TECHNISCHE UNIVERSITÄT WIEN

institute of telecommunications

DISSERTATION

CLUSTERING BY MUTUAL INFORMATION

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der technischen Wissenschaften

unter der Leitung von
Ao.Univ.-Prof. Dr. Gerald Matz
Institute of Telecommunications

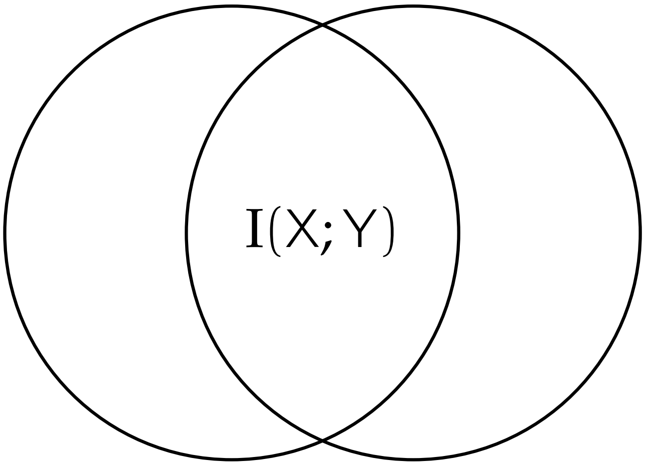
eingereicht an der Technischen Universität Wien
Fakultät für Elektrotechnik und Informationstechnik

von
Georg Pichler
Lambrechtgasse 16/8
1040 Wien

Wien, November 2017 ______________________

CLUSTERING BY MUTUAL INFORMATION

Georg Pichler



Die Beurteilung dieser Arbeit erfolgte durch:

- Prof. Pablo Piantanida
  Laboratoire des Signaux et Systèmes (L2S, UMR8506),
  CentraleSupélec-CNRS-Université Paris-Sud, Frankreich
- Prof. Gerhard Kramer
  Fakultät für Elektrotechnik und Informationstechnik
  Technische Universität München, Deutschland

Georg Pichler: *Clustering by Mutual Information,* November 2017

Dedicated to the memory of my grandfather *Wolfgang Kleibel*.

DECLARATION

I herewith declare that I have completed the present thesis independently, making use only of the specified literature and aids. This thesis in this form has not been submitted to an examination body and has not been published. This thesis draws, however, on previous publications by the author. For a complete list of the relevant scientific articles, I refer to page xi.

Vienna, November 2017

Georg Pichler

ABSTRACT

This thesis is concerned with multi-terminal source coding problems motivated by biclustering applications. We introduce the Shannon theoretic multi-clustering problem and investigate its properties, uncovering connections with many other coding problems in the literature. The figure of merit for this information-theoretic problem is mutual information, the mathematical properties of which make the multi-clustering problem amenable to techniques that could not be used in a general rate-distortion setting.

We first consider the case of two sources, where we derive single-letter bounds for the achievable region by connecting our setting to hypothesis testing and pattern recognition problems in the information theory literature. We complement these bounds with cardinality bounds for the auxiliary random variables, improving upon the results typically obtained by using the convex cover method. Applying these improved cardinality bounds to the case of a doubly symmetric binary source, we find a gap between the outer and inner bound, disproving a conjecture by Westover and O'Sullivan (2008).

We generalize the problem setup to an arbitrary number of sources and show that a CEO problem with logarithmic loss distortion, which was previously investigated by Courtade and Weissman (2014), constitutes a special case of this multi-clustering problem. This CEO problem can be extended by requiring multiple description coding. Drawing from the theory of submodular functions, we prove a tight inner and outer bound for the resulting achievable region under a suitable conditional independence assumption. The single-letter characterization of the achievable region we obtain has some interesting technical properties. In particular, the rate requirement is in general insufficient to ensure successful typicality decoding of the corresponding description.

Furthermore, we present a proof of the two-function case of a conjecture by Kumar and Courtade (2013), showing that the inequality $I\big(f(X^n); g(Y^n)\big) \leqslant I(X;Y)$ holds for any two Boolean functions f and g, where (X, Y) is a doubly symmetric binary source. We also show that the dictator functions are essentially the only functions achieving equality. The key step in the proof is a careful analysis of the Fourier spectrum of the two Boolean functions. This allows us to reduce the statement to an elementary inequality which we subsequently prove.

ZUSAMMENFASSUNG

Diese Arbeit befasst sich mit Problemen der verteilten Quellenkodierung, welche sich von Biclustering Methoden ableiten. Wir definieren dieses Shannon-theoretische verteilte Quellenkodierungsproblem, bei dem die Transinformation zwischen Codewörtern maximiert werden soll. Die mathematischen Eigenschaften der Transinformation machen dieses Problem für Techniken zugänglich, die auf allgemeine Rate-Distortion Probleme nicht anwendbar wären. Wir untersuchen die erreichbare Region und zeigen dabei Zusammenhänge mit etlichen anderen Kodierungsproblemen in der Literatur auf.

Zunächst beschränken wir uns auf zwei Quellen und finden Schranken für die erreichbare Region. Dabei stellen wir Zusammenhänge mit Problemen des Hypothesentestens und der Mustererkennung her. Wir ergänzen diese Resultate durch Abschätzungen für die Kardinalität der Hilfsvariablen, wobei wir die üblichen Abschätzungen verbessern, die durch Anwenden der Methode konvexer Überdeckungen erreicht werden. Die verbesserten Abschätzungen wenden im Fall einer doppelt symmetrischen binären Quelle an und stellen fest, dass die inneren und äußeren Schranken nicht übereinstimmen, was eine Vermutung von Westover und O'Sullivan (2008) widerlegt.

Wir verallgemeinern das Problem für eine beliebige Zahl an Quellen und zeigen, dass ein CEO-Problem mit logarithmischem Verlust als Verzerrung einen Spezialfall darstellt, welcher zuvor von Courtade und Weissman (2014) untersucht wurde. Dieses CEO-Problem kann erweitert werden, indem man mehrfache Beschreibungen berücksichtigt. Wir verwenden die Theorie submodularer Funktionen und finden eine single-letter Charakterisierung der erreichbaren Region für dieses CEO-Problem mit mehrfachen Beschreibungen unter geeigneten Annahmen bedingter Unabhängigkeit. Die resultierende Region hat einige interessante technische Eigenschaften. Insbesondere ist die benötigte Rate im Allgemeinen niedriger als für eine erfolgreiche typische Dekodierung notwendig.

Weiters präsentieren wir den Beweis einer Vermutung von Kumar und Courtade (2013) über die maximale Transinformation von zwei Boolschen Funktionen, die besagt, dass $I\big(f(X^n); g(Y^n)\big) \leqslant I(X; Y)$ für zwei beliebige Boolsche Funktionen gilt, wobei (X, Y) eine doppelt symmetrische binäre Quelle ist. Wir zeigen, dass, abgesehen von Spezialfällen, die Diktator-Funktionen die einzigen Boolschen Funktionen sind, die Gleichheit erreichen. Der Schlüsselschritt des Beweises ist eine detailierte Analyse des Fourier-Spektrums der Boolschen Funktionen, die es erlaubt, die Vermutung auf eine elementare Ungleichung zurückzuführen.

PUBLICATIONS

Material previously appeared in the following publications:

G. Pichler, G. Matz, and P. Piantanida, «A Tight Upper Bound on the Mutual Information of Two Boolean Functions,» in *Proc. Inform. Theory Workshop*, Cambridge, UK, Sep. 2016, pp. 16–20. DOI: 10.1109/ITW.2016.7606787.

G. Pichler, P. Piantanida, and G. Matz, «Distributed Information-Theoretic Biclustering of Two Memoryless Sources,» in *Proc. 53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 2015, pp. 426–433. DOI: 10.1109/ALLERTON.2015.7447035.

G. Pichler, P. Piantanida, and G. Matz, «Distributed Information-Theoretic Biclustering,» in *Proc. IEEE Int. Symp. on Inform. Theory*, Barcelona, Spain, Jul. 2016, pp. 1083–1087. DOI: 10.1109/ISIT.2016.7541466.

G. Pichler, P. Piantanida, and G. Matz, «A Multiple Description CEO Problem with Log-Loss Distortion,» in *Proc. IEEE Int. Symp. on Inform. Theory*, Aachen, Germany, Jun. 2017, pp. 111–115. DOI: 10.1109/ISIT.2017.8006500.

G. Pichler, P. Piantanida, and G. Matz, «Dictator Functions Maximize Mutual Information,» *Ann. of Applied Probability*, 2017, (submitted). [Online]. Available: http://arxiv.org/abs/1604.02109.

G. Pichler, P. Piantanida, and G. Matz, «Distributed Information-Theoretic Clustering,» *IEEE Trans. Inf. Theory*, 2017, (submitted). [Online]. Available: https://arxiv.org/abs/1602.04605.

ACKNOWLEDGMENTS

First and foremost I want to thank Gerald and Pablo for their outstanding support during these four years. Their advice and guidance made this research possible. Thank you, Gerald for offering me the possibility to pursue a Ph.D. degree in your research group. I also wish to thank the NEWCOM# (Network of Excellence in Wireless Communications) program. It was thanks to this program that I met Pablo in Lisbon at the very beginning of my Ph.D. studies in 2014, which resulted in very exciting collaboration and several research visits. Thank you, Pablo.

I am grateful that I was part of the Institute of Telecommunications, working alongside its staff, people who made the last four years truly enjoyable. Especially, I want to thank Günther for many discussions, insightful comments and advice. Of course this also extends to other members of the institute, in particular, (in alphabetical order) Andreas, Gábor, Manfred, Michael, Peter, and Stefan. These and all the other staff members, who are too numerous to all be listed here made my stay a very worthwhile experience.

También quiero dar las gracias a mi novia Violeta por su apoyo no solo durante mi doctorado sino en los últimos siete años.

Ich möchte mich auch bei meinen Freunden, meiner Familie und natürlich ganz besonders bei meinen Eltern bedanken, die mich nicht nur während des Doktorats, sondern während meiner gesamten Ausbildung unterstützt und ermutigt haben.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

CEO	chief executive officer
i.i.d.	identically and independently distributed
p.m.f.	probability mass function

Part I

INTRODUCTION

This part is organized into two chapters. In Chapter 1 we motivate the problems that will be studied in this thesis, provide an informal definition of the setup, and summarize the original findings. Chapter 2 introduces the necessary notation as well as the relevant definitions and results that will be used throughout the remainder of this thesis.

1 MOTIVATION

This thesis is concerned with the information-theoretic treatment of a data clustering technique that uses *mutual information* as its figure of merit.

The mutual information $I(X;Y)$ between two random variables X and Y is a fundamental quantity in information theory and measures the information one random variable contains about the other. It has many useful mathematical properties and its definition does not require any additional structure as it can even be defined for random variables on arbitrary probability spaces [23, Section 7.4] using the general definition of *relative entropy* [23, Section 7.1]. These properties make relative entropy and mutual information appealing candidates as objective functions in learning problems.

We adopt an information-theoretic point of view in the investigation of the *biclustering* (or *co-clustering*) technique, a data clustering algorithm. Biclustering was first explicitly considered by Hartigan [28] in 1972. A historical overview of biclustering including additional background can be found in [38, Section 3.2.4]. In general, given an $N \times M$ data matrix (a_{nm}), the goal of a biclustering algorithm [37] is to find partitions $\mathcal{B}_k \subseteq \{1,\dots,N\}$ and $\mathcal{C}_l \subseteq \{1,\dots,M\}$, $k = 1 \dots K$, $l = 1 \dots L$ such that all the "biclusters" $(a_{nm})_{n\in\mathcal{B}_k, m\in\mathcal{C}_l}$ are in a certain sense homogeneous. The measure of homogeneity of the biclusters depends on the specific application. The method received renewed attention when Cheng and Church [9] successfully applied it to gene expression data. Many biclustering algorithms have been developed since (e.g., see [62] and references therein). An introductory overview of clustering algorithms for gene expression data can be found in the lecture notes [58]. The information bottleneck method, which can be viewed as a uni-directional information-theoretic variant of biclustering, was successfully applied to gene expression data as well [59].

In 2003, Dhillon et. al. [14] adopted an information-theoretic approach to biclustering. Specifically, for the special case when the underlying matrix represents the joint probability mass function (p.m.f.) of two discrete random variables X and Y, i. e., $a_{nm} = P\{X = n, Y = m\}$, their goal was to find functions $f\colon \{1,2,\dots,N\} \to \{1,2,\dots,K\}$ and $g\colon \{1,2,\dots,M\} \to \{1,2,\dots,L\}$ that maximize $I\big(f(X);g(Y)\big)$ for specific K and L. We extend this setup from two to an arbitrary number of random variables and, using identically and independently distributed (i.i.d.) copies of those sources, we define information-theoretic *achievability*. The aim of this thesis is to characterize the resulting achievable

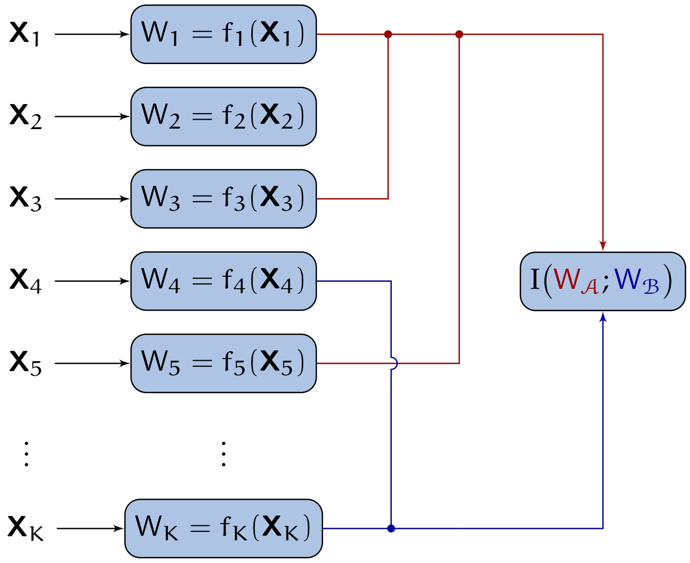

$\mathbf{X}_1$
$W_1 = f_1(\mathbf{X}_1)$
$\mathbf{X}_2$
$W_2 = f_2(\mathbf{X}_2)$
$\mathbf{X}_3$
$W_3 = f_3(\mathbf{X}_3)$
$\mathbf{X}_4$
$W_4 = f_4(\mathbf{X}_4)$
$I(W_{\mathcal{A}}; W_{\mathcal{B}})$
$\mathbf{X}_5$
$W_5 = f_5(\mathbf{X}_5)$
$\mathbf{X}_K$
$W_K = f_K(\mathbf{X}_K)$

Figure 1: Information-theoretic clustering.

region of this information-theoretic clustering problem under various constraints and connect it to other problems in network information theory. It is a multi-terminal source coding problem that offers a formidable mathematical complexity and is fundamentally different from "classical" distributed source coding problems like distributed lossy compression [16, Chapter 12]. Usually, one aims at reducing redundant information, i. e., information that is transmitted by multiple encoders, as much as possible, while still guaranteeing correct decoding. By contrast, in the clustering problem we are interested in maximizing this very redundancy. In this sense, the clustering problem is complementary to conventional distributed source coding.

1.1 PROBLEM SETUP

In this section we will present an informal introduction to the information-theoretic clustering problem. The formal definitions will follow in Part II. We will introduce the general problem as well as the special cases that are treated in this thesis and summarize the major original findings of this work in Section 1.2.

The input data is modeled by K random vectors $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_K$, which are formed by n i.i.d. copies of the discrete random variables $X_1, X_2, \dots, X_K$. The clustering is performed by K function $f_1, \dots, f_K$, where f_k takes $\mathbf{X}_k$ as its input and forms a description $W_k = f_k(\mathbf{X}_k)$. These descriptions are then collected into two disjoint, nonempty sets $\mathcal{A}, \mathcal{B} \subseteq \{1, 2, \dots, K\}$, which are compared to each other in terms of mutual information $\mu_{\mathcal{A},\mathcal{B}} = I(W_{\mathcal{A}}; W_{\mathcal{B}})$, as depicted in Figure 1.

Without any restrictions on the functions f_k, the optimal choice is the identity function $W_k = f_k(\mathbf{X}_k) = \mathbf{X}_k$, achieving the maximum mutual information $I(W_{\mathcal{A}}; W_{\mathcal{B}}) = I(\mathbf{X}_{\mathcal{A}}; \mathbf{X}_{\mathcal{B}})$. We therefore restrict the functions f_k and then characterize the achievable mutual information $I(W_{\mathcal{A}}; W_{\mathcal{B}})$ for different pairs of sets $(\mathcal{A}, \mathcal{B})$.

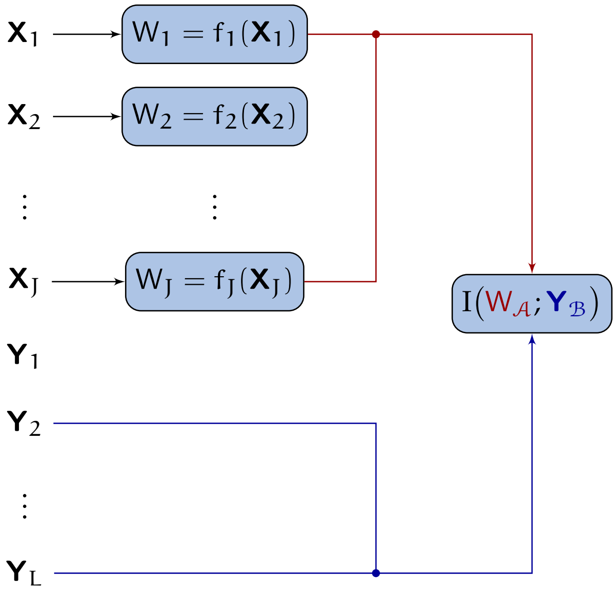
In the spirit of Shannon's work [57], the most natural restriction is bounding the rate of f_k, i.e., requiring $\frac{1}{n}\log_2|f_k| \leqslant R_k$ for $k \in \{1, 2, \dots, K\}$, where $|f_k|$ is the cardinality of the range of f_k and R_k is the rate in bit. Given the rates $R_1, R_2, \dots, R_K$, and values $\mu_{\mathcal{A},\mathcal{B}}$ for any nonempty disjoint pair $\mathcal{A}, \mathcal{B} \subseteq \{1, 2, \dots, K\}$, we are interested, whether the rates R_k are sufficient to achieve $\frac{1}{n}I(W_{\mathcal{A}}; W_{\mathcal{B}}) \geqslant \mu_{\mathcal{A},\mathcal{B}}$ for all such pairs $(\mathcal{A}, \mathcal{B})$ simultaneously. We call those values *achievable* and let $\mathcal{R}$ be the set of all achievable points. Note that the achievable region $\mathcal{R} \subseteq \mathbb{R}^{3^K - 2^{K+1} + K + 1}$ of this multi-clustering problem is a high-dimensional region that is difficult to characterize. In particular, a single-letter characterization of this region is currently out of reach, as quantize-and-bin coding schemes are known to be unable to achieve the full achievable region in this general setting. It contains a famous counterexample [34] as a special case. We will therefore provide outer and inner bounds on the achievable region in the general case and investigate several special cases of this multi-clustering problem, for which stronger statements can be made:

n is the length of the vector $\mathbf{X}_k$.

There are $3^K - 2^{K+1} + 1$ ways to choose the pair $(\mathcal{A}, \mathcal{B})$.

- For $K = 2$ sources, the multi-clustering problem turns out be equivalent to a hypothesis testing problem [25] and a pattern recognition problem [66]. We explicitly state these equivalences and exploit them, providing easy proofs of bounds on the achievable region for $K = 2$.
- If $K = 2$, the subset of the achievable region where $R_2 = \infty$ admits a single-letter characterization and the associated problem is known as the information bottleneck problem.
- We also investigate the case where $K = 2$ and (X_1, X_2) is a doubly symmetric binary source [16, Example 10.1], which was previously studied in [66, Section VII.A]. Based on novel cardinality bounds we are able to provide evidence that there is a gap between the inner and outer bounds, disproving [66, Conjecture 1].
- We will investigate a variant of the chief executive officer (CEO) problem, depicted in Figure 2, which can be obtained from the multi-clustering problem by requiring certain rates to be infinite and $\mu_{\mathcal{A},\mathcal{B}} = 0$ for certain sets $(\mathcal{A}, \mathcal{B})$. We will provide a tight single-letter characterization of the achievable region for a special case of this problem with multiple description coding.

For $K = 2$ and a doubly symmetric binary source (X_1, X_2), we also consider a different type of constraint on the functions (f_1, f_2), that is not a rate constraint. Allowing f_2 to be arbitrary, but requiring



$\mathbf{X}_1$
$W_1 = f_1(\mathbf{X}_1)$
$\mathbf{X}_2$
$W_2 = f_2(\mathbf{X}_2)$
$\mathbf{X}_J$
$W_J = f_J(\mathbf{X}_J)$
$I(W_{\mathcal{A}};\mathbf{Y}_{\mathcal{B}})$
$\mathbf{Y}_1$
$\mathbf{Y}_2$
$\mathbf{Y}_L$

Figure 2: CEO Problem.

Optimality follows from the data-processing inequality.

f_1 to be a Boolean function, i. e., $|f_1| = 2$, one can ask for the maximal value of $I\big(f_1(\mathbf{X}_1); f_2(\mathbf{X}_2)\big)$, that is attainable for any $n \in \mathbb{N}$. Clearly, choosing f_2 to be the identity function is optimal, yielding $I\big(f_1(\mathbf{X}_1); f_2(\mathbf{X}_2)\big) = I\big(f_1(\mathbf{X}_1); \mathbf{X}_2\big)$. A still open conjecture [10, Conjecture 1] claims that $I\big(f_1(\mathbf{X}_1); \mathbf{X}_2\big) \leqslant I(X_1; X_2)$ for any Boolean function f_1. We prove the weaker statement, $I\big(f_1(\mathbf{X}_1); f_2(\mathbf{X}_2)\big) \leqslant I(X_1; X_2)$ for any pair of Boolean functions (f_1, f_2). This result readily follows from the original conjecture via the data-processing inequality and was stated as an open problem in [36, Section IV] and [10, Section IV], and previously investigated in [5].

1.2 ORIGINAL CONTRIBUTIONS

The following is a summary of the original contributions of this thesis.

- We introduce the multi-clustering problem and provide inner and outer bounds on the achievable region.
- The CEO problem obtained from the general multi-clustering problem is shown to be equivalent to the CEO problem under logarithmic loss distortion.
- A single-letter characterization of the achievable region of the multiple description CEO problem is obtained.
- We provide a thorough study of the multi-clustering problem for the case of two sources, relating it to several other problems in the literature.

- Novel cardinality bounds are provided for auxiliary random variables for both the outer and the inner bound on the achievable region of the multi-clustering problem in the case of two sources. However, the technique is not limited to $K = 2$ sources and might even provide better cardinality bounds for auxiliaries in other coding theorems.
- We performed a thorough analysis of the doubly symmetric binary source case, enabled by the improved cardinality bounds on the auxiliary random variables. Our work grants more insight into the achievable region for this particular source distribution, disproving an open conjecture.
- We establish the correctness of the Courtade-Kumar conjecture for two Boolean functions [10, Section IV].

1.3 THESIS ORGANIZATION

This thesis covers two main topics, the multi-clustering problem and the Kumar-Courtade conjecture [36] for the case of two Boolean functions. Both are presented in Part II.

We first provide some preliminary material in Chapter 2, which may be safely skipped by a reader familiar with the topic. After introducing the notation in Section 2.1, Section 2.2 contains the necessary material on (network) information theory. In particular, we introduce types, typical sequences and provide results concerning various problems in network information theory. We largely follow the textbooks [12], [13], [16]. In Section 2.3, we summarize elementary results from real analysis required for subsequent proofs. For the sake of completeness, Section 2.4 contains the elementary definitions and results of Boolean analysis, taken from the textbook [43]. A small introduction to submodular functions, taken from the textbook [18], follows in Section 2.5. This material is used in Section 2.6, when proving a lemma on a sequence of convex polyhedra. As most results in Chapter 2 are given without proof, we provide a list of references to proofs in Section 2.7.

We study the multi-clustering problem in the case of two sources in Chapter 3. We published parts of the material in this chapter in [48], [52]. After introducing the problem (Section 3.1), we connect it to a hypothesis testing and to a pattern recognition problem in Section 3.2. We then exploit these connections in Section 3.3 to provide single-letter bounds for the achievable region. These bounds become tight in a special case, corresponding to the information bottleneck problem, which is itself equivalent to a source coding problem with logarithmic loss distortion. We investigate these connections in Section 3.4. Section 3.5 deals with the special case of a doubly symmetric binary source, where we find strong evidence for a gap between the outer and inner bounds that were introduced in Section 3.3.

We present a proof for the two-function case of the Kumar-Courtade conjecture in Chapter 4. Some material in this chapter was published in our previous work [47], [51]. The main results are stated in Section 4.1 and the proof is provided in Section 4.2. In Section 4.3 we further show that among the Boolean functions, the dictator functions are in fact the unique maximizers of mutual information, except in degenerate cases.

Chapter 5 extends the multi-clustering problem to an arbitrary number of sources. Part of the material in this chapter appeared in our papers [49], [50], [52]. We provide generalized outer and inner bounds in Section 5.1. The proof of the inner bound is deferred to Section 5.3. In Section 5.2 we show that the CEO problem with logarithmic loss distortion constitutes a special case of the multi-clustering problem. We generalize this to a multiple description CEO problem in Section 5.4, where we are able to provide a tight single-letter characterization of the achievable region under suitable Markov constraints.

The findings of this thesis are summarized and discussed in Chapter 6. Additionally, we provide a brief outlook and suggestions for future work on the topics.

The third and final part of this thesis is the Appendix. Proofs of some results in Chapters 3 and 5 are deferred to Appendices A and B, respectively. This is done to improve readability if the proof is rather technical and not immediately necessary to follow the presentation in Part II.

2 PRELIMINARIES

To obtain the results described in this thesis, we had to draw from several mathematical theories. In this chapter we will provide the necessary fundamentals. If the reader is already familiar with these topics, this chapter may be safely skipped.

Most proofs are omitted for brevity, however, references to proofs in the literature are listed in Section 2.7.

2.1 NOTATION AND CONVENTIONS

We will start by introducing the necessary notation. Commonly used symbols and notation can be found in the List of Symbols in Section 2.1.1. A list of acronyms is also available on page xvii.

2.1.1 *List of Symbols*

$\mathbf{0}$	all-zeros vector $\mathbf{0} \in \mathbb{R}^n$
$\varnothing$	empty set or constant random variable
$[0:n]$	interval $[0:n] := \{0, 1, 2, \ldots, n\}$ for $n \in \mathbb{N}_0$
$\mathbb{1}_{\mathcal{A}}$	indicator of the set $\mathcal{A}$; $\mathbb{1}_{\mathcal{A}}(a) = \begin{cases} 1 & a \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases}$
$2^{\mathcal{E}}$	power set of the set $\mathcal{E}$ [40, Section §1]; $2^{\mathcal{E}} := \{\mathcal{X} : \mathcal{X} \subseteq \mathcal{E}\}$
$\vert\mathcal{A}\vert$	number of elements of the set $\mathcal{A}$; $\vert\mathcal{A}\vert := \sum_{a \in \mathcal{A}} 1$
$\bar{a}$	$\bar{a} := 1 - a$ for $a \in \mathbb{R}$
$a * b$	binary convolution; $a * b := a\bar{b} + \bar{a}b$ for $a, b \in \mathbb{R}$ (cf. $\bar{a}$)
$\mathcal{A} + \mathcal{B}$	Minkowski sum of sets $\mathcal{A}$ and $\mathcal{B}$; $\mathcal{A} + \mathcal{B} := \{a + b : a \in \mathcal{A}, b \in \mathcal{B}\}$
$a \oplus b$	binary sum/exclusive or; $a \oplus b = a * b$ (cf. $a * b$)
$\mathcal{A}^c$	complement of the set (or event) $\mathcal{A}$
$\overline{\mathcal{A}}$	topological closure of the set $\mathcal{A}$ [40, Section §17]

$\mathbf{A}^{T}$	transposition of the matrix $\mathbf{A}$
$\mathcal{B}(p)$	Bernoulli distribution with parameter p
$\mathrm{cc}(\mathcal{S})$	characteristic cone of the closed, convex set $\mathcal{S}$
$\mathrm{cc}_{\mathbf{y}}(\mathcal{S})$	characteristic cone of the convex set $\mathcal{S}$ at $\mathbf{y} \in \mathcal{S}$
$\mathrm{conv}(\mathcal{A})$	convex hull of the set $\mathcal{A}$
$d_{LL}(p,x)$	logarithmic loss distortion; $d_{LL}(p,x) := -\log_2 p(x)$
$D(p \Vert q)$	Kullback-Leibler divergence; $D(p \Vert q) := \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}$
$\mathrm{DSBS}(p)$	doubly symmetric binary source; $X, Y \sim \mathcal{B}(\frac{1}{2})$ and $P\{X \neq Y\} = p$ (cf. $\mathcal{B}(p)$)
$\sqsubset e$	$\sqsubset e := \{e' \in \mathcal{E} : e' \sqsubset e\}$ for a total order $\sqsubset$ on $\mathcal{E}$; accordingly for $\sqsupseteq$, $\sqsupset$ and $\sqsubseteq$
$\boldsymbol{e}_i$	ith canonical base vector; $\boldsymbol{e}_i \in \mathbb{R}^n \colon \boldsymbol{e}_{i,j} = \mathbb{1}_i(j)$ (cf. $\mathbb{1}_{\mathcal{A}}$)
$\mathbb{E}[X]$	expectation of the random variable X
$\mathrm{ext}(\mathcal{S})$	extreme points of the convex set $\mathcal{S}$
$\mathbb{E}[X \vert Y]$	conditional expectation of the random variable X given Y
$H^{-1}(t)$	inverse of the binary entropy function; $H^{-1} \colon [0,1] \to [0, \frac{1}{2}]$ and $H\big(H^{-1}(t)\big) = t$ for all $t \in [0,1]$ (cf. Definition 2.3)
$H(p)$	binary entropy function; $H(p) := -p \log_2 p - \bar{p} \log_2 \bar{p}$ for $p \in (0,1)$ and $H(0) := H(1) := 0$ (cf. Definition 2.3)
$H(p_X)$	$H(p_X) := H(X)$ for $p_X \in \mathcal{P}(\mathcal{X})$ and $X \sim p$
$H(X)$	entropy of the random variable X in bits
$H(X \vert Y)$	conditional entropy in bits of the random variable X given the random variable Y
$I(X;Y)$	mutual information in bits between the random variables X and Y

$I(X;Y\|Z)$	conditional mutual information in bits between the random variables X and Y given the random variable Z
$\ker(\mathbf{A})$	kernel of the matrix $\mathbf{A}$
$\mathbb{N}$	natural numbers; $\mathbb{N} := \{1, 2, 3, \dots\}$
$[n]$	interval $[n] := \{1, 2, \dots, n\}$ for $n \in \mathbb{N}$
$\mathbb{N}_0$	non-negative integers; $\mathbb{N}_0 := \{0, 1, 2, \dots\}$
$N_{\mathcal{H}}(\mathbf{x})$	number of inequalities satisfied with equality at $\mathbf{x}$ cf. Definition 2.55
$N(x\|\mathbf{x})$	counting function; $N(x\|\mathbf{x}) := \sum_{i=1}^{n} \mathbb{1}_x(\mathbf{x}_i)$ (cf. $\mathbb{1}_{\mathcal{A}}$)
Ω	set of all pairs $(\mathcal{A}, \mathcal{B})$, where $\mathcal{A}, \mathcal{B} \subset [K]$ are nonempty and disjoint
$P\{\mathcal{A}\}$	probability of the event $\mathcal{A}$
$P\{\mathcal{A}\|\mathcal{B}\}$	conditional probability of events $\mathcal{A}$, $\mathcal{B}$
$P\{\mathcal{A}\|X\}$	conditional probability of the event $\mathcal{A}$ given X; $P\{\mathcal{A}\|X\} := \mathbb{E}[\mathbb{1}_{\mathcal{A}}\|X]$
Π	set of all pairs $(\mathcal{A}, \mathcal{B})$, where $\mathcal{A} \subseteq [J]$ and $\mathcal{B} \subseteq [L]$ are nonempty
$\mathcal{P}(\mathcal{X})$	set of all p.m.f.s on the finite set $\mathcal{X}$
p_X	p.m.f. of the random variable X
$\mathbb{R}$	real numbers
$\mathrm{rank}(\mathbf{A})$	rank of the matrix $\mathbf{A}$
$\mathbb{R}_-$	non-positive real numbers; $\mathbb{R}_- := \{x \in \mathbb{R} : x \leqslant 0\}$
$\mathbb{R}_+$	non-negative real numbers; $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geqslant 0\}$
$\mathrm{sgn}(x)$	sign function; $\mathrm{sgn}(x) = 1$ for $x \geqslant 0$ and $\mathrm{sgn}(x) = -1$ otherwise
$\Theta(f;g)$	co-information of f and g; cf. Definitions 3.1 and 5.1
$\mathcal{T}_X^n$	set of n-sequences with type X
$\mathcal{T}_{[X]\delta}^n$	set of δ-typical n-sequences w.r.t. X

$\mathcal{T}^n_{[Y\mid X]\delta}(\mathbf{x})$	set of conditionally δ-typical n-sequences given $\mathbf{x}$
$\mathcal{T}^n_{Y\mid X}(\mathbf{x})$	set of n-sequences with conditional type Y given $\mathbf{x}$
$\mathcal{T}_{[X]\delta}$	set of δ-typical random variables w.r.t. X
$\mathcal{U}(\mathcal{X})$	uniform distribution on the finite set $\mathcal{X}$
$\mathbf{x} \leqslant \mathbf{y}$	(partial) product order for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{E}}$; $\mathbf{x} \leqslant \mathbf{y} \iff x_e \leqslant y_e$ for all $e \in \mathcal{E}$
$X \sim p$	the random variable X is distributed according to the p.m.f. p
$\mathbf{x} \perp \mathbf{y}$	orthogonal vectors $\mathbf{x}$ and $\mathbf{y}$, i. e., $\mathbf{x}^T\mathbf{y} = 0$
$\chi_{\mathcal{S}}$	basis functions on the Boolean hypercube; $\chi_{\mathcal{S}}(\mathbf{x}) := \prod_{i\in\mathcal{S}} x_i$
$X \perp Y$	the random variables X, Y are independent
$\langle x, y\rangle$	inner product of x and y
$X \mathrel{-\!\!\circ\!\!-} Y \mathrel{-\!\!\circ\!\!-} Z$	the random variables X, Y, and Z form a Markov chain in this order

2.1.2 *Generic Conventions*

In general, we use calligraphic type to denote sets and events. We denote random quantities and their realizations by capital, sans-serif and lowercase letters, respectively. By convention, empty products are equal to 1, empty sums are regarded as 0, empty intersections are equal to the entire ambient space and empty unions are regarded as the empty set $\varnothing$. When there is no possibility of confusion we identify a singleton set with its element, e. g., we write $\{1,2,3\} \setminus 1 = \{2,3\}$. We will use superscript to indicate that a relation follows from a specific equation, e. g., the inequality $a \overset{(42)}{\leqslant} b$ follows from equation (42).

2.1.3 *Definitions*

Let Ω denote the set of all pairs $(\mathcal{A}, \mathcal{B})$, where $\mathcal{A}, \mathcal{B} \subset [K]$ are nonempty and disjoint for $K \in \mathbb{N}$. Note that $|\Omega| = 3^K - 2^{K+1} + 1$. We also define Π as the set of all pairs $(\mathcal{A}, \mathcal{B})$, where $\mathcal{A} \subseteq [J]$ and $\mathcal{B} \subseteq [L]$ are nonempty and $J, L \in \mathbb{N}$. Hence, we have $|\Pi| = 2^{J+L} - 2^J - 2^L + 1$.

For a total order $\sqsubset$ on a set $\mathcal{E}$ and $e \in \mathcal{E}$ we will use the notation $\sqsupset e := \{e' \in \mathcal{E} : e' \sqsupset e\}$ and accordingly for $\sqsupseteq$, $\sqsubset$ and $\sqsubseteq$. E. g., given the total order $\sqsubset$ on $\{1,2,3\}$ with $3 \sqsubset 1 \sqsubset 2$, we have $\sqsupset 3 = \{1,2\}$, $\sqsupset 1 = \{2\}$ and $\sqsupset 2 = \varnothing$.

2.1.4 *Vectors, Matrices and Tuples*

For an arbitrary set $\mathcal{X}$, an $\mathcal{X}$-valued vector is a function in $\mathcal{X}^{\mathcal{E}}$ for some finite set $\mathcal{E}$. Vectors are indicated by bold-face, lower-case type, e. g., $\mathbf{x} = (x_e)_{e \in \mathcal{E}} \in \mathcal{X}^{\mathcal{E}}$. Matrices are typeset in bold-face, upper-case letters, e. g., $\mathbf{A} \in \mathcal{X}^{m \times n}$. If not otherwise specified, we will deal with n-vectors, i. e., $\mathcal{E} = [n]$. Subscripts indicate parts of vectors, e. g., $\mathbf{x}_{\mathcal{A}} := (x_e)_{e \in \mathcal{A}}$ for $\mathcal{A} \subseteq \mathcal{E}$.

For $\mathbf{x} \in \mathcal{X}^n$ we further use the common abbreviations $\mathbf{x}_i^j := \mathbf{x}_{\{i,\dots,j\}}$, $\mathbf{x}^j := \mathbf{x}_1^j$ for $1 \leqslant i \leqslant j \leqslant n$ and, if a vector is already carrying a subscript, it will be separated by a comma, e. g., $\mathbf{x}_{3,1}^5 = (\mathbf{x}_3)_1^5 = (\mathbf{x}_3)^5$.

Additionally, we use subscript sets to denote tuples, e. g., $x_{[K]} \in \mathbb{R}^K$ or $x_\Omega \in \mathbb{R}^{3^K - 2^{K+1} + 1}$. Naturally, slices of tuples are indexed by subsets, e. g., $x_{\mathcal{A}} = (x_i)_{i \in \mathcal{A}}$ for a tuple $x_{\mathcal{B}}$ and $\mathcal{A} \subseteq \mathcal{B}$. This notation extends naturally to tuples of vectors, where the subscript indices are separated by a comma, e. g., for $\mathbf{x}_{[K]} \in \mathbb{R}^{nK}$, we have $\mathbf{x}_{\mathcal{A},l}^k = (\mathbf{x}_{\mathcal{A}})_l^k \in \mathbb{R}^{(k-l+1)|\mathcal{A}|}$ for $\mathcal{A} \subseteq [K]$.

2.1.5 *Probability and Information Theory*

Random variables/random vectors are assumed to be supported on finite sets. Given random variables, e. g., (X, Y), unless otherwise specified, the corresponding boldfaced random n-vectors $(\mathbf{X}, \mathbf{Y})$ are n identically and independently distributed (i.i.d.) copies of (X, Y), i. e., $(\mathbf{X}, \mathbf{Y}) = (X, Y)^n$. We use the same letter for the random variable and for its support set, e. g., Y takes values in $\mathcal{Y}$ and X_3 takes values in $\mathcal{X}_3$. Given a random variable X, we write p_X for its probability mass function (p.m.f.). We will also use the notation $X = \varnothing$ to indicate that the random variable X is void, i. e., equal to a constant with probability one. We will use the usual notation for information-theoretic quantities, given in the List of Symbols in Section 2.1.1. Information will be measured in bit.

When generating codebooks we will assume that the codebook size is an integer to keep the notation simple.

2.2 INFORMATION THEORY

We start by noting the following basic fact from probability theory.

Theorem 2.1 (Markov's inequality). *For an arbitrary random variable* X, *any function* $f : \mathcal{X} \to \mathbb{R}_+$, *and any* $\lambda > 0$ *we have*

$$P\{f(X) \geqslant \lambda\} \leqslant \frac{\mathbb{E}\left[f(X)\right]}{\lambda}. \tag{2.1}$$

Proof. We have

$$\begin{aligned}\mathbb{E}[f(X)] &= \sum_{x\in\mathcal{X}} f(x)p(x) && (2.2)\\ &\geqslant \sum_{x\in\mathcal{X}:f(x)\geqslant\lambda} f(x)p(x) && (2.3)\\ &\geqslant \sum_{x\in\mathcal{X}:f(x)\geqslant\lambda} \lambda p(x) && (2.4)\\ &= \lambda P\{f(X)\geqslant\lambda\}, && (2.5)\end{aligned}$$

where (2.3) follows from $f \geqslant 0$. □

A fundamental information-theoretic quantity is *Kullback-Leibler divergence* [35], sometimes also referred to as *relative entropy*. It measures a "distance" between p.m.f.s and can be generalized to arbitrary probability measures [23, Section 7.1]. Although it is not a metric [54, Definition 2.15], it has many convenient properties.

Definition 2.2. *Let* p *and* q *be two* p.m.f.s *on a common alphabet* $\mathcal{X}$*. The* Kullback-Leibler divergence *between* p *and* q *is defined as*

$$D(p\|q) := \sum_{x\in\mathcal{X}} p(x)\log_2\frac{p(x)}{q(x)}. \tag{2.6}$$

Note that we allow $D(p\|q) = \infty$ and adopt the usual convention $0\cdot\log_2 0 = 0\cdot\log_2\frac{0}{0} = 0$.

We can now define the fundamental quantities of information theory, entropy, mutual information and their conditional counterparts in terms of Kullback-Leibler divergence.

Definition 2.3. *For a discrete random variable* $X \sim p_X$ *on* $\mathcal{X}$*, define the* entropy *of* X *as*

$\mathcal{U}(\mathcal{X})$ is the uniform distribution on $\mathcal{X}$.

$$\begin{aligned}H(X) := H(p_X) &:= \log_2|\mathcal{X}| - D\big(p_X\big\|\mathcal{U}(\mathcal{X})\big) && (2.7)\\ &= -\sum_{x\in\mathcal{X}} p_X(x)\log_2 p_X(x). && (2.8)\end{aligned}$$

For random variables (X, Y, Z) *we define* conditional mutual information, mutual information, *and* conditional entropy *as*

$\varnothing$ denotes a random variable that is constant.

$$\begin{aligned}I(X;Y|Z) &:= H(XZ) + H(YZ) - H(XYZ) - H(Z), && (2.9)\\ I(X;Y) &:= I(X;Y|\varnothing) = H(X) + H(Y) - H(XY), && (2.10)\\ H(X|Y) &:= I(X;X|Y) = H(XY) - H(Y). && (2.11)\end{aligned}$$

$\bar{a} := 1 - a$

Slightly abusing notation, we also define the binary entropy function $H\colon [0,1] \to [0,1]$ *as* $H(p) := -p\log_2 p - \bar{p}\log_2\bar{p}$*. Accordingly, let* $H^{-1}\colon [0,1] \to [0,\frac{1}{2}]$ *be the inverse of the binary entropy function on* $[0,\frac{1}{2}]$*, defined by the relation* $H\big(H^{-1}(t)\big) = t$ *for all* $t \in [0,1]$.

Entropy and mutual information have several useful properties, which we collect in the following lemma.

Lemma 2.4. *For arbitrary random variables* X, Y, Z, $X_1, X_2, \dots, X_J$, *the following statements hold.*

1. $0 \leqslant H(X) \leqslant \log_2 |\mathcal{X}|$.
2. $I(X;Y|Z) \geqslant 0$.
3. $H(X_{[J]}|Y) = \sum_{j=1}^{J} H(X_j|X_1^{j-1}Y)$.
4. $I(X_{[J]};Y) = \sum_{j=1}^{J} I(X_j;Y|X_1^{j-1})$.

We have $X_{[J]} = (X_1, \dots, X_J)$ *and* $X_1^j = (X_1, \dots, X_j)$.

The results in Lemma 2.4 can be used to derive many other identities, e.g., $H(X|Y) \geqslant H(X|YZ)$ ("conditioning reduces entropy"). We will make extensive use of these results and apply them many times routinely throughout this thesis. We will often omit an explicit reference, unless an identity is applied in an unusual way, warranting additional explanation.

We will also make use of some additional information-theoretic inequalities, which are given in the following theorems.

Theorem 2.5 (Data-processing inequality). *If the Markov chain* $X \mathrel{-\!\!\circ\!\!-} Y \mathrel{-\!\!\circ\!\!-} Z$ *holds, then* $I(X;Y) \geqslant I(X;Z)$.

