Perspectives on Identification Systems

MINH THÀNH VŨ

Doctoral Thesis in Electrical Engineering
Stockholm, Sweden 2019

Division of Information Science and Engineering
KTH, School of Electrical Engineering and Computer Science
SE-100 44 Stockholm
SWEDEN

TRITA-EECS-AVL-2019:57
ISBN: 978-91-7873-239-5

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framläges till offentlig granskning för avläggande av teknologie doktorsexamen i telekommunikation torsdag den 29 Augusti 2019 klockan 13.15 i F3, Lindstedtsvägen 26, Stockholm.

Tryck: Universitetsservice US AB

In memory of my Grandparents.

Abstract

Identification systems such as biometric identification systems have been becoming ubiquitous. Fundamental bounds on the performance of the systems have been established in literature. In this thesis we further relax several assumptions in the identification problem and derive the corresponding fundamental regions for these settings.

The generic identification architecture is first extended so that users' information is stored in two layers. Additionally, the processing is separated in two steps where the observation sequence in the first step is a noisy, pre-processed version of the original one. This setting generalizes several known settings in the literature. Given fixed pre-processing schemes, we study optimal trade-offs in the discrete and Gaussian cases. As corollaries we also provide characterizations for related problems.

In a second aspect, the joint distribution in the identification problem is relaxed in several ways. We first assume that all users' sequences are drawn from a common distribution, which depends on a state of the system. The observation sequence is induced by a channel which has its own state. Another variant, in which the channel is fixed, however the distributions of users' sequences are not necessarily identical, is considered next. We then study the case that users' data sequence are generated independently from a mixture distribution. Optimal performance regions of these settings are provided. We further give an inner bound and an outer bound on the region when the observation channel varies arbitrarily. Additionally, we strengthen the relation between the Wyner-Ahlswede-Körner problem and the identification problem and show the equivalence of these two.

Finally, we study a binary hypothesis testing problem which decides whether or not the observation sequence is related to one user in the database. The optimal exponent of the second type of error is studied. Furthermore, we show that the single-user testing against independence problem studied by Ahlswede and Csiszár is equivalent to the identification problem as well as the Wyner-Ahlswede-Körner problem.

Sammanfattning

Identifikationssystem, som till exempel system för biometrisk identifikation, har blivit allt mer vanligt förekommande. Fundamentala begränsningar i prestanda hos sådana system har etablerats av tidigare studier. I denna avhandling reducerar vi ett antal antaganden i identifikationsproblemet samt härleder motsvarande begränsningar i prestanda under dessa förutsättningar.

Den allmänna identifikationsarkitekturen utökas så att användarnas information lagras i två lager. Vidare separeras beräkningarna i två steg där observationssekvensen i det första steget är en för-bearbetad brusig variant av originalsekvensen. Dessa förutsättningar generaliserar flera kända fall från tidigare studier av problemet. Givet valda för-bearbetningsmetoder studerar vi optimala avvägningar i det diskreta samt normalfördelade fallet. I form av följdsatser tillhandahåller vi karakteriseringar av relaterade problem.

Vidare reducerar vi på flera sätt antagandena gällande de gemensamma distributionerna i identifikationsproblemet. Först antar vi att samtliga användares sekvenser dras från en gemensam distribution som beror på systemets tillstånd. Observationssekvenserna introduceras genom en kanal med ett självständigt tillstånd. Vidare undersöker vi en variant där kanalen är fixerad, medan fördelningarna av användarnas sekvenser inte nödvändigtvis är identiska. Sedan studerar vi fallet då användarnas datasekvenser genereras oberoende från en distribution med flera komponenter. Optimala prestandaregioner härleds för dessa förutsättningar. Vi tillhandahåller även inre och yttre begränsningar för regionen när kanalobservationer varierar godtyckligt. Vi stärker öven relationen mellan Wyner-Ahlswede-Körner problemet och det studerade identifikationsproblemet samt visar att dessa är ekvivalenta.

Till sist studerar vi ett binärt hypotestest-problem vilket avser avgöra ifall observationssekvensen står i relation till någon användare i en databas. Den optimala exponenten för typ-två fel studeras. Vi visar även att testet av en enstaka användare, tidigare studerat av Ahlswede och Csiszár, samt vårt studerade identifikationsproblem är ekvivalenta med Wyner-Ahlswede-Körner problemet.

Acknowledgments

First and foremost, I'm very fortunate to have Professor Tobias Oechtering and Professor Mikael Skoglund as my supervisors. I would like to thank them for their guidance, support, patience and their overly optimistic attitude.

I wish to thank my colleagues and visitors at the ISE (former known as Communication Theory) department for creating an affable working environment. Especially, I am thankful to Zuxing and Phuong for enjoyable and joyful conversations.

I would like also to thank my collaborators, Professor Holger Boche and Dr. Moritz Wiese at TU München for their help. Further, I would like to express my gratitude towards my former supervisors Professor Marius Pesavento and Dr. Michael Muma at TU Darmstadt for their guidance during my previous studies.

Finally, I would like to thank my family and relatives for their support.

Võ Minh Thành
Stockholm, July 2019.

Contents

1

Introduction

1.1 Identification systems

The blooming numbers of smart devices and services lead to an increase in high-dimensional contents such as videos or audios. Because of the large amount of data, efficient data storage and compression mechanisms are necessary. Given an observed sequence, e.g. an image, reliable identification (recognition) of an related user inside a database is crucial in many image or video processing applications in eHealth, IoT, etc. An old but detailed overview of identification systems can be found in [PBJ00].

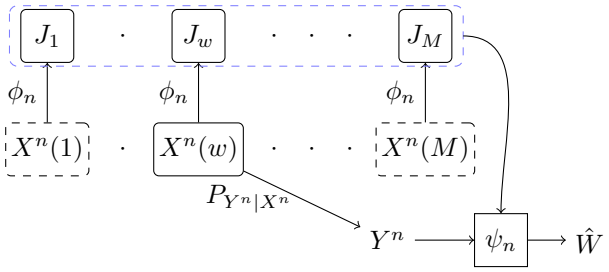


Figure 1.1: A simplified model of identification systems studied in [WKL03] and [Tun09].

The information theoretic study of an identification problem (ID) was first done by Willems in [WKL03], where he characterized the identification capacity for noisy biometric systems. The compression and distortion aspects of users' data were taken into account in [Tun09] and [TG14], where the trade-offs between compression and identification rates, and compression-distortion-capacity, respectively, were provided. In [WO08] the authors additionally considered compressing the observation and sending it to the processing center. Clustering was studied in [Wil09],

[Tun12], and [FW16] as a method to improve the search speed, where in the enrollment phase users were distributed into clusters (groups) based on their data sequences. Each user could appear in several clusters.

Generally speaking, the identification problem consists of two phases. In the first phase, *the enrollment phase*, data from M users $(x^n(i))_{i=1}^M$ are enrolled into a database as $(j_i)_{i=1}^M$, cf. Fig. 1.1, in a *compressed* or an *uncompressed* format. The users' data are not available after the enrollment phase. In the second phase, the *identification phase*, a user w is chosen uniformly at random from M users. A sequence y^n observed at the output of the observation channel $P_{Y^n|X^n}$ with the corresponding input $x^n(w)$ is available to the system. The task of the system is to identify the correct user w based on the observation y^n and the stored information in the database $(j_i)_{i=1}^M$. The *identification capacity* corresponds to the maximum number of users M such that the probability of correct identification approaches one.

The mentioned previous works provide some initial understanding about the identification systems. However, there are still significant gaps between the models and practical systems that need to be filled. For example, real data sequences are non-iid and could be generated from continuous sources. In this work, we address several of these challenges by extending and relaxing several aspects of the identification systems. We provide some motivations for each aspect below.

- *Hierarchical ID:* It can be beneficial that enrolled data are stored in parts in distributed storage nodes, e.g. reduce local storage needs when remote access to a centralized storage is available. This architecture can also be used to enhance reliability against storage failures. Furthermore, searching through the whole database and using the original observation sequence in one step can be computational expensive. We therefore propose a pre-processing scheme to address the challenges.

- *Uncertainty:* Realistic users' data and observation sequences are not necessary iid and the joint distribution might not be known exactly. However, the systems are required to be reliable regardless of these conditions. Robustness models are examined in this work.

- *Hypothesis Testing:* It is often assumed that the observation sequence is related to the data inside the system via the observation channel $P_{Y^n|X^n}$ as summarized in the previous paragraph. Again, in practice this might not be the case. Therefore, we propose a screening step to find out whether the observation sequence is legitimate or not, i.e. if y^n is related to an previously enrolled user or not.

1.2 Thesis Outline

The investigation of these perspectives in this thesis is divided into four main chapters, one technical introduction and a brief conclusion. We summarize the content

of each chapter in the following.

In Chapter 2 we review characterizations of classic settings that are relevant for latter chapters such as the lossy compression, the Wyner-Ahlswede-Körner (WAK) setting, the hypothesis testing against independence and the identification problem. The presentation focuses on showing how ideas of the achievability proofs were built up chronologically. Additionally, we discuss properties of information quantities in the mixture of discrete and continuous case which will be used in Chapter 4.

Chapter 3 studies the two-stage identification setting where users' data are stored hierarchically in two layers and the identification process is carried out in two steps to speed up the search complexity. In the first step the system outputs a list of compatible users to the second step based on a pre-processed observation sequence. Based on the information of users in the list the second step produces an estimated user and a reconstructed sequence of the true user. We present the complete trade-off between the compression rates for both layers, the list size for the first stage processing, the identification rate and the distortion level in the discrete setting. We also discuss several derivatives of the two-stage setting.

In Chapter 4 we extend the setting in Chapter 3 to the Gaussian setting. The complete trade-off characterization is also given for this setting. The proof idea extends the one of Wyner reviewed in Chapter 2 for the discrete WAK setting with a tweak in the error analysis.

Chapter 5 addresses our certainty about the underlying joint distribution of users' data sequences and the observation sequence. To facilitate the complexity of the study we focus only on the compression and identification trade-off. We consider several settings: the compound setting, the independent individual state setting, the general distribution setting as well as its specialization to mixture settings, and the arbitrary varying setting. We observe that several settings admit the same compression-identification trade-off. Additionally we show the equivalence of the identification setting and the WAK setting in the strong converse sense. Finally, we provide a strong converse line of arguments that works for both the discrete and the Gaussian identification settings.

Chapter 6 studies the problem that the observation sequence might not be related to one user in the database. We propose a binary hypothesis testing approach to address this issue based on the Neyman-Pearson framework. The characterization of the type II error exponent is given when the probability of type I of error is strictly below 1. It shows an interesting trade-off between the error exponent and the identification rate. Following the line of arguments from Chapter 5 we further show that the hypothesis testing against independence is equivalent to the WAK setting and the identification setting.

The material presented in this thesis is based on the author's joint works which are partially published in [VOS17; VOS18a; VOSB18; VOS18b; VOS19].

2

Preliminaries

This chapter recaps several classical settings in information theory. The presented ideas, results and related properties are useful for latter chapters.

2.1 Set Covering

Assume that the space $\mathcal{X}$ is given. Let $\{\mathcal{U}_\alpha\}_{\alpha\in\mathcal{I}}$ be a collection of subsets of $\mathcal{X}$ and $\mathcal{K}$ be another subset of $\mathcal{X}$. We say that $\{\mathcal{U}_\alpha\}_{\alpha\in\mathcal{I}}$ *covers* $\mathcal{K}$, if $\mathcal{K}\subset\bigcup_{\alpha\in\mathcal{I}}\mathcal{U}_\alpha$. We are often interested in finding a finite cover of a set, i.e. $|\mathcal{I}|<\infty$.
In information theory, covering arose in the context of the lossy source coding problem. The aim of this problem is to communicate a source to a destination with the minimum amount of resource (transmission rate) such that the reconstructed information is not too distorted from the original source. Formally, this action is described by a pair of compression/reconstruction mappings

$$\begin{aligned} f_n &\colon \mathcal{X}^n \to \mathcal{M}, \\ g_n &\colon \mathcal{M} \to \hat{\mathcal{X}}^n. \end{aligned} \tag{2.1}$$

The corresponding covering is given by $\{\mathcal{U}_m\}_{m\in\mathcal{M}}$, where $\mathcal{U}_m=\{x^n \mid f_n(x^n)=m\}$. The discrepancy between the original and the reconstructed sequences is measured by a function $d\colon \mathcal{X}^n\times\hat{\mathcal{X}}^n\to\mathbb{R}$, for example a distance metric. The quality of covering is usually measured in terms of the expected distortion or the excess probability for a given target distortion level. Before discussing further we make a digression to introduce some information quantities.

Definition 2.1 Given two distributions P and Q on a space $\mathcal{X}$, the KL-divergence between P and Q is defined as

$$D(P\|Q)=\begin{cases}\mathbb{E}_P\left[\log\dfrac{dP}{dQ}\right] & \text{if } P\ll Q\\ \infty & \text{otherwise}\end{cases}. \tag{2.2}$$

where $\ll$ denotes the absolute continuity relation and $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P w.r.t. Q, which is defined in detail in Theorem 2.3.

