \( \{c_1, \ldots, c_M\} \), \( Y^n \) is its output from \( P_{Y^n|X^n} = P_{Y^n|X} \), and \( \alpha \) is defined in (4.48).

**Proof** Let \( f_m := 1_{\mathcal{D}_m}, m = 1, \ldots, M \). Fix some \( t > 0 \) to be optimized later. Observe that

\[
I(X^n; Y^n) = \frac{1}{M} \sum_{m=1}^{M} D(P_{Y^n|X^n=c_m} \parallel P_{Y^n}) \tag{4.54}
\]

\[
\geq \frac{1}{M} \sum_{m=1}^{M} P_{Y^n|X^n=c_m} (\ln \Lambda_{\alpha,t}^{c_m} f_m) - \frac{1}{M} \sum_{m=1}^{M} \ln P_{Y^n}(\Lambda_{\alpha,t}^{c_m} f_m) \tag{4.55}
\]

where (4.55) is from the variational formula (4.26). We can lower bound the first term of (4.55) by

\[
\frac{1}{M} \sum_{m=1}^{M} \ln \| \Lambda_{\alpha,t}^{c_m} f_m \|_{L^0(P_{Y^n|X^n=c_m})} \geq \frac{1}{M} \sum_{m=1}^{M} \ln \| T_{c_m,t} f_m \|_{L^0(P_{Y^n|X^n=c_m})} \tag{4.56}
\]

\[
\geq - \left( \ln \frac{1}{1 - \epsilon} \right) \left( 1 + \frac{1}{t} \right), \tag{4.57}
\]
where (4.56)-(4.57) follows similarly as (4.31)-(4.34). For the second term in the right of (4.55), using Jensen’s inequality and (4.47),

\[
\frac{1}{M} \sum_{m=1}^{M} \ln P_{Y^n}(\Lambda_{\alpha,t}^n f_m) \geq - \ln P_{Y^n} \left( \frac{1}{M} \sum_{m=1}^{M} \Lambda_{\alpha,t}^n f_m \right) \quad (4.58)
\]

\[\geq \ln M - (\alpha - 1)nt. \quad (4.59)\]

The result then follows by optimizing t.

\[\blacksquare\]

**Remark 4.2.1** If \(|\mathcal{Y}| < \infty\) then (4.48) holds with \(\nu\) equiprobable and \(\alpha = |\mathcal{Y}|\). The “geometric average criterion” in (4.52) is weaker than the maximal error criterion but stronger than the the average error criterion. Under the average error criterion, one cannot expect a bound like \(I(X^n;Y^n) \geq \ln M - o(n)\) since it would contradict [149, Remark 5]. In [150] we introduce the notion of “\(\lambda\)-decodability” which subsumes the geometric average criterion, average error criterion, and the maximum error criterion as the \(\lambda = 0, 1, -\infty\) special cases, and show that \(\lambda = 0\) is the critical value for a strengthening of the Fano inequality to hold.

**Remark 4.2.2** Using the blowing-up lemma, one can get a weaker version of Theorem 4.2.2 in the case of finite alphabets under the maximal error criterion with the \(O(\sqrt{n})\) second-order term replaced by \(O(\sqrt{n}\log^{3/2} n)\). This is essentially contained in the analysis in [27][63]; see also [149, Theorem 6].

### 4.2.2 Gaussian Case

We now find an analogue of Theorem 4.2.2 for Gaussian channels, i.e., \(P_{Y|X=x} = \mathcal{N}(x,1)\). The previous argument does not immediately carry through due to the bounded density assumption (4.48). To surmount this problem, it will be convenient to replace (4.49) in the Gaussian setting by the *Ornstein-Uhlenbeck semigroup* with
stationary measure $\mathcal{N}(x^n, I_n)$:

$$T_{x^n, t}f(y^n) := \mathbb{E}[f(e^{-t}y^n + (1 - e^{-t})x^n + \sqrt{1 - e^{-2t}}V^n)]$$ (4.60)

for $f \in \mathcal{H}_+(\mathbb{R}^n)$, where $V^n \sim \mathcal{N}(0^n, I_n)$. In this setting, (4.30) holds under the even weaker assumption $t \geq \frac{1}{2} \ln \frac{1-q}{1-p}$ [148].

Figure 4.2: Illustration of the action of $T_{x^n, t}$. The original function (an indicator function) is convolved with a Gaussian measure and then dilated (with center $x^n$).

The proof proceeds a little differently than the discrete case. Here, the analogue of $\Lambda_{\alpha, t}$ in (4.44) is simply $T_{0^n, t}$. Instead of Lemma 4.2.1, we will exploit a simple change-of-variable formula: for any $f \geq 0$, $t > 0$ and $x^n \in \mathbb{R}^n$, we have

$$P_{Y^n|X^n = x^n}(\ln T_{0^n, t}f) = P_{Y^n|X^n = e^{-t}x^n}(\ln T_{e^{-t}x^n, t}f),$$ (4.61)

which can be verified from the definition in (4.60). For later applications in broadcast channels, we consider a slight extension of the setting of Theorem 4.2.2 to allow stochastic encoders.

111
Theorem 4.2.3 Let \( P_{Y|X=x} = \mathcal{N}(x, \sigma^2) \). Assume that there exist
\[
\phi: \{1, \ldots, M\} \times \{1, \ldots, L\} \rightarrow \mathbb{R}^n,
\]
and disjoint sets \((D_w)_{w=1}^M\) such that
\[
\prod_{w,v} P_{Y^n|X^n=\phi(w,v)}[D_w] \geq 1 - \epsilon
\]
where \( \epsilon \in (0, 1) \). Then
\[
I(W; Y^n) \geq \ln M - \sqrt{2n \ln \frac{1}{1-\epsilon}} - \ln \frac{1}{1-\epsilon}
\]
where \( (W, V) \) is equiprobable on \( \{1, \ldots, M\} \times \{1, \ldots, L\} \), \( X^n := \phi(W, V) \), and \( Y^n \) is the output from \( P_{Y^n|X^n} := P_{Y|X} \).

Remark 4.2.3 Under the maximal error probability criterion, one can also derive
a second-order strengthening of the Gaussian Fano inequality \((4.64)\) by applying
Poincaré’s inequality to an information spectrum bound; see [149, Theorem 8].

Proof By the scaling invariance of the bound it suffices to consider the case of
\( \sigma^2 = 1 \). Let \( f_w = 1_{D_w} \) for \( w \in \{1, \ldots, M\} \). Put \( X^n = e^t X^n \) and \( Y^n \) the corresponding
output from the same channel. Note that for each \( w \),
\[
D(P_{Y^n|W=w} \| P_{Y^n}) \geq \frac{1}{L} \sum_{v=1}^L P_{Y^n|X^n=e^t\phi(w,v)}(\ln T_{0^n,t}f_w) - \ln P_{Y^n}(T_{0^n,t}f_w) \geq \frac{1}{L} \sum_{v=1}^L P_{Y^n|X^n=\phi(w,v)}(\ln T_{\phi(w,v),t}f_w) - \ln P_{Y^n}(T_{0^n,t}f_w)
\]
where the key step \((4.66)\) used \((4.61)\). The summand in the first term of \((4.66)\) can
be bounded using reverse hypercontractivity for the Ornstein-Uhlenbeck semigroup.
\[(4.60)\text{ as}
\begin{align*}
P_{Y^n|X^n=\phi(w,v)}(\ln T_{\phi(w,v),t} f_w) &\geq \frac{1}{1-e^{-2t}} \ln P_{Y^n|X^n=\phi(w,v)}(f_w) \\
\end{align*}
\]

thus from the assumption \[(4.63)\],
\begin{align*}
\frac{1}{M L} \sum_{w,v} P_{Y^n|X^n=\phi(w,v)}(\ln T_{0^n,t} f_w) &\geq - \frac{1}{1-e^{-2t}} \ln \frac{1}{1-\epsilon}.
\end{align*}

On the other hand, by Jensen’s inequality,
\begin{align*}
\frac{1}{M} \sum_{w=1}^{M} \ln P_{Y^n}(T_{0^n,t} f_w) &\leq \ln P_{Y^n} \left( T_{0^n,t} \frac{1}{M} \sum_{w=1}^{M} f_w \right) \leq \ln \frac{1}{M}
\end{align*}

where the last step uses \(\sum_{w=1}^{M} f_w \leq 1\). Thus taking \(\frac{1}{M} \sum_{w=1}^{M} f_w\) on both sides of \[(4.66)\],
we find
\begin{align*}
I(W; \bar{Y}^n) &\geq \ln M - \frac{1}{1-e^{-2t}} \ln \frac{1}{1-\epsilon}.
\end{align*}

\[(4.67)\]

Moreover, let \(G^n \sim \mathcal{N}(0^n, \mathbf{I}_n)\) be independent of \(X^n\),
\begin{align*}
h(\bar{Y}^n) &= h(e^t X^n + G^n) \\
&= h(X^n + e^{-t} G^n) + nt \\
&\leq h(Y^n) + nt,
\end{align*}

\[(4.68), (4.69), (4.70)\]

where \[(4.70)\] can be seen from the entropy power inequality. On the other hand for each \(w\) we have
\begin{align*}
h(\bar{Y}^n|W=w) - h(Y^n|W=w) \\
= I(\bar{Y}^n; X^n|W=w) - I(Y^n; X^n|W=w) &\geq 0
\end{align*}

\[(4.71)\]
where (4.71) can be seen from [151, Theorem 1]. The result follows from (4.67), (4.70), (4.71) and by optimizing $t$.

4.3 Optimal Second-order Change-of-Measure

We now revisit a change-of-measure problem considered in [27] in proving the strong converse of source coding with side information (cf. Section 4.4.3). For a “combinatorial” formulation, which amounts to assigning an equiprobable distribution on the strongly typical set, see [8, Theorem 15.10]. Variations on such an idea have been applied to the “source and channel networks” in [8, Chapter 16] under the name “image-size characterizations”. Given a random transformation $Q_{Y|X}$, measures $\nu$ on $Y$ and $\mu_n$ on $\mathcal{X}^n$, we want to lower bound the $\nu^{\otimes n}$-measure of $A \subseteq \mathcal{Y}^n$ in terms of the $\mu_n$-measure of its “$\epsilon$-preimage” under $Q_{Y^n|X^n} := Q_{Y|X}^{\otimes n}$. More precisely, for given $c > 0$, $\epsilon \in (0, 1)$ find an upper bound on

$$
\sup_A \left\{ \ln \mu_n(x^n: Q_{Y^n|X^n} = x^n[A] > 1 - \epsilon) - c \ln \nu^{\otimes n}[A] \right\}.
$$

![Figure 4.3: A set $A \subseteq \mathcal{Y}$ and its “preimage” under $Q_{Y|X}$ in $\mathcal{X}$.

Figure 4.3: A set $A \subseteq \mathcal{Y}$ and its “preimage” under $Q_{Y|X}$ in $\mathcal{X}$.](image)
Definition 4.3.1  Fix $\mu$ on $\mathcal{X}$, $\nu$ on $\mathcal{Y}$, and $Q_{Y|X}$. For $c \in (0, \infty)$ and $P_X \rightarrow Q_{Y|X} \rightarrow P_Y$,

$$d(\mu, Q_{Y|X}, \nu, c) := \sup_{P_X : P_X \ll \mu} \{cD(P_Y \| \nu) - D(P_X \| \mu)\}$$

(4.72)

$$= \sup_{f \in \mathcal{H}_+(\mathcal{Y})} \{\ln \mu(e^{cQ_{Y|X}(\ln f)}) - c \ln \nu(f)\}.$$  

(4.73)

where (4.73) follows from, e.g. Theorem 2.2.3.

In Definitions 4.3.1 and 4.3.2 we adopt the convention $\infty - \infty = -\infty$.

Observe that given $Q_{XY}$, the largest $c > 0$ for which $d(\mu, Q_{Y|X}, \nu, c) = 0$ is the reciprocal of the strong data processing constant; see the references in [80]. If in definition (4.72) for $d(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c)$ we choose $P_{X^n}$ to be $\mu_n$ conditioned on $B := \{x^n : Q_{Y^n|X^n=x^n}(A) > 1 - \epsilon\}$, i.e., $P_{X^n}[C] := \frac{\mu_n[B \cap C]}{\mu_n[B]}$, $\forall C$, then [27, (19)-(21)] showed that when $|\nu| = 1$,

$$\ln \mu_n[x^n : Q_{Y^n|X^n=x^n}(A) > 1 - \epsilon] - c(1 - \epsilon) \ln \nu^{\otimes n}[A]$$

$$\leq d(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) + c \ln 2.$$  

(4.74)

We quickly recall the derivation of (4.74) here: define $P_{X^n}$ by

$$P_{X^n}[C] := \frac{\mu_n[B \cap C]}{\mu_n[B]}, \quad \forall C.$$  

(4.75)
Then

\[ D(P_X^n \| \mu_n) = \ln \frac{1}{\mu_n[B]} \]  \hspace{1cm} (4.76)  
\[ D(P_{Y^n} \| \nu^{\otimes n}) \geq P_{Y^n}[A] \ln \frac{P_{Y^n}[A]}{\nu^{\otimes n}[A]} + P_{Y^n}[A^c] \ln \frac{P_{Y^n}[A^c]}{\nu^{\otimes n}[A^c]} \]  \hspace{1cm} (4.77)  
\[ \geq -h(P_{Y^n}[A]) + P_{Y^n}[A] \ln \frac{1}{\nu^{\otimes n}[A]} \]  \hspace{1cm} (4.78)  
\[ \geq -\ln 2 + (1 - \epsilon) \ln \frac{1}{\nu^{\otimes n}[A]}. \]  \hspace{1cm} (4.79)  

Thus

\[ d(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}) \geq c D(P_{Y^n} \| \nu^{\otimes n}) - D(P_X^n \| \mu_n) \]  \hspace{1cm} (4.80)  
\[ \geq -c \ln 2 + c(1 - \epsilon) \ln \frac{1}{\nu^{\otimes n}[A]} - \ln \frac{1}{\mu_n[B]} \]  \hspace{1cm} (4.81)  

which is \[4.74\].

Note the undesired second \(\epsilon\) in \[4.74\], which would result in a weak converse. For finite \(\mathcal{Y}\), [27] used the blowing-up lemma to strengthen \[4.74\]. For the same reason discussed in Section 4.1 even using modern results on concentration of measure, one can only obtain \(O(\sqrt{n} \ln^{3/2} n)\) in the second order term:

\[ \ln \mu_n[x^n]: Q_{Y^n|X^n=x^n}[A] > 1 - \epsilon \]  \hspace{1cm} (4.82)  
\[ \leq d(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) + O(\sqrt{n} \ln^{3/2} n). \]  \hspace{1cm} (4.82)

If \(\mu_n = Q_X^{\otimes n}\), then by the tensorization of the BL divergence (Chapter 6),

\[ d(Q_X^{\otimes n}, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) = n d(Q_X, Q_{Y|X}, \nu, c), \]  \hspace{1cm} (4.83)  

we see the right side of \[4.74\] grows linearly with slope \(d(Q_X, Q_{Y|X}, \nu, c)\), which is larger than desired (i.e. when applied to the source coding problem in Section 4.4.3

116
would only result in an outer bound). Luckily, when \( |\mathcal{X}| < \infty \), it was noted in [27] that if \( \mu_n \) is the restriction of \( Q_X^\otimes n \) on the \( Q_X \)-strongly typical set then the linear growth rate of \( d(\mu_n, Q_{Y|X}^\otimes n, \nu^\otimes n, c) \) is the following (desired) quantity:

**Definition 4.3.2** Given \( Q_X, Q_{Y|X} \), measure \( \nu \) on \( Y \) and \( c \in (0, \infty) \), define

\[
d^*(Q_X, Q_{Y|X}, \nu, c) := \sup_{P_{UX}: P_X = Q_X} \{c D(P_{Y|U} \| \nu| P_U) - D(P_{X|U} \| Q_X| P_U)\}
\]

where \( P_{UX} := P_{UX}Q_{Y|X} \), and \( D(P_{X|U} \| Q_X| P_U) := \int D(P_{X|U} \| Q_X) dP_U \) denotes the conditional relative entropy.

It follows from Definitions 4.3.1 and 4.3.2 that \( d^*(Q_X, Q_{Y|X}, \nu, c) \leq d(Q_X, Q_{Y|X}, \nu, c) \).

For general alphabets, we can extend the idea and let \( \mu_n = Q_X^\otimes n \mid c_n \) for some \( C_n \) with \( 1 - Q_X^\otimes n[C_n] \leq \delta \) for some \( \delta \in (0, 1) \) independent of \( n \); this will not affect its information-theoretic applications in the non-vanishing error probability regime. In Chapter 3 we have shown, in the discrete and the Gaussian cases, that we can choose \( C_n \) so that

\[
d(\mu_n, Q_{Y|X}^\otimes n, \nu^\otimes n, c) \leq n d^*(Q_X, Q_{Y|X}, \nu, c) + O(\sqrt{n}). \quad (4.84)
\]

Using the semigroup method, we can also improve the \( O(\sqrt{n} \ln \frac{3}{2} n) \) term in (4.82) to \( O(\sqrt{n}) \) in the discrete and the Gaussian cases. This combined with (4.84) implies

\[
\ln \mu_n[x^n: Q_{Y^n|X^n = x^n}[A] > 1 - c] - c \ln \nu^\otimes n[A] \\
\leq n d^*(Q_X, Q_{Y|X}, \nu, c) + O(\sqrt{n}). \quad (4.85)
\]

---

The restriction \( \mu \mid \mathcal{C} \) of a measure \( \mu \) on a set \( \mathcal{C} \) is \( \mu \mid \mathcal{C}[D] := \mu(\mathcal{C} \cap D) \).
While the original proof [27] used a data processing argument, just as (4.2), to get (4.74), the present approach uses the functional inequality (4.73), just as (4.26). In the discrete case we have the following result:

**Theorem 4.3.1** Consider \( Q_X \) a probability measure on a finite set \( \mathcal{X} \), \( \nu \) a probability measure on \( \mathcal{Y} \), and \( Q_Y|X \). Let

\[
\beta_X := 1/ \min_x Q_X(x) \in [1, \infty), \tag{4.86}
\]

\[
\alpha := \sup_x \left\| \frac{dQ_Y|X=x}{d\nu} \right\| \in [1, \infty). \tag{4.87}
\]

Let \( c \in (0, \infty) \), \( \eta, \delta \in (0, 1) \) and \( n > 3\beta_X \ln \frac{|\mathcal{X}|}{\delta} \). We can choose some set \( \mathcal{C}_n \) with \( Q_X^n[\mathcal{C}_n] \geq 1 - \delta \), such that for \( \mu_n := Q_X^n|_{\mathcal{C}_n} \) we have

\[
\ln \mu_n[x^n: Q_{Y^n|X^n=x^n}(f) \geq \eta] - c \ln \nu^n(f) \\
\leq nd^*(Q_X, Q_Y|X, \nu, c) + A\sqrt{n} + c\ln \frac{1}{\eta} \tag{4.88}
\]

for any \( f \in \mathcal{H}_{[0,1]}(\mathcal{Y}^n) \), where

\[
A := \ln(\alpha^c \beta_X^{c+1}) \sqrt{3\beta_X \ln \frac{|\mathcal{Y}|}{\delta} + 2c(\alpha - 1) \ln \frac{1}{\eta}}. \tag{4.89}
\]

The proof of Theorem 4.3.1 relies on some ideas in the proof of Theorem 4.2.2 in particular the properties (4.47) and (4.50) for the operator \( \Lambda_{\alpha,t}^\nu \) play a critical role.

**Proof** We first establish the following claim: for any \( n \geq 1, f \in \mathcal{H}_{[0,1]}(\mathcal{Y}^n) \) and \( \mu_n \) a measure on \( \mathcal{X}^n \),

\[
\ln \mu_n[x^n: Q_{Y^n|X^n=x^n}(f) \geq \eta] - c \ln \nu^n(f) \\
\leq d(\mu_n, Q_{Y^n|X} \nu^n, c) + 2c(\alpha - 1)n \ln \frac{1}{\eta} + c\ln \frac{1}{\eta}, \tag{4.90}
\]
where $Q_{Y^n|X^n} := Q_{Y^n|X}^{\otimes n}$. Notice that for any $g \in \mathcal{H}_{+}(Y^n)$, from the variational formula of the relative entropy \[4.26\] we have

\[
\int \|g\| L^0(Q_{Y^n|X^n}) d\mu_n(x^n) = \int e^{Q_{Y^n|X}^{\otimes n}(\ln g)} d\mu_n
\]

\[
= e^{-D(P_{X^n} \| \mu_n) +} Q_{Y^n|X}^{\otimes n}(\ln g) dP_{X^n}
\]

\[
\leq e^{d(\mu_n, Q_{Y^n|X}^{\otimes n} \| \nu^{\otimes n}, c)} - cD(P_{Y^n} \| \nu^{\otimes n}) + cP_{X^n}(\ln g^{\frac{1}{c}})
\]

\[
\leq e^{d(\mu_n, Q_{Y^n|X}^{\otimes n} \| \nu^{\otimes n}, c)} \|g\|_{L^{1/c}(\nu^{\otimes n})}.
\]

where we defined $P_{X^n}$ via $dP_{X^n} = e^{Q_{Y^n|X}^{\otimes n}(\ln g)} \left( \int e^{Q_{Y^n|X}^{\otimes n}(\ln g)} d\mu_n \right)^{-1}$. Note that \[4.94\] can also be recovered as a particular instance of \[80, \text{Theorem 1}\]. Now take $g = (\Lambda_{\alpha, t} f)^c$ and observe that by \[4.47\],

\[
\|g\|_{L^{1/c}(\nu^{\otimes n})} \leq e^{c(\alpha - 1)nt [\nu^{\otimes n}(f)]^c}.
\]

Moreover,

\[
\int \|g\| L^0(Q_{Y^n|X^n}) d\mu_n(x^n) = \int |\Lambda_{\alpha, t} f| L^0(Q_{Y^n|X^n}) d\mu_n(x^n)
\]

\[
\geq \int \|f\|_{L^{1 - e^{-t}}(Q_{Y^n|X^n})} d\mu_n(x^n)
\]

\[
\geq \int Q_{Y^n|X^n}(f)^{1 - \frac{e^{-t}}{c}} d\mu_n(x^n)
\]

\[
\geq \eta^{1 - \frac{e^{-t}}{c}} \mu_n[Q_{Y^n|X^n}(f) \geq \eta].
\]

where \[4.97\] used Lemma \[4.2.1\] and Theorem \[4.1.3\] and \[4.98\] used the assumption $f \leq 1$. Then \[4.90\] follows by using

\[
\frac{1}{1 - e^{-t}} \leq \frac{1}{t} + 1 \quad \forall t \in (0, \infty)
\]

and optimizing over $t > 0$.  

119
Having established (4.90), the rest of the proof follows from the estimate of \(d(\mu_n, Q_{Y|X}^n, \nu^\otimes n, c)\) in Lemma 3.2.2 noting that the \(\alpha_Y\) defined therein satisfies \(\alpha_Y \leq \alpha\).

We now proceed to find an analogue of Theorem 4.3.1 in the case of Gaussian \(Q_X, Q_{Y|X}\) and Lebesgue \(\nu\). Again we consider the Ornstein-Uhlenbeck semigroup. The proof of the following uses some ideas already touched in Section 4.2.2, in particular the change-of-variable trick in (4.61).

**Theorem 4.3.2** Suppose \(Q_X = \mathcal{N}(0, \sigma^2)\), \(Q_{Y|X=x} = \mathcal{N}(x, 1)\), \(\nu\) is Lebesgue on \(\mathbb{R}\), and \(a, \delta \in (0, 1)\). For \(n \geq 24 \ln \frac{2}{\delta}\), there exists \(C_n \subseteq \mathbb{R}^n\) with \(Q_X^\otimes n[C_n] \geq 1 - \delta\) such that for \(\mu_n = Q_X^\otimes n|_{C_n}\),

\[
\ln \mu_n[x^n: Q_{Y^n|X^n=x^n}(f) > \eta] - c \ln \nu^\otimes n(f) + nd^*(Q_X, Q_{Y|X}, \nu, c) + c \sqrt{2n \ln \frac{1}{\eta}} + \sqrt{6n \ln \frac{2}{\delta}} + c \ln \frac{1}{\eta}\]

(4.101)

for any \(f \in \mathcal{H}_{[0,1]}(\mathbb{R}^n)\).

Note that the value of \(\sigma\) does not affect the bound.

**Proof** We first prove the following claim: for any nonnegative finite measure \(\mu_n\) on \(\mathbb{R}^n\),

\[
\ln \mu_n[x^n: Q_{Y^n|X^n=x^n}(f) > \eta] - c \ln \nu^\otimes n(f) + d(\mu_n, Q_{Y|X}^n, \nu^\otimes n, c) + c \sqrt{2 \left( \ln \frac{1}{\eta} \right) n} + c \ln \frac{1}{\eta}\]

(4.102)

for any \(f \in \mathcal{H}_{[0,1]}(\mathbb{R}^n)\).
Let $\bar{\mu}_n$ be the dilation of $\mu_n$ by factor $e^t$, that is,

$$\bar{\mu}_n[e^tC] = \mu_n[C],$$  \hfill (4.103)

for any measurable $C$. As before, the functional inequality gives

$$\int e^{Q_{Y|X}^{\otimes n}(\ln g)} d\bar{\mu}_n \leq \exp(d(\bar{\mu}_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c)) \|g\|_{L^{1/c}(\mathbb{R}^n)}, \quad \forall g \geq 0.$$  \hfill (4.104)

Now substitute with

$$g = (T_{0^n,t}f)^c.$$  \hfill (4.105)

Observe that

$$\|g\|_{L^{1/c}(\mathbb{R}^n)} = \|T_{0^n,t}f\|_{L^1(\mathbb{R}^n)}^c = e^{cnt} \|f\|_{L^1(\mathbb{R}^n)}^c$$  \hfill (4.106)

and

$$= e^{cnt} [\mu^{\otimes n}(f)]^c.$$  \hfill (4.108)
where (4.107) can be verified using Fubini’s theorem and (4.60).

\[
\int e^{Q_{Y^n|X}^n(\ln g)} d\bar{\mu}_n = \int e^{cQ_{Y^n|X^n=x^n}^{n}(\ln T_{y^n,T^n} f)} d\bar{\mu}_n(x^n) \\
= \int e^{cQ_{Y^n|X^n=e^{-tx^n}}^{n}(\ln T_{e^{-tx^n},t} f)} d\bar{\mu}_n(x^n) \\
= \int \|T_{e^{-tx^n},t} f\|_{L^0(Q_{Y^n|X^n=e^{-tx^n}})} d\bar{\mu}_n(x^n) \\
\geq \int \|f\|_{L^{1-e^{-2t}}(Q_{Y^n|X^n=e^{-tx^n}})} d\bar{\mu}_n(x^n) \\
\geq \eta^{1-e^{-2t}} \mu_n[x^n: Q_{Y^n|X^n=e^{-tx^n}}(f) > \eta] \\
= \eta^{1-e^{-2t}} \mu_n[x^n: Q_{Y^n|X^n=x^n}(f) > \eta]
\]  

(4.110) used (4.61); (4.112) is from (4.30); (4.113) used the fact that \( f \leq 1 \); (4.115) is from (4.103). Combining (4.104), (4.108), (4.115) we obtain

\[
\ln \mu_n[x^n: Q_{Y^n|X^n=x^n}(f) > \eta] - c \ln \nu^{\otimes n}(f) \\
\leq d(\bar{\mu}_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) + \inf_{t>0} \left\{ \left( \ln \frac{1}{\eta} \right) \frac{c}{1 - e^{-2t}} + c n t \right\} \\
\leq d(\bar{\mu}_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) + c \sqrt{2 \left( \ln \frac{1}{\eta} \right) n + c \ln \frac{1}{\eta}}
\]

(4.116) (4.117)

where the last step used (4.100). Using an argument similar to (4.71), we find

\[
d(\bar{\mu}_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) \leq d(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c).
\]

(4.118)

Indeed, by definition

\[
d(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) = \sup_{P_{X^n}} \{ c D(P_{Y^n} || \nu^{\otimes n}) - D(P_{X^n} || \mu_n) \}.
\]

(4.119)
When $P_{X^n}$ and $\mu_n$ are both scaled by $e^t$, $D(P_{X^n}\|\mu_n)$ remains unchanged. On the other hand $D(P_{Y^n}\|\nu^{\otimes n}) = -h(P_{Y^n})$, so by the same argument as (4.71) we have $D(P_{Y^n}\|\nu^{\otimes n}) \leq D(P_{Y^n}\|\nu^{\otimes n})$. This establishes (4.118). Combining (4.117), (4.118), we have established the claim (4.102). The rest of the proof of Theorem 4.3.2 follows from the bound on $d(\mu_n, Q^{\otimes n}_{Y|X}, \nu^{\otimes n}, c)$ in Lemma 3.2.6.

Remark 4.3.1 If $\nu = N(0, \tilde{\sigma}^2)$ ($\tilde{\sigma} \geq 1$) instead, then an $O(\sqrt{n})$ second-order bound still holds. Indeed, in this case notice that

\[ \|T_{0^n,t}f\|_{L^1(\nu^n)} = \mathbb{E} \left[ f(\nu^n) \right] \]

where $V^n \sim N(0^n, I_n)$ and $W^n \sim N(0^n, \tilde{\sigma}^2 I_n)$ are independent, and $U^n \sim N(0^n, (e^{-2t} \tilde{\sigma}^2 + 1 - e^{-2t}) I_n)$. Therefore the equality in (4.107) will be replaced by $\leq$, so the analysis can still go through. On the other hand, the third term on the right side of (4.101), which is the approximation error for $d(\mu_n, Q^{\otimes n}_{Y|X}, \nu^{\otimes n}, c)$, will be slightly changed, but is still of the order of $n^{1/2}$. As for (4.118), notice that

\[ D(P_{Y^n}\|\nu^{\otimes n}) - D(P_{Y^n}\|\nu^{\otimes n}) \leq \frac{1}{2\tilde{\sigma}^2} \mathbb{E}[\|Y^n\|^2] - \frac{1}{2\tilde{\sigma}^2} \mathbb{E}[\|Y^n\|^2] \]

\[ = \frac{1}{2\tilde{\sigma}^2} \mathbb{E}[\|X^n\|^2] - \frac{1}{2\tilde{\sigma}^2} \mathbb{E}[\|X^n\|^2] \]

\[ = (e^{2t} - 1) \frac{1}{2\tilde{\sigma}^2} \mathbb{E}[\|X^n\|^2]. \]
Therefore if \( \mu_n \) (hence \( P_{X^n} \)) is supported on \( \{ x^n : \| x^n \| \leq O(\sqrt{n}) \} \) then

\[
d(\mu_n, \mathcal{Q}^{\otimes n}_{Y|X}, \nu^{\otimes n}, c) \leq d(\mu_n, \mathcal{Q}^{\otimes n}_{Y|X}, \nu^{\otimes n}, c) + O(\sqrt{n}) \tag{4.129}
\]

as long as \( t = \Theta(n^{-1/2}) \).

## 4.4 Applications

### 4.4.1 Output Distribution of Good Channel Codes

Consider a stationary memoryless channel \( P_{Y|X} \) with capacity \( C \). If \( |\mathcal{Y}| < \infty \) then by the steps \cite[64)-(66)]{149} and our Theorem \ref{theorem4.2.2} we conclude that an \( (n, M, \epsilon) \) code under the maximal error criterion with deterministic encoders satisfies the converse bound

\[
D(P_{Y^n} || P_{Y^n}^*) \leq nC - \ln M + 2\sqrt{|\mathcal{Y}|n \ln \frac{1}{1 - \epsilon}} + \ln \frac{1}{1 - \epsilon}. \tag{4.130}
\]

This implies that the Burnashev condition

\[
\sup_{x, x'} \left\| \frac{dP_{Y|X=x}}{dP_{Y|X=x'}} \right\|_{\infty} < \infty \tag{4.131}
\]

in \cite[Theorem 6]{149} (cf. \cite[Theorem 3.6.6]{48}) is not necessary. Using the blowing-up lemma, \cite[Theorem 7]{149} bounded \( D(P_{Y^n} || P_{Y^n}^*) \) without requiring \( \mathbf{(4.131)} \), but with a suboptimal \( \sqrt{n}(\ln n)^{3/2} \) second-order term. Also note that in our approach \( |\mathcal{Y}| < \infty \) can be weakened to a bounded density assumption \( \mathbf{(4.48)} \), and the maximal error criterion assumption can be weakened to the geometric average criterion \( \mathbf{(4.52)} \).
4.4.2 Broadcast Channels

Consider a Gaussian broadcast channel where the SNRs in the two component channels are $S_1, S_2 \in (0, \infty)$. Suppose there exists an $(n, M_1, M_2, \epsilon)$-maximal error code. Using Theorem 4.2.3 and the same steps in the proof of the weak converse (see e.g. [18, Theorem 5.3]), we immediately obtain

$$\ln M_1 \leq n C(\alpha S_1) + \sqrt{2n \ln \frac{1}{1-\epsilon} + \ln \frac{1}{1-\epsilon}}; \quad (4.132)$$

$$\ln M_2 \leq n C\left(\frac{(1-\alpha)S_2}{\alpha S_2 + 1}\right) + \sqrt{2n \ln \frac{1}{1-\epsilon} + \ln \frac{1}{1-\epsilon}}, \quad (4.133)$$

for some $\alpha \in [0, 1]$, where $C(t) := \frac{1}{2} \ln(1 + t)$. An alternative proof of the strong converse via information spectrum, which does not give simple explicit bounds on the sublinear terms, is given in [152]. Convereses under the average error criterion can be obtained by codebook expurgation (e.g. [18, Problem 8.11]).

Next, let us prove an $O(\sqrt{n})$ second-order counterpart for the discrete broadcast channel. Consider a degraded broadcast channel $(P_{Y|X}, P_{Z|X})$ with $|Y|, |Z| < \infty$. Suppose that there exists an $(n, M_1, M_2, \epsilon)$ code under the maximal error criterion. We can generalize Theorem 4.2.2 to allow stochastic encoders as in Theorem 4.2.3 which, combined with the steps in the proofs of the weak converse of degraded broadcast channels (e.g. [18, Theorem 5.2]) show that there exist $P_{UX}$ such that

$$\ln M_1 \leq n I(X; Y|U) + 2 \sqrt{|Y|n \ln \frac{1}{1-\epsilon} + \ln \frac{1}{1-\epsilon}}; \quad (4.134)$$

$$\ln M_2 \leq n I(U; Z) + 2 \sqrt{|Z|n \ln \frac{1}{1-\epsilon} + \ln \frac{1}{1-\epsilon}}. \quad (4.135)$$

This improves the best $O(\sqrt{n \ln \frac{3}{2} n})$ second term obtained by the blowing-up argument ([27, 18, Theorem 3.6.4]). Again, with our approach the finite alphabet assumption may be weakened to (4.48). Converse bounds under the average error criterion can be obtained by an expurgation argument (e.g. [18, Problem 8.11]).
4.4.3 Source Coding with Compressed Side Information

Figure 4.4: Source coding with compressed side information

As alluded in Chapter 1, source coding with compressed side information [49][27] is a canonical example of the side-information problem, whose second-order converse was open [50]. See Figure 4.4 for the setup. Using Theorem 4.3.1 plus essentially the same arguments as [27, Theorem 3], we can show the following second-order converse:

**Theorem 4.4.1** Consider a stationary memoryless discrete source with per-letter distribution $Q_{XY}$. Let $\epsilon \in (0,1)$ and $n \geq 3\beta_X \ln \frac{2(1+\epsilon)|X|}{1-\epsilon}$, where $\beta_X$ is defined in (4.86). Suppose that there exist encoders $f: X^n \to W_1$ and $g: Y^n \to W_2$ and decoder $V: W_1 \times W_2 \to \hat{Y}^n$ such that

$$\mathbb{P}[Y^n \neq \hat{Y}^n] \leq \epsilon. \quad (4.136)$$

Then for any $c \in (0, \infty)$,

$$\ln |W_1| + c \ln |W_2|$$

$$\geq n \inf_{U: U \rightarrow X \rightarrow Y} \{cH(Y|U) - I(U;X)\}$$

$$- \sqrt{n} \left( \ln(|Y|^c\beta_X^{c+1}) + 3\beta_X \ln \frac{4|X|}{1-\epsilon} + 2c \sqrt{|Y| \ln \frac{2}{1-\epsilon}} \right)$$

$$- (1+c) \ln \frac{2}{1-\epsilon}. \quad (4.137)$$
Proof The first part of the proof parallels the proof of [27, Theorem 3]. For any $w \in \mathcal{W}_1$, define the “correctly decodable set”

$$
\mathcal{B}_w := \{ y^n : y^n = V(w, g(y^n)) \}.
$$

(4.138)

Then

$$
\mathbb{E}[Q_{Y^n|X^n}[\mathcal{B}_f(X^n)|X^n]] \geq 1 - \epsilon,
$$

(4.139)

where $X^n \sim Q^\otimes_X$. Choose any $\epsilon' \in (\epsilon, 1)$. By the Markov inequality,

$$
Q^\otimes_X[x^n : Q_{Y^n|X^n=x^n}[\mathcal{B}_f(x^n)] \geq 1 - \epsilon'] \geq 1 - \frac{\epsilon}{\epsilon'}.
$$

(4.140)

Take $\nu$ to be the equiprobable measure on $\mathcal{Y}$. Applying Theorem 4.3.1 with $\delta = \frac{\epsilon' - \epsilon}{2\epsilon'}$ and $\eta = 1 - \epsilon'$, for $n > 3\beta_X \ln \frac{2|\mathcal{X}|}{1-\epsilon/\epsilon'}$, we find $\mu_n$ such that

$$
\mu_n [x^n : Q_{Y^n|X^n=x^n}[\mathcal{B}_f(x^n)] \geq 1 - \epsilon'] \geq 1 - \frac{\epsilon}{\epsilon'} - \delta
$$

(4.141)

$$
= \frac{\epsilon' - \epsilon}{2\epsilon'}
$$

(4.142)

and

$$
\ln \mu_n [x^n : Q_{Y^n|X^n=x^n}[\mathcal{B}_f(x^n)] > a] - c \ln \nu^\otimes [\mathcal{B}_f(x^n)]
$$

$$
\leq nd^*(Q_X, Q_{Y|X}, \nu, c) + A\sqrt{n} + c \ln \frac{1}{\eta}.
$$

(4.143)

Since $f$ takes at most $|\mathcal{W}_1|$ possible values, from (4.142), there exists a $w^*$ such that

$$
\mu_n [x^n : Q_{Y^n|X^n=x^n}[\mathcal{B}_{w^*}] \geq 1 - \epsilon'] \geq \frac{\epsilon' - \epsilon}{2\epsilon'} |\mathcal{W}_1|^{-1}.
$$

(4.144)
Moreover, since $B_{w^*} \subseteq \bigcup_{w \in W_2} \{y^n : y^n = V(w^*, w)\}$, by the union bound,

$$
\nu^n[B_{w^*}] = |Y|^{-n} |B_{w^*}| \leq |W_2||Y|^{-n}.
$$

(4.145)

Comparing (4.143), (4.144) and (4.145), we obtain

$$
\ln |W_1| + c \ln |W_2| \\
\geq -n \sup_{P_{UX} : P_X = Q_X} \{cD(P_Y|v|P_U) - I(U; X)\} + nc \ln |Y| - A\sqrt{n} - c \ln \frac{1}{\eta} + \ln \frac{1 - \epsilon/\epsilon'}{2}
$$

(4.146)

$$
\geq n \inf_{U : U \to X \to Y} \{cH(Y|U) - I(U; X)\} - A\sqrt{n} - c \ln \frac{1}{1 - \epsilon'} - \ln \frac{2}{1 - \epsilon/\epsilon'}
$$

(4.147)

(4.137) then follows by taking $\epsilon' = \frac{1 + \epsilon}{2}$.

Note that the first term on the right side of (4.137) corresponds to the rate region (see e.g. [18, Theorem 10.2]). Using the BUL method [27], on the other hand, one can only bound the second term as $O(\sqrt{n} \ln \frac{3}{2} n)$, which is suboptimal.

A counterpart for Gaussian sources under the quadratic distortion is the following (which is a special case of the Berger-Tung problem), for which we also derive the $O(\sqrt{n})$ second-order converse. The setup is still as in Figure 4.4, except that we seek to approximate $Y$ by $\hat{Y}$ in quadratic error, instead of seeking exact reconstruction.