Theorem 2.6 (Fano's inequality). *If* $X \mathrel{-\!\!\circ\!\!-} Y \mathrel{-\!\!\circ\!\!-} \hat{X}$ *form a Markov chain and* $\varepsilon = P\{X \neq \hat{X}\}$, *then*

$$H(\varepsilon) + \varepsilon \log_2 |\mathcal{X}| \geqslant H(X|\hat{X}) \geqslant H(X|Y), \tag{2.12}$$

which can be weakened to

$$1 + \varepsilon \log_2 |\mathcal{X}| \geqslant H(X|Y). \tag{2.13}$$

$H(\varepsilon)$ *denotes the binary entropy function.*

Theorem 2.7 (Log-sum inequality). *Let* $a_i, b_i \in \mathbb{R}_+$ *for* $i \in [n]$. *Then,*

$$\sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geqslant \left(\sum_{i=1}^{n} a_i \right) \log_2 \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{2.14}$$

with equality if and only if there exists $c \in \mathbb{R}_+$, *such that* $a_i = cb_i$ *for all* $i \in [n]$.

$[n] = \{1, 2, \dots, n\}$.

Using the log-sum inequality, we can show that entropy $H(p)$ is not only concave [12, Theorem 2.7.3], but strictly concave.

Lemma 2.8. $H(p)$ *is strictly concave in* p, *i.e.,* $H(\lambda p + \bar{\lambda} q) > \lambda H(p) + \bar{\lambda} H(q)$ *for any* $\lambda \in (0,1)$ *and any two p.m.f.s* $p \neq q$ *on a common alphabet.*

Proof. Let p and q be two p.m.f.s on a common alphabet $\mathcal{X}$. For $\lambda \in (0, 1)$, we have

$$\begin{aligned}
&\lambda H(p) + \bar{\lambda} H(q) \\
&\overset{(2.7)}{=} \log_2 |\mathcal{X}| - \lambda D(p \| \mathcal{U}(\mathcal{X})) - \bar{\lambda} D(q \| \mathcal{U}(\mathcal{X})) && (2.15) \\
&= \log_2 |\mathcal{X}| - \sum_{x \in \mathcal{X}} \Bigg(\lambda p(x) \log_2 \frac{\lambda p(x)}{\lambda |\mathcal{X}|^{-1}} \\
&\qquad\qquad + \bar{\lambda} q(x) \log_2 \frac{\bar{\lambda} q(x)}{\bar{\lambda} |\mathcal{X}|^{-1}} \Bigg) && (2.16) \\
&\leqslant \log_2 |\mathcal{X}| - D(\lambda p + \bar{\lambda} q \| \mathcal{U}(\mathcal{X})) && (2.17) \\
&= H(\lambda p + \bar{\lambda} q). && (2.18)
\end{aligned}$$

In (2.17) we applied Theorem 2.7, which also shows that equality in (2.17) is only possible for $p = q$. □

Theorem 2.9 (Information inequality). *Let p and q be two p.m.f.s on a common alphabet, then $D(p \| q) \geqslant 0$ with equality if and only if $p = q$.*

Theorem 2.10 (Mrs. Gerber's Lemma). *The random variable U is arbitrary and X is a binary random variable, i.e., $\mathcal{X} = \{0, 1\}$. If $Z \sim \mathcal{B}(p)$, $p \in [0, 1]$, such that $Z \perp (X, U)$ and $Y := X \oplus Z$, then*

$$H(Y|U) \geqslant H\big(H^{-1}(H(X|U)) * p\big). \tag{2.19}$$

$a * b := \bar{a} b + a \bar{b}$.

$H^{-1}(t)$ *is the inverse of the binary entropy function (cf. Definition 2.3).*

We will also define the common information [19], [65], [67] of two random variables. Loosely speaking, common information is the amount of information that can be obtained from either random variable with probability one. When present, it can facilitate certain coding techniques (cf. [65]).

Definition 2.11. *For two random variables X and Y, a random variable Z is a* common component *of X and Y if and only if there exist functions ζ and ξ, such that $Z = \zeta(X) = \xi(Y)$ with probability one. The* common information *of X and Y is $\max_Z H(Z)$, where the maximum is over all common components of X and Y.*

Finally we provide the following two lemmas for future use. The first shows the existence of a code with low over-all error probability, when the individual error probabilities can be bounded arbitrarily close to zero. The second lemma is a technical result on mutual information involving four random variables.

Lemma 2.12. *For any $\delta > 0$ let C_δ be a random code and $(\mathcal{E}_i^{(\delta)})_{i \in \mathcal{I}}$ finitely many error events associated with the code C_δ. If we have $P\{\mathcal{E}_i^{(\delta)}\} \leqslant \delta$ for every $i \in \mathcal{I}$, then, for any $\varepsilon > 0$ we can find $\delta > 0$ such that there is a code c with $P\{\mathcal{E}_i^{(\delta)} | C_\delta = c\} \leqslant \varepsilon$ for every $i \in \mathcal{I}$.*

Proof. We apply Markov's inequality, Theorem 2.1, to the random variable $P\{\mathcal{E}_i^{(\delta)}|C_\delta\}$ and obtain

$$P\left\{P\{\mathcal{E}_i^{(\delta)}|C_\delta\} \geqslant \sqrt{\delta}\right\} \leqslant \frac{\delta}{\sqrt{\delta}} = \sqrt{\delta}. \tag{2.20}$$

Applying the union bound yields

$$\begin{aligned} P\left\{\bigcup_{i\in\mathcal{I}}\left\{P\{\mathcal{E}_i^{(\delta)}|C_\delta\} \geqslant \sqrt{\delta}\right\}\right\} &\leqslant \sum_{i\in\mathcal{I}} P\left\{P\{\mathcal{E}_i^{(\delta)}|C_\delta\} \geqslant \sqrt{\delta}\right\} &(2.21)\\ &\leqslant |\mathcal{I}|\sqrt{\delta}. &(2.22)\end{aligned}$$

In particular, there exists at least one code c such that $P\{\mathcal{E}_i^{(\delta)}|C_\delta = c\} < \sqrt{\delta}$ for all $i \in \mathcal{I}$ if $|\mathcal{I}|\sqrt{\delta} < 1$. Choosing $\delta = \min\left\{\varepsilon^2, \frac{1}{2|\mathcal{I}|^2}\right\}$ yields the desired result. □

Lemma 2.13. *If* $U \mathrel{-\!\!\circ\!\!-} X \mathrel{-\!\!\circ\!\!-} Z$ *and* $X \mathrel{-\!\!\circ\!\!-} Z \mathrel{-\!\!\circ\!\!-} V$, *then*

$$\begin{aligned} I(U;X) + I(V;Z) - I(UV;XZ) &= I(U;V) - I(U;V|XZ), &(2.23)\\ I(U;X) + I(V;Z) - I(UV;XZ) &\leqslant I(U;Z), &(2.24)\\ I(U;X) + I(V;Z) - I(UV;XZ) &\leqslant I(V;X). &(2.25)\end{aligned}$$

If $U \mathrel{-\!\!\circ\!\!-} X \mathrel{-\!\!\circ\!\!-} Z \mathrel{-\!\!\circ\!\!-} V$ *then* $I(U;X) + I(V;Z) - I(UV;XZ) = I(U;V)$.

Proof. We obtain (2.23) from

$$\begin{aligned} &I(U;X) + I(V;Z) - I(UV;XZ)\\ &\quad= I(U;XZ) + I(V;XZ) - I(UV;XZ) &(2.26)\\ &\quad\overset{(2.10)}{=} H(U) + H(V) - H(UV) - H(U|XZ)\\ &\qquad - H(V|XZ) + H(UV|XZ) &(2.27)\\ &\quad\overset{(2.9)}{=} I(U;V) - I(U;V|XZ), &(2.28)\end{aligned}$$

where (2.26) follows from $U \mathrel{-\!\!\circ\!\!-} X \mathrel{-\!\!\circ\!\!-} Z$ and $X \mathrel{-\!\!\circ\!\!-} Z \mathrel{-\!\!\circ\!\!-} V$. To show that (2.24) holds, note that

$$\begin{aligned} I(U;X) + I(V;Z) - I(UV;XZ) &= I(V;Z) - I(V;XZ|U) &(2.29)\\ &= I(U;Z) + I(V;Z) - I(U;Z) - I(V;XZ|U) &(2.30)\\ &= I(U;Z) + I(V;Z) - I(U;Z) - I(V;Z|U) - I(V;X|ZU) &(2.31)\\ &= I(U;Z) + I(V;Z) - I(UV;Z) - I(V;X|ZU) &(2.32)\\ &= I(U;Z) - I(U;Z|V) - I(V;X|ZU) &(2.33)\\ &\leqslant I(U;Z), &(2.34)\end{aligned}$$

using part 4 of Lemma 2.4 and applying $U \mathrel{-\!\!\circ\!\!-} X \mathrel{-\!\!\circ\!\!-} Z$ in (2.29). The inequality (2.25) can be shown by interchanging $(X,U) \leftrightarrow (Z,V)$. The last claim is a direct consequence of (2.23). □

2.2.1 Types, Typical Sequences and Related Results

Several achievability proofs in this thesis are based on the notion of robust typicality [45], also used in [25]. For convenience, the necessary notation and relevant results are summarized in this section.

We will introduce the notation for types and typical sequences and make use of the δ-convention [13, Convention 2.11].

Definition 2.14 (Type; [13, Definition 2.1]). *The* type *of a vector* $\mathbf{x} \in \mathcal{X}^n$ *is the random variable* $\hat{X} \sim p_{\hat{X}} \in \mathcal{P}(\mathcal{X})$ *defined by*

$N(x|\mathbf{x})$ *counts the number of occurrences of* x *in* $\mathbf{x}$.

$$p_{\hat{X}}(x) = \frac{1}{n} N(x|\mathbf{x}), \text{ for every } x \in \mathcal{X}. \tag{2.35}$$

For a random variable $\hat{X}$*, the set of vectors with type* $\hat{X}$ *is denoted* $\mathcal{T}^n_{\hat{X}}$.

For a pair of random variables (X, Y)*, we say that* $\mathbf{y} \in \mathcal{Y}^n$ *has* conditional type Y *given* $\mathbf{x} \in \mathcal{X}^n$ *if and only if* $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}^n_{XY}$*. The set of all vectors* $\mathbf{y} \in \mathcal{Y}^n$ *with conditional type* Y *given* $\mathbf{x}$ *will be denoted* $\mathcal{T}^n_{Y|X}(\mathbf{x})$.

A key property of types is the following result, known as type counting.

Lemma 2.15 (Type counting). *The number of different types of vectors in* $\mathcal{X}^n$ *is less than* $(n+1)^{|\mathcal{X}|}$.

Some important properties of types are listed in the following lemma.

Lemma 2.16. *For a pair of random variables* (X, Y) *and* $\mathbf{x} \in \mathcal{T}^n_{\hat{X}}$*, the following properties hold:*

1. *If* $\mathcal{Y} = \mathcal{X}$*, we have*

$\mathbf{Y}$ *are* n *i.i.d. copies of* Y.

$$p_{\mathbf{Y}}(\mathbf{x}) = 2^{-n\left(H(X) + D(X\|Y)\right)}. \tag{2.36}$$

2. *If* $\mathcal{T}^n_{Y|X}(\mathbf{x}) \neq \varnothing$*, then*

$$(n+1)^{-|\mathcal{X}||\mathcal{Y}|} 2^{nH(Y|X)} \leqslant \left|\mathcal{T}^n_{Y|X}(\mathbf{x})\right| \leqslant 2^{nH(Y|X)}. \tag{2.37}$$

Definition 2.17 (Typicality; [16, Section 2.4]). *Consider* $X \sim p_X \in \mathcal{P}(\mathcal{X})$ *and* $\delta \geqslant 0$*. We call the random variable* $Y \sim p_Y \in \mathcal{P}(\mathcal{X})$ δ-typical *if and only if* $Y \in \mathcal{T}_{[X]\delta}$ *with*

$$\mathcal{T}_{[X]\delta} := \left\{ \tilde{X} \sim p_{\tilde{X}} \in \mathcal{P}(\mathcal{X}) : \left| p_{\tilde{X}}(x) - p_X(x) \right| \leqslant \delta p_X(x), \ \forall x \in \mathcal{X} \right\}. \tag{2.38}$$

A vector $\mathbf{x} \in \mathcal{X}^n$ *is* δ-typical *if its type* $\hat{X}$ *is* δ-typical. *The set of all* δ*-typical vectors is denoted* $\mathcal{T}^n_{[X]\delta}$.

Given $p_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ we call the elements of $\mathcal{T}^n_{[XY]\delta}$ the *jointly* δ*-typical* vectors. Furthermore, we define the *conditionally typical* vectors

$\mathcal{T}^n_{[Y|X]\delta}(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathcal{Y}^n : (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{T}^n_{[XY]\delta}\}$. Typical sequences have several useful properties, which are presented in the following.

Lemma 2.18. *The following properties hold for* $(X, Y) \sim p_{X,Y}$.

1. *For any* $\delta > 0$, *we have*

$$\lim_{n\to\infty} \mathrm{P}\left\{\mathbf{X} \in \mathcal{T}^n_{[X]\delta}\right\} = 1. \tag{2.39}$$

2. *Let* $\delta' > 0$ *and for each* $n \in \mathbb{N}$, *let* $\boldsymbol{x} \in \mathcal{T}^n_{[X]\delta'}$. *If* $\delta > \delta'$, *we have*

$$\lim_{n\to\infty} \mathrm{P}\left\{\mathbf{Y} \in \mathcal{T}^n_{[Y|X]\delta}(\boldsymbol{x}) \middle| \mathbf{X} = \boldsymbol{x}\right\} = 1. \tag{2.40}$$

Lemma 2.19 (Size of typical sets)**.** *The following properties hold for random variables* X *and* Y.

1. *Using* $\varepsilon(\delta) = \delta H(X)$,

$$\left|\mathcal{T}^n_{[X]\delta}\right| \leqslant 2^{n\left(H(X)+\varepsilon(\delta)\right)}. \tag{2.41}$$

2. *For* $\delta > 0$, $\varepsilon' > 0$, n *sufficiently large (as a function of* ε' *and* p_X*), and* $\varepsilon(\delta) = \delta H(X)$,

$$\left|\mathcal{T}^n_{[X]\delta}\right| \geqslant (1-\varepsilon')2^{n(H(X)-\varepsilon(\delta))}. \tag{2.42}$$

3. *For* $\boldsymbol{x} \in \mathcal{X}^n$ *and* $\varepsilon(\delta) = \delta H(Y|X)$,

$$\left|\mathcal{T}^n_{[Y|X]\delta}(\boldsymbol{x})\right| \leqslant 2^{n\left(H(Y|X)+\varepsilon(\delta)\right)}. \tag{2.43}$$

Remark 1. We will adopt the δ-convention [13, Convention 2.11] and assume the existence of an adequate sequence $(\delta_n)_{n\in\mathbb{N}} \to 0$ for every set of random variables. We will omit δ in the notation, e. g., we will write $\mathcal{T}_{[X]}$, $\mathcal{T}^n_{[X]}$, and $\mathcal{T}_{[X|Y]}(\boldsymbol{y})$.

Lemma 2.20 (Generalized Markov lemma)**.** *Let* $X_{[K]}$ *and* $U_{[K]}$ *be such that* $U_k \multimap\!\!\!\!\circ X_k \multimap\!\!\!\!\circ (X_{[K]\setminus k}, U_{[K]\setminus k})$ *for every* $k \in [K]$ *and fix* $\varepsilon > 0$. *For* $n \in \mathbb{N}$ *and for each* $k \in [K]$ *let* $M_k \in \mathbb{N}$ *with* $M_k > 2^{nI(X_k;U_k)}$. *Furthermore, let* $\widetilde{\mathbf{U}}_k(m)_{m\in[M_k]}$ *be* M_k *mutually independent random vectors, also independent of* $X_{[K]}$, *drawn uniformly from* $\mathcal{T}^n_{[U_k]}$. *Then, for sufficiently large* n *there exist* K *functions* $f_k\colon \mathcal{X}^n_k \times (\mathcal{U}^n_k)^{M_k} \to [M_k]$ *such that, using* $W_k = f_k(\mathbf{X}_k, \widetilde{\mathbf{U}}_k(m)_{m\in[M_k]})$ *and* $\mathbf{U}^*_k = \widetilde{\mathbf{U}}_k(W_k)$, *we have*

$$\mathrm{P}\left\{(\mathbf{X}_{[K]}, \mathbf{U}^*_{[K]}) \in \mathcal{T}^n_{[X_{[K]}U_{[K]}]}\right\} \geqslant 1-\varepsilon. \tag{2.44}$$

2.2.2 (Network) Information Theory

As the multi-clustering problem is connected to several problems in (network) information theory, we will introduce the relevant topics in this section. We illustrate the problem statements and provide fundamental definitions and results.

Most problems in information theory deal with the notion of codes. These are functions, used to convert a block of input symbols into a codeword. The size of the range of such an encoding function typically scales exponentially with the length of the blocks to provide a constant number of bits per input symbol. This notion is captured by the following definition.

Definition 2.21. *Let $X_{[K]}$ be K random variables, $n \in \mathbb{N}$, and $R_{[K]} \in (\mathbb{R} \cup \{\infty\})^K$. An $(n, R_{[K]})$* code *for the source $X_{[K]}$ consists of K functions $f_1, f_2, \dots, f_K$, where $f_k \colon \mathcal{X}_k^n \to \mathcal{M}_k$ with finite sets $\mathcal{M}_{[K]}$ such that $\frac{1}{n}\log_2|\mathcal{M}_k| \leqslant R_k$ for every $k \in [K]$.*

If $R_k = \infty$, then $\mathcal{M}_k$ can be any finite set.

Hypothesis Testing with Data Compression

Hypothesis testing with communication constraints was introduced in [1] and generalized in [25]. For a good overview of the subject, the reader is also referred to [26]. We will focus on the simplest case, a hypothesis test against independence. Given the (generally dependent) sources $(X, Z) \sim p_{XZ}$, define the independent random variables $(X^*, Z^*) \sim p_X p_Z$. Two agents independently observe either the n i.i.d. copies $(\mathbf{X}, \mathbf{Z})$ or $(\mathbf{X}^*, \mathbf{Z}^*)$. They communicate a rate limited description of their observation to a decoder. Given the two descriptions, the decoder must correctly (with high probability) decide whether $(\mathbf{X}, \mathbf{Z})$ or $(\mathbf{X}^*, \mathbf{Z}^*)$ was originally observed. Formally, we state the following definition.

Definition 2.22. *An (n, R_1, R_2) hypothesis test (HT) for (X, Z) consists of an (n, R_1, R_2) code (f_n, g_n) for (X, Z) and a set $\mathcal{A}_n \subseteq \mathcal{M}_1 \times \mathcal{M}_2$. The type I and type II error probabilities of $(f_n, g_n, \mathcal{A}_n)$ are defined as*

$\mathcal{M}_1$ ($\mathcal{M}_2$) denotes the range of f_n (g_n).

$$\alpha_n := P\{(f_n(\mathbf{X}), g_n(\mathbf{Z})) \in \mathcal{A}_n\}, \text{ and} \tag{2.45}$$

$$\beta_n := P\{(f_n(\mathbf{X}^*), g_n(\mathbf{Z}^*)) \notin \mathcal{A}_n\}, \tag{2.46}$$

respectively. A triple (μ, R_1, R_2) is HT-achievable *for the source (X, Z) if and only if, for every $\varepsilon > 0$, there is a sequence of (n, R_1, R_2) hypothesis tests $(f_n, g_n, \mathcal{A}_n)$, $n \in \mathbb{N}$ such that*

$$\lim_{n\to\infty} \alpha_n \leqslant \varepsilon, \text{ and} \tag{2.47}$$

$$\lim_{n\to\infty} -\frac{1}{n}\log\beta_n \geqslant \mu. \tag{2.48}$$

Let $\mathcal{R}_{\mathrm{HT}}$ denote the set of all HT-achievable triples.

The following result provides an inner bound on the HT-achievable region.

Theorem 2.23. *We have* $(\mu, R_1, R_2) \in \overline{\mathcal{R}_{\mathrm{HT}}}$ *if there exist random variables* U *and* V *such that* $U \mathrel{-\!\!\circ\!\!-} X \mathrel{-\!\!\circ\!\!-} Z \mathrel{-\!\!\circ\!\!-} V$ *and*

$$R_1 \geqslant I(U;X), \tag{2.49}$$

$$R_2 \geqslant I(V;Z), \tag{2.50}$$

$$\mu \leqslant I(U;V). \tag{2.51}$$

Pattern Recognition

Consider the pattern recognition (PR) problem introduced in [66]. For the sake of completeness, we restate the problem here.

Let $\big(\mathbf{X}(i), \mathbf{Z}(i)\big)$ be n i.i.d. copies of (X, Z), independently generated for each $i \in \mathbb{N}$. We store rate-limited encodings of all vectors $\mathbf{X}(\mathbb{N})$, by applying the same coding function to each vector. The resulting infinite codebook is available at the decoder. Fixing $\mu > 0$, we compute a rate-limited encoding of $\mathbf{Z}(W)$, where W is chosen by nature, independently at random with $W \sim \mathcal{U}([2^{n\mu}])$. This rate-limited description, together with the infinite codebook, is presented to the decoder, which has to determine the correct index W with high probability. The following definition captures this process.

Definition 2.24. *A triple* (μ, R_1, R_2) *is said to be* PR-achievable *for the source* (X, Z) *if and only if, for any* $\varepsilon > 0$*, there exists an* (n, R_1, R_2) *code* (f, g) *for* (X, Z) *and a function* $\phi\colon (\mathcal{M}_1)^{\mathbb{N}} \times \mathcal{M}_2 \to [2^{n\mu}]$ *such that,*

$\mathcal{M}_1$ and $\mathcal{M}_1$ are the ranges of f and g.

$$P\big\{W = \phi\big(C, g(\mathbf{Z}(W))\big)\big\} \geqslant 1 - \varepsilon, \tag{2.52}$$

where $C := f(\mathbf{X}(i))_{i \in \mathbb{N}}$ *and* $\big(\mathbf{X}(i), \mathbf{Z}(i)\big)_{i \in \mathbb{N}} \perp W \sim \mathcal{U}([2^{n\mu}])$*. Let* $\mathcal{R}_{\mathrm{PR}}$ *denote the set of all PR-achievable triples.*

The following two theorems provide an inner and an outer bound on the region $\overline{\mathcal{R}_{\mathrm{PR}}}$.

Theorem 2.25. *We have* $(\mu, R_1, R_2) \in \overline{\mathcal{R}_{\mathrm{PR}}}$ *if there exist random variables* U *and* V*, such that* $U \mathrel{-\!\!\circ\!\!-} X \mathrel{-\!\!\circ\!\!-} Z \mathrel{-\!\!\circ\!\!-} V$ *and*

$$R_1 \geqslant I(U;X), \tag{2.53}$$

$$R_2 \geqslant I(V;Z), \tag{2.54}$$

$$\mu \leqslant I(U;V). \tag{2.55}$$

Theorem 2.26. *If* $(\mu, R_1, R_2) \in \overline{\mathcal{R}_{\mathrm{PR}}}$*, then there exist random variables* U *and* V*, such that* $U \mathrel{-\!\!\circ\!\!-} X \mathrel{-\!\!\circ\!\!-} Z$ *and* $X \mathrel{-\!\!\circ\!\!-} Z \mathrel{-\!\!\circ\!\!-} V$ *and*

$$R_1 \geqslant I(U;X), \tag{2.56}$$

$$R_2 \geqslant I(V;Z), \tag{2.57}$$

$$\mu \leqslant I(U;V) - I(U;V|XZ). \tag{2.58}$$

Remark 2. We want to point out that the variant of the inner bound for the pattern recognition problem stated in [66, Theorem 1] is flawed. To see this, note that the point $(R_x = 0, R_y = b, R_c = b)$ is contained in $\mathcal{R}_{\mathrm{in}}$ (choose $U = V = \varnothing$) for any $b > 0$ even if the random variables X and Y are independent. But this point is clearly not achievable in general. However, the region $\mathcal{R}'_{\mathrm{in}}$, defined in the right column of [66, p. 303], coincides with our findings and the proof given in [66, Appendix A] holds for this region.

Here we use the notation of [66].

Information Bottleneck

In the information bottleneck (IB) problem, an agent observes the random vector $\mathbf{X}$ and intends to provide a rate limited description, as informative as possible about a different, unobserved random vector $\mathbf{Y}$. The concept was introduced in [63] and the first coding theorems for the information bottleneck method were obtained in [20]. Here we will present a slightly different formulation, also used in [11, Section III.F], which is more natural to an information theorist than that of [20]. The resulting achievable region, however, is identical.

$(\mathbf{X}, \mathbf{Y})$ are n i.i.d. copies of (X, Y).

Definition 2.27. *A pair (μ, R) is* IB-achievable *for the source (X, Y) if and only if, there exists an (n, R) code f for X such that*

$$\mu \leqslant \frac{1}{n} I\big(f(\mathbf{X}); \mathbf{Y}\big). \tag{2.59}$$

Let $\mathcal{R}_{\mathrm{IB}}$ be the set of all IB-achievable pairs.

CEO Problem with Logarithmic Loss Distortion

The chief executive officer (CEO) problem was introduced in [7]. The setup is motivated by a CEO, who intends to learn random vectors $\mathbf{Y}_{[L]}$. To this end she dispatches J agents, each of them observing a different random vector $\mathbf{X}_j$, for $j \in [J]$. As the CEO is a very busy person, each agents is only allocated a certain rate R_j for reporting back to her. Based on the input of agents $\mathcal{A}$, the CEO forms an estimate of $\mathbf{Y}_{\mathcal{B}}$, subject to a distortion criterion, for every $(\mathcal{A}, \mathcal{B}) \in \Pi$.

$(\mathbf{X}_{[J]}, \mathbf{Y}_{[L]})$ are n i.i.d. copies of $(X_{[J]}, Y_{[L]})$.

Π is defined in Section 2.1.3.

We will consider the special case of logarithmic loss distortion, which was recently analyzed in [11]. For $(\mathcal{A}, \mathcal{B}) \in \Pi$ we consider a decoding function $g_{\mathcal{A},\mathcal{B}} \colon \mathcal{M}_{\mathcal{A}} \to \mathcal{P}(\mathcal{Y}^n_{\mathcal{B}})$ that produces a probabilistic estimate of $\mathbf{Y}_{\mathcal{B}}$ given the output of the encoders $\mathcal{A}$. The quality of this probabilistic estimate is measured by logarithmic loss (LL) distortion, defined as $d_{\mathrm{LL}}(p, x) := -\log_2 p(x)$ for a p.m.f. $p \in \mathcal{P}(\mathcal{X})$ and $x \in \mathcal{X}$.

Definition 2.28. *We say that the point $(D_\Pi, R_{[J]})$ is* LL-achievable *for the source $(X_{[J]}, Y_{[L]})$ if and only if, for some $n \in \mathbb{N}$, there exists an $(n, R_{[J]})$*

code $f_{[J]}$ *for* $X_{[J]}$ *such that for all* $(\mathcal{A},\mathcal{B}) \in \Pi$ *there is a decoding function* $g_{\mathcal{A},\mathcal{B}} \colon \mathcal{M}_{\mathcal{A}} \to \mathcal{P}(\mathcal{Y}_{\mathcal{B}}^n)$ *with*

$$\frac{1}{n}\mathbb{E}[d_{LL}(g_{\mathcal{A},\mathcal{B}}(W_{\mathcal{A}}),\mathbf{Y}_{\mathcal{B}})] \leqslant D_{\mathcal{A},\mathcal{B}}, \tag{2.60}$$

where $W_j := f_j(\mathbf{X}_j)$, $j \in [J]$. *Let* $\mathcal{R}_{LL}$ *be the set of all LL-achievable points.*

We note that [11] considers the case where $L = 1$ and $D_{\mathcal{A},\mathcal{B}} = \infty$ whenever $\mathcal{A} \neq [J]$. For $L = 1$ and assuming that the random variables $X_{[J]}$ are independent given Y, we obtain the following single-letter characterization of the achievable region from [11].

Theorem 2.29. *Consider the case* $L = 1$, *let* $Y := Y_1$, *and assume* $X_j \mathrel{-\!\!\circ\!\!-} Y \mathrel{-\!\!\circ\!\!-} X_{[J]\setminus j}$ *for every* $j \in [J]$. *We then have* $(D, R_{[J]}) \in \overline{\mathcal{R}_{LL}}$ *if and only if there exist random variables* Q *and* $U_{[J]}$ *such that*

We identify $D = D_{[J],1}$.

$$\sum_{j\in\mathcal{A}} R_j \geqslant I(X_{\mathcal{A}}; U_{\mathcal{A}}|U_{\mathcal{A}^c}Q) \text{ for all } \mathcal{A} \subseteq [J], \tag{2.61}$$

$\mathcal{A}^c = [J] \setminus \mathcal{A}$.

$$D \geqslant H\big(Y\big|U_{[J]}Q\big), \tag{2.62}$$

with a joint p.m.f. $p_{YX_{[J]}U_{[J]}Q} = p_Y p_Q \prod_{j\in[J]} p_{X_j|Y} p_{U_j|X_jQ}$.

An important property of logarithmic loss distortion is given in the following lemma. This simple, yet crucial result appeared in [11, Lemma 1] and is also essential for the analysis performed in [11].

Lemma 2.30. *For two random variables* (X, Y) *and two functions* $f \colon \mathcal{X} \to \mathcal{M}$ *and* $g \colon \mathcal{M} \to \mathcal{P}(\mathcal{Y})$, *where* $\mathcal{M}$ *is an arbitrary set, we have*

$$\mathbb{E}\big[d_{LL}\big(g(f(X)), Y\big)\big] \geqslant H(Y|f(X)) \tag{2.63}$$

with equality if and only if $g(m) = P\{Y = \cdot\,|f(X) = m\}$ *for all* $m \in \mathcal{M}$ *with* $P\{f(X) = m\} \neq 0$.

Proof. Define $U := f(X)$ and let $m \in \mathcal{M}$ with $P\{U = m\} \neq 0$. We then have

$$\mathbb{E}\big[d_{LL}\big(g(f(X)), Y\big)\big|U = m\big] = -\mathbb{E}\big[\log_2 g(m)(Y)\big|U = m\big] \tag{2.64}$$
$$= -\sum_{y\in\mathcal{Y}} p_{Y|U}(y|m)\log_2 g(m)(y) \tag{2.65}$$
$$= H(Y|U = m) + D\big(p_{Y|U}(\,\cdot\,|m)\big\|g(m)\big) \tag{2.66}$$
$$\geqslant H(Y|U = m), \tag{2.67}$$

where (2.67) follows from Theorem 2.9. The final result follows by calculating the expectation over U. By Theorem 2.9, $g(m) = p_{Y|U}(\,\cdot\,|m)$ is a necessary and sufficient condition for equality. □

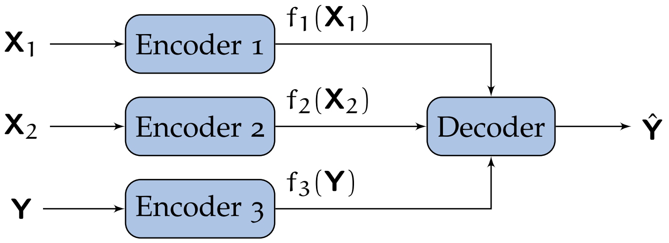


Figure 3: Two-help-one lossless source coding.

Körner-Marton Modulo-Two Sum Problem

Consider the two-help-one (TO) distributed lossless source coding problem depicted in Figure 3. Three random vectors, $\mathbf{Y}$, $\mathbf{X}_1$, and $\mathbf{X}_2$ are independently observed by agents, which communicate rate-limited descriptions to a decoder. The decoder's task is to reproduce $\mathbf{Y}$ with high accuracy. Formally we define achievability for this two-help-one problem as follows.

$\mathcal{M}_k$ is the range of the f_k.

Definition 2.31. *We say that the triple* (R_0, R_1, R_2) *is* TO-achievable *for the source* (X_1, X_2, Y) *if and only if, for any* $\varepsilon > 0$*, there exists an* (n, R_1, R_2, R_0) *code* (f_1, f_2, f_3) *for* (X_1, X_2, Y) *and a decoding function* $g\colon \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{M}_3 \to \mathcal{Y}^n$*, such that* $P\big\{\hat{\mathbf{Y}} \neq \mathbf{Y}\big\} \leqslant \varepsilon$ *where we defined* $\hat{\mathbf{Y}} := g\big(f_1(\mathbf{X}_1), f_2(\mathbf{X}_2), f_3(\mathbf{Y})\big)$*. Let* $\mathcal{R}_{\mathrm{TO}}$ *be the set of all TO-achievable triples.*

Characterizing the achievable region $\overline{\mathcal{R}_{\mathrm{TO}}}$ for arbitrary source distributions is an open problem. However, $\overline{\mathcal{R}_{\mathrm{TO}}}$ is known explicitly for a doubly symmetric binary source. Define the independent random variables $X_1 \sim \mathcal{B}(\frac{1}{2})$ and $Y \sim \mathcal{B}(p)$ and let $X_2 := X_1 \oplus Y$. We call (X_1, X_2) a doubly symmetric binary source (DSBS) [16, Example 10.1] with parameter p and write $(X_1, X_2) \sim \mathrm{DSBS}(p)$. Note that this distribution is symmetric, i. e., interchanging X_1 and X_2 yields the same joint distribution. In [34], Körner and Marton computed the achievable region $\overline{\mathcal{R}_{\mathrm{TO}}}$ for this particular input distribution. Remarkably, the resulting region cannot be achieved by a quantize-and-bin scheme (cf. [2], [6], [64], [70]) as shown in [34, Proposition 1].

Theorem 2.32. *Let* $(X_1, X_2) \sim \mathrm{DSBS}(p)$ *and* $Y := X_1 \oplus X_2$*. We then have* $(R_0, R_1, R_2) \in \overline{\mathcal{R}_{\mathrm{TO}}}$ *if and only if* $R_1, R_2, R_0 \geqslant 0$*,* $R_0 + R_1 \geqslant H(p)$*, and* $R_0 + R_2 \geqslant H(p)$*.*

Theorem 2.33. *Let $\mathcal{R}^*_{\mathrm{TO}}$ be the set of triples (R_0, R_1, R_2) such that there are random variables Q, U_1, and U_2, satisfying $Q \perp X_1X_2Y$, the Markov chains $U_1 \mathrel{-\!\!\circ\!\!-} X_1Q \mathrel{-\!\!\circ\!\!-} X_2U_2Y$ and $U_2 \mathrel{-\!\!\circ\!\!-} X_2Q \mathrel{-\!\!\circ\!\!-} X_1U_1Y$, and*

$$\begin{aligned} R_1 &\geqslant I(U_1;X_1|Q), && (2.68)\\ R_1 &\geqslant I(U_2;X_2|Q), && (2.69)\\ R_1 + R_2 &\geqslant I(U_1U_2;X_1X_2|Q), && (2.70)\\ R_0 &\geqslant H(Y|U_1U_2Q). && (2.71)\end{aligned}$$

For $(X_1, X_2) \sim \mathrm{DSBS}(p)$, $Y := X_1 \oplus X_2$, where $p \in (0, 1)$ and $p \neq \frac{1}{2}$, we have $\overline{\mathcal{R}_{\mathrm{TO}}} \neq \mathcal{R}^*_{\mathrm{TO}}$.

2.3 RESULTS FROM ANALYSIS

We begin with the fundamental definitions.

Definition 2.34. *Let $\mathcal{U} \subseteq \mathbb{R}$ be an open, half-open or closed interval. A function $f\colon \mathcal{U} \to \mathbb{R}$, is called* convex *if and only if*

$$\lambda f(u) + \bar{\lambda} f(v) \geqslant f(\lambda u + \bar{\lambda} v) \qquad (2.72)$$

for any $u, v \in \mathcal{U}$ and $\lambda \in (0, 1)$. We call f strictly convex *if (2.72) is strict whenever $u \neq v$.*

Note that $\bar{\lambda} := 1 - \lambda$.

Definition 2.35.

1. *A set $\mathcal{S} \subseteq \mathbb{R}^n$ is* convex *if and only if, for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and $\lambda \in [0, 1]$, we have $\lambda\mathbf{x} + \bar{\lambda}\mathbf{y} \in \mathcal{S}$*
2. *A point $\mathbf{x} \in \mathcal{S}$ is an* extreme point *of the convex set $\mathcal{S} \subseteq \mathbb{R}^n$ if and only if $\mathbf{y}, \mathbf{z} \in \mathcal{S}$, $\lambda \in (0, 1)$, and $\lambda\mathbf{y} + \bar{\lambda}\mathbf{z} = \mathbf{x}$ imply $\mathbf{x} = \mathbf{y} = \mathbf{z}$. We write* $\mathrm{ext}(\mathcal{S})$ *for the set of all extreme points of $\mathcal{S}$.*
3. *The* convex hull *of a set $\mathcal{A} \subseteq \mathbb{R}^n$ is defined as the set $\mathrm{conv}(\mathcal{A}) := \bigcap\{\mathcal{B} \subseteq \mathbb{R}^n : \mathcal{B}$ convex$, \mathcal{A} \subseteq \mathcal{B}\}$.*
4. *A set $\mathcal{H} \subseteq \mathbb{R}^n$ is a* closed halfspace *if and only if there exist $\mathbf{x} \in \mathbb{R}^n$ and $a \in \mathbb{R}$, such that $\mathcal{H} = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{x}^T\mathbf{y} \leqslant a\}$.*
5. *A set $\mathcal{S} \subseteq \mathbb{R}^n$ is* line-free *if and only if it does not contain a line, i. e., a set of the form $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{a} + \lambda\mathbf{b}, \lambda \in \mathbb{R}\}$ with $\mathbf{b} \neq \mathbf{0}$.*

$\mathbf{0}$ is the all-zeros vector.

Definition 2.36. *For a convex set $\mathcal{S} \subseteq \mathbb{R}^n$ and $\mathbf{y} \in \mathcal{S}$, define the* characteristic cone *of $\mathcal{S}$ at $\mathbf{y}$ as $\mathrm{cc}_{\mathbf{y}}(\mathcal{S}) := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{y} + \lambda\mathbf{x} \in \mathcal{S}, \forall\lambda \in \mathbb{R}_+\}$. If additionally $\mathcal{S} \subseteq \mathbb{R}^n$ is closed, define the* characteristic cone *of $\mathcal{S}$ as $\mathrm{cc}(\mathcal{S}) := \mathrm{cc}_{\mathbf{y}}(\mathcal{S})$ for an arbitrary $\mathbf{y} \in \mathcal{S}$. Note that $\mathrm{cc}(\mathcal{S})$ is well defined for a closed convex set $\mathcal{S}$, as $\mathrm{cc}_{\mathbf{y}}(\mathcal{S})$ is independent of $\mathbf{y}$ by [24, Theorem 2.5.2].*

We collect several classical results from (convex) analysis in the following lemma.

Lemma 2.37.

1. *Let $\mathcal{X}, \mathcal{Y}$ be two metric spaces and $f\colon \mathcal{X} \to \mathcal{Y}$ continuous. If $\mathcal{E} \subseteq \mathcal{X}$ is connected [54, Definition 2.45], then $f(\mathcal{X})$ is connected.*

2. *Let $\mathcal{X}, \mathcal{Y}$ be two topological spaces and $f\colon \mathcal{X} \to \mathcal{Y}$ continuous. If $\mathcal{E} \subseteq \mathcal{X}$ is compact [54, Definition 2.32], then $f(\mathcal{X})$ is compact.*

3. *A set $\mathcal{S} \subseteq \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

4. *A compact, convex set $\mathcal{S} \subseteq \mathbb{R}^n$ is the convex hull of its extreme points, i. e., $\mathcal{S} = \operatorname{conv}\big(\operatorname{ext}(\mathcal{S})\big)$.*

5. *If $\mathcal{A} \subseteq \mathbb{R}^n$ is compact, then $\operatorname{conv}(\mathcal{A})$ is compact.*

Closed halfspaces are defined in part 4 of Definition 2.35.

6. *A closed, convex set $\mathcal{S} \subseteq \mathbb{R}^n$ is the intersection of all closed halfspaces containing $\mathcal{S}$, i. e., $\mathcal{S} = \bigcap\{\mathcal{H} \subseteq \mathbb{R}^n : \mathcal{H}$ is a closed halfspace, $\mathcal{S} \subseteq \mathcal{H}\}$.*

$\mathcal{A} + \mathcal{B}$ denotes the Minkowski addition of sets.

7. *For a line-free, closed, convex set $\mathcal{S} \in \mathbb{R}^n$, we have the identity $\mathcal{S} = \operatorname{cc}(\mathcal{S}) + \operatorname{conv}\big(\operatorname{ext}(\mathcal{S})\big)$.*

8. *For sets $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}^n$, we have $\operatorname{conv}(\mathcal{A} + \mathcal{B}) = \operatorname{conv}(\mathcal{A}) + \operatorname{conv}(\mathcal{B})$.*

9. *If $\mathcal{A} \subseteq \mathbb{R}^n$ is closed and $\mathcal{B} \subseteq \mathbb{R}^n$ is compact, then $\mathcal{A} + \mathcal{B}$ is closed.*

We will also note the following corollary of Lemma 2.37.

Corollary 2.38. *If $\mathcal{B} \subseteq \mathbb{R}^n$ is closed and convex, and $\mathcal{C} \subseteq \mathbb{R}^n$ is compact, we have*

$$\mathcal{A} := \operatorname{conv}(\mathcal{C} + \mathcal{B}) = \operatorname{conv}(\mathcal{C}) + \mathcal{B} = \overline{\mathcal{A}}. \tag{2.73}$$

Proof. We have $\mathcal{A} = \operatorname{conv}(\mathcal{C}) + \operatorname{conv}(\mathcal{B}) = \operatorname{conv}(\mathcal{C}) + \mathcal{B}$ by part 8 of Lemma 2.37 and the convexity of $\mathcal{B}$. Note that $\operatorname{conv}(\mathcal{C})$ is compact by part 5 of Lemma 2.37. $\mathcal{A}$ is therefore the sum of a compact set and a closed set and closed by part 9 of Lemma 2.37. □

The following theorem is an extension of Carathéodory's theorem, also known as the Fenchel-Eggleston-Carathéodory theorem [16, Appendix A].

Theorem 2.39. *If $\mathbf{x} \in \operatorname{conv}(\mathcal{A})$, where $\mathcal{A} \subseteq \mathbb{R}^n$, then there exists a set of points $\mathcal{X} \subseteq \mathcal{A}$, with $|\mathcal{X}| \leqslant n+1$ such that $\mathbf{x} \in \operatorname{conv}(\mathcal{X})$. If additionally $\mathcal{A}$ is a connected set, then the statement is true with $|\mathcal{X}| \leqslant n$.*

The following lemma collects elementary facts about convex/concave functions.

Lemma 2.40. *Let $f\colon \mathcal{U} \to \mathbb{R}$ be a continuous function, defined on the compact interval $\mathcal{U} = [u_1, u_2] \subset \mathbb{R}$. Assuming that f is twice differentiable on $\mathcal{V} := (u_1, u_2)$, the following properties hold.*

1. *If* $f''(u) \geqslant 0$ *for all* $u \in \mathcal{V}$, *and* $f'(u^*) = 0$ *for some* $u^* \in \mathcal{U}$, *then* $f(u) \geqslant f(u^*)$ *for all* $u \in \mathcal{U}$. *Furthermore, if additionally* $f''(u) > 0$ *for all* $u \in \mathcal{V}$, *then* $f(u) > f(u^*)$ *for all* $u \in \mathcal{U} \backslash \{u^*\}$.

2. *If* $f''(u) \leqslant 0$ *for all* $u \in \mathcal{V}$, *then* $f(u) \geqslant \min\{f(u_1), f(u_2)\}$ *for all* $u \in \mathcal{U}$. *Furthermore, if* $f''(u) < 0$ *for all* $u \in \mathcal{V}$, *then* $f(u) > \min\{f(u_1), f(u_2)\}$ *for all* $u \in \mathcal{V}$.

Proof. To show part 1, note that f is convex on $\mathcal{V}$ [53, Theorem I.12.C] which implies convexity on $\mathcal{U}$ by continuity. Thus, $f(u^*)$ is a global minimum [42, Exercise 1.5.1]. If additionally $f''(u) > 0$ for all $u \in \mathcal{V}$, then f is strictly convex on $\mathcal{V}$ [53, Theorem I.12.C]. This implies strict convexity of f on $\mathcal{U}$ [53, Problem I.11.A (4)] and $f(u^*)$ is the unique global minimum.