Different information theoretic quantities can be related to the divergence if the above definition is relaxed. When X is a discrete random variable the entropy of X can be obtained by $H(X) = -D(P_X\|\mu_c)$ where μ_c is the counting measure. Similarly, when P_X has a density f_x w.r.t the Lebesgue measure λ then the differential entropy of X can be defined by $h(X) = -D(P_X\|\lambda)$. In the previous two examples we have abused the integral definition of divergence, since it requires two distributions on the same alphabet. We can also define the mutual information via the divergence as follows.

Definition 2.2 Let P_{XY} be a distribution on the space $\mathcal{X} \times \mathcal{Y}$ with the corresponding marginals P_X and P_Y. The mutual information is given by

$$I(X;Y) = D(P_{XY}\|P_X \times P_Y), \tag{2.3}$$

where $P_X \times P_Y$ is the product (probability) measure on $\mathcal{X} \times \mathcal{Y}$. When $P_{XY} \ll P_X \times P_Y$ holds, we define the information density as

$$\iota(x;y) = \log \frac{dP_{XY}}{d(P_X \times P_Y)}(x,y), \tag{2.4}$$

so that in this case $I(X;Y) = \mathbb{E}_{P_{XY}}[\iota(X;Y)]$ holds.

We now get back to the lossy source coding problem. Assume that the source sequence x^n is generated iid from a distribution P_X. Constructive assigning each x^n in $\mathcal{X}^n$ to a element of the representative set $\mathcal{M}$, or correspondingly a cover, is not an easy task in general. Moreover, at large block length most of the of the probability concentrates in a set of volume $\approx 2^{nH(X)}$ which is typically much smaller than the cardinality of $\mathcal{X}^n$. The probabilistic approach for this problem goes as follows. We generate a random codebook $\boldsymbol{u}^n$ where each codeword $u^n(m)$, $m \in \mathcal{M}$, is generated iid from a marginal distribution P_U. P_U is chosen as the output distribution of a test channel $P_{U|X}$ with input P_X. Then we assign each sequence x^n to a sequence $u^n(m)$ according to some deterministic rules. It can be argued that with a suitable rule the expected distortion is within the target one plus an infinitesimal quantity. A stronger and explicit statement which expresses the excess probability is given in the following.

Lemma 2.1 *[Gal68, Lemma 9.3] Fix a test channel $P_{U^n|X^n}$. Generate M codewords $u^n(m)$ independently, $m \in [1:M]$, from the marginal distribution P_{U^n} and denote them collectively by $\boldsymbol{u}^n$. Given a codebook, $\boldsymbol{u}^n$, each source sequence x^n is mapped to an index j_0 such that $d(x^n, u^n(j_0))$ is minimized. Then*

$$\Pr\{d(X^n, U^n(J_0)) > nD\} \le \Pr\Big\{\iota(\bar{X}^n;\bar{U}^n) > nR \text{ or } d(\bar{X}^n;\bar{U}^n) > nD\Big\}$$

$$+ \exp(-Me^{-nR}), \tag{2.5}$$

where $(\bar{X}^n, \bar{U}^n) \sim P_{X^nU^n}$.

The first term on the right-hand side of Lemma 2.1 can be seen as the probability of an atypical event. Consider a special case where the joint distribution $P_{X^nU^n}$ and the distortion measure $d(x^n, u^n)$ can be factorized as

$$P_{X^nU^n} \leftarrow P_{XU}^{\otimes n},\ d(x^n, u^n) \leftarrow \sum_{i=1}^{n} d_s(x_i, u_i), \tag{2.6}$$

then due to the law of large numbers it corresponds to how much the information density and the distortion measure deviates from their mean values. The approach is, however, *inflexible* due to the coupling of x^n and $\boldsymbol{u}^n$ via the function $d(\cdot, \cdot)$, i.e., it is not straightforward to extend this approach to multi-layer covering which appears in many information theory problems such as the multiple description problem.

Strongly typical approach on *finite alphabets* was used to circumvent this problem. In the following we present a definition of a joint typical set. The individual/tuple typical set can be inferred similarly.

Definition 2.3 [EK11, p. 27] Given an $\epsilon > 0$, the set of joint ϵ-(strongly) typical sequence is defined as

$$\mathcal{T}_\epsilon^n(P_{XY}) = \left\{ (x^n, y^n) \middle| \left| \frac{|\{i\colon (x_i, y_i) = (x, y)\}|}{n} - P_{XY}(x, y) \right| < \epsilon P_{XY}(x, y) \right\}. \tag{2.7}$$

We usually abbreviate strongly typical sets as $\mathcal{T}_\epsilon^n$. The following property explains why the strongly typical approach is widely used.

Lemma 2.2 *[EK11, p. 26] Given a function* $g\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ *such that* $\mathbb{E}_{P_{XY}}(g) < \infty$. *If* $(x^n, y^n) \in \mathcal{T}_\epsilon^n(P_{XY})$ *then*

$$(1 - \epsilon)\mathbb{E}_{P_{XY}}(g) \leq \frac{1}{n} \sum_{i=1}^{n} g(x_i, y_i) \leq (1 + \epsilon)\mathbb{E}_{P_{XY}}(g). \tag{2.8}$$

Applying the above lemma to the lossy compression problem we see that if $(x^n, u^n(i)) \in \mathcal{T}_\epsilon^n$ for some i and the assumptions (2.6) are fulfilled then

$$\left| \frac{1}{n} d(x^n, u^n(i)) - \mathbb{E}[d_s(X, U)] \right| < \epsilon \mathbb{E}[d_s(X, U)]. \tag{2.9}$$

In other words, we get the guarantee for the distortion for free when $P_{U|X}$ is properly chosen. The focus is hence on building a strongly typical covering only. The following lemma quantifies the quality of covering using the strongly typical approach.

Lemma 2.3 (Strongly typical covering lemma) *[EK11, Lemma 3.3] Assume that* $\Pr\{(U^n, X^n) \in \mathcal{T}_{\epsilon_1}^n\} \to 1$ *as* $n \to \infty$ *where* (U^n, X^n) *is not necessary a sequence of iid pairs of random variables. For each realization* u^n *of* U^n*, let* 2^{nR} *sequences* $\hat{X}^n(m)$ *where* $m \in [1 : 2^{nR}]$ *be iid generated via the conditional distribution* $\prod_i P_{\hat{X}|U=u_i}$*. Then if* $R > I(X;\hat{X}|U) + \gamma_n(\epsilon_2)$ *where* $\gamma_n \to 0$ *as* $n \to \infty$ *then we have*

$$\Pr\{(U^n, X^n, \hat{X}^n(m)) \notin \mathcal{T}_{\epsilon_2}^n\} \to 0, \text{ as } n \to \infty, \tag{2.10}$$

where $\epsilon_2 > \epsilon_1$ *and* $\epsilon_2 \to 0$ *as well.*

The flexibility of the strongly typical approach is explained in the following. Assume that we want to cover $\mathcal{X}^n$ using two layers of covering $\boldsymbol{u}^n$ and $\boldsymbol{v}^n$ where given a sequence $u^n(i)$, each $v^n(i,j)$ is generated iid according to $\prod_i P_{V|U=u_i}$. We first look for a codeword (sequence) $u^n(i)$ such that $(x^n, u^n(i)) \in \mathcal{T}_\epsilon^n(P_{XU})$. Then we look for a sequence $v^n(i,j)$ such that

$$(x^n, u^n(i), v^n(i,j)) \in \mathcal{T}_\epsilon^n(P_{XUV}). \tag{2.11}$$

Lemma 2.3 states how many codewords for each layer we need such that the cover is successful with high probability. As we mentioned before, the distortion constraints are automatically fulfilled with an appropriate choice of the conditional distribution $P_{UV|X}$, if required.
We now move further from the context of lossy source coding problem. Instead of producing a hard decision which assigns each x^n to a center $u^n(i)$ we can also assign each x^n to each $u^n(i)$ with a certain probability denoted by $P_{X|U}^{\otimes n}(x^n|u^n(i))$. To measure the quality of this *soft* assignment the following metric on the space of probability measures is frequently used.

Definition 2.4 The total variation distance between two distributions P and Q on the same alphabet $\mathcal{X}$ is defined as

$$\|P - Q\|_{TV} = \sup_{\mathcal{A} \subseteq \mathcal{X}} |P(\mathcal{A}) - Q(\mathcal{A})|. \tag{2.12}$$

Another "metric" which is also widely used is the KL-divergence. It should be noted that when the space $\mathcal{X}$ is finite, the metric used in (2.7) is equivalent to the variational distance as the latter can be reformulated as

$$\|P - Q\|_{TV} = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|. \tag{2.13}$$

Compared with the (strongly/weakly) typical approach in which we want to cover the space $\mathcal{X}^n$ with high probability using a suitable codebook, herein we want to cover or approximate the distribution $P_X^{\otimes n}$ with vanishing discrepancy also via a suitable codebook. We have the following lemma.

Lemma 2.4 (Soft Covering Lemma) *[Hay06, Lemma 2], [Cuf13, Corollary VII.2] Given a joint distribution P_{XU}. Generate 2^{nR} sequences $U^n(m)$ iid via the marginal distribution P_U. Denote the collection of these sequences by $\boldsymbol{U}^n$. For each realization of these sequences let $\bar{P}_{X^n}$ be the output of the channel $P_{X|U}$ where the input is chosen uniformly at random from one of these generated sequences, i.e.*

$$\bar{P}_{X^n} = \frac{1}{2^{nR}} \sum_i P_{X|U}^{\otimes n}(\cdot|u^n(i)). \tag{2.14}$$

Then we have

$$\mathbb{E}_{\mathbf{U}^n} \|\bar{P}_{X^n} - P_X^{\otimes n}\|_{TV} \leq \Pr\{\iota(X^n;U^n) > \tau\} + \frac{1}{2}\sqrt{e^\tau/2^{nR}}. \tag{2.15}$$

One may wonder what is the relation between soft-covering and covering using conditionally, strongly typical sets. To answer this question we need the following *stronger* lemma.

Lemma 2.5 (Strong soft covering lemma) *[Cuf16, Theorem 1] Assume further that the alphabets $\mathcal{X}$ and $\mathcal{U}$ are finite. Then if $R > I(X;U)$ there exists a $\gamma_1 > 0$ and $\gamma_2 > 0$ such that for sufficiently large n*

$$\Pr\{\|\bar{P}_{X^n} - P_X^{\otimes n}\|_{TV} > e^{-\gamma_1 n}\} \leq e^{-e^{\gamma_2 n}}. \tag{2.16}$$

Lemma 2.5 says that we have a vanishing soft-covering with high probability, i.e., we have

$$\|\bar{P}_{X^n} - P_X^{\otimes n}\|_{TV} \to 0, \text{in probability, as } n \to \infty. \tag{2.17}$$

If we assume that $X^n \sim P_X^{\otimes n}$, $U^n = \varnothing$ and set $P_{\hat{X}|X} \leftarrow P_{U|X}$ in Lemma 2.3, then (with high probability on the space of $\boldsymbol{u}^n$) for any $\gamma > 0$ and sufficiently large n the following two conditions are satisfied

$$\Pr\{(X^n, u^n(i)) \in \mathcal{T}_\epsilon^n\} > 1 - \gamma, \text{ and } \|\bar{P}_{X^n} - P_X^{\otimes n}\|_{TV} < \gamma. \tag{2.18}$$

In other words a good soft-covering codebook can also be used for a strongly typical covering.

2.2 The Wyner-Ahlswede-Körner setting

The Wyner-Ahlswede-Körner (WAK) setting [AK75], [Wyn75] deals with the task of recovering the correct source information with helping from another sender in the system. An illustration of the WAK setting is given in Fig. 2.1.

Formally, a code for the WAK setting consists of two encoding mappings (ϕ_{1n}, ϕ_{2n}) and a decoding mapping ψ_n which are defined as

$$\phi_{1n}\colon \mathcal{X}^n \to \mathcal{M}_1,\ \phi_{2n}\colon \mathcal{Y}^n \to \mathcal{M}_2, \text{ and } \psi_n\colon \mathcal{M}_1 \times \mathcal{M}_2 \to \mathcal{Y}^n. \tag{2.19}$$

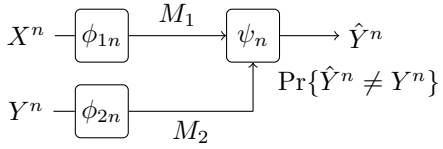


Figure 2.1: The WAK (one helper) model.

Note that both $\mathcal{X}$ and $\mathcal{Y}$ are finite in this setting. For a fixed $\epsilon \in [0,1)$ we say that a rate pair (R_1, R_2) is ϵ-achievable if there exists a WAK code $(\phi_{1n}, \phi_{2n}, \psi_n)$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_i| \leq R_i, \ i = 1, 2, \text{ and } \limsup_{n\to\infty} \Pr\{Y^n \neq \hat{Y}^n\} \leq \epsilon. \tag{2.20}$$

The WAK setup can be seen as a generalization of the Slepian-Wolf setting where $\phi_{1n} = \mathrm{id}$ and $\mathcal{M}_1 = \mathcal{X}^n$. We cite the result in the following.