**Theorem 4.4.2** Consider stationary memoryless Gaussian sources where per-letter distribution $Q_{XY}$ is jointly Gaussian with covariance matrix $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Suppose $\epsilon \in (0, 1)$ and $n \geq 24 \ln \frac{4(1+\epsilon)}{1-\epsilon}$, and that there exist encoders $f : X^n \to W_1$ and $g : Y^n \to W_2$ and decoder $V : W_1 \times W_2 \to \hat{Y}_n$ ($X = Y = \hat{Y} = \mathbb{R}$) such that for some $D \in (0, \infty)$,

$$
P[\|Y^n - \hat{Y}^n\|^2 > nD] \leq \epsilon.
$$

(4.148)
Then for any $c \in (0, \infty)$,

$$\ln |W_1| + c \ln |W_2| \geq n \inf_{\rho \in [0,1]} \left\{ \frac{c}{2} \ln \frac{1 - \rho^2 \tilde{\rho}^2}{D} + \frac{1}{2} \ln \frac{1}{1 - \tilde{\rho}^2} \right\}$$

$$- c \sqrt{2n \ln \frac{2}{1 - \epsilon}} - \sqrt{6n \ln \frac{4(1 + \epsilon)}{1 - \epsilon}}$$

$$- c \ln \frac{2}{1 - \epsilon} - \ln \frac{2(1 + \epsilon)}{1 - \epsilon}.$$  \hspace{1cm} (4.149)

**Proof** The first part of the proof parallels the proof of [27, Theorem 3], with adaptations to include the distortion function. For any $w \in W_1$, define the “correctly decodable set”

$$B_w := \{ y^n : \| y^n - V(w, g(y^n)) \|^2 \leq nD \}.$$

Then

$$\mathbb{E}[Q_{Y^n|X^n}[B_{f(X^n)}|X^n]] \geq 1 - \epsilon,$$  \hspace{1cm} (4.151)

where $X^n \sim Q_{X}^{\otimes n}$. Choose any $\epsilon' \in (\epsilon, 1)$. By the Markov inequality,

$$Q_{X}^{\otimes n}[x^n : Q_{Y^n|X^n=x^n}[B_{f(x^n)}] \geq 1 - \epsilon'] \geq 1 - \frac{\epsilon}{\epsilon'}.$$  \hspace{1cm} (4.152)

Applying Theorem 4.3.2 with $\delta = \frac{\epsilon - \epsilon'}{2\epsilon'}$ and $\epsilon' = 1 - \eta$, for $n \geq 24 \ln \frac{4}{1 - \epsilon/\epsilon'}$, we find $\mu_n$ such that

$$\mu_n[x^n : Q_{Y^n|X^n=x^n}[B_{f(x^n)}] \geq 1 - \epsilon'] \geq 1 - \frac{\epsilon}{ \epsilon'} - \eta = \frac{\epsilon' - \epsilon}{2\epsilon'}.$$  \hspace{1cm} (4.153)
and

\[
\ln \mu_n[x^n] : Q_{Y^n|X^n=x^n}^n[B_{f(x^n)}] > 1 - \epsilon' - c \ln \nu^n[B_{f(x^n)}] \\
\leq nd^*(Q_X, Q_{Y|X}, \nu, c) + c \sqrt{2n \ln \frac{1}{1-\epsilon'}} + \sqrt{6n \ln \frac{4\epsilon'}{\epsilon' - \epsilon}} + c \ln \frac{1}{1-\epsilon'},
\]

(4.154)

Since \( f \) takes at most \( |W_1| \) possible values, from (4.153), there exists a \( w^* \) such that

\[
\mu_n[x^n] : Q_{Y^n|X^n=x^n}^n[B_{w^*}] > 1 - \epsilon' \geq \frac{\epsilon' - \epsilon}{2\epsilon'} |W_1|^{-1}.
\]

(4.155)

Moreover, since \( B_{w^*} \subseteq \bigcup_{w \in W_2} \{y^n : \|y^n - V(w^*, w)\|^2 \leq nD\} \), by the union bound and the volume estimate of the \( n \)-dimensional ball,

\[
\nu^n[B_{w^*}] \leq |W_2|(2\pi e D)^{\frac{n}{2}}.
\]

(4.156)

Comparing (4.154), (4.155) and (4.156), we obtain

\[
\ln |W_1| + c \ln |W_2| \geq n \inf_{\rho \in [0,1]} \left\{ \frac{c}{2} \ln \frac{1 - \rho^2 \bar{\rho}^2}{D} + \frac{c}{2} \ln \frac{1}{1 - \bar{\rho}^2} \right\} \\
- c \sqrt{2n \ln \frac{1}{1-\epsilon'}} - \sqrt{6n \ln \frac{4\epsilon'}{\epsilon' - \epsilon}} \\
- c \ln \frac{1}{1-\epsilon'} - \ln \frac{2\epsilon'}{\epsilon' - \epsilon}.
\]

(4.157)

(4.149) then follows by taking \( \epsilon' = \frac{1+\epsilon}{2} \).

Remark: The weak converse version of (4.149) is the following result, which is a special case of [18, Theorem 12.3]. The first term on the right side of (4.149) corresponds to the known (first-order) single-letter region which can be obtained by taking \( D_2 \to \infty \) in [18, Theorem 12.3].

Theorem 4.4.3 (see [18, Theorem 12.3] and the references therein) Consider lossy source coding with compressed side information where the per-letter distribution
is jointly Gaussian with covariance matrix \( \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \) and the distortion function is quadratic. Then \((R_1, R_2, D) \in [0, \infty)^2\) is achievable if and only if

\[
R_2 \geq \frac{1}{2} \ln \frac{1 - \rho^2 + \rho^2 \exp(-2R_1)}{D}.
\] (4.158)

4.4.4 One-communicator CR Generation

We are now ready (finally!) to give a \(O(\sqrt{n})\) second-order converse to the one-communicator CR generation problem in Section 2.3.1. Armed with the optimal second-order change-of-measure lemma of the present chapter, we now only need the following result relating the change-of-measure lemma to the performance of the CR generation. For simplicity \(m = 1\) is assumed, while the method should carry over to the case of arbitrarily many receivers.

**Lemma 4.4.4** Consider the single-shot setup as depicted in Figure 1.1 with \(m = 1\). Suppose that there exist \(\delta_1, \delta_2 \in (0, 1)\), a stochastic encoder \(Q_{W|X}\), and deterministic decoders \(Q_{K|X}\) and \(Q_{\hat{K}|WY}\), such that

\[
\mathbb{P}[K = K_1] \geq 1 - \delta_1; \quad (4.159)
\]
\[
\frac{1}{2}|Q_K - T_k| \leq \delta_2, \quad (4.160)
\]

where \(T_K\) denotes the equiprobable distribution. Also, suppose that there exist \(\mu_X\), \(\delta, \epsilon' \in (0, 1)\) and \(c \in (1, \infty), d \in (0, \infty)\) such that

\[
E_1(Q_X) \leq \delta; \quad (4.161)
\]
\[
\mu_X \left(x: Q_{Y|x}(A) \geq 1 - \epsilon' \right) \leq 2^c \exp(d)Q_Y(A) \quad (4.162)
\]
for any $A \subseteq \mathcal{Y}$. Then, for any $\delta_3, \delta_4 \in (0, 1)$ such that $\delta_3 \delta_4 = \delta_1 + \delta$, we have

$$\delta_2 \geq 1 - \delta - \delta_3 - \frac{1}{|\mathcal{K}|} - \frac{2 \exp \left( \frac{d}{\epsilon'} \right) |\mathcal{W}|}{(\epsilon' - \delta_4)^{\frac{1}{2}} |\mathcal{K}|^{1 - \frac{1}{2}}}.$$  \hspace{1cm} (4.163)

**Proof** As we assumed $m = 1$, let us write $\hat{K} := K_1$. Define the joint measure

$$\mu_{XYWK\hat{K}} := \mu_XQ_Y|XQ_W|XQ_K|XQ_{\hat{K}}|YW$$

which we sometimes abbreviate as $\mu$. Since $E_1(Q\|\mu) = E_1(Q_X\|\mu_X) \leq \delta$, \hspace{1cm} (4.159)

implies

$$\mu(K \neq \hat{K}) \leq \delta_1 + \delta.$$  \hspace{1cm} (4.165)

Put

$$\mathcal{J} := \{ k: \mu_{\hat{K}|K}(k|k) \geq 1 - \delta_4 \}.$$  \hspace{1cm} (4.166)

The Markov inequality implies that $\mu_{K}(\mathcal{J}) \geq 1 - \delta_3$. Now for each $k \in \mathcal{J}$, we have

$$\begin{align*}
(1 - \delta_4)\mu_k(k) \\
\leq \mu_{K\hat{K}}(k; k) \\
\leq \int_{\mathcal{F}_k} Q_{Y|X=x} \left( \bigcup_w \mathcal{A}_{wk} \right) d\mu_X(x) \\
\leq (1 - \epsilon')\mu_{K}(k) + \mu \left( x: Q_{Y|X=x} \left( \bigcup_w \mathcal{A}_{kw} \geq 1 - \epsilon' \right) \right) \\
\leq (1 - \epsilon')\mu_{K}(k) + 2^c \exp(d)Q_{\hat{Y}} \left( \bigcup_w \mathcal{A}_{kw} \right),
\end{align*}$$

\hspace{1cm} (4.167-4.170)
where $F_k \subseteq \mathcal{X}$ is the decoding set for $K$, and $A_{kw}$ is the decoding set for $\hat{K}$ upon receiving $w$. Rearranging,

$$(\epsilon' - \delta_4)^{\frac{1}{c}} \mu_{\hat{K}}(k) \leq 2 \exp \left( \frac{d}{c} \right) Q_Y \left( \bigcup_w A_{kw} \right).$$

(4.171)

Now let $\bar{\mu}$ be the restriction of $\mu_K$ on $J$. Then summing both sides of (4.171) over $k \in J$, applying the union bound, and noting that $\{A_{kw}\}_k$ is a partition of $\mathcal{Y}$ for each $w$, we obtain

$$D_1(\bar{\mu} \| T) \leq \log |K| - \frac{1}{1 - \frac{1}{c}} \log \frac{2|\mathcal{W}|}{(\epsilon' - \delta_4)^{\frac{1}{c}}}$$

$$- \frac{d}{c - 1}.$$  (4.172)

The proof is completed invoking Proposition 4.4.5 below and noting that

$$E_1(Q_K \| \mu) \leq E_1(Q_K \| \mu) + E_1(\mu \| \bar{\mu}) \leq \delta + \delta_3.$$  (4.173)

**Proposition 4.4.5** Suppose $T$ is equiprobable on $\{1, \ldots, M\}$ and $\mu$ is a nonnegative measure on the same alphabet. For any $\alpha \in (0, 1),

$$E_1(T \| \mu) \geq 1 - \frac{1}{M} - \exp(- (1 - \alpha) D_{\alpha}(T \| \mu)).$$

(4.174)

The special case of Proposition 4.4.5 when $\mu$ is a probability measure was proved in the end of the proof of Theorem 2.3.2. The extension to unnormalized $\mu$ can be easily proved in a similar way.

Now specializing the single-shot result of Lemma 4.4.4 to the discrete memoryless case, and using Theorem 4.3.1, we obtain the following non-asymptotic second-order
converse to the CR generation problem. Note that the right side of (4.181) is essentially
\[
\frac{\exp\left(\frac{nA^*}{c} + \frac{A}{\varepsilon} \sqrt{n}\right) |W|}{|\mathcal{K}|^{1 - \frac{1}{\varepsilon}}} \tag{4.175}
\]
where the $\sqrt{n}$ term characterizes the correct second-order coding rate.

**Theorem 4.4.6** Consider a stationary memoryless source with per-letter distribution $Q_{XY}$. Let
\[
\beta_X := \frac{1}{\min_x Q_X(x)} Q_X(x) \in [1, \infty); \tag{4.176}
\]
\[
\alpha := \sup_x \frac{d Q_{Y|X=x}}{d Q_Y} \in [1, \infty); \tag{4.177}
\]
\[
c \in (0, \infty). \tag{4.178}
\]
Suppose that there exist a CR generation scheme (see the setup in Figure 1.1) for which
\[
\mathbb{P}[K = K_1] \geq 1 - \delta_1 \in (0, 1); \tag{4.179}
\]
\[
\frac{1}{2} |Q_K - T_k| \leq \delta_2 \in (0, 1). \tag{4.180}
\]
Then for any $\delta'_1 \in (\delta_1, 1)$ and $n \geq 3\beta_X \ln \frac{3|\mathcal{X}|}{\delta'_1 - \delta_1}$, we have
\[
1 - \delta_2 - \delta'_1 \leq \frac{1}{|\mathcal{K}|} + \frac{\exp\left(\frac{nA^*}{c} + \frac{A}{\varepsilon} \sqrt{n} + \ln \frac{4\delta'_1 + 2\delta_1}{\delta'_1 - \delta_1}\right) |W|}{\left(\frac{\delta'_1 - \delta_1}{4\delta'_1 + 2\delta_1}\right)^{\frac{1}{\varepsilon}} |\mathcal{K}|^{1 - \frac{1}{\varepsilon}}}, \tag{4.181}
\]
where we have defined

\[ A := \ln(\alpha^c \beta_X^{c+1}) \sqrt{\frac{3 \beta_X \ln 3 |X|}{\delta_1 - \delta_1}} + 2c \sqrt{(\alpha - 1) \ln \frac{4\delta'_1 + 2\delta_1}{\delta_1 - \delta_1}}, \quad (4.182) \]

\[ d^* := d^*(Q_X, Q_{Y|X}, Q_Y, c). \quad (4.183) \]

**Proof** Specialize Lemma 4.4.4 to the stationary memoryless setting:

\[ Q_{XY} \leftarrow Q^\otimes_{XY}. \quad (4.184) \]

Suppose we first pick the parameter \( \delta_3 \in (\delta_1, 1) \) in Lemma 4.4.4 arbitrarily, which in turn yields the following choices of the other parameters therein:

\[ \delta = \frac{\delta_3 - \delta_1}{2}; \quad (4.185) \]

\[ \delta_4 = \frac{\delta_1 + \delta_3}{2\delta_3}. \quad (4.186) \]

Moreover, let \( \epsilon' = \frac{1 + \delta_1}{2} \) in Lemma 4.4.4. Then we invoke Theorem 4.3.1 with \( \nu = Q_Y \) and

\[ \eta = 1 - \epsilon'. \quad (4.187) \]

Then (4.163) implies that

\[ 1 - \delta_2 - \frac{3\delta_3 - \delta_1}{2} \leq \frac{1}{|K|} + \exp \left( \frac{nd^*}{c} + \frac{A}{c} \sqrt{n} + \ln \frac{4\delta_1'}{\delta_3 - \delta_1} \right) |W| \left( \frac{\delta_3 - \delta_1}{4\delta_1'} \right)^{\frac{1}{2}} |K|^{1 - \frac{1}{c}} \quad (4.188) \]

which is equivalent to (4.181) upon setting \( \delta'_1 = \frac{3\delta_3 - \delta_1}{2}. \)
4.5 About Extensions to Processes with Weak Correlation

The majority of this chapter focuses on the stationary memoryless settings, and in particular, we succeeded in using the pumping-up approach to determine that the second-order rate is the square root of the blocklength in such settings for many problems from the network information theory. However, the potential of the pumping-up approach is not limited to the stationary memoryless setting.

Recall that the pumping-up approach consists of two ingredients: one is reverse hypercontractivity (4.24) and the other is to show that the integral of the function does not increase too much after applying the operator (4.25). Regarding the first ingredient, we note that reverse hypercontractivity follows from a modified log-Sobolev inequality (also known as the 1-log Sobolev inequality) [77]. There have already been results on a modified log-Sobolev inequality for “weakly correlated processes” (e.g. [153, Theorem 2.1]). The second ingredient can also be extended to certain weakly correlated processes. For example, consider the steps (4.35)-(4.38) where $P^n$ is replaced with some $P_{Y^n}$ which is not necessarily stationary memoryless. Those steps continue to hold under the assumption that

$$\alpha := \left\| \frac{d P_{Y_i|Y_{\bar{i}}=y_{\bar{i}}}}{d Q} \right\|_\infty < \infty, \quad \forall i = 1, \ldots, n, \quad y_i \in Y^{n-1}. \quad (4.189)$$
Chapter 5
More Dualities

In Chapters 2-4 we have introduced the three ingredients of the functional approach to information-theoretic converse: duality, smoothing, and reverse hypercontractivity, and we have focused on the example of the Brascamp-Lieb divergence to illustrate these ideas. The present chapter is a case study of two other information measures, their dual representations, and the applications to the achievability and the converse to coding theorems.

5.1 Air on $g^*$

The section title only bears a superficial connection to Bach’s “Air on the G String”. We will revisit the image-size characterization, a technique in the classical book [8, Chap 16] which is built on the blowing-up lemma and succeeded in showing the strong converse property of all source-channel networks with known first-order rate region in [8]. The problem considered in Section 4.3 is essentially a basic example of the image-size problem consisting only of a “forward channel”. The purpose of this section is to further demonstrate that the pumping-up argument is capable of substituting the blowing-up argument whenever the latter is used for the image-size analysis, to yield the optimal second-order bounds for the related coding theorems. To achieve this goal, we must show how the pumping-up approach applies to a more involved image size...
problem containing a “reverse channel”. In particular, we introduce an information-theoretic quantity, denoted by $g^*$, to replace the $g$-function for the image size of a set used in [8] in some intermediate steps. Informally speaking, [8] upper bounds the image-size $g(C)$ in terms of $g(B)$, where $B$ and $C$ are the images of a set $A$ under two channels as shown in Figure 5.1. In contrast, we upper bound $g^*(C)$ in terms of $g(B)$.

However the end result needed for the converse proofs of the operational problems in [8] is an upper-bound on $|A|$ in terms of $g(B)$ when $A$ is a good codebook for the $Z$-channel. We prove Theorem 5.1.9), a second-order sharpening of such an end result, which can be readily plugged into the converse proofs of the various operational problems in [8] to obtain $O(\sqrt{m})$ second-order rates.

In addition to getting the optimal second-order bounds, we find that the new approach also has the advantage of simplifying certain technical aspects in dealing with the reverse channel; for example we introduce a simpler definition of the image of a set so that the “maximal code lemma” used in [8, Chap 16] is no longer needed in the proof of certain results. Once again, this is made possible by the fact that we are looking at functions (not necessarily the indicator functions of sets) instead of restricting our attention to sets.

We start by reviewing some relevant concepts from [8] in Section 5.1.1. Then we introduce $g^*$ in Section 5.1.2. In Section 5.1.3 we show that in the discrete memoryless case, $g$ and $g^*$ differ by a sub-exponential factor.

### 5.1.1 Review of the Image-size Problem

**Definition 5.1.1 ([8, Definition 6.2])** Consider $Q_{Z|X}$, a nonnegative measure $\nu_Z$, and $\eta \in (0, 1)$. An $\eta$-image of $A \subseteq X$ is any set $B \subseteq Z$ for which $Q_{Z|X}(B)|_A \geq \eta$.\(^1\)

\(^1\) $Q_{Z|X}(B)|_A$ denotes the restriction of the function $Q_{Z|X}(1_B)$ on $A$.  

138
Define the $\eta$-image size of $A$

\[ g_{Z|X}(A, \nu_Z, \eta) := \inf_{B: \eta\text{-image of } A} \nu_Z[B]. \]  

(5.1)

Fundamental properties of $g$ are discussed in [8], which we informally summarize below (for precise statements see [8, Lemma 15.2]).

Proposition 5.1.1 ([8, Lemma 15.2]) Consider $Q_{Z|X}^{\otimes n}$, counting measure $\nu_Z^{\otimes n}$, and any $\eta \in (0, 1)$. For any $A \subseteq X$,

\[ \frac{1}{n} \log g_{Z^n|X^n}(A, \nu_Z^{\otimes n}, \eta) \geq \frac{1}{n} H(\hat{Z}^n) - o(1), \]  

(5.2)

where $\hat{Z}^n$ is the channel output of $\hat{X}^n$ which is equiprobably distributed on $A$.

Proposition 5.1.2 ([8, Lemma 15.2]) In the setting of Proposition 5.1.1, if, additionally, $A$ is the codebook of a certain $(n, \epsilon)$-fixed composition code under the average
error criterion, then
\[
\frac{1}{n} \log g_{Z^n|X^n}(A, \nu^n_Z, \eta) \leq \frac{1}{n} \log |\mathcal{Z}| + \frac{1}{n} H(\hat{Z}^n) + \epsilon \log |\mathcal{X}| + o(1).
\] (5.3)

**Proposition 5.1.3 ([8, Theorem 15.10])** Consider \( Q_X, Q_{Y|X} \) and \( Q_{Z|X} \) where \(|\mathcal{X}| < \infty \). Let \( \eta \in (0,1) \) and let \( \nu_Y \) and \( \nu_Z \) be the counting measures. Then for any \( A \subseteq \mathcal{X}^n \),

\[
\frac{1}{n} \log g_{Z^n|X^n}(A \cap C_n, \nu^n_Z, \eta) - \frac{1}{n} \log g_{Y^n|X^n}(A \cap C_n, \nu^n_Y, \eta) \leq \sup_{P_{UX}: P_X = Q_X} \{ H(Z|U) - H(Y|U) \} + o(1)
\] (5.4)

where \( C_n := \{ x^n: |P_{x^n} - Q_X| \in \omega(n^{-\frac{1}{2}}) \cap o(1) \} \) denotes the typical set.

**Remark 5.1.1** In the proof of Proposition 5.1.3, [8] used a technical result called the maximal code lemma [8, Lemma 6.3]. Roughly speaking, the proof therein proceeded by connecting \( g \) with the entropy of the output for the equiprobable distribution on the input set (Proposition 5.1.1-5.1.2). Since the bound in Proposition 5.1.1 is generally not tight when \( A \) is not a good codebook, [8] used the maximal code lemma to find a good codebook \( \tilde{A} \subseteq A \) such that \( g_{Z^n|X^n}(\tilde{A}, \nu^n_Z, \eta) \approx g_{Z^n|X^n}(A, \nu^n_Z, \eta) \), and then finished the proof by focusing on \( \tilde{A} \) instead. On the other hand, our approach below using convex duality sidesteps the maximal code lemma.

Since \( g \) is defined as the minimum size of an image, (5.4) of course continues to hold when \( g_{Y^n|X^n}(A \cap C_n, \nu^n_Y, \eta) \) is replaced by the size of any image of \( A \cap C_n \). (For example, in the proof of the strong converse of asymmetric broadcast channel with a common message [8], \( A \) is taken as the set of inputs for a fixed common message, then its decoding set will be an image of \( A \cap C_n \).) But (5.4) no longer holds when \( g_{Z^n|X^n}(A \cap C_n, \nu^n_Z, \eta) \) is replaced by the size of an arbitrary image, therefore when (5.4) is used in the proof of coding theorems, we never substitute a decoding set for
Instead, one uses the following property for the image size when $\mathcal{A}$ is a good codebook of $P$-typical sequences:

$$g_{Z^n|X^n}(\mathcal{A} \cap \mathcal{C}_n, \nu_Z^{\otimes n}, \eta) \geq |\mathcal{A}| e^{-nD(Q_{Z|X}|\nu_Z^P) + o(n)};$$

(5.5)

see [8, Lemma 6.4] for the precise statements, and see Proposition 5.1.6 below for a refinement. Using (5.5) and Proposition 5.1.3 we have the following.

**Proposition 5.1.4** In the setting of Proposition 5.1.3, let $\mathcal{C}_n$ be the $Q_X$-typical set and let $\mathcal{A}$ be an $(n, \epsilon)$-maximal error codebook for $Q_{Z|X}$. Then

$$\frac{1}{n} \log |\mathcal{A} \cap \mathcal{C}_n| - \frac{1}{n} \log g_{Y^n|X^n}(\mathcal{A} \cap \mathcal{C}_n, \nu_Y^{\otimes n}, \eta)$$

$$\leq \sup_{P_{UX}: P_X = Q_X} \{I(Z; X|U) - H(Y|U)\} + o(1)$$

(5.6)

Proposition 5.1.4 is the key argument used in [8] for the strong converse proof of coding theorems including:

- Asymmetric broadcast channel with a common message [8].
- Zigzag network [8].
- Gelfand-Pinsker channel coding [154].

In the remainder of this chapter, we use the pumping-up approach to sharpen Proposition 5.1.4 (see Theorem 5.1.9 below), which will imply converses of the above operational problems with an $O(\sqrt{n})$ second-order term.

### 5.1.2 $g^*$ and the Optimal Second-order Image Size Lemmas

We now introduce a variant of image size, which is technically more convenient (e.g. avoids the use of the maximal code lemma [8, Lemma 6.3] when we charac-
terize the exponents of the image sizes), but can be applied to the converse proof of operational problems in the same way. By the end of this subsection, we present the main result Theorem 5.1.9, which is a second order refinement of Proposition 5.1.4, allowing one to show an $O(\sqrt{n})$ second-order converse for those network information theory problems listed in the end of the previous subsection.

**Definition 5.1.2** Consider $Q_{Z|X}$ and nonnegative measure $\nu_Z$. Define

$$g^*_{Z|X}(A, \nu_Z) := \inf_{g \in \mathcal{H}_{[0, \infty)}(Z): Q_{Z|X}(\ln g) |_A \geq 0} \nu_Z(g).$$

(5.7)

In our approach to converse proofs, we will use $g^*$ instead of $g$ since the former is mathematically more convenient (e.g. it satisfies the exact tensorization property). However, in Section 5.1.3, we show that $g$ and $g^*$ differ by a factor of $e^{o(n)}$ for discrete memoryless channels.

The following exact formula of $g^*$ in terms of the relative entropy can be viewed as an improvement of Proposition 5.1.1-5.1.2.

**Proposition 5.1.5** Fix $Q_{Z|X}$, $\nu_Z$, and let $C$ be an arbitrary subset of $\mathcal{X}$. Then

$$g^*_{Z|X}(A, \nu_Z) = \exp \left( - \inf_{P_X: \text{supp}(P_X) \subseteq C} D(P_Z \| \nu_Z) \right)$$

(5.8)

where $P_X$ is supported on $C$, and $P_X \rightarrow Q_{Z|X} \rightarrow P_Z$.

While it is possible to provide a direct proof of Proposition 5.1.5, we shall derive it as a special case of Theorem 5.1.7 below. We now lower bound $g^*$ in terms of $A$ in the case of a good codebook:

**Proposition 5.1.6** Consider $|Z| < \infty$, $Q_{Z|X}$, $\nu_Z$. Let $A \subseteq \mathcal{X}^n$ be an $(n, \epsilon)$-codebook for $Q^n_{Z|X}$ under the geometric criterion (i.e. the codebook satisfies the assumption in
Theorem 4.2.2). Then

\[ \ln g^*_{Z^n|X^n}(A, \nu^n_Z) \geq \ln |A| - D(Q_{Z^n|X} \| \nu^n_Z | P^n_X) \]

\[ - 2\sqrt{(\alpha - 1)n \ln \frac{1}{1 - \epsilon} - \ln \frac{1}{1 - \epsilon}} \quad (5.9) \]

where \( P_{X^n} \) is the equiprobable distribution on \( A \).

**Proof**

\[ \ln g^*_{Z^n|X^n}(A, \nu^n_Z) \geq -D(P_{Z^n} \| \nu^n_Z) \]

\[ = -D(Q_{Z^n|X} \| \nu^n_Z | P^n_X) + I(\hat{X}^n; \hat{Z}^n) \quad (5.11) \]

\[ \geq \ln |A| - D(Q_{Z^n|X} \| \nu^n_Z | P^n_X) \]

\[ - 2\sqrt{(\alpha - 1)n \ln \frac{1}{1 - \epsilon} - \ln \frac{1}{1 - \epsilon}} \quad (5.12) \]

where (5.10) used Proposition 5.1.5 and (5.12) applied the sharp Fano inequality (Theorem 4.2.2). \( \blacksquare \)

The following key observation, which follows from convex duality, establishes the equivalence of a functional inequality and an entropic inequality. Note that the \( Z = X \) (i.e. \( Q_{Z|X} \) is the identity) special case has been considered in the previous section.

**Theorem 5.1.7** Consider finite alphabets \( \mathcal{X}, \mathcal{Y}, \mathcal{Z} \), random transformations \( Q_{Y|X}, Q_{Z|X} \), nonnegative measures \( \nu_Z \) fully supported on \( Z \) and \( \nu_Y \) fully supported on \( \mathcal{Y} \). For \( d \in \mathbb{R} \), the following statements are equivalent:

\[ \inf_{g: Q_{Z^n|X}(\ln g) = Q_{Y^n|X}(\ln f)} \nu_Z(g) \leq e^d \nu_Y(f), \quad \forall f \geq 0; \quad (5.13) \]

\[ D(P_Z \| \nu_Z) + d \geq D(P_Y \| \nu_Y), \quad \forall P_X. \quad (5.14) \]
where \( P_X \rightarrow Q_{Y|X} \rightarrow P_Y \) and \( P_X \rightarrow Q_{Z|X} \rightarrow P_Z \).

**Proof** (5.13) \( \Rightarrow \) (5.14). For any \( P_X \) and \( \epsilon \in (0, \infty) \),

\[
D(P_Y \| \nu_Y) = P_X P_{Y|X}(\ln f) - \ln \nu_Y(f) \tag{5.15}
\]

\[
\leq P_X P_{Z|X}(\ln g) - \ln \nu_Y(g) + \epsilon + d \tag{5.16}
\]

\[
= P_Z(\ln g) - \ln \nu_Y(g) + \epsilon + d \tag{5.17}
\]

\[
\leq D(P_Z \| \nu_Y) + \epsilon + d \tag{5.18}
\]

where

- In (5.15) we have chosen \( f := \frac{dP_Y}{d\nu_Y} \).

- By (5.13) we can choose \( g \) such that (5.16) holds.

- (5.18) uses the variational formula of the relative entropy.

(5.14) \( \Rightarrow \) (5.13). For the simplicity of presentation we shall assume, without loss of generality, that \( \nu_Z \) is a probability measure. We first note that the infimum in (5.13) is achieved. Let \( a := \max_y (Q_{Y|X} \ln f)(y) \in [-\infty, \infty) \). Then by considering \( g := e^a \) we see that

\[
\inf_{g:Q_{Z|X}(\ln g) \geq Q_{Y|X}(\ln f)} \nu_Z(g) \leq e^a. \tag{5.19}
\]

But if \( g(z') > e^a \max_z \nu_Z^{-1}(z) \) for some \( z' \in Z \) then \( \nu_Z(g) > e^a \). This means that

\[
\inf_{g:Q_{Z|X}(\ln g) \geq Q_{Y|X}(\ln f)} \nu_Z(g) = \inf_{g \in C} \nu_Z(g) \tag{5.20}
\]

144
where $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$ and

$$
\mathcal{C}_1 := \{ g \in \mathcal{H}_+(\mathcal{Z}) : Q_{Z|X}(\ln g) \geq Q_{Y|X}(\ln f) \}; \\
\mathcal{C}_2 := \mathcal{H}_{\left[0, e^{\max_{z} \nu_{Z}^{-1}(z)}\right]}(\mathcal{Z}).
$$

(5.21)

(5.22)

Note that the closed set $\mathcal{C}_1$ is the domain of the infimum on the left side of (5.20), and $\mathcal{C}_2$ is compact. Therefore $\mathcal{C}$ is compact, and the infimum on the right side of (5.20) is achieved at some $g^* \in \mathcal{C}$. The infimum on the left side of (5.20) is thus also achieved at $g^*$.

Next, define a probability measure $P_Z$ by $\frac{dP_Z}{d\nu_Z} = \frac{1}{\nu_Z(g^*)} g^*$, and we will argue that there exists some $P_Z$ such that $P_X \to Q_{Y|X} \to P_Z$. Indeed, apply Lagrange multiplier method to the minimization in (5.13),

$$
\min_{g} \{ \nu_Z(g) - \lambda_X [Q_{Z|X}(\ln g) - Q_{Y|X}(\ln f)] \}
$$

where $\lambda_X$ is a nonnegative measure on $\mathcal{X}$. Then there must be some $\lambda_X$ for which $g^*$ is a stationary point for the functional in (5.23). Then the claim follows since the desired $P_X$ is the $\lambda_X$ normalized to a probability measure. From the complementary slackness condition, we also see that

$$
Q_{Z|X} \ln g = Q_{Y|X} \ln f, \quad P_X - a.s.
$$

(5.24)

With those preparations the rest of the proof follows immediately:

$$
\ln \nu_Z(g^*) = P_X Q_{Z|X}(\ln g) - D(P_Z \| \nu_Z) \\
= P_X Q_{Y|X}(\ln f) - D(P_Y \| ) + d \\
\leq \ln \nu_Y(f) + d
$$

(5.25)

(5.26)

(5.27)
where (5.26) uses (5.24) and (5.14); and (5.27) uses the variational formula of the relative entropy.

**Proof of Proposition 5.1.5** We first note that it is without loss of generality to consider the case of \( \mathcal{X} = \mathcal{C} \). The claim then follows from Theorem 5.1.7 by taking \( \mathcal{Y} \) to be a singleton.

Applying the now familiar reverse hypercontractivity argument to Theorem 5.1.7, we obtain the following result, which can be viewed as a second order sharpening of Proposition 5.1.3.

**Theorem 5.1.8** For \((Q_{Z|X}, Q_{Y|X}, \nu_Z, \nu_Y, \mathcal{C})\), where \( \mathcal{C} \subseteq \mathcal{X} \), define

\[
\text{d}(Q_{Z|X}, Q_{Y|X}, \nu_Z, \nu_Y, \mathcal{C}) := \sup_{P_X : \text{supp}(P_X) \subseteq \mathcal{C}} \{D(P_Y \| \nu_Y) - D(P_Z \| \nu_Z)\}, \quad (5.28)
\]

where \( P_X \rightarrow Q_{Y|X} \rightarrow P_Y \) and \( P_X \rightarrow Q_{Z|X} \rightarrow P_Z \). Let \( \alpha_Y := \sup_x \left\| \frac{dQ_{Y|X}}{dP_X} \right\|_{\infty} \). In the stationary memoryless case, for any \( \eta \in (0, 1) \) and \( \mathcal{C}_n \subseteq \mathcal{X}^n \), we have

\[
\ln g^*_{Z^n|X^n}(\mathcal{C}_n \cap \{Q_{Y|X}^n \geq \eta\}, \nu_Z^{n*}) \leq \ln \nu_Z^{n*}(f) + \text{d}(Q_{Z^n|X^n}, Q_{Y^n|X^n}, \nu_Z^{n*}, \nu_Y^{n*}, \mathcal{C}_n) + 2\sqrt{(\alpha_Y - 1)n \ln \frac{1}{\eta} + \ln \frac{1}{\eta}}. \quad (5.29)
\]
In particular, given \( Q_X, \delta \in (0, 1) \). For \( n > 3 \beta_X \ln \frac{|X|}{\delta} \) we can find \( C_n \) with \( Q_X^{\otimes n}[C_n] \geq 1 - \delta \) so that for any \( \mathcal{A} \subseteq \mathcal{C}_n \),

\[
\ln g_{Z^n|X^n}^*(\mathcal{A}, \nu_{Z^n}^{\otimes n}) \leq \ln g_{Y^n|X^n}(\mathcal{A}, \nu_{Y^n}^{\otimes n}, \eta) \\
+ n \sup_{P_{X|U}: P_X = Q_X} \{ D(P_Y|U \| P_U) - D(P_Z|U \| P_U) \} \\
+ \sqrt{3n \beta_X \ln \frac{|X|}{\delta}} \ln \alpha_Z \beta_X \\
+ 2 \sqrt{(\alpha_Y - 1)n \ln \frac{1}{\eta} + \frac{1}{\eta}}. \\
(5.30)
\]

where \( P_{X|U = u} \to Q_{Z|X} \to P_{Z|U = u} \) and \( P_{X|U = u} \to Q_{Y|X} \to P_{Y|U = u} \) for each \( u \in \mathcal{U} \), and

\[
\beta_X := (\min_x Q_X(x))^{-1}; \\
\alpha_Z := |\nu_Z| \sup_x \left\| \frac{dQ_{Z|X = x}}{d\nu_Z} \right\|_{\infty}. \\
(5.32)
\]

**Proof** For \( t \in [0, \infty) \) and \( \alpha \in [1, \infty) \), put

\[
\Lambda_{\alpha,t}^{\nu_Y} = \bigotimes_{i=1}^n (e^{-t} + \alpha(1 - e^{-t})\nu_Y). \\
(5.33)
\]

Consider any \( f \in \mathcal{H}_{[0,1]}(Y^n) \). Denote \( d := d(Q_{Z^n|X}, Q_{Y^n|X}, \nu_{Z^n}^{\otimes n}, \nu_{Y^n}^{\otimes n}, \mathcal{C}_n) \) for simplicity.

\[
e^{(\alpha Y - 1)nt + d} \cdot \nu_{Z^n}^{\otimes n}(f) \geq e^d \cdot \nu_{Z^n}^{\otimes n}(\Lambda_{\alpha,t}^{\nu_Y} f) \\
(5.34)
\]

\[
\geq \inf_{g: Q_{Z^n|X}(\ln g)|_{C_n} \geq Q_{Y^n|X}(\ln \Lambda_{\alpha,t}^{\nu_Y})|_{C_n}} \nu_{Z^n}^{\otimes n}(g) \\
(5.35)
\]

\[
\geq \inf_{g: Q_{Z^n|X}(\ln g)|_{C_{n\cap A}} \geq \frac{1}{1 - e^{-t}} \ln \eta} \nu_{Z^n}^{\otimes n}(g) \\
(5.36)
\]

\[
= \inf_{g: Q_{Z^n|X}(\ln g)|_{C_{n\cap A}} \geq 0} \nu_{Z^n}^{\otimes n}(g) \cdot \eta \frac{1}{1 - e^{-t}} \\
(5.37)
\]
where we defined

\[ \mathcal{A} := \{ x^n : Q_{Y^n|X^n=x^n}(f) \geq \eta \}. \]  

(5.38)

This proves (5.29).

The proof of (5.30) follows by applying the single-letterization argument to (5.29).

More specifically, for any \( P_{X^n} \) supported on

\[ \mathcal{C}_n := \{ x^n : \hat{P}_{Z^n} \leq (1 + \epsilon_n)Q_X \}, \]

we have

\[
D(P_{Y^n} \| \nu_Z^n) - D(P_{Z^n} \| \nu_Z^n)
= \sum_{i=1}^{n} \left[ D(P_{Y^n}_{i+1}^n \| \nu_Y^n \otimes \nu_Z^{(n-i)}) - D(P_{Y^n}_{i+1}^n \| \nu_Y^{(i-1)} \otimes \nu_Z^{(n-i+1)}) \right]
\]

(5.40)

\[
= \sum_{i=1}^{n} \left[ D(P_{Y^n}_{i+1}^n \| \nu_Y \otimes P_{Y^{i-1}}_{Z+i+1}^n) - D(P_{Y^n}_{i+1}^n \| P_{Y^{i-1}}_{Z+i+1}^n \otimes \nu_Z) \right]
\]

(5.41)

\[
= n \left[ D(P_{Y^n}_{Y^{i-1}} Z^n_{i+1} \| \nu_Y \| P_{Y^{i-1}}_{Z^n_{i+1}}) - D(P_{Z^n}_{Y^{i-1}} Z^n_{i+1} \| \nu_Z \| P_{Y^{i-1}}_{Z^n_{i+1}}) \right]
\]

(5.42)

\[
\leq n \sup_{P_{U^n} : P_X \leq (1 + \epsilon_n)Q_X} \{ D(P_{Y|U} \| \nu_Y \| P_U) - D(P_{Z|U} \| \nu_Z \| P_U) \}
\]

(5.43)

\[
= n \sup_{P_{U^n} : P_X \leq Q_X} \{ D(P_{Y|U} \| \nu_Y \| P_U) - D(P_{Z|U} \| \nu_Z \| P_U) \}
+ n \epsilon_n \ln \beta_X \alpha_Y \alpha_Z
\]

(5.44)

where we defined \( I \) to be an equiprobable random variable on \( \{1, \ldots, n\} \) independent of \( X^nY^nZ^n \), and

- (5.42) follows from the definition of the conditional relative entropy.
- To see (5.43) we note that \( U - X_{I} - Y_{I}Z_{I} \) forms a Markov chain if \( U \leftarrow IY^{I-1}Z_{I+1}^n \), and that \( P_{X_I} \leq (1 + \epsilon_n)Q_X \).
• The proof of (5.44) follows from some simple estimates of the relative entropy. Abbreviate $\epsilon_n$ as $\epsilon$. Consider $P_{\tilde{U}X}$ where $\tilde{U} = \mathcal{U} \cup \{\ast\}$, $\tilde{X} = \mathcal{X}$, defined as the follows

$$P_{\tilde{U}} := \frac{1}{1+\epsilon}P_{U} + \frac{\epsilon}{1+\epsilon}\delta_{\bullet};$$  \hspace{1cm} (5.45)

$$P_{\tilde{X}|U=u} := P_{X|U=u}, \quad \forall u \in \mathcal{U};$$  \hspace{1cm} (5.46)

$$P_{\tilde{X}|\tilde{U}=\bullet} := \frac{1+\epsilon}{\epsilon} \left( Q_X - \frac{1}{1+\epsilon}P_X \right)$$  \hspace{1cm} (5.47)

where $\delta_{\bullet}$ is the one-point distribution on $\bullet$. Then observe that $P_{\tilde{X}} = Q_X$, and

$$D(P_{Z|U}\|\nu_Z|P_{U}) - D(P_{\tilde{Z}|\tilde{U}}\|\nu_Z|P_{\tilde{U}})$$

$$= D \left( P_{Z|U}\left\|\frac{1}{\nu_Z}P_{U} \right\|P_{U} \right) - D \left( P_{\tilde{Z}|\tilde{U}}\left\|\frac{1}{\nu_Z}P_{\tilde{U}} \right\|P_{\tilde{U}} \right)$$  \hspace{1cm} (5.48)

$$\geq -\epsilon D \left( \frac{1+\epsilon}{\epsilon}Q_Z - \frac{1}{\epsilon}P_Z \left\|\frac{1}{\nu_Z}P_Z \right\| \right)$$  \hspace{1cm} (5.49)

$$\geq -\epsilon \ln \alpha_Z.$$  \hspace{1cm} (5.50)

An upper bound on $D(P_{Y|U}\|\nu_Y|P_{U})$ can also be obtained, the steps for which we omit here.