For part 2, note that $-f$ is convex on $\mathcal{V}$ and therefore also on $\mathcal{U}$ by continuity. Thus, for any $u \in \mathcal{V}$, we choose $\lambda \in (0, 1)$ such that $u = \lambda u_1 + \bar{\lambda} u_2$ and have

$$-f(u) \leqslant -\lambda f(u_1) - \bar{\lambda} f(u_2) \tag{2.74}$$

$$\leqslant \max\{-f(u_1), -f(u_2)\}. \tag{2.75}$$

If $f''(u) < 0$ for all $u \in \mathcal{V}$, then $-f$ is strictly convex on $\mathcal{U}$ [53, Problem I.11.A (4)] and the inequality in (2.74) is strict. □

Theorem 2.41 (Cauchy-Schwarz inequality). *If* $\mathcal{X}$ *is a real inner product space, then for any* $x, y \in \mathcal{X}$ *we have*

$$\langle x, y\rangle^2 \leqslant \langle x, x\rangle\langle y, y\rangle \tag{2.76}$$

with equality if and only if x *and* y *are linearly dependent.*

We obtain the Schwarz inequality as a simple corollary with $\mathcal{X} = \mathbb{R}^n$.

Corollary 2.42 (Schwarz inequality). *For two real vectors* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$\left(\sum_{i \in [n]} a_i b_i\right)^2 \leqslant \left(\sum_{i \in [n]} a_i^2\right)\left(\sum_{i \in [n]} b_i^2\right). \tag{2.77}$$

We also note the following elementary fact for later use.

Lemma 2.43. *For* $x \in (0, 1)$ *and* $y > 0$,

$$f(x, y) := \frac{1}{x^{-y} - 1} + \log(1 - x^y) > 0. \tag{2.78}$$

Proof. Fix $y > 0$ and observe that $\lim_{x \downarrow 0} f(x, y) = 0$. It then suffices to show that $f(x, y)$ increases in x:

$$\frac{\partial f}{\partial x}(x,y) = -\frac{1}{(x^{-y}-1)^2}(-y)x^{-y-1} + \frac{1}{1-x^y}(-y)x^{y-1} \tag{2.79}$$

$$= \frac{y}{x(x^{-y}-1)}\left(\frac{1}{1-x^y} - 1\right) > 0. \tag{2.80}$$

□

2.4 BOOLEAN FUNCTIONS

The material in this section follows [43].

Let X and Y be two dependent *Rademacher* random variables, i. e., $X, Y \sim \mathcal{U}(\{-1,1\})$ are both uniformly distributed on $\mathcal{X} = \mathcal{Y} = \{-1,1\}$. Define the correlation coefficient $\rho := \mathbb{E}[XY]$, and let the random vectors $(\mathbf{X}, \mathbf{Y})$ be n i.i.d. copies of (X, Y). Consider two real functions on the Hamming cube $f, g \colon \{-1,1\}^n \to \mathbb{R}$. The set of all such functions together with the inner product

$$\langle f, g \rangle := \mathbb{E}\left[f(\mathbf{X})g(\mathbf{X})\right] = 2^{-n} \sum_{\mathbf{x} \in \{-1,1\}^n} f(\mathbf{x})g(\mathbf{x}) \tag{2.81}$$

forms a real Hilbert space. An orthonormal basis [43, Theorem 1.5] of this space is given by the 2^n functions $\chi_{\mathcal{S}}(\mathbf{x}) := \prod_{i \in \mathcal{S}} x_i$ for $\mathcal{S} \subseteq [n]$.

Remark 3. We call the function $\chi_i(\mathbf{x}) = x_i$ for $i \in [n]$ the ith dictator function [43, Definition 2.3].

The Fourier-Walsh transform and the noise operator [43, Definition 2.46] are defined as follows.

Definition 2.44. *For a function $f \colon \{-1,1\}^n \to \mathbb{R}$, define the Fourier-Walsh transform*

$$\widehat{f}_{\mathcal{S}} := \langle f, \chi_{\mathcal{S}} \rangle \overset{(2.81)}{=} 2^{-n} \sum_{\mathbf{x} \in \{-1,1\}^n} f(\mathbf{x}) \prod_{i \in \mathcal{S}} x_i \tag{2.82}$$

for every $\mathcal{S} \subseteq [n]$.

The noise operator T_ρ with parameter $\rho \in [0,1]$ maps the function f to the function $T_\rho f \colon \{-1,1\}^n \to \mathbb{R}$, with $T_\rho f(\mathbf{x}) = \mathbb{E}[f(\mathbf{Y})|\mathbf{X} = \mathbf{x}]$ for all $\mathbf{x} \in \{-1,1\}^n$, where $\rho = \mathbb{E}[XY]$ is the correlation coefficient.

We collect some important properties of the Fourier expansion and the noise operator in the following lemma.

Lemma 2.45. *For any two function $f, g \colon \{-1,1\}^n \to \mathbb{R}$, the following properties hold for all $\mathbf{x} \in \{-1,1\}^n$.*

1. $f(\mathbf{x}) = \sum_{\mathcal{S} \subseteq [n]} \widehat{f}_{\mathcal{S}} \chi_{\mathcal{S}}(\mathbf{x})$
2. $\langle f, g \rangle = \sum_{\mathcal{S} \subseteq [n]} \widehat{f}_{\mathcal{S}} \widehat{g}_{\mathcal{S}}$

3. $\widehat{T_\rho f}_S = \rho^{|S|}\widehat{f}_S$

4. $\langle f, T_\rho g\rangle = \langle T_\rho f, g\rangle = \left\langle T_{\sqrt{\rho}} f, T_{\sqrt{\rho}} g\right\rangle = \sum_{S\subseteq[n]} \rho^{|S|}\widehat{f}_S\widehat{g}_S$

If $f, g\colon \{-1,1\}^n \to \{-1,1\}$, *then the following properties also hold.*

5. $\sum_{S\subseteq[n]} \widehat{f}_S^2 = 1$

6. $\langle f, T_\rho g\rangle = 2P\{f(\mathbf{X}) = g(\mathbf{Y})\} - 1$

Proof. Part 1 is the Fourier expansion theorem [43, Theorem 1.1] and part 2 is Plancherel's Theorem [43, Section 1.4]. Part 3 is the Fourier representation of the noise operator [43, Proposition 2.47] and part 4 is a direct consequence of parts 2 and 3. Part 5 holds as $\sum_{S\subseteq[n]} \widehat{f}_S^2 = \langle f, f\rangle = \mathbb{E}\left[f(\mathbf{X})^2\right] = 1$ by part 2. To prove part 6, note that

$$\begin{aligned}
\langle f, T_\rho g\rangle &\overset{(2.81)}{=} \mathbb{E}[f(\mathbf{X})T_\rho g(\mathbf{X})] && (2.83)\\
&= \mathbb{E}\left[f(\mathbf{X})\mathbb{E}[g(\mathbf{Y})|\mathbf{X}]\right] && (2.84)\\
&= \mathbb{E}\left[f(\mathbf{X})g(\mathbf{Y})\right] && (2.85)\\
&= P\{f(\mathbf{X}) = g(\mathbf{Y})\} - P\{f(\mathbf{X}) \neq g(\mathbf{Y})\} && (2.86)\\
&= 2P\{f(\mathbf{X}) = g(\mathbf{Y})\} - 1. && (2.87)
\end{aligned}$$

□

Lemma 2.46. *Let* $f\colon \{-1,1\}^n \to \{-1,1\}$ *with* $\widehat{f}_\varnothing = 0$ *and* $\rho \in (0,1)$, *then* $\langle f, T_\rho f\rangle \leqslant \rho$, *with equality if and only if* $f = \pm\chi_i$ *for some* $i \in [n]$.

Lemma 2.47. *Let* $f, g\colon \{-1,1\}^n \to \{-1,1\}$ *with* $\widehat{f}_\varnothing = \widehat{g}_\varnothing = 0$ *and* $\rho \in (0,1)$. *If* $|\langle f, T_\rho g\rangle| = \rho$, *then* $f = \pm g = \pm\chi_i$ *for some* $i \in [n]$.

Proof. Note that

$$\begin{aligned}
\rho^2 &= \langle f, T_\rho g\rangle^2 && (2.88)\\
&= \left\langle T_{\sqrt{\rho}} f, T_{\sqrt{\rho}} g\right\rangle^2 && (2.89)\\
&\leqslant \left\langle T_{\sqrt{\rho}} f, T_{\sqrt{\rho}} f\right\rangle\left\langle T_{\sqrt{\rho}} g, T_{\sqrt{\rho}} g\right\rangle && (2.90)\\
&= \langle f, T_\rho f\rangle\langle g, T_\rho g\rangle && (2.91)\\
&\leqslant \rho^2, && (2.92)
\end{aligned}$$

where (2.89) and (2.91) follow from part 4 of Lemma 2.45. We applied Theorem 2.41 in (2.90) and (2.92) follows from Lemma 2.46. We have equality in (2.90) and thus $T_{\sqrt{\rho}} g = \lambda T_{\sqrt{\rho}} f$ for some $\lambda \in \mathbb{R}$ by Theorem 2.41. Consequently, $g = \lambda f$, which is only possible for $\lambda = \pm 1$. We also have equality in (2.92) and thus, $\langle f, T_\rho f\rangle = \langle g, T_\rho g\rangle = \rho$, yielding $g = \pm\chi_i$ by Lemma 2.46. □

2.5 SUBMODULAR FUNCTIONS

The material in this section is mainly from [18]. We will, however, not require the results in [18] in full generality and shall therefore only study the relevant special cases.

In the following let $\mathcal{E}$ be a fixed finite set.

$2^{\mathcal{E}}$ denotes the power set of $\mathcal{E}$.

Note that $(2^{\mathcal{E}}, f)$ is simple, *in the sense of [18, Section 3.2].*

Definition 2.48. *A function $f\colon 2^{\mathcal{E}} \to \mathbb{R}$ is* submodular *on $2^{\mathcal{E}}$ if and only if $f(\varnothing) = 0$ and*

$$f(\mathcal{A}) + f(\mathcal{B}) \geqslant f(\mathcal{A} \cup \mathcal{B}) + f(\mathcal{A} \cap \mathcal{B}) \quad \forall \mathcal{A}, \mathcal{B} \subseteq \mathcal{E}. \tag{2.93}$$

The submodular polyhedron *and the* base polyhedron *in $|\mathcal{E}|$ dimensional space $\mathbb{R}^{\mathcal{E}}$ are defined by*

$$P(f) := \left\{ x \in \mathbb{R}^{\mathcal{E}} : \sum_{e \in \mathcal{A}} x_e \leqslant f(\mathcal{A}), \forall \mathcal{A} \subseteq \mathcal{E} \right\}, \text{ and} \tag{2.94}$$

$$B(f) := \left\{ x \in P(f) : \sum_{e \in \mathcal{E}} x_e = f(\mathcal{E}) \right\}, \tag{2.95}$$

respectively. A function f is supermodular *if and only if $-f$ is submodular. The corresponding* supermodular polyhedron *and* base polyhedron *are defined as $P(f) := -P(-f)$ and $B(f) := -B(-f)$.*

Lemma 2.49. *The base polyhedron $B(f)$ of a submodular (supermodular) function f is compact.*

Proof. $B(f)$ is bounded by [18, Theorem 3.12] and closed by definition, i. e., compact by part 3 of Lemma 2.37. □

Lemma 2.50. *For a submodular (supermodular) function f on $2^{\mathcal{E}}$, the identity $P(f) = B(f) - \mathbb{R}^{\mathcal{E}}_{+}$ ($P(f) = B(f) + \mathbb{R}^{\mathcal{E}}_{+}$) holds.*

$x \leqslant y$ if and only if $x_e \leqslant y_e, \forall e \in \mathcal{E}$.

Proof. It suffices to prove the statement for submodular functions. From the definitions (2.94) and (2.95), $B(f) - \mathbb{R}^{\mathcal{E}}_{+} \subseteq P(f)$. By [18, Theorem 2.3], for each $x \in P(f)$, there exists $y \in B(f)$ with $x \leqslant y$, which finishes the proof. □

$\sqsubset e := \{e' : e' \sqsubset e\}$ and $\sqsubseteq e$ accordingly.

Theorem 2.51 (Extreme point theorem)**.** *For a submodular (supermodular) function f on $2^{\mathcal{E}}$ and $x \in B(f)$, we have $x \in \operatorname{ext}(B(f))$ if and only if there is a total order $\sqsubset$ of $\mathcal{E}$, such that $x_e = f(\sqsubset e) - f(\sqsubseteq e)$ for all $e \in \mathcal{E}$.*

2.6 CONVEX POLYHEDRA

We will conclude this chapter with some technical results on convex polyhedra, noted here for further use in Section 5.4.

Let $\mathcal{H}$ be the closed, convex polyhedron $\mathcal{H} := \{x \in \mathbb{R}^n : \mathbf{A}x \geqslant \mathbf{b}\}$ for an $m \times n$ matrix $\mathbf{A} = (\mathbf{a}_{(1)}, \mathbf{a}_{(2)}, \dots, \mathbf{a}_{(m)})^T$ and $\mathbf{b} \in \mathbb{R}^m$, where $\mathbf{a}^T_{(j)}$ is the jth row of $\mathbf{A}$.

Lemma 2.52. *For $\mathbf{x} \in \mathbb{R}^n$ we have $\mathbf{x} \in \mathrm{cc}(\mathcal{H})$ if and only if $\mathbf{A}\mathbf{x} \geqslant 0$.*

Proof. If $\mathbf{A}\mathbf{x} \geqslant 0$, $\mathbf{y} \in \mathcal{H}$ and $\lambda \geqslant 0$, $\mathbf{A}(\mathbf{y}+\lambda\mathbf{x}) \geqslant \mathbf{A}\mathbf{y} \geqslant \mathbf{b}$. On the other hand, if $\mathbf{a}_{(i)}^T\mathbf{x} < 0$, we have $\mathbf{a}_{(i)}^T(\mathbf{y}+\lambda\mathbf{x}) < b_i$ for $\lambda > \frac{b_i - \mathbf{a}_{(i)}^T\mathbf{y}}{\mathbf{a}_{(i)}^T\mathbf{x}} > 0$. □

Lemma 2.53. *If, for every $i \in [n]$, there exists $j \in [m]$ such that $\mathbf{a}_{(j)} = \mathbf{e}_i$ and $\mathbf{a}_{(j)} \geqslant 0$ for every $j \in [m]$, then $\mathcal{H}$ is line-free and $\mathrm{cc}(\mathcal{H}) = \mathbb{R}^n_+$.*

$\mathbf{e}_i$ is the ith canonical basis vector.

Proof. For any $\mathbf{y} \in \mathbb{R}^n_+$, clearly $\mathbf{A}\mathbf{y} \geqslant 0$ and hence $\mathbf{y} \in \mathrm{cc}(\mathcal{H})$ by Lemma 2.52. If $\mathbf{y} \notin \mathbb{R}^n_+$, let $y_i < 0$ and choose $j \in [m]$ such that $\mathbf{a}_{(j)} = \mathbf{e}_i$. We have $\mathbf{a}_{(j)}^T\mathbf{y} = y_i < 0$ and therefore $\mathbf{y} \notin \mathrm{cc}(\mathcal{H})$ by Lemma 2.53.

To show that $\mathcal{H}$ is line-free assume that $\mathbf{x}+\lambda\mathbf{y} \in \mathcal{H}$ for all $\lambda \in \mathbb{R}$. This implies $\pm\mathbf{y} \in \mathrm{cc}(\mathcal{H}) = \mathbb{R}^n_+$, i.e., $\mathbf{y} = 0$. □

Definition 2.54. *A point $\mathbf{x}$ is on an* extreme ray *of the cone $\mathrm{cc}(\mathcal{H})$ if and only if the decomposition $\mathbf{x} = \mathbf{y}+\mathbf{z}$ with $\mathbf{y}, \mathbf{z} \in \mathrm{cc}(\mathcal{H})$ implies that $\mathbf{y} = \lambda\mathbf{z}$ for some $\lambda \in \mathbb{R}$.*

It is easy to see that the points on extreme rays of $\mathbb{R}^n_+$ are $\lambda\mathbf{e}_i$ for any $\lambda \geqslant 0$ and $i \in [n]$.

Definition 2.55. *For $\mathbf{x} \in \mathcal{H}$, we define the number of inequalities that are satisfied with equality at $\mathbf{x}$ as $N_{\mathcal{H}}(\mathbf{x}) := \mathrm{rank}(\mathbf{A}_\mathbf{x})$, where $\mathbf{A}_\mathbf{x} := (\mathbf{a}_{(j)}^T)_{j\in[m]:\mathbf{a}_{(j)}^T\mathbf{x}=b_j}$.*

Lemma 2.56. *For $\mathbf{x} \in \mathcal{H}$, we have $\mathbf{x} \in \mathrm{ext}(\mathcal{H})$ if and only if $N_{\mathcal{H}}(\mathbf{x}) = n$.*

Proof. Assuming that $N_{\mathcal{H}}(\mathbf{x}) < n$, we find $0 \neq \mathbf{r} \in \ker(\mathbf{A}_\mathbf{x})$ and thus $\mathbf{x} \pm \varepsilon\mathbf{r} \in \mathcal{H}$ for some $\varepsilon > 0$, showing that $\mathbf{x} \notin \mathrm{ext}(\mathcal{H})$.

Conversely, let without loss of generality $N_{\mathcal{H}}(\mathbf{x}) \neq 0$ and assume $\mathbf{x} \notin \mathrm{ext}(\mathcal{H})$, i.e., $\mathbf{x} = \lambda\mathbf{x}_1 + \bar{\lambda}\mathbf{x}_2$ for $\lambda \in (0,1)$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{H}$, $\mathbf{x}_1 \neq \mathbf{x}_2$. With $\mathbf{b}_\mathbf{x} := (b_j)_{j\in[m]:\mathbf{a}_{(j)}^T\mathbf{x}=b_j}$, we have $\mathbf{A}_\mathbf{x}(\lambda\mathbf{x}_1 + \bar{\lambda}\mathbf{x}_2) = \mathbf{b}_\mathbf{x}$, which implies $\mathbf{A}_\mathbf{x}\mathbf{x}_1 = \mathbf{A}_\mathbf{x}\mathbf{x}_2 = \mathbf{b}_\mathbf{x}$ and therefore $0 \neq \mathbf{x}_1 - \mathbf{x}_2 \in \ker(\mathbf{A}_\mathbf{x})$. □

Lemma 2.57. *Let $\mathbf{x} \in \mathcal{H}$ with $N_{\mathcal{H}}(\mathbf{x}) = n-1$ and assume that $\mathcal{H}$ is line-free. Then either $\mathbf{x} = \lambda\mathbf{x}_1 + \bar{\lambda}\mathbf{x}_2$ where $\lambda \in (0,1)$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathrm{ext}(\mathcal{H})$ or $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ where $\mathbf{x}_1 \in \mathrm{ext}(\mathcal{H})$ and $\mathbf{x}_2 \neq 0$ lies on an extreme ray of $\mathrm{cc}(\mathcal{H})$.*

Proof. We obtain $0 \neq \mathbf{r} \in \ker(\mathbf{A}_\mathbf{x})$. Define $\lambda_1 := \inf\{\lambda : \mathbf{x}+\lambda\mathbf{r} \in \mathcal{H}\}$ and $\lambda_2 := \sup\{\lambda : \mathbf{x}+\lambda\mathbf{r} \in \mathcal{H}\}$. Clearly $\lambda_1 \leqslant 0 \leqslant \lambda_2$. As $\mathcal{H}$ is line-free, we may assume without loss of generality $\lambda_1 = -1$ and set $\mathbf{x}_1 = \mathbf{x} - \mathbf{r}$. We now have $\mathbf{x}_1 \in \mathrm{ext}(\mathcal{H})$ as otherwise $\mathbf{x}_1 - \varepsilon\mathbf{r} \in \mathcal{H}$ for some small $\varepsilon > 0$.

If $\lambda_2 < \infty$, define $\mathbf{x}_2 = \mathbf{x} + \lambda_2\mathbf{r}$ which yields $\mathbf{x}_2 \in \mathrm{ext}(\mathcal{H})$ and $\mathbf{x} = \lambda\mathbf{x}_1 + \bar{\lambda}\mathbf{x}_2$ with $\lambda = \frac{\lambda_2}{\lambda_2+1}$. Note that $\lambda_2 \neq 0$ as $\mathbf{x} \notin \mathrm{ext}(\mathcal{H})$.

If $\lambda_2 = \infty$ we have $\mathbf{x} - \mathbf{x}_1 = \mathbf{r} \in \mathrm{cc}_{\mathbf{x}_1}(\mathcal{H}) = \mathrm{cc}(\mathcal{H})$. We need to show that $\mathbf{r}$ is also on an extreme ray of $\mathrm{cc}(\mathcal{H})$. Assume $\mathbf{r} = \mathbf{r}_1 + \mathbf{r}_2$ with $\mathbf{r}_1, \mathbf{r}_2 \in \mathrm{cc}(\mathcal{H})$. Then $\mathbf{A}_\mathbf{x}\mathbf{r}_1 + \mathbf{A}_\mathbf{x}\mathbf{r}_2 = 0$ and hence, $\mathbf{r}_1, \mathbf{r}_2 \in \ker(\mathbf{A}_\mathbf{x})$ by

Lemma 2.52. Noting that $\ker(\mathbf{A}_x)$ is a one-dimensional space finishes the proof. □

$[0:J] = \{0, 1, \ldots, J\}$.

Let $J \in \mathbb{N}$ and $K := J+1$. For each $j \in [0:J]$, define the closed convex polyhedron $\mathcal{H}^{(j)} := \{\mathbf{x} \in \mathbb{R}^{K+j} : \mathbf{A}^{(j)}\mathbf{x} \geqslant \mathbf{b}^{(j)}\}$, where $\mathbf{A}^{(j)}$ is a matrix and $\mathbf{b}^{(j)}$ a vector of appropriate dimension. We make the following three assumptions:

1. $\mathbf{A}^{(j)}$ and $\mathbf{b}^{(j)}$ are defined recursively as

$$\mathbf{A}^{(j)} := \begin{pmatrix} \mathbf{A}^{(j-1)} & 0 \\ 0^T & 1 \\ \mathbf{e}_j^T & 1 \end{pmatrix}, \qquad \mathbf{b}^{(j)} = \begin{pmatrix} \mathbf{b}^{(j-1)} \\ c_1^{(j)} \\ c_2^{(j)} \end{pmatrix}, \tag{2.96}$$

for $j \in [J]$, where $\mathbf{e}_j$ is the jth unit vector of appropriate dimension and $c_1^{(j)}, c_2^{(j)} \in \mathbb{R}$ are arbitrary.

2. Each entry of $\mathbf{A}^{(0)}$ equals 0 or 1 and for all $k \in [K]$ at least one row of $\mathbf{A}^{(0)}$ is equal to $\mathbf{e}_k^T$. Due to assumption 1, this also implies that each entry of $\mathbf{A}^{(j)}$ is either 0 or 1 and for all $k \in [K+j]$ at least one row of $\mathbf{A}^{(j)}$ is equal to $\mathbf{e}_k^T$.

3. For any extreme point $\mathbf{x} \in \operatorname{ext}(\mathcal{H}^{(0)})$ and any $j \in [J]$, assume $x_j \leqslant c_2^{(j)} - c_1^{(j)}$.

Lemma 2.58. *Under assumptions 1 to 3, for every $k \in [0:J]$ and every extreme point $\mathbf{y} \in \operatorname{ext}(\mathcal{H}^{(k)})$ there is an extreme point $\mathbf{x} \in \operatorname{ext}(\mathcal{H}^{(0)})$ and a subset $\mathcal{E}_k \subseteq [k]$ such that $y_K = x_K$ and for every $j \in [J]$,*

$$y_j = \begin{cases} x_j, & j \notin \mathcal{E}_k, \\ c_2^{(j)} - c_1^{(j)}, & j \in \mathcal{E}_k, \end{cases} \tag{2.97}$$

and for every $j \in [k]$,

$$y_{K+j} = \begin{cases} c_2^{(j)} - x_j, & j \notin \mathcal{E}_k, \\ c_1^{(j)}, & j \in \mathcal{E}_k. \end{cases} \tag{2.98}$$

Proof. For every $k \in [0:J]$, $\mathcal{H}^{(k)}$ is line-free by assumption 2 and Lemma 2.53, and can be written as $\mathcal{H}^{(k)} = \operatorname{cc}(\mathcal{H}^{(k)}) + \operatorname{conv}\big(\operatorname{ext}(\mathcal{H}^{(k)})\big)$ by Part 7 of Lemma 2.37. Lemma 2.53 also implies $\operatorname{cc}(\mathcal{H}^{(k)}) = \mathbb{R}_+^{K+k}$.

Let us proceed inductively over $k \in [0:J]$. For $k = 0$ the statement is trivial. Given any $\mathbf{y} \in \operatorname{ext}(\mathcal{H}^{(k)})$, we need to obtain $\mathbf{x} \in \operatorname{ext}(\mathcal{H}^{(0)})$ and $\mathcal{E}_k$ such that $\mathbf{y}$ is given according to (2.97) and (2.98). Let $\mathbf{z} = \mathbf{y}_1^{K+k-1}$ be the truncation of $\mathbf{y}$. We have $N_{\mathcal{H}^{(k)}}(\mathbf{y}) = K+k$ by Lemma 2.56, which is possible in only two different ways:

- *Construction I:* We assume that $N_{\mathcal{H}^{(k-1)}}(\boldsymbol{z}) = K+k-1$, i. e., $\boldsymbol{z} \in \operatorname{ext}(\mathcal{H}^{(k-1)})$ by Lemma 2.56, and at least one of

  $$y_{K+k} \geqslant c_1^{(k)}, \tag{2.99}$$

  $$y_k + y_{K+k} \geqslant c_2^{(k)}, \tag{2.100}$$

  is satisfied with equality.

  As $\boldsymbol{z} \in \operatorname{ext}(\mathcal{H}^{(k-1)})$, there exists $\boldsymbol{x} \in \operatorname{ext}(\mathcal{H}^{(0)})$ and $\mathcal{E}_{k-1}$ such that (2.97) holds for every $j \in [J]$ and (2.98) holds for $j \in [k-1]$ by the induction hypothesis. In particular $y_k = x_k$. Assuming that (2.100) holds with equality, we have $y_{K+k} = c_2^{(k)} - x_k$. Thus, the point $\boldsymbol{x}$ together with $\mathcal{E}_k = \mathcal{E}_{k-1}$ yields $\boldsymbol{y}$ by (2.97) and (2.98). Equality in (2.99) implies equality in (2.100) by assumption 3.

- *Construction II:* We assume that both (2.99) and (2.100) are satisfied with equality and we have $N_{\mathcal{H}^{(k-1)}}(\boldsymbol{z}) = K+k-2$. This can occur in two different ways by Lemma 2.57.

  First consider the case where $\boldsymbol{z} = \lambda\boldsymbol{x} + \bar{\lambda}\hat{\boldsymbol{x}}$ for $\boldsymbol{x}, \hat{\boldsymbol{x}} \in \operatorname{ext}(\mathcal{H}^{(k-1)})$, $\boldsymbol{x} \neq \hat{\boldsymbol{x}}$ and $\lambda \in (0,1)$. This implies $y_{K+k} = c_1^{(k)}$ and $y_k = \lambda x_k + \bar{\lambda} x_k' = c_2^{(k)} - c_1^{(k)}$, which by assumption 3 leads to $x_k = x_k' = c_2^{(k)} - c_1^{(k)}$. Thus, (2.99) and (2.100) are satisfied (with equality) for every $\lambda \in [0,1]$ and $\boldsymbol{y}$ cannot be an extreme point as it can be written as a non-trivial convex combination.

  We can thus focus on the second option which is that $\boldsymbol{z}$ is on an extreme ray of $\mathcal{H}^{(k-1)}$, i. e., $\boldsymbol{z} = \boldsymbol{x} + \lambda \boldsymbol{e}_{k'}$ for some $\boldsymbol{x} \in \operatorname{ext}(\mathcal{H}^{(k-1)})$, $\lambda > 0$ and $k' \in [K+k-1]$. If $k' \neq k$, (2.99) and (2.100) are satisfied for all $\lambda > 0$ and thus $\boldsymbol{y}$ cannot be an extreme point because it can be written as a non-trivial convex combination. For $k' = k$, the point $\boldsymbol{x}$ with $\mathcal{E}_k = \mathcal{E}_{k-1} \cup k$ yields the desired extreme point. □

2.7 REFERENCES

If a result in this chapter was given without proof, Table 1 lists a reference to the literature, where a proof can be found.

RESULT	REFERENCE
Part 1 of Lemma 2.4	[12, Lemma 2.1.1, Theorem 2.6.4]
Part 2 of Lemma 2.4	[12, Equation (2.92)]
Part 3 of Lemma 2.4	[12, Theorem 2.5.1]
Part 4 of Lemma 2.4	[12, Theorem 2.5.2]
Theorem 2.5	[12, Theorem 2.8.1]

RESULT	REFERENCE
Theorem 2.6	[12, Theorem 2.8.1]
Theorem 2.7	[12, Theorem 2.7.1]
Theorem 2.9	[12, Theorem 2.6.3]
Theorem 2.10	[16, Section 2.1]
Lemma 2.15	[13, Lemma 2.2]
Lemma 2.16	[13, Lemmas 2.5 and 2.6]
Lemmas 2.18 and 2.19	[16, Sections 2.4 and 2.5]
Lemma 2.20	[27, Lemma 3.4]
Theorem 2.23	[25, Corollary 6]
Theorem 2.25	[66, Appendix A]
Theorem 2.26	[66, Appendix B]
Theorem 2.29	[11, Appendix B]
Theorem 2.32	[34, Theorem 1]
Theorem 2.33	[34, Proposition 1]
Part 1 of Lemma 2.37	[54, Theorem 4.22]
Part 2 of Lemma 2.37	[40, Lemma 26.5]
Part 3 of Lemma 2.37	[54, Theorem 2.41]
Part 4 of Lemma 2.37	[53, Theorem III.32.D]
Part 5 of Lemma 2.37	[4, Corollary 5.33]
Part 6 of Lemma 2.37	[24, Theorem 2.2.3]
Part 7 of Lemma 2.37	[24, Theorem 2.5.6]
Part 8 of Lemma 2.37	[56, Theorem 1.1.2]
Part 9 of Lemma 2.37	[55, Exercise 1.3(e)]
Theorem 2.39	[15, Theorem 18]
Theorem 2.41	[4, Lemma 6.49]
Corollary 2.42	[54, Theorem 1.35]
Theorem 2.51	[18, Theorem 3.22]
Lemma 2.46	[43, Proposition 2.50]

Table 1: References to proofs.

Part II

INFORMATION THEORETIC CLUSTERING

In this part we investigate two different clustering problems and connect them to well-known problems in the information theory literature. Chapter 3 focuses on the multi-clustering problem with two sources. We provide bounds for the resulting achievable region and investigate special cases. In Chapter 4 we positively resolve the two function case of the Kumar-Courtade conjecture concerning the mutual information between Boolean functions. Finally in Chapter 5, we revisit the multi-clustering problem and extend it to multiple sources. In particular, we introduce a multiple description CEO problem and provide a single-letter characterization of its achievable region under a suitable Markov constraint.

3 CLUSTERING WITH TWO SOURCES

In this chapter we will investigate the case of two sources in the multi-clustering problem, which was informally introduced in Section 1.1.

3.1 PROBLEM STATEMENT

A schematic overview is given in Figure 4. In order to keep notation simpler, we use X and Z for the random variables and (f, g) for the code.

Definition 3.1. *For an* (n, R_1, R_2) *code* (f, g) *for* (X, Z)*, we define the* co-information of f and g *as*

Codes *are defined in Definition 2.21.*

$$\Theta(f; g) := \frac{1}{n} I\big(f(\mathbf{X}); g(\mathbf{Z})\big). \tag{3.1}$$

This co-information serves as a measure of the mutual relevance of the two encodings $f(\mathbf{X})$ and $g(\mathbf{Z})$. The idea is to find functions f and g that extract a compressed version of the common randomness in the observed data $\mathbf{X}$ and $\mathbf{Z}$.

Definition 3.2. *A triple* $(\mu, R_1, R_2) \in \mathbb{R}^3$ *is* achievable *for the source* (X, Z) *if and only if, for some* $n \in \mathbb{N}$*, there exists an* (n, R_1, R_2) *code* (f, g) *for* (X, Z) *such that*

$$\Theta(f; g) \geqslant \mu. \tag{3.2}$$

The achievable region $\overline{\mathcal{R}}$ *is defined as the closure of the set* $\mathcal{R}$ *of achievable triples.*

Remark 4. Note that a standard time-sharing argument shows that $\overline{\mathcal{R}}$ is a convex set (see, e. g., [16, Section 4.4]).

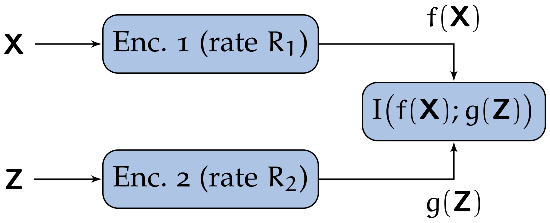


Figure 4: Clustering with two sources.

We also point out that stochastic encodings cannot enlarge the achievable region as any stochastic encoding can be represented as the convex combination of deterministic encodings and $\overline{\mathcal{R}}$ is convex.

3.2 CONNECTION WITH OTHER PROBLEMS

The multi-clustering problem with two sources turns out to be connected to a hypothesis testing problem and a pattern recognition problem. In this section we will clarify these connections explicitly, using the "multi-letter" region $\mathcal{R}_*$.

Definition 3.3. *Let $\mathcal{R}_*$ be the set of triples (μ, R_1, R_2) such that there exist $n \in \mathbb{N}$ and random variables U, V satisfying $U \mathrel{-\!\!\circ\!\!-} \mathbf{X} \mathrel{-\!\!\circ\!\!-} \mathbf{Z} \mathrel{-\!\!\circ\!\!-} V$ and*

$$
\begin{aligned}
nR_1 &\geqslant I(U;\mathbf{X}), && (3.3)\\
nR_2 &\geqslant I(V;\mathbf{Z}), \text{ and} && (3.4)\\
n\mu &\leqslant I(U;V). && (3.5)
\end{aligned}
$$

We will now show that $\overline{\mathcal{R}_*}$ is in fact the achievable region of the multi-clustering problem, a hypothesis testing problem, and a pattern recognition problem. As a first step, we show that $\mathcal{R}_*$ is the multi-letter region of the hypothesis testing problem introduced in Section 2.2.2.

Theorem 3.4. $\overline{\mathcal{R}_{\mathrm{HT}}} = \overline{\mathcal{R}_*}$.

Proof. Assume $(\mu, R_1, R_2) \in \mathcal{R}_{\mathrm{HT}}$. For $\varepsilon > 0$, pick an (n, R_1, R_2) hypothesis test $(f_n, g_n, \mathcal{A}_n)$ such that $\alpha_n \leqslant \varepsilon$ and $\log \beta_n \leqslant -n(\mu - \varepsilon)$. The random variables $U := f_n(\mathbf{X})$ and $V := g_n(\mathbf{Z})$ satisfy the required Markov chain as well as (3.3) and (3.4). We apply the log-sum inequality and obtain for any $\varepsilon' > 0$, provided that ε is small enough and n is large enough,

$$
\begin{aligned}
I(U;V) &= \sum_{u,v \in \mathcal{M}_1 \times \mathcal{M}_2} p_{UV}(u,v) \log_2 \frac{p_{UV}(u,v)}{p_U(u)p_V(v)} && (3.6)\\
&= \sum_{u,v \in \mathcal{A}_n} p_{UV}(u,v) \log_2 \frac{p_{UV}(u,v)}{p_U(u)p_V(v)} \\
&\qquad + \sum_{u,v \in \mathcal{A}_n^c} p_{UV}(u,v) \log_2 \frac{p_{UV}(u,v)}{p_U(u)p_V(v)} && (3.7)\\
&\geqslant (1-\alpha_n) \log \frac{1-\alpha_n}{\beta_n} + \alpha_n \log \frac{\alpha_n}{1-\beta_n} && (3.8)\\
&= -H(\alpha_n) + (1-\alpha_n) \log \frac{1}{\beta_n} + \alpha_n \log \frac{1}{1-\beta_n} && (3.9)\\
&\geqslant -(1-\varepsilon) \log \beta_n - \varepsilon' && (3.10)\\
&\geqslant (1-\varepsilon) n(\mu - \varepsilon) - \varepsilon' && (3.11)\\
&\geqslant n(\mu - \varepsilon'), && (3.12)
\end{aligned}
$$

where Theorem 2.7 was applied twice in (3.7). This shows that $(\mu - \varepsilon', R_1, R_2) \in \mathcal{R}_*$ and consequently $(\mu, R_1, R_2) \in \overline{\mathcal{R}_*}$.

The bound in Theorem 2.23 shows that $(n\mu, nR_1, nR_2)$ is asymptotically HT-achievable for the vector source $(\mathbf{X}, \mathbf{Z})$ if $(\mu, R_1, R_2) \in \mathcal{R}_*$. I. e., for any $\varepsilon, \varepsilon' > 0$, there is a sequence of $(k, nR_1 + \varepsilon', nR_2 + \varepsilon')$ hypothesis tests $(f_k, g_k, \mathcal{A}_k)$ for $(\mathbf{X}, \mathbf{Z})$, $k \in \mathbb{N}$ such that

$$\lim_{k\to\infty} \alpha_k \leqslant \varepsilon, \text{ and} \tag{3.13}$$

$$\lim_{k\to\infty} -\frac{1}{k}\log\beta_k \geqslant n\mu - \varepsilon'. \tag{3.14}$$

This shows that $\left(\mu - \frac{\varepsilon'}{n}, R_1 + \frac{\varepsilon'}{n}, R_2 + \frac{\varepsilon'}{n}\right) \in \mathcal{R}_{\mathrm{HT}}$ for the source (X, Z) and as ε' was arbitrary, this completes the proof. □

We can leverage this equivalence to show that Definition 3.3 is indeed a multi-letter characterization of $\overline{\mathcal{R}}$.

Corollary 3.5. $\overline{\mathcal{R}} = \overline{\mathcal{R}_*}$.

Proof. To prove $\mathcal{R} \subseteq \mathcal{R}_*$, assume $(\mu, R_1, R_2) \in \mathcal{R}$ and choose n, f, and g according to Definition 3.2. Defining $U := f(\mathbf{X})$ and $V := g(\mathbf{Z})$ yields inequalities (3.3)–(3.5) and satisfies the required Markov chain.

We will show $\overline{\mathcal{R}_{\mathrm{HT}}} \subseteq \overline{\mathcal{R}}$, which is equivalent to $\overline{\mathcal{R}_*} \subseteq \overline{\mathcal{R}}$ by Theorem 3.4. Assuming $(\mu, R_1, R_2) \in \mathcal{R}_{\mathrm{HT}}$, choose an arbitrary $\varepsilon > 0$ and pick an (n, R_1, R_2) hypothesis test $(f_n, g_n, \mathcal{A}_n)$ such that $\alpha_n \leqslant \varepsilon$ and $-\log\beta_n \geqslant n(\mu - \varepsilon)$. Pick $\varepsilon' > 0$ and apply the same reasoning as in (3.12). Provided that ε is small enough and n is large enough, the (n, R_1, R_2) code (f_n, g_n) achieves $\Theta(f_n; g_n) \geqslant \mu - \varepsilon'$, implying $(\mu, R_1, R_2) \in \overline{\mathcal{R}}$. □

The multi-clustering problem and the pattern recognition problem given in Section 2.2.2, also share a multi-letter region.

Proposition 3.6. $\overline{\mathcal{R}_{\mathrm{PR}}} = \overline{\mathcal{R}_*}$.

Proof. Assume $(\mu, R_1, R_2) \in \mathcal{R}_{\mathrm{PR}}$ and for an arbitrary $\varepsilon > 0$ and sufficiently large $n \in \mathbb{N}$ choose appropriate functions f, g, ϕ satisfying (2.52). The random variables $U := f(\mathbf{X})$ and $V := g(\mathbf{Z})$ satisfy the required Markov chain as well as (3.3) and (3.4). Furthermore,

$$\begin{aligned}
I(U;V) &= I\big(f(\mathbf{X}); g(\mathbf{Z})\big) && (3.15)\\
&= I\big(C; g(\mathbf{Z}(W))\big|W\big) && (3.16)\\
&= I\big(C; g(\mathbf{Z}(W)), W\big) && (3.17)\\
&\geqslant I\big(C; W\big|g(\mathbf{Z}(W))\big) && (3.18)\\
&= H\big(W\big|g(\mathbf{Z}(W))\big) - H\big(W\big|C, g(\mathbf{Z}(W))\big) && (3.19)\\
&\geqslant n\mu - H\Big(W\Big|\phi\big(C, g(\mathbf{Z}(W))\big)\Big) && (3.20)\\
&\overset{(2.52)}{\geqslant} n\mu - H(\varepsilon) - \varepsilon n\mu. && (3.21)
\end{aligned}$$

The equality in (3.16) holds as $\mathbf{X}(i) \perp \mathbf{Z}(j)$ for $i \neq j$, (3.17) follows from $W \perp C$, (3.20) follows from $W \perp \mathbf{Z}(W)$, the fact that $H(W) = n\mu$ and the data processing inequality, Theorem 2.5. Fano's inequality, Theorem 2.6, was used in (3.21). This shows $(\mu, R_1, R_2) \in \overline{\mathcal{R}_*}$ as ε was arbitrary.

To show the other direction, we apply the achievability result from Theorem 2.25 to the multi-letter source $(\mathbf{X}, \mathbf{Z})$. Assuming $(\mu, R_1, R_2) \in \mathcal{R}_*$, we know that for some $n \in \mathbb{N}$ there are random variables (U, V) satisfying the Markov chain $U -\!\!\!\circ\!\!\!- \mathbf{X} -\!\!\!\circ\!\!\!- \mathbf{Z} -\!\!\!\circ\!\!\!- V$ and (3.3)–(3.5) hold. By Theorem 2.25, the triple $(n\mu, nR_1, nR_2)$ is asymptotically PR-achievable for the source $(\mathbf{X}, \mathbf{Z})$ with an arbitrary error probability $\varepsilon > 0$. For any $\varepsilon' > 0$ we can find $k \in \mathbb{N}$, a $(k, nR_1 + \varepsilon', nR_2 + \varepsilon')$-code (f, g) for $(\mathbf{X}, \mathbf{Z})$, and a function ϕ, such that (2.52) is satisfied with $W \sim \mathcal{U}\Big([e^{k(n\mu - \varepsilon')}]\Big)$. Thus, $\Big(\mu - \frac{\varepsilon'}{n}, R_1 + \frac{\varepsilon'}{n}, R_2 + \frac{\varepsilon'}{n}\Big) \in \mathcal{R}_{\mathrm{PR}}$ for the source (X, Z) and as ε' was arbitrary, this completes the proof. □

3.3 BOUNDS ON THE ACHIEVABLE REGION

We first provide outer bounds on the set of achievable triples.

Theorem 3.7. *We have $\mathcal{R} \subseteq \mathcal{R}_o \subseteq \mathcal{R}'_o$, where the two regions $\mathcal{R}_o$ and $\mathcal{R}'_o$ are given by*

$$\mathcal{R}_o := \bigcup_{U,V} \big\{(\mu, R_1, R_2) : R_1 \geqslant I(U;X), R_2 \geqslant I(V;Z), \textit{ and } \mu \leqslant I(V;Z) + I(U;X) - I(UV;XZ)\big\}, \quad (3.22)$$

$$\mathcal{R}'_o := \bigcup_{U,V} \big\{(\mu, R_1, R_2) : R_1 \geqslant I(U;X), R_2 \geqslant I(V;Z), \textit{ and } \mu \leqslant \min\{I(U;Z), I(V;X)\}\big\}, \quad (3.23)$$

with U and V any pair of random variables satisfying $U -\!\!\!\circ\!\!\!- X -\!\!\!\circ\!\!\!- Z$ and $X -\!\!\!\circ\!\!\!- Z -\!\!\!\circ\!\!\!- V$.

Theorem 3.7 follows from the outer bound for the pattern recognition problem, Theorem 2.26, via the equivalence shown in Proposition 3.6. Nonetheless, we provide a short, self-contained proof in Appendix A.1 for the sake of completeness.

The regions $\mathcal{R}_o$ and $\mathcal{R}'_o$ are both convex since a time-sharing random variable can be incorporated into U and V. Furthermore, $\mathcal{R}'_o$ remains unchanged when U and V are required to satisfy the complete Markov chain $U -\!\!\!\circ\!\!\!- X -\!\!\!\circ\!\!\!- Z -\!\!\!\circ\!\!\!- V$.