Theorem 2.1 *[SW73] Given discrete memoryless source $\{(X_i, Y_i)\}_{i=1}^{\infty}$. Let $0 \leq \epsilon < 1$ and $\gamma > 0$ be arbitrary but given. Then for all sufficiently large n, there exists encoding functions $f_n \colon \mathcal{X}^n \to \mathcal{M}_1$ and $g_n \colon \mathcal{Y}^n \to \mathcal{M}_2$ and a decoding function $h_n \colon \mathcal{M}_1 \times \mathcal{M}_2 \to \mathcal{Y}^n$ such that $\Pr\{h_n(f_n(X^n), g_n(Y^n)) \neq Y^n\} \leq \epsilon$ if*

$$\frac{1}{n} \log |\mathcal{M}_1| \geq H(X) + \gamma, \text{ and } \frac{1}{n} \log |\mathcal{M}_2| \geq H(Y|X) + \gamma. \tag{2.21}$$

A simplified achievability proof for the Slepian-Wolf setting is provided by Cover [Cov75] which we explain briefly in the following. We pre-select an alphabet $\mathcal{M}_2$. Then each sequence $y^n \in \mathcal{Y}^n$ is assigned independently and uniformly at random to an index $m_2 \in \mathcal{M}_2$. Upon observing y^n the second encoder sends the corresponding bin index m_2. Then based on the given information x^n and m_2, the decoder look for a unique sequence $\hat{y}^n$ which has the bin index m_2 such that $(x^n, \hat{y}^n)$ is jointly typical.

The conditionally typical lemma, which is a standard, contemporary tool to study source coding with side information settings, was not available at that time when the problem was put forth. To show the achievability of a rate pair (R_1, R_2), Wyner [Wyn75] used a piggyback code inside a Slepian-Wolf code. We explain the idea in the following. Assume that we use a test channel $P_{U|X}$, where U takes values in $\mathcal{U}$, to encode the information from $\mathcal{X}^n$. Assume further that there exists a sequence of functions $\tilde{\psi}_n$ such that if $(Y^n, U^n) \sim P_{YU}^{\otimes n}$ then

$$\delta_n = \mathbb{E}[\tilde{\psi}_n(Y^n, U^n)] \to 0, \text{ as } n \to \infty. \tag{2.22}$$

A toy example of $\tilde{\psi}_n$ would be $\mathbf{1}\{(y^n, u^n) \notin \mathcal{T}_\epsilon^n(P_{YU})\}$. Wyner used the following function

$$\tilde{\psi}_n(y^n, u^n) = d_H(y^n, \psi_n^{\mathrm{SW}}(\phi_n^{\mathrm{SW}}(y^n), u^n)), \tag{2.23}$$

where $d_H(\cdot,\cdot)$ is the normalized Hamming distance. ϕ_n^{SW} and ψ_n^{SW} are encoding and decoding functions of the Slepian-Wolf setting with the source (Y^n, U^n) which exist if $R_2 = H(Y|U) + \gamma$ where $\gamma > 0$ is arbitrary. In the WAK setting they correspond to ϕ_{2n} and ψ_n, respectively. Let $f_n\colon \mathcal{X}^n \to \mathcal{U}^n$ be an arbitrary map. Then, after some derivation steps it can be shown that

$$\mathbb{E}[\tilde{\psi}_n(Y^n, f_n(X^n))] \le \Pr\{(X^n, f_n(X^n)) \notin \mathcal{S}_n\} + \delta_n^{1/2} \tag{2.24}$$

as $0 \le \tilde{\psi}_n \le 1$ where

$$\mathcal{S}_n = \{(x^n, u^n) \mid \mathbb{E}[\tilde{\psi}_n(Y^n, u^n)|X^n = x^n] \le \delta_n^{1/2}\}. \tag{2.25}$$

If we use our toy function $\tilde{\psi}_n$ then $\mathcal{S}_n$ is the set of (x^n, u^n) such that the (conditional given x^n) probability that Y^n is not jointly typical with u^n is small. By Markov's inequality $\Pr\{(\bar{X}^n, \bar{U}^n) \notin \mathcal{S}_n\} \le \delta_n^{1/2}$ where $(\bar{X}^n, \bar{U}^n) \sim P_{XU}^{\otimes n}$.

Wyner noticed that in Lemma 2.1 any measurable function $d(\cdot,\cdot)$, which is not necessary a distortion measure, can be used. Hence Lemma 2.1 can be used to bound the first term in (2.24) by defining $d(x^n, u^n) = \mathbf{1}\{(x^n, u^n) \notin \mathcal{S}_n\}$ and setting the threshold $D = 0$. Note further that f_n is the corresponding encoding rule when $\boldsymbol{u}^n$ is fixed. Therefore, we get

$$\begin{aligned}\mathbb{E}[\tilde{\psi}_n(Y^n, f_n(X^n))] \le{}& \Pr\{\iota(\bar{X}^n; \bar{U}^n) > nR\} + \Pr\{(\bar{X}^n; \bar{U}^n) \notin \mathcal{S}_n\} \\ &+ \exp(-Me^{-nR}) + \delta_n^{1/2}.\end{aligned} \tag{2.26}$$

By choosing $R_1 = I(X;U) + \gamma$ and $M = e^{n(R+\gamma)}$, we can drive $\mathbb{E}[\tilde{\psi}_n(Y^n, f_n(X^n))]$ to 0. The existence of f_n or ϕ_{1n} hence follows. The expression (2.26) says that although $(Y^n, f_n(X^n))$ is not iid, there exists a piggyback mapping $f_n\colon \mathcal{X}^n \to \mathcal{M} \cong \{u^n(i)\}_{i=1}^{|\mathcal{M}|} \subset \mathcal{U}^n$ such that the overall effect is similar to (2.22), i.e., as if $(Y^n, f_n(X^n))$ were iid distributed.

Ahlswede and Körner approached this setting with the same idea but with a different method [AK75]. Let $f_n\colon \mathcal{X}^n \to \mathcal{M}$ be an arbitrary function. If we apply f_n to consecutive blocks of length n: $X_1^n, X_{n+1}^{2n}, ...$, then the output indicies $\{f_n(X_{kn+1}^{(k+1)n})\}_{k=0}^{\infty}$ are mutually independent and identically distributed. Hence we can view the sequence of pairs

$$\{Y_{kn+1}^{(k+1)n}, f_n(X_{kn+1}^{(k+1)n})\}_{k=0}^{\infty} \tag{2.27}$$

as a new iid *super-source*. Applying the Slepian-Wolf theorem for this super-source we can see that the decoding error goes to zero if

$$R_1 \ge \frac{1}{n} H(f_n(X^n)),\ R_2 \ge \frac{1}{n} H(Y^n|f_n(X^n)), \tag{2.28}$$

where the normalizing factor $1/n$ is because each element of the super-source has length n. Since f_n and n are arbitrary, the following region is achievable

$$\left\{(R_1, R_2) \middle| R_1 \ge \frac{1}{n} H(f_n(X^n)),\ R_2 \ge \frac{1}{n} H(Y^n|f_n(X^n)),\ f_n\colon \mathcal{X}^n \to \mathbb{N}\right\}. \tag{2.29}$$

Fano's inequality can be used to show that the given region is indeed the optimal region. The key technical step in Ahlswede and Körner's work is showing that (2.29) equals to the following region

$$\big\{(R_1, R_2) \mid R_1 \geq I(X;U),\ R_2 \geq H(Y|U),\ U - X - Y,\ |\mathcal{U}| \leq \mathcal{X} + 1\big\}, \tag{2.30}$$

which is called the image-characterization approach. The reader is referred to [AK75] for a detailed exposition.
The contemporary approach simplifies the proof by combining the following lemma, the covering lemma, and the random binning argument by Cover, cf. [CT12, Section 15.4].

Lemma 2.6 (Conditional Typicality Lemma) *[EK11, p. 27] Given* $(X^n, Y^n) \sim P_{XY}^{\otimes n}$ *and* $x^n \in \mathcal{T}_{\epsilon_1}^n(P_X)$. *Then for any* $\epsilon_2 \geq \epsilon_1$ *we have*

$$\Pr\{(x^n, Y^n) \in \mathcal{T}_{\epsilon_2}^n(P_{XY}) | X^n = x^n\} \to 1, \text{ as } n \to \infty. \tag{2.31}$$

It should be mentioned that the conditionally typical lemma is not available in the weak typicality approach. We also digress a bit from the context of the WAK problem and discuss the applicability of Wyner's approach when $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{U}$ have continuous supports. Assume that[2.1] $0 \leq \tilde{\psi}_n \leq 1$, we observe that the relation (2.26) also holds in this case under the same choice of $d(x^n, u^n)$. Additionally, the right-hand side of (2.26) goes to zero if there exists a sequence of functions $\tilde{\psi}_n$ such that (2.22) is satisfied and $\mathbb{E}[\iota(X;U)] < \infty$ holds.
Finally, the strong converse result for the WAK problem is proved based on the *blowing-up* lemma developed by Ahlswede, Gacs and Körner in [AGK76].

2.3 Hypothesis testing

In many inference application one would like to distinguish between two hypotheses H_0 and H_1 under which the data sample x is generated from the distributions P_{H_0} and P_{H_1}, respectively. We define an acceptance region as a subset $\mathcal{A}$ of $\mathcal{X}$ such that when $x \in \mathcal{A}$ we declare that H_0 is true. The false alarm (error of type I) and the miss detection (error of type II) probabilities are given by

$$\alpha = P_{H_0}(\mathcal{A}^c),\ \beta = P_{H_1}(\mathcal{A}). \tag{2.32}$$

The following important lemma which comes from the data-processing inequality for the divergence is often used to prove the weak converse result.

Lemma 2.7 $d(1-\alpha\|\beta) \leq D(P\|Q)$ *where* $d(1-\alpha\|\beta) = (1-\alpha)\log\frac{1-\alpha}{\beta} + \alpha\log\frac{\alpha}{1-\beta}$ *is the binary divergence.*

We state another fundamental result in the following[2.2].

[2.1] This assumption can be relaxed to $\tilde{\psi}_n \in [0, a]$.

[2.2] Both Lemma 2.7 and 2.8 hold when the test is random, i.e., each sample x is assigned to H_0 with probability p_x and H_1 with probability $1 - p_x$.

Lemma 2.8 *[Han03, Lemma 4.1.2] For any $\gamma > 0$ and any $\mathcal{A}$ the following inequality holds*

$$\alpha + \gamma\beta \geq \Pr\left\{ \frac{dP_{H_0}}{dP_{H_1}}(X) \leq \gamma \right\}, \tag{2.33}$$

where $X \sim P_{H_0}$.

Although Lemma 2.8 is not difficult to be proven, recalling it is another story. Perhaps, the following interpretation helps to memorize this inequality:

> *false alarm probability + a scaled (by γ) miss detection probability $\geq$ the false alarm probability of the likelihood ratio test with threshold γ.*

Several approaches can be used to study the hypothesis testing problem such as the Neyman-Pearson and the Bayesian frameworks. In the Bayesian approach one, for example, aims to minimize the total probability of errors, while in the Neyman-Pearson approach, the goal is to minimize the miss detection probability given a false alarm level. We mainly focus on the Neyman-Pearson approach in this thesis. As the result of Neyman-Pearson lemma, an optimal acceptance region for a given false alarm threshold α_0 can be given in the following form, assuming that $P_{H_0} \ll P_{H_1}$,

$$\mathcal{A} = \left\{ x \in \mathcal{X} \middle| \frac{dP_{H_0}}{dP_{H_1}}(x) > t \right\}, \tag{2.34}$$

for some threshold t such that $P_{H_0}(\mathcal{A}^c) = \alpha_0$.

As a standard scenario, it is often assumed that under hypothesis H_0 the observation x^n is generated iid from a distribution P_X while under hypothesis H_1, x^n is also generated iid but from a distribution Q_X. In this case, the acceptance region is also denoted by $\mathcal{A}_n$ which induces the corresponding error probabilities denoted by α_n and β_n. Since we also assume that the block length n tends to ∞, it would be natural to inspect the convergence to 0 rate of the miss detection probability for a given false alarm threshold in the Neyman-Pearson framework.
We say that E is an ϵ-achievable exponent of type II of error if there exists a decision region sequence $\mathcal{A}_n$ such that

$$\limsup_{n\to\infty} \alpha_n \leq \epsilon, \ \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_n} \geq E. \tag{2.35}$$

Let $E_\epsilon^\star = \sup\{E \mid E \text{ is an } \epsilon\text{-achievable exponent of type II of error}\}$ be the maximum ϵ-achievable error exponent of type II.
The following result summarizes the asymptotic behavior of the miss detection probability when the false alarm probability is constrained to be less than an ϵ.