The proof of (5.30) is completed by choosing $\epsilon_n = \sqrt{\frac{3\beta_X}{n}} \ln \left\| \frac{|X|}{\delta} \right\|$, and using

$$\mathbb{P}[\hat{P}_{X^n} \leq (1 + \epsilon_n)Q_X] \geq 1 - \delta.$$  \hspace{1cm} (5.51)

Finally, applying Proposition 5.1.6 to (5.30), we obtain a second-order sharpening of Proposition 5.1.4.

**Theorem 5.1.9** Fix $Q_X$, $Q_{Y|X}$, $Q_{Z|X}$, $\eta$, $\delta$, $\epsilon \in (0,1)$. Then for $n > 3\beta_X \ln \frac{2|X|}{\delta}$, there exists $C_n \subseteq \mathcal{X}^n$ with $Q_X^n[C_n] \geq 1 - \delta$ such that for any $A \subseteq C_n$ which is an
$(n, \varepsilon)$-codebook under the geometric error criterion,

$$\log |A| - \log \mathfrak{g}_{Y^n|X^n}(A, \nu_{\mathcal{Y}^n})$$

$$\leq n \sup_{P_{UX}: P_X = Q_X} \{I(Z; X|U) + D(P_{Y|U}||\nu_Y|P_U)\} + D_{\frac{1}{2}} \sqrt{n} + D_0. \quad (5.52)$$

where

$$D_{\frac{1}{2}} := \sqrt{3n\beta_X \ln \frac{2|\mathcal{X}|}{\delta} \ln (\alpha_Y \alpha_Z \beta_X)} + 2\sqrt{(\alpha_Y - 1)n \ln \frac{1}{\eta}} + \sqrt{n \frac{2}{\delta} \ln \alpha_Z + 2\sqrt{(\alpha_Z - 1)n \ln \frac{1}{1 - \epsilon}}}.$$ \quad (5.53)

and

$$D_0 := \ln \frac{1}{1 - \epsilon} + \ln \frac{1}{\eta}. \quad (5.54)$$

and $\alpha_Z$, $\alpha_Y$ and $\beta_X$ are as defined in Theorem 5.1.8.

**Proof** Consider

$$\mathcal{D}_n := \left\{ x^n: \sum_{i=1}^{n} D(Q_{Z|X=x_i}\|\nu_Z) \leq nD(Q_{Z|X}\|\nu_Z|Q_X) + \sqrt{\frac{n}{2} \ln \frac{1}{\delta} \ln \alpha_Z} \right\}. \quad (5.55)$$

In Proposition 5.1.6 for $A \subseteq \mathcal{D}_n$, we have

$$\ln \mathfrak{g}_{Y^n|X^n}^*(A, \nu_{\mathcal{Y}^n})$$

$$\geq \ln |A| - nD(Q_{Z|X}\|\nu_Z|Q_X) - \sqrt{\frac{n}{2} \ln \frac{1}{\delta} \ln \alpha_Z}$$

$$- 2\sqrt{(\alpha_Z - 1)n \ln \frac{1}{1 - \epsilon}} - \ln \frac{1}{1 - \epsilon}. \quad (5.56)$$

Moreover, since

$$\sup_{x} D(Q_{Z|X}(\cdot|x)\|\nu_Z) - \inf_{x} D(Q_{Z|X}(\cdot|x)\|\nu_Z) \leq \ln \alpha_Z, \quad (5.57)$$
by Hoeffding’s inequality we have

\[ Q_X^{\otimes n} [\mathcal{D}_n] \geq 1 - \delta. \]  

(5.58)

If \( \mathcal{C}_n \) is as in Theorem 5.1.8 for which (5.30) holds, then for any \( \mathcal{A} \in \mathcal{C}_n \cap \mathcal{D}_n \), by (5.56) and (5.30) we have

\[
\ln |\mathcal{A}| \leq \ln g_{Y^n|X^n}(\mathcal{A}, \nu_Z^{\otimes n}, \eta)
+ n \sup_{P_{X^n}: P_X = Q_X} \left\{ D(P_Y | U \| \nu_Y | U) - D(P_Z | U \| \nu_Z | U) \right\}
+ 3n\beta_X \ln \frac{|\mathcal{X}|}{\delta} \ln \alpha_Y \alpha_Z \beta_X
+ 2\sqrt{(\alpha_Y - 1)n \ln \frac{1}{\eta} + \ln \alpha_Y \sqrt{n \ln \frac{1}{\delta}}}
+ 2\sqrt{(\alpha_Z - 1)n \ln \frac{1}{1 - \epsilon} + \ln \frac{1}{1 - \epsilon} + \ln \frac{1}{\eta}}.
\]

(5.59)

\[ \blacksquare \]

5.1.3 Relationship between \( g \) and \( g^* \)

**Proposition 5.1.10** Consider \( Q_{Z|X} \), the counting measure \( \nu_Z \) on the finite \( Z \), \( \eta \in (0, 1) \), \( \lambda \in (0, 1 - \eta) \), and \( \mathcal{A} \subseteq \mathcal{X}^n \). We claim that in the definition of \( g_{Z|X} \), the assumption \( g \in H_{(0,1]} \) can be changed to \( g \in H_{[0,1]} \) without essential difference. More precisely,

\[
\inf_{g \in H_{[0,1]}(Z) \cap Q_{Z|X}(g) } \nu_Z(g) \leq g_{Z|X}(\mathcal{A}, \nu_Z, \eta) \leq \lambda^{-1} \inf_{g \in H_{[0,1]}(Z) \cap Q_{Z|X}(g) } \nu_Z(g).
\]

(5.60)

(5.61)
Moreover, \( g_{Z|X} \) would be upper-bounded by \( g^*_{Z|X} \) if the assumption \( g \in \mathcal{H}_{[0,z]} \) in the definition of \( g^*_{Z|X} \) were replaced by \( \mathcal{H}_{[0,1]} \):

\[
\inf_{g \in \mathcal{H}_{[0,1]}(\mathcal{Z}) : Q_{Z|X}(g) \geq \eta} \nu_Z(g) \leq \eta \inf_{g \in \mathcal{H}_{[0,1]}(\mathcal{Z}) : Q_{Z|X}(\ln g) \geq 0} \nu_Z(g). \tag{5.62}
\]

Further, in the stationary memoryless setting with \( |\mathcal{X}| < \infty \), \( g_{Z^n|X^n} \) and \( g^*_{Z^n|X^n} \) are equivalent up to a sub-exponential factor. More precisely, we also have

\[
g^*_{Z^n|X^n}(\mathcal{A}, \nu_Z^{\otimes n}) \leq \frac{1}{\eta} \exp\left(\frac{(\alpha-1)n \ln \frac{1}{\eta}}{\eta}\right) \inf_{g \in \mathcal{H}_{[0,1]}(\mathcal{Z}^n) : Q_{Z|X}(g) \geq 0} \nu_Z^{\otimes n}(g), \tag{5.63}
\]

where \( \alpha := \sup_x \left\| \frac{dP_{Z=X=x}}{d\nu_Z} \right\|_\infty \), and

\[
g_{Z^n|X^n}(\mathcal{A}, \nu_Z^{\otimes n}, \eta) \leq e^{o(n)} \cdot g^*_{Z^n|X^n}(\mathcal{A}, \nu_Z^{\otimes n}). \tag{5.64}
\]

**Proof** By definition,

\[
g_{Z|X}(\mathcal{A}, \nu_Z, \eta) = \inf_{g \in \mathcal{H}_{[0,1]}(\mathcal{Z}) : Q_{Z|X}(g) \geq \eta} \nu_Z(g) \tag{5.65}
\]

so \(5.60) follows because \( \mathcal{H}_{[0,1]}(\mathcal{Z}) \subseteq \mathcal{H}_{[0,1]}(\mathcal{Z}) \).

To see \(5.61\), consider an arbitrary \( g \in H_{[0,1]}(\mathcal{Z}) \), and put \( \tilde{g} := 1\{z : g(z) \geq \lambda\} \). Note that for any \( z \) such that \( Q_{Z|X=x}(g) \geq \eta + \lambda \), we have

\[
Q_{Z|X=x}(\tilde{g}) \geq Q_{Z|X=x}(g) - Q_{Z|X=x}(g - \tilde{g}) \tag{5.66}
\]

\[
\geq \eta + \lambda - \sup_z (g(z) - \tilde{g}(z)) \tag{5.67}
\]

\[
\geq \eta. \tag{5.68}
\]

But \( \lambda \tilde{g} \leq g \), which establishes \(5.61\).
The proof of (5.62) follows from the fact that for any $g \in H_{[0,1]}(Z)$ and $x$,

$$Q_{Z|X=x}(\eta g) \geq \eta e^{Q_{Z|X}(x \ln g)}$$

(5.69)

and obviously $\eta g \in H_{[0,1]}(Z)$.

To see (5.63), given $g \in H_{[0,1]}(Z^n)$ such that $Q_{Z^n|X}(g) \geq \eta$, let $\tilde{g} := \left(\frac{1}{\eta}\right)^{1+\frac{1}{\eta}} L_{\alpha,t} \Lambda_{\alpha} g$ (see (4.44)), with $t := \sqrt{\frac{1}{\alpha-1} \ln \frac{1}{\eta}}$. Then by the same argument as in the proof of Theorem 4.2.2

$$Q_{Z|X}(\ln \tilde{g}) \geq 0,$$

(5.70)

$$\nu_{Z}^{\otimes n} (\tilde{g}) \leq \frac{1}{\eta} e^{\sqrt{(\alpha-1) \ln \frac{1}{\eta}}} \nu_{Z}^{\otimes n} (g).$$

(5.71)

Finally, to see (5.64), consider

$$g_{Z^n|X^n}(A, \nu_{Z}^{\otimes n}, \eta)$$

$$\leq \sum_{P: n\text{-type}} g_{Z^n|X^n}(A \cap T_{P}^{n}, \nu_{Z}^{\otimes n}, \eta)$$

(5.72)

$$\leq \text{poly}(n) \cdot \max_{P} g_{Z^n|X^n}(A \cap T_{P}^{n}, \nu_{Z}^{\otimes n}, \eta)$$

(5.73)

$$\leq \text{poly}(n) \cdot M \cdot e^{nH(Q_{Z|X})(P) + n(\epsilon - \eta)}$$

(5.74)

$$\leq \text{poly}(n) \cdot e^{H(\tilde{Z}^n)} + n(\epsilon - \eta)$$

(5.75)

$$\leq \text{poly}(n) \cdot e^{H(\tilde{Z}^n) + O(\sqrt{\eta}) + n(\epsilon - \eta)}$$

(5.76)

where

- In (5.74) we used [8, Lemma 6.3] to find a $\epsilon$-maximal error code where the codebook $\tilde{A} \subseteq A \cap T_P^n$ is of size $M$ and $\epsilon \in (\eta, 1)$ is arbitrary. We have chosen $P$ to be one that achieves the maximum in (5.73).
• In (5.75) we set $\tilde{X}^n$ to be equiprobable on $\tilde{A}$ and let $\tilde{Z}^n$ be the corresponding output.

• (5.76) used the sharp Fano’s inequality (Theorem 4.2.2).

The proof is completed noting that $\epsilon$ can be arbitrarily close to $\eta$ and using

$$e^{H(\tilde{Z}^n)} \leq \sup_{P_{X^n} : \text{supp}(P_{X^n}) \subseteq \tilde{A}} e^{H(P_{Z^n})}$$

(5.77)

$$= g_{Z^n|X^n}(\tilde{A}, \nu_{Z^n})$$

(5.78)

$$\leq g_{Z^n|X^n}(\tilde{A}, \nu_{Z^n}).$$

(5.79)

5.2 Variations on $E_\gamma$

Fundamental to the information spectrum approach \[44\] to information theory is the quantity $\mathbb{P}[t_{P|Q}(X) > \lambda]$, where $X \sim P$ and $\lambda \in \mathbb{R}$. The knowledge of $\mathbb{P}[t_{P|Q}(X) > \lambda]$ for all parameters $\lambda$ is equivalent to the knowledge of the binary hypothesis region for the probability measures $P$ and $Q$ (see e.g. \[43\]). In this section, we study another information theoretic quantity, the knowledge of which (for all parameters) is also equivalent to the knowledge of the binary hypothesis testing region. We show how its variational formula (i.e. the functional representation of this quantity) is used in the achievability and converse proofs of some operational problems.

We define $E_\gamma$ and discuss its mathematical properties in Section 5.2.1. Section 5.2.2 discusses the relationship between $E_\gamma$ and the smooth Rényi divergence, which suggests the potential of $E_\gamma$ in non-asymptotic information theory and the large deviation analysis. To illustrate such a potential, in Section 5.2.3 we generalize the channel resolvability problem \[155\] by replacing the total variation metric with the
$E_\gamma$ metric, and show how results on $E_\gamma$-resolvability can be applied in source coding, the mutual covering lemma \[156\] and the wiretap channels in Sections \[5.2.4\,5.2.6\].

### 5.2.1 $E_\gamma$: Definition and Dual Formulation

**Definition 5.2.1** Given probability distributions $P \ll Q$ on the same alphabet and $\gamma \geq 1$, define\(^3\)

\[
E_\gamma(P\|Q) := \mathbb{P}[i_{P\|Q}(X) > \log \gamma] - \gamma \mathbb{P}[i_{P\|Q}(Y) > \log \gamma] \tag{5.80}
\]

where $X \sim P$ and $Y \sim Q$.

Then we see that $E_\gamma$ is an $f$-divergence \[158\] with

\[
f(x) = (x - \gamma)^+ \tag{5.81}
\]

We now find a variational formula for $E_\gamma$ via convex duality. Recall that the Legendre-Fenchel dual of a convex functional $\phi(\cdot)$ on the set of measurable functions is defined by

\[
\phi^*(\pi) := \sup_f \left\{ \int f \, d\pi - \phi(f) \right\} \tag{5.82}
\]

for any finite measure $\pi$, and under mild regularity conditions (see e.g. \[105\, Lemma 4.5.8\]), we have

\[
\phi(f) := \sup_{\pi} \left\{ \int f \, d\pi - \phi^*(\pi) \right\} \tag{5.83}
\]

\(^3\)We found the $E_\gamma$ notation in an email communication with Yury Polyanskiy \[157\]. However, we are uncertain about the genesis of such a notation.
Now, definition (5.80) can be extended to the set of all finite (not necessarily non-negative and normalized) measures by

\[ E_\gamma(\pi \| Q) = \int \left( \frac{d\pi}{dQ} - \gamma \right)^+ dQ \]  \hspace{1cm} (5.84)

which is convex in \( \pi \). Then for any bounded function \( f \), it is easy to verify that

\[
\sup_{\pi} \left\{ \int f d\pi - E_\gamma(\pi \| Q) \right\} = \sup_{\pi} \left\{ \int \left[ \frac{d\pi}{dQ} f - \left( \frac{d\pi}{dQ} - \gamma \right)^+ \right] dQ \right\} = \begin{cases} \gamma \int f dQ, & 0 \leq f \leq 1 \text{ a.e.} ; \\ +\infty, & \text{otherwise.} \end{cases} \]  \hspace{1cm} (5.85)

We thus obtain

**Theorem 5.2.1** For a given \( \gamma \in [1, \infty) \) and a probability measure \( Q \), \( E_\gamma(\cdot \| Q) \) is the Legendre-Fenchel dual of the functional

\[
f \mapsto \begin{cases} \gamma \int f dQ, & 0 \leq f \leq 1 \text{ a.e.} \\ +\infty, & \text{otherwise.} \end{cases} \]  \hspace{1cm} (5.86)

Thus

\[
E_\gamma(\pi \| Q) = \sup_{f : 0 \leq f \leq 1} \left\{ \int f d\pi - \gamma \int f dQ \right\}. \hspace{1cm} (5.87)
\]

The variational formula \( (5.88) \) can also be derived from \( (5.80) \) by other means; the present derivation emphasizes the dual optimization viewpoint of this thesis.

Below, we prove some basic properties of \( E_\gamma \) useful for later sections. Additional properties of \( E_\gamma \) can be found in [132][146] Theorem 21][43].

**Proposition 5.2.2** Assume that \( P \ll S \ll Q \) are probability distributions on the same alphabet, and \( \gamma, \gamma_1, \gamma_2 \geq 1 \).
Figure 5.2: $E_\gamma(P\|Q)$ as a function of $\gamma$ where $P = \text{Ber}(0.1)$ and $Q = \text{Ber}(0.5)$.

1. $[1, \infty) \to [0, \infty)$: $\gamma \mapsto E_\gamma(P\|Q)$ is convex, non-increasing, and continuous.

2. For any event $\mathcal{A}$,

$$Q(\mathcal{A}) \geq \frac{1}{\gamma} (P(\mathcal{A}) - E_\gamma(P\|Q)).$$  \hfill (5.89)

3. Triangle inequalities:

$$E_{\gamma_1\gamma_2}(P\|Q) \leq E_{\gamma_1}(P\|S) + \gamma_1 E_{\gamma_2}(S\|Q),$$  \hfill (5.90)

$$E_\gamma(P\|Q) + E_\gamma(P\|S) \geq \frac{\gamma}{2} |S - Q| + 1 - \gamma.$$  \hfill (5.91)

4. Monotonicity: if $P_{XY} = P_X P_{Y|X}$ and $Q_{XY} = Q_X Q_{Y|X}$ are joint distributions on $\mathcal{X} \times \mathcal{Y}$, then

$$E_\gamma(P_X\|Q_X) \leq E_\gamma(P_{XY}\|Q_{XY})$$  \hfill (5.92)

where equality holds for all $\gamma \geq 1$ if and only if $P_{Y|X} = Q_{Y|X}$. 

157
5. Given $P_X$, $P_{Y|X}$ and $Q_{Y|X}$, define

$$E_\gamma(P_{Y|X}\|Q_{Y|X}|P_X) := \mathbb{E}[E_\gamma(P_{Y|X}(\cdot|X)\|Q_{Y|X}(\cdot|X))] \quad (5.93)$$

where the expectation is w.r.t. $X \sim P_X$. Then we have

$$E_\gamma(P_XP_{Y|X}\|P_XQ_{Y|X}) = E_\gamma(P_{Y|X}\|Q_{Y|X}|P_X). \quad (5.94)$$

6.

$$1 - \gamma \left( 1 - \frac{1}{2} |P - Q| \right) \leq E_\gamma(P\|Q) \leq \frac{1}{2} |P - Q|. \quad (5.95)$$

**Proof** The proofs of 1), 2), 4), 5) and the second inequality in (5.95) are omitted since they follow either directly from the definition of $E_\gamma$ or are similar to the proofs of the corresponding properties for total variation distance.

For 3), observe that

$$E_{\gamma_1\gamma_2}(P\|Q) = \max_A (P(A) - \gamma_1 S(A) + \gamma_1 S(A) - \gamma_1 \gamma_2 Q(A)) \quad (5.96)$$

$$\leq \max_A (P(A) - \gamma_1 S(A)) + \max_A (\gamma_1 S(A) - \gamma_1 \gamma_2 Q(A)) \quad (5.97)$$

$$= E_{\gamma_1}(P\|S) + \gamma_1 E_{\gamma_2}(S\|Q), \quad (5.98)$$
and that

\[ E_\gamma(P\|Q) + E_\gamma(P\|S) = \max_{\mathcal{A}}(P(\mathcal{A}) - \gamma Q(\mathcal{A})) \]
\[ + \max_{\mathcal{A}}(1 - P(\mathcal{A}) - \gamma + \gamma S(\mathcal{A})) \]
\[ \geq \max_{\mathcal{A}}(P(\mathcal{A}) - \gamma Q(\mathcal{A})) \]
\[ + 1 - P(\mathcal{A}) - \gamma + \gamma S(\mathcal{A})) \]
\[ = \gamma \max_{\mathcal{A}}(S(\mathcal{A}) - Q(\mathcal{A})) + 1 - \gamma \]
\[ = \frac{\gamma}{2} |S - Q| + 1 - \gamma. \]

(5.99)  (5.100)  (5.101)  (5.102)

As far as 6) note that the left inequality in (5.95) follows by setting \( S = P \) in (5.91).

The \( E_\gamma \) metric provides a very convenient tool for change-of-measure purposes: if \( E_\gamma(P\|Q) \) is small, and the probability of some event is large under \( P \), then the probability of this event under \( Q \) is essentially lower-bounded by \( \frac{1}{\gamma} \). In the special case of \( \gamma = 1 \) (total variation distance), this change-of-measure trick has been widely used, see [159] [160], but the general \( \gamma \geq 1 \) case is more versatile, since \( P \) and \( Q \) need not be essentially the same.

5.2.2 The Relationship of \( E_\gamma \) with Excess Information and Smooth Rényi Divergence

This subsection shows that \( E_\gamma \) (as a function of \( \gamma \)) is the inverse function of the smooth Rényi divergence [73] of order \(+\infty\). Various bounds among \( E_\gamma \) and other popular information measures used in non-asymptotic information theory are presented. These facts buttress the potential of \( E_\gamma \) in non-asymptotic analysis and large deviation theory.

159
Definition 5.2.2 For $\gamma > 0$, a probability measure $P$, and a nonnegative $\sigma$-finite measure $\mu$, where $P \ll \mu$, we define the complementary relative information spectrum metric with threshold $\gamma$,

$$
\tilde{F}_\gamma(P\|\mu) := \mathbb{P}[t_{P\|\mu}(X) > \log \gamma]
$$

(5.103)

where $X \sim P$.

As such, the complementary relative information spectrum can be expressed in terms of the relative information spectrum (see [43]) as

$$
\tilde{F}_\gamma(P\|\mu) = 1 - F_{P\|\mu}(\log \gamma).
$$

(5.104)

Also, it is clear from (5.103) that $\tilde{F}_\gamma$ upper-bounds $E_\gamma$.

Note that (5.103) is nonnegative and vanishes when $P = \mu$ provided that $\gamma > 1$, and can therefore be considered as a measure of the discrepancy between $P$ and $\mu$. In addition to playing an important role in one-shot analysis (see [161]), (5.103) provides richer information than the relative entropy measure since

$$
D(P\|\mu) := \mathbb{E}[t_{P\|\mu}(X)]
$$

(5.105)

$$
= \int_{[0, +\infty)} \mathbb{P}[t_{P\|\mu}(X) > \tau]d\tau
$$

$$
- \int_{(-\infty, 0]} (1 - \mathbb{P}[t_{P\|\mu}(X) > \tau])d\tau.
$$

(5.106)

Moreover, the complementary relative information spectrum is also related to total variation distance since for probability measures $P$ and $Q$,

$$
\frac{1}{2}|P - Q| = \mathbb{P}[t_{P\|\mu}(X) > 0] - \mathbb{P}[t_{P\|\mu}(Y) > 0],
$$

(5.107)
where $X \sim P$ and $Y \sim Q$. However, perhaps surprisingly, the complementary relative information spectrum metric does not satisfy the data processing inequality, in contrast to the relative entropy and total variation distance; the proof of such a result, along with other properties of the complementary relative information spectrum, can be found in [143].

Another information measure closely related to $E_\gamma$ is the smooth Rényi. Let us first generalize some of our definitions to allow nonnegative finite measures that are not necessarily probability measures:

**Definition 5.2.3** For $\gamma \geq 1$, nonnegative finite measures $\mu$ and $\nu$ on $\mathcal{X}$, $\mu \ll \nu$,

$$
|\mu - \nu| := \int |d\mu - d\nu|,
$$

(5.108)

and\footnote{Following established usage in measure theory, we use $\lambda(\{x : i_{\mu|\nu}(x) > \log \gamma\})$ as an abbreviation of $\lambda(\{x : i_{\mu|\nu}(x) > \log \gamma\})$ for an arbitrary signed measure $\lambda$.}

$$
E_\gamma(\mu\|\nu) := \sup_{A} \{\mu(A) - \gamma \nu(A)\}
$$

(5.109)

$$
= (\mu - \gamma \nu)(i_{\mu|\nu} > \log \gamma)
$$

(5.110)

$$
= \frac{1}{2}(\mu - \gamma \nu)(\mathcal{X}) - \frac{1}{2}(\mu - \gamma \nu)(i_{\mu|\nu} \leq \log \gamma)
$$

$$
+ \frac{1}{2}(\mu - \gamma \nu)(i_{\mu|\nu} > \log \gamma)
$$

(5.111)

$$
= \frac{1}{2}\mu(\mathcal{X}) - \frac{\gamma}{2}\nu(\mathcal{X}) + \frac{1}{2}|\mu - \gamma \nu|.
$$

(5.112)

Note that $E_1(P\|\mu) \neq \frac{1}{2}|P - \mu|$ when $\mu$ is not a probability measure.

The following result is a generalization of the $\gamma_1 = 1$ case of (5.90) to unnormalized measures, and the proof is immediate from the definition of (5.109) and the subadditivity of the sup operator.
Proposition 5.2.3 Triangle inequality: if $\mu$, $\nu$ and $\theta$ are nonnegative finite measures on the same alphabet, $\mu \ll \theta \ll \nu$, then

$$E_\gamma(\mu\|\nu) \leq E_1(\mu\|\theta) + E_\gamma(\theta\|\nu). \quad (5.113)$$

The Rényi divergence is not an $f$-divergence, but is a monotonic function of the Hellinger distance [42]. We have seen the definition of the Rényi divergence between two probability measures in Section 2.5.1. Below we extend the definition to the case of unnormalized measures:

Definition 5.2.4 (Rényi $\alpha$-divergence) Let $\mu$ be a nonnegative finite measure and $Q$ a probability measure on $\mathcal{X}$, $\mu \ll Q$, $X \sim Q$. For $\alpha \in (0, 1) \cup (1, +\infty)$,

$$D_\alpha(\mu\|Q) := \frac{1}{\alpha - 1} \log \mathbb{E} \left[ \left( \frac{d\mu}{dQ}(X) \right)^{\alpha} \right], \quad (5.114)$$

and

$$D_0(\mu\|Q) := \log \frac{1}{Q(\mu\|Q > -\infty)}, \quad (5.115)$$

$$D_\infty(\mu\|Q) := \mu\text{-ess sup } \mu\|Q \quad (5.116)$$

which agree with the the limits as $\alpha \downarrow 0$ and $\alpha \uparrow \infty$.

$D_\alpha(P\|Q)$ is non-negative and monotonically increasing in $\alpha$. More properties about the Rényi divergence can be found, e.g. in [163] [130] [43].

Definition 5.2.5 (smooth Rényi $\alpha$-divergence) For $\alpha \in (0, 1)$, $\epsilon \in (0, 1)$,

$$D_\alpha^{+\epsilon}(P\|Q) := \sup_{\mu \in B^\epsilon(P)} D_\alpha(\mu\|Q); \quad (5.117)$$
for $\alpha \in (1, \infty]$, $\epsilon \in (0, 1)$,

$$D^{-\epsilon}_\alpha (P \| Q) := \inf_{\mu \in B^\epsilon (P)} D_\alpha (\mu \| Q),$$

(5.118)

where $B^\epsilon (P) := \{ \mu \text{ nonnegative} : E_1 (P \| \mu) \leq \epsilon \}$ is the $\epsilon$-neighborhood of $P$ in $E_1$.

The definition of smooth $\infty$-divergence given here agrees with the smooth max Rényi divergence in [164], although the definitions look different. However, our smooth 0-divergence is different from the smooth min Rényi divergences in [165] and [164, Definition 1] except for non-atomic measures.

**Remark 5.2.1** The smooth Rényi divergence is a natural extension of the smooth Rényi entropy $H^\epsilon_\alpha$ defined in [53] (which can be viewed as a special case where the reference measure is the counting measure). In [164] the smooth min and max Rényi divergences are introduced, which correspond to the $\alpha = 0$ and $\alpha = +\infty$ cases of Definition 5.2.5. Moreover, we have introduced $+/-$ in the notation to emphasize the difference between the two possible ways of smoothing in (5.117) and (5.118).

The following quantity, which characterizes the binary hypothesis testing error, is a $g$-divergence but not a monotonic function of any $f$-divergence [42]. We will see in Proposition 5.2.4 that it is a monotonic function of the 0-smooth Rényi divergence.

**Definition 5.2.6** For nonnegative finite measures $\mu$ and $\nu$ on $\mathcal{X}$, define

$$\beta_\alpha (\mu, \nu) := \min_{\mathcal{A} : \mu (\mathcal{A}) \geq \alpha} \nu (\mathcal{A}).$$

(5.119)

---

5With the definition of smooth min Rényi divergence in [165] and [164, Definition 1], Proposition 5.2.4 would hold with the $\beta_\alpha$ as defined in (5.120) rather than (5.119).
Remark 5.2.2 In the literature, the definition of $\beta_\alpha(P, Q)$ is usually restricted to probability measures and allows randomized tests:

$$\beta_\alpha(P_W, Q_W) := \min \int P_{Z|W}(1|w)dQ_W(w)$$

(5.120)

where the minimization is over all random transformations $P_{Z|W}: Z \to \{0, 1\}$ such that

$$\int P_{Z|W}(1|w)dP_W(w) \geq \alpha.$$  

(5.121)

In contrast to $E_\gamma$, allowing randomization in the definition can change the value of $\beta_\alpha$ except for non-atomic measures. Nevertheless, many important properties of $\beta_\alpha$ are not affected by this difference.

We conclude this section with some inequalities relating those distance measures we have discussed. These results show the asymptotic equivalence among these distance metrics: $E_\gamma$ is essentially equivalent to $\bar{F}_\gamma$, and the value of $\gamma$ for which they vanish is essentially the exponential of the relative entropy; the smooth Rényi entropy is also asymptotically equivalent to the relative entropy.

**Proposition 5.2.4** Suppose $\mu$ is a finite nonnegative measure and $P$ and $Q$ are probability measures, all on $\mathcal{X}$, $X \sim P$, $\epsilon \in (0, 1)$, and $\gamma \geq 1$.

1. For $a > 1$,

$$E_\gamma(P\|Q) \leq \bar{F}_\gamma(P\|Q) \leq \frac{a}{a-1}E_\gamma^+(P\|Q).$$  

(5.122)

2. $D_\alpha^\epsilon(P\|Q)$ is increasing in $\alpha \in [0, 1]$ and $D_\alpha^{-\epsilon}(P\|Q)$ is increasing in $\alpha \in [1, \infty]$. 

164
3. If $P$ is a probability measure, then

$$D(P\|Q) \leq \int_{0}^{\infty} \mathbb{P}[r_{P\|Q}(X) > \tau] \, d\tau; \quad (5.123)$$

$$D(P\|Q) \geq E_{\gamma}(P\|Q) \log \gamma - 2e^{-1} \log e. \quad (5.124)$$

4. If $\alpha \in [0, 1)$ then

$$D_{\alpha}(\mu\|Q) \leq \log \gamma - \frac{1}{1-\alpha} \log (\mu(X) - E_{\gamma}(\mu\|Q)). \quad (5.125)$$

If $\alpha \in (1, \infty)$ then

$$D_{\alpha}(\mu\|Q) \geq \log \gamma + \frac{1}{\alpha - 1} \log E_{\gamma}(\mu\|Q). \quad (5.126)$$

5. Suppose $\alpha \in [0, 1)$, $E_{\gamma}(P\|Q) < 1 - \epsilon$, then

$$D_{\alpha}^{+\epsilon}(P\|Q) \leq \log \gamma - \frac{1}{1-\alpha} \log (1 - \epsilon - E_{\gamma}(P\|Q)). \quad (5.127)$$

Suppose $\alpha \in (1, \infty]$, $E_{\gamma}(P\|Q) > \epsilon$, then

$$D_{\alpha}^{-\epsilon}(P\|Q) \geq \log \gamma + \frac{1}{\alpha - 1} \log (E_{\gamma}(P\|Q) - \epsilon). \quad (5.128)$$

6. $D_{\infty}^{-\epsilon}(P\|Q) \leq \log \gamma \iff E_{\gamma}(P\|Q) \leq \epsilon$. That is, $D_{\infty}^{-\epsilon}(P\|Q) = \log \inf\{\gamma : E_{\gamma}(P\|Q) \leq \epsilon\}.$

7. $D_{0}^{+\epsilon}(P\|Q) = -\log \beta_{1-\epsilon}(P, Q).$

---

6 The special case of (5.125) for $\alpha = \frac{1}{2}$, $\mu(X) = 1$ and $\gamma = 1$ is equivalent to a well-known bound on a quantity called fidelity using total variation distance, see [166].
8. Fix \( \tau \in \mathbb{R} \) and \( \epsilon \geq \mathbb{P}\{t_P Q(X) \leq \tau\} \), where \( X \sim P \) and \( P \) is non-atomic. Then

\[
D_0^{+\epsilon}(P\|Q) \geq \tau + \log \frac{1}{1-\epsilon}.
\]

(5.129)

Moreover, \( D_0^{+\epsilon}(P\|Q) \geq \tau \) holds when \( P \) is not necessarily non-atomic.

**Proof**

1. The first inequality in (5.122) is evident from the definition. For the second inequality in (5.122), consider the event

\[
\mathcal{A} := \{x \in X : t_P Q(x) > \log \gamma\}.
\]

(5.130)

Then

\[
E_2 \gamma(P\|Q) \\
\geq P(\mathcal{A}) - \frac{\gamma}{a} Q(\mathcal{A}) \\
\geq \mathbb{P}\{t_P Q(X) > \log \gamma\} - \frac{\gamma}{a} \cdot \frac{1}{\gamma} \mathbb{P}\{t_P Q(X) > \log \gamma\} \\
\geq \frac{a-1}{a} \mathbb{P}\{t_P Q(X) > \log \gamma\}
\]

(5.131)

(5.132)

and the result follows by rearrangement.

2. Direct from the monotonicity of Rényi divergences in \( \alpha \) and the definitions of their smooth versions.

3. The bound (5.123) follows from \( D(P\|Q) = \mathbb{E}[t_P Q(\tilde{X})] \) and (5.124) can be seen from

\[
D(P\|Q) \geq \mathbb{E}[t_P Q(\tilde{X})] - 2e^{-1} \log e
\]

\[
\geq \log \gamma \mathbb{P}\{t_P Q(\tilde{X}) > \log \gamma\} - 2e^{-1} \log e
\]

\[
\geq \log \gamma E_\gamma(P\|Q) - 2e^{-1} \log e
\]

(5.133)

(5.134)

(5.135)
where (5.133) is due to Pinsker [167 (2.3.2)], (5.134) uses Markov’s inequality, and (5.135) is from the definition (5.80).

4. By considering the \( \frac{d\mu}{dQ} > \gamma \) and \( \frac{d\mu}{dQ} \leq \gamma \) cases separately, we can check the following (homogeneous) inequalities: for each \( \alpha < 1 \),

\[
\left| \frac{d\mu}{dQ} - \gamma \right| \geq \frac{d\mu}{dQ} - 2\gamma^{1-\alpha} \left( \frac{d\mu}{dQ} \right)^\alpha + \gamma, \tag{5.136}
\]

and for each \( \alpha > 1 \),

\[
\left| \frac{d\mu}{dQ} - \gamma \right| \leq -\frac{d\mu}{dQ} + 2\gamma^{1-\alpha} \left( \frac{d\mu}{dQ} \right)^\alpha + \gamma. \tag{5.137}
\]

Integrating with respect to \( dQ \) both sides of (5.136), we obtain

\[
|\mu - \gamma Q| \geq \mu(\mathcal{X}) - 2\gamma^{1-\alpha} \int \left( \frac{d\mu}{dQ} \right)^\alpha dQ + \gamma \tag{5.138}
\]

and (5.125) follows by rearrangement. Integrating with respect to \( dQ \) on both sides of (5.137), we obtain

\[
|\mu - \gamma Q| \leq -\mu(\mathcal{X}) + 2\gamma^{1-\alpha} \int \left( \frac{d\mu}{dQ} \right)^\alpha dQ + \gamma \tag{5.139}
\]

and (5.126) follows by rearrangement.

5. Immediate from the previous result, the definition of smooth Rényi divergence, and the triangle inequality (5.90).
6. $\Rightarrow$: By assumption there exists a nonnegative finite measure $\mu$ such that $E_1(P\|\mu) \leq \epsilon$ and $\mu \leq \gamma Q$. Then from Proposition 5.2.3

$$E_\gamma(P\|Q) \leq E_1(P\|\mu) + E_\gamma(\mu\|Q) \leq \epsilon + 0.$$ 

$\Leftarrow$: Define $\mu$ by $\frac{d\mu}{dQ} := \min\{\frac{dP}{dQ}, \gamma\}$. Since $\left\{\frac{dP}{dQ} > \gamma\right\} = \left\{\frac{dP}{d\mu} > 1\right\}$,

$$E_1(P\|\mu) = P\left(\frac{dP}{dQ} > \gamma\right) - \mu\left(\frac{dP}{dQ} > \gamma\right) = P\left(\frac{dP}{dQ} > \gamma\right) - \gamma Q\left(\frac{dP}{dQ} > \gamma\right) \leq E_\gamma(P\|Q).$$ 

Then $D_\infty(\mu\|Q) \leq \log \gamma$ implies that $D_{\infty}^{-\epsilon}(P\|Q) \leq \log \gamma$.

7. "$\Rightarrow$": let $A$ be a set achieving the minimum in (5.119). Let $\mu$ be the restriction of $P$ on $A$. Then by definition (5.119) we have

$$E_1(P\|\mu) = P(A^c) \leq \epsilon \quad (5.140)$$

therefore

$$D_0^\pm(P\|Q) \geq -\log (1 - \epsilon)(P, Q). \quad (5.141)$$

"$\Leftarrow$": fix arbitrary $\delta > 0$ and let $\mu$ be a nonnegative measure satisfying

$$D_0^{-\epsilon}(P\|Q) < D_0(\mu\|Q) + \delta. \quad (5.142)$$
Define $\mathcal{A} := \text{supp}(\mu)$. Then

$$P(\mathcal{A}^c) = P(\mathcal{A}^c) - \mu(\mathcal{A}^c)$$

$$\leq E_1(P\|\mu)$$

$$\leq \epsilon,$$  \hspace{1cm} (5.143 - 5.145)

hence

$$D_0^+\epsilon(P\|Q) - \delta \leq D_0(\mu\|Q)$$

$$= -\log Q(\mathcal{A})$$

$$\leq -\log \beta_{1-\epsilon}(P, Q)$$  \hspace{1cm} (5.146 - 5.148)

and the result follows by setting $\delta \downarrow 0$.

8. In the case of non-atomic $P$, there exists a set $\mathcal{A} \subseteq \{x : \nu_{P\|Q}(x) > \tau\}$ such that $P(\mathcal{A}) = 1 - \epsilon$. Then

$$\beta_{1-\epsilon}(P, Q) \leq Q(\mathcal{A})$$

$$\leq \exp(-\tau)P(\mathcal{A}).$$  \hspace{1cm} (5.149 - 5.150)

We obtain (5.129) by taking the logarithms on both sides of the above and invoking Part 7. When $P$ is not necessarily non-atomic, since $P(\nu_{P\|Q} > \tau) \geq 1 - \epsilon,$

$$\beta_{1-\epsilon}(P, Q) \leq Q(\nu_{P\|Q} > \tau)$$

$$\leq \exp(-\tau)P(\nu_{P\|Q} > \tau)$$

$$\leq \exp(-\tau)$$  \hspace{1cm} (5.151 - 5.153)
and again the result follows by taking the logarithms on both sides of the above.

5.2.3 $E_\gamma$-Resolvability

Given the close connection between $E_\gamma$ and the smooth Rényi divergence, and the variational formula (5.88), we expect that $E_\gamma$ has a great potential in various venues of the non-asymptotic information theory. However, in this thesis we only discuss the usage of $E_\gamma$ in the setting of channel resolvability.

After an introduction to the resolvability problem and the formulation of its extension to the $E_\gamma$ setting, we present a general one-shot achievability bound on $E_\gamma$-resolvability, which is the basis of all information-theoretic of $E_\gamma$ that we will discuss in this thesis. Then, a refinement of such a result using concentration inequalities leads to a one-shot tail bound on the approximation error for random codebooks, which is useful for some problems in the information-theoretic literature [168]. Finally, we argue the asymptotic tightness of the one-shot bounds in the discrete memoryless settings using the method of types.

Later, we will discuss the applications of the results on $E_\gamma$-resolvability in non-asymptotic information theory (Sections 5.2.4-5.2.6).