The numerical computation of the outer bounds requires the cardinalities of the auxiliary random variables to be bounded. We therefore complement Theorem 3.7 with the following result, whose proof is provided in Appendix A.2.

Proposition 3.8. *We have $\mathcal{R}_o = \mathrm{conv}(\mathcal{S}_o)$ and $\mathcal{R}'_o = \mathrm{conv}(\mathcal{S}'_o)$, where the regions $\mathcal{S}_o$ and $\mathcal{S}'_o$ are defined as $\mathcal{R}_o$ and $\mathcal{R}'_o$, respectively, but with the additional cardinality bounds $|\mathcal{U}| \leqslant |\mathcal{X}|$ and $|\mathcal{V}| \leqslant |\mathcal{Z}|$.*

The cardinality bounds in this result are tighter than the usual bounds obtained with the convex cover method (cf. [16, Appendix C], [2], [71]), where the cardinality has to be increased by one. Thus, when dealing with binary sources in Section 3.5, binary auxiliaries suffice. The smaller cardinalities come at the cost of convexification in Proposition 3.8 since the regions $\mathcal{S}_o$ and $\mathcal{S}'_o$ themselves are not necessarily convex.

We next state an inner bound for the achievable region. A more general inner bound will be proved in Section 5.1 (cf. Theorem 5.4) for the multi-clustering problem with an arbitrary number of sources.

Theorem 3.9. *We have $\mathcal{R}_i \subseteq \overline{\mathcal{R}}$ where*

$$\mathcal{R}_i := \bigcup_{U,V} \{(\mu, R_1, R_2) : R_1 \geqslant I(U;X), R_2 \geqslant I(V;Z), \text{ and } \\ \mu \leqslant I(U;V)\}, \tag{3.24}$$

with auxiliary random variables U, V *satisfying* $U \multimap X \multimap Z \multimap V$.

Theorem 3.9 directly follows from Theorem 2.23, leveraging the equivalence detailed in Theorem 3.4 and Corollary 3.5. Alternatively, it also follows from Theorem 2.25, using Corollary 3.5 and Proposition 3.6. As usual for Berger-Tung type bounds, the main differences between the outer and the inner bound lies in the Markov conditions (cf. [64, Chapter 7] or [16, Section 12.2]). Note that $\mathcal{R}_o$ and $\mathcal{R}_i$ would coincide if the Markov condition $U \multimap X \multimap Z \multimap V$ were imposed in the definition of $\mathcal{R}_o$.

Employing a binning scheme would not enlarge the inner bound $\mathcal{R}_i$. The intuition is that binning reduces redundant information transmitted by both encoders. In the multi-clustering problem, however, this quantity should actually be maximized.

A tight bound on the achievable region can be obtained if μ is not greater than the common information (cf. Definition 2.11) of X and Z, as stated in the following corollary.

Corollary 3.10. *If $Y = \zeta(X) = \xi(Z)$ is a common component of X and Z and $0 \leqslant \mu \leqslant H(Y)$ then we have $(\mu, R_1, R_2) \in \overline{\mathcal{R}}$ if and only if $\mu \leqslant \min\{R_1, R_2\}$ holds.*

Common components are defined in Definition 2.11.

Proof. Theorem 3.7 entails $\mu \leqslant \min\{R_1, R_2\}$ for any $(\mu, R_1, R_2) \in \overline{\mathcal{R}}$. With $U = V = Y$, Theorem 3.9 implies $(H(Y), H(Y), H(Y)) \in \overline{\mathcal{R}}$. Using time-sharing with $0 \in \overline{\mathcal{R}}$ we obtain $(\mu, \mu, \mu) \in \overline{\mathcal{R}}$ for $0 \leqslant \mu \leqslant H(Y)$ and hence $(\mu, R_1, R_2) \in \overline{\mathcal{R}}$ if $\mu \leqslant \min\{R_1, R_2\}$. □

The inner bound $\mathcal{R}_i$ can be improved by convexification. Furthermore, we incorporate the same strong cardinality bounds as for the outer bound (cf. Proposition 3.8), thereby enabling us to use binary auxiliaries also for the inner bound, when dealing with binary sources in Section 3.5.

Proposition 3.11. *We have* $\mathcal{S}_i' := \operatorname{conv}(\mathcal{S}_i) = \operatorname{conv}(\mathcal{R}_i) \subseteq \overline{\mathcal{R}}$ *where* $\mathcal{S}_i$ *is defined as* $\mathcal{R}_i$*, but with the additional cardinality bounds* $|\mathcal{U}| \leqslant |\mathcal{X}|$*, and* $|\mathcal{V}| \leqslant |\mathcal{Z}|$*. Furthermore,* $\mathcal{S}_i'$ *can be explicitly expressed as*

$$\mathcal{S}_i' = \bigcup_{U,V,Q} \{(\mu, R_1, R_2) : \; R_1 \geqslant I(U;X|Q), R_2 \geqslant I(V;Z|Q), \text{ and } \\ \mu \leqslant I(U;V|Q)\}, \tag{3.25}$$

where U*,* V*, and* Q *are random variables with* $|\mathcal{U}| \leqslant |\mathcal{X}|$*,* $|\mathcal{V}| \leqslant |\mathcal{Z}|$*,* $|\mathcal{Q}| \leqslant 3$*, and a p.m.f. of the form* $p_{XZUVQ} = p_Q \, p_{XZ} \, p_{U|XQ} \, p_{V|ZQ}$*.*

The proof of this result is given in Appendix A.3.

3.4 THE INFORMATION BOTTLENECK METHOD

The information-theoretic problem posed by the information bottleneck method (cf. Section 2.2.2) can be obtained as a special case from the multi-clustering problem. The information bottleneck problem in turn is equivalent to a CEO problem with logarithmic loss distortion (cf. Section 2.2.2). In total, we can show the following equivalences.

Proposition 3.12. *For a source* (X, Z) *and a pair* $(\mu, R) \in \mathbb{R}^2$*, the following are equivalent:*

1. $(\mu, R) \in \overline{\mathcal{R}_{\mathrm{IB}}}$ *for the source* (X, Z)*.*
2. $(H(Z) - \mu, R) \in \overline{\mathcal{R}_{\mathrm{LL}}}$ *for the source* (X, Z)*.*
3. $(\mu, R, \infty) \in \overline{\mathcal{R}}$ *for the source* (X, Z)*.*
4. *There exists a random variable* U *such that* $U -\!\!\circ\!\!- X -\!\!\circ\!\!- Y$*,* $I(X;U) \leqslant R$*, and* $I(Z;U) \geqslant \mu$*.*

We apply Definition 2.28 with $J = L = 1$.

Proof. The equivalence "1 ⇔ 3" holds as Definition 3.2 collapses to Definition 2.27 for $R_2 = \infty$. "2 ⇔ 4" is a special case of Theorem 2.29. To show "3 ⇒ 4" we apply the outer bound $\mathcal{R}_o'$ of Theorem 3.7 to obtain a random variable U satisfying the conditions in part 4. The direction "4 ⇒ 3" can be deduced from Theorem 3.9, choosing the auxiliary $V = Y$. □

This tradeoff between relevance (μ) and complexity (R) can equivalently be characterized by the IB function (cf. [11], [20]), defined as $\mu_{\mathrm{IB}}(R) := \sup\{\mu : (\mu, R) \in \overline{\mathcal{R}_{\mathrm{IB}}}\}$. Proposition 3.12 provides

$$\mu_{\mathrm{IB}}(R) = \max_{\substack{U \,:\, I(U;X) \leqslant R \\ U -\!\circ\!- X -\!\circ\!- Z}} I(U;Z). \tag{3.26}$$

Interestingly, the function (3.26) is the solution to a variety of different problems in information theory. As mentioned in [20], (3.26) is the solution to the problem of loss-less source coding with one helper [2], [70]. Witsenhausen and Wyner [68] investigated a lower bound for a conditional entropy when simultaneously requiring another conditional entropy to fall below a threshold. Their work was a generalization of [73] and furthermore related to [2], [3], [69], [72]. The conditional entropy bound in [68] turns out to be an equivalent characterization of (3.26). Furthermore, μ_{IB} characterizes the optimal error exponent, when testing against independence with one-sided data compression [1, Theorem 2]. Also in the context of gambling in the horse race market, (3.26) occurs as the maximum incremental growth in wealth when rate-limited side-information is available to the gambler [17, Theorem 3].

3.5 DOUBLY SYMMETRIC BINARY SOURCE

We will consider the special case where $(X, Z) \sim \mathrm{DSBS}(p)$ is a doubly symmetric binary source. The cardinality bounds in Propositions 3.8 and 3.11 will enable us to use binary auxiliaries. We first show that the inner bound $\mathcal{S}'_i$ and the outer bound $\mathcal{R}'_o$ do not coincide.

$\mathrm{DSBS}(p)$ is defined in Section 2.2.2.

Proposition 3.13. *For the source $(X, Z) \sim \mathrm{DSBS}(p)$ with $p \in (0, 1)$ and $p \neq \frac{1}{2}$, we have $\mathcal{S}'_i \neq \mathcal{R}'_o$.*

The proof of this proposition is given in Appendix A.4. We conjecture that there is also a gap between $\mathcal{S}'_i$ and the stronger outer bound $\mathcal{R}_o$.

Conjecture 3.14. *There exists $p \in [0, 1]$, such that $\mathcal{S}'_i \neq \mathcal{R}_o$ for the source $(X, Z) \sim \mathrm{DSBS}(p)$.*

To support Conjecture 3.14, we will introduce a region $\mathcal{S}_b \subseteq \mathcal{S}_i$ and show that $\mathrm{conv}(\mathcal{S}_b) \neq \mathcal{R}_o$. Let $\mathcal{S}_b$ be defined as

$$\mathcal{S}_b := \bigcup_{0 \leqslant \alpha, \beta \leqslant \frac{1}{2}} \big\{ (\mu, R_1, R_2) : R_1 \geqslant 1 - H(\alpha), \; R_2 \geqslant 1 - H(\beta), \text{ and } \mu \leqslant 1 - H(\alpha * p * \beta) \big\}. \tag{3.27}$$

$a * b = \bar{a}b + a\bar{b}$.

By choosing $U = X \oplus N_1$ and $V = Z \oplus N_2$, where $N_1 \sim \mathcal{B}(\alpha)$ and $N_2 \sim \mathcal{B}(\beta)$ are independent of (X, Z) and of each other, it follows that $\mathcal{S}_b \subseteq \mathcal{S}_i$. To illustrate the tradeoff between complexity (R_1, R_2) and relevance (μ), the upper boundary of $\mathcal{S}_b$ is depicted in Figure 5 for $p = 0.1$.

Based on numerical experiments, we conjecture the following.

Conjecture 3.15. *For the source $(X, Z) \sim \mathrm{DSBS}(p)$ with $p \in [0, 1]$ we have $\mathcal{S}'_i = \mathrm{conv}(\mathcal{S}_b)$, or equivalently $\mathcal{S}_i \subseteq \mathrm{conv}(\mathcal{S}_b)$.*

The natural, stronger conjecture that $\mathcal{S}_b = \mathcal{S}_i$ already appeared in [66, Conjecture 1, Eq. (14)]. However, there is a counterexample [8].

Proposition 3.16. *For the source $(X, Z) \sim \mathrm{DSBS}(0)$ we have $\mathcal{S}_b \neq \mathcal{S}_i$.*

Note that $X = Z$ for $p = 0$.

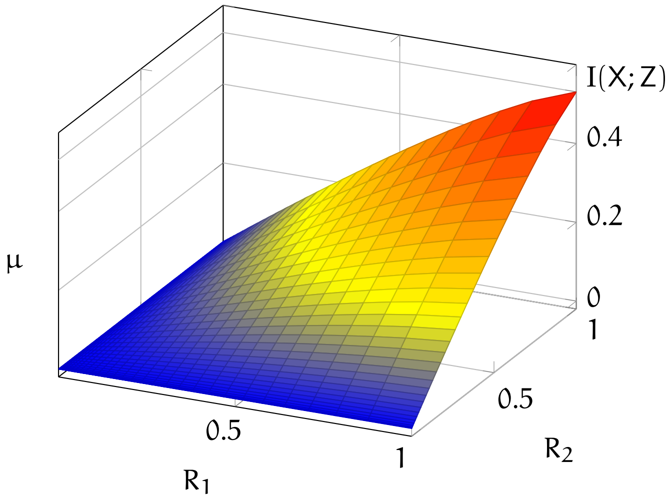


Figure 5: Boundary of $\mathcal{S}_b$ for $p = 0.1$.

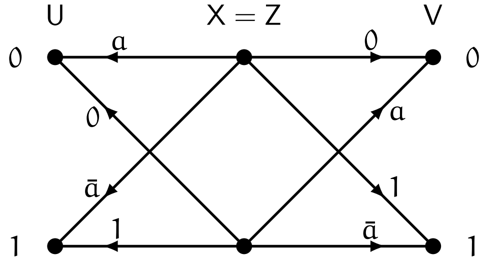


Figure 6: Binary channels for the proof of Proposition 3.16.

Proof. For $a \in [0, 1]$ we define (U, V) by the binary channels depicted in Figure 6, satisfying $U \mathrel{-\!\!\circ\!\!-} X \mathrel{-\!\!\circ\!\!-} Z \mathrel{-\!\!\circ\!\!-} V$. We obtain $(\mu, R, R) \in \mathcal{S}_i$ with $R = I(U;X) = I(V;Z) = H\left(\frac{a}{2}\right) - \frac{1}{2}H(a)$ and $\mu = I(U;V) = 2R - a$. For $a = 0.8$ we have $\mu \approx 0.419973$ and $R \approx 0.609987$. On the other hand, we obtain $\mu_b := \max\{\hat{\mu} : (\hat{\mu}, R, R) \in \mathcal{S}_b\} < 0.412025$ using (3.27) with $\alpha = \beta \approx 0.07658$. As $\mu_b < \mu$ we have $(\mu, R, R) \notin \mathcal{S}_b$.

This argument can be verified numerically using interval arithmetic [39]. Code written in the Octave Programming Language [21] using its interval package [29] can be found at [46]. □

Note that Proposition 3.16 concerns the case $p = 0$ and therefore does not impact Conjecture 3.15. For $p = 0$ we have $X = Z$ and Corollary 3.10 implies $\overline{\mathcal{R}} = \big\{(\mu, R_1, R_2) : R_1, R_2 \geqslant 0 \text{ and } \mu \leqslant \min\{R_1, R_2, 1\}\big\}$. It is easily verified that $\overline{\mathcal{R}} = \operatorname{conv}(\mathcal{S}_b)$ and thus Conjecture 3.15 holds for $p = 0$ by Proposition 3.11.

In fact, the entire statement [66, Conjecture 1] does not hold. The second part [66, Conjecture 1, Eq. (15)] claims that $\operatorname{conv}(\mathcal{S}_b) = \mathcal{R}_o$. In what follows, we show how to construct a counterexample.

Proposition 3.17. *For* $(X, Z) \sim \mathrm{DSBS}(0.1)$, *we have* $\operatorname{conv}(\mathcal{S}_b) \neq \mathcal{R}_o$.

Proposition 3.17 shows that Conjecture 3.14 follows directly from Conjecture 3.15.

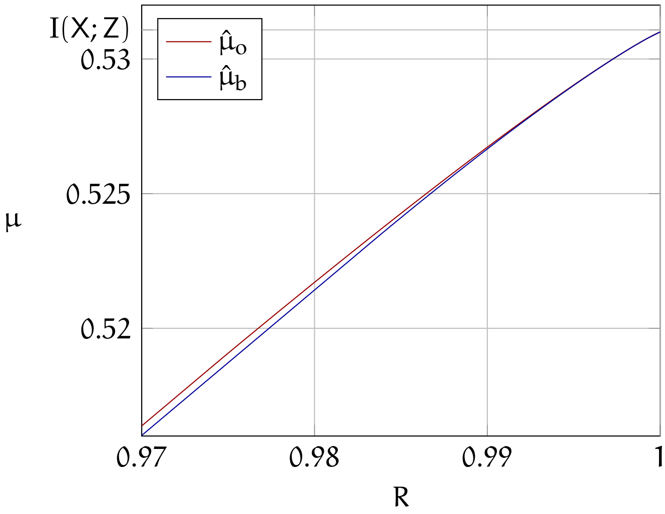


Figure 7: Numerical evaluation of $\hat{\mu}_o$ and $\hat{\mu}_b$ for $p = 0.1$.

To prove Proposition 3.17 we construct a point $(\mu, R, R) \in \mathcal{R}_o$ that satisfies $(\mu, R, R) \notin \operatorname{conv}(\mathcal{S}_b)$. To this end, define the concave functions $\hat{\mu}_b(R) := \max\{\mu : (\mu, R, R) \in \operatorname{conv}(\mathcal{S}_b)\}$ and $\hat{\mu}_o(R) := \max\{\mu : (\mu, R, R) \in \mathcal{R}_o\}$ for $R \in [0, 1]$. In order to show $\operatorname{conv}(\mathcal{S}_b) \neq \mathcal{R}_o$, it suffices to find $\hat{R} \in [0, 1]$ with $\hat{\mu}_b(\hat{R}) < \hat{\mu}_o(\hat{R})$.

It is straightforward to compute an upper bound of $\hat{\mu}_b$ numerically: For $\alpha, \beta \in [0, \frac{1}{2}]$, we compute

$$\widetilde{R}_1 := 1 - H(\alpha) \tag{3.28}$$

$$\widetilde{R}_2 := 1 - H(\beta) \tag{3.29}$$

$$\widetilde{\mu} := 1 - H(\alpha * p * \beta) \tag{3.30}$$

on a suitably fine grid and numerically bound the upper concave hull of the implicitly defined function $\widetilde{\mu}(\widetilde{R}_1, \widetilde{R}_2)$. Evaluating it at $R = \widetilde{R}_1 = \widetilde{R}_2$ yields an upper bound of $\hat{\mu}_b(R)$.

On the other hand, we can obtain a lower bound on $\hat{\mu}_o$ by numerically computing (3.22) for specific probability mass functions that satisfy the Markov constraints in Theorem 3.7. Note that based on the cardinality bound in Proposition 3.8, we can restrict the auxiliaries U and V to be binary. We randomly sample the binary probability mass functions that satisfy the Markov constraints in Theorem 3.7 (but not necessarily the long Markov chain $U \multimap X \multimap Z \multimap V$) and in doing so encountered points strictly above the graph of $\hat{\mu}_b$. Figure 7 shows the resulting bounds for $p = 0.1$ in the vicinity of $R = 1$. Albeit small, there is clearly a gap between $\hat{\mu}_b$ and $\hat{\mu}_o$ outside the margin of numerical error. This shows that the bounds are not tight and [66, Eq. (15), Conjecture 1] does not hold.

Proof of Proposition 3.17. We observed the largest gap between the two bounds at a rate of $\hat{R} \approx 0.974795$. The particular distribution of UV

u	v	x	z	$P\{U = u, V = v \vert X = x, Z = z\}$
0	0	0	0	0.9953581462173534O6525
0	0	0	1	0.00249767559844423319075
0	0	1	0	0.002498344003957310643325
0	0	1	1	0.00034313919194834475
0	1	0	0	0.002142603857654094275
0	1	0	1	0.99500307447656326760925
0	1	1	0	0.000000905921035188556675
0	1	1	1	0.00215611073304415445
1	0	0	0	0.002142603857654094275
1	0	0	1	0.00000157432654826600925
1	0	1	0	0.995002406071050190156675
1	0	1	1	0.00215611073304415445
1	1	0	0	0.000356646067338404925
1	1	0	1	0.00249767559844423319075
1	1	1	0	0.002498344003957310643325
1	1	1	1	0.99534463934196334635

Table 2: Distribution resulting from random search.

at this rate, resulting from optimizing over the distributions that satisfy the Markov constraints in Theorem 3.7 is given in Table 2 for reference. Note that this is an exact conditional p.m.f. satisfying the Markov chains $U \multimap\!\!- X \multimap\!\!- Z$ and $X \multimap\!\!- Z \multimap\!\!- V$. It achieves $I(V;Z) + I(U;X) - I(UV;XZ) \approx 0.518966$ which is $\Delta \approx 2.86472 \cdot 10^{-4}$ above the inner bound. Thus, this distribution provides a point $\boldsymbol{x} \in \mathcal{R}_o$ with $\boldsymbol{x} \notin \operatorname{conv}(\mathcal{S}_b)$.

Using interval arithmetic [39] this claim can be verified numerically. Code written in the Octave Programming Language [21] using its interval package [29] can be found at [46]. It uses the distribution given in Table 2. □

We firmly believe that a tight characterization of the achievable region requires an improved outer bound. However, it appears very difficult to find a manageable outer bound based on the full Markov chain $U \multimap\!\!- X \multimap\!\!- Z \multimap\!\!- V$.

3.6 CONCLUSION

We studied the multi-clustering problem with two sources and connected it to a hypothesis testing and a pattern recognition problem. Exploiting these connections, we provided an outer and an inner bound on the achievable region. In case one rate is large enough, the problem degenerates to the information bottleneck problem and we obtained tight bounds. This case is also equivalent to a rate-distortion problem with logarithmic loss distortion. We were able to prove novel cardinality bounds by combining the convex cover method with the perturbation method and leveraging ideas similar to [41]. This allowed us to restrict our attention to the extreme points of the achievable region. The resulting cardinality bounds enabled a thorough study of the doubly symmetric binary source, where we were able to disprove the conjecture [66, Conjecture 1] and found a gap between the outer and inner bound. The existence of this gap, however, rests on the unproven Conjecture 3.15.

We believe that the improved cardinality bounds in Propositions 3.8 and 3.11 can be applied to many bounds in information theory. They offer a cardinality reduction by one compared to bounds obtained via the convex cover method, at the cost of additional convexification. In many cases, as, e. g., in the binary example that was studied here, numerically computing the upper concave envelope can be computationally cheaper than optimizing over random variables with larger cardinality.

Regarding the tightness of the bounds in Theorems 3.7 and 3.9, we believe that the inner bound $\operatorname{conv}(\mathcal{R}_i)$ coincides with the achievable region $\overline{\mathcal{R}}$ of the multi-clustering problem and that the outer bound $\mathcal{R}_o$ is loose and needs to be improved. However, obtaining a good upper bound for the mutual information between two arbitrary encodings solely based on their rates is a difficult task. Standard information-theoretic manipulations appear incapable of handling this dependence well.

4 MUTUAL INFORMATION BETWEEN TWO BOOLEAN FUNCTIONS

4.1 INTRODUCTION AND MAIN RESULTS

Recently, Kumar and Courtade introduced a conjecture [10], [36] concerning Boolean functions that maximize mutual information. Their work was inspired by a similar problem in computational biology [33].

Conjecture 4.1 ([36, Conjecture 1]). *Let $(X, Y) \sim \mathrm{DSBS}(p)$ be a doubly symmetric binary source with $p \in [0, 1]$. Then, for any $n \in \mathbb{N}$ and any Boolean function $f\colon \{0, 1\}^n \to \{0, 1\}$, we have*

$$I\big(f(\mathbf{X}); \mathbf{Y}\big) \leqslant I(X; Y) = 1 - H(p). \tag{4.1}$$

$(\mathbf{X}, \mathbf{Y})$ are n i.i.d. copies of (X, Y).

This result appears innocent at first sight and indeed it is trivial to see that the functions $f_i(\mathbf{x}) = x_i$ for $i \in [n]$ achieve equality in (4.1). However, the conjecture turns out to be much more involved and cannot be established by standard information-theoretic arguments or by induction over n (cf. [32, Section 2]). Furthermore, Conjecture 4.1 can hold only for doubly symmetric binary sources, i. e., a generalization to arbitrary binary sources is impossible [5, Section I.A]. Conjecture 4.1 has received significant interest and resisted several efforts to find a proof (see the discussion "Recent Progress" in [10, Section IV]). More recently, Ordentlich et. al. [44] used Fourier-analytic techniques and leveraged hypercontractivity to improve upon previously known bounds for $I(f(\mathbf{X}); \mathbf{Y})$. This group also recently published on a complementary problem concerning quantization to $n - 1$ bits [30]. Kindler et. al. [32] studied an analogous problem in Gaussian spaces.

The main result of this chapter is a relaxed version of Conjecture 4.1, involving two Boolean functions. To prove this statement, we will make use of Fourier analysis for Boolean functions (cf. Section 2.4). Therefore, it will be will be more convenient to state the problem in terms of Rademacher random variables. Let X, Y be two dependent Rademacher random variables (cf. Section 2.4), with correlation coefficient $\rho := \mathbb{E}[XY] \in [-1, 1]$.

We have $X, Y \sim \mathcal{U}(\{-1, 1\})$.

Theorem 4.2. *For any two Boolean functions $f, g\colon \{-1, 1\}^n \to \{-1, 1\}$,*

$$I\big(f(\mathbf{X}); g(\mathbf{Y})\big) \leqslant I(X; Y). \tag{4.2}$$

Remark 5. Note that Theorem 4.2 is equivalent to

$$I\big(\tilde{f}(\tilde{\mathbf{X}});\tilde{g}(\tilde{\mathbf{Y}})\big) \leqslant I(\tilde{X};\tilde{Y}), \tag{4.3}$$

for any pair of functions $\tilde{f},\tilde{g}\colon \{0,1\}^n \to \{0,1\}^n$, where $(\tilde{X},\tilde{Y}) \sim \mathrm{DSBS}(p)$. This can be shown from (4.2) by introducing the bijective transformation $T(w) := \frac{1-w}{2}$ and identifying $p = T(\rho)$, $\tilde{X} = T(X)$, $\tilde{Y} = T(Y)$, $\tilde{f} = T(f)$, and $\tilde{g} = T(g)$.

Assuming that Conjecture 4.1 holds and taking into account the equivalence discussed in Remark 5, Theorem 4.2 readily follows by noting that

$$\begin{aligned} I\big(f(\mathbf{X});g(\mathbf{Y})\big) &\leqslant I\big(f(\mathbf{X});\mathbf{Y}\big) && (4.4)\\ &\overset{(4.1)}{\leqslant} I(X;Y), && (4.5)\end{aligned}$$

where (4.4) follows from the data processing inequality (Theorem 2.5). This shows that Theorem 4.2 is indeed weaker than Conjecture 4.1. However, Theorem 4.2 was stated as an open problem in [36, Section IV] and [10, Section IV], and separately investigated in [5]. A proof of (4.2) was previously available only under the additional restrictive assumptions that f and g are equally biased (i. e., $\mathbb{E}[f(\mathbf{X})] = \mathbb{E}[g(\mathbf{X})]$) and satisfy the condition

$$P\{f(\mathbf{X}) = 1, g(\mathbf{X}) = 1\} \geqslant P\{f(\mathbf{X}) = 1\}P\{g(\mathbf{X}) = 1\}. \tag{4.6}$$

The reader is invited to see [10, Section IV] for further details. Our proof of Theorem 4.2 builds on these ideas. Using Fourier-analytic tools we prove (4.2) any additional restrictions on f and g. First, we suitably bound the Fourier coefficients of f and g and thereby reduce (4.2) to an elementary inequality, which is subsequently established. Like Conjecture 4.1, Theorem 4.2 does not hold for arbitrary binary asymmetric sources: using the counterexample from [5, Section I.A] it follows that our Fourier-analytic proof will not carry over to p-biased Fourier analysis [60], [61]. We note that Anantharam et al. [5] showed that Theorem 4.2 would follow from a conjectured result concerning the hypercontractivity ribbon of two binary random variables; however, that conjecture itself remains unproven to date.

The i*th dictator is the function* $\chi_i(x) = x_i$.

A careful inspection of our proof of Theorem 4.2 reveals that in general, up to sign changes, the dictator functions (cf. Remark 3) χ_i, $i \in [n]$ are the unique maximizers of $I\big(f(\mathbf{X});g(\mathbf{Y})\big)$.

Proposition 4.3. *If* $0 < |\rho| < 1$, *equality in (4.2) is achieved if and only if* $f = \pm g = \pm\chi_i$ *for some* $i \in [n]$.

For the degenerate case $\rho = 0$, the upper bound $I\big(f(\mathbf{X});g(\mathbf{Y})\big) = 0$ is trivially achieved by any two Boolean functions f, g. Similarly, for the case $\rho = \pm 1$ (deterministically dependent sources), the upper bound

$I\big(f(\mathbf{X}); g(\mathbf{Y})\big) = 1$ is achieved by any two unbiased Boolean functions f and g that satisfy $g(x) = \pm f(\rho x)$.

4.2 PROOF OF THEOREM 4.2

Let f and g be two arbitrary functions on the Boolean hypercube, i. e., $f, g\colon \{-1,1\}^n \to \{-1,1\}$. Define

$$a := \frac{1+\widehat{f}_\varnothing}{2} = P\{f(\mathbf{X}) = 1\}, \quad b := \frac{1+\widehat{g}_\varnothing}{2} = P\{g(\mathbf{X}) = 1\}, \tag{4.7}$$

$$\theta_\rho := \frac{1}{4}\big(\langle f, T_\rho g\rangle - \widehat{f}_\varnothing \widehat{g}_\varnothing\big) = \frac{1}{4} \sum_{\mathcal{S}:|\mathcal{S}|\geqslant 1} \rho^{|\mathcal{S}|} \widehat{f}_\mathcal{S} \widehat{g}_\mathcal{S}. \tag{4.8}$$

$\widehat{f}$ and $\widehat{g}$ denote the Fourier transforms (cf. Section 2.4).

Without loss of generality, we may assume $\widehat{f}_\varnothing, \widehat{g}_\varnothing \in [0,1]$ (or equivalently $a, b \in \big[\frac{1}{2}, 1\big]$), as

$$I\big(f(\mathbf{X}); g(\mathbf{Y})\big) = I\big(\operatorname{sgn}(\widehat{f}_\varnothing) f(\mathbf{X}); \operatorname{sgn}(\widehat{g}_\varnothing) g(\mathbf{Y})\big) \tag{4.9}$$

$$= I\big(f^*(\mathbf{X}); g^*(\mathbf{Y})\big), \tag{4.10}$$

where $f^* := \operatorname{sgn}(\widehat{f}_\varnothing) f$ and $g^* := \operatorname{sgn}(\widehat{g}_\varnothing) g$ with $\widehat{f}^*_\varnothing, \widehat{g}^*_\varnothing \in [0,1]$. We may further assume $\frac{1}{2} \leqslant a \leqslant b \leqslant 1$ as f and g can be interchanged. We also restrict $\rho \in [0,1]$ as

Mutual information is symmetric, i. e., $I(A;B) = I(B;A)$.

$$I\big(f(\mathbf{X}); g(\mathbf{Y})\big) = I\big(f(\mathbf{X}); g(\operatorname{sgn}(\rho)\mathbf{Y}^*)\big) = I\big(f(\mathbf{X}); g^*(\mathbf{Y}^*)\big), \tag{4.11}$$

where we defined $Y^* := \operatorname{sgn}(\rho) Y$ and $g^*(y) = g(\operatorname{sgn}(\rho) y)$, and hence have $\mathbb{E}[XY^*] = |\rho|$.

Part 6 of Lemma 2.45 allows us to rewrite the probabilities

$$P\{f(\mathbf{X}) = g(\mathbf{Y}) = 1\} = ab + \theta_\rho, \tag{4.12}$$

$$P\{f(\mathbf{X}) = 1, g(\mathbf{Y}) = -1\} = a\bar{b} - \theta_\rho, \tag{4.13}$$

$$P\{f(\mathbf{X}) = -1, g(\mathbf{Y}) = 1\} = \bar{a}b - \theta_\rho, \text{ and} \tag{4.14}$$

$$P\{f(\mathbf{X}) = g(\mathbf{Y}) = -1\} = \bar{a}\bar{b} + \theta_\rho. \tag{4.15}$$

$\bar{t} := 1 - t.$

This can be seen as follows. Fist note that for an arbitrary pair of random variables (A, B) on $\{-1,1\}^2$, we have

$$2P\{A = B = 1\} = 2P\{A = B = 1\} + P\{A = 1, B = -1\}$$

$$+ P\{A = -1, B = 1\} - P\{A \neq B\} \tag{4.16}$$

$$= P\{A = 1\} + P\{B = 1\} - P\{A \neq B\} \tag{4.17}$$

and thus

$$\begin{aligned}
P\{f(\mathbf{X}) = g(\mathbf{Y}) = 1\} &\overset{(4.17)}{=} \frac{1}{2}\big(a + b - P\{f(\mathbf{X}) \neq g(\mathbf{Y})\}\big) && (4.18)\\
&= \frac{1}{2}\left(a + b - \frac{1}{2}\big(1 - \langle f, T_\rho g\rangle\big)\right) && (4.19)\\
&\overset{(4.8)}{=} \frac{1}{2}(a+b) - \frac{1 - \widehat{f}_\varnothing \widehat{g}_\varnothing}{4} + \theta_\rho && (4.20)\\
&\overset{(4.7)}{=} \frac{1}{2}(a+b) + \frac{4ab - 2a - 2b}{4} + \theta_\rho && (4.21)\\
&= ab + \theta_\rho, && (4.22)
\end{aligned}$$

where (4.19) follows from part 6 of Lemma 2.45. The other identities (4.13)–(4.15) can be obtained similarly. Using (4.12)–(4.15), we obtain $I\big(f(\mathbf{X}); g(\mathbf{Y})\big) = \xi(\theta_\rho, a, b)$ with the continuous function

$H(p) = H(Z)$, *where* $Z \sim p$.

$$\begin{aligned}
\xi(\theta, a, b) := {}& H(a) + H(b)\\
&- H\big(ab + \theta, a\bar{b} - \theta, \bar{a}b - \theta, \bar{a}\bar{b} + \theta\big). && (4.23)
\end{aligned}$$

By the non-negativity of probabilities (4.13) and (4.15), for any $\rho \in [0, 1]$,

$$-\bar{a}\bar{b} \leqslant \theta_\rho \leqslant a\bar{b}. \qquad (4.24)$$

Defining $\mathcal{P} := \{\mathcal{S} \subseteq [n] : \widehat{f}_\mathcal{S} \widehat{g}_\mathcal{S} > 0\} \setminus \{\varnothing\}$ and $\mathcal{N} := \{\mathcal{S} \subseteq [n] : \widehat{f}_\mathcal{S} \widehat{g}_\mathcal{S} < 0\}$, we have $\theta_1 = \tau^+ + \tau^-$ with

$$\tau^+ := \frac{1}{4} \sum_{\mathcal{S} \in \mathcal{P}} \widehat{f}_\mathcal{S} \widehat{g}_\mathcal{S}, \qquad \tau^- := \frac{1}{4} \sum_{\mathcal{S} \in \mathcal{N}} \widehat{f}_\mathcal{S} \widehat{g}_\mathcal{S}. \qquad (4.25)$$

Using the Schwarz inequality we can show

$$\begin{aligned}
(\tau^+ - \tau^-)^2 &\overset{(4.25)}{=} \frac{1}{16}\left(\sum_{\mathcal{S}:|\mathcal{S}|\geqslant 1} |\widehat{f}_\mathcal{S} \widehat{g}_\mathcal{S}|\right)^2 && (4.26)\\
&\leqslant \frac{1}{16}\left(\sum_{\mathcal{S}:|\mathcal{S}|\geqslant 1} \widehat{f}_\mathcal{S}^2\right)\left(\sum_{\mathcal{S}:|\mathcal{S}|\geqslant 1} \widehat{g}_\mathcal{S}^2\right) && (4.27)\\
&= \frac{1}{16}(1 - \widehat{f}_\varnothing^2)(1 - \widehat{g}_\varnothing^2) && (4.28)\\
&\overset{(4.7)}{=} a\bar{a}b\bar{b}, && (4.29)
\end{aligned}$$

where we applied Corollary 2.42 in (4.27), and (4.28) follows from part 5 of Lemma 2.45. Combining these results, we obtain

$$
\begin{aligned}
2\tau^{+} &= \theta_1 + \tau^{+} - \tau^{-} \tag{4.30}\\
&\overset{(4.29)}{\leqslant} \theta_1 + \sqrt{a\bar{a}b\bar{b}} \tag{4.31}\\
&\overset{(4.24)}{\leqslant} a\bar{b} + \sqrt{a\bar{a}b\bar{b}} \tag{4.32}
\end{aligned}
$$

and, using (4.24) and (4.29), one similarly obtains $2\tau^{-} \geqslant -\bar{a}\bar{b} - \sqrt{a\bar{a}b\bar{b}}$ from $2\tau^{-} = \theta_1 + \tau^{-} - \tau^{+}$. From the definition of θ_ρ we have

$$
\begin{aligned}
\theta_\rho &\overset{(4.8)}{=} \frac{1}{4}\sum_{\mathcal{S}:|\mathcal{S}|\geqslant 1} \rho^{|\mathcal{S}|}\widehat{f}_{\mathcal{S}}\widehat{g}_{\mathcal{S}} \tag{4.33}\\
&\leqslant \frac{1}{4}\sum_{\mathcal{S}\in\mathcal{P}} \rho^{|\mathcal{S}|}\widehat{f}_{\mathcal{S}}\widehat{g}_{\mathcal{S}} \tag{4.34}\\
&\leqslant \rho\frac{1}{4}\sum_{\mathcal{S}\in\mathcal{P}} \widehat{f}_{\mathcal{S}}\widehat{g}_{\mathcal{S}} \tag{4.35}\\
&= \rho\tau^{+}, \tag{4.36}
\end{aligned}
$$

and similarly one can also show $\theta_\rho \geqslant \rho\tau^{-}$. Combining these bounds with (4.24) yields $\theta_\rho \in \big[-\hat{\theta}(\rho,\bar{a},b), \hat{\theta}(\rho,a,b)\big]$, where

$$
\begin{aligned}
\hat{\theta}(\rho,a,b) &:= \min\big\{a\bar{b}, \rho C_{a,b}\big\}, \text{ with} \tag{4.37}\\
C_{a,b} &:= \frac{a\bar{b} + \sqrt{a\bar{a}b\bar{b}}}{2}. \tag{4.38}
\end{aligned}
$$

The function $\xi(\theta,a,b)$, defined in (4.23), is convex in θ by the concavity of entropy, Lemma 2.8, and consequently

$$
\begin{aligned}
I\big(f(\mathbf{X});g(\mathbf{Y})\big) &= \xi(\theta_\rho,a,b) \tag{4.39}\\
&= \xi(t\hat{\theta}(\rho,a,b) - \bar{t}\hat{\theta}(\rho,\bar{a},b),a,b) \tag{4.40}\\
&\leqslant t\xi(\hat{\theta}(\rho,a,b),a,b) + \bar{t}\xi(-\hat{\theta}(\rho,\bar{a},b),a,b) \tag{4.41}\\
&\leqslant \max_{\theta\in\{-\hat{\theta}(\rho,\bar{a},b),\hat{\theta}(\rho,a,b)\}} \xi(\theta,a,b), \tag{4.42}
\end{aligned}
$$

where we used the convexity of ξ in (4.41) and defined t such that $\theta_\rho = t\hat{\theta}(\rho,a,b) - \bar{t}\hat{\theta}(\rho,\bar{a},b)$. Thus, Theorem 4.2 can be proved by establishing

$$
1 - H\left(\frac{\rho+1}{2}\right) - \xi(\theta,a,b) \geqslant 0 \tag{4.43}
$$

for $\theta \in \{-\hat{\theta}(\rho, \bar{a}, b), \hat{\theta}(\rho, a, b)\}$. Furthermore, it suffices to consider $\frac{1}{2} < a < b < 1$ by continuity of ξ. Define

$$\phi(\rho, a, b) := 1 - H\left(\frac{\rho+1}{2}\right) - \xi(\rho C_{a,b}, a, b). \tag{4.44}$$

Theorem 4.2 can now be reduced to the following lemma.

Lemma 4.4. *For* $0 < \alpha < \beta < 1$ *and* $\rho \in \left[0, \frac{\alpha\bar{\beta}}{C_{\alpha,\beta}}\right]$, $\phi(\rho, \alpha, \beta) \geqslant 0$ *with equality if and only if* $\rho = 0$.

To see this, first observe that we have the identity $\phi(\rho, a, b) = 1 - H\left(\frac{\rho+1}{2}\right) - \xi\big(\hat{\theta}(\rho, a, b), a, b\big)$ for $\rho \in \left[0, \frac{a\bar{b}}{C_{a,b}}\right]$, and for $\rho \in \left[\frac{a\bar{b}}{C_{a,b}}, 1\right]$ we have

$$1 - H\left(\frac{\rho+1}{2}\right) - \xi\big(\hat{\theta}(\rho, a, b), a, b\big) \tag{4.45}$$

$$\overset{(4.37)}{=} 1 - H\left(\frac{\rho+1}{2}\right) - \xi(a\bar{b}, a, b) \tag{4.46}$$

$$\geqslant 1 - H\left(\frac{\frac{a\bar{b}}{C_{a,b}}+1}{2}\right) - \xi(a\bar{b}, a, b) \tag{4.47}$$

$$\overset{(4.44)}{=} \phi\left(\frac{a\bar{b}}{C_{a,b}}, a, b\right), \tag{4.48}$$

where we used the monotonicity of the binary entropy function in (4.47). Thus, for $\theta = \hat{\theta}(\rho, a, b)$ we obtain (4.43) from Lemma 4.4 with $\alpha = a$ and $\beta = b$. Using the fact that in general $\xi(-\theta, \alpha, \beta) = \xi(\theta, \bar{\alpha}, \beta)$, we obtain for $\rho \in \left[0, \frac{\bar{a}\bar{b}}{C_{\bar{a},b}}\right]$

cf. *(4.23) and use* $H(\alpha) = H(\bar{\alpha})$.

$$1 - H\left(\frac{\rho+1}{2}\right) - \xi\big(-\hat{\theta}(\rho, \bar{a}, b), a, b\big) \tag{4.49}$$

$$\overset{(4.37)}{=} 1 - H\left(\frac{\rho+1}{2}\right) - \xi(\rho C_{\bar{a},b}, \bar{a}, b) \tag{4.50}$$

$$\overset{(4.44)}{=} \phi(\rho, \bar{a}, b) \tag{4.51}$$

and for $\rho \in \left[\frac{\bar{a}\bar{b}}{C_{\bar{a},b}}, 1\right]$ we have

$$1 - H\left(\frac{\rho+1}{2}\right) - \xi\big(-\hat{\theta}(\rho, \bar{a}, b), a, b\big) \tag{4.52}$$

$$\overset{(4.37)}{=} 1 - H\left(\frac{\rho+1}{2}\right) - \xi(\bar{a}\bar{b}, \bar{a}, b) \tag{4.53}$$

$$\geqslant 1 - H\left(\frac{\frac{\bar{a}\bar{b}}{C_{\bar{a},b}}+1}{2}\right) - \xi(\bar{a}\bar{b}, \bar{a}, b) \tag{4.54}$$

$$\overset{(4.44)}{=} \phi\left(\frac{\bar{a}\bar{b}}{C_{\bar{a},b}}, \bar{a}, b\right), \tag{4.55}$$

where the monotonicity of the binary entropy function was used in (4.54). Thus, for $\theta = \hat{\theta}(\rho, a, b)$ we obtain (4.43) from Lemma 4.4 with $\alpha = \bar{a}$ and $\beta = b$.

Recall the assumption $\frac{1}{2} < a < b < 1$.