Theorem 2.2 (Stein's lemma) *[Han03, Corollary 4.2.1] Assume that $D(P_X\|Q_X)$ is finite then for all $\epsilon \in [0, 1)$ we have*

$$E_\epsilon^\star = D(P_X\|Q_X). \tag{2.36}$$

The proof of Theorem 2.2 is quite short, we discuss it in somewhat detail here. Since we are considering the Neyman-Pearson framework, it is intuitively reasonable to declare an acceptance region as

$$\mathcal{A}_n = \left\{ x^n \middle| \frac{dP_X^{\otimes n}}{dQ_X^{\otimes n}}(x^n) > t \right\}, \text{ for some } t. \tag{2.37}$$

The law of large numbers indicates that as $n \to \infty$, $\frac{1}{n} \log \frac{dP_X^{\otimes n}}{dQ_X^{\otimes n}}(X^n) \to D(P_X \| Q_X)$ when $X^n \sim P_X^{\otimes n}$ with probability 1 (or almost surely). Therefore if we select t such that $\log t = n(D(P_X \| Q_X) - \gamma)$ where $\gamma > 0$ is arbitrary, then the false alarm probability can be forced to 0. The miss detection can be calculated based on the *change of measure* argument as follows.

$$\begin{aligned} Q_X^{\otimes n}(\mathcal{A}_n) &= \int_{\mathcal{A}_n} dQ_X^{\otimes n}(x^n) \overset{(a)}{\leq} \int_{\mathcal{A}_n} \frac{1}{t} \frac{dP_X^{\otimes n}}{dQ_X^{\otimes n}}(x^n) dQ^{\otimes n}(x^n) \\ &= e^{-n(D(P_X \| Q_X) - \gamma)} \int_{\mathcal{A}_n} dP_X^{\otimes n}(x^n) \leq e^{-n(D(P_X \| Q_X) - \gamma)} \end{aligned} \tag{2.38}$$

where (a) is valid due to the definition of $\mathcal{A}_n$. The argument works by changing the underlying measure from $Q^{\otimes n}$ to $P^{\otimes n}$ based on the definition of the event $\mathcal{A}_n$. Hence we can conclude that $E_\epsilon^\star \geq D(P_X \| Q_X)$.

Lemma 2.7 can be used to show that the converse for $\epsilon = 0$ holds, i.e. $E^\star = D(P_X \| Q_X)$. To show the converse direction for $\epsilon \in (0, 1)$, or the strong converse, we use Lemma 2.8. Assume that for a given pair (ϵ, E), there exists a sequence of decision regions $\{\mathcal{A}_n\}$ such that all the conditions in (2.35) are satisfied. Then setting $\gamma = e^{n(D(P_X \| Q_X) + \xi)}$ where $\xi > 0$ is arbitrary, we obtain by Lemma 2.8

$$\begin{aligned} \alpha_n + e^{n(D(P_X \| Q_X) + \xi)} \beta_n &\geq \Pr\left\{ \frac{1}{n} \log \frac{dP_X^{\otimes n}}{dQ_X^{\otimes n}}(X^n) \leq D(P_X \| Q_X) + \xi \right\}, \\ &\qquad X^n \sim P_X^{\otimes n}, \ \forall n, \\ \Rightarrow \alpha_n + e^{n(D(P_X \| Q_X) + \xi)} e^{-n(E - \xi)} &\geq \Pr\left\{ \frac{1}{n} \log \frac{dP_X^{\otimes n}}{dQ_X^{\otimes n}}(X^n) \leq D(P_X \| Q_X) + \xi \right\}, \\ &\qquad \forall n \geq n_0(\gamma). \end{aligned} \tag{2.39}$$

We can see that the right-hand side goes to 1 due to the law of large numbers. Since $\alpha_n < 1$ for all sufficiently large n, we must have $D(P_X \| Q_X) + \xi \geq E - \xi$ for all ξ. Hence $D(P_X \| Q_X) \geq E$ which implies that $D(P_X \| Q_X) \geq E_\epsilon^\star$ for all ϵ.

In an attempt to connect the fields of information theory and statistics, Ahlswede and Csiszár considered a distributed hypothesis testing problem with communication constraint in [AC86]. We consider first the testing against independence scenario. The hypotheses are $H_0 \colon P_{XY}^{\otimes n}$ and $H_1 \colon P_X^{\otimes n} \times P_Y^{\otimes n}$ where P_{XY} is a probability measure on a finite space $\mathcal{X} \times \mathcal{Y}$. Due to the communication limits, the decision center has only access to (y^n, m_1) where m_1 is a compressed version of x^n

at some rate R_c. The designer is allowed to select a compression mapping ϕ_n and a decision mapping ψ_n which are defined as follows:

$$\phi_n \colon \mathcal{X}^n \to \mathcal{M}_1, \text{ and } \psi_n \colon \mathcal{Y}^n \times \mathcal{M}_1 \to \{0,1\}. \tag{2.40}$$

An illustration of this setting is given in Fig. 2.2.

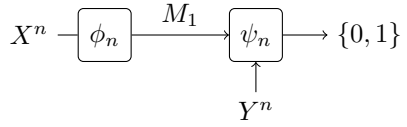


Figure 2.2: The distributed hypothesis against independence model.

The decision region is given accordingly as

$$\mathcal{A}_n = \{(y^n, \phi_n(x^n)) \mid \psi_n(y^n, \phi_n(x^n)) = 0\}. \tag{2.41}$$

The probabilities of error of type I and II, α_n and β_n, can be calculated accordingly. Similarly, Ahlswede and Csiszár considered the Neyman-Pearson framework and studied the convergence rate of the miss detection probability. We say that a rate-exponent pair (R_c, E) is ϵ-achievable if there exists a testing scheme (ϕ_n, ψ_n) such that

$$R_c \geq \limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_1|, \ \limsup_{n\to\infty} \alpha_n \leq \epsilon, \text{ and } E \leq \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_n}. \tag{2.42}$$

The achievability can be shown as follows. Note that for any mapping $f_n \colon \mathcal{X}^n \to \mathbb{N}$ we can form a super-source as in (2.27). Then we can apply Stein's lemma to the super-source to show that the following set of rate-exponent pairs (R_c, E) is achievable

$$\left\{ (R_c, E) \middle| R_c \geq \frac{1}{n} H(f_n(X^n)), \right. \\ \left. E \geq \frac{1}{n} D(P_{Y^n f_n(X^n)} \| P_{Y^n} \times P_{f_n(X^n)}), \ f_n \colon \mathcal{X}^n \to \mathbb{N} \right\}. \tag{2.43}$$

As a consequence of the above entropy characterization, the maximum achievable error exponent is given as

$$\theta(R_c) = \max\{I(Y;U) \mid U - X - Y, \ |\mathcal{U}| \leq |\mathcal{X}| + 1, I(X;U) \leq R_c\}. \tag{2.44}$$

$\theta(R_c)$ is a continuous, monotonically increasing and concave function of R_c, cf. [AC86; WW75]. We observe the similarity in the characterizations (2.29) and (2.43). It can also be shown that (2.43) is indeed optimal. Since both are the optimal achievable regions, it can be inferred that if a sequence of compression

mappings $\phi_n \colon \mathcal{X}^n \to \mathbb{N}$ achieves the optimal characterization, i.e., there exists a corresponding sequence of mappings, for one problem then it also attains the optimality for the other problem, cf. also [TC08]. The achievability proof can also be simplified by using the covering lemma and the conditionally typical lemma.
The strong converse, which also holds for a more general setting in which the distribution under H_1, Q_{XY}, is arbitrary such that $Q_{XY}(x,y) > 0$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, is shown by the blowing up lemma in [AC86].

2.4 Identification problem

As mentioned in Chapter 1 the identification problem [WKL03] consists of two phases: the *enrollment* and the *identification* phases. In the enrollment phase, we map the data sequences from M users into a database. For simplicity we discuss the discrete setting exclusively. Assume that each user sequence $x^n(i)$ is generated iid from the distribution P_X for all $i \in [1 : M]$. Then the enrollment process is described by the following mapping

$$\phi_n \colon \mathcal{X}^n \to \mathcal{M}_1. \tag{2.45}$$

The uncompressed case corresponds to $\phi_n = \mathrm{id}$ and $\mathcal{M}_1 = \mathcal{X}^n$. In the identification phase a user w is chosen uniformly at random from M users. An observation sequence y^n which is an output of the channel $P_{Y|X}^{\otimes n}$ with an input sequence $x^n(w)$ is provided to the system. Let $\mathbf{x}^n$ and $\phi_n(\mathbf{x}^n)$ be abbreviations of $(x^n(m))_{m=1}^M$ and $(\phi_n(x^n(m)))_{m=1}^M$, respectively. The joint distribution of the whole system can be written as

$$P_{Y^n \mathbf{X}^n W}(y^n, \mathbf{x}^n, w) = \frac{1}{M} P_{Y|X}^{\otimes n}(y^n | x^n(w)) \prod_{i=1}^{M} P_X^{\otimes n}(x^n(i)). \tag{2.46}$$

An identification mapping ψ_n is designed to find the true user w based on the information inside the database $\phi_n(\mathbf{x}^n)$ and the observation sequence y^n which is given as

$$\psi_n \colon \mathcal{M}_1^M \times \mathcal{Y}^n \to \mathcal{W} \cup \{e\}, \tag{2.47}$$

where $\mathcal{W} = [1 : M]$.

We say that a rate pair (R_c, R) is achievable if there exists a pair of mappings (ϕ_n, ψ_n) such that

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_1| \le R_c, \ \liminf_{n\to\infty} \frac{1}{n} \log M \ge R, \text{ and } \lim_{n\to\infty} \Pr\{W \ne \hat{W}\} = 0. \tag{2.48}$$

The closure of the set of all achievable rate pairs is the rate region. Using the strongly typical approach, the covering lemma, the conditionally typical lemma and the analysis of the channel decoding problem, we can characterize the achievable rate region as [Tun09, Theorem 1]

$$\big\{(R_c, R) \mid R_c \ge I(X;U),\ R \le I(Y;U),\ U - X - Y,\ |\mathcal{U}| \le |\mathcal{X}| + 1\big\}. \tag{2.49}$$

The identification capacity, which corresponds to the maximum number of users that the system can support, is given by $C = I(X;Y)$. By looking at (2.49), (2.30), we see that there is some relation between the identification problem and the WAK setting.

2.5 Formalities

In this part, we discuss some formal results which are needed in the study of identification systems where the underlying joint distribution is Gaussian. We first cite some important results in probability theory. Using these tools we justify the definition and necessary properties of information quantities to ensure that everything is just as planned.
Assume that ν and μ are two measures on a measurable space $(\Omega, \mathcal{A})$. We say ν is *absolutely continuous* w.r.t. μ (written as $\nu \ll \mu$) if $\nu(A) = 0$ whenever $\mu(A) = 0$ for $A \in \mathcal{A}$.

Theorem 2.3 (Radon-Nikodym) *[Dur10, Section 5.1] Let μ and ν be σ-finite measures on a measurable space $(\Omega, \mathcal{A})$. Then $\nu \ll \mu$ iff there exists a Borel measurable function $f\colon (\Omega, \mathcal{A}) \to (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ such that*

$$\nu(A) = \int_A f d\mu, \ \forall A \in \mathcal{A}. \tag{2.50}$$

Examples of σ-finite measures are probability measures on $\mathbb{R}$, the Lebesgue measure λ on $\mathbb{R}$, the counting measure μ_c on any countable spaces. We denote the underlying probability space in this work by $(\Omega, \mathcal{A}, \mathbb{P})$.

Definition 2.5 (Conditional expectation) [Dur10, Section 5.1]
We say that Y is a version of the conditional expectation of a integrable random variable X given a sub-σ-algebra $\mathcal{F} \subset \mathcal{A}$, where the latter is denoted by $\mathbb{E}[X|\mathcal{F}]$, if

- Y is $\mathcal{F}$ measurable.
- For any $A \in \mathcal{F}$ we have $\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[Y\mathbf{1}_A]$.

We also define the conditional probability of $H \in \mathcal{A}$ given $\mathcal{F}$ as $\mathbb{P}(H|\mathcal{F}) = \mathbb{E}[\mathbf{1}_A|\mathcal{F}]$.

Definition 2.6 (Regular conditional distribution) [Dur10, Section 5.1.3]
Let $X\colon (\Omega, \mathcal{A}) \to (S, \mathcal{S}_X)$ be a random variable. We say a mapping $\kappa\colon \Omega \times \mathcal{S}_X \to \mathbb{R}_+$ is a regular conditional distribution of X given $\mathcal{F} \subset \mathcal{A}$ if

- For each $A \in \mathcal{S}_X$, $\kappa(\cdot, A)$ is a version of $\mathbb{P}(X \in A|\mathcal{F})$.
- For $\mathbb{P}$- almost all ω we have $\kappa(\omega, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{S}_X)$.

A *complete, separable* metric space, for example $(\mathbb{R}^n, \|\cdot\|_2)$, is a Polish space. Let E be a Polish space and $\mathcal{E}$ be the corresponding σ-algebra generated from open subsets of $\mathcal{E}$. If X is a random variable taking values in $\mathcal{E}$ then a regular conditional distribution of X given any sub σ-algebra $\mathcal{F} \subset \mathcal{A}$ exists.

An important property of conditional expectation, often called tower property, is that if $\mathbb{E}[|X|] < \infty$ then we have $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. However, in our problem we usually need to go deeper, for example to bound the differential entropy by the minimum mean square error in the following Lemma 2.12, a density function is needed. Therefore, we need to use a regular conditional distribution. The following result helps to simplify the derivation.