The Resolvability Problem: An Introduction

Channel resolvability, introduced by Han and Verdú [155], is defined as the minimum randomness rate required to synthesize an input so that its corresponding output distribution approximates a target output distribution. While the resolvability problem itself differs from classical topics in information theory such as data compression and transmission, [155] unveils its potential utility in operational problems through the solution of the strong converse problem of identification coding [22]. Other applications of distribution approximation in information theory include common random-
ness of two random variables [6], strong converse in identification through channels [155], random process simulation [169], secrecy [170] [171] [172] [173], channel synthesis [174] [159], lossless and lossy source coding [169] [155] [160], and the empirical distribution of a capacity-achieving code [175] [149]. The achievability part of resolvability (also known as the soft-covering lemma in [159]) is particularly useful, and coding theorems via resolvability have certain advantages over what is obtained from traditional typicality-based approaches (see e.g. [172]).

If the channel is stationary memoryless and the target output distribution is induced by a stationary memoryless input, then the resolvability is the minimum mutual information over all input distributions inducing the (per-letter) target output distribution, no matter when the approximation error is measured in total variation distance [155] [176], Theorem 1], normalized relative entropy [6], Theorem 6.3] [155], or unnormalized relative entropy [173]. In contrast, relatively few measures for the quality of the approximation of output statistics have been proposed for which the resolvability can be strictly smaller than mutual information. As shown by Steinberg and Verdú [169], one exception is the Wasserstein distance measure, in which case the finite precision resolvability for the identity channel can be related to the rate-distortion function of the source where the distortion is the metric in the definition of the Wasserstein distance [169].

We remark that in view of convex duality, channel resolvability in the relative entropy is equivalent to (our new definition of) the image-size problem in a precise sense (Proposition 5.1.5): there exist an input distribution supported on a set of a given size whose output approximates a given output distribution if and only if the “image” of that set takes up almost all of that output distribution.

In this section we generalize the theory of resolvability by considering a distance measure, \( E_\gamma \), defined in Section 5.2.1, of which the total variation distance is a special
case where $\gamma = 1$. The $E_\gamma$ metric\footnote{"Metric" or "distance" are used informally since, other than nonnegativity, $E_\gamma$, in general, does not satisfy any of the other three requirements for a metric.} is more forgiving than total variation distance when $\gamma > 1$, and the larger $\gamma$ is, the less randomness is needed at the input for approximation in $E_\gamma$. Various achievability and converse bounds for resolvability in the $E_\gamma$ metric will be derived.

- In the case of a discrete memoryless channel with a given stationary memoryless target output, we provide a single-letter characterization of the minimum exponential growth rate of $\gamma$ to achieve approximation in $E_\gamma$.

- In addition to achievability results in terms of the expectation of the approximation error over a random codebook, we also prove achievability under the high probability criteria\footnote{More precisely, by "the high probability criteria" we that mean that the approximation error satisfies a Gaussian concentration result (which ensures a doubly exponential decay in blocklength when the single-shot result is applied to the multi-letter setting).} for a random codebook. The implications of the latter problem in secrecy has been noted by several authors [168][177][178]. Here we adopt a simple non-asymptotic approach based on concentration inequalities, dispensing with the finiteness or stationarity assumptions on the alphabet required by the previous proof method [168] based on Chernoff bounds.

- An asymptotically matching converse bound is derived for the stationary memoryless case using the method of types. The analysis is different from the previous proof for the $\gamma = 1$ (total variation) special case and, in particular, we need to introduce a slightly new definition of conditional typicality [18] for this proof.

We remark that while the conventional resolvability theory ($\gamma = 1$) also considers the amount of randomness needed to approximate the worst case output distribution, we do not discuss the worst case counterpart for general $\gamma$ in this thesis. However, upper and lower bounds for the worst case $E_\gamma$-resolvability can be found in our journal version [143].
Problem Formulation

The setup in Figure 5.3 is the same as in the original paper on channel resolvability under total variation distance [155]. Given a random transformation $Q_{X|U}$ and a target distribution $\pi_X$, we wish to minimize the size $M$ of a codebook $c^M = (c_m)_{m=1}^M$ such that when the codewords are equiprobably selected, the output distribution

$$Q_{X[c^M]} := \frac{1}{M} \sum_{m=1}^M Q_{X|U=c_m}$$  \hspace{1cm} (5.154)

approximates $\pi_X$. The difference from [155] is that we use $E_\gamma$ (and other metrics) to measure the level of the approximation.

Figure 5.3: Setup for channel resolvability.

The fundamental one-shot tradeoff is the minimum $M$ required for a prespecified degree of approximation. One-shot bounds are general in that no structural assumptions on either the random transformation or the target distribution are imposed. The corresponding asymptotic results can usually be recovered quite simply from the one-shot bounds using, say, the law of large numbers in the memoryless case. Asymptotic limits are of interest because of the compactness of the expressions, and because good one-shot bounds are not always known, especially in the converse parts.

Unless otherwise stated, we do not restrict the alphabets to be finite or countable. All one-shot achievability bounds for $E_\gamma$-resolvability apply to general alphabets; some asymptotic converse bounds assume finite input alphabets or discrete memoryless channels (DMC).

---

9A DMC is a stationary memoryless channel whose input and output alphabets are finite, which is denoted by the corresponding per-letter random transformation (such as $Q_{X|U}$) in this chapter.
Next, we define the achievable regions in the general asymptotic setting (when sources and channels are arbitrary, see [155][171]) as well as the case of stationary memoryless channels and memoryless outputs. Boldface letters such as $X$ denote a general sequence of random variables $(X^n)_{n=1}^\infty$, and sanserif letters such as $X$ denote the generic distributions in iid settings.

**Definition 5.2.7** Given a channel $10 (Q_{X^n|U^n})_{n=1}^\infty$ and a sequence of target distributions $(\pi_{X^n})_{n=1}^\infty$, the triple $(G, R, X)$ is $\epsilon$-achievable ($0 < \epsilon < 1$) if there exists $(c^M_n)_{n=1}^\infty$, where $c^M_n := (c_1, \ldots, c_{M_n})$ and $c_m \in \mathcal{U}^n$ for each $m = 1, \ldots, M_n$, and $(\gamma_n)_{n=1}^\infty$, so that

\[
\limsup_{n \to \infty} \frac{1}{n} \log M_n \leq R; \quad (5.155)
\]
\[
\limsup_{n \to \infty} \frac{1}{n} \log \gamma_n \leq G; \quad (5.156)
\]
\[
\sup_n E_{\gamma_n} (Q_{X^n|c^M_n} \| \pi_{X^n}) \leq \epsilon. \quad (5.157)
\]

Moreover, $(G, R, X)$ is said to be achievable if it is $\epsilon$-achievable for all $0 < \epsilon < 1$.

Define the asymptotic fundamental limits $11$

\[
G_\epsilon(R, X) := \min \{ g : (g, R, X) \text{ is } \epsilon\text{-achievable} \}; \quad (5.158)
\]
\[
G(R, X) := \sup_{\epsilon > 0} G_\epsilon(R, X), \quad (5.159)
\]

and

\[
S_\epsilon(G, X) := \min \{ r : (G, r, X) \text{ is } \epsilon\text{-achievable} \}; \quad (5.160)
\]
\[
S(G, X) := \sup_{\epsilon > 0} S_\epsilon(G, X), \quad (5.161)
\]

\[10]\text{In this setting a “channel” refers to a sequence of random transformations.}\]
\[11]\text{We can write } \min \text{ instead of } \inf \text{ in (5.159) and (5.160) since for fixed } X \text{ the set of } (G, R) \text{ such that } (G, R, X) \text{ is } \epsilon\text{-achievable is necessarily closed.}\]
which, in keeping with [155], we refer to as the resolvability function. In the special case of $Q_{X|U^n} = Q_{X|U}^{\otimes n}$ and target $\pi_X = \pi_X^{\otimes n}$, we may write the quantities in (5.159) and (5.160) as $G(R, \pi_X)$ and $S(G, \pi_X)$.

Note that by [155], $(0, R, \pi_X)$ is achievable if and only if $R \geq I(P_U, Q_{X|U})$ for some $P_U$ satisfying $P_U \rightarrow Q_{X|U} \rightarrow \pi_X$.

A useful property is that the approximation error (5.157) actually converges uniformly. A similar observation was made in the proof of [155, Lemma 6] in the context of resolvability in total variation distance.

**Proposition 5.2.5** If $(G, R)$ is $\epsilon$-achievable for $(Q_{X^n|U^n})_{n=1}^{\infty}$, then there exists $(\gamma_n)_{n=1}^{\infty}$ and $(M_n)_{n=1}^{\infty}$ such that

$$\limsup_{n \to \infty} \frac{1}{n} \log \gamma_n \leq G;$$

$$\limsup_{n \to \infty} \frac{1}{n} \log M_n \leq R;$$

$$\sup_n \sup_{P_U, c_{M_n}} \inf_{Q_{X^n}} E_{\gamma_n} (Q_{X^n[Q_{X^n}]} \| Q_{X^n}) \leq \epsilon,$$

where $Q_{U^n} \rightarrow Q_{X^n|U^n} \rightarrow Q_{X^n}$.

**Proof** Fix arbitrary $G' > G$ and $R' > R$. Observe that the sequence

$$\sup_{Q_{U^n}} \inf_{c_{M_n}} E_{\exp(nG')} (Q_{X^n[Q_{X^n}]} \| Q_{X^n})$$

(5.165)

where $M_n := \exp(nR')$, must be upper-bounded by $\epsilon$ for large enough $n$. For if otherwise, there would be a sequence $(Q_{U^n})_{n=1}^{\infty}$ such that the infimum in (5.165) is not upper-bounded by $\epsilon$ for large enough $n$, which is a contradiction since we can find $(c_{M_n})_{n=1}^{\infty}$ and $(\gamma_n)_{n=1}^{\infty}$ in Definition 5.2.7 such that $M_n < \exp(nR')$ and $\gamma_n \leq \exp(nG')$ for $n$ large enough, and apply the monotonicity of $E_{\gamma}$ in $\gamma$. Finally, since $(G', R')$ can be arbitrarily close to $(G, R)$ and the $\epsilon$-achievable region is a closed set, we conclude that $(G, R)$ is $\epsilon$-achievable.
One-shot Achievability Bounds

While our goal is to bound $E_\gamma$ in the resolvability problem, we consider instead the problem of bounding the complementary relative information spectrum $\overline{F}$, which, according to Proposition 5.2.4, is (up to constant factors) equivalent to bounding $E_\gamma$.

First, we study a simple special case where the target distribution matches the input distribution according to which the codewords are generated.

Theorem 5.2.6 (Softer-covering Lemma) Fix $Q_{UX} = Q_UQ_{X|U}$. For an arbitrary codebook $[c_1, \ldots, c_M]$, define

$$Q_X[c^M] := \frac{1}{M} \sum_{m=1}^M Q_{X|U=c_m}. \quad (5.166)$$

Then for any $\gamma, \epsilon, \sigma > 0$ satisfying $\gamma - 1 > \epsilon + \sigma$ and $\tau \in \mathbb{R}$,

$$\mathbb{E}[\overline{F}_\gamma (Q_{X|U^M} \| Q_X)] \leq \mathbb{P}[u_{U;X}(U;X) > \log M\sigma]$$

$$+ \frac{1}{\epsilon} \mathbb{P}[u_{U;X}(U;X) > \log M - \tau]$$

$$+ \frac{\exp(-\tau)}{(\gamma - 1 - \epsilon - \sigma)^2} \quad (5.167)$$

where $U^M \sim Q_U \times \cdots \times Q_U$, $(U;X) \sim Q_UQ_{X|U}$, and the information density

$$u_{U;X}(u; x) := \log \frac{dQ_{X|U=u}}{dQ_X}(x). \quad (5.168)$$

Remark 5.2.3 Theorem 5.2.6 implies (the general asymptotic version of) the soft-covering lemma based on total variation (see [155], [171], [159]). Indeed if $M_n = \exp(nR)$ at blocklength $n$ where $R > \overline{I}(U;X)$, we can select $\tau_n \leftarrow \frac{nR}{2}(R - \overline{I}(U;X))$. Moreover for any $\gamma > 1$ we can pick constant $\epsilon, \sigma > 0$ such that $\gamma - 1 > \epsilon + \sigma$ in the
theorem, to show that

$$
\lim_{n \to \infty} \mathbb{E}[\tilde{F}_\gamma(Q_{X^n[U^M,\alpha]} \| Q_{X^n})] = 0
$$

(5.169)

which, by (5.95) and by taking $\gamma \downarrow 1$, implies that

$$
\lim_{n \to \infty} \mathbb{E}|Q_{X^n[U^M,\alpha]} - Q_{X^n}| = 0.
$$

(5.170)

We refer to Theorem 5.2.6 as the softer-covering lemma since for a larger $\gamma$ it allows us to use a smaller codebook to cover the output distribution more softly (i.e. approximate the target distribution under a weaker metric).

**Proof** Define the “atypical” set

$$
\mathcal{A}_\tau := \{(u, x) : \nu_{U; X}(u; x) \leq \log M - \tau\}.
$$

(5.171)

Now, let $X^M$ be such that $(U^M, X^M) \sim Q_{UX} \times \cdots \times Q_{UX}$, The joint distribution of $(U^M, \hat{X})$ is specified by letting $\hat{X} \sim Q_{X[c^M]}$ conditioned on $U^M = c^M$. We perform a change-of-measure step using the symmetry of the random codebook:

$$
\mathbb{E}[\tilde{F}_\gamma(Q_{X[U^M]} \| Q_X)] = \mathbb{P}\left[\frac{dQ_{X[U^M]}}{dQ_X}(\hat{X}) > \gamma\right]
$$

(5.172)

$$
= \frac{1}{M} \sum_m \mathbb{P}\left[\frac{dQ_{X[U^M]}}{dQ_X}(X_m) > \gamma\right]
$$

(5.173)

$$
= \mathbb{P}\left[\frac{dQ_{X[U^M]}}{dQ_X}(X_1) > \gamma\right]
$$

(5.174)

where (5.173) is because of (5.166), and (5.174) is because the summands in (5.173) are equal. Note that $X_1$ is correlated with only the first codeword $U_1$. 

177
Next, because of the relation

\[
\frac{dQ_{X_c^{\cap M}}}{dQ_X}(x) = \frac{1}{M} \sum_{m} \exp(i_{U,X}(c_m, x)),
\]  

(5.175)

we can upper-bound (5.174) by the union bound as

\[
\mathbb{P}[\exp(i_{U,X}(U_1; X_1)) > M\sigma] \\
+ \mathbb{P} \left[ \frac{1}{M} \sum_{m=2}^{M} \exp(i_{U,X}(U_m; X_1)) 1_{A^c}(U_m, X_1) > \epsilon \right] \\
+ \mathbb{P} \left[ \frac{1}{M} \sum_{m=2}^{M} \exp(i_{U,X}(U_m; X_1)) 1_{A}(U_m, X_1) > \gamma - \epsilon - \sigma \right]
\]  

(5.176)

where we used the fact that \(1_{A} + 1_{A^c} = 1\). Notice that the first term of (5.176) may be regarded as the probability of “atypical” events and accounts for the first term in (5.167). The second term of (5.176) can be upper-bounded with Markov’s inequality:

\[
\frac{1}{M\epsilon} \sum_{m=2}^{M} \mathbb{E}[\exp (i_{U,X}(U_m; X_1)) 1_{A^c}(U_m, X_1)] \\
\leq \frac{1}{M\epsilon} \sum_{m=2}^{M} \mathbb{E}[1_{A^c}(U, X)] \\
\leq \frac{1}{\epsilon} \mathbb{P} [i_{U,X}(U; X) > \log M - \tau]
\]  

(5.177)

accounting for the second term in (5.167) where (5.177) is a change-of-measure step using the fact that \((U_1, X_1) \sim Q_U \times Q_X\) for \(m \geq 2\). 

178
Finally we take care of the last term in (5.176), again using the independence of $U_m$ and $X_1$ for $m \geq 2$. Observe that for any $x \in \mathcal{X}$,

$$
\mu := \mathbb{E} \left[ \frac{1}{M} \sum_{m=2}^{M} \exp(i U(X_m; x)) 1_{A_r}(U_m, x) \right] \\
\leq \frac{1}{M} \sum_{m=2}^{M} \mathbb{E} [\exp(i U(X_m; x))] \\
= \frac{M - 1}{M} \\
\leq 1,
$$

whereas

$$
\text{Var} \left( \frac{1}{M} \sum_{m=2}^{M} \exp(i U(X_m; x)) 1_{A_r}(U_m, x) \right) \\
= \frac{1}{M^2} \sum_{m=2}^{M} \text{Var} (\exp(i U(X_m; x)) 1_{A_r}(U_m, x)) \\
\leq \frac{1}{M} \text{Var} (\exp(i U; x)) 1_{A_r}(U, x) \\
\leq \frac{1}{M} \mathbb{E} [\exp(2i U; x)) 1_{A_r}(U, x)] \\
\leq \exp(-\tau) \mathbb{E} [\exp(i U; x))] \\
= \exp(-\tau),
$$

where the change of measure step (5.186) uses (5.171). It then follows from Chebyshev’s inequality that

$$
\mathbb{P} \left[ \frac{1}{M} \sum_{m=2}^{M} \exp(i U(X_m; x)) 1_{A_r}(U_m, x) > \gamma - \epsilon - \sigma \right] \\
\leq \mathbb{P} \left[ \frac{1}{M} \sum_{m=2}^{M} \exp(i U(X_m; x)) 1_{A_r}(U_m, x) - \mu > \gamma - \epsilon - \sigma - 1 \right] \\
\leq \exp(-\tau) \frac{\exp(-\tau)}{(\gamma - \epsilon - \sigma - 1)^2}.
$$
For the asymptotic analysis, we are interested in the regime where $M$ and $\gamma$ are growing exponentially. In this case, the right hand side of \( 5.167 \) can be regarded as essentially

\[
\mathbb{P}[I_{U;X}(U;X) > \log M \gamma]
\]

modulo nuisance parameters. This can be seen from the choice of parameters in Corollary \[5.2.8\]. Thus the sum rate of $M$ and $\gamma$ has to exceed the sup information rate in order that the approximation error vanishes asymptotically.

Extending Theorem \[5.2.6\] to the more general scenario where the target distribution may not have any relation with the input distribution, we have the following result, where we allow $\pi_X$ to be an arbitrary positive measure,

**Theorem 5.2.7 (Softer-covering Lemma: Unmatched Target Distribution)**

Fix $\pi_X$ and $Q_{UX} = Q_U Q_{X|U}$. For an arbitrary codebook $[c_1, \ldots, c_M]$, define $Q_{X|M}$ as in \( 5.166 \). Then for any $\gamma, \epsilon, \sigma > 0$ satisfying $\gamma > \epsilon + \sigma$, $\tau \in \mathbb{R}$ and $0 < \delta < 1$,

\[
\mathbb{E}[\tilde{F}_{\gamma}(Q_{X|M}\|\pi_X)] \\
\leq \mathbb{P}[i_{Q_X|\pi_X}(X) > \log(\gamma - \sigma - \epsilon) \text{ or } I_{U;X}(U;X) \\
+ i_{Q_X|\pi_X}(X) > \log \delta M \sigma] \\
+ \frac{\gamma - \epsilon - \sigma}{\epsilon} \mathbb{P}[I_{U;X}(U;X) > \log M - \tau] \\
+ \frac{\exp(-\tau)(\gamma - \epsilon - \sigma)^2}{(1 - \delta)^2 \sigma^2}
\]

(5.192)

where $U^M \sim Q_U \times \cdots \times Q_U$ and $(U, X) \sim Q_U Q_{X|U}$.

**Proof Sketch** Similarly to the proof of Theorem \[5.2.6\], we first use a symmetry argument and change-of-measure step so that the random variable of the channel
output is correlated only with the first codeword, to obtain

\[ \mathbb{E}[\tilde{F}_\gamma (Q_{X[U^M]} \| \pi_X)] \leq \mathbb{P}\left[ \frac{dQ_{X[U^M]} }{d\pi_X}(X_1) > \gamma \right]. \] (5.193)

Then in the next union bound step we have to take care of another “atypical” event that \( \frac{dQ_x}{d\pi_x}(X_1) > \gamma_2 \), where

\[ \gamma_2 := \gamma - \epsilon - \sigma. \] (5.194)

More precisely, we have

\[ \mathbb{P}\left[ \frac{dQ_{X[U^M]} }{d\pi_X}(X_1) > \gamma \right] \]

\[ \leq \mathbb{P}[\xi(X_1) > \gamma_2 \text{ or } \eta(U_1, X_1) > \delta M \sigma] \]

\[ + \mathbb{P}\left[ \frac{1}{M} \sum_{m=2}^{M} \eta(U_m, X_1)1_{A_\xi}(U_m, X_1) > \epsilon, \xi(X_1) \leq \gamma_2 \right] \]

\[ + \mathbb{P}\left[ \frac{1}{M} \sum_{m=2}^{M} \eta(U_m, X_1)1_{A_\xi}(U_m, X_1) > \gamma - \epsilon - \delta \sigma, \xi(X_1) \leq \gamma_2 \right] \] (5.195)

where we have defined

\[ \eta(u, x) := \frac{dQ_{X[U^M]} }{d\pi_X}(x); \] (5.196)

\[ \xi(x) := \frac{dQ_X }{d\pi_X}(x). \] (5.197)

As before the first term of (5.195) may be regarded as the probability of “atypical” events and accounts for the first term in (5.192). The second and the third terms of
(5.195) can be upper-bounded by

\[ P \left[ \frac{1}{M} \sum_{m=2}^{M} \frac{\eta(U_m, X_1)}{\xi(X_1)} 1_{A_{r'}}(U_m, X_1) > \frac{\epsilon}{\gamma^2} \right] \]

\[ \leq P \left[ \frac{1}{M} \sum_{m=2}^{M} \exp(\eta(U_m; X_1)) 1_{A_{r'}}(U_m, X_1) > \frac{\epsilon}{\gamma^2} \right] \] (5.198)

and

\[ P \left[ \frac{1}{M} \sum_{m=2}^{M} \frac{\eta(U_m, X_1)}{\xi(X_1)} 1_{A_{r'}}(U_m, X_1) > \frac{\gamma - \epsilon - \delta \sigma}{\gamma^2} \right] \]

\[ \leq P \left[ \frac{1}{M} \sum_{m=2}^{M} \exp(\eta(U_m; X_1)) 1_{A_{r'}}(U_m, X_1) > 1 + \frac{(1 - \delta)\sigma}{\gamma^2} \right] . \] (5.199)

The rest of the proof is similar to that of Theorem 5.2.6 and is omitted. ■

For the purpose of asymptotic analysis in the stationary memoryless setting, the right hand side of (5.192) can be regarded as essentially

\[ P \left[ I_{Q_X \| \pi_X} (X) > \log \gamma \right] \]

\[ + P \left[ I_{U; X} (U; X) + I_{Q_X \| \pi_X} (X) > \log M \gamma \right] \] (5.200)

modulo nuisance parameters.

**Remark 5.2.4** By setting \( \tau \leftarrow +\infty \) and letting \( \delta \uparrow 1 \), the bound in Theorem 5.2.7 can be weakened in the following slightly simpler form:

\[ \mathbb{E} [ F_{\gamma}(Q_{X[U^M]} \| \pi_X)] \leq P \left[ \frac{dQ_X}{d\pi_X} (X) > \gamma - \sigma - \epsilon \right] \]

\[ + P \left[ \frac{dQ_{X|U}}{d\pi_X} (X|U) \geq M\sigma \right] \]

\[ + \frac{\gamma - \sigma - \epsilon}{\epsilon} . \] (5.201)
In fact, assuming $\tau = +\infty$ we can simplify the proof of the theorem and strengthen \((5.201)\) to

$$
\mathbb{E}[\hat{F}_\gamma(Q_{X[U^M]}\|\pi_X)] \leq \mathbb{P}
\left[
\frac{dQ_X}{d\pi_X}(X) > \gamma_2
\right] + \mathbb{P}
\left[
\frac{dQ_X}{d\pi_X}(X|U) > M(\gamma - \epsilon)
\right] + \frac{\gamma_2}{\epsilon}
\tag{5.202}
$$

for any $\gamma_2 > 0$ and $0 < \epsilon < \gamma$. As we show in Corollary \(5.2.5\), the weakened bounds \((5.201)\) and \((5.202)\) are still asymptotically tight provided that the exponent with which the threshold $\gamma$ grows is strictly positive. However, when the exponent is zero (corresponding to the total variation case), we do need $\tau$ in the bound for asymptotic tightness.

**Remark 5.2.5** In the case of $Q_X = \pi_X$, we can set $\gamma_2 \leftarrow 1$ and $\epsilon \leftarrow \gamma_2$ in \(5.202\) to obtain the simplification

$$
\mathbb{E}[\hat{F}_\gamma(Q_{X[U^M]}\|Q_X)] \leq \mathbb{P}
\left[
\exp\{\nu_U(X;X)\} > \log \frac{M\gamma}{2}
\right] + \frac{2}{\gamma}
\tag{5.203}
$$

The general one-shot achievability bound in Theorem \(5.2.7\) implies the following asymptotic result, which we will later prove to be tight.

**Corollary 5.2.8** Fix per-letter distributions $\pi_X$ on $\mathcal{X}$ and $Q_{UX} = Q_UQ_{X|U}$ on $\mathcal{U} \times \mathcal{X}$, and $E,R \in (0,\infty)$. For each $n$, define $\gamma_n := \exp(nE)$ and $M_n = \lceil\exp(nR)\rceil$; let $U^{M_n} = (U_1,\ldots,U_{M_n})$ have independent coordinates each distributed according to $Q_{U^n}$. 

183
Given any $c^{M_n} = (c_m)_{m=1}^{M_n}$, where each $c_m \in \mathcal{U}^n$, define

$$Q_X[c^{M_n}] := \frac{1}{M_n} \sum_{m=1}^{M_n} Q_{X|U}^m(d|c_m).$$

(5.204)

Then

$$\lim_{n \to \infty} \mathbb{E}[E_{\gamma_n}(Q_X[U^{M_n}]||\pi_X^{\otimes n})] = \lim_{n \to \infty} \mathbb{E}[\bar{F}_{\gamma_n}(Q_X[U^{M_n}]||\pi_X^{\otimes n})] = 0$$

(5.205)

provided that

$$E > D(Q_X||\pi_X) + [I(Q_U, Q_X|U) - R]^+.$$  

(5.206)

Proof Choose $E'$ such that

$$E > E' > D(Q_X||\pi_X) + [I(Q_U, Q_X|U) - R]^+.$$  

(5.207)

Set $\delta = \frac{1}{2}$, $\epsilon_n = \exp(nE) - \exp(nE')$ and $\sigma_n = \frac{1}{2}(\gamma_n - \epsilon_n) = \frac{1}{2}\exp(nE')$, and apply (5.201). Notice that

$$\mathbb{E} \left[ \log \frac{dQ_X|U}{d\pi_X}(X|U) \right] = n[I(Q_U, Q_X|U) + D(Q_X||\pi_X)]$$

(5.209)

12We define $Q_{X|U}^n$ by $Q_{X|U}^n(d|u^n) := \prod_{i=1}^{n} Q_{X|U=u_i}$ for any $u^n$. Also note that in this chapter we differentiates per-letter symbols such as $U$ between one-shot/block symbols such as $U$ (so that $U = U^n$ in this corollary).
where \((X, U) \sim Q_{X\mid U}^m\), \(Q_{X\mid U} := Q_{X\mid U}^m\), and \(\pi_X := \pi_X^m\). By the law of large numbers, the first and second terms in (5.201) vanish because

\[
\begin{align*}
D(Q_X\|\pi_X) &< E'; \\ I(Q_U, Q_{X|U}) + D(Q_X\|\pi_X) &< E' + R
\end{align*}
\]

are satisfied.

The \(Q_U\) that minimizes the right hand side of (5.207) generally does not satisfy \(Q_U \rightarrow Q_{X|U} \rightarrow Q_X\). This means that in the large deviation analysis, for the best approximation of a target distribution in \(E_\gamma\), we generally should not generate the codewords according to a distribution that corresponds to the target through the channel. This is a remarkable distinction from approximation in total variation distance, in which case an unmatched input distribution would result in the maximal total variation distance asymptotically. However, if we stick to matching input codeword distributions, then a simple and general asymptotic achievability bound can be obtained. Recall that [155] defined the sup-information rate

\[
\bar{I}(U; X) := \inf \left\{ R : \lim_{n \to \infty} \mathbb{P} \left[ \frac{1}{n} I_{U^n; X^n}(U^n; X^n) > R \right] = 0 \right\}
\]

and the inf-information rate

\[
I(U; X) := \sup \left\{ R : \lim_{n \to \infty} \mathbb{P} \left[ \frac{1}{n} I_{U^n; X^n}(U^n; X^n) < R \right] = 0 \right\}.
\]
Theorem 5.2.9 For any channel $W = (Q_{X^n|U^n})_{n=1}^{\infty}$, sequence of inputs $U = (U^n)_{n=1}^{\infty}$, and $G > 0$, we have

$$S(G, X) \leq \tilde{I}(U; X) - G$$  \hspace{1cm} (5.214)

where $X$ is the output of $U$ through the channel $W$. As a consequence,

$$S(G) \leq \sup_U \tilde{I}(U; X) - G$$  \hspace{1cm} (5.215)

For channels satisfying the strong converse property, the right hand side of (5.215) can be related to the channel capacity because of the relations $\sup_U \tilde{I}(U; X) = \sup_U I(U; X) = C(W)$.

(5.216)

We conclude the subsection by remarking that had we used the soft-covering lemma (see Remark 5.2.3) to bound total variation distance and in turn, bounded the complementary relative information spectrum with total variation distance, we would not have obtained Theorem 5.2.9. Indeed, consider $M_n = \exp(nR)$ and let $V_1, \ldots, V_{M_n}$ be i.i.d. according to $Q_{U^n}$. Regardless of how fast $\gamma_n$ grows, we cannot conclude from that

$$\mathbb{E}[\tilde{F}_{\gamma_n}(Q_{X^n[V_{M_n}]}||\pi_{X^n})]$$  \hspace{1cm} (5.217)

vanishes unless we show that $\mathbb{E}|Q_{X^n[V_{M_n}]} - \pi_{X^n}|$ vanishes (the general and tight bounds between $E_{\gamma}$ and the total variation distance are characterized in [143]). Since the total variation distance vanishes only when $R > \tilde{I}(U; X)$ by the conventional resolvability theorem, the conventional resolvability theorem only gives an upper bound $\tilde{I}(U; X)$ which is looser than Theorem 5.2.9 when $G > 0$. 

186
Tail Bound of Approximation Error for Random Codebooks

For applications such as secrecy and channel synthesis, it is sometimes desirable to prove that the approximation error vanishes not only in expectation (e.g. Theorem 5.2.7), but also with high probability (see Footnote 8), in the case of a random codebook [168] [177] [178]. If the probability that the approximation error exceeds an arbitrary positive number vanishes doubly exponentially in the blocklength, then the analyses in these applications carry through because a union bound argument can be applied to exponentially many events. Previous proofs (e.g. [168]) based on carefully applying Chernoff bounds to each $Q_{X|U}^r(x) - Q_X(x)$ and then taking the union bound over $x$ require finiteness of the alphabets.

Here we adopt a different approach. Using concentration inequalities we can directly bound the probability that the error $E_\gamma(Q_{X|U}^r\|Q_X)$ deviates from its expectation, without any restrictions on the alphabet and in fact the bound only depends on the number of codewords. Therefore if the rate is high enough for the approximation error to vanish in expectation (by Theorem 5.2.7), we can also conclude that the error vanishes with high probability. The crux of the matter is thus resolved by the following one-shot result:

**Theorem 5.2.10** Fix $\pi_X$ and $Q_{UX} = Q_UQ_X|U$. For an arbitrary codebook $[c_1, \ldots, c_M]$, define $Q_{X[c^M]}$ as in (5.166). Then, for any $r > 0$,

$$\mathbb{P}[E_\gamma(Q_{X|U}^r\|\pi_X) - \mathbb{E}[E_\gamma(Q_{X|U}^r\|\pi_X)] > r] \leq \exp(-2Mr^2) \quad (5.218)$$

where the probability and the expectation are with respect to $U^M \sim Q_X \times Q_X$.

**Proof** Consider $f: c^M \mapsto E_\gamma(Q_{X[c^M]}\|\pi_X)$. By the definition (5.166) and the triangle inequality (5.90), we have the following uniform bound on the discrete derivative:

$$\sup_{c, c' \in X} |f(c_i, c^M) - f(c_i, c^M')| \leq \frac{1}{M}, \quad \forall i, c^M. \quad (5.219)$$
The result then follows by McDiarmid’s inequality (see e.g. [48, Theorem 2.2.3]). □

**Remark 5.2.6** If we are interested in bounding both the upper and the lower tails then the right side of (5.218) gains a factors of 2. Other concentration inequalities may also be applied here; the transportation method gives the same bound in this example.

**Converse for Stationary Memoryless Outputs**

We now establish a matching asymptotic converse about Corollary 5.2.8 for the $E_{\gamma}$-resolvability rate for stationary memoryless outputs and discrete memoryless channels.

**Theorem 5.2.11 (Resolvability for Stationary Memoryless Outputs)** For a DMC $Q_{X|U}$ and a nonnegative finite measure $\pi_X$, 

$$G_\epsilon(R, \pi_X) = \min_{Q_U} \left\{ D(Q_X\|\pi_X) + \left[ I(Q_U, Q_{X|U}) - R \right]^+ \right\}$$ \hspace{1cm} (5.220) 

where $Q_U \rightarrow Q_{X|U} \rightarrow Q_X$, for any $0 < \epsilon < 1$.

**Remark 5.2.7** When resolvability was introduced in [155], the resolvability rate (under total variation distance) was formulated for outputs of stationary memoryless inputs, rather than all the tensor power distributions on the output alphabet, because otherwise there is no guarantee that the output process can be approximated under total variation distance even with an arbitrarily large codebook. Here we can extend the scope because all stationary memoryless distributions on the output alphabet (satisfying the mild condition of being absolutely continuous with respect to some output) can be approximated under $E_\gamma$ as long as $\gamma$ is sufficiently large.

The achievability part of Theorem 5.2.11 is already shown in Corollary 5.2.8. For the converse, we need a notion of conditional typicality specially tailored for our problem.
which differs from the definitions of conditional typicality in [25] or [180] (see also [18]). This can be viewed as an intermediate of the those two definitions.

**Definition 5.2.8 (Moderate Conditional Typicality)** The $\delta$-typical set of $u^n \in \mathcal{U}^n$ with respect to the discrete memoryless channel with per-letter conditional distribution $Q_{X|U}$ is defined as

$$T^n_{[Q_{X|U}]\delta}(u^n) := \{x^n : \forall a, b, |\hat{P}_{u^n,x^n}(a, b) - Q_{X|U}(b|a)\hat{P}_{u^n}(a)| \leq \delta Q_{X|U}(b|a)\}$$  \hspace{1cm} (5.221)

where $\hat{P}_{u^n,x^n}$ denotes the empirical distribution of $(u^n, x^n)$.

**Remark 5.2.8** In addition to its broad interest, Definition 5.2.8 plays an important role in obtaining the uniform bound in Lemma 35, as well as in Lemma 36. This definition of conditional typicality is of broad interest because of Lemma 5.2.12 and Lemma 5.2.13 ahead, and in particular the uniform bound in Lemma 5.2.12. Note that the definition in [25] corresponds to replacing the term $\delta Q_{X|U}(b|a)$ in (5.221) with $\delta$, in which case we cannot bound the probability of a sequence in the typical set as in Lemma 5.2.13. The “robust typicality” definition of [180] (see also [18]) corresponds to replacing this term with $\delta Q_{X|U}(b|a)\hat{P}_{u^n}(a)$, which does not give the uniform lower bound on the probability of conditional typical set as in Lemma 5.2.12.

**Lemma 5.2.12** For fixed $\delta > 0$ and $Q_{X|U}$, there exists a sequence $(\gamma_n)$ such that $\lim_{n \to \infty} \gamma_n = 0$ and

$$Q^n_{X|U}(T^n_{[Q_{X|U}]\delta}(u^n)|u^n) \geq 1 - \gamma_n$$  \hspace{1cm} (5.222)

for all $u^n \in \mathcal{U}^n$.  

189
Proof We show that the statement holds with

\[ \gamma_n = \frac{|\mathcal{U}|r}{n\delta^2} \left( \frac{1}{q} - 1 \right), \]  

(5.223)

where

\[ q := \min_{(a,b) : Q_{X|U}(b|a) \neq 0} Q_{X|U}(b|a). \]  

(5.224)

The number of occurrences \( N(a,b|u^n,X^n) \) of \((a,b) \in \mathcal{U} \times \mathcal{X}\) in \((u^n,X^n)\), where \( X^n \sim \prod_{i=1}^n Q_{X|U=U_i} \), is binomial with mean \( N(a|u^n)Q_{X|U}(b|a) \) and variance \( N(a|u^n)Q_{X|U}(b|a)(1 - Q_{X|U}(b|a)) \). If \( Q_{X|U}(b|a) = 0 \) the condition that defines the set in (5.221) is automatically true. Otherwise, by Chebyshev’s inequality we have for each \((a,b)\),

\[
\mathbb{P}[|N(a,b|u^n,X^n) - N(a|u^n)Q_{X|U}(b|a)| > n\delta Q_{X|U}(b|a)] \\
\leq \frac{N(a|u^n)Q_{X|U}(b|a)(1 - Q_{X|U}(b|a))}{n^2\delta^2 Q_{X|U}^2(b|a)} \\
\leq \frac{1}{n\delta^2} \left( \frac{1}{q} - 1 \right)
\]

(5.225)

(5.226)

and the claim follows by taking the union bound.

\[ \square \]

Lemma 5.2.13 For each \( u^n \), and \( x^n \in T_{[Q_{X|U}]}(u^n) \), we have the bound

\[ Q_{X|U}^{\otimes n}(x^n|u^n) \geq \exp(-n[H(Q_{X|U}|\hat{P}_{u^n}) + \delta|\mathcal{U}| \log |\mathcal{X}|]). \]  

(5.227)

Proof Since

\[ Q_{X|U}^{\otimes n}(x^n|u^n) = \prod_{a \in \mathcal{U}, b \in \mathcal{X}} Q_{X|U}(b|a)^{N(a,b|u^n,x^n)}, \]  

(5.228)
we have

\[
\begin{align*}
\frac{1}{n} \log \frac{1}{Q_{X|U}^{\otimes n}(x^n|u^n)} &= \sum_{a,b} P_{u^n;x^n}(a,b) \log \frac{1}{Q_{X|U}(b|a)} \\
&\leq \sum_{a,b} [Q_{X|U}(b|a)P_{u^n}(a) + \delta Q_{X|U}(b|a)] \log \frac{1}{Q_{X|U}(b|a)} \\
&= H(Q_{X|U}|P_{u^n}) + \delta \sum_a H(Q_{X|U}=a) \\
&\leq H(Q_{X|U}|P_{u^n}) + \delta |U| \log |\mathcal{X}|. 
\end{align*}
\]

Lemma 5.2.14 For any type \( P_X \) and sequence \( u^n \),

\[
|T_{[Q_{X|U}]\delta}(u^n) \cap T_{P_X}| \leq \exp(n[H_{[P_X]\delta} + \delta|U| \log |\mathcal{X}|])
\]

where we have defined

\[
H_{[P_X]\delta} := \max_{Q_U|Q_X-P_X \leq \delta|U|} H(Q_{X|U}|Q_U)
\]

where \( Q_U \rightarrow Q_{X|U} \rightarrow Q_X \), and the maximum in [5.234] is understood as \(-\infty\) if the set \( \{Q_U : |Q_X-P_X| \leq \delta|U|\} \) is empty.