Proof of Lemma 4.4. Let $\mathcal{I} := \{(\alpha, \beta) \in \mathbb{R}^2 : 0 < \alpha < \beta < 1\}$, fix $(\alpha, \beta) \in \mathcal{I}$ and define

$$\rho_- := \frac{\max\{\alpha\beta, \bar{\alpha}\bar{\beta}\}}{C_{\alpha,\beta}}, \quad \rho_\circ := \frac{\min\{\alpha\beta, \bar{\alpha}\bar{\beta}\}}{C_{\alpha,\beta}}, \quad \rho_+ := \frac{\alpha\bar{\beta}}{C_{\alpha,\beta}}. \tag{4.56}$$

We shall adopt the simplified notation $\phi(\rho) := \phi(\rho, \alpha, \beta)$, suppressing the fixed parameters (α, β). For $\rho \in [0, \rho_+)$, we have the derivatives

$\alpha\bar{\beta} < \bar{\alpha}\beta$ follows from $\alpha < \beta$.

$$\begin{aligned}\phi'(\rho) &= \frac{1}{2}\log_2\left(\frac{1+\rho}{1-\rho}\right) \\ &\quad + C_{\alpha,\beta}\log_2\left(\frac{(\bar{\alpha}\beta - C_{\alpha,\beta}\rho)(\alpha\bar{\beta} - C_{\alpha,\beta}\rho)}{(\alpha\beta + C_{\alpha,\beta}\rho)(\bar{\alpha}\bar{\beta} + C_{\alpha,\beta}\rho)}\right) \qquad (4.57)\\ \phi''(\rho) &= \frac{C_{\alpha,\beta}^2}{\log 2}\left(\frac{1}{C_{\alpha,\beta}^2(1-\rho^2)} - \frac{1}{\bar{\alpha}\beta - C_{\alpha,\beta}\rho} - \frac{1}{\alpha\bar{\beta} - C_{\alpha,\beta}\rho}\right. \\ &\qquad \left. - \frac{1}{\bar{\alpha}\bar{\beta} + C_{\alpha,\beta}\rho} - \frac{1}{\alpha\beta + C_{\alpha,\beta}\rho}\right). \qquad (4.58)\end{aligned}$$

Note, that both $\phi'(\rho_+)$ and $\phi''(\rho_+)$ are undefined, but

$$\lim_{\rho \uparrow \rho_+} \phi'(\rho) = \lim_{\rho \uparrow \rho_+} \phi''(\rho) = -\infty. \tag{4.59}$$

Moreover, we have

$$\begin{aligned}\phi''(0) &\overset{(4.58)}{=} \frac{C_{\alpha,\beta}^2}{\log 2}\left(\frac{1}{C_{\alpha,\beta}^2} - \frac{1}{\bar{\alpha}b} - \frac{1}{\alpha\bar{\beta}} - \frac{1}{\bar{\alpha}\bar{\beta}} - \frac{1}{\alpha\beta}\right) && (4.60)\\ &= \frac{C_{\alpha,\beta}^2}{\log 2}\left(\frac{1}{C_{\alpha,\beta}^2} - \frac{\alpha\bar{\beta} + \bar{\alpha}\beta + \alpha\beta + \bar{\alpha}\bar{\beta}}{\alpha\bar{\alpha}\beta\bar{\beta}}\right) && (4.61)\\ &= \frac{C_{\alpha,\beta}^2}{\log 2}\left(\frac{1}{C_{\alpha,\beta}^2} - \frac{1}{\alpha\bar{\alpha}\beta\bar{\beta}}\right) && (4.62)\\ &= \frac{1}{\log 2}\left(1 - \frac{C_{\alpha,\beta}^2}{\alpha\bar{\alpha}\beta\bar{\beta}}\right) && (4.63)\\ &= \frac{1}{\log 2}\left(1 - \left(\frac{\sqrt{\alpha\bar{\beta}} + \sqrt{\bar{\alpha}\beta}}{\sqrt{\bar{\alpha}\beta} + \sqrt{\bar{\alpha}\beta}}\right)^2\right) > 0. && (4.64)\end{aligned}$$

We write $\phi''(\rho) = \frac{p(\rho)}{q(\rho)}$, where both p and q are polynomials in ρ and choose

$$\begin{aligned}q(\rho) &= \log(2)(1-\rho^2)(\bar{\alpha}\beta - C_{\alpha,\beta}\rho) \\ &\qquad \times (\alpha\bar{\beta} - C_{\alpha,\beta}\rho)(\bar{\alpha}\bar{\beta} + C_{\alpha,\beta}\rho)(\alpha\beta + C_{\alpha,\beta}\rho), \qquad (4.65)\end{aligned}$$

satisfying $q(\rho) > 0$ for $\rho \in [0, \rho_+)$. By (4.58), $p(\rho)$ is given by

$$\begin{aligned} p(\rho) = {} & (\bar{\alpha}\beta - C_{\alpha,\beta}\,\rho)(\alpha\bar{\beta} - C_{\alpha,\beta}\,\rho)(\bar{\alpha}\bar{\beta} + C_{\alpha,\beta}\,\rho)(\alpha\beta + C_{\alpha,\beta}\,\rho) \\ & - C_{\alpha,\beta}^2(1-\rho^2)\Big((\alpha\bar{\beta} - C_{\alpha,\beta}\,\rho)(\bar{\alpha}\bar{\beta} + C_{\alpha,\beta}\,\rho)(\alpha\beta + C_{\alpha,\beta}\,\rho) \\ & + (\bar{\alpha}\beta - C_{\alpha,\beta}\,\rho)(\bar{\alpha}\bar{\beta} + C_{\alpha,\beta}\,\rho)(\alpha\beta + C_{\alpha,\beta}\,\rho) \\ & + (\bar{\alpha}\beta - C_{\alpha,\beta}\,\rho)(\alpha\bar{\beta} - C_{\alpha,\beta}\,\rho)(\alpha\beta + C_{\alpha,\beta}\,\rho) \\ & + (\bar{\alpha}\beta - C_{\alpha,\beta}\,\rho)(\alpha\bar{\beta} - C_{\alpha,\beta}\,\rho)(\bar{\alpha}\bar{\beta} + C_{\alpha,\beta}\,\rho)\Big). \end{aligned} \tag{4.66}$$

This entails $\deg(p) \leqslant 5$ and careful calculation of the coefficients reveals $\deg(p) \leqslant 3$.

We will now demonstrate that there is a unique point $\rho^* \in (0, \rho_+)$ with $p(\rho^*) = 0$. To this end, reinterpret $\phi''(\rho)$ as a rational function of ρ on $\mathbb{R}$. By $q(\rho) > 0$, (4.59) and (4.64), we know that the number of zeros of p in $(0, \rho_+)$ is odd and at most equal to its degree, i.e., either one or three. We next show that p has at least one zero in $(-\infty, 0)$, ensuring that there is only one zero in $(0, \rho_+)$. Depending on $\rho_\circ$ (cf. (4.56)), we distinguish three cases:

1. $\rho_\circ < 1$: We have $q(\rho) > 0$ for $\rho \in (-\rho_\circ, 0)$, $\phi''(0) > 0$ and $\lim_{\rho \downarrow -\rho_\circ} \phi''(\rho) = -\infty$. Thus, there is an odd number of zeros in $(-\rho_\circ, 0)$.

2. $\rho_\circ = 1$: Observe that $p(-1) = 0$.

3. $\rho_\circ > 1$: Let $\mathcal{U} := (-\rho_-, -\rho_\circ)$ and observe that $q(\rho) > 0$ for $\rho \in \mathcal{U}$. Thus, there needs to be an odd number of zeros in $\mathcal{U}$ as $\lim_{\rho \downarrow -\rho_-} \phi''(\rho) = -\infty$ and $\lim_{\rho \uparrow -\rho_\circ} \phi''(\rho) = \infty$.

Figures 8a and 8b qualitatively illustrate the behavior of $p(\rho)$ and $\phi''(\rho)$ for cases 1 and 3, respectively.

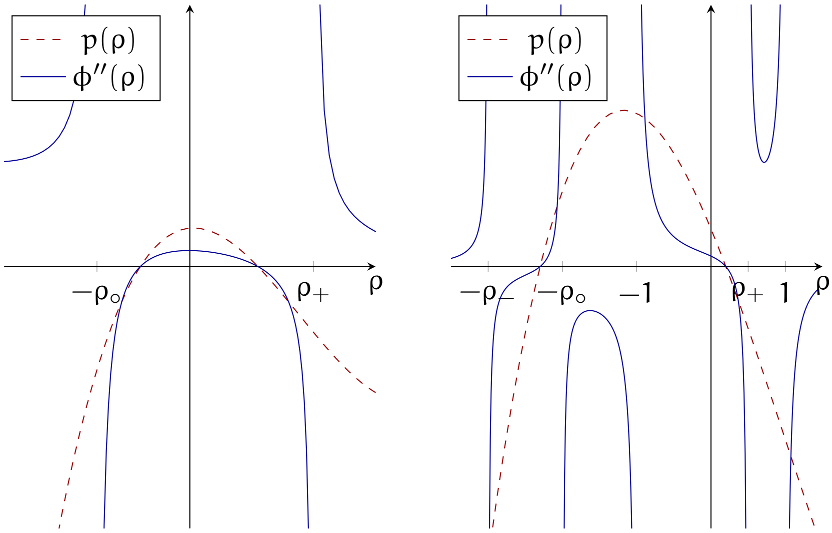
Consequently, $\phi''(\rho) > 0$ for $\rho \in (0, \rho^*)$. By part 1 of Lemma 2.40, $\phi(\rho) > \phi(0) = 0$ for $\rho \in (0, \rho^*]$ as $\phi'(0) = 0$. Since $\phi''(\rho) < 0$ for $\rho \in (\rho^*, \rho_+)$, we have $\phi(\rho) > \min\{\phi(\rho^*), \phi(\rho_+)\}$ for all $\rho \in (\rho^*, \rho_+)$, by part 2 of Lemma 2.40. In total, $\phi(\rho) > \min\{0, \phi(\rho_+)\}$ for $\rho \in (0, \rho_+)$.

As $\phi(0) = 0$, it remains to show $\phi(\rho_+, \alpha, \beta) > 0$ for $(\alpha, \beta) \in \mathcal{I}$. To this end, introduce the transformation

$$(\alpha, \beta) \longmapsto (c, x) := \left(\frac{\log \frac{\alpha}{\beta}}{\log \frac{\alpha\bar{\beta}}{\bar{\alpha}\beta}}, \sqrt{\frac{\alpha\bar{\beta}}{\bar{\alpha}\beta}} \right), \tag{4.67}$$

a bijective mapping from $\mathcal{I}$ to $(0,1)^2$ with the inverse

$$(c, x) \longmapsto (\alpha, \beta) = \left(\frac{x^{2c} - x^2}{1 - x^2}, \frac{1 - x^{2-2c}}{1 - x^2} \right). \tag{4.68}$$



(a) $\rho_\circ < 1$

(b) $\rho_\circ > 1$

Figure 8: Sketch of $p(\rho)$ and $\phi''(\rho)$.

In terms of c and x, we have $\phi(\rho_+, \alpha, \beta) = \psi(c, x)$, where

$$\begin{aligned}\psi(c,x) := 1 - H\left(\frac{1}{2} + \frac{x}{1+x}\right)& \\ - H\left(\frac{x^{2c} - x^2}{1 - x^2}\right) + \frac{1 - x^{2-2c}}{1 - x^2} H\left(x^{2c}\right).& \qquad (4.69)\end{aligned}$$

We fix a particular $x \in (0,1)$ and use the simplified notation $\psi(c) := \psi(c,x)$, obtaining the derivatives

$$\begin{aligned}\psi'(c) &= \frac{2\log(x)}{(x^2-1)\log(2)}\big(2x^{2c}c\log(x) \\ &\qquad + x^{2(1-c)}\log(1-x^{2c}) - x^{2c}\log(x^{2c}-x^2)\big) \qquad (4.70)\\ \psi''(c) &= \frac{4\log(x)^2 x^{2c}}{(1-x^2)\log(2)}\left[\left(\frac{1}{x^{-2(1-c)}-1} + \log(1-x^{2(1-c)})\right)\right. \\ &\qquad \left. + \frac{x^2}{x^{4c}}\left(\log(1-x^{2c}) + \frac{1}{x^{-2c}-1}\right)\right]. \qquad (4.71)\end{aligned}$$

Two applications of Lemma 2.43 yield $\psi''(c) > 0$. Thus, $\psi(c) > \psi(\frac{1}{2})$ by part 1 of Lemma 2.40 as $\psi'(\frac{1}{2}) = 0$. It remains to show, that $\gamma(x) := \psi(\frac{1}{2}, x) > 0$. Note that $\gamma(0) = \gamma(1) = 0$ and

$$\gamma'(x) = \frac{1}{(1+x)^2}\log_2\left((1+3x)(1-x)\right) \qquad (4.72)$$

for $x \in [0,1)$. If $\gamma(x) \leqslant 0$ for any $x \in (0,1)$ then f necessarily attains its minimum in $(0,1)$ and there exists $x^* \in (0,1)$ with $\gamma(x^*) \leqslant 0$ and

$\gamma'(x^*) = 0$. As $x^* = \frac{2}{3}$ is the only point in $(0,1)$ with $\gamma'(x^*) = 0$ and $\gamma\big(\frac{2}{3}\big) = \log_2\big(\frac{27}{25}\big) > 0$, this concludes the proof. □

4.3 PROOF OF PROPOSITION 4.3

We may assume $0 < \rho < 1$ and $\frac{1}{2} \leqslant a \leqslant b \leqslant 1$, using the same reasoning as in (4.10) and (4.11). Clearly, $g = \pm f = \pm \chi_i$ for some $i \in [n]$ is a sufficient condition to maximize $I\big(f(\mathbf{X}); g(\mathbf{Y})\big)$. We will now show that this condition is also necessary.

In the following, we will use the notation of Section 4.2. As $b = 1$ implies $I\big(f(\mathbf{X}); g(\mathbf{Y})\big) = 0 < 1 - H\Big(\frac{1+\rho}{2}\Big)$, we assume $\frac{1}{2} \leqslant a \leqslant b < 1$. For equality in Theorem 4.2, we need

$$1 - H\left(\frac{\rho+1}{2}\right) - \xi(\theta, a, b) = 0 \tag{4.73}$$

Equality requires $\alpha = \beta$ by Lemma 4.4.

for either $\theta = -\hat{\theta}(\rho, \bar{a}, b)$ or $\theta = \hat{\theta}(\rho, a, b)$. For $\theta = -\hat{\theta}(\rho, \bar{a}, b)$, we apply Lemma 4.4 with $\alpha = \bar{a}$ and $\beta = b$ and see that (4.73) can only hold if $a = b = \frac{1}{2}$. For $\theta = \hat{\theta}(\rho, a, b)$ on the other hand, Lemma 4.4 with $\alpha = a$ and $\beta = b$ shows that $a = b$ is necessary for (4.73) to hold. In this case we have $1 - H\Big(\frac{\rho+1}{2}\Big) - \xi(\theta, a, a) = \phi(\rho, a, a)$ and to show that (4.73) additionally implies $a = \frac{1}{2}$, assume $a \neq \frac{1}{2}$, leading to

$$\phi'(\rho) \overset{(4.57)}{=} \frac{1}{2}\log_2\left(\frac{1+\rho}{1-\rho}\right) - a\bar{a}\log_2\left(\frac{\rho}{a\bar{a}\bar{\rho}^2} + 1\right), \tag{4.74}$$

$$\phi''(\rho) \overset{(4.58)}{=} \frac{\rho(1-2a)^2}{\log(2)(a+\rho\bar{a})(1-a\bar{\rho})(1-\rho^2)} > 0. \tag{4.75}$$

Part 1 of Lemma 2.40 now yields $0 = \phi(0, a, a) < \phi(\rho, a, a)$ as $\phi(0) = 0$. The function $\xi(\theta, \frac{1}{2}, \frac{1}{2})$ is strictly convex in θ by Lemma 2.8 and therefore $\theta_\rho = \frac{\langle f, T_\rho g\rangle}{4} \in \{\theta_\rho^+, \theta_\rho^-\} = \pm\frac{\rho}{4}$. Apply Lemma 2.47 to finish the proof.

4.4 DISCUSSION

The key idea underlying the proof of Theorem 4.2 is to split $\theta_1 = \tau^+ + \tau^-$ into its positive and negative part (see Section 4.2). After the problem was reduced to the inequality in Lemma 4.4, the remaining proof is routine analysis. However, Lemma 4.4 might turn out to be useful in the context of other converse proofs concerning the optimization of rate regions with binary random variables.

Although we provided a conclusive and complete proof for the tight upper bound on the mutual information of two Boolean functions, Conjecture 4.1 remains open. Our proof might provide some insight into the general problem. However, it seems unlikely that the idea behind our proof can be applied to fully resolve Conjecture 4.1.

5 CLUSTERING WITH MULTIPLE SOURCES

We now consider the general version of the multi-clustering problem introduced in Section 3.1. Recall that the case of $K = 2$ sources was addressed in Chapter 3. In this chapter we extend the multi-clustering problem to the case of multiple sources and provide bounds on the associated achievable region. We will also discuss several special cases of the problem.

5.1 PROBLEM STATEMENT AND MAIN RESULTS

We start with a formal definition of the multi-clustering problem with K sources. To this end, we reuse some notation by providing natural generalizations of the quantities defined in Chapter 3. A schematic illustration of the problem is shown in Figure 9. Formally, we require the following definitions.

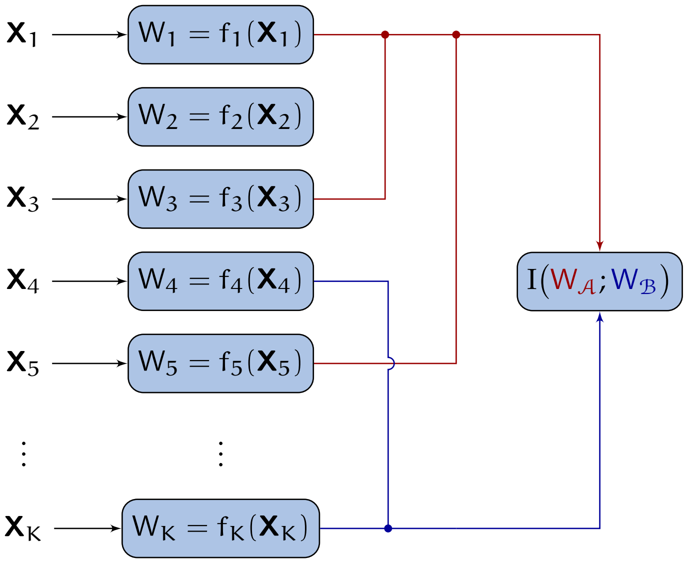

$\mathbf{X}_1$
$W_1 = f_1(\mathbf{X}_1)$
$\mathbf{X}_2$
$W_2 = f_2(\mathbf{X}_2)$
$\mathbf{X}_3$
$W_3 = f_3(\mathbf{X}_3)$
$\mathbf{X}_4$
$W_4 = f_4(\mathbf{X}_4)$
$\mathbf{X}_5$
$W_5 = f_5(\mathbf{X}_5)$
⋮
⋮
$\mathbf{X}_K$
$W_K = f_K(\mathbf{X}_K)$
$I(W_{\mathcal{A}}; W_{\mathcal{B}})$

Figure 9: Clustering of multiple sources.

Definition 5.1. *Consider an* $(n, R_{[K]})$ *code* $f_{[K]}$ *for* $X_{[K]}$ *and define* $W_k := f_k(\mathbf{X}_k)$ *for* $k \in [K]$. *For any* $(\mathcal{A}, \mathcal{B}) \in \Omega$ *we define the co-information of* $f_{\mathcal{A}}$ *and* $f_{\mathcal{B}}$ *as*

Ω is the set of pairs $(\mathcal{A}, \mathcal{B})$ with $\mathcal{A}, \mathcal{B} \subset [K]$ nonempty and disjoint.

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) := \frac{1}{n} I(W_{\mathcal{A}}; W_{\mathcal{B}}). \tag{5.1}$$

Definition 5.2. *A point* $(\mu_\Omega, R_{[K]})$ *is called* achievable *for the sources* $X_{[K]}$ *if and only if, for some* $n \in \mathbb{N}$*, there exists an* $(n, R_{[K]})$ *code* $f_{[K]}$ *for* $X_{[K]}$ *such that for all* $(\mathcal{A}, \mathcal{B}) \in \Omega$*,*

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \geqslant \mu_{\mathcal{A},\mathcal{B}}. \tag{5.2}$$

The achievable region $\overline{\mathcal{R}}$ *is the closure of the set* $\mathcal{R}$ *of all achievable points.*

Remark 6. The region $\mathcal{R}$ is $3^K - 2^{K+1} + K + 1$-dimensional. By exploiting the symmetry of mutual information, i. e., requiring $\mu_{\mathcal{A},\mathcal{B}} = \mu_{\mathcal{B},\mathcal{A}}$, the dimension could be reduced to $\frac{3^K+1}{2} - 2^K + K$. However, we will not make use of this simplification, to keep the formulation of our results shorter and more concise.

Remark 7. A standard time-sharing argument can be used to show that $\overline{\mathcal{R}}$ is a convex set (see, e. g., [16, Section 4.4]).

We first state an outer bound for the achievable region, whose proof is provided in Appendix B.1.

Theorem 5.3. *We have the outer bounds* $\mathcal{R} \subseteq \mathcal{R}_o \subseteq \mathcal{R}'_o$*. Here, the region* $\mathcal{R}'_o$ *is defined as*

$$\mathcal{R}'_o := \bigcup_{U_{[K]}} \Big\{ (\mu_\Omega, R_{[K]}) : \sum_{k \in \mathcal{A}} R_k \geqslant I\big(U_{\mathcal{A}}; X_{[K]} \big| U_{\mathcal{C}}\big) \text{ for } \mathcal{A}, \mathcal{C} \subseteq [K],$$
$$\mu_{\mathcal{A},\mathcal{B}} \leqslant I(U_{\mathcal{A}}; X_{\mathcal{B}}) \text{ for } (\mathcal{A}, \mathcal{B}) \in \Omega \Big\} \tag{5.3}$$

where the auxiliary random variables $U_{[K]}$ *satisfy* $U_{\mathcal{A}} \mathrel{-\!\!\circ\!\!-} X_{\mathcal{A}} \mathrel{-\!\!\circ\!\!-} X_{[K]\setminus\mathcal{A}}$ *for every* $\mathcal{A} \subseteq [K]$*. The region* $\mathcal{R}_o$ *is defined as* $\mathcal{R}'_o$ *only that the inequality for the relevance* $\mu_{\mathcal{A},\mathcal{B}}$ *is replaced with*

$$\mu_{\mathcal{A},\mathcal{B}} \leqslant I(U_{\mathcal{A}}; X_{\mathcal{A}}) + I(U_{\mathcal{B}}; X_{\mathcal{B}}) - I(U_{\mathcal{A}} U_{\mathcal{B}}; X_{\mathcal{A}} X_{\mathcal{B}}). \tag{5.4}$$

Remark 8. In particular we have the Markov chains $U_{\mathcal{A}} \mathrel{-\!\!\circ\!\!-} X_{\mathcal{A}} \mathrel{-\!\!\circ\!\!-} X_{\mathcal{B}}$ and $U_{\mathcal{B}} \mathrel{-\!\!\circ\!\!-} X_{\mathcal{B}} \mathrel{-\!\!\circ\!\!-} X_{\mathcal{A}}$. Using Lemma 2.13 we can write (5.4) equivalently as

$$\mu_{\mathcal{A},\mathcal{B}} \leqslant I(U_{\mathcal{A}}; U_{\mathcal{B}}) - I(U_{\mathcal{A}}; U_{\mathcal{B}} | X_{\mathcal{A}} X_{\mathcal{B}}). \tag{5.5}$$

The next result provides an inner bound for $\overline{\mathcal{R}}$.

Theorem 5.4. *We have* $\mathcal{R}_i \subseteq \overline{\mathcal{R}}$ *where the region* $\mathcal{R}_i$ *consists of all points* $(\mu_\Omega, R_{[K]})$ *for which there exist random variables* $U_{[K]}$ *satisfying* $U_k \mathrel{-\!\!\circ\!\!-} X_k \mathrel{-\!\!\circ\!\!-} (X_{[K]\setminus k}, U_{[K]\setminus k})$ *for all* $k \in [K]$ *and for all* $(\mathcal{A}, \mathcal{B}) \in \Omega$ *there exist subsets* $\widetilde{\mathcal{A}} \subseteq \mathcal{A}$ *and* $\widetilde{\mathcal{B}} \subseteq \mathcal{B}$ *such that*

The index sets $\widetilde{\mathcal{A}}$ *and* $\widetilde{\mathcal{B}}$ *implicitly depend on the pair* $(\mathcal{A}, \mathcal{B})$*.*

$$\sum_{k \in \hat{\mathcal{A}}} R_k \geqslant I\Big(X_{\hat{\mathcal{A}}}; U_{\hat{\mathcal{A}}} \Big| U_{\mathcal{A}\setminus\hat{\mathcal{A}}}\Big) \text{ for all } \hat{\mathcal{A}} \subseteq \mathcal{A} \text{ with } \hat{\mathcal{A}} \cap \widetilde{\mathcal{A}} \neq \varnothing, \tag{5.6}$$

$$\sum_{k \in \hat{\mathcal{B}}} R_k \geqslant I\Big(X_{\hat{\mathcal{B}}}; U_{\hat{\mathcal{B}}} \Big| U_{\mathcal{B}\setminus\hat{\mathcal{B}}}\Big) \text{ for all } \hat{\mathcal{B}} \subseteq \mathcal{B} \text{ with } \hat{\mathcal{B}} \cap \widetilde{\mathcal{B}} \neq \varnothing, \tag{5.7}$$

$$\mu_{\mathcal{A},\mathcal{B}} \leqslant I\big(U_{\widetilde{\mathcal{A}}}; U_{\widetilde{\mathcal{B}}}\big). \tag{5.8}$$

The proof of Theorem 5.4 will be provide in Section 5.3.

In contrast to the case of two sources, binning does help for $K > 2$ sources. For illustration, consider the case $K = 3$ and assume we are only interested in maximizing $\Theta(f_1, f_2; f_3)$. Then any information encoded by both f_1 and f_2 is redundant as it does not increase $\Theta(f_1, f_2; f_3)$. The corresponding rate loss can be reduced by a quantize-and-bin scheme (cf. [2], [6], [64], [70]).

The proof that $\mathcal{R}_i$ is indeed achievable uses typicality coding and binning. The conditions (5.6) and (5.7) ensure that the messages of encoders $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ can be correctly decoded from the output of the encoders $\mathcal{A}$ and $\mathcal{B}$, respectively. By (5.8), these suffice to ensure that $\mu_{\mathcal{A},\mathcal{B}}$ is achievable. Intuitively, the encoders $\mathcal{A} \setminus \widetilde{\mathcal{A}}$ and $\mathcal{B} \setminus \widetilde{\mathcal{B}}$ act as helpers. The special case $\widetilde{\mathcal{A}} = \mathcal{A}$, $\widetilde{\mathcal{B}} = \mathcal{B}$ for every $(\mathcal{A}, \mathcal{B}) \in \Omega$ corresponds to no binning at all, as (5.6) and (5.7) then imply $R_k \geqslant I(X_k; U_k)$ for all $k \in [K]$.

Finally, note that in general $\mathcal{R}_i$ is not convex and thus Theorem 5.4 can be strengthened to $\operatorname{conv}(\mathcal{R}_i) \subseteq \overline{\mathcal{R}}$. However, it is tedious to characterize $\operatorname{conv}(\mathcal{R}_i)$ using a time-sharing random variable due to the freedom of choosing the index sets $\widetilde{\mathcal{A}}$, and $\widetilde{\mathcal{B}}$ for each $(\mathcal{A}, \mathcal{B}) \in \Omega$.

The following cardinality bounds show that $\mathcal{R}_i$ is computable (see Appendix B.2 for the proof).

Proposition 5.5. *The region $\mathcal{R}_i$ remains unchanged if the cardinality bound $|\mathcal{U}_k| \leqslant |\mathcal{X}_k| + 4^K$ is imposed for every $k \in [K]$.*

5.2 A SPECIAL CASE: THE CEO PROBLEM

In this section we study a special case of the clustering problem that corresponds to a variant of the CEO problem [7] in which the usual distortion criterion is replaced with mutual information. This problem turns out to be equivalent to the classical CEO problem with logarithmic loss distortion as analyzed in [11]. The equivalence follows in the same fashion as the equivalence between the information bottleneck problem and lossy source coding with logarithmic loss distortion, shown in Section 3.4. Using results from [11], we will show that our inner bound becomes tight in this special case.

We consider a CEO problem under a mutual information (MI) constraint where random variables $X_{[J]}$ are encoded to be maximally informative about another set of random variables $Y_{[L]}$.

Definition 5.6. *We say that the point $(\nu_\Pi, R_{[J]})$ is* MI-achievable *for the source $(X_{[J]}, Y_{[L]})$ if and only if, for some $n \in \mathbb{N}$, there exists an $(n, R_{[J]})$ code $f_{[J]}$ for $X_{[J]}$ such that for all $(\mathcal{A}, \mathcal{B}) \in \Pi$, we have*

Π is the set of all pairs $(\mathcal{A}, \mathcal{B})$ of nonempty sets $\mathcal{A} \subseteq [J]$ and $\mathcal{B} \subseteq [L]$.

$$\frac{1}{n} I(W_{\mathcal{A}}; \mathbf{Y}_{\mathcal{B}}) \geqslant \nu_{\mathcal{A},\mathcal{B}}, \tag{5.9}$$

where $W_j := f_j(\mathbf{X}_j)$ for $j \in [J]$. The set of all MI-achievable points is $\mathcal{R}_{MI}$.

Remark 9. The achievable region $\overline{\mathcal{R}_{\mathrm{MI}}}$ is convex by a standard time-sharing argument (see, e. g., [16, Section 4.4]).

We can show that $\mathcal{R}_{\mathrm{MI}}$ corresponds to a subset of $\mathcal{R}$.

Proposition 5.7. *Letting* $K := J + L$ *and* $X_{[K]} := (X_{[J]}, Y_{[L]})$*, we have*

$R_{[J]}$ *is a slice of* $R_{[K]}$.

$$\mathcal{R}_{\mathrm{MI}} = \bigcup_{(\mu_\Omega, R_{[K]}) \in \mathcal{R}} \big\{ (\nu_\Pi, R_{[J]}) : \nu_{\mathcal{A},\mathcal{B}} = \mu_{\mathcal{A},J+\mathcal{B}}, \ \ (\mathcal{A},\mathcal{B}) \in \Pi \big\}. \tag{5.10}$$

Proof. Let $(\nu_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}$ be achieved by the $(n, R_{[J]})$ code $f_{[J]}$ in the sense of Definition 5.6. We extend $f_{[J]}$ to an $(n, R_{[K]})$ code $f_{[K]}$ for $X_{[K]}$ by setting $f_k(\mathbf{x}_k) = \mathbf{x}_k$ and choosing $R_k = \infty$ for $k > J$. Set $\mu_{\mathcal{A},J+\mathcal{B}} = \nu_{\mathcal{A},\mathcal{B}}$ whenever $(\mathcal{A},\mathcal{B}) \in \Pi$ and $\mu_{\mathcal{A},J+\mathcal{B}} = 0$ otherwise. The code $f_{[K]}$ achieves $(\mu_\Omega, R_{[K]}) \in \mathcal{R}_{\mathrm{MI}}$ (cf. Definition 5.2).

Let $(\mu_\Omega, R_{[K]}) \in \mathcal{R}$ be achieved by the $(n, R_{[K]})$ code $f_{[K]}$ (cf. Definition 5.2). The restriction $f_{[J]}$ then provides $(\nu_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}$ with $\nu_{\mathcal{A},\mathcal{B}} = \mu_{\mathcal{A},J+\mathcal{B}}$ for $(\mathcal{A},\mathcal{B}) \in \Pi$ (cf. Definition 5.6). □

To shorten notation we will introduce the set of random variables

$$\begin{aligned}\mathcal{P}_* := \big\{ & (U_{[J]}, Q) : Q \perp X_{[J]} Y_{[L]}, \\ & U_j \multimap X_j Q \multimap X_{[J]\setminus j} Y_{[L]} U_{[J]\setminus j} \ \text{ for all } j \in [J] \big\}.\end{aligned} \tag{5.11}$$

Using the extended source $X_{[K]} = (X_{[J]}, Y_{[L]})$, we can obtain an inner bound on $\overline{\mathcal{R}_{\mathrm{MI}}}$ directly from Theorem 5.4 as stated in the following corollary.

Corollary 5.8. *An inner bound for the achievable region is* $\mathcal{R}_i^{(\mathrm{MI})} \subseteq \overline{\mathcal{R}_{\mathrm{MI}}}$ *where the region* $\mathcal{R}_i^{(\mathrm{MI})}$ *is obtained directly from* $\mathcal{R}_i$ *by*

$$\mathcal{R}_i^{(\mathrm{MI})} := \bigcup_{(\mu_\Omega, R_{[K]}) \in \mathcal{R}_i} \big\{ (\nu_\Pi, R_{[J]}) : \nu_{\mathcal{A},\mathcal{B}} = \mu_{\mathcal{A},J+\mathcal{B}}, \ \ (\mathcal{A},\mathcal{B}) \in \Pi \big\}. \tag{5.12}$$

Furthermore, $\mathcal{R}_i^{(\mathrm{MI})}$ *consists of all points* $(\nu_\Pi, R_{[J]})$ *such that there exist random variables* $(U_{[J]}, \varnothing) \in \mathcal{P}_*$ *and for every* $(\mathcal{A},\mathcal{B}) \in \Pi$ *there is a set* $\widetilde{\mathcal{A}} \subseteq \mathcal{A}$ *with*

$$\sum_{k \in \hat{\mathcal{A}}} R_k \geqslant I(X_{\hat{\mathcal{A}}}; U_{\hat{\mathcal{A}}} | U_{\mathcal{A} \setminus \hat{\mathcal{A}}}) \ \textit{for all } \hat{\mathcal{A}} \subseteq \mathcal{A} \textit{ with } \hat{\mathcal{A}} \cap \widetilde{\mathcal{A}} \neq \varnothing, \tag{5.13}$$

$$\nu_{\mathcal{A},\mathcal{B}} \leqslant I(U_{\widetilde{\mathcal{A}}}; Y_{\mathcal{B}}). \tag{5.14}$$

Proof. The result follows from Theorem 5.4 and Proposition 5.7 with auxiliaries $U_k = X_k$ for $k > J$. □

Using Theorem 5.3 and Proposition 5.7, one can also formulate a corresponding outer bound.

We next argue that the CEO problem introduced in Definition 5.6 is equivalent to the logarithmic loss distortion approach described in Section 2.2.2.

Lemma 5.9. *We have* $(D_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{LL}}$ *if and only if* $(\nu_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}$, *where* $\nu_{\mathcal{A},\mathcal{B}} = H(Y_\mathcal{B}) - D_{\mathcal{A},\mathcal{B}}$ *for all* $(\mathcal{A}, \mathcal{B}) \in \Pi$.

Proof. Assume that $(\nu_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}$ is achieved by the $(n, R_{[J]})$ code $f_{[J]}$, i. e., (5.9) holds for all $(\mathcal{A}, \mathcal{B}) \in \Pi$ with $W_j := f_j(\mathbf{X}_j)$. We choose the decoding functions $g_{\mathcal{A},\mathcal{B}}(w_\mathcal{A}) = p_{\mathbf{Y}_\mathcal{B}|W_\mathcal{A}}(\,\cdot\,|w_\mathcal{A})$ and Lemma 2.30 shows

$$\frac{1}{n}\mathbb{E}\left[d_{\mathrm{LL}}(g_{\mathcal{A},\mathcal{B}}(W_\mathcal{A}), \mathbf{Y}_\mathcal{B})\right] = \frac{1}{n}H(\mathbf{Y}_\mathcal{B}|W_\mathcal{A}) \tag{5.15}$$

$$\overset{(5.9)}{\leqslant} D_{\mathcal{A},\mathcal{B}}, \tag{5.16}$$

implying $(D_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{LL}}$. To show $\mathcal{R}_{\mathrm{LL}} \subseteq \mathcal{R}_{\mathrm{MI}}$, let $(D_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{LL}}$ be achieved by the $(n, R_{[J]})$ code $f_{[J]}$ and the decoding function $g_{\mathcal{A},\mathcal{B}}$, i. e., (2.60) holds for all $(\mathcal{A}, \mathcal{B}) \in \Pi$ with $W_j := f_j(\mathbf{X}_j)$. Lemma 2.30 then implies $I(W_\mathcal{A}; \mathbf{Y}_\mathcal{B}) \geqslant n\nu_{\mathcal{A},\mathcal{B}}$ and hence $(\nu_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}$. □

For the rest of this section, we assume $L = 1$ and for brevity write $Y := Y_1$ and $\nu_\mathcal{A} := \nu_{\mathcal{A},1}$. We first note the following connection between the Körner-Marton modulo-two sum problem (cf. Section 2.2.2) and the CEO problem with logarithmic loss distortion (cf. Section 2.2.2).

Theorem 5.10. *For* $J = 2$, $(X_1, X_2) \sim \mathrm{DSBS}(p)$, *and* $Y = X_1 \oplus X_2$, *we have*

$$\begin{aligned}\overline{\mathcal{R}_{\mathrm{MI}}} = \big\{&(\nu_1, \nu_2, \nu_{\{1,2\}}, R_1, R_2) \in \mathbb{R}^5 : \nu_1, \nu_2 \leqslant 0,\\ &R_1, R_2 \geqslant 0, \textit{ and } \nu_{\{1,2\}} \leqslant \min\{R_1, R_2, H(Y)\}\big\}.\end{aligned} \tag{5.17}$$

Proof. Assuming $(\nu_\Pi, R_1, R_2) \in \mathcal{R}_{\mathrm{MI}}$, we immediately have $\nu_1, \nu_2 \leqslant 0$, which follows from $Y \perp X_1$ and $Y \perp X_2$. Applying Definition 5.6, choose an (n, R_1, R_2) code (f_1, f_2) for (X_1, X_2) that achieves $n\nu_{\{1,2\}} \leqslant I(f_1(\mathbf{X}_1)f_2(\mathbf{X}_2); \mathbf{Y}) \leqslant nH(Y)$. Introducing $W_1 := f_1(\mathbf{X}_1)$ and $W_2 := f_2(\mathbf{X}_2)$, we have

$$\begin{aligned}
n\nu_{\{1,2\}} &\leqslant I(W_1W_2; \mathbf{Y}) && (5.18)\\
&\leqslant I(W_1\mathbf{X}_2; \mathbf{Y}) && (5.19)\\
&= I(\mathbf{X}_2; \mathbf{Y}) + I(W_1; \mathbf{Y}|\mathbf{X}_2) && (5.20)\\
&\leqslant H(W_1) && (5.21)\\
&\leqslant nR_1, && (5.22)
\end{aligned}$$

where (5.19) follows from $\mathbf{Y} \mathbin{-\!\!\circ\!\!-} W_1\mathbf{X}_2 \mathbin{-\!\!\circ\!\!-} W_2$ and (5.21) from $\mathbf{Y} \perp \mathbf{X}_2$. The inequality $\nu_{\{1,2\}} \leqslant R_2$ can be proved the same way, by interchanging X_1 and X_2.

On the other hand, assume $\nu_{\{1,2\}} \leqslant \min\{R_1, R_2, H(Y)\}$. By Theorem 2.32 this shows $(H(Y) - \nu_{\{1,2\}}, R_1, R_2) \in \mathcal{R}_{\mathrm{TO}}$. We pick an arbitrary $\varepsilon > 0$ and, using Definition 2.31, we can find an $(n, H(Y) - \nu_{\{1,2\}}, R_1, R_2)$ code (f_0, f_1, f_2) for (Y, X_1, X_2) and a decoding function

$\mathcal{R}_{\mathrm{TO}}$ is defined in Section 2.2.2.

g such that $P\{g(f_0(\mathbf{Y}), f_1(\mathbf{X}_1), f_2(\mathbf{X}_2)) \neq \mathbf{Y}\} \leqslant \varepsilon$. Applying Fano's inequality, Theorem 2.6, yields

$$\begin{aligned} nH(Y) - H(\varepsilon) - \varepsilon &\leqslant I\big(\mathbf{Y}; f_0(\mathbf{Y}) f_1(\mathbf{X}_1) f_2(\mathbf{X}_2)\big) && (5.23)\\ &= I\big(\mathbf{Y}; f_1(\mathbf{X}_1) f_2(\mathbf{X}_2)\big) + I\big(\mathbf{Y}; f_0(\mathbf{Y}) \big| f_1(\mathbf{X}_1) f_2(\mathbf{X}_2)\big) && (5.24)\\ &\leqslant I\big(\mathbf{Y}; f_1(\mathbf{X}_1) f_2(\mathbf{X}_2)\big) + n\big(H(Y) - \nu_{\{1,2\}}\big). && (5.25) \end{aligned}$$

As ε was arbitrary, this completes the proof. □

We can now show that the inner bound in Theorem 5.4 cannot be tight due to the Körner-Marton counterexample. For $J = 2$, consider $(X_1, X_2) \sim \mathrm{DSBS}(p)$ and $Y = X_1 \oplus X_2$ with $p \in (0, 1)$ and $p \neq \frac{1}{2}$. The inner bound $\mathcal{R}_{\mathrm{i}}^{(\mathrm{MI})}$ in Corollary 5.8 specializes to all points $(\nu_1, \nu_2, \nu_{\{1,2\}}, R_1, R_2)$ with

$$\begin{aligned} R_1 &\geqslant I(X_1; U_1 | U_2), && (5.26)\\ R_2 &\geqslant I(X_2; U_2 | U_1), && (5.27)\\ R_1 + R_2 &\geqslant I(X_1 X_2; U_1 U_2), && (5.28)\\ \nu_{\{1,2\}} &\leqslant I(U_1 U_2; Y), && (5.29)\\ \nu_1 &\leqslant 0, && (5.30)\\ \nu_2 &\leqslant 0, && (5.31) \end{aligned}$$

for random variables U_1 and U_2, satisfying $U_1 \mathrel{-\!\!\circ\!\!-} X_1 \mathrel{-\!\!\circ\!\!-} X_2 \mathrel{-\!\!\circ\!\!-} U_2$. Using the characterization of $\overline{\mathcal{R}_{\mathrm{MI}}}$ in Theorem 5.10, we have $\overline{\mathcal{R}_{\mathrm{MI}}} \neq \mathrm{conv}\big(\mathcal{R}_{\mathrm{i}}^{(\mathrm{MI})}\big)$ by Theorem 2.33. Inequality in this special case shows that also $\overline{\mathcal{R}} \neq \mathrm{conv}(\mathcal{R}_{\mathrm{i}})$.

If we additionally assume $X_j \mathrel{-\!\!\circ\!\!-} Y_1 \mathrel{-\!\!\circ\!\!-} X_{[J]\setminus j}$ for all $j \in [J]$, the results in [11] directly apply to the CEO problem with a mutual information constraint as consequence of Lemma 5.9.

Lemma 5.11. *Assume* $\nu_{\mathcal{A}} = 0$ *whenever* $\mathcal{A} \neq [J]$*. Then* $(\nu_\Pi, R_{[J]}) \in \overline{\mathcal{R}_{\mathrm{MI}}}$ *if and only if there exist random variables* $(U_{[J]}, Q) \in \mathcal{P}_*$ *and the following inequalities hold:*

$$\begin{aligned} \sum_{k \in \hat{\mathcal{A}}} R_k &\geqslant I\big(X_{\hat{\mathcal{A}}}; U_{\hat{\mathcal{A}}} \big| U_{[J]\setminus\hat{\mathcal{A}}}, Q\big) \quad \textit{for all } \hat{\mathcal{A}} \subseteq [J], && (5.32)\\ \nu_{[J]} &\leqslant I\big(U_{[J]}; Y \big| Q\big). && (5.33) \end{aligned}$$

Proof. $(\nu_\Pi, R_{[J]}) \in \overline{\mathcal{R}_{\mathrm{MI}}}$ follows by applying Corollary 5.8 with $\widetilde{\mathcal{A}} = \varnothing$ for $\mathcal{A} \neq [J]$ and $\widetilde{\mathcal{A}} = [J]$ for $\mathcal{A} = [J]$, taking into account the convexity of $\overline{\mathcal{R}_{\mathrm{MI}}}$ (Remark 9). The converse follows from Theorem 2.29 and Lemma 5.9. □

Remark 10. The achievable region of the multiterminal source coding problem with logarithmic loss distortion, introduced in [11, Section II], can be obtained as a special case of $\mathcal{R}_{\mathrm{MI}}$ as well. Choose

$J = L = 2$ and set $Y_j = X_j$, $j \in \{1, 2\}$. The inner bound $\mathcal{R}_i^{(\mathrm{MI})}$ is also tight (up to convexification) due to the results in [11].

5.3 PROOF OF THEOREM 5.4

The proof of Theorem 5.4 extends the methods developed in [25] for the hypothesis testing problem (cf. Section 2.2.2) to a setup with multiple sources. We begin by extending [25, Lemma 8] and incorporating a binning strategy.