Theorem 2.4 (Disintegration) *[Kal06, Theorem 5.4] Fix two measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, a sub σ-algebra $\mathcal{F} \subset \mathcal{A}$, and a random variable X taking values in $\mathcal{X}$ such that $\mathbb{P}[X \in \cdot|\mathcal{F}]$ has a regular version ν. Further, consider an $\mathcal{F}$-measurable random variable Y in $\mathcal{Y}$ and a measurable function f on $\mathcal{X} \times \mathcal{Y}$ with $\mathbb{E}[|f(X,Y)|] < \infty$. Then*

$$\mathbb{E}[f(X,Y)|\mathcal{F}](\omega) = \int d\nu(\omega, x) f(x, Y(\omega)),\ \mathbb{P} - a.s. \tag{2.51}$$

If κ is a regular conditional distribution of X given Y and $\mathcal{F} = \sigma(Y)$ then a consequence of the above theorem is

$$\mathbb{E}[f(X,Y)] = \int \int f(x,y) d\kappa(y,x) dP_Y(y). \tag{2.52}$$

Before discussing information quantities such as mutual information, differential entropy in more details we make an important *convention*. In our problem we restrict our encoding, processing, reconstruction mappings to the following form

$$f\colon (A_1, \mathcal{A}_1) \to (A_2, \mathcal{A}_2) \tag{2.53}$$

where $\{(A_i, \mathcal{A}_i)\}_{i=1}^2$ are measurable spaces with $\mathcal{A}_i$ being the corresponding Borel σ-algebra. The Borel σ-algebra of $\mathbb{R}$ equipped with the Euclidean distance is $\mathcal{B}(\mathbb{R})$, while the Borel σ-algebra of a discrete set A equipped with the discrete metric is its power set 2^A. If a mapping takes multiple arguments as input or output, in which each argument's range can be either discrete or $\mathbb{R}$, then the corresponding (Borel) σ-algebra is the product of the (Borel) σ-algebra of each individual argument. This particular assumption is a consequence of the following result which says that for a countable product of separable metric spaces, the (big) Borel σ-algebra and the product of Borel σ-algebras agree

Lemma 2.9 *[Kal06, Lemma 1.2] Let S_1, S_2, ... be separable metric spaces. Then*

$$\mathcal{B}(S_1 \times S_2 \times \cdots) = \mathcal{B}(S_1) \times \mathcal{B}(S_2) \times \cdot. \tag{2.54}$$

On the product space the finite product of measures has also the associate property [Tao15]. Working with complete measurable space like $(\mathbb{R}, \mathcal{L}(\mathbb{R}))$ where

$\mathcal{L}(\mathbb{R})$ is the Lebesgue σ-algebra could be *out of hand*, since even the complete measure λ_n on $\mathcal{L}(\mathbb{R})$ is not equal to $\lambda^{\otimes n}$ and some properties hold only *almost everywhere*, cf. Fubini's theorem. A further explanation is beyond the scope of the recapitulation of this thesis.

We present in the following the calculation of information quantities needed in Chapter 4. For a more comprehensive consideration about general properties of these information quantities, the reader is referred to for examples [Gra13; PW17]. We say Y is a continuous random variable if it is $(\Omega, \mathcal{A})/(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ measurable and the distribution P_Y has a density function w.r.t. $\lambda^{\otimes n}$. For simplicity, we carry out the derivation with $n = 1$ only. An extension to any finite n can be inferred directly.

Let (X, Y) be a pair of random variables taking values in $\mathbb{R}^2$. Assume that $P_{XY} \ll \lambda^{\otimes 2}$. Let $f_{XY} = dP_{XY}/d\lambda^{\otimes 2}$ be the corresponding density function. We can define f_X, f_Y and $f_{X|Y} = f_{XY}/f_Y$ whenever $f_Y(y) \neq 0$. Then, the differential/conditional differential entropies are defined as

$$h(X) = -\int \log f_X dP_X, \quad h(Y) = -\int \log f_Y dP_Y,$$
$$h(X|Y) = -\int \log f_{X|Y} dP_{XY}, \tag{2.55}$$

whenever the corresponding integrals exist. It is more tricky to define conditional entropies when P_{XY} is not homogeneous, for example when it is a distribution of a mixture of discrete and continuous random variables.

Lemma 2.10 *Let P_{XY} be the joint distribution of a finite alphabet random variable X and a continuous random variable Y then:*

1. *$P_{XY} \ll P_X \times P_Y \ll \mu_c \times P_Y$ and $P_{XY} \ll P_X \times P_Y \ll P_X \times \lambda$*
2. *Let $P_{X|Y}$ be a regular conditional distribution of X given Y then $I(X;Y) = \mathbb{E}_{P_{XY}}[\log \frac{P_{X|Y}}{P_X}] \leq \log |\mathcal{X}|$. Hence $H(X|Y) = \mathbb{E}_{P_{XY}}[-\log P_{X|Y}]$ is finite and $I(X;Y) = H(X) - H(X|Y)$ holds.*
3. *Let $f_{Y|X}$ be the density of the regular conditional distribution*

$$\mathbb{P}(Y \in B | X = x) = \frac{P_{YX}(Y \in B, X = x)}{P_X(x)}. \tag{2.56}$$

Then $I(X;Y) = \mathbb{E}_{P_{XY}}[\log \frac{f_{Y|X}}{f_Y}]$. Hence if $h(Y)$ exists then

$$h(Y|X) = \mathbb{E}_{P_{XY}}[-\log f_{Y|X}] \tag{2.57}$$

also exists and hence can also be calculated as

$$h(Y|X) = \mathbb{E}_{P_X}[\mathbb{E}_{P_{Y|X}}[-\log f_{Y|X}]] = \mathbb{E}_{P_X}[h(Y|X = x)].$$

The assumption that Y is continuous can be lifted in 2. and the first part of 1..

Proof. 1. It is clear that $P_X \times P_Y \ll P_X \times \lambda$ and $P_X \times P_Y \ll \mu_c \times P_Y$. We show now that $P_{XY} \ll P_X \times P_Y$ is valid. Assume that $P_X \times P_Y(A) = 0$ for some $A \in 2^{|\mathcal{X}|} \times \mathcal{B}(\mathbb{R})$. Then as A can be decomposed as $\bigcup_x \{x\} \times A_x$. Then $P_X(x) \times P_Y(A_x) = 0$. Hence for such x either $P_Y(A_x) = 0$ or $P_X(x) = 0$ holds which implies that $P_{XY}(\{x\} \times A_x) = 0$. $P_{XY}(A) = 0$ then follows.

2. For any $A \in 2^{|\mathcal{X}|} \times \mathcal{B}(\mathbb{R})$ we have by Fubini's theorem

$$\begin{aligned}\int_A P_{X|Y}(x|y)d(\mu_c \times P_Y) &= \int_x \int_{A_x} P_{X|Y}(x|y)dP_Y(y)d\mu_c(x) \\ &\overset{(a)}{=} \int_x P_{XY}(\{x\} \times A_x)d\mu_c(x) = P_{XY}(A), \quad (2.58)\end{aligned}$$

where (a) holds since $P_{X|Y}$ is a regular conditional distribution. The derivation implies that $P_{X|Y} = dP_{XY}/d(\mu_c \times P_Y)$. Then we have

$$\begin{aligned}I(X;Y) = \mathbb{E}_{P_{XY}}[\log \frac{dP_{XY}}{d(P_X \times P_Y)}] &\overset{(b)}{=} \mathbb{E}_{P_{XY}}[\log \frac{P_{X|Y}}{P_X}] \\ &\overset{(c)}{\leq} \mathbb{E}_{P_{XY}}[\log \frac{1}{P_X}] \leq \log |\mathcal{X}|. \quad (2.59)\end{aligned}$$

(b) is valid since $\frac{dP_{XY}}{d(P_X \times P_Y)} = \frac{dP_{XY}}{d(\mu_c \times P_Y)} / \frac{d(P_X \times P_Y)}{d(\mu_c \times P_Y)} = \frac{P_{X|Y}}{P_X}$. (c) holds since for each y, $P_{X|Y}(\cdot|y)$ is a distribution on $\mathcal{X}$ hence $0 \leq P_{X|Y}(x|y) \leq 1$.

3. That $\mathbb{P}(Y \in B|X = x)$ is a regular conditional distribution follows since $\mathcal{X}$ is finite. We also observe that for each x, $\mathbb{P}(Y \in \cdot|X = x) \ll P_Y \ll \lambda$. Hence $f_{Y|X}$ exists and is a measurable function on $\mathbb{R}$ for each x. Since $|\mathcal{X}|$ is finite $f_{Y|X}$ is jointly measurable function of (x, y). Similarly, we have

$$\begin{aligned}\int_A f_{Y|X}(y|x)d(\lambda \times dP_X) &= \sum_x \int_{\{x\}} \int_{A_x} f_{Y|X}(y|x)d\lambda(y)dP_X(x) \\ &= \sum_x \int_{\{x\}} \mathbb{P}(Y \in A_x|X = x)dP_X(x) = \sum_x P_{XY}(\{x\} \times A_x) = P_{XY}(A). \quad (2.60)\end{aligned}$$

Therefore, $I(X;Y) = \mathbb{E}_{P_{XY}}[\log \frac{f_{Y|X}}{f_Y}]$ follows. Since $I(X;Y)$ is finite, if $h(Y)$ is finite then $h(Y|X)$ is also finite. The last expression follows from the disintegration.

□

We extend the consideration to the conditionally differential entropy of a continuous random variable given a mixture of a continuous random variable and a

finite random variable. This quantity arises in a standard step when we relate the minimum mean square error of estimating the source given the compressed information and the continuous side information with the corresponding conditionally differential entropy term.

Lemma 2.11 *Let P_{XYZ} be a joint distribution of continuous random variables (X, Y) and a finite random variable Z. There exists a non-negative jointly measurable function $f_{X|YZ}$ such that $\nu(B|y,z) = \int_B f_{X|YZ}(x)d\lambda$ is a regular conditional distribution of X given (Y,Z). Furthermore if $h(Y)$ and $h(X|Y)$ are finite then we can define $h(X|Y,Z)$ as*

$$h(X|Y,Z) = \mathbb{E}_{P_{XYZ}}[-\log f_{X|YZ}], \tag{2.61}$$

i.e., the right-hand side is finite. It also follows that

$$h(X|Y,Z) = \mathbb{E}_{P_{YZ}}[\mathbb{E}_{P_{X|YZ}}[-\log f_{X|YZ}]], \tag{2.62}$$

and $h(X,Y|Z) = h(Y|Z) + h(X|Y,Z)$ as well as $h(X|Y,Z) \le h(X|Y)$ hold.

Proof. Let $\mathbb{P}((X,Y) \in \cdot | Z = z)$ be a regular conditional distribution defined similar as in (2.56) with the corresponding density function $f_{XY|Z}$. Then we can define

$$f_{X|YZ} = \begin{cases} f_{XY|Z}/f_{Y|Z} & \text{if } f_{Y|Z}(y|z) \neq 0 \\ f_X & \text{otherwise} \end{cases}.$$

It is clear that for each (y,z), $\nu(\cdot|y,z)$ is a probability measure on $\mathbb{R}$ and for each B, $\nu(B|y,z)$ is a measurable function of (y,z). Furthermore

$$\begin{aligned} \int_{C_z \times \{z\}} \nu(B|y,z) dP_{YZ} &\overset{(a)}{=} \int_{C_z \times \{z\}} \int_B f_{X|YZ} f_{Y|Z} d\lambda^{\otimes 2} dP_Z \\ &\overset{(b)}{=} \int_{\{z\}} \int_{B \times C_z} f_{XY|Z} d\lambda^{\otimes 2} dP_Z = P_{XYZ}(B \times C_z \times \{z\}), \end{aligned} \tag{2.63}$$

where (a) follows from the Fubini's theorem, the definition of $\nu(B|y,z)$ and $f_{Y|Z} = dP_{YZ}/d(\lambda \times P_Z)$. (b) holds since the set $\{(y,z) \mid f_{Y|Z}(y|z) = 0\}$ has probability zero. The derivation further implies that $\int_C \nu(B|y,z) dP_{YZ} = \mathbb{P}(X \in B, (Y,Z) \in C)$ since $|\mathcal{Z}|$ is finite. Hence $\nu(B|y,z)$ is a regular conditional distribution.
Next, we have

$$\begin{aligned} I(X,Y;Z) = \mathbb{E}_{P_{XYZ}}[\log \frac{f_{XY|Z}}{f_{XY}}] &= \mathbb{E}_{P_{XYZ}}[\log \frac{f_{Y|Z}}{f_Y} + \log \frac{f_{X|YZ}}{f_{X|Y}}] \\ &\overset{(c)}{=} \mathbb{E}_{P_{XYZ}}[\log \frac{f_{Y|Z}}{f_Y}] + \mathbb{E}_{P_{XYZ}}[\log \frac{f_{X|YZ}}{f_{X|Y}}]. \end{aligned} \tag{2.64}$$

Since $h(Y)$ is defined, the first term in the last expression is finite. Therefore (c) holds since the left-hand side is also finite. Since $h(X|Y)$ is finite, the right-hand

side in (2.61) is finite as well. We can identify the second term in the last expression to be the conditional mutual information $I(X;Z|Y)$. Furthermore, $I(X;Z|Y) \geq 0$ by applying the disintegration property and the non-negativity of divergence. □

The above step-by-step calculation can be repeated to show that $I(X,Y;Z) = I(X;Z) + I(Y;Z|X)$ when X and Y are finite. Next, we revisit a standard lemma which relates the differential entropy with the MMSE in estimating with side information.