Proof For any \( u^n \), we have the upper bound

\[
|T_{[Q_{X|U}]\delta}(u^n) \cap T_{P_X}| \leq |T_{[Q_{X|U}]\delta}(u^n)| \leq \left( \min_{x^n \in T_{[Q_{X|U}]\delta}(u^n)} Q_{X|U}^{\otimes n}(x^n|u^n) \right)^{-1} \\
\leq \exp(n[H(Q_{X|U}|P_{u^n}) + \delta|U| \log |\mathcal{X}|])
\]
where we used Lemma 5.2.13 in (5.237). Moreover, if \( u^n \) satisfies \( |P_X - Q_{X|U} \circ \hat{P}_{u^n}| > \delta |U| \) where \( Q_{X|U} \circ \hat{P}_{u^n} := \int Q_{X|U=a} d\hat{P}_{u^n}(a) \), then \( T_{[Q_{X|U}]\delta}(u^n) \cap T_{P_X} \) is empty, because

\[
|\hat{P}_{x^n} - Q_{X|U} \circ \hat{P}_{u^n}| = \sum_{a,b} |\hat{P}_{u^n,a^n}(a,b) - Q_{X|U}(b|a)\hat{P}_{u^n}(a)| \quad (5.238)
\]

\[
\leq \sum_{a,b} \delta Q_{X|U}(b|a) \quad (5.239)
\]

\[
= \delta |U| \quad (5.240)
\]

implies that any \( x^n \) in \( T_{[Q_{X|U}]\delta}(u^n) \) does not have the type \( P_X \). Therefore the desired result follows by taking the maximum of (5.237) over type \( Q_U \) satisfying \( |Q_X - P_X| \leq \delta |U| \).

\[\blacksquare\]

**Proof** [Proof of Converse of Theorem 5.2.11] Fix a codebook \((c_1, \ldots, c_M)\) and type \( P_X \). Define

\[
\mathcal{A}_n := \bigcup_{m=1}^M T_{[Q_{X|U}]\delta}(c_m). \quad (5.241)
\]

Then

\[
\pi_X^n(\mathcal{A}_n \cap T_{P_X}) = \sum_{x^n \in \mathcal{A}_n \cap T_{P_X}} \pi_X^n(x^n) \quad (5.242)
\]

\[
= \exp(-n[H(P_X) + D(P_X||\pi_X)]) \cdot |\mathcal{A}_n \cap T_{P_X}| \quad (5.243)
\]

\[
\leq \exp(-n[H(P_X) + D(P_X||\pi_X)]) \cdot \sum_{m=1}^M |T_{[Q_{X|U}]\delta}(c_m) \cap T_{P_X}| \quad (5.244)
\]

\[
\leq \exp(-n[H(P_X) + D(P_X||\pi_X)]) \cdot M \exp(n[H_{[P_X]\delta} + \delta |U| \log |\mathcal{X}|]) \quad (5.245)
\]

\[
= \exp(-n[D(P_X||\pi_X) + H(P_X) - H_{[P_X]\delta} - R - \delta |U| \log |\mathcal{X}|]). \quad (5.246)
\]
where (5.245) is from Lemma 5.2.14. Whence (5.246) and the trivial bound

\[ \pi_X^{\otimes n}(A_n \cap T_{P_X}) \leq \pi_X^{\otimes n}(T_{P_X}) \leq \exp(-n D(P_X \| \pi_X)) \leq \exp(-n[D(P_X \| \pi_X) - \delta|U| \log |X|]) \]  

yield the bound

\[ \pi_X^{\otimes n}(A_n \cap T_{P_X}) \leq \exp(-n[f(\delta, P_X) - \delta|U| \log |X|]), \]  

where we have defined the function

\[ f(\delta, P_X) := D(P_X \| \pi_X) + [H(P_X) - H(P_X) - R]^+ \]  

for \( \delta > 0 \) and \( P_X \ll \pi_X \). Define\(^\text{13} \)

\[ g(\delta) := \min_{P_X} f(\delta, P_X), \]  

Then

\[ \pi_X^{\otimes n}(A_n) = \sum_{P_X} \pi_X^{\otimes n}(A_n \cap T_{P_X}) \leq \sum_{P_X} \exp(-n[g(\delta) - \delta|U| \log |X|]) \leq (n + 1)^{|X|} \exp(-n[g(\delta) - \delta|U| \log |X|]) \]  

where the summation is over all type \( P_X \) absolutely continuous with respect to \( \pi_X \), and (5.254) is from (5.250). Then for any real number \( G < g(\delta) - \delta|U| \log |X| \) we

\(^{13}\)The reason why we can write minimum in (5.252) is explained in Remark 5.2.9.
have

\[
E_{\exp(nG)}(P_{X^n[U]}^\otimes |\pi_{X^n[U]}) \\
\geq P_{X^n[U]}(A_n) - \exp(nG)\pi_{X^n[U]}(A_n) \tag{5.256}
\]

\[
\geq \frac{1}{M} \sum_{m=1}^{M} Q_{X|U}|c_m|c_m) - \exp(nG)\pi_{X^n[U]}(A_n) \tag{5.257}
\]

\[
\geq 1 - \gamma_n - \exp(nG)\pi_{X^n[U]}(A_n) \tag{5.258}
\]

\[
\rightarrow 1, \quad n \rightarrow \infty. \tag{5.259}
\]

where

- [5.257] uses \( T^n_{Q|U} |c_m) \subseteq A_n, \) and we used the notation of the tensor power for the conditional law \( Q^n_{X|U} (\cdot |u^n) := \prod^n_{i=1} Q_{X|U = u_i}. \)

- [5.258] is from Lemma 5.2.12.


Since \( \delta > 0 \) was arbitrary, we thus conclude

\[
G_\epsilon(R, \pi_X) \geq \sup_{\delta > 0} \{ g(\delta) - \delta |U| \log |X| \} \tag{5.260}
\]

\[
\geq \lim \inf_{\delta \rightarrow 0} g(\delta) \tag{5.261}
\]

\[
\geq g(0) \tag{5.262}
\]

where [5.262] is from Lemma 5.2.15. Since \( g(0) \) is the right side of (5.220), the converse bound is established.

**Lemma 5.2.15** The functions \( f \) and \( g \) defined in (5.251) and (5.252) are both lower semicontinuous.

**Remark 5.2.9** We can write minimum instead of infimum in (5.252) and hence (5.220) because of the lower semicontinuity of \( f \).
**Proof** Consider a lower semicontinuous function \( \chi \) where \( \chi(\delta, P_X, Q_U) \) equals

\[
D(P_X\|\pi_X) + [H(P_X) - H(Q_{X|U}|Q_U) - R]^+
\]  
(5.263)

if \(|Q_X - P_X| \leq \delta|U|\) and \(+\infty\) otherwise. Then \( f(\delta, P_X) = \min_{Q_X} \chi(\delta, P_X, Q_U) \) is lower semicontinuous, as it is the pointwise infimum of a lower semicontinuous functions over a compact set (see for example the proof in \[181\, Lemma 9\]). The lower semicontinuity of \( g \) follows for the same reason.

The function \( G(R, \pi_X) \) in \[5.220\] satisfies some nice properties.

**Proposition 5.2.16** Denote \( G(R, \pi_X) \) as \( G(R, \pi_X, Q_{X|U}) \). Let \( \mathcal{X} \) and \( \mathcal{U} \) be arbitrary alphabets (measurable spaces).

1. The function being minimized in \[5.220\], denoted as \( F(Q_U, R, \pi_X, Q_{X|U}) \), is convex in \( Q_U \).

2. Additivity: for any \( R > 0 \), \( \pi_X_i \) and \( Q_{X_i|U_i} \) \((i = 1, 2)\),

\[
G(R_i, \pi_{X_1}, \pi_{X_2}, Q_{X_i|U_i}Q_{X_{i'}|U_{i'}})
= \min_{R_1, R_2: R_1 + R_2 \leq R} \{G(R_1, \pi_{X_1}, Q_{X_1|U_1})
+ G(R_2, \pi_{X_2}, Q_{X_2|U_2})\},
\]  
(5.264)

where we have abbreviated \( \pi_{X_1} \times \pi_{X_2} \) and \( Q_{X_1|U_1} \times Q_{X_2|U_2} \) as \( \pi_{X_1} \pi_{X_2} \) and \( Q_{X_1|U_1}Q_{X_2|U_2} \).

3. \( G(R, \pi_X, Q_{X|U}) \) is continuous in \( R \).

4. \( G(R, \pi_X, Q_{X|U}) \) is convex in \( R \).
Proof 1. The function of interest is the maximum of the two functions $D(Q_x\|\pi_x)$ and

$$D(Q_x\|\pi_x) + I(Q_U, Q_{X|U}) - R = \mathbb{E} \left[ i_{Q_{X|U}\|\pi_x}(X|U) \right] - R \quad (5.265)$$

where $(U, X) \sim Q_{UX}$, $Q_U \rightarrow Q_{X|U} \rightarrow Q_X$, and the conditional relative information

$$i_{Q_{X|U}\|\pi_x}(x|u) := \log \frac{dQ_{X|U=x,u}}{d\pi_x}(x), \quad \forall u, x. \quad (5.266)$$

The former is convex and the latter is linear in $Q_U$.

2. The $\leq$ direction is immediate from the single-letter formula $(5.220)$ and the inequality

$$[a]^+ + [b]^+ \geq [a + b]^+ \quad (5.267)$$

for any $a, b \in \mathbb{R}$. For the $\geq$ direction, suppose $Q_{U_1U_2}$ achieves the minimum in the single-letter formula of $G(R, \pi_{X_1}, \pi_{X_2}, Q_{X_1|U_1}, Q_{X_2|U_2})$. Observe that

$$F(Q_{U^2}, R, \pi_{X_1}, \pi_{X_2}, Q_{X_1|U_1}, Q_{X_2|U_2})$$

$$- F(Q_{U_1}, Q_{U_2}, R, \pi_{X_1}, \pi_{X_2}, Q_{X_1|U_1}, Q_{X_2|U_2})$$

$$= I(X_1; X_2) + I(Q_{U_1U_2}, Q_{X_1|U_1}, Q_{X_2|U_2}) - R]^+$$

$$- \left[ \sum_{i=1}^2 I(Q_{U_i}, Q_{X_i|U_i}) - R \right]^+ \quad (5.268)$$

$$\geq [I(X_1; X_2) + I(Q_{U_1U_2}, Q_{X_1|U_1}, Q_{X_2|U_2}) - R]^+$$

$$- \left[ \sum_{i=1}^2 I(Q_{U_i}, Q_{X_i|U_i}) - R \right]^+ \quad (5.269)$$

$$= 0 \quad (5.270)$$
where (5.268) uses $D(Q_{X_1 X_2} || \pi_{X_1 \pi X_2}) - D(Q_{X_1} || \pi_{X_1}) - D(Q_{X_2} || \pi_{X_2}) = I(X_1; X_2)$, and (5.269) follows from (5.267). Therefore

\[
G(R, \pi_{X_1}, \pi_{X_2}, Q_{X_1|U_1}, Q_{X_2|U_2})
= F(Q_{U_1}, Q_{U_2}, R, \pi_{X_1}, \pi_{X_2}, Q_{X_1|U_1}, Q_{X_2|U_2}).
\]

(5.271)

But clearly there exists $R_1$ and $R_2$ summing to $R$ such that

R.H.S. of (5.271)

\[
= F(Q_{U_1}, R_1, \pi_{X_1}, Q_{X_1|U_1}) + F(Q_{U_2}, R_2, \pi_{X_2}, Q_{X_2|U_2})
\geq F(R_1, \pi_{X_1}, Q_{X_1|U_1}) + F(R_2, \pi_{X_2}, Q_{X_2|U_2})
\]

(5.272)

(5.273)

and the result follows.

3. Fix any two numbers $0 \geq R' < R$. Choose $Q_U$ such that

\[
G(R, \pi_{X}, Q_{X|U}) = F(Q_{U}, R, \pi_{X}, Q_{X|U}).
\]

(5.274)

Then

\[
0 \leq G(R', \pi_{X}, Q_{X|U}) - G(R, \pi_{X}, Q_{X|U})
\leq F(Q_{U}, R', \pi_{X}, Q_{X|U}) - F(Q_{U}, R, \pi_{X}, Q_{X|U})
= [I(Q_{U}, Q_{X|U}) - R']^+ - [I(Q_{U}, Q_{X|U}) - R]^+
\leq [R - R']^+
\]

(5.275)

(5.276)

(5.277)

(5.278)

where (5.275) follows because $G(\cdot, \pi_{X}, Q_{X|U})$ is non-increasing, and (5.278) uses (5.267) again. Thus $G(R, \pi_{X}, Q_{X|U})$ is actually 1-Lipschitz continuous in $R$. 

197
4. Fix $R_1, R_2 \geq 0$, $\alpha \in [0, 1]$, and let $Q_{u_i}$ maximize $F(\cdot, R_i, \pi_X, Q_{X|u})$ for $i = 1, 2$.

Define

$$R_\alpha := (1 - \alpha)R_0 + \alpha R_1;$$

$$Q_{u_\alpha} := (1 - \alpha)Q_{u_0} + \alpha Q_{u_1}.$$  \hfill (5.279)

In both $I(Q_{u_\alpha}, Q_{X|u}) > R_\alpha$ and $I(Q_{u_\alpha}, Q_{X|u}) \leq R_\alpha$ cases one can explicitly calculate that

$$F(Q_{u_\alpha}, R_\alpha, \pi_X, Q_{X|u}) \leq (1 - \alpha)F(Q_{u_0}, R_0, \pi_X, Q_{X|u})$$

$$+ \alpha F(Q_{u_1}, R_1, \pi_X, Q_{X|u})$$  \hfill (5.281)

and the convexity follows.

\[ \square \]

### 5.2.4 Application to Lossy Source Coding

The simplest application of the new resolvability result (in particular, the softer-covering lemma) is to derive a one-shot achievability bound for lossy source coding, which is most fitting in the regime of low rate and exponentially decreasing success probability. The method is applicable to general sources. In the special case of i.i.d. sources, it recovers the “success exponent” in lossy source coding originally derived by the method of types \[25\] for discrete memoryless sources. The achievability bound in \[160\] can be viewed as the $\gamma = 1$ special case, which is capable of recovering the rate-distortion function, but cannot recover the exact rate-distortion-exponent tradeoff.

**Theorem 5.2.17** Consider a source with distribution $\pi_X$ and a distortion function $d(\cdot, \cdot)$ on $U \times X$. For any joint distribution $Q_U Q_{X|U}$, $\gamma \geq 1$, $d > 0$ and integer $M$,
there exists a random transformation $\pi_{U|X}$ (stochastic encoder) whose output takes at most $M$ values, and

$$\mathbb{P}[d(\hat{U}, \hat{X}) \leq d] \geq \frac{1}{\gamma} \left( \mathbb{P}[d(U, X) \leq d] - \mathbb{E}[E_\gamma(Q_{X|U^M}\|\pi_X)] \right)$$  \hfill (5.282)

where $(U, X) \sim \pi_{U|X}\pi_X$, $(U, X) \sim Q_{UX}$, and $U^M \sim Q_U^\otimes M$.

**Proof** Given a codebook $(c_1, \ldots, c_M) \in \mathcal{U}$, let $P_U$ be the equiprobable distribution on $(c_1, \ldots, c_M)$ and set

$$P_{UX} := Q_{X|U}P_U.$$  \hfill (5.283)

The likelihood encoder is then defined as a random transformation

$$\pi_{U|X} := P_{U|X}$$  \hfill (5.284)

so that the joint distribution of the codeword selected and the source realization $X$ is

$$\pi_{UX} = \pi_X P_{U|X}$$  \hfill (5.285)

From Proposition 5.2.2\[2\] and Proposition 5.2.2\[4\] we obtain

$$\gamma \pi_{UX}(d(\cdot, \cdot) \leq d) \geq P_{UX}(d(\cdot, \cdot) \leq d) - E_\gamma(P_{X|U}\|\pi_X)$$  \hfill (5.286)

$$= P_{UX}(d(\cdot, \cdot) \leq d) - E_\gamma(P_X\|\pi_X)$$  \hfill (5.287)

where $(\hat{U}, \hat{X}) \sim P_{UX}$. Note that $P_{UX}$ and $\pi_{U|X}$ depend on the codebook $c^M$. Now consider a random codebook $c^M \leftarrow U^M$. Taking the expectation on both sides of
with respect to $U^M$, we have

$$
\gamma \mathbb{E}[\pi_{UX}(d(\cdot, \cdot) \leq d)] \\
\geq \mathbb{E}[P_{UX}(d(\cdot, \cdot) \leq d)] - \mathbb{E}[\gamma(Q_{X|U^M}\|\pi_X)] \\
= \mathbb{P}[d(U, X) \leq d] - \mathbb{E}[\gamma(Q_{X|U^M}\|\pi_X)]
$$

(5.288)

(5.289)

where in (5.289) we used the fact that $\mathbb{E}[P_{UX}] = Q_{UX}$. Finally we can choose one codebook (corresponding to one $\pi_{U|X}$) such that $\pi_{UX}(d(\cdot, \cdot) \leq d)$ is at least its expectation.

Remark 5.2.10 In the i.i.d. setting, let $R(\pi_X, d)$ be the rate-distortion function when the source has per-letter distribution $\pi_X$. The distortion function for the block is derived from the per-letter distortion by

$$
d^n(u^n, x^n) := \frac{1}{n} \sum_{i=1}^{n} d(u_i, x_i).
$$

(5.290)

Let $(\bar{X}^n, \bar{U}^n)$ be the source-reconstruction pair distributed according to $\pi_{X^n|U^n}$. If $0 \leq R < R(\pi_X, d)$, the maximal probability that the distortion does not exceed $d$ converges to zero with the exponent

$$
\lim_{n \to \infty} \frac{1}{n} \log \frac{1}{\mathbb{P}[d^n(\bar{U}^n, \bar{X}^n) \leq d]} = G(R, d)
$$

(5.291)

where

$$
G(R, d) := \min_Q [D(Q\|P) + [R(Q, d) - R]^+].
$$

(5.292)

A weaker achievability result than (5.292) was proved in [182, p168], whereas the final form (5.292) is given in [25, p158, Ex6] based on method of types. Here we can easily prove the achievability part of (5.292) using Theorem 5.2.17 and Corollary 5.2.8 by
setting $Q_X$ to be the minimizer of \((5.292)\) and $Q_{U|X}$ to be such that

\begin{align*}
\mathbb{E}[d(U, X)] &\leq d, \quad (5.293) \\
I(Q_u, Q_{X|u}) &\leq R. \quad (5.294)
\end{align*}

Then $\gamma_n = \exp(nE)$ with

\begin{align*}
E > D(Q_X||\pi_X) + [I(Q_u, Q_{X|u}) - R]^+, \quad (5.295)
\end{align*}

ensures that

\begin{align*}
\mathbb{P}[d^n(U^n, \bar{X}^n) \leq d] \geq \frac{1}{2} \exp(-nE) \quad (5.296)
\end{align*}

for $n$ large enough, by the law of large numbers.

**Remark 5.2.11** Since the $E_\gamma$ metric reduces to total variation distance when $\gamma = 1$, Theorem 5.2.17 generalizes the likelihood source encoder based on the standard soft-covering/resolvability lemma \([160]\). In \([160]\), the error exponent for the likelihood source encoder at rates above the rate-distortion function is analyzed using the exponential decay of total variation distance in the approximation of output statistics, and the exponent does not match the optimal exponent found in \([23]\). It is also possible to upper-bound the success exponent of the total variation distance-based likelihood encoder at rates below the rate-distortion function by analyzing the exponential convergence to 2 of total variation distance in the approximation of output statistics; however that does not yield the optimal exponent \((5.292)\) either. This application illustrates one of the nice features of the $E_\gamma$-resolvability method: it converts a large deviation analysis into a law of large numbers analysis, that is, we only care about whether $E_\gamma$ converges to 0, but not the speed, even when dealing with error exponent problems.
5.2.5 Application to Mutual Covering Lemma

Another application of the softer-covering lemma is a one-shot generalization of the *mutual covering lemma* in network information theory \[156\]. The asymptotic mutual covering lemma says, fixing a (per-letter) joint distribution \( P_{UV} \), if enough \( U^n \)-sequences and \( V^n \)-sequences are independently generated according to \( P_{U^n} \) and \( P_{V^n} \) respectively, then with high probability we will be able to find one pair jointly typical with respect to \( P_{UV} \). In the one-shot version, the “typical set” is replaced with an arbitrarily high probability (under the given joint distribution) set. The original proof of the asymptotic mutual covering lemma \([156,18]\) used the second-order moment method.

The one-shot mutual covering lemma can be used to prove a one-shot version of Marton’s inner bound for the broadcast channel with a common message\(^{14}\) without time-sharing, improving the proof in \([161]\) based on the basic covering lemma where time-sharing is necessary. More discussions about the background and the derivation of the one-shot Marton’s inner bound can be found in our conference paper \([186]\). For general discussions on single-shot covering lemmas, see \([161,187]\). To avoid using time-sharing in the single-shot setting, \([188]\) pursued a different approach to derive a single-shot Marton’s inner bound. Moreover, a version of one-shot mutual covering lemma can be distilled from their approach \([189]\). We compare their approach and ours at the end of the section.

In our conference paper \([186]\) we applied the one-shot mutual covering lemma to derive a one-shot version of Marton’s inner bound for the broadcast channel with a common message, without using time-sharing/common randomness. In our conference paper \([190]\) we use a concentration inequality of Talagrand to prove that the

\(^{14}\)More precisely, we are referring to the three auxiliary random variables version due to Liang and Kramer \([183, \text{Theorem 5}]\) (see also \([18, \text{Theorem 8.4}]\)), which is equivalent to an inner bound obtained by Gelfand and Pinsker \([184]\) upon optimization (see \([185]\) or \([18, \text{Remark 8.6}]\)).
covering error probability is doubly exponentially vanishing in the blocklength. Since those results are not directly related to \( E_\gamma \), we exclude them from this thesis.

Now, we proceed to provide a simple derivation of a mutual covering using the softer-covering lemma.

**Lemma 5.2.18** Fix \( P_{UV} \) and let

\[
P_{U^M V^L} := \left( \prod_{m} P_{U} \times \cdots \times P_{U} \right) \left( \prod_{l} P_{V} \times \cdots \times P_{V} \right).
\]

Then

\[
\mathbb{P} \left[ \bigcap_{m=1}^{M} \bigcap_{l=1}^{L} \{(U_m, V_l) \notin \mathcal{F} \} \right] \\
\leq \mathbb{P}[(U, V) \notin \mathcal{F}] + \mathbb{P}[i_{U;V}(U;V) \geq \log ML - \tau] \\
+ \frac{\exp(\tau)}{\max\{M, L\}} + e^{-\frac{1}{2}\exp(\tau)}. \tag{5.298}
\]

for all \( \tau > 0 \) and event \( \mathcal{F} \).

**Proof** Assume without loss of generality that \( L \geq M \). For any \( u \in \mathcal{U} \), define

\[
\mathcal{F}_u := \{ v : (u, v) \in \mathcal{F} \}, \tag{5.299}
\]

and for any \( u^M \in \mathcal{U}^M \), define

\[
\mathcal{A}_{u^M} := \bigcup_{m=1}^{M} \mathcal{F}_{u_m}. \tag{5.300}
\]
Now fix a $U$-codebook $c^M$ and observe that

\[
\gamma P_V(A_{c^M}) \geq P_{V[c^M]}(A_{c^M}) - E_\gamma(P_{V[c^M]})P_V
\]

\[
\geq \frac{1}{M} \sum_{m=1}^{M} P_{V|U=c_m}(F_{c_m}) - E_\gamma(P_{V[c^M]})P_V
\]

where we recall that

\[
P_{V[c^M]} := \frac{1}{M} \sum_{m=1}^{M} P_{V|U=c_m}.
\]

(5.301) is from the definition of $E_\gamma$ and (5.302) is because $F_{c_m} \subseteq \bigcup_{m=1}^{M} F_{c_m} = A_{c^M}$.

Denote by $\Gamma(c_1, \ldots, c_M)$ the right side of (5.302), which is trivially upper-bounded by 1. Next, we show that

\[
P\left[ \bigcap_{m=1, l=1}^{M, L} \{(U_m, V_l) \notin F\} \left| U^M = c^M \right. \right] \leq 1 - \Gamma(c^M) + e^{-\frac{L}{\gamma}}
\]

(5.304) which is trivial when $\Gamma(c^M) < 0$. In the case of $\Gamma(c^M) \in [0, 1]$,

\[
P\left[ \bigcap_{m=1, l=1}^{M, L} \{(U_m, V_l) \notin F\} \left| U^M = c^M \right. \right] = [1 - P_V(A_{c^M})]^L
\]

\[
\leq \left[ 1 - \frac{\Gamma(c^M) \frac{L}{\gamma}}{L} \right]^L
\]

\[
\leq 1 - \Gamma(c^M) + e^{-\frac{L}{\gamma}}
\]

(5.307) where (5.305) is from the definition of $A_{c^M}$, and (5.306) is from (5.302). The last step (5.307) uses the basic inequality

\[
\left(1 - \frac{p\alpha}{M}\right)^M \leq 1 - p + e^{-\alpha}
\]

(5.308)
for $M, \alpha > 0$ and $0 \leq p \leq 1$, which has been useful in the proofs of the basic covering lemma (see [3][161][13]). Integrating both sides of (5.304) over $e^M$ with respect to $P_U \times \cdots \times P_U$,

$$
\mathbb{P} \left[ \bigcap_{m=1,1=1}^{M,L} \{ (U_m, V_i) \notin \mathcal{F} \} \right] 
\leq P_{UV}(\mathcal{F}^c) + \mathbb{E} \left[ E_\gamma(P_{V|U^M}\|P_V) \right] + e^{-\frac{L}{\gamma}},
$$

(5.309)

where we have used the fact that $\mathbb{E}[P_{V|U}(\mathcal{F}_{U_m}|U_m)] = P_{UV}(\mathcal{F})$ for each $m$. Applying the “softer-covering lemma” as in Remark 5.2.5, the middle term on the right hand side of (5.309) is upper-bounded by

$$
\mathbb{P} \left[ \nu_{V;U}(V;U) \geq \log \frac{M\gamma}{2} \right] + \frac{2}{\gamma}
$$

(5.310)

and the result follows by $\gamma \leftarrow 2L \exp(-\tau)$.

**Remark 5.2.12** From the above derivation we see that for the proof of the basic covering lemma ($M = 1$ case) we will need the “softest-covering lemma” (the case of one codeword) rather than the soft-covering lemma (case of $\gamma = 1$ and $L > 1$ codewords). However, it is still possible to prove the basic covering lemma using the soft-covering lemma using a different argument; see the discussion in [189], which is essentially based on the idea in [159].

Lemma 5.2.19 below is a strengthened version of the one-shot-mutual covering lemma, which improves Lemma 5.2.18 in terms of the error exponent. The proof of Lemma 5.2.19 essentially combines the proof the achievability part of resolvability and the proof Lemma 5.2.18 and the improvement results from not treating the two steps separately. The proof is not as conceptually simple as Lemma 5.2.18 since the complexities are no longer buried under the softer-covering lemma.
Lemma 5.2.19 Under the same assumptions as Lemma 5.2.18,

\[
\mathbb{P} \left[ \bigcap_{m=1}^{M} \bigcap_{l=1}^{L} \{(U_m, V_l) \notin \mathcal{F}\} \right] \\
\leq \mathbb{P} \left[ (U, V) \notin \mathcal{F} \text{ or } \exp(\nu_{U;V}(U;V)) > ML \exp(-\gamma) - \delta \right] \\
+ \frac{\min\{M, L\} - 1}{\delta} + e^{-\exp(\gamma)}. \tag{5.311}
\]

for all \(\delta, \gamma > 0\) and event \(\mathcal{F}\).

Proof See Lemma 1 and Remark 4 in the conference version [191]. \hfill \blacksquare

Remark 5.2.13 An advantage of Lemma 5.2.19 over Lemma 5.2.18 is that the upper-bound in the former contains a probability of a union of two events, rather than the sum of the probability of the two events. This yields a strict improvement in the second order rate analysis. Moreover, by setting \(\delta \downarrow 0\) and \(M = 1\) we exactly recover the basic one-shot covering lemma in [161].

In terms of the second order rates, the one-shot Marton’s inner bound for broadcast obtained from our one-shot mutual covering lemma ([186, Theorem 10]) is equivalent to the achievability bound claimed in [37, Theorem 4] based on the stochastic likelihood encoder. However, although it is not demonstrated explicitly in [186, Theorem 10], we can improve the analysis of [186, Theorem 10] by using various nuisance parameters rather than a single \(\gamma\), to obtain a one-shot Marton’s bound which gives strictly better error exponents than [37, Theorem 4]. The reason for such improvement is that the third term in (5.311) is doubly exponential and the second term converges to zero with a large exponent. On the other hand, the approach of [37, Theorem 4] has the advantage of being easily extendable to the case of more than two users (which would correspond to a multivariate mutual covering lemma).
5.2.6 Application to Wiretap Channels

Our final application of the $E_\gamma$-resolvability (in particular, the softer-covering lemma) is in the wiretap channel, whose setup is as depicted in Figure 5.4. The receiver and the eavesdropper observe $y \in Y$ and $z \in Z$, respectively. Given a codebook $c^{ML}$, the input to $P_{YZ|X}$ is $c_{wl}$ where $w \in \{1, \ldots, M\}$ is the message to be sent, and $l$ is equiprobably chosen from $\{1, \ldots, L\}$ to randomize the eavesdropper’s observation. We call such a $c^{ML}$ an $(M, L)$-code. Moreover, the eavesdropper’s observation has the distribution $\pi_Z$ when no message is sent. In this setup, we don’t need to assume a prior distribution on the message/non-message. We wish to design the codebook such that the receiver can decode the message (reliability) whereas the eavesdropper cannot detect whether a message is sent nor guess which message is sent (security).

For general wiretap channels the performance may be enhanced by appending a conditioning channel $Q_{X|U}$ at the input of the original channel [171]. In that case the same analysis can be carried out for the new wiretap channel $Q_{YZ|U}$. Thus the model in Figure 5.4 entails no loss of generality.

![Figure 5.4: The wiretap channel](Image)

In Wyner’s setup (see for example [13]), secrecy is measured in terms of the conditional entropy of the message given the eavesdropper observation. In contrast, we measure secrecy in terms of the size of the list that the eavesdropper has to declare for the message to be included with high probability. Practically, the message $W$ is the compressed version of the plaintext. Assuming that the attacker knows which compression algorithm is used, the plaintext can be recovered by running each of the
items in the eavesdropper output list through the decompressor and selecting the one that is intelligible.

While there have been previous proofs for wiretap channels using the conventional resolvability (soft-covering lemma) [171][172], the conventional resolvability based on the total variation metric is only suitable when the communication rate is low enough to achieve perfect secrecy. In contrast, $E_{\gamma}$-resolvability yields lower bounds on the minimum size of the eavesdropper list for an arbitrary rate of communication. This interpretation of security in terms of list size is reminiscent of equivocation [192], and indeed we recover the same formula in the asymptotic setting, even though there is no direct correspondence between both criteria. Moreover, we also consider a more general case where the eavesdropper wishes to reliably detect whether a message is sent, while being able to produce a list including the actual message if it decides it is present. This is a practical setup because “no message” may be valuable information which the eavesdropper wants to ascertain reliably. The idea is reminiscent of the stealth communication problem (see [193][194] and the references therein) which also involves a hypothesis test on whether a message is sent. However, the setup and the analysis (including the covering lemma) are quite different from [193] and [194]. In comparison, our results are more suitable for the regime with higher communication rates and lower secrecy demands. In the discrete memoryless case, we obtain single-letter expressions of the tradeoff between the transmission rate, eavesdropper list, and the exponent of the false-alarm probability for the eavesdropper (i.e. declaring the presence of a message when there is none).

**Summary of The Asymptotic Results**

We need the following definitions to quantify the eavesdropper ability to detect/decode messages.
Definition 5.2.9  For a fixed codebook and channel $P_{Z|X}$, we say the eavesdropper can perform $(A,T,\epsilon)$-decoding if upon observing $Z$, it outputs a list which is either empty or of size $T$, such that

- $\Pr[\text{list} \neq \emptyset | \text{no message}] \leq A^{-1}$.

- There exists $\epsilon_m \in [0, 1]$, $m = 1, \ldots, M$ satisfying $\epsilon = \frac{1}{M} \sum_{m=1}^{M} \epsilon_m$ such that

$$\Pr[m \notin \text{list}|W = m] \leq \epsilon_m. \tag{5.312}$$

Although the decoder in Definition 5.2.9 is reminiscent of erasure and list decoding [21], for the former it is possible that actually no message is sent, and we treat the undetected and detected errors together.

The logarithm of $T$ can be intuitively understood as the equivocation $H(W|Z)$ [192]. However, $\log T$ can be much smaller than $H(W|Z)$: a distribution can have 99% of its mass supported on a very small set, and yet have an arbitrarily large entropy.

The quantity $A > 0$ characterizes how well the eavesdropper can detect that no message is sent, which is related the notion of stealth communication [193] [194]. The “non-stealth” is measured by $D(P_Z \parallel \pi_Z)$ in [193], and is measured by $|P_Z - \pi_Z|$ in [194]. Although both the relative entropy and total variation are related to error probability hypothesis testing, their results cannot be directly compared with ours, since they are interested in the regime where non-stealth vanishes while the transmission rate is below the secrecy capacity. In contrast, we are mainly interested in the regime where $A$ grows exponentially (so that the “non-stealth” in their definition grows) in the blocklength, but the transmission rate is above the secrecy capacity.

The asymptotic version of the eavesdropper achievability is as follows.

Definition 5.2.10  Fix a sequence of codebooks and a eavesdropper channel $(P_{Z^n|X^n})_{n=1}^{\infty}$. The rate pair $(\alpha, \tau)$ is $\epsilon$-achievable by the eavesdropper if there
exist sequences \((A_n)\) and \((T_n)\) with

\[
\begin{align*}
\lim \inf_{n \to \infty} \frac{1}{n} \log A_n & \geq \alpha \\
\lim \sup_{n \to \infty} \frac{1}{n} \log T_n & \leq \tau
\end{align*}
\]

(5.313) (5.314)

such that for sufficiently large \(n\), the eavesdropper can achieve \((A_n, T_n, \epsilon)\)-decoding.

By the diagonalization argument [44, P56], the set of \(\epsilon\)-achievable \((\alpha, \tau)\) is closed.

An \((M, L, Q_X)\)-random code is defined as the ensemble of the codebook \(c^{ML}\) where each codeword \(c_{wl}\) is i.i.d. chosen according to \(Q_X, w \in \{1, \ldots, M\}, m \in \{1, \ldots, M\}\). We shall focus on random codes, for which reliability is guaranteed by channel coding theorems, so we only need to consider the security condition.

First, we extend the notions of achievability to the case of a random ensemble of codes by taking the average: we say for a random ensemble of codes the eavesdropper can perform \((A, T, \epsilon)\)-decoding, if there exists \(\epsilon(c^{ML})\) such that for each \(c^{ML}\) the eavesdropper can perform \((A, T, \epsilon(c^{ML}))\)-decoding (in the sense of Definition 5.2.9), and the average of \(\epsilon(c^{ML})\) with respect to the codebook distribution is upper-bounded by \(\epsilon\). Similarly, Definition 5.2.10 can be extended to random codes. Then, the following is our main result which characterizes the set of eavesdropper achievable pairs for stationary memoryless channels.

**Theorem 5.2.20** Fix any \(Q_X, R, R_L\) and \(0 < \epsilon < 1\). Consider \((\exp(nR), \exp(nR_L), Q_X^{\otimes n})\)-random codes and stationary memoryless channel with per-letter conditional distribution \(Q_{Z|X}\). Then the pair \((\alpha, \tau)\) is \(\epsilon\)-achievable by the eavesdropper if and only if

\[
\begin{align*}
\alpha & \leq D(Q_Z \| \pi_z) + [I(Q_X, P_{Z|X}) - R - R_L]^+; \\
\tau & \geq R - [I(Q_X, P_{Z|X}) - R_L]^+
\end{align*}
\]

(5.315)

where \(Q_X \to Q_{Z|X} \to Q_Z\).
From the noisy channel coding theorem, the supremum randomization rate $R_L$ such that the sender can reliably transmit messages at the rate $R$ is $I(Q_X, P_{Y|X}) - R$. The larger $R_L$ the less reliably the eavesdropper can decode, so the optimal encoder chooses $R_L$ as close to this supremum as possible. Thus Theorem 5.2.20 implies the following result:

**Theorem 5.2.21** Given a stationary memoryless wiretap channel with per-letter conditional distribution $P_{YZ|X}$, there exists a sequence of codebooks such that messages at the rate $R$ can be reliably transmitted to the intended receiver and that $(\alpha, \tau)$ is not $\epsilon$-achievable, for any $\epsilon \in (0, 1)$, by the eavesdropper if there exists some $Q_X$ such that

$$R < I(Q_X, P_{Y|X})$$  \hspace{1cm} (5.316)

and either

$$\alpha > D(Q_Z\|\pi_Z) + [I(Q_X, P_{Z|X}) - I(Q_X, P_{Y|X})]^+$$  \hspace{1cm} (5.317)

or

$$\tau < R - [I(Q_X, P_{Z|X}) - I(Q_X, P_{Y|X}) + R]^+.$$  \hspace{1cm} (5.318)

**Remark 5.2.14** In general the sender-receiver want to minimize $\alpha$ and maximize $\tau$ obeying the tradeoff (5.317), (5.318) by selecting $Q_X$. In the special case where $\alpha$ has no importance and $R$ is larger than the secrecy capacity $C := \sup_{Q_X} \{I(Q_X, P_{Y|X}) - I(Q_X, P_{Z|X})\}$, we see from (5.318) that the supremum $\tau$ is $C$. The formula for the supremum of $\tau$ is the same as the equivocation measure defined as $\frac{1}{n}H(W|Z^n)$ \[192\], but technically our result does not follow directly from the lower bound on equivocation, since it may be possible that the a posterior distribution of $W$ is concentrated on
a small list but has a tail spread over an exponentially large set, resulting a large equivocation.

Next, we need to prove the “only if” and “if” parts of Theorem 5.2.20

Converse for the Eavesdropper: One-Shot Bounds

The “only if” part (the eavesdropper converse) of Theorem 5.2.20 follows by applying the following non-asymptotic bounds to different regions and invoking Corollary 5.2.8.

Theorem 5.2.22 In the wiretap channel, fix an arbitrary distribution $\mu_Z$. Suppose the eavesdropper can either detect that no message is sent upon observing $z \in D_0$ with

$$\mu_Z(D_0) \geq 1 - A^{-1}$$

for some $A \in [1, \infty)$, or outputs a list of $T(z)$ messages upon observing $z \notin D_0$ that contains the actual message $m \in \{1, \ldots, M\}$ with probability at least $1 - \epsilon_m$, for some $\epsilon_m \in [0, 1]$. Define the average quantities

$$T := \frac{1}{\mu_Z(D_0)} \int_{D_0^c} T(z) d\mu_Z(z),$$

$$\epsilon := \frac{1}{M} \sum_{m=1}^{M} \epsilon_m.$$ 

Then for any $\gamma \in [1, +\infty)$,

$$\frac{1}{A} \geq \frac{1}{\gamma} (1 - \epsilon - E_\gamma(P_Z \parallel \pi_Z)),$$

where we recall that $\pi_Z$ is the non-message distribution, $P_Z := \frac{1}{M} \sum_{m=1}^{M} P_{Z|W=m}$, and $P_{Z|W=m}$ is the distribution of the eavesdropper observation for the message $m$.
(assuming an arbitrary codebook is used). Moreover,

$$\frac{T}{MA} \geq \frac{1}{\gamma} \left( 1 - \epsilon - \frac{1}{M} \sum_{m=1}^{M} E_{\gamma}(P_{Z|W=m}\|\mu_Z) \right).$$

(5.323)

We will choose $\mu_Z = P_Z$ when we use Theorem 5.2.22 to prove Theorem 5.2.20 although (5.323) holds for any $\mu_Z$.

From the eavesdropper viewpoint, a larger $A$ and a smaller $T$ is more desirable since it will then be able to find out that no message is sent with smaller error probability or narrow down to a smaller list when a message is sent. This observation agrees with (5.322) and (5.323): a smaller $\gamma$ implies a higher degree of approximation, and hence higher indistinguishability of output distributions which is to the eavesdropper disadvantage.

**Proof** To see (5.322),

$$\frac{1}{A} \geq \pi_Z(D_0^c)$$

(5.324)

$$\geq \frac{1}{\gamma} (P_Z(D_0^c) - E_{\gamma}(P_Z\|\pi_Z))$$

(5.325)

$$= \frac{1}{\gamma} \left( \frac{1}{M} \sum_{m=1}^{M} P_{Z|W=m}(D_0^c) - E_{\gamma}(P_Z\|\pi_Z) \right)$$

(5.326)

$$\geq \frac{1}{\gamma} (1 - \epsilon - E_{\gamma}(P_Z\|\pi_Z)).$$

(5.327)
To see (5.323), let $\mathcal{D}_m$ be the set of outputs $z \in \mathcal{Z}$ for which the eavesdropper list contains $m \in \{1, \ldots, M\}$. Then

\[
\frac{T}{MA} \geq \frac{T}{M} \mu_Z(\mathcal{D}_0^c) = \frac{1}{M} \int_{\mathcal{D}_0^c} T(z) \, d\mu_Z(z) \geq \frac{1}{M} \sum_{m=1}^M 1\{z \in \mathcal{D}_m\} \, d\mu_Z(z) = \frac{1}{M} \sum_{m=1}^M \mu_Z(\mathcal{D}_m) \geq \frac{1}{M} \sum_{m=1}^M \left( P_{Z|W=m}(\mathcal{D}_m) - E_\gamma(P_{Z|W=m}\|\mu_Z) \right) \geq \frac{1}{\gamma} \left( 1 - \epsilon - \frac{1}{M} \sum_{m=1}^M E_\gamma(P_{Z|W=m}\|\mu_Z) \right).
\]

Next, we particularize Theorem 5.2.22 to the asymptotic setting.