Lemma 5.12. *Let* $\varepsilon > 0$ *and assume* $U_k \multimap X_k \multimap (X_{[K]\setminus k}, U_{[K]\setminus k})$ *for all* $k \in [K]$, *and* $R_{[K]} \in \mathbb{R}_+^K$. *Then, for sufficiently large* $n \in \mathbb{N}$ *we can obtain an* $(n, R_{[K]} + \varepsilon)$ *code* $f_{[K]}$ *with* $W_k := f_k(\mathbf{X}_k)$ *and decoding functions* $g_{\mathcal{A},\widetilde{\mathcal{A}}} \colon \mathcal{M}_{\mathcal{A}} \to \mathcal{U}_{\widetilde{\mathcal{A}}}^n$ *for each* $\mathcal{A}, \widetilde{\mathcal{A}} \subseteq [K]$ *with* $\varnothing \neq \widetilde{\mathcal{A}} \subseteq \mathcal{A}$ *such that the following two properties hold.*

For every $(\mathcal{A}, \mathcal{B}) \in \Omega$ *and* $\varnothing \neq \widetilde{\mathcal{A}} \subseteq \mathcal{A}$ *as well as* $\varnothing \neq \widetilde{\mathcal{B}} \subseteq \mathcal{B}$ *we have*

$$\left|\left(g_{\mathcal{A},\widetilde{\mathcal{A}}}(\mathcal{M}_{\mathcal{A}}) \times g_{\mathcal{B},\widetilde{\mathcal{B}}}(\mathcal{M}_{\mathcal{B}})\right) \cap \mathcal{T}^n_{[U_{\widetilde{\mathcal{A}}} U_{\widetilde{\mathcal{B}}}]}\right| \leqslant 2^{n\left(I(U_{\widetilde{\mathcal{A}}} U_{\widetilde{\mathcal{B}}}; X_{\widetilde{\mathcal{A}}} X_{\widetilde{\mathcal{B}}}) + \varepsilon\right)}. \tag{5.34}$$

Furthermore, if (5.6) *and* (5.7) *hold, then*

$$P\left\{\left(g_{\mathcal{A},\widetilde{\mathcal{A}}}(W_{\mathcal{A}}), \mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{B}}, g_{\mathcal{B},\widetilde{\mathcal{B}}}(W_{\mathcal{B}})\right) \notin \mathcal{T}^n_{[U_{\widetilde{\mathcal{A}}} X_{\mathcal{A}} X_{\mathcal{B}} U_{\widetilde{\mathcal{B}}}]}\right\} \leqslant \varepsilon. \tag{5.35}$$

The proof of Lemma 5.12 is provided in Appendix B.3.

Furthermore we will need the following set of random variables.

Definition 5.13. *For random variables* (A, B, C, D) *and* $\delta \geqslant 0$, *define the set of random variables*

$$\begin{aligned} \mathcal{S}_\delta(A, B, C, D) := \big\{\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D} : (\widetilde{A}, \widetilde{B}) \in \mathcal{T}_{[AB]\delta}, \\ (\widetilde{C}, \widetilde{D}) \in \mathcal{T}_{[CD]\delta}, (\widetilde{A}, \widetilde{D}) \in \mathcal{T}_{[AD]\delta}\big\}. \end{aligned} \tag{5.36}$$

Consider $(\mu_\Omega, R_{[K]}) \in \mathcal{R}_i$ and choose $U_{[K]}$ as in Theorem 5.4. Fix $\varepsilon > 0$ and apply Lemma 5.12 to obtain encoding functions $f_{[K]}$ and decoding functions $g_{\mathcal{A},\widetilde{\mathcal{A}}}$. For any pair $(\mathcal{A}, \mathcal{B}) \in \Omega$, find the nonempty subsets $\widetilde{\mathcal{A}} \subseteq \mathcal{A} \subseteq \mathcal{A}$ and $\widetilde{\mathcal{B}} \subseteq \mathcal{B} \subseteq \mathcal{B}$ such that (5.6)–(5.8) hold. (The case $\widetilde{\mathcal{A}} = \varnothing$ or $\widetilde{\mathcal{B}} = \varnothing$ can be ignored since due to (5.8) it leads to $\mu_{\mathcal{A},\mathcal{B}} \leqslant 0$, which is achieved by any code.) Define the functions $h_1 := g_{\mathcal{A},\widetilde{\mathcal{A}}} \circ f_{\mathcal{A}}$ and $h_2 := g_{\mathcal{B},\widetilde{\mathcal{B}}} \circ f_{\mathcal{B}}$. To analyze $\Theta(f_{\mathcal{A}}; f_{\mathcal{B}})$, we define $\mathcal{D}_1 := h_1(\mathcal{X}^n_{\mathcal{A}})$ and partition $\mathcal{X}^n_{\mathcal{A}}$ as $\mathcal{X}^n_{\mathcal{A}} = \bigcup_{\mathbf{u}_{\widetilde{\mathcal{A}}} \in \mathcal{D}_1} h_1^{-1}(\mathbf{u}_{\widetilde{\mathcal{A}}})$. We may assume without loss of generality that $h_1^{-1}(\mathbf{u}_{\widetilde{\mathcal{A}}}) \subseteq \mathcal{T}^n_{[X_{\mathcal{A}}|U_{\widetilde{\mathcal{A}}}]}(\mathbf{u}_{\widetilde{\mathcal{A}}})$ whenever $\mathbf{u}_{\widetilde{\mathcal{A}}} \in \mathcal{T}^n_{[U_{\widetilde{\mathcal{A}}}]}$ as this does not interfere with the properties (5.34) and (5.35) of the code. Defining $\mathcal{D}_2$ accordingly, we set $\mathcal{F} := (\mathcal{D}_1 \times \mathcal{D}_2) \cap \mathcal{T}^n_{[U_{\widetilde{\mathcal{A}}} U_{\widetilde{\mathcal{B}}}]}$. Let us use the shorthand notation $\hat{\mathbf{U}}_1 := h_1(\mathbf{X}_{\mathcal{A}})$ and $\hat{\mathbf{U}}_2 := h_2(\mathbf{X}_{\mathcal{B}})$, and define $p_{\mathcal{F}} := P\{(\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2) \in \mathcal{F}\}$ and

$(\mathbf{X}^*_{\mathcal{A}}, \mathbf{X}^*_{\mathcal{B}})$ are n i.i.d. copies of $(X^*_{\mathcal{A}}, X^*_{\mathcal{B}})$.

$(X^*_{\mathcal{A}}, X^*_{\mathcal{B}})$ are independent, with the same marginals as $(X_{\mathcal{A}}, X_{\mathcal{B}})$.

$q_{\mathcal{F}} := P\{(\mathbf{U}^*_1, \mathbf{U}^*_2) \in \mathcal{F}\}$ with $\mathbf{U}^*_1 := h_1(\mathbf{X}^*_{\mathcal{A}})$, $\mathbf{U}^*_2 := h_2(\mathbf{X}^*_{\mathcal{B}})$, where $(X^*_{\mathcal{A}}, X^*_{\mathcal{B}}) \sim p_{X_{\mathcal{A}}} p_{X_{\mathcal{B}}}$. We then have

$$n\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \geqslant n\Theta(h_1; h_2) = I(h_1(\mathbf{X}_{\mathcal{A}}); h_2(\mathbf{X}_{\mathcal{B}})) \tag{5.37}$$

$$= \sum_{\mathbf{u}_{\tilde{\mathcal{A}}} \in \mathcal{D}_1, \mathbf{u}_{\tilde{\mathcal{B}}} \in \mathcal{D}_2} p_{\hat{U}_1 \hat{U}_2}(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}) \log_2 \frac{p_{\hat{U}_1 \hat{U}_2}(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}})}{p_{\hat{U}_1}(\mathbf{u}_{\tilde{\mathcal{A}}}) p_{\hat{U}_2}(\mathbf{u}_{\tilde{\mathcal{B}}})} \tag{5.38}$$

$$= \sum_{(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}) \in \mathcal{F}} p_{\hat{U}_1 \hat{U}_2}(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}) \log_2 \frac{p_{\hat{U}_1 \hat{U}_2}(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}})}{p_{\hat{U}_1}(\mathbf{u}_{\tilde{\mathcal{A}}}) p_{\hat{U}_2}(\mathbf{u}_{\tilde{\mathcal{B}}})}$$

$$+ \sum_{(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}) \in \mathcal{F}^c} p_{\hat{U}_1 \hat{U}_2}(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}) \log_2 \frac{p_{\hat{U}_1 \hat{U}_2}(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}})}{p_{\hat{U}_1}(\mathbf{u}_{\tilde{\mathcal{A}}}) p_{\hat{U}_2}(\mathbf{u}_{\tilde{\mathcal{B}}})} \tag{5.39}$$

$$\geqslant p_{\mathcal{F}} \log_2 \frac{p_{\mathcal{F}}}{q_{\mathcal{F}}} + (1 - p_{\mathcal{F}}) \log_2 \frac{1 - p_{\mathcal{F}}}{1 - q_{\mathcal{F}}} \tag{5.40}$$

$$= -H(p_{\mathcal{F}}) - p_{\mathcal{F}} \log_2 q_{\mathcal{F}} - (1 - p_{\mathcal{F}}) \log_2 (1 - q_{\mathcal{F}}) \tag{5.41}$$

$$\geqslant -1 - p_{\mathcal{F}} \log_2 q_{\mathcal{F}} \tag{5.42}$$

$$\overset{(5.35)}{\geqslant} -1 - (1 - \varepsilon) \log_2 q_{\mathcal{F}}. \tag{5.43}$$

Furthermore, (5.37) follows from Theorem 2.5 and (5.40) is a consequence of Theorem 2.7. For each $\mathbf{u}_{\tilde{\mathcal{A}}} \in \mathcal{D}_1$ and $\mathbf{u}_{\tilde{\mathcal{B}}} \in \mathcal{D}_2$ define

$$\mathcal{S}(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}) := \{\mathbf{u}_{\tilde{\mathcal{A}}}\} \times h_1^{-1}(\mathbf{u}_{\tilde{\mathcal{A}}}) \times h_2^{-1}(\mathbf{u}_{\tilde{\mathcal{B}}}) \times \{\mathbf{u}_{\tilde{\mathcal{B}}}\} \tag{5.44}$$

and

$$\mathcal{S} := \bigcup_{(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}) \in \mathcal{F}} \mathcal{S}(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}). \tag{5.45}$$

Now, pick any $(\hat{\mathbf{u}}_{\tilde{\mathcal{A}}}, \hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{x}}_{\mathcal{B}}, \hat{\mathbf{u}}_{\tilde{\mathcal{B}}}) \in \mathcal{S}$. Let $\hat{U}_{\tilde{\mathcal{A}}}$, $\hat{X}_{\mathcal{A}}$, $\hat{X}_{\mathcal{B}}$, and $\hat{U}_{\tilde{\mathcal{B}}}$ be the type variables corresponding to $\hat{\mathbf{u}}_{\tilde{\mathcal{A}}}$, $\hat{\mathbf{x}}_{\mathcal{A}}$, $\hat{\mathbf{x}}_{\mathcal{B}}$, and $\hat{\mathbf{u}}_{\tilde{\mathcal{B}}}$, respectively. From part 1 of Lemma 2.16 we know

$$p_{\mathbf{X}^*_{\mathcal{A}} \mathbf{X}^*_{\mathcal{B}}}(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{x}}_{\mathcal{B}}) = 2^{-n\left(H(\hat{X}_{\mathcal{A}} \hat{X}_{\mathcal{B}}) + D(\hat{X}_{\mathcal{A}} \hat{X}_{\mathcal{B}} \| X^*_{\mathcal{A}} X^*_{\mathcal{B}})\right)}. \tag{5.46}$$

We use $\kappa(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}; \hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}})$ to denote the number of elements in $\mathcal{S}(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}})$ that have type $(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}})$. Then, by applying part 2 of Lemma 2.16

$$\kappa(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}; \hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}) \leqslant 2^{nH(\hat{X}_{\mathcal{A}} \hat{X}_{\mathcal{B}} | \hat{U}_{\tilde{\mathcal{A}}} \hat{U}_{\tilde{\mathcal{B}}})}. \tag{5.47}$$

Letting $\kappa(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}})$ be the number of elements of $\mathcal{S}$ with type $(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}})$, we have

$$\begin{aligned}
\kappa(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}) &= \sum_{(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}) \in \mathcal{F}} \kappa(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}; \hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}) && (5.48)\\
&\overset{(5.47)}{\leqslant} \sum_{(\mathbf{u}_{\tilde{\mathcal{A}}}, \mathbf{u}_{\tilde{\mathcal{B}}}) \in \mathcal{F}} 2^{nH(\hat{X}_{\mathcal{A}} \hat{X}_{\mathcal{B}} | \hat{U}_{\tilde{\mathcal{A}}} \hat{U}_{\tilde{\mathcal{B}}})} && (5.49)\\
&\overset{(5.34)}{\leqslant} 2^{n\left(I(U_{\tilde{\mathcal{A}}} U_{\tilde{\mathcal{B}}}; X_{\mathcal{A}} X_{\mathcal{B}}) + H(\hat{X}_{\mathcal{A}} \hat{X}_{\mathcal{B}} | \hat{U}_{\tilde{\mathcal{A}}} \hat{U}_{\tilde{\mathcal{B}}}) + \varepsilon\right)}. && (5.50)
\end{aligned}$$

Thus,

$$\begin{aligned}
q_{\mathcal{F}} &\overset{(5.46)}{=} \sum_{\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}} \kappa(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}})\\
&\qquad \times 2^{-n\left(H(\hat{X}_{\mathcal{A}} \hat{X}_{\mathcal{B}}) + D(\hat{X}_{\mathcal{A}} \hat{X}_{\mathcal{B}} \| X^*_{\mathcal{A}} X^*_{\mathcal{B}})\right)} && (5.51)\\
&\overset{(5.50)}{\leqslant} \sum_{\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}} 2^{-n\left(k(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}) - \varepsilon\right)}, && (5.52)
\end{aligned}$$

where the sum is over all types that occur in $\mathcal{S}$ and

$$\begin{aligned}
k(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}) := I(\hat{U}_{\tilde{\mathcal{A}}} \hat{U}_{\tilde{\mathcal{B}}}; \hat{X}_{\mathcal{A}} \hat{X}_{\mathcal{B}}) - I(U_{\tilde{\mathcal{A}}} U_{\tilde{\mathcal{B}}}; X_{\mathcal{A}} X_{\mathcal{B}})&\\
+ D(\hat{X}_{\mathcal{A}} \hat{X}_{\mathcal{B}} \| X^*_{\mathcal{A}} X^*_{\mathcal{B}}).& \qquad (5.53)
\end{aligned}$$

Using a type counting argument (Lemma 2.15) we can further bound

$$\begin{aligned}
q_{\mathcal{F}} &\overset{(5.52)}{\leqslant} (n+1)^{|\mathcal{U}_{\tilde{\mathcal{A}}}||\mathcal{X}_{\mathcal{A}}||\mathcal{X}_{\mathcal{B}}||\mathcal{U}_{\tilde{\mathcal{B}}}|}\\
&\qquad \times \max_{\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}} 2^{-n\left(k(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}) - \varepsilon\right)}, && (5.54)
\end{aligned}$$

where the maximum is over all types occurring in $\mathcal{S}$. For any type $(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}})$ in $\mathcal{S}$, we have by construction $(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}) \in \mathcal{S}_\delta(U_{\tilde{\mathcal{A}}}, X_{\mathcal{A}}, X_{\mathcal{B}}, U_{\tilde{\mathcal{B}}})$ (following the δ-convention, Remark 1) and we thus conclude

$$\begin{aligned}
q_{\mathcal{F}} &\overset{(5.54)}{\leqslant} (n+1)^{|\mathcal{U}_{[K]}||\mathcal{X}_{[K]}|}\\
&\times \max_{(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}) \in \mathcal{S}_\delta(U_{\tilde{\mathcal{A}}}, X_{\mathcal{A}}, X_{\mathcal{B}}, U_{\tilde{\mathcal{B}}})} 2^{-n\left(k(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}) - \varepsilon\right)}. && (5.55)
\end{aligned}$$

Combining (5.43) and (5.55) we have shown that for n large enough

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \overset{(5.43)}{\geqslant} -\frac{1}{n} - \frac{1-\varepsilon}{n} \log_2 q_{\mathcal{F}} \tag{5.56}$$

$$\overset{(5.55)}{\geqslant} -2\varepsilon + (1-\varepsilon) \cdot \min k(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}) \tag{5.57}$$

$$\geqslant \min k(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}) - \big(2 + I(X_{\mathcal{A}}; X_{\mathcal{B}})\big)\varepsilon, \tag{5.58}$$

where the minimum is over all random variables $(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}})$ in $\mathcal{S}_\delta(U_{\tilde{\mathcal{A}}}, X_{\mathcal{A}}, X_{\mathcal{B}}, U_{\tilde{\mathcal{B}}})$. To justify the inequality (5.58), observe that $\min k(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}) \leqslant I(X_{\mathcal{A}}; X_{\mathcal{B}})$ by setting $(\hat{U}_{\tilde{\mathcal{A}}}, \hat{X}_{\mathcal{A}}, \hat{X}_{\mathcal{B}}, \hat{U}_{\tilde{\mathcal{B}}}) = (U_{\tilde{\mathcal{A}}}, X_{\mathcal{A}}, X_{\mathcal{B}}, U_{\tilde{\mathcal{B}}})$ in (5.53). The expression $k(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}})$ is a continuous function of $p_{\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}}$. By letting $n \to \infty$,

We have $\delta \to 0$ as $n \to \infty$ by the δ-convention (cf. Remark 1).

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \geqslant \min k(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}) - C\varepsilon \tag{5.59}$$

for some fixed constant C, where the minimum is over all random variables $(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}) \in \mathcal{S}_0(U_{\tilde{\mathcal{A}}}, X_{\mathcal{A}}, X_{\mathcal{B}}, U_{\tilde{\mathcal{B}}})$. Observe that for $(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}) \in \mathcal{S}_0(U_{\tilde{\mathcal{A}}}, X_{\mathcal{A}}, X_{\mathcal{B}}, U_{\tilde{\mathcal{B}}})$ we have

$$k(\tilde{U}_{\tilde{\mathcal{A}}}, \tilde{X}_{\mathcal{A}}, \tilde{X}_{\mathcal{B}}, \tilde{U}_{\tilde{\mathcal{B}}}) = I(\tilde{U}_{\tilde{\mathcal{A}}} \tilde{X}_{\mathcal{A}}; \tilde{X}_{\mathcal{B}} \tilde{U}_{\tilde{\mathcal{B}}}) \tag{5.60}$$

$$\geqslant I(\tilde{U}_{\tilde{\mathcal{A}}}; \tilde{U}_{\tilde{\mathcal{B}}}) = I(U_{\tilde{\mathcal{A}}}; U_{\tilde{\mathcal{B}}}). \tag{5.61}$$

Combining (5.59) and (5.61), we have

$$\Theta(f_{\mathcal{A}}; f_{\mathcal{B}}) \geqslant I\big(U_{\tilde{\mathcal{A}}}; U_{\tilde{\mathcal{B}}}\big) - C\varepsilon \overset{(5.8)}{\geqslant} \mu_{\mathcal{A},\mathcal{B}} - C\varepsilon. \tag{5.62}$$

We hence obtain $(\mu_\Omega - C\varepsilon, R_{[K]} + \varepsilon) \in \mathcal{R}$; since ε was arbitrary, this completes the proof.

5.4 THE MULTIPLE DESCRIPTION CEO PROBLEM

We continue the discussion of the CEO problem of Section 5.2 and assume $L = 1$ as well as $X_j -\!\!\circ\!\!- Y := Y_1 -\!\!\circ\!\!- X_{[J]\setminus j}$ for all $j \in [J]$. To simplify notation we will again use $\nu_{\mathcal{A}} := \nu_{\mathcal{A},1}$. Extending the setup discussed in Section 5.2, we will allow $\nu_j > 0$ for any $j \in [J]$. Loosely speaking, this requires a multiple description code for the CEO problem, enabling the CEO to obtain valuable information from the message of the jth agent alone. Surprisingly, this extension also permits a single-letter characterization. In particular, for the case $J = 2$, this allows us to give a full single-letter characterization of the achievable region, which will be explicitly stated in Corollary 5.16.

Definition 5.14. *For a total order $\sqsubset$ on $[J]$ and a set $\mathcal{E} \subseteq [J]$, let the region $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}$ be the set of points $(\nu_{\Pi}, R_{[J]})$ such that there exist random variables $(U_{[J]}, \varnothing) \in \mathcal{P}_*$ with*

Notation regarding total orders is given in Section 2.1.3.

$$\begin{aligned}
R_j &\geqslant I\big(U_j; X_j \big| U_{\sqsupset j}\big), && j \in [J], && (5.63)\\
R_j &\geqslant I\big(U_j; X_j\big), && j \in \mathcal{E}, && (5.64)\\
\nu_j &\leqslant I\big(U_j; Y \big| U_{\sqsupset j}\big), && j \notin \mathcal{E}, && (5.65)\\
\nu_j &\leqslant I\big(U_j; Y\big), && j \in \mathcal{E}, && (5.66)\\
\nu_{[J]} &\leqslant I\big(U_{[J]}; Y\big), && && (5.67)\\
\nu_{\mathcal{A}} &\leqslant 0, && 1 < |\mathcal{A}| < J. && (5.68)
\end{aligned}$$

Remark 11. The purpose of the order $\sqsubset$ is to determine the order of the messages for successive decoding. Equivalently, Definition 5.14 could be rephrased using a permutation of $[J]$ in place of a total order.

We are now able to state the single-letter characterization of $\overline{\mathcal{R}_{\mathrm{MI}}}$ with the additional condition that (5.68) holds.

Theorem 5.15. *We have the equality* $\{(\nu_{\Pi}, R_{[J]}) \in \overline{\mathcal{R}_{\mathrm{MI}}} : (5.68)$ *holds*$\} = \operatorname{conv}\Big(\bigcup_{\sqsubset,\mathcal{E}} \mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}\Big)$*, where the union is over all total orders $\sqsubset$ on $[J]$ and all sets $\mathcal{E} \subseteq [J]$.*

The proof of Theorem 5.15 is provided at the end of this section. In particular, Theorem 5.15 provides a single-letter characterization of $\overline{\mathcal{R}_{\mathrm{MI}}}$ for the case of $J = 2$ agents. We state this special case separately in the following corollary to showcase some interesting features of this single-letter region.

Corollary 5.16. *For $J = 2$, we have $\overline{\mathcal{R}_{\mathrm{MI}}} = \operatorname{conv}\Big(\mathcal{R}_{\mathrm{MI}}^{(1)} \cup \mathcal{R}_{\mathrm{MI}}^{(2)} \cup \mathcal{R}_{\mathrm{MI}}^{(3)}\Big)$, where $(\nu_{\Pi}, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}^{(i)}$ if, for some $(U_{[J]}, \varnothing) \in \mathcal{P}_*$, the following inequalities are satisfied*

$$\begin{array}{lll}
\mathcal{R}_{\mathrm{MI}}^{(1)}: & \mathcal{R}_{\mathrm{MI}}^{(2)}: & \mathcal{R}_{\mathrm{MI}}^{(3)}: \\
\quad \nu_1 \leqslant I(Y;U_1) & \quad \nu_1 \leqslant I(Y;U_1|U_2) & \quad \nu_1 \leqslant I(Y;U_1) \\
\quad \nu_2 \leqslant I(Y;U_2|U_1) & \quad \nu_2 \leqslant I(Y;U_2) & \quad \nu_2 \leqslant I(Y;U_2) \\
\nu_{\{1,2\}} \leqslant I(Y;U_1U_2) & \nu_{\{1,2\}} \leqslant I(Y;U_1U_2) & \nu_{\{1,2\}} \leqslant I(Y;U_1U_2) \\
\quad R_1 \geqslant I(U_1;X_1) & \quad R_1 \geqslant I(U_1;X_1|U_2) & \quad R_1 \geqslant I(U_1;X_1) \\
\quad R_2 \geqslant I(U_2;X_2|U_1) & \quad R_2 \geqslant I(U_2;X_2) & \quad R_2 \geqslant I(U_2;X_2).
\end{array}$$

Proof. Assuming $1 \sqsubset 2$, we obtain $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})} = \mathcal{R}_{\mathrm{MI}}^{(2)}$ if $1 \notin \mathcal{E}$ and otherwise $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})} = \mathcal{R}_{\mathrm{MI}}^{(3)}$. On the other hand, if $2 \sqsubset 1$, we obtain $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})} = \mathcal{R}_{\mathrm{MI}}^{(1)}$ if $2 \notin \mathcal{E}$ and otherwise also $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})} = \mathcal{R}_{\mathrm{MI}}^{(3)}$. □

Remark 12. Note that the total available rate of encoder 2 is $R_2 = I(X_2;U_2|U_1)$ to achieve a point in $\mathcal{R}_{\mathrm{MI}}^{(1)}$. Interestingly, this rate is in general less than the rate required to ensure successful typicality decoding of U_2. However, $\nu_2 = I(Y;U_2|U_1)$ can still be achieved.

Remark 13. On the other hand, fixing the random variables U_1, U_2 in the definition of $\mathcal{R}_{\mathrm{MI}}^{(i)}$ shows another interesting feature of this region. The achievable values for ν_1 and ν_2 vary across $i \in \{1,2,3\}$ and hence do not only depend on the chosen random variables U_1 and U_2, but also on the specific rates R_1 and R_2.

It is worth mentioning that by setting $\nu_1 = \nu_2 = 0$ the region $\overline{\mathcal{R}_{\mathrm{MI}}}$ reduces to the rate region in Lemma 5.11.

The following proposition shows that $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}$ is computable, at least in principle. The given cardinality bound is not optimal, but it implies $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})} = \overline{\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}}$. The proof of Proposition 5.17 is provided in Appendix B.4.

Proposition 5.17. *The region $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}$ remains unchanged if the cardinality bound $|\mathcal{U}_j| \leqslant |X_j| + 4^J$ is imposed for every $j \in [J]$.*

The following two theorems provide an inner and an outer bound for $\overline{\mathcal{R}_{\mathrm{MI}}}$. In order to show that Theorem 5.15 holds, we subsequently prove that these bounds are indeed tight, assuming that (5.68) holds.

Theorem 5.18. *We have $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})} \subseteq \overline{\mathcal{R}_{\mathrm{MI}}}$ for any $\mathcal{E} \subseteq [J]$ and any total order $\sqsubset$ on $[J]$.*

Theorem 5.19. *If $(\nu_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}$ then*

$$\sum_{j\in\mathcal{B}} R_j - \nu_{\mathcal{A}} \geqslant I(X_{\mathcal{B}}; U_{\mathcal{B}}|YQ) - I\big(Y; U_{\mathcal{A}\setminus\mathcal{B}}\big|Q\big) \tag{5.69}$$

for all $\mathcal{A}, \mathcal{B} \subseteq [J]$ and some random variables $(U_{[J]}, Q) \in \mathcal{P}_*$.

The proof of Theorems 5.18 and 5.19 are given in Appendices B.5 and B.6, respectively. We will, however, only require the following simple corollary of Theorem 5.19.

Corollary 5.20. *For any $(\nu_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}$ there are random variables $(U_{[J]}, Q) \in \mathcal{P}_*$ with*

$$\begin{aligned}
R_j &\geqslant 0, && \textit{for all } j \in [J], && (5.70)\\
\sum_{j\in\mathcal{A}} R_j - \nu_{[J]} &\geqslant I(X_{\mathcal{A}}; U_{\mathcal{A}}|YQ) && && \\
&\qquad - I\big(Y; U_{[J]\setminus\mathcal{A}}\big|Q\big), && \textit{for all } \mathcal{A} \subseteq [J], && (5.71)\\
R_j - \nu_j &\geqslant I\big(X_j; U_j\big|YQ\big), && \textit{for all } j \in [J], && (5.72)\\
\nu_j &\leqslant I\big(Y; U_j\big|Q\big), && \textit{for all } j \in [J]. && (5.73)
\end{aligned}$$

Proof of Theorem 5.15. We will make use of some rather technical results on convex polyhedra, derived in Section 2.6.

Assume $(\nu_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}$. We can then find $(U_{[J]}, Q) \in \mathcal{P}_*$ such that (5.70)–(5.73) hold. We define $\tilde{\nu}_\Pi := -\nu_\Pi$ to simplify notation. It is straightforward to check that the inequalities (5.70)–(5.73) define a sequence of closed convex polyhedra $\mathcal{H}^{(k)}$, $k \in [0\!:\!J]$ in the variables $(R_{[J]}, \tilde{\nu}_\Pi)$ that satisfy assumptions 1 and 2 of Lemma 2.58. $\mathcal{H}^{(0)}$

is defined by (5.70) and (5.71) alone, and for $k \in [J]$ the polyhedron $\mathcal{H}^{(k)}$ is given in the $K+k$ variables $(R_{[J]}, \tilde{\nu}_{[J]}, \tilde{\nu}_1, \tilde{\nu}_2, \dots, \tilde{\nu}_k)$ by adding constraints (5.72) and (5.73) for each $j \in [k]$. The set $\mathcal{H}^{(0)}$ is a supermodular polyhedron (Definition 2.48) in the K variables $(R_{[J]}, \tilde{\nu}_{[J]})$ on $([K], 2^{[K]})$ with rank function

$$f(\mathcal{A}) = \begin{cases} 0, & K \notin \mathcal{A}, \\ I(X_{\mathcal{A}\setminus K}; U_{\mathcal{A}\setminus K} | YQ) - I(Y; U_{[J]\setminus\mathcal{A}} | Q), & K \in \mathcal{A}, \end{cases} \tag{5.74}$$

where supermodularity follows via standard information-theoretic arguments. By Theorem 2.51, every extreme point of $\mathcal{H}^{(0)}$ is associated with a total order $\sqsubset$ on $[K]$. Such an extreme point is given by

$$R_j^{(\sqsubset)} = 0 \text{ for } j \sqsubset K, \tag{5.75}$$

$$R_j^{(\sqsubset)} = I(U_j; X_j | U_{\sqsupset j} Q) \text{ for } j \sqsupset K, \tag{5.76}$$

$$\nu_{[J]}^{(\sqsubset)} = I(Y; U_{\sqsupset K}|Q) - I(Y; U_{\sqsubset K}|YQ). \tag{5.77}$$

Assumption 3 of Lemma 2.58 is now verified by

$$R_j^{(\sqsubset)} \leqslant I(X_j; U_j | YQ) + I(Y; U_j | Q) = I(X_j; U_j | Q). \tag{5.78}$$

By applying Lemma 2.58 we find that every extreme point of $\mathcal{H}^{(J)}$ is given by a subset $\mathcal{E} \subseteq [J]$ and an order $\sqsubset$ of $[K]$ as

$$R_j^{(\sqsubset,\mathcal{E})} = I(X_j; U_j | Q), \qquad j \in \mathcal{E}, \tag{5.79}$$

$$R_j^{(\sqsubset,\mathcal{E})} = 0, \qquad j \notin \mathcal{E} \text{ and } j \sqsubset K, \tag{5.80}$$

$$R_j^{(\sqsubset,\mathcal{E})} = I(U_j; X_j | U_{\sqsupset j} Q), \qquad j \notin \mathcal{E} \text{ and } j \sqsupset K, \tag{5.81}$$

$$\nu_j^{(\sqsubset,\mathcal{E})} = I(U_j; Y | Q), \qquad j \in \mathcal{E}, \tag{5.82}$$

$$\nu_j^{(\sqsubset,\mathcal{E})} = -I(U_j; X_j | YQ), \qquad j \notin \mathcal{E} \text{ and } j \sqsubset K, \tag{5.83}$$

$$\nu_j^{(\sqsubset,\mathcal{E})} = I(U_j; Y | U_{\sqsupset j} Q), \qquad j \notin \mathcal{E} \text{ and } j \sqsupset K. \tag{5.84}$$

$$\nu_{[J]}^{(\sqsubset,\mathcal{E})} = I(Y; U_{\sqsupset K}|Q) - I(Y; U_{\sqsubset K}|YQ), \tag{5.85}$$

For each $q \in \mathcal{Q}$ with $P\{Q = q\} > 0$ let the point $(\nu_{\Pi}^{(\sqsubset,\mathcal{E},q)}, R_{[J]}^{(\sqsubset,\mathcal{E},q)})$ be defined by (5.79)–(5.85), but given $\{Q = q\}$. By substituting $U_j \to \varnothing$ if $j \notin \mathcal{E}$ and $j \sqsubset K$, we see that $(\nu_{\Omega}^{(\sqsubset,\mathcal{E},q)}, R_{[J]}^{(\sqsubset,\mathcal{E},q)}) \in \mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}$ and consequently $(\nu_{\Pi}^{(\sqsubset,\mathcal{E})}, R_{[J]}^{(\sqsubset,\mathcal{E})}) \in \operatorname{conv}\big(\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}\big)$. Defining the orthant $\mathcal{O} := \big\{(\nu_{\Pi}, R_{[J]}) : \nu_{\Pi} \leqslant 0, R_{[J]} \geqslant 0\big\}$, this implies

$$(\nu_{[K]}, R_{[J]}) \in \operatorname{conv}\left(\bigcup_{\sqsubset,\mathcal{E}} \operatorname{conv}\Big(\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}\Big)\right) + \mathcal{O} \tag{5.86}$$

$$= \operatorname{conv}\left(\bigcup_{\sqsubset,\mathcal{E}} \mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}\right) + \operatorname{conv}(\mathcal{O}) \tag{5.87}$$

$$= \operatorname{conv}\left(\bigcup_{\sqsubset,\mathcal{E}} \mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})} + \mathcal{O}\right) \tag{5.88}$$

$$= \operatorname{conv}\left(\bigcup_{\sqsubset,\mathcal{E}} \mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}\right), \tag{5.89}$$

where (5.88) follows from part 8 of Lemma 2.37 and in (5.89) we used that $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})} + \mathcal{O} = \mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}$ by definition. Together with Theorem 5.18 and the convexity of $\overline{\mathcal{R}_{\mathrm{MI}}}$ (Remark 9) we obtain

$$\mathcal{R}_{\mathrm{MI}} \subseteq \operatorname{conv}\left(\bigcup_{\sqsubset,\mathcal{E}} \mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}\right) \subseteq \overline{\mathcal{R}_{\mathrm{MI}}}. \tag{5.90}$$

It remains to show that $\operatorname{conv}\Big(\bigcup_{\sqsubset,\mathcal{E}} \mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}\Big)$ is closed. Using Proposition 5.17, we can write $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})} = \mathbf{F}^{(\sqsubset,\mathcal{E})}(\mathcal{P}'_*) + \mathcal{O}$, where

$$\mathcal{P}'_* := \Big\{p_{YX_{[J]}U_{[J]}} : (U_{[J]}, \varnothing) \in \mathcal{P}_*, |\mathcal{U}_j| = |\mathcal{X}_j| + 4^J, j \in [J]\Big\} \tag{5.91}$$

is a compact subset of the probability simplex and $\mathbf{F}^{(\sqsubset,\mathcal{E})}$ is a continuous function, given by the definition of $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}$, (5.63)–(5.68). We can thus write

$$\operatorname{conv}\left(\bigcup_{\sqsubset,\mathcal{E}} \mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{E})}\right) = \operatorname{conv}\left(\bigcup_{\sqsubset,\mathcal{E}} \mathbf{F}^{(\sqsubset,\mathcal{E})}(\mathcal{P}'_*) + \mathcal{O}\right), \tag{5.92}$$

which is closed by Corollary 2.38. □

5.5 CONCLUSION

We extended the multi-clustering problem to the case of an arbitrary number of sources. As in the case of two sources, we provided outer and inner bounds for the resulting achievable region. However, these bounds cannot be tight since the famous Körner-Marton problem constitutes a counterexample. For an analogue of the well-known

CEO problem, we showed that our bounds are tight in a special case, leveraging existing results from multiterminal lossy source coding.

Furthermore we considered a multiple description CEO problem under a suitable Markov constraint, which surprisingly also permits a single-letter characterization of the achievable region. In deriving this single-letter characterization we made use of submodularity and polymatroid theory. The resulting region has the remarkable feature that it allows to exploit rate that is in general insufficient to guarantee successful typicality decoding of the corresponding description.

6 DISCUSSION AND OUTLOOK

We presented a thorough study of two information-theoretic clustering problems, inspired by biclustering algorithms.

- A novel multi-terminal source coding problem termed the multi-clustering problem was introduced. It appears to be fundamentally different from "classical" distributed source coding problems, where the encoders usually aim at reducing redundancy as much as possible. In the multi-clustering problem, however, the encoders strive to maximize this redundant information.

  In the case of two sources, the multi-clustering problem turned out to be equivalent to two other problems in the information theory literature. Even for two sources, the multi-clustering problem is of formidable mathematical complexity and an exact characterization of the achievable region remains elusive. Already for one of the simplest cases, a doubly symmetric binary source, we were unable to provide a single-letter characterization of the achievable region. Even more so, we found strong evidence that our inner and outer bounds are loose. In doing so, we were able to disprove [66, Conjecture 1].

  We extended the multi-clustering problem, as well as the outer and inner bounds for its achievable region, to an arbitrary number of sources. Using the Körner-Marton binary sum problem [34] as a counterexample, we showed that these bounds cannot be tight. However, under a suitable Markov constraint, tight bounds are known for the CEO problem with logarithmic loss distortion [11], which constitutes a special case of the multi-clustering problem. We extended the CEO problem by requiring a multiple description code and were able to provide a single-letter characterization of the resulting achievable region using the same Markov constraint. This characterization also had some remarkable technical properties.

- We furthermore provided a proof for the two-function case of the Kumar-Courtade conjecture [10, Section IV]. Building upon previous results, we used Fourier analysis to reduce the conjecture to an elementary inequality, which we subsequently established. Furthermore we were able to show that the dictator functions are essentially the unique maximizers, achieving equality. Although the strategy employed heavily builds upon the joint properties of two Boolean functions, we hope that the proof can provide some insight into the general conjecture.

There are several questions this thesis left unanswered. In the context of the multi-clustering problem with a doubly symmetric binary source, Conjecture 3.14 claims that there is indeed a gap between the inner bound $\mathcal{S}_i'$ and the outer bound $\mathcal{R}_o$. This would follow from the stronger statement in Conjecture 3.15, which asserts that the inner bound $\mathcal{S}_i'$ is equal to the region $\mathcal{S}_b$. However, both statements remain open. Unfortunately, we were only recently made aware of the counterexample [8] provided in Proposition 3.13 and therefore unable to investigate this problem much further. In Remarks 12 and 13 we point out some interesting technical properties of the region $\overline{\mathcal{R}_{\mathrm{MI}}}$. In particular, it appears that it is possible to exploit rate that is in general insufficient to assure correct typicality decoding. However, it remains unclear whether these properties have a fundamental technical reason or are merely an artifact of our formulation. Evaluating $\overline{\mathcal{R}_{\mathrm{MI}}}$ for a specific source distribution could improve our understanding of this region. Further investigation is also necessary to explore whether a single-letter characterization as in Theorem 5.15 can be retained for $\overline{\mathcal{R}_{\mathrm{MI}}}$ when allowing $\nu_{\mathcal{A}} > 0$ for arbitrary $\mathcal{A} \subseteq [J]$, i.e., when lifting the constraint (5.68). Finally, a resolution for the original Kumar-Courtade conjecture on the mutual information between Boolean functions (Conjecture 4.1) would be appreciated.

It could also be worthwhile to apply the strategy used for deriving the cardinality bounds in Propositions 3.8 and 3.11 to other problems in information theory. This could aid the numerical evaluation of achievable regions.

Part III

APPENDIX

A

PROOFS FROM CHAPTER 3

A.1 PROOF OF THEOREM 3.7

For $(\mu, R_1, R_2) \in \mathcal{R}$, let (f, g) be an (n, R_1, R_2) code for (X, Z) with $\Theta(f; g) \geqslant \mu$. Defining the random variables $U_i := \big(f(\mathbf{X}), \mathbf{X}^{i-1}\big)$ and $V_i := \big(g(\mathbf{Z}), \mathbf{Z}^{i-1}\big)$ for $i \in [n]$, we have

$\Theta(f; g)$ is introduced in Definition 3.1.

$\mathbf{X}^i = (X_1, \dots, X_i)$.

$$
\begin{aligned}
nR_1 &\geqslant H\big(f(\mathbf{X})\big) = I\big(f(\mathbf{X}); \mathbf{X}\big) && \text{(A.1)}\\
&= \sum_{i=1}^{n} I\big(X_i; f(\mathbf{X}) \big| \mathbf{X}^{i-1}\big) && \text{(A.2)}\\
&= \sum_{i=1}^{n} I(X_i; U_i), && \text{(A.3)}
\end{aligned}
$$

where (A.3) holds because $\mathbf{X}$ are i.i.d. and we used part 3 of Lemma 2.4 in (A.2). Analogously, we obtain

$$
nR_2 \geqslant \sum_{i=1}^{n} I(Z_i; V_i). \tag{A.4}
$$

From Lemma 2.13 and the Markov chain $f(\mathbf{X}) \mathrel{-\!\!\circ\!\!-} \mathbf{X} \mathrel{-\!\!\circ\!\!-} \mathbf{Z} \mathrel{-\!\!\circ\!\!-} g(\mathbf{Z})$, we obtain

$$
\begin{aligned}
n\mu &\leqslant I\big(f(\mathbf{X}); g(\mathbf{Z})\big) && \text{(A.5)}\\
&\overset{(2.23)}{=} I\big(f(\mathbf{X}); \mathbf{X}\big) + I\big(g(\mathbf{Z}); \mathbf{Z}\big) - I\big(f(\mathbf{X})g(\mathbf{Z}); \mathbf{X}\mathbf{Z}\big) && \text{(A.6)}\\
&= \sum_{i=1}^{n} \Big[I(U_i; X_i) + I(V_i; Z_i) - I(U_i V_i; X_i Z_i)\Big]. && \text{(A.7)}
\end{aligned}
$$

Now a standard time-sharing argument (see, e. g., [16, Section 4.5.2]) shows $\mathcal{R} \subseteq \mathcal{R}_o$. Lemma 2.13 provides $\mathcal{R}_o \subseteq \mathcal{R}'_o$.

A.2 PROOF OF PROPOSITION 3.8

Most steps in the proof apply to both $\mathcal{R}_o$ and $\mathcal{R}'_o$. We thus state the proof for $\mathcal{R}_o$ and point out the modifications required for $\mathcal{R}'_o$ where appropriate.