Lemma 2.12 (Estimation error and differential entropy) *Let X be a continuous random variable, and Y be an arbitrary random variable such that $\mathbb{E}_{P_{XY}}[(X - \mathbb{E}[X|Y])^2] < \infty$. Assume that a jointly measurable density $f_{X|Y}$ corresponding to a regular conditional distribution $P(X \in \cdot|Y = y)$ exists and $h(X|Y)$ is defined. Then*

$$h(X|Y) \leq \frac{1}{2} \log 2\pi e \mathbb{E}_{P_{XY}}[(X - \mathbb{E}[X|Y])^2]. \tag{2.65}$$

An example for Y in the current lemma is the mixture of a continuous Y and a discrete Z in the previous lemma.

Proof.

$$\begin{aligned}
\mathbb{E}_{P_{XY}}[(X - \mathbb{E}[X|Y])^2] &\overset{(a)}{=} \int (x - \mathbb{E}[X|y])^2 dP_{X|Y}(x|y) dP_Y(y) \\
&= \int (x - \mathbb{E}[X|y])^2 f_{X|Y}(x|y) d\lambda(x) dP_Y(y) \\
&\overset{(b)}{\geq} \int \frac{1}{2\pi e} e^{2h(X|Y=y)} dP_Y(y) \overset{(c)}{\geq} \frac{1}{2\pi e} e^{2h(X|Y)}.
\end{aligned} \tag{2.66}$$

where (a) follows due to the disintegration, (b) follows since

$$\int (x - \mathbb{E}[X|y])^2 f_{X|Y}(x|y) d\lambda(x)$$

as well as $h(X|Y = y)$ are finite P_Y-almost everywhere and Gaussian distribution maximizes the conditional entropy for a given variance. Finally (c) follows due to the Jensen's inequality and $h(X|Y)$ is finite.

□

We discuss herein the formal definition and property of a Markov chain.

Definition 2.7 (Conditional independence) [Kal06, p. 86] Let $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{G}$ be sub σ-algebras of $\mathcal{A}$. We say that $\mathcal{F}_1$ and $\mathcal{F}_2$ are conditionally independent of $\mathcal{G}$ if for all $B_1 \in \mathcal{F}_1$ and $B_2 \in \mathcal{F}_2$ we have

$$\mathbb{P}[B_1 \cap B_2|\mathcal{G}] = \mathbb{P}[B_1|\mathcal{G}]\mathbb{P}[B_2|\mathcal{G}], \ \mathbb{P} - a.s. \tag{2.67}$$

It is usually not easy to verify the condition in the above definition. If $\mathcal{F}_1 = \sigma(X)$, $\mathcal{F}_2 = \sigma(Y)$ and $\mathcal{G} = \sigma(Z)$ where X and Y take values in Polish spaces, then we can *formally* check the above condition by using regular conditional distributions of X and Y given Z when these are simple to calculate. In our work, we assume that the condition is fulfilled due to *actions* of encoding and decoding processes.

Lemma 2.13 *[Kal06, Proposition 5.6] For any σ-algebra $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$, we have $\mathcal{F}$ and $\mathcal{H}$ are conditionally independent given $\mathcal{G}$ iff*

$$P[H|\mathcal{F},\mathcal{G}] = P[H|\mathcal{G}],\ \mathbb{P}-a.s,\ H \in \mathcal{H}. \tag{2.68}$$

The conditioning on the left-hand side of (2.68) should be understood as w.r.t. the joint σ-algebra $\sigma(\mathcal{F},\mathcal{G})$. We state in the following corollaries of the above lemma which might be of independent interest.

Corollary 2.1 *Assume that $\mathcal{F}$ and $\mathcal{H}$ are conditionally independent given $\mathcal{G}$. Let f be a nonnegative $\mathcal{H}$-measurable, integrable function. Then*

$$\mathbb{E}[f|\mathcal{F},\mathcal{G}] = \mathbb{E}[f|\mathcal{G}]\ ,\ \mathbb{P}-a.s.. \tag{2.69}$$

Proof. We note that for a given $H \in \mathcal{H}$, $\big\{\omega \mid P[H|\mathcal{F},\mathcal{G}](\omega) \neq P[H|\mathcal{G}](\omega)\big\} \in \sigma(\mathcal{F},\mathcal{G})$. Lemma 2.13 implies that for any positive simple function $\chi = \sum_{i=1}^{k} a_i \chi_{A_i}$ where $A_i \in \mathcal{H}$, and $a_i > 0$, $\forall i$,

$$\mathbb{E}[\chi|\mathcal{F},\mathcal{G}] = \mathbb{E}[\chi|\mathcal{G}] \text{ a.s..} \tag{2.70}$$

Since f is a nonnegative $\mathcal{H}$-measurable, integrable function, there is a sequence of increasing nonnegative, $\mathcal{H}$-measurable, simple functions χ_n that converges pointwise to f. We have by monotone convergence theorem

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}[\chi_n|\mathcal{F},\mathcal{G}] &= \mathbb{E}[f|\mathcal{F},\mathcal{G}] \text{ a.s.}, \\ \lim_{n\to\infty} \mathbb{E}[\chi_n|\mathcal{G}] &= \mathbb{E}[f|\mathcal{G}] \text{ a.s..} \end{aligned} \tag{2.71}$$

Denote $B_1 = \big\{\omega \mid \lim_{n\to\infty} \mathbb{E}[\chi_n|\mathcal{F},\mathcal{G}](\omega) \neq \mathbb{E}[f|\mathcal{F},\mathcal{G}](\omega)\big\}$,

$$B_2 = \big\{\omega \mid \lim_{n\to\infty} \mathbb{E}[\chi_n|\mathcal{G}](\omega) \neq \mathbb{E}[f|\mathcal{G}](\omega)\big\}$$

and

$$C_i = \big\{\omega \mid \mathbb{E}[\chi_i|\mathcal{F},\mathcal{G}](\omega) \neq \mathbb{E}[\chi_i|\mathcal{G}](\omega)\big\},\ i = 1,\ldots. \tag{2.72}$$

Define $B = B_1 \cup B_1 \bigcup_i C_i$. We observe that $B \in \sigma(\mathcal{F},\mathcal{G})$ and $\mathbb{P}(B) = 0$. For $\omega \in B^c$ then

$$\mathbb{E}[f|\mathcal{F},\mathcal{G}](\omega) = \lim_{n\to\infty} \mathbb{E}[\chi_n|\mathcal{F},\mathcal{G}](\omega) = \lim_{n\to\infty} \mathbb{E}[\chi_n|\mathcal{G}](\omega) = \mathbb{E}[f|\mathcal{G}](\omega). \tag{2.73}$$

Hence, the conclusion follows. □

Corollary 2.2 *Assume that the conditions of Theorem 2.4 are fulfilled. Furthermore, assume that* $\mathbb{P}[X \in \cdot|\mathcal{F},\mathcal{G}]$ *has a regular version* ν_1 *and* $\sigma(X)$ *is conditionally independent of* $\mathcal{G}$ *given* $\mathcal{F}$ *then*

$$\mathbb{E}[f(X,Y)|\mathcal{F},\mathcal{G}](\omega) = \int d\nu(\omega,x) f(x,Y(\omega)),\ \mathbb{P}-a.s. \tag{2.74}$$

Proof. By comparing ν and ν_1 via Lemma 2.13 and using Definition 2.6 we have ν is a version of $\mathbb{P}[X \in \cdot|\mathcal{F},\mathcal{G}]$. Therefore the conclusion follows from Theorem 2.4. □

Lemma 2.14 (Data processing inequality) *Assume that* X, Y *and* Z *take values in Polish spaces. Furthermore* $Y - X - Z$ *and* Z *is a finite alphabet random variable then* $I(X,Y;Z) = I(X;Z)$ *hence* $I(X;Z) \geq I(Y;Z)$. *The conclusions also hold if both* X *and* Y *are finite random variables while* Z *is arbitrary.*

Proof. Since Z is a finite alphabet random variable we have

$$\begin{aligned} I(X,Y;Z) &= \mathbb{E}_{P_{XYZ}}[\log \frac{P_{Z|XY}}{P_Z}] = \mathbb{E}_{P_{XYZ}}[\log \frac{P_{Z|XY}}{P_{Z|X}} + \log \frac{P_{Z|X}}{P_Z}] \\ &= \mathbb{E}_{P_{XYZ}}[\log \frac{P_{Z|XY}}{P_{Z|X}}] + I(X;Z), \end{aligned} \tag{2.75}$$

holds since the exception set $\{(x,y,z) \mid P_{Z|XY}(z|x,y) = 0, \text{ or } P_{Z|X}(z|x) = 0\}$ has probability zero. Now by Lemma 2.13 we have for each z, $P_{Z|XY}(z|x,y) = P_{Z|X}(z|x)$ for almost all (x,y). Hence the first conclusion follows. The second holds by exchanging X to Y in the above derivation and using the non-negativity property of divergence.

When X and Y are finite then we have $I(X,Y;Z) = I(X;Z) + \mathbb{E}_{P_{XYZ}}[\log \frac{P_{Y|XZ}}{P_{Y|X}}]$. Again using Lemma 2.13 we have for any $y \in \mathcal{Y}$, $P_{Y|XZ}(y|x,z) = P_{Y|X}(y|x)$ for almost all (x,z). Hence $I(X,Y;Z) = I(X;Z)$. □

3

Discrete Hierarchical ID

IDENTIFICATION using high-dimensional data is essential in many applications in eHealth, IoT, etc.. However, using high-dimensional observations directly puts a heavy toll on the system. We propose a pre-processing procedure, e.g. a letter-wise quantization, to reduce the search complexity.

As a motivating example for our work, consider designing a visual search app for smartphones which takes as its input a picture of an object. Due to memory, power restrictions on handheld devices and communication bandwidth restriction, the app first returns a list of similar objects obtaining from a local database. This operation can be modeled as a query to low resolution database using a low resolution image or few important features. Additionally, it has an option of exact identification and returning images of the same object by accessing to the refined information in the cloud. The option can be viewed as a refined processing step and reconstruction of the original information based on the list of objects returned in the first stage and the full captured image information. The designer's task is to make the trade-off among different constraints feasible. Practical examples of visual search apps could be Google Lens [Goo] or Amazon Flow [Ama].

Another example would be: in some areas, such as in forensics or surveillance, one would like to identify the suspects as quickly as possible and view their criminal records. In these scenarios, we can also reduce the search complexity by first providing a list of possible suspects. Then the search is refined inside the given list to provide an identity of the suspect and his/her reconstructed record. The records are stored in a second node which might be even a legal requirement, e.g. if only further details about suspicious people are stored.

Motivated by these presented examples a two-stage discrete identification problem with pre-processing to enable efficient data retrieval and reconstruction is studied in this chapter. In the enrollment phase, users' data are stored into the database in two layers. The first layer stores some representative features of the sequence as in [TKR04]. The second layer contains refinement information. This informa-

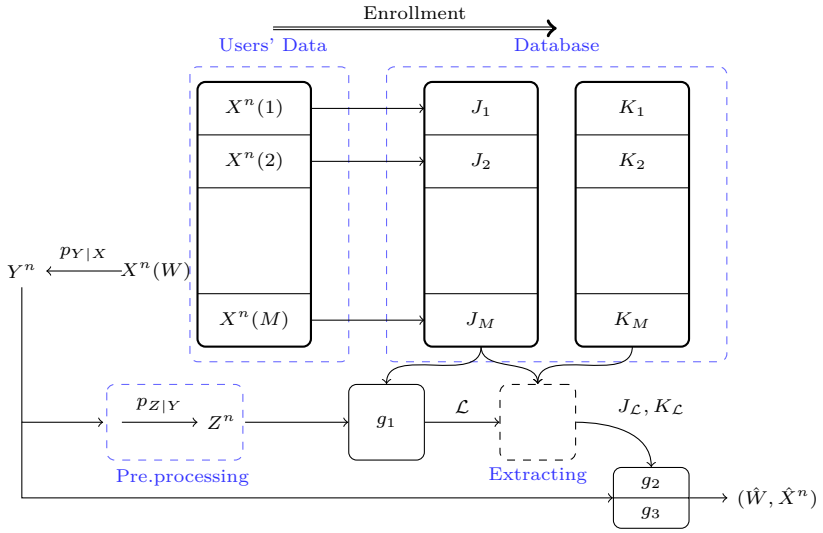


Figure 3.1: An overview of the two stage identification system. We assume that there always exists a user W which has been enrolled previously and to which the observation Y^n is the output of a memoryless channel $P_{Y|X}$ with the input $X^n(W)$. Furthermore, W is uniformly distributed over $[1 : M]$ and independent of users' data. The first and second layer information are represented by the collections $(J_i)_{i=1}^M$ and $(K_i)_{i=1}^M$, respectively.

tion layer helps to identify the user exactly and reconstruct the corresponding data sequence. This approach becomes interesting when querying information in the second layer is costly and therefore the system needs to limit the number of queries. An observation y^n is provided to the processing unit which needs to return the correct user's identity and its corresponding reconstruction sequence. To facilitate the processing time and power, the observation is first passed through a fixed channel $P_{Z|Y}$, which can be thought of as a feature extraction operation, or a fixed observation compression scheme. In the first stage, the observation is pre-processed, and the result is then used in combination with the stored first layer information in the database to output a list of compatible users to the second stage. Then, the second step uses the information of users contained in the list from both layers and the original observation sequence to return the exact user identity and a corresponding reconstruction sequence. The setting is depicted in Fig. 3.1.