**Proof** [Proof of “only if” in Theorem 5.2.20]

- Fix an arbitrary

\[
\alpha > D(Q_Z\|\pi_Z) + [I(Q_X, P_{Z|X}) - R - R_L]^+.
\]  

We will show that $(\alpha, \tau)$ is not $\epsilon$-achievable by the eavesdropper for any $\tau > 0$ and $\epsilon \in (0, 1)$. Pick $\sigma > 0$ such that

\[
\alpha > D(Q_Z\|\pi_Z) + [I(Q_X, P_{Z|X}) - R - R_L]^+ + 2\sigma
\]  

214
and define

\[
\begin{aligned}
A_n &= \exp(n(\alpha - \sigma)) \\
T_n &= \exp(n(\tau + \sigma)).
\end{aligned}
\] (5.337)

Assuming the eavesdropper can perform \((A_n, T_n, \epsilon(c^{ML}))\)-decoding for a particular realization of the codebook \(c^{ML}\), then applying Theorem 5.2.22 with

\[\gamma_n = \exp(n(D(Q_Z \| \pi_Z) + [I(Q_X, P_{Z|X}) - R - R_L]^+ + \sigma)),\] (5.338)

we obtain

\[
\exp(n(D(Q_Z \| \pi_Z) + [I(Q_X, P_{Z|X}) - R - R_L]^+ - \alpha + 2\sigma)) = \frac{\gamma_n}{A_n} \geq 1 - \epsilon(c^{ML}) - E_{\gamma_n}(P_{Z^n[c^{ML}]\|\pi_Z^{\otimes n}}). \] (5.339)

From (5.336), the above implies

\[E_{\gamma_n}(P_{Z^n[c^{ML}]\|\pi_Z^{\otimes n}}) \geq \frac{1 - \epsilon(c^{ML})}{2}.\] (5.340)

For sufficiently large \(n\). By Corollary 5.2.8 and (5.338), the average of the left side converges to zero as \(n \to \infty\), thus the average of the right side cannot be lower bounded by \(\frac{1 - \epsilon}{2}\).

- Fix an arbitrary

\[\tau < R - [I(Q_X, P_{Z|X}) - R_L]^+.\] (5.342)
We will show that $(\alpha, \tau)$ is not $\epsilon$-achievable by the eavesdropper for any $\alpha > 0$ and $\epsilon \in (0, 1)$. Pick $\sigma > 0$ such that

$$\tau + 2\sigma < R - [I(Q_X, P_{Z|X}) - R_L]^+ \quad (5.343)$$

and again define $A_n$ and $T_n$ as in (5.337). Assuming the eavesdropper can perform $(A_n, T_n, \epsilon(c^{ML}))$-decoding for a particular realization of the codebook $c^{ML}$, then applying Theorem 5.2.22 with

$$\mu_Z = Q_Z^\otimes n, \quad (5.344)$$

$$\gamma_n = \exp(n([I(Q_X, P_{Z|X}) - R_L]^+ + \sigma)), \quad (5.345)$$

and noting that $A_n \geq 1$, we obtain

$$\exp(n(\tau - R + [I(Q_X, P_{Z|X}) - R_L]^+ + 2\sigma))$$

$$= \frac{T_n \gamma_n}{M_n} \quad (5.346)$$

$$\geq 1 - \epsilon(c^{ML}) - \frac{1}{M} \sum_{m=1}^{M} E_{\gamma_n}(P_{Z^n|W=m}||Q_Z^\otimes n) \quad (5.347)$$

$$= 1 - \epsilon(c^{ML}) - \frac{1}{M} \sum_{m=1}^{M} E_{\gamma_n}(P_{Z^n[c_m L]}||Q_Z^\otimes n) \quad (5.348)$$

where $c_m^L := (c_{m i})_{i=1}^L$. From (5.343), the above implies

$$\frac{1}{M} \sum_{m=1}^{M} E_{\gamma_n}(P_{Z^n[c_m L]}||Q_Z^\otimes n) \geq \frac{1 - \epsilon(c^{ML})}{2}, \quad (5.349)$$

for sufficiently large $n$. Invoking Corollary 5.2.8, we see the average of the right side with respect to the codebook converges zero as $n \to \infty$, and in particular cannot be lower-bounded by $\frac{1-\epsilon}{2}$.  

\[\blacksquare\]
Achievability of the Eavesdropper: Ensemble Tightness

The (eavesdropper) achievability part of Theorem 5.2.20 follows by analyzing the eavesdropper list decoding ability for different cases of the rates \((R, R_L)\). First, consider the following one-shot achievability bounds for channel coding with possibly no message sent:

**Theorem 5.2.23** Consider a random transformation \(P_{Z|X}\) and a \((M, L, Q_X)\)-random code. Let \(Q_X \rightarrow P_{Y|X} \rightarrow Q_Y\), and let \(\pi_Z\) be the distribution of the eavesdropper observation when no message is sent. Define

\[
\bar{\bar{1}}_{Z;X}(z; x) := \log \frac{dP_{Z|X=x}(z)}{d\pi_Z}; \quad \bar{\bar{1}}_{Z;X}(z; x) := \log \frac{dP_{Z|X=x}(z)}{dQ_Z}.
\]

Let \(\delta, \beta, A, T > 0\). Then, there exist three list decoders such that for Decoder 1,

\[
\mathbb{E}_C \mathbb{P}[\text{error|no message}] \leq \frac{1}{A} \exp(-\delta), \quad \text{(5.352)}
\]

\[
\mathbb{E}_C \mathbb{P}[\text{error|message is } m] \leq \mathbb{P}[\bar{\bar{1}}_{Z;X}(Z; X) \leq \log(LMA) + \delta] + \mathbb{P}[\bar{\bar{1}}_{Z;X}(Z; X) \leq \log \frac{LM}{T} + \delta] + \frac{1}{1+\beta} + e^{-(\beta+1)} + \beta \exp(-\delta). \quad \text{(5.353)}
\]

*Here an error in the case of no message means that a non-empty list is produced. An error in the case of message \(m\) means either the list does not contain \(m\), or the list*
For Decoder 2,

\[ \mathbb{E}_c \mathbb{P}[\text{error}|\text{no message}] \leq \frac{1}{A}; \]
\[ \mathbb{E}_c \mathbb{P}[\text{error}|\text{message is m}] \leq \mathbb{P} \left[ i_{Z;X}(Z;X) \leq \log \frac{LM}{T} + \delta \right] + \mathbb{P}[i_{QZ|\pi Z}(Z) \leq \log A] + \frac{1}{1+\beta} + e^{-(\beta+1)} + \beta \exp(-\delta), \]  

where the error events are defined similarly to Decoder 1. Decoder 3 either output an empty list or a list of all messages, and

\[ \mathbb{E}_c \mathbb{P}[\text{error}|\text{no message}] \leq \frac{1}{A}; \]
\[ \mathbb{E}_c \mathbb{P}[\text{error}|\text{message is m}] \leq \mathbb{P}[i_{QZ|\pi Z}(Z) \leq \log A]. \]

**Proof** We defer the proof to the end of this Section.

Under various conditions, one out of the three decoders are asymptotically optimal. By choosing appropriate parameters \( \delta, \beta, A, T > 0 \), it is clear that Theorem 5.2.23 implies the following:

**Corollary 5.2.24** Fix any \( Q_X, R, R_L \) and \( 0 < \epsilon < 1 \). Consider \( (\exp(nR), \exp(nR_L), Q^{\epsilon n}_X) \)-random codes and stationary memoryless channel with per-letter conditional distribution \( Q_{Z|X} \).

\(^{15}\)Such a decoder is a variable list-size decoder. However, we can add a post processor which declares no message if the list is empty, or outputs a list of fixed size \( T \) otherwise (by arbitrarily deleting or adding messages to the list), resulting a new decoder as considered in Definition 5.2.9 and the two types of error probability for the new decoder (i.e. the best values of \( \frac{1}{A} \) and \( \epsilon_m \) in Definition 5.2.9) do not exceed the two types of error probability for the original variable list-size decoder.
• When $R + R_L < I(Q_X, P_{Z|X})$, the rate pair $(\alpha, \tau)$ is $\epsilon$-achievable by a Decoder 1 if

$$\begin{cases}
    D(Q_Z \| \pi_Z) + I(Q_X, P_{Z|X}) > R_L + R + \alpha; \\
    I(Q_X, P_{Z|X}) > R + R_L - \tau.
\end{cases}$$ (5.358)

• When $R + R_L \geq I(Q_X, P_{Z|X})$ but $R_L < I(Q_X, P_{Z|X})$, the rate pair $(\alpha, \tau)$ is $\epsilon$-achievable by a Decoder 2 if

$$\begin{cases}
    I(Q_X, P_{Z|X}) > R_L + R - \tau; \\
    \alpha < D(Q_Z \| \pi_Z).
\end{cases}$$ (5.359)

• When $R_L \geq I(Q_X, P_{Z|X})$, the rate pair $(\alpha, \tau)$ is $\epsilon$-achievable by a Decoder 3 if

$$\begin{cases}
    \tau \geq R; \\
    \alpha < D(Q_Z \| \pi_Z).
\end{cases}$$ (5.360)

**Proof** [Proof Sketch] Consider the first case. To see the achievability of $(\alpha, \tau)$ satisfying (5.358), choose

$$\delta_n := n^{0.9},$$ (5.361)

$$\beta_n := n,$$ (5.362)

$$A_n := \exp(n\alpha),$$ (5.363)

$$T_n := \exp(n\tau).$$ (5.364)

$^{16}$By which we mean it is possible to choose the $\delta, \beta, A, T$ parameters for the decoder to achieve the desired performance.
Then the right sides of (5.352) and (5.353) converges to zero as $n \to \infty$. The analyses of the other two cases are similar using the same choice of the parameters as above.

The eavesdropper’s achievability (“if ” part) of Theorem 5.2.20 then follows from Corollary 5.2.24 and an application of the standard diagonalization argument to show that the achievable region is closed (see [44]).

**Proof of Theorem 5.2.23**

- Codebook generation: $(c_{ij})_{1 \leq i \leq M, 1 \leq j \leq L}$ according to $Q^{\otimes ML}_{X}$.

- Decoders: Fix an arbitrary constant $\delta > 0$. Upon observing $z$, Decoder 1 outputs as a list all $1 \leq i \leq M$ such that there exists $1 \leq j \leq L$ satisfying

\[
\begin{align*}
\bar{t}_{Z,X}(z; c_{ij}) &> \log(AMA) + \delta \\
\underline{t}_{Z,X}(z; c_{ij}) &> \frac{LM}{T} + \delta
\end{align*}
\] (5.365)

if there is at least one such an $i$, or declares that no message is sent (i.e. outputs an empty list) if otherwise. Decoder 2 outputs as a list all $1 \leq i \leq M$ such that there exists $1 \leq j \leq L$ satisfying

\[
\underline{t}_{Z,X}(z; c_{ij}) > \log \frac{LM}{T} + \delta
\] (5.366)

if there exists at least one such $i$ and in addition,

\[
\underline{t}_{Q_{Z}||\pi_{Z}}(z) > \log A,
\] (5.367)

or declares that no message is sent if otherwise. Decoder 3 outputs $\{1, \ldots, M\}$ as the list if (5.367) holds (so that the list size equals $M$), or otherwise declares no message.
- Error analysis: we denote by $\mathcal{L}$ the list of messages recovered by the eavesdropper.

Decoder 1:

$$
\mathbb{P}[\mathcal{L} \neq \emptyset | \text{no message}] \\
\leq \mathbb{P} \left[ \max_{1 \leq m \leq M, 1 \leq l \leq L} i_{Z;X}(\bar{Z}; X_m) > \log(LMA) + \delta \right] \\
\leq \frac{1}{A} \exp(-\delta) \\
$$  \hspace{1cm} (5.368)

where the probability is averaged over the codebook, $(X^{ML}, Z) \sim Q_X^{\otimes ML} \times \pi_Z$, and (5.369) used the packing lemma [161]. Moreover

$$
\mathbb{P}[1 \notin \mathcal{L} \text{ or } \mathcal{L} = \emptyset | W = 1] \\
\leq \mathbb{P}[i_{Z;X}(Z; X) \leq \log(LMA) + \delta] \\
+ \mathbb{P} \left[ i_{Z;X}(Z; X) \leq \log \frac{LM}{T} + \delta \right] \\
$$  \hspace{1cm} (5.371)
where \( W \) denotes the message sent, and \((X, Z) \sim Q_{XZ}\). Further,

\[
\mathbb{P}[|L| \geq T + 1|W = 1] \\
\leq \mathbb{P} \left[ |L| \geq T + 1, L \cap \left\{2, \ldots, \frac{\beta M}{T} + 1\right\} = \emptyset \middle| W = 1 \right] \\
+ \mathbb{P} \left[ L \cap \left\{2, \ldots, \frac{\beta M}{T} + 1\right\} \neq \emptyset \middle| W = 1 \right] \\
\leq \left(1 - \frac{\beta M/T}{M}\right)^T \\
+ \mathbb{P} \left[ L \cap \left\{2, \ldots, \frac{\beta M}{T} + 1\right\} \neq \emptyset \middle| W = 1 \right] \\
\leq \frac{1}{1 + \beta} + e^{-(\beta+1)} \\
+ \mathbb{P} \left[ \max_{2 \leq m \leq \frac{\beta M}{T} + 1, 1 \leq t \leq L} \tau_{Z,X}(\hat{Z}; X_{ml}) > \log \frac{LM}{T} + \delta \right] \\
\leq \frac{1}{1 + \beta} + e^{-(\beta+1)} + \beta \exp(-\delta)
\tag{5.374}
\]

where

- To see (5.373), note that by the symmetry among the messages \(2, \ldots, M\), for any \( t \geq T \),

\[
\mathbb{P} \left[ L \cap \left\{2, \ldots, \frac{\beta M}{T} + 1\right\} = \emptyset \middle| W = 1, |L\backslash\{1\}| = t \right] \\
= \left(1 - \frac{\beta M/T}{M-1}\right)^t \\
\leq \left(1 - \frac{\beta M/T}{M}\right)^T 
\tag{5.376}
\]

- In (5.374) \((X^{ML}, \hat{Z}) \sim Q_X^{ML} \times Q_{Z,ML} \times Q_{Z}\), and we used the inequality (5.308).

- (5.375) used the packing lemma [161].
In summary,

\[ \mathbb{P}[\text{error}|\text{no message}] \leq \frac{1}{A} \exp(-\delta), \quad (5.379) \]

and for each \( m = 1, \ldots, M \), by the union bound and by the symmetry in codebook generation we have

\[
\begin{align*}
\mathbb{P}[\text{error}|W = m] & \leq \mathbb{P}[i_{Z;X}(Z;X) \leq \log(LMA) + \delta] \\
& \quad + \mathbb{P}
\left[
\begin{array}{c}
\log \left( \frac{LM}{T} \right) + \delta \\
\end{array}
\right] \\
& \quad + \frac{1}{1 + \beta} + e^{-(\beta+1)} + \beta \exp(-\delta). 
\end{align*}
\]

(5.380)

Decoder 2:

\[
\begin{align*}
\mathbb{P}[\mathcal{L} \neq \emptyset|\text{no message}] & \leq \mathbb{P}[i_{QZ|\pi_Z}(\tilde{Z}) > \log A] \\
& \leq \frac{1}{A} \mathbb{P}[i_{QZ|\pi_Z}(Z) > \log A] \quad (5.381) \\
& \leq \frac{1}{A} 
\end{align*}
\]

(5.382)

(5.383)

where \( \tilde{Z} \sim \pi_Z \) and \( Z \sim Q_Z \), and (5.382) used the change of measure. On the other hand,

\[
\begin{align*}
\mathbb{P}[\mathcal{L} = \emptyset, 1 \notin \mathcal{L}|W = 1] & \leq \mathbb{P}[i_{Z;X}(Z;X) \leq \log \left( \frac{LM}{T} \right) + \delta] \\
& \leq \mathbb{P}[i_{QZ|\pi_Z}(Z) \leq \log A], 
\end{align*}
\]

(5.384)

and

\[
\mathbb{P}[\mathcal{L} = \emptyset|W = 1] \leq \mathbb{P}[i_{QZ|\pi_Z}(Z) \leq \log A],
\]

(5.385)
Moreover, as in (5.375), we have

\[
\mathbb{P}[|L| \geq T + 1 | W = 1] \\
\leq \frac{1}{1 + \beta} + e^{-(\beta + 1)} + \beta \exp(-\delta) \tag{5.386}
\]

By union bound,

\[
\mathbb{P}[\text{error|no message}] \leq \frac{1}{A}; \tag{5.387}
\]

and for each \( m = 1, \ldots, M, \)

\[
\mathbb{P}[\text{error}\mid W = m] \leq \mathbb{P}\left[ i_{Z;X}(Z; X) \leq \log \frac{LM}{T} + \delta \right] \\
+ \mathbb{P}[i_{Q_Z\Pi_Z}(Z) \leq \log A] \\
+ \frac{1}{1 + \beta} + e^{-(\beta + 1)} + \beta \exp(-\delta). \tag{5.388}
\]

- Decoder 3:

The analysis is similar to that of Decoder 2 and the result follows from (5.382) and (5.385).
Chapter 6

The Counterpoint

While the previous chapters have demonstrated the usages of functional formulations of information measures and functional-analytic methods in the achievability and converse proofs of operational problems in information theory, the present chapter is devoted to the antithesis, that is, how entropic formulations and information-theoretic methods can be used to prove functional inequalities. Most notably, the Gaussian rotational invariance implies the Gaussian optimality in the Brascamp-Lieb inequality and its variations, and the proofs appear to be conveniently carried out in the information-theoretic formulations than in their traditional functional versions [89].

Data processing property, tensorization, and convexity are studied in Section 6.1. Section 6.2 proves the Gaussian optimality in a generalization of the Brascamp-Lieb inequality where the deterministic linear maps are generalized by Gaussian random transformations\(^1\). In most cases, we are able to prove the Gaussian extremality and uniqueness of the minimizer under a certain non-degenerate assumption, while establishing the Gaussian exhaustibility in full generality\(^2\). In Section 6.3 we further establish the Gaussian optimality in the forward-reverse Brascamp-Lieb inequality.

Section 6.4 discusses several implications of the Gaussian optimality results: some quantities/rate regions arising in information theory can be efficiently computed by...

---

\(^1\)That is, a random transformation \(x \mapsto Ax + w\) where \(A\) is deterministic and \(w\) is a Gaussian vector independent of \(x\).

\(^2\)See the beginning of Section 6.2 for precise definitions of extremisability and exhaustibility.
solving a finite dimensional optimization problem in the Gaussian cases. Examples include multi-variate hypercontractivity, Wyner’s common information for multiple variables, and certain secret key or common randomness generation problems. The relationship between the Gaussian optimality in the forward-reverse Brascamp-Lieb inequality and the transportation-cost inequalities for Gaussian measures is also discussed.

6.1 Data Processing, Tensorization and Convexity

Given $Q_X$ and $(Q_{Y_j|X})$, denote by $G_{BL}(Q_X,(Q_{Y_j|X}))$ the set of $(d,(c_j))$ in Theorem 2.2.3 (forward Brascamp-Lieb inequality) such that (2.14) holds. In this section we show that some elementary properties of $G_{BL}(Q_X,(Q_{Y_j|X}))$ follows conveniently from the information-theoretic characterization (2.15).

6.1.1 Data Processing

Loosely speaking, the set $G_{BL}(Q_X,(Q_{Y_j|X}))$ characterizes the level of “uncorrelatedness” between $X$ and $(Y_1, \ldots, Y_m)$. The following data processing property captures this intuition:

Proposition 6.1.1 1. Given $Q_W$, $Q_{X|W}$ and $(Q_{Y_j|X})_{j=1}^m$, assume that $Q_{WXY_j} = Q_WQ_{X|W}Q_{Y_j|X}$ for each $j$. If $(0,(c_j)) \in G_{BL}(Q_X,(Q_{Y_j|X}))$, then $(0,(c_j)) \in G_{BL}(Q_W,(Q_{Y_j|W}))$.

2. Given $Q_X$, $(Q_{Y_j|X})_{j=1}^m$ and $(Q_{Z_j|Y_j})_{j=1}^m$, assume that $Q_{XY_jZ_j} = Q_XQ_{Y_j|X}Q_{Z_j|Y_j}$ for each $j$. Then $G_{BL}(Q_X,(Q_{Y_j|X})) \subset G_{BL}(Q_X,(Q_{Z_j|X}))$.}

The proof is omitted since it follows immediately from the monotonicity of the relative entropy and (2.15).
6.1.2 Tensorization

The term "tensorization" refers to the phenomenon of additivity/multiplicativity in certain functional inequalities under tensor products. In information theory this is a central feature of many converse proofs, and is closely related to the fact that some operational problems admit single-letter solutions. In functional analysis, this provides a “particularly cute” [195] tool for proving many inequalities in arbitrary dimensions. As a close example, Lieb’s proof [89] of the Brascamp Lieb inequality relies on a special case of Proposition 6.1.2 below, where the proof uses the (functional version of) Brascamp-Lieb inequality and the Minkowski inequality. The original proof of Brascamp-Lieb inequality [83] is also based on a tensor power construction.

**Proposition 6.1.2** Suppose \((d^{(i)}, (c_j)) \in G_{BL}(Q_{X}^{(i)}, (Q_{Y_{j}|X})^{(i)})\) for \(i = 1, 2\). Then

\[
(d^{(1)} + d^{(2)}, (c_j)) \in G_{BL}\left(Q_{X}^{(1)} \times Q_{X}^{(2)}, \left(Q_{Y_{j}|X}^{(1)} \times Q_{Y_{j}|X}^{(2)}\right)\right)
\]

where \(d^{(1)} + d^{(2)}\) is defined as the function

\[
X^{(1)} \times X^{(2)} \rightarrow \mathbb{R};
\]

\[
(x_1, x_2) \mapsto d^{(1)}(x_1) + d^{(2)}(x_2).
\]

We provide a simple information-theoretic proof using the chain rules of the relative entropy. Note that the algebraic expansions here are similar to the ones in the proof of Gaussian optimality in Section 6.2 or the converse proof for the key generation problem in Section 6.4.5.

**Proof** For any arbitrary \(P_{X^{(1)}X^{(2)}}\), define \(P_{X^{(1)}X^{(2)}Y^{(1)}Y^{(2)}} := P_{X^{(1)}X^{(2)}}Q_{Y_{j}|X}^{(1)}Q_{Y_{j}|X}^{(2)}\).

Observe that

\[
D(P_{X^{(1)}X^{(2)}}\|Q_{X}^{(1)} \times Q_{X}^{(2)}) = D(P_{X^{(1)}}\|Q_{X}^{(1)}) + D(P_{X^{(2)}|X^{(1)}}\|Q_{X}^{(2)}|P_{X^{(1)}}).
\]
\[ D(P_{Y_j^{(1)}} Y_j^{(2)} \| Q_{Y_j}^{(1)} \times Q_{Y_j}^{(2)}) = D(P_{Y_j^{(1)}} Y_j^{(2)} \| Q_{Y_j}^{(1)}) + D(P_{Y_j^{(2)} | Y_j^{(1)}} Q_{Y_j}^{(2)} | P_{Y_j^{(1)}}) \] (6.4)
\[ \leq D(P_{Y_j^{(1)}} Q_{Y_j}^{(1)}) + D(P_{Y_j^{(2)} | X^{(1)} Y_j^{(1)}} Q_{Y_j}^{(2)} | P_{X^{(1)} Y_j^{(1)}}) \] (6.5)
\[ = D(P_{Y_j^{(1)}} Q_{Y_j}^{(1)}) + D(P_{Y_j^{(2)} | X^{(1)}} Q_{Y_j}^{(2)} | P_{X^{(1)}}) \] (6.6)

where (6.5) uses Jensen’s inequality, and (6.6) is from the Markov chain \( \hat{Y}_j^{(2)} - \hat{X}^{(1)} - \hat{Y}_j^{(1)} \), wherein \( (\hat{X}^{(i)}, \hat{Y}_j^{(i)}) \sim P_{X^{(i)} Y_j^{(i)}} \) for \( i = 1, 2, j = 1, \ldots, m \). By the assumption and the law of total expectation,

\[ D(P_{X^{(1)}} Q_{X}^{(1)}) + \mathbb{E}[d(\hat{X}^{(1)})] \geq \sum_{j=1}^{m} c_j D(P_{Y_j^{(1)}} Q_{Y_j}^{(1)}); \] (6.7)
\[ D(P_{X^{(2)} | X^{(1)}} Q_{X}^{(2)} | P_{X^{(1)}}) + \mathbb{E}[d(\hat{X}^{(2)})] \geq \sum_{j=1}^{m} c_j D(P_{Y_j^{(2)} | X^{(1)}} Q_{Y_j}^{(2)} | P_{X^{(1)}}). \] (6.8)

Adding up (6.7) and (6.8) and applying (6.3) and (6.6), we obtain

\[ D(P_{X^{(1)} X^{(2)}} Q_{X}^{(1)} \times Q_{X}^{(2)}) + \mathbb{E}[(d^{(1)} + d^{(2)}) (X^{(1)}, X^{(2)})] \geq \sum_{j=1}^{m} c_j D(P_{Y_j^{(1)} Y_j^{(2)}} Q_{Y_j}^{(1)} \times Q_{Y_j}^{(2)}) \] (6.9)

as desired.

A functional proof of the tensorization of reverse Brascamp-Lieb inequalities can be given by generalizing the proof of the tensorization of the Prékopa-Leindler inequality (see for example [196]). Alternatively, information-theoretic proofs of the tensorization of these reverse-type inequalities can be extracted from the proof of the Gaussian optimality in Theorem 6.3.1 ahead, and we omit the repetition here.

### 6.1.3 Convexity

Another property which follows conveniently from the information-theoretic characterization of \( G_{BL}(\cdot) \) is convexity:
Proposition 6.1.3 If \((d^i, (c_j^i)) \in \mathcal{G}_{BL}(Q_X, (Q_{Yj|X}))\) for \(i = 0, 1\), then \((d^\theta, (c_j^\theta)) \in \mathcal{G}_{BL}(Q_X, (Q_{Yj|X}))\) for \(\theta \in [0, 1]\), where we have defined

\[
d^\theta := (1 - \theta)d^0 + \theta d^1, \\
c_j^\theta := (1 - \theta)c_j^0 + \theta c_j^1, \quad \forall j \in \{1, \ldots, m\}.
\]

Proof Follows immediately from (2.36) and taking convex combinations.

Note that by taking \(m = 2, X = (Y_1, Y_2), d^i(\cdot) = 0\) and \(Q_{Yj|X}\) to be the projection to the coordinates, we recover the Riesz-Thorin theorem on the interpolation of operator norms in the special case of nonnegative kernels. This information-theoretic proof (for this special case) is much simpler than the common proof of the Riesz-Thorin theorem in functional analysis based on the three-lines lemma, because the \(c_j\)'s only affect the right side of (2.15) as linear coefficients, rather than as tilting of the distributions or functions.

6.2 Gaussian Optimality in the Forward Brascamp-Lieb Inequality

In this section we prove the Gaussian extremality in several information-theoretic inequalities related to the forward Brascamp-Lieb inequality. Specifically, we first establish this for an inequality involving conditional differential entropies, which immediately implies the variants involving conditional mutual informations or differential entropies; the latter is directly connected to the Brascamp-Lieb inequality, as Theorem 2.2.3 showed. These extremal inequalities have implications for certain operational problems in information theory, and quite interestingly, the essential steps in the proofs of these extremal inequalities follow the same patterns as the reverse proofs for the corresponding operational problems.
Of course, in the functional analysis literature there has been a lot of work supplying various proofs of the Brascamp-Lieb inequality (Theorem 2.2.1), based on different properties of the Gaussian distribution/function which we summarize below:

- The tensor power of a one-dimensional Gaussian distribution is a multidimensional Gaussian distribution, which is stable under Schwarz symmetrization (i.e. spherically decreasing rearrangement).

- Rotational invariance: if \( f \) is a one-dimensional Gaussian function, then

\[
f(x)f(y) = f \left( \frac{x-y}{\sqrt{2}} \right) f \left( \frac{x+y}{\sqrt{2}} \right).
\]  

(6.12)

- The convolution of Gaussian functions is Gaussian.

- If a real valued random variable is added to an independent Gaussian noise, then the derivative of the differential entropy of the sum with respect to the variance of the noise is half the Fisher information (de Bruijn’s identity), and of course the non-Gaussianness of the sum eventually disappears as the variance goes to infinity.

Among these, the rotation invariance principle also forms the basis of the proof strategy in this thesis. Rotation invariance is a characterizing property of the Gaussian distribution: two independent random variables are both Gaussian if their sum is independent of their difference (i.e. Cramer’s theorem \[198\]). This rotation invariance argument\[3\] has been used in establishing Gaussian extremality by Lieb \[89\], Carlen \[197\] and recently in information theory by Geng-Nair \[92\] \[199\], Courtade-Jiao \[200\] and Courtade \[201\]. Some related ideas have also appeared in the literature on the Brascamp-Lieb inequality, such as the observation that convolution preserves the ex-

\[3\]This argument was referred to as “O(2)-invariance” in \[98\] and “doubling trick” in \[197\].
tremizers of Brascamp-Lieb inequality [90, Lemma 2] due to Ball. However, as keenly noted in [92], applying the rotation invariance/doubling trick on the information-theoretic formulation has certain advantages. For example, the chain rules provide convenient tools, and the establishment of the extremality usually follows similar steps as the converse proofs of the corresponding operational problems in information theory. Since the optimization problems we consider involve many information-theoretic terms, we introduce a simplification/strengthening of the Geng-Nair approach by perturbing the coefficients in the objective function (see Remark 6.2.2), thus giving rise to some identities which become handy in the proof. A similar idea was used in [200], and this should be applicable to a wide range of other problems.

In this section, $X, Y_1, \ldots, Y_m$ are assumed to be Euclidean spaces of dimensions $n, n_1, \ldots, n_m$. To be specific about the notions of Gaussian optimality, we adopt some terminologies from [98]:

**Definition 6.2.1**

• **Extremisability:** a certain supremization/infimization is finitely attained by some argument.

• **Gaussian extremisability:** a certain supremization/infimization is finitely attained by Gaussian function/Gaussian distributions.

• **Gaussian exhaustibility:** the value of a certain supremization/infimization does not change when the arguments are restricted to the subclass of Gaussian functions/Gaussian distributions.

Most of the times, we will be able to prove Gaussian extremisability in a certain non-degenerate case, while showing Gaussian exhaustibility in general.
6.2.1 Optimization of Conditional Differential Entropies

Fix \( M \geq 0, c_0 \in [0, \infty), c_1, \ldots, c_m \in (0, \infty) \), and Gaussian random transformations \( Q_{Y_j|X} \) for \( j \in \{1, \ldots, m\} \). For each \( P_{XU} \)\footnote{In the case of standard Borel space, the conditional distribution \( P_{X|U} \) can be uniquely defined from the joint distribution \( P_{XU}, P_U \)-almost surely; see e.g. [43].} define

\[
F(P_{XU}) := h(X|U) - \sum_{j=1}^{m} c_j h(Y_j|U) - c_0 \text{Tr}[M \Sigma_{X|U}],
\]

(6.13)

where \( X \sim P_X \) (the marginal of \( P_{XU} \)) and \( Y_j \) has distribution induced by \( P_X \rightarrow Q_{Y_j|X} \rightarrow P_{Y_j} \). We have defined the differential entropy and the conditional differential entropies as

\[
h(X) := -D(P_X \| \lambda);
\]

(6.14)

\[
h(X|U = u) := -D(P_{X|U=u} \| \lambda), \quad \forall u \in U;
\]

(6.15)

\[
h(X|U) := \int h(X|U = u) dP_U;
\]

(6.16)

where \( \lambda \) is the Lebesgue measure (with the same dimension as \( X \)), and (6.16) is defined whenever the integral exists. Moreover, we have used the notation \( \Sigma_{X|U} := E[\text{Cov}(X|U)] \) for the expectation of the conditional covariance matrix.

**Definition 6.2.2** We say \( (Q_{Y_1|X}, \ldots, Q_{Y_m|X}) \) is non-degenerate if each \( Q_{Y_j|X=0} \) is a \( n_j \)-dimensional Gaussian distribution with invertible covariance matrix.

In the non-degenerate case, we can show an extremal result for the following optimization with a regularization on the covariance of the input.

**Theorem 6.2.1** If \( (Q_{Y_1|X}, \ldots, Q_{Y_m|X}) \) is non-degenerate, then \( \sup_{P_{XU}} \{ F(P_{XU}) : \Sigma_{X|U} \leq \Sigma \} \) is finite and is attained by a Gaussian \( X \) and constant \( U \). Moreover, the covariance of such \( X \) is unique.
In Theorem 6.2.1 we assume that the supremum is over $P_{X|U}$ such that $P_{X|U = u}$ is absolutely continuous with respect to the Lebesgue measure (hence having a density function) for almost every $u$. Additionally, we adopt the following convention in all the optimization problems in Section 6.2 and Section 6.4 unless otherwise specified. This eliminates situations such as $\infty + \infty$ or $a + \infty$ which can be considered as legitimate calculations but are technically difficult to deal with.

**Convention 1** The sup or inf are taken over all arguments such that each term in the objective function (e.g. (6.13)) is well-defined and finite.

**Proof of Theorem 6.2.1** Assume that both $P_{X^{(1)}|U^{(1)}}$ and $P_{X^{(2)}|U^{(2)}}$ are maximizers of (6.13) subject to $\Sigma_{X|U} \leq \Sigma$; the proof of the existence of maximizer is deferred to Appendix 6.5. Let $(U^{(1)}, X^{(1)}, Y^{(1)}_1, \ldots, X^{(1)}_m) \sim P_{X^{(1)}|U^{(1)}} Q_{Y|X} \prod Q_{Y_m|X}$ and $(U^{(2)}, X^{(2)}, Y^{(2)}_1, \ldots, X^{(2)}_m) \sim P_{X^{(2)}|U^{(2)}} Q_{Y|X} \prod Q_{Y_m|X}$ be mutually independent. Define

$$X^+ = \frac{1}{\sqrt{2}} (X^{(1)} + X^{(2)}) \quad X^- = \frac{1}{\sqrt{2}} (X^{(1)} - X^{(2)}). \quad (6.17)$$

Define $Y^+_j$ and $Y^-_j$ similarly for $j = 1, \ldots, m$, and put $\hat{U} = (U^{(1)}, U^{(2)})$. We now make three important observations:

1. First, due to the Gaussian nature of $Q_{Y_j|X}$, it is easily seen that $Y^+_j | (X^+, X^- = x^-, \hat{U} = \hat{u}) \sim Q_{Y_j|X = x^+}$ is independent of $x^-$. Thus $Y^+_j | (X^+ = x, \hat{U} = \hat{u}) \sim Q_{Y_j|X = x}$ as well. Similarly, $Y^-_j | (X^- = x, \hat{U} = u) \sim Q_{Y_j|X = x}$ for $j = 1, \ldots, m$.

---

\footnote{Since the integral in (6.16) may not be well-defined in general, some authors have restricted the attention to finite $\mathcal{U}$ in the optimization problems. In this chapter, we are allowed to drop this restriction as long as the integral in (6.16) is well-defined. These distinctions do not appear to make an essential difference for our purpose; see Footnote 13}
2. Second, observe that for each $\hat{u} = (u_1, u_2)$ we can verify the algebra

$$
\Sigma_{X^+|\hat{U} = \hat{u}} = \mathbb{E}[(X^+ - \mu_{X^+|\hat{U}})(X^+ - \mu_{X^+|\hat{U}})^\top | \hat{U} = \hat{u}]
= \frac{1}{2}\mathbb{E}[(X^{(1)} - \mu_{X^{(1)}|\hat{U}})(X^{(1)} - \mu_{X^{(1)}|\hat{U}})^\top | \hat{U} = \hat{u}]
+ \frac{1}{2}\mathbb{E}[(X^{(2)} - \mu_{X^{(2)}|\hat{U}})(X^{(2)} - \mu_{X^{(2)}|\hat{U}})^\top | \hat{U} = \hat{u}]
+ \mathbb{E}[(X^{(1)} - \mu_{X^{(1)}|\hat{U}})(X^{(2)} - \mu_{X^{(2)}|\hat{U}})^\top | \hat{U} = \hat{u}]
= \frac{1}{2}\mathbb{E}[(X^{(1)} - \mu_{X^{(1)}|U^{(1)}})(X^{(1)} - \mu_{X^{(1)}|U^{(1)}})^\top | \hat{U} = \hat{u}]
+ \frac{1}{2}\mathbb{E}[(X^{(2)} - \mu_{X^{(2)}|U^{(2)}})(X^{(2)} - \mu_{X^{(2)}|U^{(2)}})^\top | \hat{U} = \hat{u}]
+ \mathbb{E}[(X^{(1)} - \mu_{X^{(1)}|U^{(1)}})(X^{(2)} - \mu_{X^{(2)}|U^{(2)}})^\top | \hat{U} = \hat{u}],
$$

(6.18)

The last term above vanishes upon averaging over $(u_1, u_2)$ because of the independence $(U^{(1)}, X^{(1)}) \perp (U^{(2)}, X^{(2)})$. Thus

$$
\Sigma_{X^+|\hat{U}} = \frac{1}{2}\Sigma_{X^{(1)}|U^{(1)}} + \frac{1}{2}\Sigma_{X^{(2)}|U^{(2)}} \leq \Sigma.
$$

(6.19)

By the same token,

$$
\Sigma_{X^-|\hat{U}} = \frac{1}{2}\Sigma_{X^{(1)}|U^{(1)}} + \frac{1}{2}\Sigma_{X^{(2)}|U^{(2)}} \leq \Sigma.
$$

(6.20)

These combined with $\Sigma_{X^-|X^+\hat{U}} \leq \Sigma_{X^-|\hat{U}}$ justify that both $P_{X^+\hat{U}}$ and $P_{X^-\hat{U}X^+}$ satisfy the covariance constraint $\Sigma_{X|\hat{U}} \leq \Sigma$ in the theorem.
3. Third, we have

\[
\sum_{k=1}^{2} \left[ h(X^{(k)}|U^{(k)}) - \sum_{j=1}^{m} c_j h(Y_j^{(k)}|U^{(k)}) \right] = h(X^{(1)}, X^{(2)}|\hat{U}) - \sum_{j=1}^{m} c_j h(Y_j^{(1)}, Y_j^{(2)}|\hat{U}) \tag{6.21}
\]

\[
h(X^+, X^-|\hat{U}) - \sum_{j=1}^{m} c_j h(Y_j^+, Y_j^-|\hat{U}) \tag{6.22}
\]

\[
h(X^+|\hat{U}) - \sum_{j=1}^{m} c_j h(Y_j^+|\hat{U}) + h(X^-|X^+, \hat{U}) - \sum_{j=1}^{m} c_j h(Y_j^-|X^+, \hat{U}) \tag{6.23}
\]

\[
h(X^+|\hat{U}) - \sum_{j=1}^{m} c_j h(Y_j^+|\hat{U}) + h(X^-|X^+, \hat{U}) - \sum_{j=1}^{m} c_j h(Y_j^-|X^+, \hat{U}), \tag{6.24}
\]

where the final inequality follows from the Markov chain \( Y_j^- - X^+ \hat{U} - Y_j^+ \), which is because the joint distribution factorizes as \( P_{\hat{U}X^+X^-Y_j^+Y_j^-} = P_{\hat{U}X^+X^-Q_{Y_j}XQ_{Y_j}X} \).

Thus, we can conclude that

\[
\sum_{i=1}^{2} F(P_{X_{(i)}U_{(i)}}) = \sum_{i=1}^{2} \left[ h(X^{(k)}|U^{(k)}) - \sum_{j=1}^{m} c_j h(Y_j^{(k)}|U^{(k)}) - c_0 \text{Tr}[M\Sigma_{X^{(k)}|U^{(k)}}] \right] \tag{6.26}
\]

\[
h(X^+|\hat{U}) - \sum_{j=1}^{m} c_j h(Y_j^+|\hat{U}) - c_0 \text{Tr}[M\Sigma_{X^+|\hat{U}}] + h(X^-|X^+, \hat{U}) - \sum_{j=1}^{m} c_j h(Y_j^-|X^+, \hat{U}) - c_0 \text{Tr}[M\Sigma_{X^-|X^+, \hat{U}}] \tag{6.27}
\]

\[
\leq \sum_{i=1}^{2} F(P_{X_{(i)}U_{(i)}}), \tag{6.28}
\]

where

- (6.27) follows from (6.19), (6.20) and (6.25);
• (6.28) follows since $P_{X^+}U$ and $P_{X^-}U$ are candidate optimizers of (6.13) subject to the given covariance constraint whereas $P_{X(i)U(i)}$ are the optimizers by assumption $(i = 1, 2)$.

Then, the equalities in (6.26)-(6.28) must be achieved throughout, so both $P_{X^+}U$ and $P_{X^-}U$ (and also $P_{X^+}U$ by symmetry of the argument) are maximizers of (6.13) subject to $\Sigma_{X|U} \leq \Sigma$.