Define the set of p.m.f.s (with finite, but arbitrarily large support)

$$
\mathcal{Q} := \{p_{UVXZ} : U \mathrel{-\!\!\circ\!\!-} X \mathrel{-\!\!\circ\!\!-} Z \text{ and } X \mathrel{-\!\!\circ\!\!-} Z \mathrel{-\!\!\circ\!\!-} V\} \tag{A.8}
$$

and the compact set of p.m.f.s with fixed alphabet size

$$\mathcal{Q}(a,b) := \{p_{UVXZ} \in \mathcal{Q} : |\mathcal{U}| = a, |\mathcal{V}| = b\}. \tag{A.9}$$

Define the continuous vector valued function $\mathbf{F} := (F_1, F_2, F_3)$ as

$$F_1(p_{UVXZ}) := I(U;X) + I(V;Z) - I(UV;XZ), \tag{A.10}$$
$$F_2(p_{UVXZ}) := I(U;X), \tag{A.11}$$
$$F_3(p_{UVXZ}) := I(V;Z). \tag{A.12}$$

In the proof of $\mathcal{R}_o' = \mathrm{conv}(\mathcal{S}_o')$ let $F_1(p_{UVXZ}) := \min\big\{I(U;Z), I(V;X)\big\}$. We can now write $\mathcal{R}_o = \mathbf{F}(\mathcal{Q}) + \mathcal{O}$ and $\mathcal{S}_o = \mathbf{F}\big(\mathcal{Q}(|\mathcal{X}|,|\mathcal{Z}|)\big) + \mathcal{O}$ where $\mathcal{O} := (\mathbb{R}_- \times \mathbb{R}_+ \times \mathbb{R}_+)$. Noting that $\mathcal{R}_o$ is a convex set, we define the function $\psi(\boldsymbol{\lambda}) := \inf_{\mathbf{x} \in \mathcal{R}_o} \boldsymbol{\lambda}^T \mathbf{x}$ and, using part 6 of Lemma 2.37, we obtain

We allow $\psi(\lambda) = -\infty$.

$$\overline{\mathrm{conv}(\mathcal{R}_o)} = \overline{\mathcal{R}_o} = \bigcap_{\boldsymbol{\lambda} \in \mathbb{R}^3} \big\{\mathbf{x} \in \mathbb{R}^3 : \boldsymbol{\lambda}^T \mathbf{x} \geqslant \psi(\boldsymbol{\lambda})\big\}. \tag{A.13}$$

From the definition of $\mathcal{R}_o$, we have $\psi(\boldsymbol{\lambda}) = -\infty$ if $\boldsymbol{\lambda} \notin \mathcal{O}$, and $\psi(\boldsymbol{\lambda}) = \inf_{p \in \mathcal{Q}} \boldsymbol{\lambda}^T \mathbf{F}(p)$ everywhere else. Thus,

$$\overline{\mathcal{R}_o} = \bigcap_{\boldsymbol{\lambda} \in \mathcal{O}} \Big\{\mathbf{x} \in \mathbb{R}^3 : \boldsymbol{\lambda}^T \mathbf{x} \geqslant \psi(\boldsymbol{\lambda})\Big\}. \tag{A.14}$$

Using the same argument, one can show

$$\overline{\mathrm{conv}(\mathcal{S}_o)} = \bigcap_{\boldsymbol{\lambda} \in \mathcal{O}} \Big\{\mathbf{x} \in \mathbb{R}^3 : \boldsymbol{\lambda}^T \mathbf{x} \geqslant \widetilde{\psi}(\boldsymbol{\lambda})\Big\}, \text{ with} \tag{A.15}$$
$$\widetilde{\psi}(\boldsymbol{\lambda}) = \min_{p \in \mathcal{Q}(|\mathcal{X}|,|\mathcal{Z}|)} \boldsymbol{\lambda}^T \mathbf{F}(p). \tag{A.16}$$

The minimum in (A.16) is justified by compactness. We next show that $\psi(\boldsymbol{\lambda}) = \widetilde{\psi}(\boldsymbol{\lambda})$ for any $\boldsymbol{\lambda} \in \mathcal{O}$. For any $\delta > 0$, we can find random variables $(\widetilde{U}, X, Z, \widetilde{V}) \sim \widetilde{p} \in \mathcal{Q}$ with $\boldsymbol{\lambda}^T \mathbf{F}(\widetilde{p}) \leqslant \psi(\boldsymbol{\lambda}) + \delta$. By compactness of $\mathcal{Q}(a,b)$ and continuity of $\mathbf{F}$, there is $p \in \mathcal{Q}(|\widetilde{\mathcal{U}}|, |\widetilde{\mathcal{V}}|)$ with

$$\boldsymbol{\lambda}^T \mathbf{F}(p) = \min_{\widetilde{p} \in \mathcal{Q}(|\widetilde{\mathcal{U}}|,|\widetilde{\mathcal{V}}|)} \boldsymbol{\lambda}^T \mathbf{F}(\widetilde{p}) \leqslant \boldsymbol{\lambda}^T \mathbf{F}(\widetilde{p}) \leqslant \psi(\boldsymbol{\lambda}) + \delta. \tag{A.17}$$

We now show that there exists $\hat{p} \in \mathcal{Q}(|\mathcal{X}|,|\mathcal{Z}|)$ with

$$\boldsymbol{\lambda}^T \mathbf{F}(\hat{p}) = \boldsymbol{\lambda}^T \mathbf{F}(p). \tag{A.18}$$

As a consequence of the inequalities $F_1 \leqslant F_2$ and $F_1 \leqslant F_3$ we have $\boldsymbol{\lambda}^T \mathbf{F}(p) = 0$ if $\lambda_1 + \max\{\lambda_2, \lambda_3\} \geqslant 0$. Thus, we only need to show (A.18) for $\boldsymbol{\lambda} \in \mathcal{O}$ with $\lambda_1 + \lambda_2 < 0$ and $\lambda_1 + \lambda_3 < 0$. To this end we use the perturbation method [22], [31] and perturb p, obtaining the candidate

$$(\hat{U}, X, Z, \hat{V}) \sim \hat{p}(u,x,z,v) = p(u,x,z,v)\big(1 + \varepsilon\phi(u)\big). \tag{A.19}$$

We require

$$\begin{aligned} 1+\varepsilon\phi(u) &\geqslant 0, && \text{for every } u\in\mathcal{U}, && \text{(A.20)}\\ \mathbb{E}[\phi(U)] &= 0, && && \text{(A.21)}\\ \mathbb{E}[\phi(U)|X=x,Z=z] &= 0, && \text{if } p(x,z)>0. && \text{(A.22)} \end{aligned}$$

The conditions (A.20) and (A.21) ensure that $\hat{p}$ is a valid p.m.f. and (A.22) implies $\hat{p}\in\mathcal{Q}$. Observe that for any ϕ, there is an $\varepsilon_0>0$ such that (A.20) is satisfied for $\varepsilon\in[-\varepsilon_0,\varepsilon_0]$. Furthermore, (A.22) is equivalent to

$$\mathbb{E}[\phi(U)|X=x]=0, \quad \text{for every } x\in\mathcal{X} \tag{A.23}$$

because of the Markov chain $U \multimap\!\!- X \multimap\!\!- Z$. Note also that (A.23) already implies (A.21). If $|\mathcal{U}|\geqslant|\mathcal{X}|+1$ there is a non-trivial solution to (A.23), which means there exists $\phi\not\equiv 0$ such that (A.20)–(A.22) are satisfied. We have

$$\begin{aligned} \lambda^T F(\hat{p}) = \lambda_1 \Big[& I(X;U) - I(UV;XZ) + H(Z) + \varepsilon H_\phi(U) - \varepsilon H_\phi(UX) \\ & - \varepsilon H_\phi(UV) + \varepsilon H_\phi(UXZV) + H(\hat{V}) - H(Z\hat{V}) \Big] \\ & + \lambda_2[I(X;U) + \varepsilon H_\phi(U) - \varepsilon H_\phi(UX)] \\ & + \lambda_3[H(Z) + H(\hat{V}) - H(Z\hat{V})]. \end{aligned} \tag{A.24}$$

Here, we used the shorthand

$$H_\phi(UX) := -\sum_{u,x} p(u,x)\phi(u)\log_2 p(u,x) \tag{A.25}$$

and analogous for other combinations of random variables. By (A.17), we have $\frac{\partial^2}{\partial\varepsilon^2}\lambda^T F(\hat{p})\big|_{\varepsilon=0}\geqslant 0$.

Observe that

$$\frac{\partial}{\partial\varepsilon}\big(H(\hat{V}) - H(Z\hat{V})\big) = \frac{\partial}{\partial\varepsilon}\sum_{z,v}\hat{p}(z,v)\log_2\frac{\hat{p}(z,v)}{\hat{p}(v)} \tag{A.26}$$

$$\begin{aligned} &= \frac{1}{\log 2}\sum_{z,v}\frac{\partial\hat{p}(z,v)}{\partial\varepsilon}\log\frac{\hat{p}(z,v)}{\hat{p}(v)} \\ &\qquad + \hat{p}(z,v)\frac{\hat{p}(v)}{\hat{p}(z,v)}\frac{\hat{p}(v)\frac{\partial\hat{p}(z,v)}{\partial\varepsilon} - \hat{p}(z,v)\frac{\partial\hat{p}(v)}{\partial\varepsilon}}{\hat{p}(v)^2} \end{aligned} \tag{A.27}$$

$$= \frac{1}{\log 2}\sum_{z,v}\frac{\partial\hat{p}(z,v)}{\partial\varepsilon}\log\frac{\hat{p}(z,v)}{\hat{p}(v)} + \frac{\partial\hat{p}(z,v)}{\partial\varepsilon} - \frac{\hat{p}(z,v)\frac{\partial\hat{p}(v)}{\partial\varepsilon}}{\hat{p}(v)} \tag{A.28}$$

and consequently

$$\frac{\partial^2}{\partial\varepsilon^2}\boldsymbol{\lambda}^T\mathbf{F}(\hat{p}) = (\lambda_1+\lambda_3)\frac{\partial^2}{\partial\varepsilon^2}\big(H(\hat{V}) - H(Z\hat{V})\big) \tag{A.29}$$

$$= \frac{\lambda_1+\lambda_3}{\log 2}\sum_{z,v}\frac{\partial\hat{p}(z,v)}{\partial\varepsilon}\frac{\hat{p}(v)}{\hat{p}(z,v)}\frac{\hat{p}(v)\frac{\partial\hat{p}(z,v)}{\partial\varepsilon} - \hat{p}(z,v)\frac{\partial\hat{p}(v)}{\partial\varepsilon}}{\hat{p}(v)^2} - \frac{\partial\hat{p}(v)}{\partial\varepsilon}\frac{\frac{\partial\hat{p}(z,v)}{\partial\varepsilon}\hat{p}(v) - \hat{p}(z,v)\frac{\partial\hat{p}(v)}{\partial\varepsilon}}{\hat{p}(v)^2} \tag{A.30}$$

$$= \frac{\lambda_1+\lambda_3}{\log 2}\sum_{z,v}\left(\frac{\partial\hat{p}(z,v)}{\partial\varepsilon}\right)^2\frac{1}{\hat{p}(z,v)} - 2\frac{\partial\hat{p}(z,v)}{\partial\varepsilon}\frac{\partial\hat{p}(v)}{\partial\varepsilon}\frac{1}{\hat{p}(v)} + \left(\frac{\partial\hat{p}(v)}{\partial\varepsilon}\right)^2\frac{\hat{p}(z,v)}{\hat{p}(v)^2}. \tag{A.31}$$

Here we already used that $\frac{\partial^2\hat{p}(v)}{\partial\varepsilon^2} \equiv \frac{\partial^2\hat{p}(z,v)}{\partial\varepsilon^2} \equiv 0$. It is straightforward to calculate

$$\frac{\partial\hat{p}(v)}{\partial\varepsilon} = p(v)\mathbb{E}[\phi(U)|V=v], \tag{A.32}$$

$$\frac{\partial\hat{p}(z,v)}{\partial\varepsilon} = p(z,v)\mathbb{E}[\phi(U)|V=v, Z=z], \tag{A.33}$$

$$\hat{p}(z,v)|_{\varepsilon=0} = p(z,v), \tag{A.34}$$

$$\hat{p}(v)|_{\varepsilon=0} = p(v), \tag{A.35}$$

and thus, taking into account that $\lambda_1+\lambda_3 < 0$, we have

$$0 \geqslant \sum_{z,v} p(z,v)\big(\mathbb{E}[\phi(U)|V=v, Z=z] - \mathbb{E}[\phi(U)|V=v]\big)^2, \tag{A.36}$$

which implies for any $(z,v) \in \mathcal{Z}\times\mathcal{V}$ with $p(z,v) > 0$ that

$$\sum_u p(u|z,v)\phi(u) = \sum_u p(u|v)\phi(u). \tag{A.37}$$

We conclude that

$$H(\hat{V}) - H(Z\hat{V}) = \sum_{z,v}\hat{p}(z,v)\log_2\frac{\hat{p}(z,v)}{\hat{p}(v)} \tag{A.38}$$

$$= \sum_{z,v,u} p(u,z,v)\big(1+\varepsilon\phi(u)\big)\log_2\frac{\hat{p}(z,v)}{\hat{p}(v)} \tag{A.39}$$

$$= \sum_{z,v,u} p(u,z,v)\big(1+\varepsilon\phi(u)\big)\log_2\frac{p(z,v)}{p(v)} \tag{A.40}$$

$$= H(V) - H(ZV) + \varepsilon H_\phi(V) - \varepsilon H_\phi(ZV), \tag{A.41}$$

where we used

$$H_\phi(V) := -\sum_{u,v} p(u,v)\phi(u)\log_2 p(v) \tag{A.42}$$

$$H_\phi(ZV) := -\sum_{u,z,v} p(u,z,v)\phi(u)\log_2 p(z,v). \tag{A.43}$$

and the equality in (A.40) follows from (A.37) by

$$\begin{aligned}
\frac{\hat{p}(z,v)}{\hat{p}(v)} &= \frac{\sum_{\hat{u}} p(\hat{u},z,v)\big(1+\varepsilon\phi(\hat{u})\big)}{\sum_{\hat{u}} p(\hat{u},v)\big(1+\varepsilon\phi(\hat{u})\big)} && \text{(A.44)}\\
&= \frac{p(z,v)\sum_{\hat{u}} p(\hat{u}|z,v)\big(1+\varepsilon\phi(\hat{u})\big)}{p(v)\sum_{\hat{u}} p(\hat{u}|v)\big(1+\varepsilon\phi(\hat{u})\big)} && \text{(A.45)}\\
&= \frac{p(z,v)\big(1+\varepsilon\sum_{\hat{u}} p(\hat{u}|z,v)\phi(\hat{u})\big)}{p(v)\big(1+\varepsilon\sum_{\hat{u}} p(\hat{u}|v)\phi(\hat{u})\big)} && \text{(A.46)}\\
&\overset{\text{(A.37)}}{=} \frac{p(z,v)\big(1+\varepsilon\sum_{\hat{u}} p(\hat{u}|v)\phi(\hat{u})\big)}{p(v)\big(1+\varepsilon\sum_{\hat{u}} p(\hat{u}|v)\phi(\hat{u})\big)} && \text{(A.47)}\\
&= \frac{p(z,v)}{p(v)}. && \text{(A.48)}
\end{aligned}$$

Substituting (A.41) in (A.24) shows that $\boldsymbol{\lambda}^T\mathbf{F}(\hat{p})$ is linear in ε, and by the optimality of p it must therefore be constant. We may now choose ε maximal, i.e., such that there is at least one $u \in \mathcal{U}$ with $p(u)(1+\varepsilon\phi(u)) = 0$. This effectively reduces the cardinality of $\hat{\mathcal{U}}$ by at least one and may be repeated until $|\hat{\mathcal{U}}| = |\mathcal{X}|$ (as then $\phi \equiv 0$). The same process can be carried out for V and yields $\hat{p} \in \mathcal{Q}(|\mathcal{X}|,|\mathcal{Z}|)$ such that (A.18) holds.

In the proof of $\mathcal{R}_o' = \operatorname{conv}(\mathcal{S}_o')$, we show (A.18) by applying the support lemma [16, Appendix C] with $|\mathcal{X}|-1$ test functions $f_x(p_{\hat{X}}) := p_{\hat{X}}(x)$ $(x \in \mathcal{X})$ and with the function

$$\begin{aligned}
f(p_{\hat{X}}) := \lambda_1 \min\big\{I(V;X), H(Z)-H(\hat{Z})\big\}\\
+\lambda_2\big(H(X)-H(\hat{X})\big)+\lambda_3 I(Z;V),
\end{aligned} \tag{A.49}$$

where $(\hat{Z},\hat{X}) \sim p_{\hat{X}}p_{Z|X}$. Consequently, there exists a random variable $\hat{U}$ with $(\hat{U},X,Z,V) \sim \hat{p} \in \mathcal{Q}(|\mathcal{X}|,|\widetilde{\mathcal{V}}|)$ and $\boldsymbol{\lambda}^T\mathbf{F}(\hat{p}) = \boldsymbol{\lambda}^T\mathbf{F}(p)$. By applying the same argument to V, we obtain $\hat{p} \in \mathcal{Q}(|\mathcal{X}|,|\mathcal{Z}|)$ such that (A.18) holds.

By combining (A.17) and (A.18) we obtain

$$\boldsymbol{\lambda}^T\mathbf{F}(\hat{p}) = \boldsymbol{\lambda}^T\mathbf{F}(p) \leqslant \psi(\boldsymbol{\lambda}) + \delta. \tag{A.50}$$

As $\delta > 0$ was arbitrary and $\mathcal{Q}(|\mathcal{X}|,|\mathcal{Z}|)$ is compact, we proved $\psi(\boldsymbol{\lambda}) = \widetilde{\psi}(\boldsymbol{\lambda})$, which implies $\overline{\mathcal{R}_o} = \overline{\operatorname{conv}(\mathcal{S}_o)}$ using (A.14) and (A.16). We find

$$
\begin{aligned}
\overline{\mathcal{R}_o} &= \overline{\operatorname{conv}(\mathcal{S}_o)} && \text{(A.51)}\\
&= \overline{\operatorname{conv}\big(\mathbf{F}\big(\mathcal{Q}(|\mathcal{X}|,|\mathcal{Z}|)\big) + \mathcal{O}\big)} && \text{(A.52)}\\
&= \operatorname{conv}\big(\mathbf{F}\big(\mathcal{Q}(|\mathcal{X}|,|\mathcal{Z}|)\big)\big) + \mathcal{O} && \text{(A.53)}\\
&= \operatorname{conv}(\mathcal{S}_o) && \text{(A.54)}\\
&\subseteq \mathbf{F}(\mathcal{Q}) + \mathcal{O} && \text{(A.55)}\\
&= \mathcal{R}_o, && \text{(A.56)}
\end{aligned}
$$

where (A.53) follows from Corollary 2.38. The relation (A.55) is a consequence of $\mathcal{Q}(|\mathcal{X}|,|\mathcal{Z}|) \subseteq \mathcal{Q}$ and the convexity of $\mathbf{F}(\mathcal{Q})$.

A.3 PROOF OF PROPOSITION 3.11

We only need to show $\operatorname{conv}(\mathcal{S}_i) = \operatorname{conv}(\mathcal{R}_i)$ as the cardinality bound $|\mathcal{Q}| \leqslant 3$ follows directly from Theorem 2.39 because $\operatorname{conv}(\mathcal{R}_i)$ is the convex hull of a connected set in $\mathbb{R}^3$ and hence connected by part 1 of Lemma 2.37. We will only show the cardinality bound $|\mathcal{U}| \leqslant |\mathcal{X}|$ as the corresponding bound for $|\mathcal{V}|$ follows analogously. Note that the weaker bounds $|\mathcal{U}| \leqslant |\mathcal{X}| + 1$ and $|\mathcal{V}| \leqslant |\mathcal{Z}| + 1$ can be shown directly using the convex cover method (cf. [16, Appendix C], [2], [71]), i. e., by applying Theorem 2.39. Define the continuous vector-valued function

$$
\mathbf{F}(p_{UXZV}) := \big(\mathrm{I}(U;V), \mathrm{I}(X;U), \mathrm{I}(Z;V)\big) \tag{A.57}
$$

and the compact, connected sets of p.m.f.s

$[0:n] = \{0,1,2,\dots,n\}$ and $[n] = \{1,2,\dots,n\}$.

$$
\begin{aligned}
\mathcal{Q} &:= \Big\{p_{UXZV} : U \mathrel{-\!\circ\!-} X \mathrel{-\!\circ\!-} Z \mathrel{-\!\circ\!-} V, \mathcal{U} = \big[0:|\mathcal{X}|\big], \mathcal{V} = \big[0:|\mathcal{Z}|\big]\Big\}, && \text{(A.58)}\\
\mathcal{Q}' &:= \Big\{p_{UXZV} \in \mathcal{Q} : \mathcal{U} = \big[|\mathcal{X}|\big]\Big\}. && \text{(A.59)}
\end{aligned}
$$

To complete the proof of the proposition, it suffices to show

$$
\operatorname{conv}\big(\mathbf{F}(\mathcal{Q})\big) \subseteq \operatorname{conv}\big(\mathbf{F}(\mathcal{Q}')\big), \tag{A.60}
$$

since we then have

$$
\begin{aligned}
\operatorname{conv}(\mathcal{R}_i) &= \operatorname{conv}\big(\mathbf{F}(\mathcal{Q}) + \mathcal{O}\big) && \text{(A.61)}\\
&= \operatorname{conv}\big(\mathbf{F}(\mathcal{Q})\big) + \mathcal{O} && \text{(A.62)}\\
&\overset{\text{(A.60)}}{\subseteq} \operatorname{conv}\big(\mathbf{F}(\mathcal{Q}')\big) + \mathcal{O} && \text{(A.63)}\\
&= \operatorname{conv}\big(\mathbf{F}(\mathcal{Q}') + \mathcal{O}\big) && \text{(A.64)}\\
&= \operatorname{conv}(\mathcal{S}_i), && \text{(A.65)}
\end{aligned}
$$

where (A.62) and (A.64) follow from Corollary 2.38 and we used $\mathcal{O} := (\mathbb{R}_- \times \mathbb{R}_+ \times \mathbb{R}_+)$. The region $\mathbf{F}(\mathcal{Q}) \subseteq \mathbb{R}^3$ is the continuous image of a compact set and hence compact by part 2 of Lemma 2.37. Therefore, its convex hull $\operatorname{conv}\big(\mathbf{F}(\mathcal{Q})\big)$ is compact by part 5 of Lemma 2.37 and can be represented as an intersection of halfspaces using part 6 of Lemma 2.37: For $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^3$ we define the function $\psi(\boldsymbol{\lambda}) := \min_{\mathbf{x}\in\mathbf{F}(\mathcal{Q})} \boldsymbol{\lambda}^T\mathbf{x}$ and have

$$\operatorname{conv}\big(\mathbf{F}(\mathcal{Q})\big) = \bigcap_{\boldsymbol{\lambda}\in\mathbb{R}^3} \Big\{\mathbf{x} \in \mathbb{R}^3 : \boldsymbol{\lambda}^T\mathbf{x} \geqslant \psi(\boldsymbol{\lambda})\Big\}. \tag{A.66}$$

With the same reasoning we obtain

$$\operatorname{conv}\big(\mathbf{F}(\mathcal{Q}')\big) = \bigcap_{\boldsymbol{\lambda}\in\mathbb{R}^3} \Big\{\mathbf{x} \in \mathbb{R}^3 : \boldsymbol{\lambda}^T\mathbf{x} \geqslant \tilde{\psi}(\boldsymbol{\lambda})\Big\}, \tag{A.67}$$

where $\tilde{\psi}(\boldsymbol{\lambda}) := \min_{\mathbf{x}\in\mathbf{F}(\mathcal{Q}')} \boldsymbol{\lambda}^T\mathbf{x}$. We next show $\tilde{\psi}(\boldsymbol{\lambda}) \leqslant \psi(\boldsymbol{\lambda})$ which already implies (A.60) due to (A.66) and (A.67).

Let $\mathcal{X}' := \mathcal{X} \setminus \{x\}$ where $x \in \mathcal{X}$ is arbitrary. Define the test functions $\mathbf{t} = (t_x)_{x\in\mathcal{X}'}$ with $t_x(p_{\hat{X}}) := p_{\hat{X}}(x)$. Choose any $\boldsymbol{\lambda} \in \mathbb{R}^3$ and fix $(U, X, Z, V) \sim p \in \mathcal{Q}$ that achieve $\boldsymbol{\lambda}^T\mathbf{F}(p) = \psi(\boldsymbol{\lambda})$. Define the continuous function

$$\begin{aligned} f(p_{\hat{X}}) := \lambda_1 &\big(H(V) - H(\hat{V})\big) \\ &+ \lambda_2\big(H(X) - H(\hat{X})\big) + \lambda_3 I(Z;V) \end{aligned} \tag{A.68}$$

where $(\hat{V}, \hat{Z}, \hat{X}) \sim p_{V|Z}p_{Z|X}p_{\hat{X}}$. The point $\big(p_X(x)_{x\in\mathcal{X}'}, \psi(\boldsymbol{\lambda})\big)$ lies in the convex hull of $(\mathbf{t}, f)(\mathcal{Q})$ which is compact and connected due to parts 1 and 2 of Lemma 2.37. Theorem 2.39 hence implies that $|\mathcal{X}|$ points suffice, i. e., there exists a random variable $\tilde{U}$ with $|\tilde{\mathcal{U}}| = |\mathcal{X}|$ (i. e., $p_{\tilde{U}XZV} \in \mathcal{Q}'$) such that $\mathbb{E}_{\tilde{U}}\big[f\big(p_{X|\tilde{U}}(\,\cdot\,|\tilde{U})\big)\big] = \boldsymbol{\lambda}^T\mathbf{F}(p_{\tilde{U}XZV}) = \psi(\boldsymbol{\lambda})$. This shows $\tilde{\psi}(\boldsymbol{\lambda}) \leqslant \psi(\boldsymbol{\lambda})$.

By applying the same reasoning to V, one can show that $|\mathcal{V}| = |\mathcal{Z}|$ also suffices.

A.4 PROOF OF PROPOSITION 3.13

With $\tilde{U} = X \oplus N_1$ and $\tilde{V} = Z \oplus N_2$, where $N_1, N_2 \sim \mathcal{B}(\alpha)$ are independent of (X, Z) and of each other, it follows that $(\mu, R, R) := \big(1 - H(\alpha * p), 1 - H(\alpha), 1 - H(\alpha)\big) \in \mathcal{R}'_o$ for $\alpha \in (0, \frac{1}{2})$. Assume $(\mu, R, R) \in \mathcal{S}'_i$ and choose U, V, and Q according to Proposition 3.11. We then have

$H(p)$ is the binary entropy function and $a * b$ denotes binary convolution.

$$H(X|UQ) \geqslant H(\alpha), \tag{A.69}$$

$$H(Z|VQ) \geqslant H(\alpha), \tag{A.70}$$

$$I(U;V|Q) \geqslant 1 - H(\alpha * p). \tag{A.71}$$

Using Mrs. Gerber's Lemma, Theorem 2.10, we obtain

$$\mathrm{H}(X|VQ) \geqslant \mathrm{H}\big(\mathrm{H}^{-1}\big(\mathrm{H}(Z|VQ)\big) * p\big) \overset{(A.70)}{\geqslant} \mathrm{H}(\alpha * p). \tag{A.72}$$

$\mathrm{H}^{-1}(t)$ is the inverse of the binary entropy function (cf. Definition 2.3).

Thus, $\mathrm{I}(X;V|Q) \leqslant 1 - \mathrm{H}(\alpha * p)$ and furthermore $\mathrm{I}(X;V|Q) \geqslant \mathrm{I}(U;V|Q)$ due to $U \mathrel{-\!\!\circ\!\!-} XQ \mathrel{-\!\!\circ\!\!-} V$. These two inequalities in combination with (A.71) imply $\mathrm{I}(X;V|Q) = \mathrm{I}(U;V|Q)$, which amounts to $X \mathrel{-\!\!\circ\!\!-} UQ \mathrel{-\!\!\circ\!\!-} V$. We can therefore write the joint p.m.f. of (U, X, V, Q) in two ways, as

$$\begin{aligned} & p_{UXVQ}(u, x, v, q) \\ &\quad = p_X(x) p_Q(q) p_{U|XQ}(u|x, q) p_{V|XQ}(v|x, q) && \text{(A.73)} \\ &\quad = p_X(x) p_Q(q) p_{U|XQ}(u|x, q) p_{V|UQ}(v|u, q). && \text{(A.74)} \end{aligned}$$

Assume without loss of generality that $p_Q(q) > 0$ for all $q \in \mathcal{Q}$. If $p_{U|XQ}(u|x, q) > 0$, then (A.74) necessitates

$$p_{V|UQ}(v|u, q) = p_{V|XQ}(v|x, q) \tag{A.75}$$

for $v \in \{0, 1\}$. Next, we partition $\mathcal{Q}$ into three disjoint subsets

$$\begin{aligned} \mathcal{Q}_1 &:= \big\{q \in \mathcal{Q} : \mathrm{P}\{U = X|Q = q\} \in \{0, 1\}\big\}, && \text{(A.76)} \\ \mathcal{Q}_2 &:= \big\{q \in \mathcal{Q} : \mathrm{H}(U|Q = q) = 0\big\}, && \text{(A.77)} \\ \mathcal{Q}_3 &:= \big\{q \in \mathcal{Q} : \mathrm{H}(U|X, Q = q) \neq 0\big\}. && \text{(A.78)} \end{aligned}$$

For $q \in \mathcal{Q}_3$ there is some $x \in \mathcal{X}$ such that for both $u \in \{0, 1\}$ we have $p_{U|XQ}(u|x, q) > 0$. We apply (A.75) twice and obtain

$$p_{V|UQ}(v|0, q) = p_{V|XQ}(v|x, q) = p_{V|UQ}(v|1, q), \tag{A.79}$$

i.e., $\mathrm{I}(U;V|Q = q) = 0$, which also holds for $q \in \mathcal{Q}_2$. We can thus write

$$\begin{aligned} 1 - \mathrm{H}(\alpha * p) &\overset{(A.71)}{\leqslant} \mathrm{I}(U;V|Q) && \text{(A.80)} \\ &= \sum_{q \in \mathcal{Q}_1} p_Q(q) \mathrm{I}(U;V|Q = q) && \text{(A.81)} \\ &\leqslant \sum_{q \in \mathcal{Q}_1} p_Q(q) \mathrm{I}(X;Z|Q = q) && \text{(A.82)} \\ &= \mathrm{P}\{Q \in \mathcal{Q}_1\}\big(1 - \mathrm{H}(p)\big), && \text{(A.83)} \end{aligned}$$

where (A.82) follows from the Markov chains $U \mathrel{-\!\!\circ\!\!-} XQ \mathrel{-\!\!\circ\!\!-} V$ and $X \mathrel{-\!\!\circ\!\!-} ZQ \mathrel{-\!\!\circ\!\!-} V$. On the other hand we have

$$\begin{aligned} H(\alpha) &\overset{(A.69)}{\leqslant} H(X|UQ) && \text{(A.84)} \\ &= 1 - I(X;U|Q) && \text{(A.85)} \\ &\leqslant 1 - \sum_{q \in \mathcal{Q}_1} p_Q(q) I(X;U|Q=q) && \text{(A.86)} \\ &\overset{(A.76)}{=} 1 - P\{Q \in \mathcal{Q}_1\}. && \text{(A.87)} \end{aligned}$$

Combining the previous two inequalities leads to

$$\frac{1 - H(\alpha * p)}{1 - H(p)} \overset{(A.83)}{\leqslant} P\{Q \in \mathcal{Q}_1\} \overset{(A.87)}{\leqslant} 1 - H(\alpha), \tag{A.88}$$

which is a contradiction since $\frac{1-H(\alpha * p)}{1-H(p)} > 1 - H(\alpha)$.

B PROOFS FROM CHAPTER 5

B.1 PROOF OF THEOREM 5.3

If $(\mu_\Omega, R_{[K]}) \in \mathcal{R}$ we obtain an $(n, R_{[K]})$ code $f_{[K]}$ for $X_{[K]}$ such that (5.2) holds. Define $W_k := f_k(\mathbf{X}_k)$ and the auxiliary random variables $U_{k,i} := (W_k, \mathbf{X}^{i-1}_{[K],1})$ for $k \in [K]$ and $i \in [n]$. For any two sets $\mathcal{A}, \mathcal{C} \subseteq [K]$ we have

$$n \sum_{k \in \mathcal{A}} R_k \geqslant H(W_{\mathcal{A}}) \tag{B.1}$$

$$= I(W_{\mathcal{A}}; \mathbf{X}_{[K]}) \tag{B.2}$$

$$\geqslant I(W_{\mathcal{A}}; \mathbf{X}_{[K]} | W_{\mathcal{C}}) \tag{B.3}$$

$$= \sum_{i=1}^{n} I(W_{\mathcal{A}}; X_{[K],i} | W_{\mathcal{C}} \mathbf{X}^{i-1}_{[K],1}) \tag{B.4}$$

$$= \sum_{i=1}^{n} I(U_{\mathcal{A},i}; X_{[K],i} | U_{\mathcal{C},i}), \tag{B.5}$$

where (B.3) follows from $W_{\mathcal{A}} \mathrel{-\!\!\circ\!\!-} \mathbf{X}_{[K]} \mathrel{-\!\!\circ\!\!-} W_{\mathcal{C}}$. Furthermore, for any pair $(\mathcal{A}, \mathcal{B}) \in \Omega$, we have by Lemma 2.13 and $W_{\mathcal{A}} \mathrel{-\!\!\circ\!\!-} \mathbf{X}_{\mathcal{A}} \mathrel{-\!\!\circ\!\!-} \mathbf{X}_{\mathcal{B}} \mathrel{-\!\!\circ\!\!-} W_{\mathcal{B}}$ that

$$n\mu_{\mathcal{A},\mathcal{B}} \leqslant I(W_{\mathcal{A}}; W_{\mathcal{B}}) \tag{B.6}$$

$$= I(W_{\mathcal{A}}; \mathbf{X}_{\mathcal{A}}) + I(W_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) - I(W_{\mathcal{A}} W_{\mathcal{B}}; \mathbf{X}_{\mathcal{A}} \mathbf{X}_{\mathcal{B}}) \tag{B.7}$$

$$= \sum_{i=1}^{n} \Big[I(U_{\mathcal{A},i}; X_{\mathcal{A},i}) + I(U_{\mathcal{B},i}; X_{\mathcal{B},i}) - I(U_{\mathcal{A},i} U_{\mathcal{B},i}; X_{\mathcal{A},i} X_{\mathcal{B},i}) \Big]. \tag{B.8}$$

Now a standard time-sharing argument shows $\mathcal{R} \subseteq \mathcal{R}_o$ (see, e. g., [16, Section 4.5.2]). Lemma 2.13 implies $\mathcal{R}_o \subseteq \mathcal{R}'_o$.

B.2 PROOF OF PROPOSITION 5.5

Pick an arbitrary $k \in [K]$. For nonempty $\mathcal{A}, \mathcal{B} \subseteq [K]$ with $k \in \mathcal{B}$ we can write $H(X_{\mathcal{A}} | U_{\mathcal{B}}) = \mathbb{E}_{U_k}\big[f_{\mathcal{A},\mathcal{B}}(p_{X_k|U_k}(\,\cdot\,|U_k))\big]$ where

$$f_{\mathcal{A},\mathcal{B}}(p_{X_k|U_k}(\,\cdot\,|u_k)) := H(X_{\mathcal{A}} | U_{\mathcal{B}\setminus k}, U_k = u_k). \tag{B.9}$$

Furthermore, $H(U_{\mathcal{A}}|U_{\mathcal{B}}) = \mathbb{E}_{U_k}\big[g_{\mathcal{A},\mathcal{B}}\big(p_{X_k|U_k}(\,\cdot\,|U_k)\big)\big]$ where

$$g_{\mathcal{A},\mathcal{B}}\big(p_{X_k|U_k}(\,\cdot\,|u_k)\big) := H\big(U_{\mathcal{A}}\big|U_{\mathcal{B}\setminus k}, U_k = u_k\big). \tag{B.10}$$

Observe that both $f_{\mathcal{A},\mathcal{B}}$ and $g_{\mathcal{A},\mathcal{B}}$ are continuous functions of the p.m.f. $p_{X_k|U_k}(\,\cdot\,|u_k)$. Apply the support lemma [16, Appendix C] with the functions $f_{\mathcal{A},\mathcal{B}}$ and $g_{\mathcal{A},\mathcal{B}}$ for all nonempty $\mathcal{A},\mathcal{B} \subseteq [K]$ such that $k \in \mathcal{B}$, and $|\mathcal{X}_k| - 1$ test functions, which guarantee that the marginal distribution p_{X_k} does not change. We obtain a new random variable $\hat{U}_k$ with $H\big(X_{\mathcal{A}}\big|U_{\mathcal{B}\setminus k}\hat{U}_k\big) = H(X_{\mathcal{A}}|U_{\mathcal{B}})$ and $H\big(U_{\mathcal{A}}\big|U_{\mathcal{B}\setminus k}\hat{U}_k\big) = H(U_{\mathcal{A}}|U_{\mathcal{B}})$. By rewriting (5.6)–(5.8) in terms of conditional entropies, it is evident that the defining inequalities for $\mathcal{R}_i$ remain the same when replacing U_k by $\hat{U}_k$. The support of $\hat{U}_k$ satisfies the required cardinality bound:

There are (2^K-1) ways to choose $\mathcal{A}$ and 2^{K-1} ways to choose $\mathcal{B}$.

$$\begin{aligned} |\hat{\mathcal{U}}_k| &\leqslant |\mathcal{X}_k| - 1 + 2(2^K-1)2^{K-1} && \text{(B.11)}\\ &= |\mathcal{X}_k| - 1 + 2^{2K} - 2^K && \text{(B.12)}\\ &\leqslant |\mathcal{X}_k| + 4^K. && \text{(B.13)} \end{aligned}$$

The same process is repeated for every $k \in [K]$.

B.3 PROOF OF LEMMA 5.12

Fix $0 < \varepsilon', \varepsilon'' < \varepsilon$ and set $\widetilde{R}_k = I(X_k;U_k) + \varepsilon''/2$ for each $k \in [K]$.

- **Encoding:** For $n \in \mathbb{N}$ define $\widetilde{M}_k := 2^{n\widetilde{R}_k}$. We apply the generalized Markov lemma (Lemma 2.20) and consider the random codebooks $C_k := (\mathbf{V}_m^{(k)})_{m\in[\widetilde{M}_k]}$, which are drawn independently uniform from $\mathcal{T}^n_{[U_k]}$ for each $k \in [K]$. Denote the resulting randomized coding functions as $\widetilde{W}_k = \widetilde{f}_k(\mathbf{X}_k, C_k)$ and the corresponding decoded value as $\widetilde{\mathbf{U}}_k := \mathbf{V}^{(k)}_{\widetilde{W}_k}$. If n is chosen large enough we therefore have

  $$P\Big\{(\widetilde{\mathbf{U}}_{[K]}, \mathbf{X}_{[K]}) \notin \mathcal{T}^n_{[U_{[K]}X_{[K]}]}\Big\} \leqslant \varepsilon'. \tag{B.14}$$

  Next, we introduce (deterministic) binning. If $R_k < I(X_k;U_k)$, partition $[\widetilde{M}_k]$ into $M_k := 2^{n(R_k+\varepsilon'')}$ equally sized, consecutive bins, each of size $2^{n\Delta_k}$ with

  $$\Delta_k := \widetilde{R}_k - R_k - \varepsilon'' = I(X_k;U_k) - R_k - \frac{\varepsilon''}{2}. \tag{B.15}$$

  The deterministic function $\beta_k\colon [\widetilde{M}_k] \to [M_k]$ maps a codeword index to the index of the bin, it belongs to. In total the encoding function becomes $f_k := \beta_k \circ \widetilde{f}_k$. If $R_k \geqslant I(X_k;U_k)$, we do not require binning and let β_k be the identity on $[\widetilde{M}_k]$ and hence $f_k := \widetilde{f}_k$.

- **Decoding:** Given the codebooks, we define a decoding procedure $g_{\mathcal{A},\tilde{\mathcal{A}}}\colon [M_{\mathcal{A}}] \to \mathcal{U}^n_{\tilde{\mathcal{A}}}$ for each $\varnothing \neq \tilde{\mathcal{A}} \subseteq \mathcal{A} \subseteq [K]$, to be carried out as follows: Given $w_{\mathcal{A}} \in [M_{\mathcal{A}}]$, let $\tilde{m}_{\mathcal{A}} := \beta_k^{-1}(w_k)_{k\in\mathcal{A}} \subseteq [\widetilde{M}_{\mathcal{A}}]$ be all indices that belong to the bins $w_{\mathcal{A}}$. Consider only the typical sequences $\mathbf{V}^{(\mathcal{A})}_{\tilde{m}_{\mathcal{A}}} \cap \mathcal{T}^n_{[U_{\mathcal{A}}]} =: \Phi \subseteq \mathcal{U}^n_{\mathcal{A}}$. If $\Phi \neq \varnothing$, choose the lexicographically smallest element of $\Phi_{\tilde{\mathcal{A}}}$, otherwise choose the lexicographically smallest element of $\mathbf{V}^{(\tilde{\mathcal{A}})}_{\tilde{m}_{\tilde{\mathcal{A}}}}$.

We use $[M_{\mathcal{A}}] = [M_k]_{k\in\mathcal{A}}$.

$\Phi_{\tilde{\mathcal{A}}}$ *is the restriction of* Φ *to the indices* $\tilde{\mathcal{A}}$.