3.1 Formal Problem Formulation & Result

We assume that all alphabets in this chapter are finite. The data sequence $x^n(i) \in \mathcal{X}^n$, where[3.1] $i \in \mathcal{W} = [1 : M]$ with $M = |\mathcal{W}|$, is compressed and stored hierarchically in two layers. The enrollment can be described by (possibly stochastic) mappings

$$\phi_{kn} \colon \mathcal{X}^n \to \mathcal{M}_k,\ k = 1, 2. \tag{3.1}$$

We denote database indices $\phi_{1n}(x^n(i))$ and $\phi_{2n}(x^n(i))$ as $j_i \in \mathcal{M}_1$ and $k_i \in \mathcal{M}_2$ for all $i \in \mathcal{W}$.

An observer obtains information y^n about a user in the database from the output of the memoryless channel $P_{Y|X}$ with input $x^n(w)$, where w is an instance of a uniformly distributed random variable W over the set $\mathcal{W}$, which is independent of the users' data. The observer sends y^n to a processing unit to identify w and obtains a reconstruction $\hat{x}^n$ of $x^n(w)$ within the distortion D.

In the processing unit, the observation y^n is first pre-processed. The pre-processing is modeled by a *fixed* channel $P_{Z|Y}$ to produce a noisy version z^n, which can be linked to a quantization or a feature extraction process. Then, based on z^n and the first layer database $(j_i)_{i=1}^M$, a list $\mathcal{L} \in \mathfrak{L}$ of at most $2^{n\Delta}$ possible matching indices of a given size, is produced. This action can be described by a processing mapping

$$\begin{aligned} g_1 &\colon \mathcal{Z}^n \times \mathcal{M}_1^M \to \mathfrak{L}, \\ g_1(z^n, (j_i)_{i=1}^M) &\mapsto \mathcal{L}, \end{aligned} \tag{3.2}$$

where

$$\mathfrak{L} = \left\{ \mathcal{S} \middle| \mathcal{S} \subseteq \mathcal{W},\ |\mathcal{S}| \le 2^{n\Delta} \right\} \cup \left\{ \{e\} \right\}$$

is the set of subsets of users in $\mathcal{W}$ with cardinality at most $2^{n\Delta}$ and the set $\{e\}$, which describes an error event. This means that we allow the mapping g_1 to declare an error. The *extracting* action, which takes as its inputs the index list $\mathcal{L}$ and the stored information of all users in both layers $((j_i)_{i=1}^M, (k_i)_{i=1}^M)$ to return the information of all chosen users inside the list along with the list $((j_i)_{i\in\mathcal{L}}, (k_i)_{i\in\mathcal{L}}, \mathcal{L})$, can be described formally by a projection mapping

$$\begin{aligned} &\pi \colon \mathcal{M}_1^M \times \mathcal{M}_2^M \times \mathfrak{L} \to \mathfrak{M}_{12}, \\ &\pi((j_i)_{i=1}^M, (k_i)_{i=1}^M, \mathcal{L}) \mapsto \begin{cases} ((j_i)_{i\in\mathcal{L}}, (k_i)_{i\in\mathcal{L}}, \mathcal{L}) & \text{if } \mathcal{L} \neq \{e\} \\ (1, 1, \{e\}) & \text{otherwise} \end{cases}, \end{aligned} \tag{3.3}$$

where

$$\mathfrak{M}_{12} = \bigcup_{\mathcal{L}\neq\{e\}} \left\{ ((j_i)_{i\in\mathcal{L}}, (k_i)_{i\in\mathcal{L}}, \mathcal{L}) \middle| (j_i)_{i\in\mathcal{L}} \in \mathcal{M}_1^{|\mathcal{L}|},\ (k_i)_{i\in\mathcal{L}} \in \mathcal{M}_2^{|\mathcal{L}|},\ \mathcal{L} \in \mathfrak{L} \right\}$$

[3.1] For $a \in \mathbb{Z}$ we use the shorthand notation $[1 : a]$ for the set $\{1, \ldots, a\}$.

$$\cup \{(1,1,\{e\})\}.$$

It should be clear from the definition of $\mathfrak{M}_{12}$ that the vectors $(j_i)_{i\in\mathcal{L}}$ and $(k_i)_{i\in\mathcal{L}}$ can contain repeated elements. Therefore, the inclusion of $\mathcal{L}$ at the output of π helps to pinpoint which combination of users the output information belongs to[3.2]. For brevity, elements of $\mathfrak{M}_{12}$ are denoted by $(j_{\mathcal{L}}, k_{\mathcal{L}}, \mathcal{L})$. In the second stage the processing unit returns an estimate of the index $\hat{w}$ which is the output of a deterministic processing mapping $g_2(\cdot)$ where

$$\begin{aligned} g_2 \colon \mathcal{Y}^n \times \mathfrak{M}_{12} &\to \mathcal{W} \cup \{e\}, \\ g_2(y^n, (j_{\mathcal{L}}, k_{\mathcal{L}}, \mathcal{L})) &\mapsto \hat{w}, \end{aligned} \tag{3.4}$$

i.e., g_2 can declare an error event as well. Furthermore, the processing unit needs to output a reconstruction sequence $\hat{x}^n$ of the data sequence $x^n(w)$. To describe the reconstruction processing mapping, first define a second projection mapping

$$\begin{aligned} \hat{\pi} \colon \mathfrak{M}_{12} \times \Big(\mathcal{W} \cup \{e\}\Big) &\to \hat{\mathfrak{M}}_{12} = \mathcal{M}_1 \times \mathcal{M}_2 \times \Big(\mathcal{W} \cup \{e\}\Big) \\ \hat{\pi}((j_{\mathcal{L}}, k_{\mathcal{L}}, \mathcal{L}), \hat{w}) &\mapsto \begin{cases} (j_{\hat{w}}, k_{\hat{w}}, \hat{w}) & \text{if } \hat{w} \in \mathcal{L} \neq \{e\} \\ (1,1,e) & \text{otherwise} \end{cases}. \end{aligned} \tag{3.5}$$

Similarly, we denote elements of $\hat{\mathfrak{M}}_{12}$ as $(j_{\hat{w}}, k_{\hat{w}}, \hat{w})$, then the reconstruction mapping is given by

$$\begin{aligned} g_3 \colon \mathcal{Y}^n \times \hat{\mathfrak{M}}_{12} &\to \hat{\mathcal{X}}^n \\ g_3(y^n, (j_{\hat{w}}, k_{\hat{w}}, \hat{w})) &\mapsto \hat{x}^n. \end{aligned} \tag{3.6}$$

The two projection mappings π and $\hat{\pi}$ are inherent, hence need not to be designed explicitly.

Definition 3.1 For a given pre-processing channel $P_{Z|Y}$, an *identification scheme of length n* consists of two enrollment mappings $\{\phi_{kn}\}_{k=1}^{2}$ and three processing mappings $\{g_k\}_{k=1}^{3}$.

Definition 3.2 For a given pre-processing scheme $P_{Z|Y}$, a rate-distortion tuple (R, R_1, R_2, R_L, D) is *achievable* if for every $\epsilon > 0$, there exists an identification scheme of length n such that[3.3]

$$\begin{aligned} \frac{1}{n}\log M > R - \epsilon, \quad \frac{1}{n}\log|\mathcal{M}_1| < R_1 + \epsilon \\ \frac{1}{n}\log|\mathcal{M}_2| < R_2 + \epsilon, \quad \Delta < R_L + \epsilon, \quad \Pr(W \notin \boldsymbol{\mathcal{L}}) < \epsilon, \end{aligned}$$

[3.2] The choice of $(1,1)$ as the output information when $\mathcal{L} = \{e\}$ is inconsequential.

[3.3] With abuse of notation, the distortion measure herein is already normalized compared with (2.6). Furthermore, $d \colon \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}$ is assumed to be bounded.

$$\Pr(W \neq \hat{W}) < \epsilon, \quad \mathbb{E}[d(X^n(W), \hat{X}^n)] < D + \epsilon, \tag{3.7}$$

for *all* sufficiently large n. The set of all achievable tuples is denoted by $\mathcal{R}$.

Note that given $\Pr\{W \neq \hat{W}\} < \epsilon$ in the finite alphabet case our constraint

$$\mathbb{E}[d(X^n(W), \hat{X}^n)] < D + \epsilon$$

is equivalent[3.4] to the constraint $\mathbb{E}[d(X^n(W), \hat{X}^n)|\hat{W} = W] < D' + \epsilon'$ for an appropriate parameter pair (D', ϵ'), which is considered in [TG14], since the distortion measure is bounded.

Definition 3.3 Let $\mathcal{R}^\star$ be the collection of tuples (R, R_1, R_2, R_L, D) such that there exist random variables U and V defined on finite alphabets $\mathcal{U}$ and $\mathcal{V}$ which satisfy

$$|\mathcal{U}| \leq |\mathcal{X}| + 5, |\mathcal{V}| \leq (|\mathcal{X}| + 5)(|\mathcal{X}| + 2) \tag{3.8}$$

and a deterministic reconstruction mapping $f\colon \mathcal{U} \times \mathcal{V} \times \mathcal{Y} \to \hat{\mathcal{X}}$ such that the followings expressions are fulfilled:

$$\begin{aligned}
&U - V - X - Y - Z, \\
&R_1 \geq I(X;U), && \text{(3.9a)}\\
&R_1 + R_2 \geq I(X;U) + I(X;V|U,Y), && \text{(3.9b)}\\
&R_1 + R_2 - R \geq I(X;U,V|Y), && \text{(3.9c)}\\
&R \leq \min\{R_L + I(Z;U), I(Y;U,V)\}, && \text{(3.9d)}\\
&D \geq \mathbb{E}[d(X, f(U,V,Y))]. && \text{(3.9e)}
\end{aligned}$$

The above definitions imply that both $\mathcal{R}$ and $\mathcal{R}^\star$ are closed subsets of $\mathbb{R}^5$ w.r.t. ℓ_1 metric. Furthermore, $\mathcal{R}^\star$ is not empty since it contains $(0, 0, 0, 0, d_{\max})$. We state our first result in the following theorem.

Theorem 3.1 *For a given pre-processing strategy $P_{Z|Y}$, memoryless data source P_X, and observation model $P_{Y|X}$, the rate-distortion region for our setting is given by*

$$\mathcal{R} = \mathcal{R}^\star. \tag{3.10}$$

The proof of Theorem 3.1 is given in Subsection 3.3.

Remark 3.1 For a given choice of auxiliary random variables U, V such that the distortion constraint (3.9e) is fulfilled, the first inequality (3.9a) shows the minimum compression rate for the first layer. The second inequality (3.9b) indicates the trade-off between total compression rate in both layers. One notices that the second term

[3.4] This follows from a chain of inequalities $\Pr\{W = \hat{W}\}\mathbb{E}[d(X^n(W), \hat{X}^n)|\hat{W} = W] < \mathbb{E}[d(X^n(W), \hat{X}^n)] < \mathbb{E}[d(X^n(W), \hat{X}^n)|\hat{W} = W] + d_{\max}\Pr\{W \neq \hat{W}\}$.

on the right-hand side of (3.9b) suggests the use of binning for the stored data on the second storage node. (3.9c) shows the trade-off between the total compression rate and the identification rate. Namely, the identification rate is strictly smaller than the total compression rate if the right-hand side of (3.9c) is positive. Lastly, the first term on the right-hand side of (3.9d) is the maximum identification rate resulting from the first layer information and pre-processed information. The second term in (3.9d) is the maximum identification rate when the identification process is performed jointly, i.e., with the original observation and information from both layers.

Remark 3.2 In the special case where $R_L = R$, i.e. the first processing stage returns all possible users, we notice that U can be set to a deterministic value, e.g. $U = \varnothing$, so that the rate-distortion region (3.9) reduces to the one given in [TG14, Theorem 1].

3.2 Related problems

3.2.A The identification problem

When the distortion level $D = d_{\max}$, i.e., the distortion constraint can be removed, then binning for the second layer codewords is not necessary. We obtain the following corollary.

Corollary 3.1 *For a fixed $P_{Z|Y}$, the rate region of our identification setting, i.e., $D = d_{\max}$, is given by the set of tuples (R, R_1, R_2, R_L) such that*

$$\begin{aligned} &U - V - X - Y - Z, \\ &R_1 \geq I(X;U), \quad R_1 + R_2 \geq I(X;U,V), \\ &R \leq \min\{R_L + I(Z;U), I(Y;U,V)\}, \end{aligned} \tag{3.11}$$

where U and V are random variables taking values on alphabets $\mathcal{U}$ and $\mathcal{V}$, respectively, with $|\mathcal{U}| \leq |\mathcal{X}| + 4$ and $|\mathcal{V}| \leq (|\mathcal{X}| + 4)(|\mathcal{X}| + 1)$.

The proof of Corollary 3.1 is given in Appendix 3.A.