So far, we have considered fixed coefficients $c_0^m = (c_0, \ldots, c_m)$. The same argument applies for coefficients on a line:

$$c_0^m(t) := ta_0^m, \quad t > 0$$

for any fixed $a_0 \in [0, \infty)$, $a_1, \ldots, a_m \in (0, \infty)$, and we next show several properties for a dense subset of this line. Applying Lemma 6.2.3 (following this proof) with

$$p(P_{X|U}) \leftarrow h(X|U);$$

$$q(P_{X|U}) \leftarrow -\sum_{j=1}^m a_j h(Y_j|U) - a_0 \text{ Tr}[M\Sigma_{X|U}],$$

$$f(t) \leftarrow \max_{P_{X|U}}[p(P_{X|U}) + tq(P_{X|U})],$$

we obtain from the optimality of $P_{X^+}U$, $P_{X^-}U$ and $P_{X^+}U$ that

$$h(X^+|\hat{U}) = h(X^-|X^+,\hat{U}) = h(X^+|X^-,\hat{U})$$

for almost all $t \in (0, \infty)$, where $P_{X^+X^-}$ depends implicitly on $t$. Note that (6.33) implies that $I(X^+;X^-|\hat{U}) = 0$ hence $X^+$ and $X^-$ are independent conditioned on $\hat{U}$. Recall the following Skitovic-Darmois characterization of Gaussian distributions (with the extension to the vector Gaussian case in [92]):

236
Lemma 6.2.2  Let $A_1$ and $A_2$ be mutually independent $n$-dimensional random vectors. If $A_1 + A_2$ is independent of $A_1 - A_2$, then $A_1$ and $A_2$ are normally distributed with identical covariances.

Using Lemma 6.2.2 we can conclude that for almost all $t \in (0, \infty)$, $X^{(i)}$ must be Gaussian, with covariance not depending on $U^{(i)}$, thus $U^{(i)}$ can be chosen as a constant $(i = 1, 2)$. Thus for all such $t$,

$$f(t) = \max_{P_{X_U} : U = \text{const.}, X \text{ Gaussian}} [p(P_{X_U}) + tq(P_{X_U})].$$

(6.34)

Since both sides of (6.34) are concave in $t$, hence continuous on $(0, \infty)$, we see (6.34) actually holds for all $t \in (0, \infty)$. The proof is completed since $a_0^m$ can be arbitrarily chosen.

Lemma 6.2.3  Let $p$ and $q$ be real-valued functions on an arbitrary set $D$. If $f(t) := \max_{x \in D} \{p(x) + tq(x)\}$ is always attained, then for almost all $t$, $f'(t)$ exists and

$$f'(t) = q(x^*), \quad \forall x^* \in \text{arg max}_{x \in D} \{p(x) + tq(x)\}.$$  

(6.35)

In particular, for all such $t$, $q(x^*) = q(\tilde{x}^*)$ and $p(x^*) = p(\tilde{x}^*)$ for all $x^*, \tilde{x}^* \in \text{arg max}_{x \in D} \{p(x) + tq(x)\}$.

Geometrically, $f(t)$ is the support function of the set $S := \{(p(x), q(x))\}_{x \in D}$ evaluated at $(1, t)$. Hence $f(\cdot)$ is convex, and the left and the right derivatives are determined by the two extreme points of the intersection between $S$ and the supporting hyperplane.

Proof of Lemma 6.2.3  The function $f$ is convex since it is a pointwise supremum of linear functions, and is therefore differentiable almost everywhere. Moreover, $f'(t)$ (which is well-defined in the a.e. sense) is monotone increasing by convexity, and is therefore continuous almost everywhere.
Let \( x^*_i \) denote an arbitrary element of \( \arg \max_{x \in D} \{ p(x) + tq(x) \} \). By definition, for any \( s \in \mathbb{R} \), \( f(s) \geq p(x^*_i) + sq(x^*_i) \). Thus, for \( \delta > 0 \),

\[
\frac{f(t + \delta) - f(t)}{\delta} \geq q(x^*_i) \quad \text{and} \quad \frac{f(t) - f(t - \delta)}{\delta} \leq q(x^*_i). \tag{6.36}
\]

If \( f'(t) \) exists, that is, the left sides of the two inequalities above have the same limit \( f'(t) \) as \( \delta \downarrow 0 \), then \( f'(t) = q(x^*_i) \). The second claim of the lemma follows immediately from the first.

**Remark 6.2.1** Let us remark on an interesting connection between the above proof of the optimality of Gaussian random variable and Lieb’s proof of that Gaussian functions maximize Gaussian kernels. Recall that [89, Theorem 3.2] wants to show that for an operator \( G \) given by a two-variate Gaussian kernel function and \( p,q > 1 \), the ratio \( \frac{\|Gf\|_p}{\|f\|_p} \) is maximized by Gaussian \( f \). First, a tensorization property is proved, implying that \( f^*(x_1)f^*(x_2) \) is a maximizer for \( G \otimes G \) if \( f^* \) is any maximizer of \( G \).

Then, Lieb made two important observations:

1. By a rotation invariance property of the Lebesgue measure/isotropic Gaussian measure, \( f^*(\frac{z_1+z_2}{\sqrt{2}})f^*(\frac{z_1-z_2}{\sqrt{2}}) \) is also a maximizer of \( G \otimes G \).

2. An examination of the equality condition in the proof of tensorization property reveals that any maximizer for \( G \otimes G \) must be of a product form.

Thus Lieb concluded that \( f^*(\frac{z_1+z_2}{\sqrt{2}})f^*(\frac{z_1-z_2}{\sqrt{2}}) = \alpha(x_1)\beta(x_2) \) for some functions \( \alpha \) and \( \beta \), and \( f^* \) must be a Gaussian function. This is very similar to the above proof, once we think of \( f^* \) as the density function of \( P^*_{X|U=u} \) in our proof.

**Remark 6.2.2** Our proof technique is essentially following ideas of Geng and Nair [92][199] who established the Gaussian optimality for several information-theoretic regions. However, we also added the important ingredient of Lemma 6.2.3\(^6\). That is,

---

\(^6\)The similar idea of differentiating the coefficients has been used in [200].
by differentiating with respect to the linear coefficients, we can conveniently obtain
information-theoretic identities which helps us to conclude the conditional indepen-
dence of $X^+$ and $X^-$ quickly. For fixed $m$, in principle, this may be avoided by trying
various expansions of the two-letter quantities manually (e.g. as done in [192]), but
that approach will become increasingly complicated and unstructured as $m$ increases.
Finally, we also note that a simple rotational invariance argument/doubling trick has
been used for proving that the capacity achieving distribution for an additive Gaussian
channel is Gaussian (cf. [202, P36]), which does not involve Lemma 6.2.2 and whose
extension to problems involving auxiliary random variables is not clear.

If we do not have the non-degenerate assumption and the regularization $\Sigma_{X|U} \preceq \Sigma$, it is very well possible that the optimization in Theorem 6.2.1 is nonfinite and/or not attained by any $P_{U|X}$. In this case, we can show that the optimization is exhausted by Gaussian distributions. To state the result conveniently, for any $P_X$, define

$$F_0(P_X) := h(X) - \sum_{j=1}^m c_j h(Y_j) - c_0 \text{Tr}[M\Sigma_X], \quad (6.37)$$

where $(X, Y_j) \sim P_X Q_{Y_j|X}$. Apparently, $F(P_{X|U}) = F_0(P_X)$ when $U$ is constant.

**Theorem 6.2.4** In the general (possibly degenerate) case,

1. For any given positive semidefinite $\Sigma$,

$$\sup_{P_{X|U}, \Sigma_{X|U} \preceq \Sigma} F(P_{X|U}) = \sup_{P_X \text{ Gaussian}, \Sigma_X \preceq \Sigma} F_0(P_X) \quad (6.38)$$

where the left side of (6.38) follows Convention [1].

2.

$$\sup_{P_{X|U}} F(P_{X|U}) = \sup_{P_X \text{ Gaussian}} F_0(P_X). \quad (6.39)$$
Remark 6.2.3 Theorem 6.2.4 reduces an infinite dimensional optimization problem to a finite dimensional one. In particular, in the degenerate case we can verify that the left side of (6.38) to be extremisable (Definition 6.2.1) if the the right side of (6.38) is extremisable.

Proof Theorem 6.2.4 Note that the Gaussian random transformation $Q_{Y_j|X}$ can be realized as a linear transformation $B_j$ to $\mathbb{R}^{n_j}$ followed by adding an independent Gaussian noise of covariance $\Sigma_j$. We will assume that $\Sigma_j$ is non-degenerate in the orthogonal complement of the image of $B_j$, since otherwise both sides of (6.38) are $+\infty$. In particular, if $Q_{Y_j|X}$ is a deterministic linear transform $B_j$, then it must be onto $\mathbb{R}^{n_j}$.

1. We can use the argument at the beginning of Appendix 6.5 to show that the left side of (6.38) equals

$$
\sup_{P_{X|U}: U=\{1,\ldots, D\}, \Sigma_{X|U} \leq \Sigma} F(P_{X|U})
$$

(6.40)

where $D$ is a fixed integer (depending only on the dimension of $\Sigma$).

Next, we argue that it is without loss of generality to add a restriction to (6.40) that $P_{X|U=u}$ has a smooth density with compact support for each $u \in \{1, \ldots, D\}$, in which case each $P_{Y_j|U=u}$ will have a smooth density by the assumption made at the beginning of the proof. For this, define $F_0(\cdot)$ as in (6.37), and we will show that for any $P_X$ absolutely continuous with respect to the Lebesgue measure and any $\epsilon > 0$, we can choose a distribution $P_{X^*}$ whose density is smooth with compact support such that

$$
F_0(P_{X^*}) \geq F_0(P_X) - \epsilon
$$

(6.41)
and

\[ \Sigma_{X''} \leq (1 + \epsilon) \Sigma_X. \]  

(6.42)

First, by conditioning \( X \) on a large enough compact set and applying dominated convergence theorem, there exists \( P_{X'} \), which is supported on a compact set and whose density \( f_{X'} \) is bounded, such that each term in the definition of \( F_0(P_X) \) is well-approximated when \( X \) is replaced with \( X' \), so that

\[ F_0(P_{X'}) \geq F_0(P_X) - \frac{\epsilon}{2} \]  

(6.43)

and

\[ \Sigma_{X'} \leq \sqrt{1 + \epsilon} \Sigma_X. \]  

(6.44)

Second, we can pick \( P_{X''} \) with smooth density with compact support such that \( \| f_{X''} - f_{X'} \|_1 \) can be made arbitrarily small, which implies that \( \| f_{Y''} - f_{Y'} \|_1 \) is small by Jensen’s inequality, and that \( \operatorname{Tr}[M \Sigma_{X''}] - \operatorname{Tr}[M \Sigma_{X'}] \) is small by the fact that the densities are supported on a compact set. Since \( x \mapsto x \log x \) is Lipschitz on a bounded interval, the differential entropy terms can be made small as well (here we used the boundedness of \( f_{X'} \)). As a result, we can ensure that

\[ F_0(P_{X''}) \geq F_0(P_X') - \frac{\epsilon}{2} \]  

(6.45)

and

\[ \Sigma_{X''} \leq \sqrt{1 + \epsilon} \Sigma_{X'}, \]  

(6.46)
which, combined with (6.43) and (6.44), yield (6.41)-(6.42). This and the observation in (6.40) imply that

\[
\inf_{\epsilon > 0} \sup_{P_{XU} \in C_\epsilon} F(P_{XU}) \leq \sup_{P_{XU} : \Sigma_{X|U} \leq \Sigma} F(P_{XU}) \tag{6.47}
\]

where \( C_\epsilon \) denotes the set of \( P_{XU} \) such that \( U = \{1, \ldots, D\} \), \( \Sigma_{X|U} \leq (1 + \epsilon)\Sigma \), and the density of \( P_{X|U=u} \) is smooth with compact support for each \( u \). In the final steps, we write \( F \) as \( F^Q \) to indicate its dependence on \( (Q_{Y_1|X}, \ldots, Q_{Y_m|X}) \), and define \( \tilde{Q} \) the set of non-degenerate \( (Q_{\tilde{Y}_1|X}, \ldots, Q_{\tilde{Y}_m|X}) \) where each \( \tilde{Y}_j \) is obtained by adding an independent Gaussian noise with covariance \( \tilde{\Sigma}_j \) to \( Y_j \) (here the random transformation from \( X \) to \( Y_j \) is fixed and \( \tilde{\Sigma}_j \) runs over the set of positive definite matrices of the given dimension, \( j = 1, \ldots, m \)). Then

\[
\sup_{P_{XU} \in C_\epsilon} F^Q(P_{XU}) = \sup_{P_{XU} \in C_\epsilon} \sup_{\tilde{Q} \in \tilde{Q}} F^\tilde{Q}(P_{XU}) \tag{6.48}
\]

\[
= \sup_{\tilde{Q} \in \tilde{Q}} \sup_{P_{XU} \in C_\epsilon} F^\tilde{Q}(P_{XU}) \tag{6.49}
\]

\[
= \sup_{\tilde{Q} \in \tilde{Q}} \sup_{P_{X} \text{ Gaussian, } \Sigma_{X} \leq (1 + \epsilon)\Sigma} F^\tilde{Q}_0(P_{X}) \tag{6.50}
\]

\[
= \sup_{P_{X} \text{ Gaussian, } \Sigma_{X} \leq (1 + \epsilon)\Sigma} \sup_{\tilde{Q} \in \tilde{Q}} F^\tilde{Q}_0(P_{X}) \tag{6.51}
\]

\[
= \sup_{P_{X} \text{ Gaussian, } \Sigma_{X} \leq (1 + \epsilon)\Sigma} F^Q_0(P_{X}) \tag{6.52}
\]

where

- (6.48) and (6.52) are because, first, \( h(\tilde{Y}_j) \geq h(Y_j) \) by the entropy power inequality; second, \( \inf_{\tilde{\Sigma}_j > 0} h(\tilde{Y}_j) = h(Y_j) \). The proof of the second claim is standard, since \( \tilde{Y}_j \) has smooth density with fast (Gaussian like) decay,
and we can obtain pointwise convergence of the density function and apply dominated convergence theorem\footnote{Such a continuity in the variance of the additive noise can fail terribly when $f_X$ does not have the decay properties; in \cite[Proposition 4]{BobkovChistyakov} Bobkov and Chistyakov provided an example of random vector $X$ with finite differential entropy such that $h(X + Z) = \infty$ for each $Z$ independent of $X$ and having finite differential entropy. In their example, $\sup_{\|X\| \geq r} f_X(x) = 1$ for every $r > 0$, and we cannot obtain a dominating function for $|f_{X+Z} \log f_{X+Z}|$ to apply the dominated convergence theorem.}

- (6.50) is from Gaussian extremality in non-degenerate case.

Then (6.47) and (6.52) gives

\[
\sup_{P_{XU} : \Sigma_{X|U} \leq \Sigma} F(P_{XU}) \leq \sup_{\epsilon > 0} \sup_{P_X \text{ Gaussian}, \Sigma_X \leq (1+\epsilon) \Sigma} F_0^Q(P_X) \tag{6.53}
\]

\[
= \sup_{P_X \text{ Gaussian}, \Sigma_X \leq \Sigma} F_0^Q(P_X) \tag{6.54}
\]

where (6.54) is a property that can be verified without much difficulty for Gaussian distributions. Thus the $\leq$ part of (6.38) is established. The other direction is immediate from the definition.

2. First, observe that it is without loss of generality to assume that $U$ is constant, that is,

\[
\sup_{P_{XU}} F(P_{XU}) = \sup_{P_X} F_0(P_X). \tag{6.55}
\]

Next, by the same argument as [1], for any $P_X$ and $\epsilon > 0$, there exists $P_X$ such that $\|X\| < M$ with probability one for some finite $M > 0$, and that

\[
F_0(P_X') \geq F_0(P_X) - \epsilon. \tag{6.56}
\]
Then

\[
F_0(P_{X'}) \leq \sup_{\Sigma \succeq 0} \sup_{P_{X'}: \Sigma_{X'} \succeq \Sigma} F(P_{X'}) \tag{6.57}
\]

\[
\leq \sup_{\Sigma \succeq 0} \sup_{P_X: \text{Gaussian}, \Sigma_{X} \succeq \Sigma} F_0(P_X) \tag{6.58}
\]

\[
= \sup_{P_X: \text{Gaussian}} F_0(P_X) \tag{6.59}
\]

where (6.58) was established in (6.54). Finally, (6.55), (6.56), (6.59) and arbitrariness of \(\epsilon\) give

\[
\sup_{P_{X'}} F(P_{X'}) \leq \sup_{P_X: \text{Gaussian}} F_0(P_X). \tag{6.60}
\]

Thus the \(\leq\) part of (6.39) is established. The other direction is trivial from the definition.

\[\blacksquare\]

6.2.2 Optimization of Mutual Informations

Let \(X, Y_1, \ldots, Y_m\) be jointly Gaussian vectors, \(a_0, \ldots, a_m\) be nonnegative real numbers, and \(M\) be a positive-semidefinite matrix. We are interested in minimizing \(I(U; X)\) over \(P_{U|X}\) (that is, \(U - X - Y^m\) must hold) subject to \(I(U; Y_i) \geq a_i\), \(i \in \{1, \ldots, m\}\) and \(\operatorname{Tr}[M\Sigma_{X|U}] \leq a_0\). This is relevant to many problems in information theory including the Gray-Wyner network [6] and some common randomness/key generation problems [101] (to be discussed in Section 6.4.5). We have the following result for the Lagrange dual of such an optimization problem:
Theorem 6.2.5  Fix $M \geq 0$, positive constants $c_0, c_1, \ldots, c_m$, and jointly Gaussian vectors $(X, Y_1, \ldots, Y_m) \sim Q_{XY_1,\ldots,Y_m}$. Define the function

$$G(P_{U|X}) := \sum_{j=1}^{m} c_j I(Y_j; U) - I(X; U) - c_0 \text{Tr}[M\Sigma_{X|U}].$$  \hfill (6.61)

Then

1. If $(Q_{Y_1|X}, \ldots, Q_{Y_m|X})$ is non-degenerate, then $\sup_{P_{U|X}} G(P_{U|X})$ is achieved by $U^*$ for which $X|\{U^* = u\}$ is normal for $P_{U^*}$-a.e. $u$, with covariance not depending on $u$. This can be realized by a Gaussian vector $U^*$ with the same dimension as $X$.

2. In general, $\sup_{P_{U|X}} G(P_{U|X}) = \sup_{P_{U|X}: (U, X) \text{ is Gaussian}} G(P_{U|X})$.

Proof  Set $\Sigma := \Sigma_X$. Note that for any $P_{U|X}$, we have

$$\sum_{j=1}^{m} c_j I(Y_j; U) - I(X; U) - c_0 \text{Tr}[M\Sigma_{X|U}] \leq \sum_{j=1}^{m} c_j h(Y_j) - h(X) + \left( h(X|U) - \sum_{j=1}^{m} c_j h(Y_j|U) - c_0 \text{Tr}[M\Sigma_{X|U}] \right) \leq \sum_{j=1}^{m} c_j h(Y_j) - h(X) + \sup_{P_{X|U'}: \Sigma_{X|U'} \geq \Sigma} F(P_{X|U'}). \hfill (6.64)$$

In the non-degenerate case, by Theorem 6.2.1, the supremum is attained in the last line by constant $U'$ and $X' \sim N(0, \Sigma')$, where $\Sigma' \leq \Sigma$. Hence, taking $U^* \sim N(0, \Sigma - \Sigma')$ and $P_{X|U^*} \sim N(u, \Sigma')$, we have $P_X \sim N(0, \Sigma)$ as required, and (6.64) reads as

$$\sum_{j=1}^{m} c_j I(Y_j; U) - I(X; U) - c_0 \text{Tr}[M\Sigma_{X|U}] \leq \sum_{j=1}^{m} c_j I(Y_j; U^*) - I(X; U^*) - c_0 \text{Tr}[M\Sigma_{X|U^*}], \hfill (6.65)$$
and follows since $P_{U|X}$ is arbitrary. The claim for the general (possibly degenerate) case follows by invoking Theorem 6.2.4 and using a similar argument.

### 6.2.3 Optimization of Differential Entropies

The $F_0(·)$ defined in Section 6.2.1 is a linear combination of differential entropies and second order moments, which is closely related to the Brascamp-Lieb inequality. Immediately from Theorem 6.2.1 and Theorem 6.2.4, we have the following result regarding maximization of $F_0(·)$:

**Corollary 6.2.6** *For any $\Sigma \geq 0$,*

\[
\sup_{P_X: \Sigma_X \leq \Sigma} F_0(P_X) = \sup_{P_X: \text{Gaussian}, \Sigma_X \leq \Sigma} F_0(P_X). \tag{6.66}
\]

\[
\sup_{P_X} F_0(P_X) = \sup_{P_X: \text{Gaussian}} F_0(P_X), \tag{6.67}
\]

where $F_0(·)$ is defined in (6.37). In addition, in the non-degenerate case (6.66) is always finite and (up to a translation) uniquely achieved by a Gaussian $P_X$; (6.67) is finite and (up to a translation) uniquely achieved by a Gaussian $P_X$ if $M$ in (6.37) is positive definite.

**Remark 6.2.4** *Since the left side of (6.38) is obviously concave as a function of $\Sigma$, (6.38) and (6.66) combined shows that the left side of (6.66) is concave in $\Sigma$. This is not obvious from the definition; in particular it is not clear whether the concavity still holds for non-gaussian $Q_X$ and $Q_{Y|X}$.***
6.3 Gaussian Optimality in the Forward-Reverse Brascamp-Lieb Inequality

In this section, some notations and terminologies from Sections 2.4 and 6.2 will be used. Moreover, consider the following parameters/data:

- Fix Lebesgue measures $\mu_j$ and Gaussian measures $\nu_i$ on $\mathbb{R}$;
- non-degenerate (Definition 6.2.2) linear Gaussian random transformation $(P_{Y_j|X})_{j=1}^m$ (where $X := (X_1, \ldots, X_l)$) associated with conditional expectation operators $(T_j)_{j=1}^m$;
- positive $(c_j)$ and $(b_i)$.

Given Borel measures $P_{X_i}$ on $\mathbb{R}$, $i = 1, \ldots, l$, define

$$F_0((P_{X_i})) := \inf_{P_X} \sum_{j=1}^m c_j D(P_{Y_j} \| \mu_j) - \sum_{i=1}^l b_i D(P_{X_i} \| \nu_i)$$  \hspace{1cm} (6.68)

where the infimum is over Borel measures $P_X$ that has $(P_{X_i})$ as marginals. The aim of this section is to prove the following:

**Theorem 6.3.1** $\sup_{(P_{X_i})} F_0((P_{X_i}))$, where the supremum is over Borel measures $P_{X_i}$ on $\mathbb{R}$, $i = 1, \ldots, l$, is achieved by some Gaussian $(P_{X_i})_{i=1}^l$.

Naturally, one would expect that Gaussian optimality can be established when $(\mu_j)_{j=1}^m$ and $(\nu_i)_{i=1}^l$ are either Gaussian or Lebesgue. We made the assumption that the former is Lebesgue and the latter is Gaussian so that certain technical conditions can be justified conveniently, while the crux of the matter—the tensorization steps can still be demonstrated. More precisely, we have the following observation:
Proposition 6.3.2 \( \sup_{(P_{X_i})} F_0((P_{X_i})) \) is finite and there exist \( \sigma_i^2 \in (0, \infty) \), \( i = 1, \ldots, l \) such that it equals

\[
\sup_{(P_{X_i}): \mathbb{E}[X_i^2] \leq \sigma_i^2} F_0((P_{X_i})).
\]

Proof when \( \mu_j \) is Lebesgue and \( P_{Y_j|X} \) is non-degenerate, \( D(P_{Y_j}\|\mu_j) = -h(P_{Y_j}) \leq -h(P_{Y_j}|X) \) is bounded above (in terms of the variance of additive noise of \( P_{Y_j|X} \)). Moreover, \( D(P_{X_i}\|\nu_i) \geq 0 \) when \( \nu_i \) is Gaussian, so \( \sup_{(P_{X_i})} F_0((P_{X_i})) < \infty \). Further, choosing \( (P_{X_i}) = (\nu_i) \) and applying a covariance argument to lower bound the first term in (6.68) shows that \( \sup_{(P_{X_i})} F_0((P_{X_i})) > -\infty \).

To see (6.69), notice that

\[
D(P_{X_i}\|\nu_i) = D(P_{X_i}\|\nu_i') + \mathbb{E}[\nu_i'|\nu_i(X)]
\]

(6.70)

\[
= D(P_{X_i}\|\nu_i') + D(\nu_i'|\nu_i)
\]

(6.71)

\[
\geq D(\nu_i'|\nu_i)
\]

(6.72)

where \( \nu_i' \) is a Gaussian distribution with the same first and second moments as \( X_i \sim P_{X_i} \). Thus \( D(P_{X_i}\|\nu_i) \) is bounded below by some function of the second moment of \( X_i \) which tends to \( \infty \) as the second moment of \( X_i \) tends to \( \infty \). Moreover, as argued in the preceding paragraph the first term in (6.68) is bounded above by some constant depending only on \( (P_{Y_j}|X) \). Thus, we can choose \( \sigma_i^2 > 0, i = 1, \ldots, l \) large enough such that if \( \mathbb{E}[X_i^2] > \sigma_i^2 \) for some of \( i \) then \( F_0((P_{X_i})) < \sup_{(P_{X_i})} F_0((P_{X_i})) \), irrespective of the choices of \( P_{X_1}, \ldots, P_{X_{i-1}}, P_{X_{i+1}}, \ldots, P_{X_l} \). Then these \( \sigma_1, \ldots, \sigma_l \) are as desired in the proposition.

As we saw in Section 6.2 the regularization ensures that the supremum is achieved, although it might be possible to prove Gaussian exhaustibility results by taking limits.
Proposition 6.3.3  
1. For any $(P_{X_i})_{i=1}^l$, the infimum in (6.68) is attained.

2. If $(P_{Y_j|X_i})_{j=1}^m$ are non-degenerate (Definition 6.2.2), then the supremum in (6.69) is achieved by some $(P_{X_i})_{i=1}^l$.

Proof  
1. For any $\epsilon > 0$, by the continuity of measure there exists $K > 0$ such that

\[ P_{X_i}([-K, K]) \geq 1 - \epsilon/l, \quad i = 1, \ldots, l. \]  

(6.73)

By the union bound,

\[ P_X([-K, K]^l) \geq 1 - \epsilon \]  

(6.74)

wherever $P_X$ is a coupling of $(P_{X_i})$. Now let $P_{X}^{(n)}$, $n = 1, 2, \ldots$ be a such that

\[ \lim_{n \to \infty} \sum_{j=1}^m c_j D(P_{Y_j}^{(n)} \| \mu_j) = \inf_{P_X} \sum_{j=1}^m c_j D(P_{Y_j} \| \mu_j) \]  

(6.75)

where $P_{Y_j} := T_j^* P_X$, $j = 1, \ldots, m$. The sequence $(P_{X}^{(n)})$ is tight by (6.74), Thus invoking Prokhorov theorem and by passing to a subsequence, we may assume that $(P_{X}^{(n)})$ converges weakly to some $P_X^*$. Therefore $P_{Y_j}^{(n)}$ converges to $P_{Y_j}^*$ weakly, and by the semicontinuity property in Lemma 6.5.2 we have

\[ \sum_{j=1}^m c_j D(P_{Y_j}^* \| \mu_j) \leq \lim_{n \to \infty} \sum_{j=1}^m c_j D(P_{Y_j}^{(n)} \| \mu_j) \]  

(6.76)

establishing that $P_X^*$ is an infimizer.
2. Suppose \((P_{X_i}^{(n)})_{1 \leq i \leq l, n \geq 1}\) is such that \(\mathbb{E}[X_i^2] \leq \sigma_i^2\), \(X_i \sim P_{X_i}^{(n)}\), where \((\sigma_i)\) is as in Proposition 6.3.2 and

\[
\lim_{n \to \infty} F_0 \left( (P_{X_i}^{(n)})_{i=1}^l \right) = \sup_{(P_{X_i}^{(n)}) : \Sigma_{X_i} \leq \sigma_i^2} F_0 \left( (P_{X_i}^{(n)})_{i=1}^l \right). \tag{6.77}
\]

The regularization on the covariance implies that for each \(i\), \((P_{X_i}^{(n)})_{n \geq 1}\) is a tight sequence. Thus upon the extraction of subsequences, we may assume that for each \(i\), \((P_{X_i}^{(n)})_{n \geq 1}\) converges to some \(P_{X_i}^*\), and a simple truncation and min-max inequality argument (see e.g. (6.169)) shows that \(\mathbb{E}[X_i^2] \leq \sigma_i^2\), \(X_i \sim P_{X_i}^*\). Then by Lemma 6.5.2,

\[
\sum_i b_i D(P_{X_i}^* \| \nu_i) \leq \lim_{n \to \infty} \sum_i b_i D(P_{X_i}^{(n)} \| \nu_i) \tag{6.78}
\]

Under the covariance regularization and the non-degenerateness assumption, we showed in Proposition 6.3.2 that the value of \((6.69)\) cannot be \(+\infty\) or \(-\infty\). This implies that we can assume (by passing to a subsequence) that \(P_{X_i}^{(n)} \ll \lambda\), \(i = 1, \ldots, l\) since otherwise \(F((P_{X_i})) = -\infty\). Moreover, since \(\left( \sum_j c_j D(P_{Y_j}^{(n)} \| \mu_j) \right)_{n \geq 1}\) is bounded above under the non-degenerateness assumption, the sequence \(\left( \sum_i b_i D(P_{X_i}^{(n)} \| \nu_i) \right)_{n \geq 1}\) must also be bounded from above, which implies, using \((6.78)\), that

\[
\sum_i b_i D(P_{X_i}^* \| \nu_i) < \infty. \tag{6.79}
\]

In particular, we have \(P_{X_i}^* \ll \lambda\) for each \(i\). Let \(P_{X}^*\) be an infimizer as in Part 1) for marginals \((P_{X_i}^*)\). In view of Lemma 6.3.4, by possibly passing to subsequences we can assume that each \((P_{X_i}^{(n)})_{i=1}^l\) admits a coupling \(P_{X}^{(n)}\), such that \(P_{X}^{(n)} \rightarrow P_{X}^*\) weakly as \(n \to \infty\). In the non-degenerate case the output differential entropy is weakly continuous in the input distribution under the covariance constraint
(see for example [92, Proposition 18]), which establishes that

\[
\inf_{P_X : s^*_n P_X = P^{(n)}_{X_i}} \sum_j c_j D(T^*_j P_X \| \mu_j) = \sum_j c_j D(P^{(n)}_{Y_j} \| \mu_j) = \lim_{n \to \infty} \sum_j c_j D(P^{(n)}_{Y_j} \| \mu_j) \geq \lim_{n \to \infty} \inf_{P_X : s^*_n P_X = P^{(n)}_{X_i}} \sum_j c_j D(T^*_j P_X \| \mu_j)
\]

Thus (6.78) and (6.82) show that \(P^{*}_{X_i}\) is in fact a maximizer.

\[\text{Lemma 6.3.4} \text{ Suppose that for each } i = 1, \ldots, l \ (l \geq 2), P_{X_i} \text{ is a Borel measure on } \mathbb{R} \text{ and } P^{(n)}_{X_i} \text{ converges weakly to } P_{X_i} \text{ as } n \to \infty. \text{ If } P_X \text{ is a coupling of } (P_{X_i})_{1 \leq i \leq l}, \text{ then, upon extraction of a subsequence, there exist couplings } P^{(n)}_X \text{ for } (P^{(n)}_{X_i})_{1 \leq i \leq l} \text{ which converge weakly to } P_X \text{ as } n \to \infty.\]

\[\text{Remark 6.3.1} \text{ Lemma 6.3.4 will be used to show that the infimum of an upper semicontinuous (w.r.t. the joint distribution) functional over couplings is also upper semicontinuous (w.r.t the marginal distributions), which will be the key to the proof of Proposition 6.3.3. Another application is to prove the weak continuity of the optimal transport cost (the lower semicontinuity part being trivial) in the theory of optimal transportation, which may also be proved using the “stability of optimal transport” in [57, Theorem 5.20]. However, we note that the approach in [57, Theorem 5.20] relies on cyclic monotonicity, and hence cannot be extended to the setting of general upper semicontinuous functionals such as in Corollary 6.3.5.}\]

\[\text{Corollary 6.3.5 (Weak stability of optimal coupling) In the case of non-degenerate } (P_{Y_j | X}), \text{ suppose for each } n \geq 1, P^{(n)}_{X_i} \text{ is a Borel measure on } \mathbb{R}, i = 1, \ldots, l, \text{ whose second moment is bounded by } \sigma_i^2 < \infty. \text{ Assume that } P^{(n)}_X \text{ is a coupling of } (P^{(n)}_{X_i}) \text{ that minimizes } \sum_{j=1}^l c_j D(P^{(n)}_{Y_j} \| \mu_j). \text{ If } P^{(n)}_X \text{ converges weakly to some } P^*_X, \text{ then } P^*_X \text{ minimizes } \sum_{j=1}^l c_j D(P^{*}_{Y_j} \| \mu_j) \text{ given the marginals } (P^{*}_{X_i}).\]
Proof Lemma 6.3.4 and the fact that the differential entropy of the output of a non-degenerate Gaussian channel is weakly semicontinuous with respect to the input distribution under moment constraint (see e.g. [92, Proposition 18], [204, Theorem 7], or [205, Theorem 1, Theorem 2]) imply that there exists a strictly increasing sequence of positive integers $(n_k)_{k \geq 1}$ such that

$$\min_{P_X: S^*_P_X = P_{X_i}} \sum_j c_j D(T_j^*_P X \| \mu_j) \geq \limsup_{k \to \infty} \min_{P_X: S^*_P_X = P_{X_i}^{(n_k)}} \sum_j c_j D(T_j^*_P X \| \mu_j).$$

(6.83)

On the other hand, the assumption on the convergence of the optimal couplings and Lemma 6.5.2 imply that

$$\sum_j c_j D(T_j^*_P X \| \mu_j) \leq \liminf_{k \to \infty} \min_{P_X: S^*_P_X = P_{X_i}^{(n_k)}} \sum_j c_j D(T_j^*_P X \| \mu_j)$$

(6.84)

so the left side of (6.84) is no larger than the left side of (6.83).

Proof of Lemma 6.3.4 For each integer $k \geq 1$, define the random variable $W_i^{[k]} := \phi_k(X_i)$ where $\phi_k$ is the following “dyadic quantization function”:

$$\phi_k: x \in \mathbb{R} \mapsto \begin{cases} 2^k x & |x| \leq k, x \notin 2^{-k} \mathbb{Z}; \\ e & \text{otherwise}, \end{cases}$$

(6.85)

and let $W^{[k]} := (W_i^{[k]})_{i=1}^l$. Denote by $W^{[k]} := \{-k2^k, \ldots, k2^k - 1, e\}$ the alphabet of $W_i^{[k]}$.

For simplicity of the presentation, we shall assume that the set of “dyadic points” has measure zero:

$$P_{X_i} \left( \bigcup_{k=1}^{\infty} 2^{-k} \mathbb{Z} \right) = 0, \quad i = 1, \ldots, l.$$
This is not an essential restriction, since the property of $2^{-k}\mathbb{Z}$ used in the proof is that it is a countable dense subset of $\mathbb{R}$. In general, $P_{X_i}$ can have a positive mass only on countably many points, and in particular there exists a point of zero mass in any open interval. Thus, we can always choose a countable dense subset of $\mathbb{R}$ with $P_{X_i}$ measure zero ($i = 1$) and appropriately modify (6.85) with points in this set for quantization instead.

Since $P_{X_i}^{(n)} \to P_{X_i}$ weakly and the assumption in the preceding paragraph precluded any positive mass on the quantization boundaries under $P_{X_i}$, for each $k \geq 1$ there exists some $n := n_k$ large enough such that

$$P_{W_i^{[k]}}^{(n)}(w) \geq (1 - \frac{1}{k})P_{W_i^{[k]}}(w),$$

(6.87)

for each $i$ and $w \in W_i^{[k]}$. Now define a coupling $P_{W_i^{[k]}}^{(n)}$ compatible with the $\left(P_{W_i^{[k]}}^{(n)}\right)_{i=1}^{l}$ induced by $\left(P_{X_i}^{(n)}\right)_{i=1}^{l}$, as follows:

$$P_{W_i^{[k]}}^{(n)} := (1 - \frac{1}{k})P_{W_i^{[k]}} + k^{l-1} \prod_{i=1}^{l} \left(P_{W_i^{[k]}}^{(n)} - (1 - \frac{1}{k})P_{W_i^{[k]}}\right).$$

(6.88)

Observe that this is a well-defined probability measure because of (6.87), and indeed has $\left(P_{W_i^{[k]}}^{(n)}\right)_{i=1}^{l}$ as the marginals. Moreover, by triangle inequality we have the following bound on the total variation distance

$$\left|P_{W_i^{[k]}}^{(n)} - P_{W_i^{[k]}}\right| \leq \frac{2}{k}.$$  

(6.89)
Next, construct\(^8\) \(P^n_x\):

\[
P^n_x := \sum_{w^l \in \mathcal{W}[k]} \frac{P^n_{\mathcal{W}[k]}(w^l)}{\prod_{i=1}^l P^n_{\mathcal{W}[i]}(w_i)} \prod_{i=1}^l P^n_{X_i | \phi^{-1}_x(w_i)}. \tag{6.90}
\]

Observe that the \(P^n_x\) so defined is compatible with the \(P^n_{\mathcal{W}[k]}\) defined in (6.88), and indeed has \((P^n_x)_i = \sum_{w^l \in \mathcal{W}[k]} \frac{P^n_{\mathcal{W}[k]}(w^l)}{\prod_{i=1}^l P^n_{\mathcal{W}[i]}(w_i)} \prod_{i=1}^l P^n_{X_i | \phi^{-1}_x(w_i)}\) the marginals. Since \(n := n_k\) can be made increasing in \(k\), we have constructed the desired sequence \((P^n_{X_k})_{k=1}^\infty\) converging weakly to \(P_X\). Indeed, for any bounded open dyadic cube\(^9\) \(A\), using (6.89) and the assumption (6.86), we conclude

\[
\liminf_{k \to \infty} P^n_{X_k}(A) \geq P_X(A). \tag{6.91}
\]

Moreover, since bounded open dyadic cubes form a countable basis of the topology in \(\mathbb{R}^l\), we see (6.91) actually holds for any open set \(A\) (by writing \(A\) as a countable union of dyadic cubes, using the continuity of measure to pass to a finite disjoint union, and then apply (6.91)), as desired. \(\blacksquare\)

Next, we need the following tensorization result:

**Lemma 6.3.6** Fix \((P_{X_i^{(1)}}), (P_{X_i^{(2)}}), (\mu_j), (T_j), (c_j) \in [0, \infty)^m\), and let \(S_j\) be induced by coordinate projections. Then

\[
\inf_{P_{X_i^{(1)}} : S_i \otimes P_{X_i^{(1)}} = P_{X_i^{(1)} \times X_i^{(2)}}} \sum_{j=1}^m c_j D(P_{Y_j^{(1,2)}} || \mu_j) = \sum_{t=1,2} \sum_{j=1}^m c_j \inf_{P_{X_i^{(t)}} : S_i \otimes P_{X_i^{(t)}} = P_{X_i^{(t)}}} D(P_{Y_j^{(t)}} || \mu_j) \tag{6.92}
\]

\(^8\)We use \(P|_A\) to denote the restriction of a probability measure \(P\) on measurable set \(A\), that is, \(P|_A(B) := P(A \cap B)\) for any measurable \(B\).

\(^9\)That is, a cube whose corners have coordinates being multiples of \(2^{-k}\) where \(k\) is some integer.
where for each $j$,

$$
P_{Y_j^{(1,2)}} := T_j^* \otimes^2 P_{X^{(1,2)}} \quad (6.93)
$$
on the left side and

$$
P_{Y_j^{(t)}} := T_j^* \otimes^2 P_{X^{(t)}} \quad (6.94)
$$
on the right side, $t = 1, 2$.

**Proof**  We only need to prove the nontrivial $\geq$ part. For any $P_{X^{(1,2)}}$ on the left side, choose $P_{X^{(t)}}$ on the right side by marginalization. Then

$$
D(P_{Y_j^{(1,2)}} \| \mu_j^{(2)}) - \sum_t D(P_{Y_j^{(t)}} \| \mu_j) = I(Y_j^{(1)}; Y_j^{(2)}) \quad (6.95)
$$

$$
\geq 0 \quad (6.96)
$$

for each $j$.

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

**Proof of Theorem 6.3.1**

1. Assume that $(P_{X_i^{(1)}})$ and $(P_{X_i^{(2)}})$ are maximizers of $F_0$ (possibly equal). Let $P_{X_i^{1,2}} := P_{X_i^{(1)}} \times P_{X_i^{(2)}}$. Define

$$
X^+ := \frac{1}{\sqrt{2}} (X^{(1)} + X^{(2)}) ; \quad (6.97)
$$

$$
X^- := \frac{1}{\sqrt{2}} (X^{(1)} - X^{(2)}) . \quad (6.98)
$$

Define $(Y_j^+)$ and $(Y_j^-)$ analogously. Then $Y_j^+|\{X^+ = x^+, X^- = x^-\} \sim Q_{Y_j|X=x^+}$ is independent of $x^-$ and $Y_j^-|\{X^+ = x^+, X^- = x^-\} \sim Q_{Y_j|X=x^-}$ is independent of $x^+$. 