Let $\mathcal{A}, \tilde{\mathcal{A}}, \mathcal{B}, \tilde{\mathcal{B}} \subseteq [K]$ be sets of indices such that the conditions (5.6) and (5.7) are satisfied. Using $W_k := f_k(\mathbf{X}_k, C_k)$ and the randomized decodings $\hat{\mathbf{U}}_1 := g_{\mathcal{A},\tilde{\mathcal{A}}}(W_{\mathcal{A}}, C_{\mathcal{A}})$, $\hat{\mathbf{U}}_2 := g_{\mathcal{B},\tilde{\mathcal{B}}}(W_{\mathcal{B}}, C_{\mathcal{B}})$, consider the error event $\mathcal{E}_0 := \{(\hat{\mathbf{U}}_1, \mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{B}}, \hat{\mathbf{U}}_2) \notin \mathcal{T}^n_{[U_{\tilde{\mathcal{A}}} X_{\mathcal{A}} X_{\mathcal{B}} U_{\tilde{\mathcal{B}}}]}\}$. Define the other events

$$\mathcal{E}_1 := \{(\tilde{\mathbf{U}}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{B}}, \tilde{\mathbf{U}}_{\mathcal{B}}) \notin \mathcal{T}^n_{[U_{\tilde{\mathcal{A}}} X_{\mathcal{A}} X_{\mathcal{B}} U_{\tilde{\mathcal{B}}}]}\}, \tag{B.16}$$

$$\mathcal{E}_2 := \left\{ \left| \left(\mathbf{V}^{(\mathcal{A})}_{\mathfrak{W}_{\mathcal{A}}} \cap \mathcal{T}^n_{[U_{\mathcal{A}}]} \right)_{\tilde{\mathcal{A}}} \right| > 1 \right\}, \tag{B.17}$$

$$\mathcal{E}_3 := \left\{ \left| \left(\mathbf{V}^{(\mathcal{B})}_{\mathfrak{W}_{\mathcal{B}}} \cap \mathcal{T}^n_{[U_{\mathcal{B}}]} \right)_{\tilde{\mathcal{B}}} \right| > 1 \right\}, \tag{B.18}$$

where we used the random sets of indices $\mathfrak{W}_{\mathcal{A}} := \beta_k^{-1}(W_k)_{k\in\mathcal{A}}$ and $\mathfrak{W}_{\mathcal{B}} := \beta_k^{-1}(W_k)_{k\in\mathcal{B}}$. We clearly have $\mathcal{E}_0 \subseteq \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3$ and thus

$$P\{\mathcal{E}_0\} \leqslant P\{\mathcal{E}_1\} + P\{\mathcal{E}_2 | \mathcal{E}_1^c\} + P\{\mathcal{E}_3 | \mathcal{E}_1^c\} \tag{B.19}$$

$$\overset{\text{(B.14)}}{\leqslant} P\{\mathcal{E}_2 | \mathcal{E}_1^c\} + P\{\mathcal{E}_3 | \mathcal{E}_1^c\} + \varepsilon'. \tag{B.20}$$

We can partition $\mathfrak{W}_{\mathcal{A}} = \bigcup_{\hat{\mathcal{A}} \subseteq \mathcal{A}} \mathfrak{D}_{\hat{\mathcal{A}}}$ into (random) subsets

$$\mathfrak{D}_{\hat{\mathcal{A}}} := \left\{ \tilde{w}_{\mathcal{A}} \in \mathfrak{W}_{\mathcal{A}} : \tilde{w}_{\hat{\mathcal{A}}^c} = \tilde{W}_{\hat{\mathcal{A}}^c} \text{ and } \tilde{w}_k \neq \tilde{W}_k, \forall k \in \hat{\mathcal{A}} \right\}, \tag{B.21}$$

where we used $\hat{\mathcal{A}}^c := \mathcal{A} \setminus \hat{\mathcal{A}}$. Observe that $\mathfrak{D}_{\varnothing} = \{\tilde{W}_{\mathcal{A}}\}$. For each set $\varnothing \neq \hat{\mathcal{A}} \subseteq \mathcal{A}$ we define the error event

$$\mathcal{E}_{\hat{\mathcal{A}}} := \left\{ \mathbf{V}^{(\mathcal{A})}_{\mathfrak{D}_{\hat{\mathcal{A}}}} \cap \mathcal{T}^n_{[U_{\mathcal{A}}]} \neq \varnothing \right\} \tag{B.22}$$

and obtain

$$\mathcal{E}_2 \subseteq \bigcup_{\substack{\hat{\mathcal{A}} \subseteq \mathcal{A}: \\ \hat{\mathcal{A}} \cap \tilde{\mathcal{A}} \neq \varnothing}} \mathcal{E}_{\hat{\mathcal{A}}}, \tag{B.23}$$

which implies

$$P\{\mathcal{E}_2 | \mathcal{E}_1^c\} \leqslant \sum_{\substack{\hat{\mathcal{A}} \subseteq \mathcal{A}: \\ \hat{\mathcal{A}} \cap \tilde{\mathcal{A}} \neq \varnothing}} P\left\{\mathcal{E}_{\hat{\mathcal{A}}} \middle| \mathcal{E}_1^c\right\}. \tag{B.24}$$

By construction, $\mathfrak{D}_{\hat{\mathcal{A}}}$ has $\prod_{k\in\hat{\mathcal{A}}}(2^{n\Delta_k}-1)$ elements. For $\tilde{w}_{\mathcal{A}} \in \mathfrak{D}_{\hat{\mathcal{A}}}$ we have that $\mathbf{V}_{\tilde{w}_{\hat{\mathcal{A}}}}^{(\hat{\mathcal{A}})}$ are uniformly distributed on $\prod_{k\in\hat{\mathcal{A}}} \mathcal{T}^n_{[U_k]}$ and $\tilde{w}_{\hat{\mathcal{A}}^c} = \widetilde{W}_{\hat{\mathcal{A}}^c}$. Given $\mathcal{E}_1^c$, we have in particular $\widetilde{\mathbf{U}}_{\mathcal{A}} \in \mathcal{T}^n_{[U_{\mathcal{A}}]}$. Thus, for any $\mathbf{u}_{\hat{\mathcal{A}}^c} \in \mathcal{T}^n_{[U_{\hat{\mathcal{A}}^c}]}$, we can conclude,

$$P\left\{\mathcal{E}_{\hat{\mathcal{A}}}\middle|\mathcal{E}_1^c, \widetilde{\mathbf{U}}_{\hat{\mathcal{A}}^c} = \mathbf{u}_{\hat{\mathcal{A}}^c}\right\} \tag{B.25}$$

$$= P\left\{\bigcup_{\tilde{w}_{\mathcal{A}}\in\mathfrak{D}_{\hat{\mathcal{A}}}} \{\mathbf{V}_{\tilde{w}_{\mathcal{A}}}^{(\mathcal{A})} \in \mathcal{T}^n_{[U_{\mathcal{A}}]}\}\middle|\mathcal{E}_1^c, \widetilde{\mathbf{U}}_{\hat{\mathcal{A}}^c} = \mathbf{u}_{\hat{\mathcal{A}}^c}\right\} \tag{B.26}$$

$$\leqslant \sum_{\tilde{w}_{\mathcal{A}}\in\mathfrak{D}_{\hat{\mathcal{A}}}} P\left\{\mathbf{V}_{\tilde{w}_{\mathcal{A}}}^{(\mathcal{A})} \in \mathcal{T}^n_{[U_{\mathcal{A}}]}\middle|\mathcal{E}_1^c, \widetilde{\mathbf{U}}_{\hat{\mathcal{A}}^c} = \mathbf{u}_{\hat{\mathcal{A}}^c}\right\} \tag{B.27}$$

$$\leqslant 2^{n\left(\sum_{k\in\hat{\mathcal{A}}}\Delta_k\right)} \frac{\left|\mathcal{T}^n_{[U_{\hat{\mathcal{A}}}|U_{\hat{\mathcal{A}}^c}]}(\mathbf{u}_{\hat{\mathcal{A}}^c})\right|}{\prod_{k\in\hat{\mathcal{A}}}\left|\mathcal{T}^n_{[U_k]}\right|} \tag{B.28}$$

$$\leqslant 2^{n\left(\sum_{k\in\hat{\mathcal{A}}}\Delta_k\right)} \frac{2^{n\left(H(U_{\hat{\mathcal{A}}}|U_{\hat{\mathcal{A}}^c})+\varepsilon_0(n)\right)}}{2^{n\left(\sum_{k\in\hat{\mathcal{A}}}H(U_k)-\varepsilon_k(n)\right)}} \tag{B.29}$$

$$\leqslant 2^{n\left(\varepsilon(n)+H(U_{\hat{\mathcal{A}}}|U_{\hat{\mathcal{A}}^c})+\sum_{k\in\hat{\mathcal{A}}}(\Delta_k-H(U_k))\right)}, \tag{B.30}$$

where $\varepsilon(n) = \sum_{k\in\hat{\mathcal{A}}\cup 0}\varepsilon_k(n)$ goes to zero as $n \to \infty$. Here, (B.29) follows from parts 2 and 3 of Lemma 2.19. We observe that the definition of $\widetilde{R}_k$ and (5.6) imply for any $\varnothing \neq \hat{\mathcal{A}} \subseteq \mathcal{A}$ with $\hat{\mathcal{A}} \cap \widetilde{\mathcal{A}} \neq \varnothing$ that

$$\sum_{k\in\hat{\mathcal{A}}}\Delta_k \leqslant -\frac{\varepsilon''}{2} - H(U_{\hat{\mathcal{A}}}|U_{\hat{\mathcal{A}}^c}) + \sum_{k\in\hat{\mathcal{A}}} H(U_k). \tag{B.31}$$

Marginalize over $\widetilde{\mathbf{U}}_{\hat{\mathcal{A}}^c}$ in (B.30) and use (B.31) to obtain

$$P\{\mathcal{E}_{\hat{\mathcal{A}}}|\mathcal{E}_1^c\} \leqslant 2^{n\left(\varepsilon(n)-\frac{\varepsilon''}{2}\right)} \leqslant \varepsilon' \tag{B.32}$$

for n large enough. Applying the same arguments to $P\{\mathcal{E}_3|\mathcal{E}_1^c\}$ and combining (B.20), (B.24) and (B.32), we have

$$P\{\mathcal{E}_0\} \leqslant \varepsilon' + 2^{|\mathcal{A}|}\varepsilon' + 2^{|\mathcal{B}|}\varepsilon' \leqslant 2^K\varepsilon'. \tag{B.33}$$

For a set $\varnothing \neq \mathcal{A} \subseteq [K]$, we next analyze the random quantity $L_{\mathcal{A}} := \left|C_{\mathcal{A}} \cap \mathcal{T}^n_{[U_{\mathcal{A}}]}\right|$. For n large enough, we have for any $\widetilde{\mathbf{V}}_{\mathcal{A}} \in C_{\mathcal{A}}$

$$\begin{aligned}
\mathbb{E}[L_{\mathcal{A}}] &\leqslant \mathbb{E}\left[\sum_{\mathbf{V}_{\mathcal{A}} \in C_{\mathcal{A}}} \mathbb{E}\left[\mathbb{1}_{\mathcal{T}^n_{[U_{\mathcal{A}}]}}(\mathbf{V}_{\mathcal{A}}) \middle| C_{\mathcal{A}}\right]\right] && \text{(B.34)}\\
&= \left(\prod_{k \in \mathcal{A}} \widetilde{M}_k\right) \mathbb{E}\left[\mathbb{1}_{\mathcal{T}^n_{[U_{\mathcal{A}}]}}(\widetilde{\mathbf{V}}_{\mathcal{A}})\right] && \text{(B.35)}\\
&= \left(\prod_{k \in \mathcal{A}} \widetilde{M}_k\right) \frac{\left|\mathcal{T}^n_{[U_{\mathcal{A}}]}\right|}{\prod_{k \in \mathcal{A}} \left|\mathcal{T}^n_{[U_k]}\right|} && \text{(B.36)}\\
&\leqslant \left(\prod_{k \in \mathcal{A}} \widetilde{M}_k\right) \frac{2^{n\left(H(U_{\mathcal{A}}) + \varepsilon_0(n)\right)}}{2^{n\left(\sum_{k \in \mathcal{A}} H(U_k) - \varepsilon_k(n)\right)}} && \text{(B.37)}\\
&\leqslant \left(\prod_{k \in \mathcal{A}} \widetilde{M}_k\right) 2^{n\left(H(U_{\mathcal{A}}) - \sum_{k \in \mathcal{A}} H(U_k) + \varepsilon(n)\right)} && \text{(B.38)}\\
&= 2^{n\left(H(U_{\mathcal{A}}) + \varepsilon(n) + \sum_{k \in \mathcal{A}} I(U_k;X_k) + \frac{\varepsilon''}{2} - H(U_k)\right)} && \text{(B.39)}\\
&= 2^{n\left(H(U_{\mathcal{A}}) + \varepsilon(n) + |\mathcal{A}|\frac{\varepsilon''}{2} - \sum_{k \in \mathcal{A}} H(U_k|X_k)\right)} && \text{(B.40)}\\
&= 2^{n\left(I(U_{\mathcal{A}};X_{\mathcal{A}}) + \varepsilon(n) + |\mathcal{A}|\frac{\varepsilon''}{2}\right)}. && \text{(B.41)}
\end{aligned}$$

where $\varepsilon(n) = \sum_{k \in \mathcal{A} \cup 0} \varepsilon_k(n)$ goes to zero as $n \to \infty$. Here, (B.37) follows from parts 1 and 2 of Lemma 2.19. Assume that ε'' is such that $K\varepsilon''/2 < \varepsilon$. Defining the error event $\mathcal{E}_4 = \left\{L_{\mathcal{A}} \geqslant 2^{n(I(U_{\mathcal{A}};X_{\mathcal{A}})+\varepsilon)}\right\}$, we apply Markov's inequality, Theorem 2.1, and obtain for n large enough

$$P\{\mathcal{E}_4\} \leqslant 2^{n\left(\varepsilon(n) - \varepsilon + |\mathcal{A}|\frac{\varepsilon''}{2}\right)} \leqslant \varepsilon'. \qquad \text{(B.42)}$$

Using (B.33) and (B.42) we can apply Lemma 2.12 and obtain deterministic encoding functions $f_k \colon \mathcal{X}_k^n \to \mathcal{M}_k$, and deterministic decoding functions $g_{\mathcal{A},\widetilde{\mathcal{A}}} \colon \mathcal{M}_{\mathcal{A}} \to \mathcal{U}^n_{\widetilde{\mathcal{A}}}$ such that (5.35) holds whenever the conditions (5.6) and (5.7) are satisfied. Taking into account that $g_{\mathcal{A},\widetilde{\mathcal{A}}}(\mathcal{M}_{\mathcal{A}}) \times g_{\mathcal{B},\widetilde{\mathcal{B}}}(\mathcal{M}_{\mathcal{B}}) \subseteq C_{\widetilde{\mathcal{A}} \cup \widetilde{\mathcal{B}}}$, we also have (5.34). (Note that, given a specific code, the condition $P\left\{\mathcal{E}_4 \middle| C_{[K]} = c_{[K]}\right\} < 1$ already implies $P\left\{\mathcal{E}_4 \middle| C_{[K]} = c_{[K]}\right\} = 0$ as the event $\mathcal{E}_4$ is determined by the code $C_{[K]}$ alone.)

B.4 PROOF OF PROPOSITION 5.17

Pick arbitrary $j, k \in [J]$. For nonempty $\mathcal{B} \subseteq [J]$ with $j \in \mathcal{B}$ we can write $H(X_k|U_{\mathcal{B}}) = \mathbb{E}_{U_j}\big[f_{k,\mathcal{B}}\big(p_{X_j|U_j}(\,\cdot\,|U_j)\big)\big]$ as well as $H(Y|U_{\mathcal{B}}) = \mathbb{E}_{U_j}\big[g_{\mathcal{B}}\big(p_{X_j|U_j}(\,\cdot\,|U_j)\big)\big]$, where

$$f_{k,\mathcal{B}}\big(p_{X_j|U_j}(\,\cdot\,|u_j)\big) := H\big(X_k\big|U_{\mathcal{B}\setminus j}, U_j = u_j\big), \tag{B.43}$$

$$g_{\mathcal{B}}\big(p_{X_j|U_j}(\,\cdot\,|u_j)\big) := H\big(Y\big|U_{\mathcal{B}\setminus j}, U_j = u_j\big). \tag{B.44}$$

Observe that $f_{k,\mathcal{B}}$ and $g_{\mathcal{B}}$ are continuous functions of $p_{X_j|U_j}(\,\cdot\,|u_j)$. Apply the support lemma [16, Appendix C] with the functions $f_{k,\mathcal{B}}$ and $g_{\mathcal{B}}$ for all $k \in [J]$, $j \in \mathcal{B} \subseteq [J]$, and $|\mathcal{X}_j| - 1$ test functions, which guarantee that the marginal distribution p_{X_j} does not change. We obtain a new random variable $\hat{U}_j$ with $H(X_k|U_{\mathcal{B}\setminus j}\hat{U}_j) = H(X_k|U_{\mathcal{B}})$ and $H(Y|U_{\mathcal{B}\setminus j}\hat{U}_j) = H(Y|U_{\mathcal{B}})$. By rewriting (5.63)–(5.68) in terms of conditional entropies, it is evident that the defining inequalities for $\mathcal{R}_{\mathrm{MI}}^{(\sqsubset,\mathcal{J})}$ remain the same when replacing U_j by $\hat{U}_j$. The support of $\hat{U}_j$ satisfies the required cardinality bound

There are J ways to choose k and 2^{J-1} ways to choose $\mathcal{B}$.

$$|\hat{\mathcal{U}}_j| \leqslant |\mathcal{X}_j| - 1 + J2^{J-1} + 2^{J-1} \tag{B.45}$$

$$\leqslant |\mathcal{X}_j| + 4^J. \tag{B.46}$$

The same process is repeated for every $j \in [J]$.

B.5 PROOF OF THEOREM 5.18

$\mathcal{P}_*$ is defined in (5.11).

Pick a total order $\sqsubset$ on $[J]$, a set $\mathcal{E} \subseteq [J]$ and $(U_{[J]}, \varnothing) \in \mathcal{P}_*$. To obtain a code we apply Lemma 5.12 with $K = J+1$, $X_K = U_K = Y$, $\mathcal{B} = \widetilde{\mathcal{B}} = \{K\}$, $\widetilde{\mathcal{A}} = \mathcal{A}$ for all $\varnothing \neq \mathcal{A} \subseteq [J]$, and rates $R_j = I(U_j; X_j|U_{\sqsupset j})$, $R_K = \log_2|\mathcal{Y}|$, as suggested by Proposition 5.7. As in the proof of Lemma 5.12 let $\tilde{f}_j$ denote the encoding function without binning and with rate $n^{-1}\log_2|\tilde{f}_j| \leqslant I(U_j; X_j) + \frac{\varepsilon}{2}$. Furthermore, let f'_j be the encoding function including binning, obtaining a rate of $n^{-1}\log_2|f'_j| \leqslant I(U_j; X_j|U_{\sqsupset j}) + \varepsilon$. Finally we obtain the $(n, R_{[J]} + \varepsilon)$ code $f_{[J]}$ by setting $f_j := \tilde{f}_j$ for $j \in \mathcal{E}$ and $f_j := f'_j$ for $j \notin \mathcal{E}$. Let the decoding functions be $g_{\mathcal{A}} := g_{\mathcal{A},\mathcal{A}}$ for all $\varnothing \neq \mathcal{A} \subseteq [J]$. Furthermore, for each $j \in [J]$, we define the decoding function $\tilde{g}_j$, which maps $\widetilde{W}_j := \tilde{f}_j(\mathbf{X}_j)$ onto its codebook entry, i.e., $\tilde{g}_j(w) = \mathbf{V}_w^{(j)}$ (using the notation from Appendix B.3). Also let $W_j := f_j(\mathbf{X}_j)$ and $W'_j := f'_j(\mathbf{X}_j)$. For later use, we note that W'_j is a function of W_j, which is in turn a function of $\widetilde{W}_j$.

Let the event $\mathcal{S}'_{\mathcal{A}}$ be the success event that $\big(\mathbf{Y}, \mathbf{X}_{\mathcal{A}}, g_{\mathcal{A}}(W'_{\mathcal{A}})\big) \in \mathcal{T}^n_{[YX_{\mathcal{A}}U_{\mathcal{A}}]}$ holds. Also let $\widetilde{\mathcal{S}}_j$ be the event that $\big(\mathbf{Y}, \mathbf{X}_j, \widetilde{g}_j(\widetilde{W}_j)\big) \in \mathcal{T}^n_{[YX_jU_j]}$. For any $\mathcal{A} = \sqsupseteq k$, $k \in [J]$, and $\hat{\mathcal{A}} \subseteq \mathcal{A}$, we have

$$
\begin{aligned}
\sum_{j \in \hat{\mathcal{A}}} R_j &= \sum_{j \in \hat{\mathcal{A}}} I\big(U_j; X_j \big| U_{\sqsupset j}\big) && \text{(B.47)}\\
&\geqslant \sum_{j \in \hat{\mathcal{A}}} I\big(U_j; X_{\hat{\mathcal{A}}} \big| U_{\sqsupset j}, U_{\mathcal{A} \setminus \hat{\mathcal{A}}}\big) && \text{(B.48)}\\
&= I\big(U_{\hat{\mathcal{A}}}; X_{\hat{\mathcal{A}}} \big| U_{\mathcal{A} \setminus \hat{\mathcal{A}}}\big). && \text{(B.49)}
\end{aligned}
$$

Thus, condition (5.6) is satisfied and for n large enough we have $P\{\mathcal{S}'_{\mathcal{A}}\} \geqslant 1 - \varepsilon$ by Lemma 5.12. Clearly also $P\{\widetilde{\mathcal{S}}_j\} \geqslant 1 - \varepsilon$ for each $j \in [J]$ and n large enough, using parts 1 and 2 of Lemma 2.18.

Pick an arbitrary $\varepsilon' > 0$. Provided that n is large enough and ε small enough, we have for any $\mathcal{A} = \sqsupseteq k$

In what follows, we will routinely merge expressions that can be made arbitrarily small (for n large and ε sufficiently small) and bound them by ε'.

$$
\begin{aligned}
\frac{1}{n} I(\mathbf{Y}; W_{\mathcal{A}}) &\geqslant \frac{1}{n} I\big(\mathbf{Y}; W'_{\mathcal{A}}\big) && \text{(B.50)}\\
&\geqslant \frac{1}{n} I\big(\mathbf{Y}; g_{\mathcal{A}}(W'_{\mathcal{A}})\big) && \text{(B.51)}\\
&= H(Y) - \frac{1}{n} H\big(\mathbf{Y} \big| g_{\mathcal{A}}(W'_{\mathcal{A}})\big) && \text{(B.52)}\\
&\geqslant H(Y) - \frac{1}{n} H\Big(\mathbf{Y}, \mathbb{1}_{\mathcal{S}'_{\mathcal{A}}} \Big| g_{\mathcal{A}}(W'_{\mathcal{A}})\Big) && \text{(B.53)}\\
&= H(Y) - \frac{1}{n} H\Big(\mathbb{1}_{\mathcal{S}'_{\mathcal{A}}} \Big| g_{\mathcal{A}}(W'_{\mathcal{A}})\Big) - \frac{1}{n} H\Big(\mathbf{Y} \Big| g_{\mathcal{A}}(W'_{\mathcal{A}}), \mathbb{1}_{\mathcal{S}'_{\mathcal{A}}}\Big) && \text{(B.54)}\\
&\geqslant H(Y) - \varepsilon' - \frac{1}{n}(1 - \varepsilon) H\big(\mathbf{Y} \big| g_{\mathcal{A}}(W'_{\mathcal{A}}), \mathcal{S}'_{\mathcal{A}}\big) - \varepsilon H(Y) && \text{(B.55)}\\
&\geqslant H(Y) - \varepsilon' - \frac{1}{n} H\big(\mathbf{Y} \big| g_{\mathcal{A}}(W'_{\mathcal{A}}), \mathcal{S}'_{\mathcal{A}}\big) && \text{(B.56)}\\
&\geqslant H(Y) - \varepsilon' \\
&\qquad - \frac{1}{n} \sum_{\mathbf{u}_{\mathcal{A}}} P\big\{g_{\mathcal{A}}(W'_{\mathcal{A}}) = \mathbf{u}_{\mathcal{A}} \big| \mathcal{S}'_{\mathcal{A}}\big\} \log_2 \Big|\mathcal{T}^n_{[Y|U_{\mathcal{A}}]}(\mathbf{u}_{\mathcal{A}})\Big| && \text{(B.57)}\\
&\geqslant H(Y) - H(Y|U_{\mathcal{A}}) - \varepsilon' && \text{(B.58)}\\
&= I(U_{\mathcal{A}}; Y) - \varepsilon'. && \text{(B.59)}
\end{aligned}
$$

Here, (B.50) and (B.51) follow from the data processing inequality, Theorem 2.5. We applied part 1 of Lemma 2.4 in (B.57), and part 3 of Lemma 2.19 in (B.58). For $\mathcal{A} = [J]$ we specifically obtain

$$
\frac{1}{n} I\big(\mathbf{Y}; W_{[J]}\big) \geqslant I\big(U_{[J]}; Y\big) - \varepsilon' \overset{(5.67)}{\geqslant} \nu_{[J]} - \varepsilon'. \tag{B.60}
$$

For $k \in [J]$ and $\mathcal{A} = \sqsupset k$ we obtain the following chain of inequalities, where (B.65) and (B.66) will be justified subsequently.

$$
\begin{aligned}
\frac{1}{n}I(\mathbf{Y};W_k) \geqslant \frac{1}{n}I(\mathbf{Y};W'_k) &\geqslant \frac{1}{n}I(\mathbf{Y};W'_k|W'_{\mathcal{A}}) && \text{(B.61)}\\
&= \frac{1}{n}I(\mathbf{Y};W'_kW'_{\mathcal{A}}) - \frac{1}{n}I(\mathbf{Y};W'_{\mathcal{A}}) && \text{(B.62)}\\
&\overset{\text{(B.59)}}{\geqslant} I(U_{\mathcal{A}}U_k;Y) - \varepsilon' - \frac{1}{n}I(\mathbf{Y};W'_{\mathcal{A}}) && \text{(B.63)}\\
&= I(U_{\mathcal{A}}U_k;Y) - \varepsilon' - \frac{1}{n}I(\mathbf{X}_{\mathcal{A}};W'_{\mathcal{A}}) + \frac{1}{n}I(\mathbf{X}_{\mathcal{A}};W'_{\mathcal{A}}|\mathbf{Y}) && \text{(B.64)}\\
&\geqslant I(U_{\mathcal{A}}U_k;Y) - \varepsilon' - I(X_{\mathcal{A}};U_{\mathcal{A}}) \\
&\quad + H(X_{\mathcal{A}}|Y) - \frac{1}{n}H(\mathbf{X}_{\mathcal{A}}|W'_{\mathcal{A}},\mathbf{Y}) && \text{(B.65)}\\
&\geqslant I(U_{\mathcal{A}}U_k;Y) - \varepsilon' - I(X_{\mathcal{A}};U_{\mathcal{A}}) \\
&\quad + H(X_{\mathcal{A}}|Y) - H(X_{\mathcal{A}}|U_{\mathcal{A}},Y) && \text{(B.66)}\\
&= I(U_k;Y|U_{\mathcal{A}}) - \varepsilon' && \text{(B.67)}\\
&\overset{\text{(5.65)}}{\geqslant} \nu_k - \varepsilon'. && \text{(B.68)}
\end{aligned}
$$

Equality in (B.64) follows from the Markov chain $W'_{\mathcal{A}} \multimap\!\!\!\!- \mathbf{X}_{\mathcal{A}} \multimap\!\!\!\!- \mathbf{Y}$. In (B.65), we used that for ε small and n large enough, we have

$$
\begin{aligned}
\frac{1}{n}I(\mathbf{X}_{\mathcal{A}};W'_{\mathcal{A}}) &= \frac{1}{n}H(W'_{\mathcal{A}}) && \text{(B.69)}\\
&\leqslant \frac{1}{n}\sum_{j\in\mathcal{A}} H(W'_j) && \text{(B.70)}\\
&\leqslant \sum_{j\in\mathcal{A}} \Big(I(U_j;X_j|U_{\sqsupset j}) + \varepsilon\Big) && \text{(B.71)}\\
&\leqslant I(U_{\mathcal{A}};X_{\mathcal{A}}) + \varepsilon', && \text{(B.72)}
\end{aligned}
$$

where (B.70) follows from part 3 of Lemma 2.4 and Theorem 2.5, and (B.71) follows from part 1 of Lemma 2.4 and the fact that $n^{-1}\log_2|f'_j| \leqslant I(U_j;X_j|U_{\sqsupset j}) + \varepsilon$. The inequality (B.66) follows similar to (B.59) as for n large enough and ε small enough,

$$
\begin{aligned}
\frac{1}{n}H(\mathbf{X}_{\mathcal{A}}|W'_{\mathcal{A}},\mathbf{Y}) &\leqslant \frac{1}{n}H(\mathbf{X}_{\mathcal{A}}|g_{\mathcal{A}}(W'_{\mathcal{A}}),\mathbf{Y}) && \text{(B.73)}\\
&\leqslant \frac{1}{n}H(\mathbf{X}_{\mathcal{A}},\mathbb{1}_{\mathcal{S}'_{\mathcal{A}}}|g_{\mathcal{A}}(W'_{\mathcal{A}}),\mathbf{Y}) && \text{(B.74)}\\
&\leqslant \varepsilon' + \frac{1}{n}H(\mathbf{X}_{\mathcal{A}}|g_{\mathcal{A}}(W'_{\mathcal{A}}),\mathbf{Y},\mathcal{S}'_{\mathcal{A}}) && \text{(B.75)}\\
&\leqslant \varepsilon' + \frac{1}{n}\sum_{\mathbf{u}_{\mathcal{A}},\mathbf{y}} P\{g_{\mathcal{A}}(W'_{\mathcal{A}}) = \mathbf{u}_{\mathcal{A}}, \mathbf{Y} = \mathbf{y}|\mathcal{S}'_{\mathcal{A}}\}\\
&\qquad \times \log_2\left|\mathcal{T}^n_{[X_{\mathcal{A}}|U_{\mathcal{A}},Y]}(\mathbf{u}_{\mathcal{A}},\mathbf{y})\right| && \text{(B.76)}\\
&\leqslant \varepsilon' + H(X_{\mathcal{A}}|U_{\mathcal{A}},Y). && \text{(B.77)}
\end{aligned}
$$

For $k \in \mathcal{E}$, we have similarly to (B.59) that

$$\begin{aligned}
\frac{1}{n} I(\mathbf{Y}; W_k) &= \frac{1}{n} I\big(\mathbf{Y}; \widetilde{W}_k\big) && \text{(B.78)} \\
&\geqslant \frac{1}{n} I\big(\mathbf{Y}; \tilde{g}_k(\widetilde{W}_k)\big) && \text{(B.79)} \\
&= H(Y) - \frac{1}{n} H\big(\mathbf{Y} \big| \tilde{g}_k(\widetilde{W}_k)\big) && \text{(B.80)} \\
&\geqslant H(Y) - \frac{1}{n} H\big(\mathbf{Y}, \mathbb{1}_{\tilde{\mathcal{S}}_k} \big| \tilde{g}_k(\widetilde{W}_k)\big) && \text{(B.81)} \\
&\geqslant H(Y) - \frac{1}{n} H\big(\mathbb{1}_{\tilde{\mathcal{S}}_k}\big) - \frac{1}{n} H\big(\mathbf{Y} \big| \tilde{g}_k(\widetilde{W}_k), \mathbb{1}_{\tilde{\mathcal{S}}_k}\big) && \text{(B.82)} \\
&\geqslant H(Y) - \varepsilon' - \frac{1}{n} H\big(\mathbf{Y} \big| \tilde{g}_k(\widetilde{W}_k), \tilde{\mathcal{S}}_k\big) && \text{(B.83)} \\
&\geqslant H(Y) - \varepsilon' - \frac{1}{n} \sum_{\mathbf{u}_k} P\Big\{\tilde{g}_k(\widetilde{W}_k) = \mathbf{u}_k \Big| \tilde{\mathcal{S}}_k\Big\} \\
&\qquad\qquad \times \log_2 \Big|\mathcal{T}^n_{[Y|U_k]}(\mathbf{u}_k)\Big| && \text{(B.84)} \\
&\geqslant H(Y) - \varepsilon' - H(Y|U_k) && \text{(B.85)} \\
&= I(U_k; Y) - \varepsilon' \overset{(5.66)}{\geqslant} \nu_k - \varepsilon'. && \text{(B.86)}
\end{aligned}$$

B.6 PROOF OF THEOREM 5.19

For $(\nu_\Pi, R_{[J]}) \in \mathcal{R}_{\mathrm{MI}}$ we apply Definition 5.6, choosing an $(n, R_{[J]})$ code $f_{[J]}$ for $X_{[J]}$ and define $W_j := f_j(\mathbf{X}_j)$ for $j \in [J]$. For any $\mathcal{A} \subseteq [J]$ we thus have

$$\frac{1}{n} I(W_{\mathcal{A}}; \mathbf{Y}) \geqslant \nu_{\mathcal{A}}. \tag{B.87}$$

With $U_{j,i} := (W_j, \mathbf{X}_{j,1}^{i-1})$ and $Q_i := (\mathbf{Y}^{i-1}, \mathbf{Y}_{i+1}^{n})$ we have

$$\begin{align}
n \sum_{j \in \mathcal{B}} R_j &\geqslant H(W_{\mathcal{B}}) \tag{B.88}\\
&= I(W_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}}) \tag{B.89}\\
&= I(W_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}} \mathbf{Y}) \tag{B.90}\\
&= I(W_{\mathcal{B}}; \mathbf{Y}) + I(W_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}} | \mathbf{Y}) \tag{B.91}\\
&= I(W_{\mathcal{A}} W_{\mathcal{B}}; \mathbf{Y}) - I\big(W_{\mathcal{A} \setminus \mathcal{B}}; \mathbf{Y} \big| W_{\mathcal{B}}\big) + I(W_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}} | \mathbf{Y}) \tag{B.92}\\
&= I(W_{\mathcal{A}}; \mathbf{Y}) + I\big(W_{\mathcal{B} \setminus \mathcal{A}}; \mathbf{Y} \big| W_{\mathcal{A}}\big) \nonumber\\
&\qquad - I\big(W_{\mathcal{A} \setminus \mathcal{B}}; \mathbf{Y} \big| W_{\mathcal{B}}\big) + I(W_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}} | \mathbf{Y}) \tag{B.93}\\
&\overset{\text{(B.87)}}{\geqslant} n\nu_{\mathcal{A}} + I\big(W_{\mathcal{B} \setminus \mathcal{A}}; \mathbf{Y} \big| W_{\mathcal{A}}\big) \nonumber\\
&\qquad - I\big(W_{\mathcal{A} \setminus \mathcal{B}}; \mathbf{Y} \big| W_{\mathcal{B}}\big) + I(W_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}} | \mathbf{Y}) \tag{B.94}\\
&\geqslant n\nu_{\mathcal{A}} - I\big(W_{\mathcal{A} \setminus \mathcal{B}}; \mathbf{Y}\big) + I(W_{\mathcal{B}}; \mathbf{X}_{\mathcal{B}} | \mathbf{Y}) \tag{B.95}\\
&= \sum_{i=1}^{n} \Big[\nu_{\mathcal{A}} - I\big(W_{\mathcal{A} \setminus \mathcal{B}}; Y_i \big| \mathbf{Y}^{i-1}\big) + I\big(W_{\mathcal{B}}; X_{\mathcal{B},i} \big| \mathbf{Y} \mathbf{X}_{\mathcal{B}}^{i-1}\big)\Big] \tag{B.96}\\
&\geqslant \sum_{i=1}^{n} \Big[\nu_{\mathcal{A}} - I\big(W_{\mathcal{A} \setminus \mathcal{B},i}; Y_i \big| Q_i\big) + I\big(W_{\mathcal{B}}; X_{\mathcal{B},i} \big| \mathbf{Y} \mathbf{X}_{\mathcal{B}}^{i-1}\big)\Big] \tag{B.97}\\
&= \sum_{i=1}^{n} \Big[\nu_{\mathcal{A}} - I\big(U_{\mathcal{A} \setminus \mathcal{B},i}; Y_i \big| Q_i\big) + I(U_{\mathcal{B},i}; X_{\mathcal{B},i} | Y_i Q_i)\Big]. \tag{B.98}
\end{align}$$

The result now follows by a standard time-sharing argument (see, e. g., [16, Section 4.5.2]). Note that the required Markov chain and the independence are satisfied.

BIBLIOGRAPHY

[1] R. Ahlswede and I. Csiszár, «Hypothesis Testing with Communication Constraints,» *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 533–542, Jul. 1986. DOI: 10.1109/TIT.1986.1057194.

[2] R. Ahlswede and J. Körner, «Source Coding with Side Information and a Converse for Degraded Broadcast Channels,» *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629–637, Nov. 1975. DOI: 10.1109/TIT.1975.1055469.

[3] R. Ahlswede and J. Körner, «On the connection between the entropies of input and output distributions of discrete memoryless channels,» in *Proc. 5th Conf. Probability Theory, Sep. 1974*, Brasov, Romania, 1977, pp. 13–23.

[4] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd ed. Springer, 2006.

[5] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, «On Hypercontractivity and the Mutual Information between Boolean Functions,» in *Proc. 51st Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Oct. 2013, pp. 13–19. DOI: 10.1109/ALLERTON.2013.6736499.

[6] T. Berger, «Multiterminal Source Coding,» in *The Information Theory Approach to Communications*, G. Longo, Ed., Springer, 1977, pp. 171–231.

[7] T. Berger, Z. Zhang, and H. Viswanathan, «The CEO Problem,» *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 887–902, May 1996. DOI: 10.1109/18.490552.

[8] C. Chapman, personal communication, Aug. 2017. [Online]. Available: https://mathoverflow.net/questions/213084/do-binary-symmetric-channels-maximize-mutual-information.

[9] Y. Cheng and G. M. Church, «Biclustering of Expression Data,» in *Proc. 8th Int. Conf. Intelligent Syst. for Molecular Biology*, vol. 8, San Diego, CA, Aug. 2000, pp. 93–103.

[10] T. A. Courtade and G. R. Kumar, «Which Boolean Functions Maximize Mutual Information on Noisy Inputs?» *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4515–4525, Aug. 2014. DOI: 10.1109/TIT.2014.2326877.

[11] T. A. Courtade and T. Weissman, «Multiterminal Source Coding Under Logarithmic Loss,» *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014. DOI: 10.1109/TIT.2013.2288257.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2006. DOI: 10.1002/047174882X.

[13] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, Aug. 2011. DOI: 10.1017/CBO9780511921889.

[14] I. S. Dhillon, S. Mallela, and D. S. Modha, «Information-theoretic Co-clustering,» in *Proc. 9^{th} ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Washington, DC, Aug. 2003, pp. 89–98. DOI: 10.1145/956750.956764.

[15] H. G. Eggleston, *Convexity*, P. Hall and F. Smithies, Eds. Cambridge University Press, 1958. DOI: 10.1017/CBO9780511566172.

[16] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[17] E. Erkip and T. M. Cover, «The Efficiency of Investment Information,» *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998. DOI: 10.1109/18.669153.

[18] S. Fujishige, *Submodular Functions and Optimization*, 2nd ed., ser. Annals of Discrete Mathematics. Elsevier Science, 2005.

[19] P. Gács and J. Körner, «Common Information Is Far Less Than Mutual Information,» *Problems of Control and Inform. Theory*, vol. 2, pp. 149–162, 1973.

[20] R. Gilad-Bachrach, A. Navot, and N. Tishby, «An Information Theoretic Tradeoff between Complexity and Accuracy,» in *Learning Theory and Kernel Machines*, Springer, 2003, pp. 595–609. DOI: 10.1007/978-3-540-45167-9_43.

[21] *GNU Octave*, Free Software Foundation. [Online]. Available: https://www.gnu.org/software/octave/.

[22] A. A. Gohari and V. Anantharam, «Evaluation of Marton's Inner Bound for the General Broadcast Channel,» *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 608–619, Feb. 2012. DOI: 10.1109/TIT.2011.2169537.

[23] R. M. Gray, *Entropy and Information Theory*, 1st and corrected ed. Springer-Verlag, 2013. [Online]. Available: https://ee.stanford.edu/~gray/it.html.

[24] B. Grünbaum, *Convex Polytopes*. Springer, New York, 2003. DOI: 10.1007/978-1-4613-0019-9.

[25] T. S. Han, «Hypothesis Testing with Multiterminal Data Compression,» *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, Nov. 1987. DOI: 10.1109/TIT.1987.1057383.

[26] T. S. Han and S. Amari, «Statistical Inference under Multiterminal Data Compression,» *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2300–2324, Oct. 1998. DOI: 10.1109/18.720540.

[27] T. S. Han and K. Kobayashi, «A Unified Achievable Rate Region for a General Class of Multiterminal Source Coding Systems,» *IEEE Trans. Inf. Theory*, vol. 26, no. 3, pp. 277–288, May 1980. DOI: 10.1109/TIT.1980.1056192.

[28] J. A. Hartigan, «Direct Clustering of a Data Matrix,» *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, Mar. 1972. DOI: 10.1080/01621459.1972.10481214.

[29] O. Heimlich, *GNU Octave Interval Package*. [Online]. Available: https://octave.sourceforge.io/interval/.

[30] W. Huleihel and O. Ordentlich, «How to Quantize n Outputs of a Binary Symmetric Channel to $n-1$ Bits?» In *Proc. IEEE Int. Symp. on Inform. Theory*, Jun. 2017, pp. 91–95. DOI: 10.1109/ISIT.2017.8006496.

[31] V. Jog and C. Nair, «An information inequality for the BSSC broadcast channel,» in *Inform. Theory and Applicat. Workshop (ITA)*, San Diego, CA, Feb. 2010, pp. 1–8. DOI: 10.1109/ITA.2010.5454102.

[32] G. Kindler, R. O'Donnell, and D. Witmer, «Remarks on the Most Informative Function Conjecture at Fixed Mean,» *arXiv preprint*, 2016. [Online]. Available: http://arxiv.org/abs/1506.03167.

[33] J. G. Klotz, D. Kracht, M. Bossert, and S. Schober, «Canalizing Boolean Functions Maximize Mutual Information,» *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2139–2147, Apr. 2014. DOI: 10.1109/TIT.2014.2304952.

[34] J. Körner and K. Marton, «How to Encode the Modulo-Two Sum of Binary Sources,» *IEEE Trans. Inf. Theory*, vol. 25, no. 2, pp. 219–221, Mar. 1979. DOI: 10.1109/TIT.1979.1056022.

[35] S. Kullback and R. A. Leibler, «On Information and Sufficiency,» *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951. DOI: 10.1214/AOMS/1177729694.

[36] G. R. Kumar and T. A. Courtade, «Which Boolean Functions are Most Informative?» In *Proc. IEEE Int. Symp. on Inform. Theory*, Istanbul, Turkey, Jul. 2013, pp. 226–230. DOI: 10.1109/ISIT.2013.6620221.

[37] S. C. Madeira and A. L. Oliveira, «Biclustering Algorithms for Biological Data Analysis: A Survey,» *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 1, no. 1, pp. 24–45, Aug. 2004. DOI: 10.1109/TCBB.2004.2.

[38] B. Mirkin, *Mathematical Classification and Clustering*. Kluwer Academic Publisher, 1996. DOI: 10.1007/978-1-4613-0457-9.

[39] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*. SIAM, 2009. DOI: 10.1137/1.9780898717716.

[40] J. R. Munkres, *Topology*. Prentice Hall, 2000.

[41] C. Nair, «Upper concave envelopes and auxiliary random variables,» *Int. J. of Advances in Eng. Sciences and Appl. Math.*, vol. 5, no. 1, pp. 12–20, Mar. 2013. DOI: 10.1007/S12572-013-0081-7.

[42] C. P. Niculescu and L.-E. Persson, *Convex Functions and Their Applications: A Contemporary Approach*. New York, NY: Springer, 2006. DOI: 10.1007/0-387-31077-0.

[43] R. O'Donnell, *Analysis of Boolean Functions*. Cambridge University Press, Jul. 2014. DOI: 10.1017/CBO9781139814782.

[44] O. Ordentlich, O. Shayevitz, and O. Weinstein, «An Improved Upper Bound for the Most Informative Boolean Function Conjecture,» *arXiv preprint*, 2015. [Online]. Available: http://arxiv.org/abs/1505.05794.

[45] A. Orlitsky and J. R. Roche, «Coding for Computing,» *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903–917, Mar. 2001.

[46] G. Pichler, *DSBS-MutInf-counterexample*, Program code, 2017. DOI: 10.5281/ZENODO.1042588. [Online]. Available: https://github.com/m3phisto/DSBS-MutInf-counterexample.

[47] G. Pichler, G. Matz, and P. Piantanida, «A Tight Upper Bound on the Mutual Information of Two Boolean Functions,» in *Proc. Inform. Theory Workshop*, Cambridge, UK, Sep. 2016, pp. 16–20. DOI: 10.1109/ITW.2016.7606787.

[48] G. Pichler, P. Piantanida, and G. Matz, «Distributed Information-Theoretic Biclustering of Two Memoryless Sources,» in *Proc. 53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 2015, pp. 426–433. DOI: 10.1109/ALLERTON.2015.7447035.

[49] G. Pichler, P. Piantanida, and G. Matz, «Distributed Information-Theoretic Biclustering,» in *Proc. IEEE Int. Symp. on Inform. Theory*, Barcelona, Spain, Jul. 2016, pp. 1083–1087. DOI: 10.1109/ISIT.2016.7541466.

[50] G. Pichler, P. Piantanida, and G. Matz, «A Multiple Description CEO Problem with Log-Loss Distortion,» in *Proc. IEEE Int. Symp. on Inform. Theory*, Aachen, Germany, Jun. 2017, pp. 111–115. DOI: 10.1109/ISIT.2017.8006500.

[51] G. Pichler, P. Piantanida, and G. Matz, «Dictator Functions Maximize Mutual Information,» *Ann. of Applied Probability*, 2017, (submitted). [Online]. Available: http://arxiv.org/abs/1604.02109.

[52] G. Pichler, P. Piantanida, and G. Matz, «Distributed Information-Theoretic Clustering,» *IEEE Trans. Inf. Theory*, 2017, (submitted). [Online]. Available: https://arxiv.org/abs/1602.04605.

[53] A. W. Roberts and D. E. Varberg, *Convex Functions*. Academic Press, 1973. DOI: 10.2307/2319679.

[54] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. McGraw-Hill, 1976. DOI: 10.2307/3608793.

[55] W. Rudin, *Functional Analysis*, 2nd ed. McGraw-Hill, 1991.

[56] R. Schneider, *Convex Bodies: The Brunn-Minkowski Theory*, 2nd ed. Cambridge University Press, 2014.

[57] C. E. Shannon, «A Mathematical Theory of Communication,» *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948. DOI: 10.1002/J.1538-7305.1948.TB01338.X.

[58] R. Sharan, *Analysis of Biological Networks: Network Modules – Clustering and Biclustering*, lecture notes, 2006. [Online]. Available: http://www.cs.tau.ac.il/~roded/courses/bnet07.html.

[59] N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, «Information-based clustering,» *Proc. of the Nat. Academy of Sciences of the United States of America*, vol. 102, no. 51, pp. 18 297–18 302, Dec. 2005. DOI: 10.1073/PNAS.0507432102.

[60] M. Talagrand, «Isoperimetry, Logarithmic Sobolev Inequalities on the Discrete Cube, and Margulis' Graph Connectivity Theorem,» *Geometric & Functional Analysis*, vol. 3, no. 3, pp. 295–314, May 1993. DOI: 10.1007/BF01895691.

[61] M. Talagrand, «On Russo's Approximate Zero-one Law,» *Ann. of Probability*, pp. 1576–1587, 1994. DOI: 10.1214/AOP/1176988612.

[62] A. Tanay, R. Sharan, and R. Shamir, «Biclustering Algorithms: A Survey,» *Handbook of Computational Molecular Biology*, vol. 9, no. 1-20, pp. 122–124, 2005. DOI: 10.1201/9781420036275.CH26.

[63] N. Tishby, F. C. Pereira, and W. Bialek, «The information bottleneck method,» in *Proc. 37th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 1999, pp. 368–377. [Online]. Available: https://arxiv.org/abs/physics/0004057.

[64] S.-Y. Tung, «Multiterminal Source Coding,» PhD thesis, Cornell University, May 1978.

[65] A. B. Wagner, B. G. Kelly, and Y. Altug, «Distributed Rate-distortion with Common Components,» *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4035–4057, 2011. DOI: 10.1109/TIT.2011.2145570.

[66] M. B. Westover and J. A. O'Sullivan, «Achievable Rates for Pattern Recognition,» *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 299–320, Jan. 2008. DOI: 10.1109/TIT.2007.911296.

[67] H. S. Witsenhausen, «On Sequences of Pairs of Dependent Random Variables,» *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 100–113, Jan. 1975. DOI: 10.1137/0128010.

[68] H. S. Witsenhausen and A. D. Wyner, «A Conditional Entropy Bound for a Pair of Discrete Random Variables,» *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, Sep. 1975. DOI: 10.1109/TIT.1975.1055437.

[69] H. Witsenhausen, «Entropy inequalities for discrete channels,» *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 610–616, Sep. 1974. DOI: 10.1109/TIT.1974.1055285.

[70] A. D. Wyner, «On Source Coding with Side Information at the Decoder,» *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 294–300, May 1975. DOI: 10.1109/TIT.1975.1055374.

[71] A. D. Wyner and J. Ziv, «The Rate-distortion Function for Source Coding with Side Information at the Decoder,» *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976. DOI: 10.1109/TIT.1976.1055508.

[72] A. Wyner, «A Theorem on the Entropy of Certain Binary Sequences and Applications: Part II,» *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 772–777, Nov. 1973. DOI: 10.1109/TIT.1973.1055108.

[73] A. Wyner and J. Ziv, «A Theorem on the Entropy of Certain Binary Sequences and Applications: Part I,» *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 769–772, Nov. 1973. DOI: 10.1109/TIT.1973.1055107.

colophon

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of November 13, 2017 (commit `5d2a59842f`).