255
2. Next we perform the same algebraic expansion as in the proof of tensorization:

$$
\sum_{t=1}^{2} F_0\left( P_{X_t^{(t)}} \right) = \inf_{P_{X(1,2)} : S_{X(1,2)}^{\otimes 2} P_{X(1,2)} = P_{X_j(1,2)}^{\otimes 2}} \sum_{j} c_j D(P_{Y_j(1,2)} \| \mu_j^{\otimes 2}) - \sum_{i} b_i D(P_{X_i(1,2)} \| \nu_i^{\otimes 2})
$$

(6.99)

$$
= \inf_{P_{X^+X^-} : S_{X^+X^-}^{\otimes 2} P_{X^+X^-} = P_{X_j^+X_j^-}^{\otimes 2}} \sum_{j} c_j D(P_{Y_j^+X_j^-} \| \mu_j^{\otimes 2}) - \sum_{i} b_i D(P_{X_i^+X_i^-} \| \nu_i^{\otimes 2})
$$

(6.100)

$$
\leq \inf_{P_{X^+X^-} : S_{X^+X^-}^{\otimes 2} P_{X^+X^-} = P_{X_j^+X_j^-}^{\otimes 2}} \sum_{j} c_j \left[ D(P_{Y_j^+} \| \mu_j) + D(P_{Y_j^-X_j^+} \| \nu_j|P_{X_j^+}) \right]
- \sum_{i} b_i \left[ D(P_{X_i^+} \| \nu_i) + D(P_{X_i^-|X_i^+} \| \nu_i|P_{X_i^+}) \right]
$$

(6.101)

$$
\leq \sum_{j} c_j \left[ D(P_{Y_j^+}^* \| \mu_j) + D(P_{Y_j^-}^{*}|X_j^+ \| \mu_j|P_{X_j^+}) \right]
- \sum_{i} b_i \left[ D(P_{X_i^+}^* \| \nu_i) + D(P_{X_i^-|X_i^+}^* \| \nu_i|P_{X_i^+}^*) \right]
$$

(6.102)

$$
= F_0\left( P_{X_i^+}^* \right) + \int F_0\left( P_{X_i^-|X_i^+}^* \right) dP_{X_i^+}^*
$$

(6.103)

$$
\leq \sum_{t=1}^{2} F_0\left( P_{X_t^{(t)}} \right)
$$

(6.104)

where

- [6.99] uses Lemma 6.3.6
- [6.101] is because of the Markov chain $Y_j^+ - X^+ - Y_j^-$ (for any coupling).
- In [6.102] we selected a particular instance of coupling $P_{X^+X^-}$, constructed as follows: first we select an optimal coupling $P_{X^+}$ for given marginals $(P_{X_i^+})$. Then, for any $x^+ = (x_i^+)^{t}_{i=1}$, let $P_{X^-|X^+=x^+}$ be an optimal coupling of $(P_{X_i^-|X_i^+=x_i^+})$. With this construction, it is apparent that $X_i^+ - X^+ -$

---

Here we need to justify that we can select optimal coupling $P_{X^-|X^+=x^+}$ in a way that $P_{X^-|X^+}$ is indeed a regular conditional probability distribution, or equivalently, $P_{X^-|X^+=x^+}$ is Borel measurable, where we endow the weak topology on the space of probability measures. This is justified by two observations: (a) $x^+ \mapsto (P_{X_i^-|X_i^+=x_i^+})$ is Borel measurable; (b) the marginalization map
\[ X_i^- \] and hence
\[ D(P_{X_i^-|X_i^+}\|\nu_i|P_{X_i^+}) = D(P_{X_i^-|X_i^+}\|\nu_i|P_{X_i^+}). \]  (6.105)

- (6.103) is because in the above we have constructed the coupling optimally.
- (6.104) is because \( (P_X^{(t)}) \) maximizes \( F_0, t = 1, 2. \)

3. Thus in the expansions above, equalities are attained throughout. Using the differentiation technique as in the case of forward inequality, for almost all \((b_i), (c_j)\), we have
\[ D(P_{X_i^-|X_i^+}\|\nu_i|P_{X_i^+}) = D(P_{X_i^+}\|\nu_i) \]
\[ = D(P_{X_i^-}\|\nu_i), \quad \forall i \]  (6.107)

where the last equality is because by symmetry we can perform the algebraic expansions in a different way to show that \( P_{X_i^-} \) is also a maximizer of \( F_0 \).

Then \( I(X_i^+; X_i^-) = 0 \), which, combined with \( I(X_i^{(1)}; X_i^{(2)}) \), shows that \( X_i^{(1)} \) and \( X_i^{(2)} \) are Gaussian with the same covariance. Lastly, using Lemma 6.3.6 and the doubling trick one can show that the optimal coupling is also Gaussian.

\( \phi: \mathcal{O} \rightarrow \prod_i \mathcal{P}(X_i), P_X \rightarrow (P_{X_i}) \) admits a Borel right inverse, where \( \mathcal{O} \) is the set of optimal couplings (w.r.t. the relative entropy functional) whose marginals satisfy the second moment constraint. Part (a) is equivalent to the regularity of \( P_{X_i^-|X_i^+} \), and the existence of such a regular conditional distribution is guaranteed for joint distributions on Polish spaces (whose measurable space structure is isomorphic to a standard Borel space); see e.g. [43]. Part (b) is justified by measurable selection theorems (see e.g. references in [57, Corollary 5.22]). In particular, a similar argument as [57, Corollary 5.22] can also be applied here, since Corollary 6.3.5 implies that \( \mathcal{O} \) is closed and the pre-images \( \phi^{-1}(P_{X_i}) \) are all compact, and Proposition 6.3.3.1 justifies that \( \phi \) is onto.
6.4 Consequences of the Gaussian Optimality

In this section, we demonstrate several implications of the entropic Gaussian optimality results in Sections 6.2-6.3 to functional inequalities, transportation-cost inequalities, and network information theory.

6.4.1 Brascamp-Lieb Inequality with Gaussian Random transformations: an Information-Theoretic Proof

We give a simple proof of an extension of the Brascamp-Lieb inequality using Corollary 6.2.6. That is, we give an information-theoretic proof of the following result:

**Theorem 6.4.1** Suppose $\mu$ is either a Gaussian measure or the Lebesgue measure on $\mathcal{X} = \mathbb{R}^n$. Let $Q_{Y_j|X}$ be Gaussian random transformations where $Y_j = \mathbb{R}^{n_j}$, and $c_j \in (0, \infty)$ for $j \in \{1, \ldots, m\}$. For any non-negative measurable functions $f_j: Y_j \rightarrow \mathbb{R}$, $j \in \{1, \ldots, m\}$, define

$$
H(f_1, \ldots, f_m) = \log \int \exp \left( \sum_{j=1}^{m} \mathbb{E} \left[ \log f_j(Y_j) | X = x \right] \right) \mu(x) - \log \prod_{j=1}^{m} \| f_j \|_{\frac{1}{c_j}} \tag{6.108}
$$

where the norm $\| f_j \|_{\frac{1}{c_j}}$ is with respect to the Lebesgue measure and the expectation is with respect to $Q_{Y_j|X=x}$. Then

$$
\sup_{f_1, \ldots, f_m} H(f_1, \ldots, f_m) = \sup_{\text{Gaussian } f_1, \ldots, f_m} H(f_1, \ldots, f_m). \tag{6.109}
$$

Moreover, if $\mu$ is Gaussian, $(Q_{Y_j|X})$ is non-degenerate in the sense of Definition 6.2.2 then (6.109) is finite and uniquely attained by a set of Gaussian functions.

**Proof** We only prove the case of Gaussian $\mu$, since the proof for the Lebesgue case is similar. By the translation and scaling invariance, it suffices to consider the case where $\mu$ is centered Gaussian with density $\exp(-x^T M x)$ for some $w \in \mathbb{R}$ and $M \succeq 0$. 258
Then

\[ \iota_{\lambda|\mu}(x) = x^\top Mx \]  

(6.110)

where \( \lambda \) denotes the Lebesgue measure on \( \mathbb{R}^n \). Therefore,

\[
D(P_X||\mu) + \sum_{j=1}^{m} c_j h(P_{Y_j}) = -h(P_X) + \sum_{j=1}^{m} c_j h(P_{Y_j}) + \mathbb{E}[\iota_{\lambda|\mu}(\hat{X})] 
\]  

(6.111)

\[
= -h(P_X) + \sum_{j=1}^{m} c_j h(P_{Y_j}) + \mathbb{E}[\hat{X}^\top \hat{M} \hat{X}]. 
\]  

(6.112)

where \( \hat{X} \sim P_X \). Then (6.109) is established by

\[
\sup_{f_1, \ldots, f_m} H(f_1, \ldots, f_m) = - \inf_{P_X} \left\{ D(P_X||\mu) + \sum_{j=1}^{m} c_j h(P_{Y_j}) \right\} 
\]  

(6.113)

\[
= - \inf_{\text{Gaussian } P_X} \left\{ D(P_X||\mu) + \sum_{j=1}^{m} c_j h(P_{Y_j}) \right\} 
\]  

(6.114)

\[
= \sup_{\text{Gaussian } f_1, \ldots, f_m} H(f_1, \ldots, f_m) 
\]  

(6.115)

where

- (6.113) is from Remark 2.2.3
- (6.114) is from (6.112) and Corollary 6.2.6
- (6.115) is essentially a “Gaussian version” of the equivalence of the two inequalities in Theorem 2.2.3, which is easily shown with the same steps in the proof of Theorem 2.2.3, noting that (2.17) sends Gaussian distributions to Gaussian functions and (2.24) sends Gaussian functions to Gaussian distributions.

In the case of Gaussian \( \mu \) and non-degenerate \( (Q_{Y_j}|X) \), Corollary 6.2.6 implies that (6.114) is finitely attained by a unique Gaussian \( P_X \). By Remark 2.2.3 (6.109) is finitely attained by a unique set of Gaussian functions.
Remark 6.4.1 The proof of the Brascamp-Lieb inequality by Carlen and Erasquin [70] also relies on dual information-theoretic formulation. However, their proof of the differential entropy inequality uses a different approach based on superadditivity of Fisher information. That approach applies to the case where each $Q_{Y_j|X}$ is deterministic rank-one linear map, and it requires the problem to be first reduced to a special case called the geometric Brascamp-Lieb inequality (proposed by K. Ball [206]).

6.4.2 Multi-variate Gaussian Hypercontractivity

In this section we show a multivariate extension of Gaussian hypercontractivity. An $m$-tuple of random variables $(X_1, \ldots, X_m) \sim Q_X^m$ is said to be $(p_1, \ldots, p_m)$-hypercontractive for $p_l \in [1, \infty]$, $l \in \{1, \ldots, m\}$ if

$$
\mathbb{E} \left[ \prod_{l=1}^m f_l(X_l) \right] \leq \prod_{l=1}^m \|f_l(X_l)\|_{p_l}
$$

(6.116)

for all bounded real-valued measurable functions $f_l$ defined on $X_l$, $l \in \{1, \ldots, m\}$. Define the hypercontractivity region\footnote{Note that this definition is similar to the hypercontractivity ribbon defined in \cite{9} but without taking the H"{o}lder conjugate of one of the two exponent, which is more symmetrical and convenient in the multivariate setting.}

$$
\mathcal{R}(Q_X^m) := \{(p_1, \ldots, p_m) \in \mathbb{R}_+^m : (X_1, \ldots, X_m) \text{ is } (p_1, \ldots, p_m) \text{ hypercontractive}\}.
$$

(6.117)

By Theorem 2.2.3, the inequality (6.116) is true if and only if

$$
D(P_X^m||Q_X^m) \geq \sum_{j=1}^m \frac{1}{p_j} D(P_{X_j}||Q_{X_j})
$$

(6.118)
holds for any $P_{X^m} \ll Q_{X^m}$. In fact, [9] showed that (6.118) is equivalent to the following (which hinges on the fact that the constant term $d = 0$ in (6.118)):

$$I(U; X^m) \geq \sum_{j=1}^{m} \frac{1}{p_j} I(U; X_j), \quad \forall U.$$  \hspace{1cm} (6.119)

In the case of Gaussian $Q_{X^m}$, by Theorem [6.2.5] the inequality (6.119) holds if it holds for all $U$ jointly Gaussian with $X^m$ and having dimension at most $m$. When restricted to such $U$, (6.118) becomes an inequality involving the covariance matrices, and some elementary computations show that:

**Proposition 6.4.2** Suppose $Q_{X^m} = \mathcal{N}(0, \Sigma)$ where $\Sigma$ is a positive semidefinite matrix whose diagonal values are all 1. Then $p^m \in \mathcal{R}(Q_{X^m})$ if and only if

$$P \geq \Sigma$$  \hspace{1cm} (6.120)

where $P$ is a diagonal matrix with $p_j$, $j \in \{1, \ldots, m\}$ as its diagonal entries.

**Proof** Let $A$ be the covariance matrix of $X^m$ conditioned on $U$, and put $C := P^{-1}$. We will use the lowercase letters such as $a_1, \ldots, a_m$ to denote the diagonal entries of the corresponding matrices. Then in view of (6.119), we see that the goal is to show that (6.120) is a necessary and sufficient condition for

$$\log \frac{|\Sigma|}{|A|} \geq \sum_j c_j \log \frac{1}{a_i}, \quad \forall 0 \leq A \leq \Sigma.$$  \hspace{1cm} (6.121)

Let us first assume that $\Sigma$ is invertible. Define $D := \Sigma - A$, then rewrites as

$$|I - \Sigma^{-1} D| \leq \prod_j (1 - d_j)^{c_j}, \quad \forall 0 \leq D \leq \Sigma.$$  \hspace{1cm} (6.122)
If (6.120), then $C \preceq \Sigma^{-1}$, and we have

$$|I - \Sigma^{-1}D| \leq |I - CD|$$  \hspace{1cm} (6.123)

$$\leq \prod_j (1 - c_j d_j)$$  \hspace{1cm} (6.124)

$$\leq \prod_j (1 - d_j)^{c_j}$$  \hspace{1cm} (6.125)

where (6.124) is because \{1\( - c_j d_j\} is the diagonal entries of $I - CD$, which is majorized by the eigenvalues of that matrix. Inequality (6.125) is because $C \preceq \Sigma^{-1}$ and $D \preceq \Sigma$ imply the nonnegativity of $1 - c_j d_j$.

Conversely, if (6.122) holds, we can apply Taylor expansions on both sides of (6.122):

$$1 - \text{Tr}(\Sigma^{-1}D) + o(\|D\|) \leq 1 - \sum_j c_j d_j + o(\|D\|),$$  \hspace{1cm} (6.126)

where $\|D\|$ can be chosen as, say, the trace norm. Comparing the first order terms, we see that

$$\text{Tr}((\Sigma^{-1} - C)D) \geq 0$$  \hspace{1cm} (6.127)

must hold for any $0 \leq D \leq \Sigma$. Therefore (6.120) holds.

More generally, if $\Sigma$ is not necessarily invertible, we can consider the inverse of its restriction $\Sigma|_V$ where $V$ is its column space. Then the above arguments still carry
through with trivial modifications, and we can show that (6.121) holds if and only if

\[ \Sigma_{V^{-1}} \geq C_{V} \]
\[ \iff \Sigma_{V^{1/2}C_{V}} \Sigma_{V^{1/2}} \preceq I \] (6.128)
\[ \iff \Sigma^{1/2}C^{1/2} = \Sigma^{1/2}P^{-1/2}\Sigma^{1/2} \preceq I \] (6.130)
\[ \iff P^{-1/2}\Sigma P^{-1/2} \preceq I \] (6.131)
\[ \iff \Sigma \preceq P \] (6.132)

where (6.131) is because \( \Sigma^{1/2}P^{-1/2}\Sigma^{1/2} \) and \( P^{-1/2}\Sigma P^{-1/2} \) have the same set of eigenvalues.

\[ \blacksquare \]

**Remark 6.4.2** When \( m = 2 \), Proposition 6.4.2 reduces to Nelson’s hypercontractivity theorem for a pair of Gaussian scalar random variables \( X_1 \) and \( X_2 \), that is, \((p_1, p_2) \in \mathcal{R}(Q_{X^2}) \) if and only if

\[ (p_1 - 1)(p_2 - 1) \geq \rho^2(X_1; X_2), \]

where \( \rho^2 \) denotes the squared Pearson correlation.

### 6.4.3 \( T_2 \) Inequality for Gaussian Measures

Consider the Euclidean space endowed with \( \ell_2 \)-norm \((\mathbb{R}^n, \| \cdot \|_2)\). Talagrand [207] showed that the standard Gaussian measure \( Q = \mathcal{N}(0, I_n) \) satisfies the \( T_2(1) \) inequality (see Definition 2.5.1). Below we give a new proof using the Gaussian optimality in the forward-reverse Brascamp-Lieb inequality. By the tensorization of \( T_2 \) inequality [137], it suffices to prove the \( n = 1 \) case. By continuity, it suffices to prove \( T_2(\lambda) \) for any \( \lambda > 1 \). Moreover, as one can readily check, when \( \lambda > 1 \) and \( Q = \mathcal{N}(0, 1) \), (2.168) is satisfied for all Gaussian \( P \) (in which case the optimal coupling in (2.168) is also Gaussian), so it suffices to prove Gaussian extremisability in (2.168).
Let $\lambda \in (1, +\infty)$, $b_2 \in (0, +\infty)$, and

$$F_0(P_{X_1}, P_{X_2}) := \inf_{P_{X_1} \times P_{X_2}} \mathbb{E}[|X_1 - X_2|^2] - 2\lambda D(P_{X_1} \| Q) - b_2 D(P_{X_2} \| Q), \quad (6.134)$$

where the infimum is over coupling $P_{X_1, X_2}$ of $P_{X_1}$ and $P_{X_2}$. Using the rotation invariance argument/doubling trick in the proof of Theorem 6.3.1, we can show that if $(P_{X_1}, P_{X_2})$ maximizes $F_0(P_{X_1}, P_{X_2})$, then they must be Gaussian, and the optimal coupling $P_{X_1, X_2}$ is also Gaussian\(^{12}\). Letting $b_2 \to \infty$, we see the Gaussian optimality in the $T_2$ inequality. To make the above argument rigorous, we need to take care of two technical issues:

(a) We approximated the functional (2.175) with $b_2 D(P_{X_2} \| Q)$ and let $b_2 \to \infty$, but we want that Gaussian optimality continue to hold when the last term in (6.134) is exactly (2.175).

(b) We want to show the existence of a maximizer $(P_{X_1}, P_{X_2})$ for (6.134).

While it might be possible to provide a formal justification of the limit argument in (a), a slicker way is to circumvent it by directly working with (2.175) instead of $b_2 D(P_{X_2} \| Q)$ with $b_2 \to \infty$. From the tensorization of the functional (6.134), it is relatively easy to distill the tensorization of the $T_2$ inequality (see also [137] for a direct proof of tensorization of $T_2$ inequality), and then use the rotation invariance argument/doubling trick to conclude Gaussian optimality.

\(^{12}\)In Theorem 6.3.1 the first term of the objective function is the infimum of the relative entropy, rather than the infimum of the expectation of a quadratic cost function. However, the argument in Theorem 6.3.1 also works in the latter case, since the expectation functional has a similar tensorization property, and the quadratic cost function also has a rotational invariance property.
As for (b), note that if \( b_2 D(P_{X_2} \| Q) \) is replaced with \( 2.175 \), we want to show the existence of a maximizer \((P_{X_1}, Q)\) for \( 6.134 \). If \( \sigma^2 := \mathbb{E}[|X_1|^2] \), then

\[
\inf_{P_{X_1}X_2} \mathbb{E}[|X_1 - X_2|^2] \leq \sigma^2 + \mathbb{E}[|X_2|^2]
\]

\[
= \sigma^2 + 1. \quad (6.136)
\]

On the other hand,

\[
D(P_{X_1} \| Q) = D(P_{X_1} \| \mathcal{N}(0, \sigma^2)) + \mathbb{E}[\mathcal{I}_{\mathcal{N}(0, \sigma^2)} \| Q](X_1) = \frac{\sigma^2}{2} + o(\sigma^2). \quad (6.137)
\]

Therefore, if \( \lambda > 1 \) and \( (P^{(t)}_{X_1})_{t=1}^{\infty} \) is a supremizing sequence, then \( (P^{(t)}_{X_1})_{t=1}^{\infty} \) must have bounded second moment, hence must be tight. Thus Prokhorov Theorem implies the existence of a subsequence weakly converging to some \( P^{\ast}_{X_1} \), and a semicontinuity argument similar to Proposition \( 6.3.3 \) shows that \( P^{\ast}_{X_1} \) is in fact a maximizer.

### 6.4.4 Wyner’s Common Information for \( m \) Dependent Random Variables

Wyner’s common information for \( m \) dependent random variables \( X_1, \ldots, X_m \) is commonly defined as \( 208 \)

\[
C(X^m) := \inf_{U} I(U; X^m)
\]

where the infimum is over \( P_{U \| X^m} \) such that \( X_1, \ldots, X_m \) are independent conditioned on \( U \). Previously, to the best of our knowledge, the common information for \( m \) Gaussian scalars \( X_1, \ldots, X_m \) could only be obtained in the special case where the correlation coefficient between \( X_i \) and \( X_j \) are equal for all \( 1 \leq i, j \leq m \) \( 208, \) Corollary 1\] via a different approach.

265
Using Theorem 6.2.5 and setting $X \leftarrow X^m$ and $Y_j \leftarrow X_j$, we immediately obtain the following characterization of the multivariate common information for Gaussian sources:

**Theorem 6.4.3** The common information of $m$ Gaussian scalar random variables $X_1, \ldots, X_m$ with covariance matrix $\Sigma > 0$ is given by

$$C(X^m) = \frac{1}{2} \inf_{\Lambda} \log \frac{|\Sigma|}{|\Lambda|}$$  \hspace{1cm} (6.139)

where the infimum is over all diagonal matrices $\Lambda$ satisfying $\Lambda \preceq \Sigma$. More generally when $\Sigma$ is not necessarily invertible, then $\Sigma$ and $\Lambda$ in (6.139) should be replaced by the restrictions $\Sigma|_V$ and $\Lambda|_V$ to the column space $V$ of $\Sigma$.

**Remark 6.4.3** After we completed a draft of the journal version of this work [93], Jun Chen and Chandra Nair independently found and showed us a simple proof of Theorem 6.4.3 without invoking Theorem 6.2.5 using the fact that Gaussian distribution maximizes differential entropy given a covariance constraint and the concavity of the log-determinant function,

$$I(U; X^m) = h(X^m) - h(X^m|U) \geq \frac{1}{2} \log |\Sigma| - \mathbb{E} \left[ \frac{1}{2} \log |\text{Cov}(X|U)| \right]$$ \hspace{1cm} (6.140)

$$\geq \frac{1}{2} \log |\Sigma| - \frac{1}{2} \log |\Sigma_{X|U}|.$$ \hspace{1cm} (6.141)

Since $\Sigma_{X|U} \leq \Sigma$ and $\Sigma_{X|U}$ is diagonal, we establishes the nontrivial “$\geq$” part of (6.139) in the case of invertible $\Sigma$. This argument is also related to the proof of Theorem 6.2.5 in the sense that both convert a mutual information optimization problem (without a covariance constraint) to a conditional differential entropy optimization problem with a covariance constraint.
6.4.5 Key Generation with an Omniscient Helper

As an example of applications in the network information theory, we give a simple characterization of the achievable rate region for secret key generation with an omniscient helper [101] in the case of stationary memoryless Gaussian sources. This problem can be viewed as the omniscient CR generation problem in Chapter 2 with an additional secrecy constraint. For completeness we formula the problem below: Let $Q_{X^m}$ be the per-letter joint distribution of sources $X_1, \ldots, X_m$. As in Figure 6.1, the Terminals $T_1, \ldots, T_m$ observe i.i.d. realizations of $X_1, \ldots, X_m$, respectively, whereas the omniscient helper $T_0$ has access to all the sources. Suppose the terminals perform block coding with length $n$. The communicator computes the integers $W_1((X^m)^n), \ldots, W_m((X^m)^n)$ possibly stochastically and sends them to $T_1, \ldots, T_m$, respectively. Then, all the terminals calculate integers $K((X^m)^n), K_1((X_1)^n,W_1), \ldots, K_m((X_m)^n,W_m)$ possibly stochastically. The goal is to make $K = K_1 = \cdots = K_m$ with high probability and $K$ almost equiprobable and independent of each message $W_i$. In other words, we want to minimize the following quantities:

$$\epsilon_n = \max_{1 \leq l \leq m} \mathbb{P}[K \neq K_l], \quad (6.143)$$

$$\nu_n = \max_{1 \leq l \leq m} \{ \log |K| - H(K|W_l) \}. \quad (6.144)$$
Thus, the difference with the omniscient helper CR generation problem is that here $K$ needs to be (asymptotically) independent of $W_l$, $l = 1, \ldots, m$.

An $(m + 1)$-tuple $(R, R_1, \ldots, R_m)$ is said to be achievable if a sequence of key generation schemes can be designed to fulfill the following conditions:

\begin{align}
\liminf_{n \to \infty} \frac{1}{n} \log |K| &\geq R; \quad (6.145) \\
\limsup_{n \to \infty} \frac{1}{n} \log |W_l| &\leq R_l, \quad l \in \{1, \ldots, m\}; \quad (6.146) \\
\lim_{n \to \infty} \epsilon_n &= 0; \quad (6.147) \\
\lim_{n \to \infty} \nu_n &= 0. \quad (6.148)
\end{align}

Notice that a small $\nu_n$ does not imply that $K$ is nearly independent with all the messages $W^m$; the problem appears to be harder to solve if a the stronger requirement that

\begin{align}
\lim_{n \to \infty} (\log |K| - H(K|W^m)) = 0 \quad (6.149)
\end{align}

is imposed in place of (6.148).

**Theorem 6.4.4** [101] The set of achievable rates is the closure of

\[
\bigcup_{Q_U|X^m} \left\{ (R, R_1, \ldots, R_m) : \begin{array}{l}
R \leq \min\{I(U; X^m), H(X_1), \ldots, H(X_m)\}; \\
R_l \geq I(U; X^m) - I(U; X_l), \quad 1 \leq l \leq m
\end{array} \right\}. \quad (6.150)
\]

A priori, computing the rate region from (6.150) requires solving an optimization with possibly infinite dimensions. However, using Theorem 6.2.5 we easily see that the problem can be reduced to a matrix optimization in the case of Gaussian sources:
Theorem 6.4.5 For stationary memoryless sources where the per-letter joint dis-
btribution is jointly Gaussian with non-degenerate covariance matrix $\Sigma$, the achievable
region for omniscient helper key generation can be represented as the closure of

$$
\bigcup_{0 \leq \Sigma' \leq \Sigma} \left\{ (R, R_1, \ldots, R_m) : \\
R \leq \frac{1}{2} \log |\Sigma|; \\
R_l \geq \frac{1}{2} \log |\Sigma| - \frac{1}{2} \log |\Sigma_{ll}|, \quad 1 \leq l \leq m \right\}.
$$ (6.151)

As alluded, omniscient helper CR generation is a related problem where there is
no secrecy (independence) assumption. For CR generation, \[16, \text{Theorem 4.2}\] showed
that the achievable region is similar to that of Theorem 6.4.4 except that the first
bound in (6.150) needs to be replaced by

$$
R \leq I(U; X^m).
$$ (6.152)

Note that such a characterization is equivalent to the supporting hyperplane charac-
terization we gave in (2.40). In particular, for continuous sources which has infinite
Shannon entropy, the achievable region is the same for CR generation and key gen-
eration!

We remark that despite the superficial similarity of the achievable regions for CR
and key generation, the achievability part of Theorem 6.4.4 requires more sophisti-
cated coding technique to guarantee secrecy. We observe that Theorem 6.4.5 also
applies to CR generation from Gaussian sources, because in that case $H(X_j) = \infty$
and hence does not effectively change the bound on $R$.

As we introduced in Chapter 2 a generalization of the omniscient helper problem
is called the one communicator problem in \[101\], where in Figure 6.1 the terminal
$T_0$ does not see all the random variables $X^m$ but instead another random variable
$Z$ which can be arbitrarily correlated with $X^m$. The achievable rate region for one
communicator CR generation is known \[16, \text{Theorem 4.2}\] to be the closure of

$$
\bigcup_{Q_{U|Z}} \left\{ \begin{array}{l}
(R, R_1, \ldots, R_m) : \\
R \leq I(U; Z) ; \\
R_l \geq I(U; Z) - I(U; X_l), \quad 1 \leq l \leq m
\end{array} \right\}
$$

(6.153)

and obviously we can use Theorem 6.2.5 to reduce (6.153) to a matrix optimization problem in the case of Gaussian \((Z, X_1, \ldots, X_m)\). The achievable rate region is also derived for key generation with one communicator in [101]. But the expression of that region is more complicated involving \(m + 1\) auxiliary random variables, and it is not immediate to conclude that Gaussian auxiliary random variables suffice merely using Theorem 6.2.5.

### 6.5 Appendix: Existence of Maximizer in Theorem 6.2.1

**Proposition 6.5.1** In the non-degenerate case, for any \(\Sigma \succeq 0\),

$$
\phi(\Sigma) := \sup_{P_{U|X}: \Sigma_{X|U} \preceq \Sigma} F(P_{U|X})
$$

(6.154)

is finite and is attained by some \(P_{U|X}\) with \(|U| < \infty\).

**Proof** First, observe that if we let \(\tilde{\phi}(\cdot)\) be the supremum in (6.154) with the additional restriction that \(|U| < \infty\), then \(\tilde{\phi}(\cdot)\) is a concave function on a convex set of finite dimension. Hence Jensen’s inequality\(^{13}\) implies that \(\phi(\cdot) \leq \tilde{\phi}(\cdot)\), while \(\tilde{\phi}(\cdot) \leq \phi(\cdot)\) is obvious from the definition. Thus \(\tilde{\phi}(\cdot) = \phi(\cdot)\).

\(^{13}\)Luckily, \(\phi\) is defined on a finite dimensional set of matrices (rather than a possibly infinite dimensional set of distributions \(P_X\)). In the infinite dimensional case without further continuity assumptions, Jensen’s inequality can fail; see the example in [209] equation (1.3)].
The set
\[ C := \bigcup_{P_X} \{(F_0(P_X), \text{Cov}(X))\} \]  
(6.155)

lies in a linear space of dimension \( 1 + \frac{\dim(X)(\dim(X) + 1)}{2} \). By Carathéodory’s theorem \cite[Theorem 17.1]{94}, each point in the convex hull of \( C \) is a convex combination of at most \( D := 2 + \frac{\dim(X)(\dim(X) + 1)}{2} \) points in \( C \):

\[ \bigcup_{P_{UX} : \mathcal{U} \text{ finite}} \{(F(P_{UX}), \Sigma_{X|U})\} = \bigcup_{P_{UX} : |\mathcal{U}| \leq D} \{(F(P_{UX}), \Sigma_{X|U})\} \]  
(6.156)

hence

\[ \phi(\Sigma) = \sup_{P_{UX} : \mathcal{U} = \{1, \ldots, D\}, \Sigma_{X|U} \leq \Sigma} F(P_{UX}). \]  
(6.157)

Now suppose \( \{P_{U_nX_n}\}_{n \geq 1} \) is a sequence satisfying \( \Sigma_{X_n|U_n} \leq \Sigma, |\mathcal{U}_n| = \{1, \ldots, D\} \) for each \( n \), and

\[ \lim_{n \to \infty} F(P_{U_nX_n}) = (6.157). \]  
(6.158)

We can assume without loss of generality that \( P_{U_n} \) converges to some \( P_{U^*} \), since otherwise we can pass to one convergent subsequence instead. Moreover, by the translation invariance we can assume without loss of generality that

\[ \mathbb{E}[X_n|U_n = u] = 0 \]  
(6.159)

for each \( u \) and \( n \).
If \( u \in \{1, \ldots, D\} \) is such that \( P_{U^*}(u) > 0 \), then for \( n \) sufficiently large, we have 
\[
P_{U_n} > \frac{P_{U^*}(u)}{2}
\] and
\[
\text{Cov}(X_n | U_n = u) \leq \frac{2\Sigma}{P_{U^*}(u)}. \tag{6.160}
\]

Thus \( \{P_{X_n | U_n = u}\}_{n \geq 1} \) is a tight sequence of measures by Chebyshev’s inequality, and Prokhorov’s theorem \[109\] guarantees the existence of a subsequence of \( \{P_{X_n | U_n = u}\}_{n \geq 1} \) converging weakly to some Borel measure \( P_{X_n^*} \). We might as well assume that \( \{P_{X_n | U_n = u}\}_{n \geq 1} \) converges to \( P_{X_n^*} \) since otherwise we pass to a convergent subsequence instead. This argument can applied to each \( u \in \{1, \ldots, D\} \) satisfying \( P_{U^*}(u) > 0 \) iteratively, hence we can assume the existence of the weak limits
\[
\lim_{n \to \infty} P_{X_n | U_n = u} = P_{X_n^*} \tag{6.161}
\]
for all such \( u \). Next, we show that
\[
\limsup_{n \to \infty} F_0(P_{X_n | U_n = u}) \leq F_0(P_{X_n^*}) \tag{6.162}
\]
for all such \( u \). Using Lemma \[6.5.2\] below, we obtain
\[
\limsup_{n \to \infty} h(X_n | U_n = u) \leq h(X_n^*), \tag{6.163}
\]
Because of the moment constraint \( (6.160) \), the differential entropy of the output distribution, which is smoothed by the Gaussian kernel, enjoys weak continuity in the input distribution (see e.g. \[92\] Proposition 18], \[204\] Theorem 7], or \[205\] Theorem 1, Theorem 2]):
\[
\lim_{n \to \infty} h(Y_{j_n} | U_n = u) = h(Y_{j_n}^* | U^* = u) \text{ for each } u \in \{1, \ldots, D\} \tag{6.164}
\]
where \((U, X_n, Y_{jn}) \sim P_{X_n U_n} Q_{Y_j | X}\) and \((U^*, X^*, Y^*_j) \sim P_{X^* U^*} Q_{Y_j | X^*}\). As for the trace term, consider

\[
\liminf_{n \to \infty} \text{Tr}[M \text{Cov}(X_n | U_n = u)] = \liminf_{n \to \infty} \mathbb{E}[\text{Tr}[MX_n X_n^T] | U_n = u] = \liminf_{n \to \infty} \sup_{K > 0} \mathbb{E}[\text{Tr}[MX_n X_n^T] \land K | U_n = u] \geq \sup_{K > 0} \liminf_{n \to \infty} \mathbb{E}[\text{Tr}[MX_n X_n^T] \land K] \geq \sup_{K > 0} \mathbb{E}[\text{Tr}[MX_u X_u^T] \land K] = \mathbb{E}[\text{Tr}[MX_u X_u^T]]
\] (6.165)

where “\(\land\)” takes the minimum of two numbers, (6.166) and (6.169) are from monotone convergence theorem, and (6.168) uses the weak convergence (6.161). The proof of (6.162) is finished by combining (6.163) (6.164) and (6.169).

The final step deals with any \(u \in \{1, \ldots, D\}\) satisfying \(P_{U^*}(u) = 0\). The variance constraint implies that

\[
\text{Cov}(X_n | U_n = u) \leq \frac{1}{P_{U_n}(u)} \Sigma,
\] (6.170)

hence by the fact that Gaussian distribution maximizes the differential entropy under a covariance constraint, we have the bound

\[
h(X_n | U_n = u) \leq \frac{\dim(X)}{2} \log(2\pi) + \frac{1}{2} \log e + \frac{1}{2} \log \left| \frac{1}{P_{U_n}(u)} \Sigma \right| = \frac{\dim(X)}{2} \log(2\pi) + \frac{1}{2} \log e + \frac{1}{2} \log |\Sigma| + \frac{\dim(X)}{2} \log \frac{1}{P_{U_n}(u)}.
\] (6.171)

This combined with the fact that \(h(Y_{jn} | U = u) \geq h(Y_{jn} | X_n)\) is bounded below, \(j = 1, \ldots, m\) in the non-degenerate case, implies that if \(P_{U_n}(u)\) converges to zero,
then

\[
\limsup_{n \to \infty} P_{U_n}(u) F_0(P_{X_n|U_n=u}) \leq 0. \tag{6.173}
\]

Combining (6.162) and (6.173), we see

\[
F(P_{U^*X^*}) \geq \limsup_{n \to \infty} F(P_{U_n X_n}),
\tag{6.174}
\]

where \(P_{X^*|U^*=u} := P_{X^*_u}\) for each \(u = 1, \ldots, D\).

Lemma 6.5.2 Suppose \((P_{X_n})\) is a sequence of distributions on \(\mathbb{R}^d\) converging weakly to \(P_{X^*}\), and

\[
\mathbb{E}[X_n X'_n] \leq \Sigma \tag{6.175}
\]

for all \(n\). Then

\[
\limsup_{n \to \infty} h(X_n) \leq h(X^*). \tag{6.176}
\]

The result fails without the condition (6.175). Also, related results when the weak convergence is replaced with pointwise convergence of density functions and certain additional constraints was shown in [205, Theorem 1, Theorem 2] (see also the proof of [92, Theorem 5]). Those results are not applicable here since the density functions of \(X_n\) do not converge pointwise. They are applicable for the problems discussed in [92] because the density functions of the output of the Gaussian random transformation enjoy many nice properties due to the smoothing effect of the “good kernel”.

Proof It is well known that in metric spaces and for probability measures, the relative entropy is weakly lower semicontinuous (cf. [43]). This fact and a scaling
argument immediately show that, for any $r > 0$,

$$\limsup_{n \to \infty} h(X_n \| X_n \leq r) \leq h(X^\star \| X^\star \leq r). \quad (6.177)$$

Let $p_n(r) := \mathbb{P}[\|X_n\| > r]$, then (6.175) implies

$$\mathbb{E}[XX^\top \| X_n \| > r] \leq \frac{1}{p_n(r)} \Sigma. \quad (6.178)$$

Therefore, since the Gaussian distribution maximizes differential entropy given a second moment upper bound, we have

$$h(X_n \| X_n \leq r) \leq \frac{1}{2} \log \frac{(2\pi)^d e |\Sigma|}{p_n(r) p_n(r)}. \quad (6.179)$$

Since $\lim_{r \to \infty} \sup_n p_n(r) = 0$ by (6.175) and Chebyshev’s inequality, the above implies that

$$\limsup_{r \to \infty} p_n(r) h(X_n \| X_n \leq r) = 0. \quad (6.180)$$

The desired result follows from (6.177), (6.180) and the fact that

$$h(X_n) = p_n(r) h(X_n \| X_n \leq r) + (1 - p_n(r)) h(X_n \| X_n \leq r) + h(p_n(r)). \quad (6.181)$$
Chapter 7

Conclusion and Future Work

Let us recall the main idea of the thesis: information measures (or their linear combination) generally have two equivalent characterizations: the functional representation and the entropic representation. The existence of such equivalent characterizations can be understood in terms of the duality of convex optimizations, even though technically speaking the optimization problem is not always convex (consider the case where the linear combination of convex information measures may contain both positive and negative coefficients). As far as we have seen, the functional representation is very useful in proving non-asymptotic bounds of operational problems. Especially, the reverse hypercontractivity argument introduced in Chapter 4 appears to be a tailor-made for the functional representation. On the other hand, the entropic formulation appears more convenient in proving certain abstract properties such as data processing, tensorization and the Gaussian optimality, leveraging tools familiar to information theorists such as the chain rule.

Below, we mention two scenarios of common randomness generation where the extension of the proposed functional approach appears challenging. These scenarios are related to some of the author’s publications which have not been discussed in the main text. The strong converse counterpart of these results appears to be open.

• Rate limited interactive communication. In the literature, a very common protocol for common randomness generation allows terminals to communicate to
each other \cite{210}, in contrast to the one-communicator protocol we discussed (Figure 1.1) where only one terminal is allowed to send messages to others. We remark that without communication rate constraints, the maximum secret key rate has known single-letter expressions \cite{210,211} and second-order rate results \cite{55}. However when the communication rate is limited, we can only characterize the tradeoff between the communication rate and the secret key (or common randomness) rate using the same number of auxiliary random variables as the number of rounds of communications \cite{102}. Does the strong converse hold in this setting? The challenge appears to be that no corresponding functional inequality is known when two or more auxiliary random variables are involved. A related open problem is the source-channel network with two or more helpers: not only the strong converse but even the first-order rate region is generally unknown \cite{8}.

- **Demanding secrecy.** If we further impose that the common randomness generated in the one-communicator problem (Figure 1.1) is independent of the public communications (i.e. the secret key generation problem), can we still show a strong converse or even a second-order result? The secret key generation problem is formally related to the wiretap channel (see the remark in \cite{18}). We note that the second-order rate for wiretap channel has been open for non-degenerate channels \cite{212}, in which case the rate region involves an auxiliary random variable, and is very similar to the rate region of the one-communicator problem whose second-order rate we solved in Chapter 4.
Bibliography


287