3.2.B A two observer problem

A related problem is stated in the following. The data sequence $x^n(w)$ is observed through the channel $P_{ZY|X}$ by two Observers 1 and 2, which obtain y^n and z^n, respectively. Moreover, Observer 2 has only access to the information stored in the first layer and is interested in obtaining a list of users in the database only, for instance due to complexity or due to privilege restriction. Accordingly, the decoding mapping and the requirement for the second observer are given by

$$\mathcal{L} = g_2(z^n, (j_i)_{i=1}^M), \text{ and } \Pr(W \notin \boldsymbol{\mathcal{L}}) < \epsilon, \tag{3.12}$$

where the list size is similarly constrained as $|\mathcal{L}| \leq 2^{n\Delta}$. Observer 2 in the current setting corresponds to the first processing stage in the previous settings. In contrast, Observer 1 has access to both layers and wants to identify the user correctly, i.e., the decoding mapping and the requirement of the first observer are

$$\hat{w} = g_1(y^n, (j_i)_{i=1}^M, (k_i)_{i=1}^M), \text{ and } \Pr(W \neq \hat{W}) < \epsilon. \tag{3.13}$$

In other words, the identification processes for two observers are separated. Note that there is no reconstruction requirement in the current problem. The rate region for this problem can be described by the following proposition.

Proposition 3.1 *The optimal rate region for the stated problem is the set of tuples* (R, R_1, R_2, R_L) *such that*

$$\begin{aligned} &U - V - X - (Y, Z), \\ &R_1 \geq I(X;U), \quad R_1 + R_2 \geq I(X;U,V), \\ &R \leq \min\{R_L + I(Z;U), I(Y;U,V)\}, \end{aligned} \tag{3.14}$$

where U *and* V *are random variables taking values on finite alphabets* $\mathcal{U}$ *and* $\mathcal{V}$*, respectively, with* $|\mathcal{U}| \leq |\mathcal{X}| + 4$ *and* $|\mathcal{V}| \leq (|\mathcal{X}| + 4)(|\mathcal{X}| + 1)$.

Note that the Markov condition $X - Y - Z$ is not needed since the two identification processes work independently. This means that our original problem can be viewed as a sequential cooperation scheme between two identification processes. Note further that if the pair $(P_{Y|X}, P_{Z|X})$ in Proposition 3.1 is equal to the one in Corollary 3.1 then the regions are identical. The proof of Proposition 3.1 is given in Appendix 3.B.

3.3 Proof of Theorem 3.1

3.3.A Achievability

Fix a conditional pmf $P_{UV|X}$ where $U - V - X$ and a deterministic reconstruction mapping f such that we have

$$\mathbb{E}[d(X, f(U, V, Y))] = D. \tag{3.15}$$

Additionally, for a fixed $\epsilon > 0$, we assume that the number of enrolled users is given by $M = 2^{n\hat{R}}$ where $\hat{R} = R - \epsilon/2$ and the actual list size is $\hat{\Delta} = R_L + \epsilon/2$. Also let $\hat{R}_U = R_1 + \epsilon/2$, $\hat{R}_V = R_2 + \epsilon/2$ and $\hat{R}'_V = R'_V - \epsilon/4$ be the actual code rates. The set of suitable tuples (R, R_L, R_1, R_2, R'_V) will be specified later in (3.35). We also select an $\bar{\epsilon} > 0$ for the strongly typical set, which depends on n and $\bar{\epsilon} \to 0$ as $n \to \infty$.

Codebook generation: The codebook used in the enrollment process is *identical* for all users and constructed as follows: Generate $2^{n\hat{R}_U}$ iid codewords $u^n(j)$ where

$j \in [1 : 2^{n\hat{R}_U}]$ according to the marginal distribution P_U. For each j, we draw $2^{n(\hat{R}_V+\hat{R}'_V)}$ codewords $v^n(j,l)$ where $l \in [1 : 2^{n(\hat{R}_V+\hat{R}'_V)}]$ iid via the conditional distribution $P_{V|U}$. Each index l is parsed into a unique tuple $l = (k, k')$ where $k \in [1 : 2^{n\hat{R}_V}]$ and $k' \in [1 : 2^{n\hat{R}'_V}]$. Denote by

$$\mathfrak{B}(k) = \{l \mid l = (k, k'), \text{ for some } k'\}, \tag{3.16}$$

the k-th bin, where $k \in [1 : 2^{n\hat{R}_V}]$. Additionally, we include a *fixed* pair of codewords (u_e^n, v_e^n) corresponding to the error event. The codebook is known in the whole system.

Enrollment: For each user index $i \in \mathcal{M}$, a codeword $u^n(j_i)$ is looked for such that $(x^n(i), u^n(j_i)) \in \mathcal{T}_{\bar{\epsilon}}^n(XU)$. The chosen j_i is stored in the first layer. Next, a codeword $v^n(j_i, l_i)$ is searched for such that

$$(x^n(i), u^n(j_i), v^n(j_i, l_i)) \in \mathcal{T}_{\bar{\epsilon}}^n(XUV). \tag{3.17}$$

The chosen bin index k_i is stored in the second layer. We note that in both steps if there is more than one suitable index, we select one of them uniformly at random. If there is none, an index is selected from the corresponding set of all indices uniformly at random.

Identification and Reconstruction: The observation y^n is first passed through the memoryless pre-processing channel given by $P_{Z|Y}$ to produce z^n which is used in the first stage of our identification and reconstruction process.

First stage: We look for all indices $i \in \mathcal{M}$ such that

$$(z^n, u^n(j_i)) \in \mathcal{T}_{\bar{\epsilon}}^n(ZU), \tag{3.18}$$

and put them into the list $\mathcal{L}$. If there are more than $2^{n\hat{\Delta}}$ suitable indices then an error is declared, i.e., we output the set $\mathcal{L} = \{e\}$. In this way, our list always meets the size constraint in (3.2).

Second stage: If $\mathcal{L} = \{e\}$, then we set $\hat{w} = e$. Otherwise, if $\mathcal{L} \neq \{e\}$, we find a unique index $\hat{w}$ in $\mathcal{L}$ such that

$$(y^n, u^n(j_{\hat{w}}), v^n(j_{\hat{w}}, \tilde{l})) \in \mathcal{T}_{\bar{\epsilon}}^n(YUV) \tag{3.19}$$

for some $\tilde{l}$, where $\tilde{l} \in \mathfrak{B}(k_{\hat{w}})$, and $j_{\hat{w}}$ and $k_{\hat{w}}$ are the stored information of the $\hat{w}$-th user. If there is no such $\hat{w}$ or there is more than one then we also set $\hat{w} = e$. In the next step, if $\hat{w} \neq e$ then we search for a unique $\tilde{l} \in \mathfrak{B}(k_{\hat{w}})$ such that

$$(y^n, u^n(j_{\hat{w}}), v^n(j_{\hat{w}}, \tilde{l})) \in \mathcal{T}_{\bar{\epsilon}}^n. \tag{3.20}$$

If there is more than one $\tilde{l}$ or there is none then we set $\tilde{l} = e$. Moreover, if $\hat{w} = e$ or $\tilde{l} = e$ then we set $u^n(j_{\hat{w}}) = u_e^n$ and $v^n(j_{\hat{w}}, \tilde{l}) = v_e^n$. The reconstruction sequence is given as $\hat{x}_\tau = f(u_\tau(j_{\hat{w}}), v_\tau(j_{\hat{w}}, \tilde{l}), y_\tau)$ for all $\tau = [1 : n]$.

Note that the search for the unique pair $(\hat{w}, \tilde{l})$ could be done in a single step,

however, to mitigate the complexity of describing (g_2, g_3) we choose the separate descriptions, cf. the Gaussian setting in Chapter 4 for more information.
Analysis: Let J_i and L_i, $i \in \mathcal{M}$, be the chosen indices for the i-th user. Furthermore, let $\boldsymbol{\mathcal{L}}_1$ be the list of indices $i \in \mathcal{M}$ that satisfy (3.18) in the first stage of the identification process, while the return list is denoted by $\boldsymbol{\mathcal{L}}$. Consider the following events

$$\begin{aligned}
\mathcal{E}_u &= \{(X^n(W), U^n(J_W)) \notin \mathcal{T}_{\bar{\epsilon}}^n\},\\
\mathcal{E}_v &= \{(X^n(W), U^n(J_W), V^n(J_W, L_W)) \notin \mathcal{T}_{\bar{\epsilon}}^n\},\\
\mathcal{E}_{yz} &= \Big\{(Y^n, Z^n, X^n(W), U^n(J_W), V^n(J_W, L_W)) \notin \mathcal{T}_{\bar{\epsilon}}^n\Big\},\\
\mathcal{E}_1 &= \{|\boldsymbol{\mathcal{L}}_1| > 2^{n\hat{\Delta}}\},\\
\mathcal{E}_2 &= \{\exists l \neq L_W, l \in \mathfrak{B}(K_W), (Y^n, U^n(J_W), V^n(J_W, l)) \in \mathcal{T}_{\bar{\epsilon}}^n\},\\
\mathcal{E}_3 &= \Big\{\exists (w', l_{w'}), w' \neq W, w' \in \boldsymbol{\mathcal{L}}_1,\\
&\qquad (Y^n, U^n(J_{w'}), V^n(J_{w'}, l_{w'})) \in \mathcal{T}_{\bar{\epsilon}}^n,\ l'_w \in \mathfrak{B}(K_{w'})\Big\}. \qquad (3.21)
\end{aligned}$$

Define

$$\mathcal{E} = \mathcal{E}_u \cup \mathcal{E}_v \cup \mathcal{E}_{yz} \bigcup_{i=1}^{3} \mathcal{E}_i, \qquad (3.22)$$

to be the event that summarizes all "errors." By the strongly typical covering lemma [EK11, Lemma 3.3] we obtain

$$\Pr(\mathcal{E}_u) = \frac{1}{M} \sum_i \Pr\big((X^n(i), U^n(J_i)) \notin \mathcal{T}_{\bar{\epsilon}}^n\big) \to 0, \qquad (3.23)$$

if $\hat{R}_U > I(X;U) + \gamma_n$, where $\gamma_n > 0$ and $\gamma_n \to 0$ as $n \to \infty$, since W is independent of both $\big(X^n(i)\big)_{i=1}^M$ and the codebook. Similarly, we have $\Pr(\mathcal{E}_u^c \cap \mathcal{E}_v) \to 0$ if $\hat{R}_V + \hat{R}'_V > I(X;V|U) + \gamma_n$. Due to the Markov (conditional typicality) lemma [EK11, p.27], we have

$$\Pr(\mathcal{E}_u^c \cap \mathcal{E}_v^c \cap \mathcal{E}_{yz}) \to 0. \qquad (3.24)$$

For the sake of simplicity in the analysis of the last three events we use the symmetric property of our problem. Due to symmetry, it is sufficient to condition on the event $\{W = 1\}$. Following the analysis in [EK11, Section 11.3] we have

$$\Pr(\mathcal{E}_2 | W = 1) \to 0 \qquad (3.25)$$

as $n \to \infty$ if $\hat{R}'_V < I(Y;V|U) - \gamma_n$. We focus on the two remaining events $\mathcal{E}_1$ and $\mathcal{E}_3$. For each $i \in \mathcal{M}$ define an indicator random variable

$$B_i = \chi_{\left\{(Z^n, U^n(J_i)) \in \mathcal{T}_{\bar{\epsilon}}^n\right\}}. \qquad (3.26)$$

Note that $\Pr(B_1 = 1|W = 1) \to 1$ as $n \to \infty$. Hence, it is sufficient to consider the following probability

$$\begin{aligned}\Pr(B_1 = 1,\ |\boldsymbol{\mathcal{L}}_1| > 2^{n\hat{\Delta}}|W = 1) &= \Pr\left\{B_1 = 1, \sum_{i=1}^{2^{n\hat{R}}} B_i > 2^{n\hat{\Delta}} \middle| W = 1\right\} \\ &\le \Pr\left\{\sum_{i=2}^{2^{n\hat{R}}} B_i > 2^{n\hat{\Delta}} - 1 \middle| W = 1\right\} \\ &\le \frac{\sum_{i=2}^{2^{n\hat{R}}} \mathbb{E}[B_i|W = 1]}{2^{n\hat{\Delta}} - 1} = \frac{\sum_{i=2}^{2^{n\hat{R}}} \Pr\{B_i|W = 1\}}{2^{n\hat{\Delta}} - 1} \\ &\overset{(\star)}{\le} \xi 2^{n(\hat{R}-\hat{\Delta})} 2^{-n(I(Z;U)-\gamma_n)} \to 0 \end{aligned} \tag{3.27}$$

if $\hat{R} - \hat{\Delta} < I(Z;U) - \gamma_n$ where $\xi = (1 - 1/2^{n\hat{R}})/(1 - 1/2^{n\hat{\Delta}}) \to 1$ as $n \to \infty$. $(\star)$ is valid since conditioning on $W = 1$, Z^n is independent[3.5] of $U^n(J_i)$ for $i \in [2 : M]$. Therefore, for $i \ge 2$

$$\begin{aligned}\Pr\{B_i|W = 1\} &= \sum_{j_i}\sum_{u^n}\sum_{z^n \in \mathcal{T}_{\tilde{\epsilon}}^n(Z|u^n)} P_{J_i}(j_i) P_{U^n(J_i)|J_i}(u^n|j_i) P(Z^n = z^n|W = 1) \\ &\overset{(\star\star)}{\le} \sum_{j_i}\sum_{u^n}\sum_{z^n \in \mathcal{T}_{\tilde{\epsilon}}^n(Z|u^n)} P_{J_i}(j_i) P_{U^n(J_i)|J_i}(u^n|j_i) 2^{-n(H(Z)-\gamma_n)} \\ &\le 2^{-n(I(Z;U)-\gamma_n)}, \end{aligned} \tag{3.28}$$

where $(\star\star)$ holds since W is independent of Z^n and Z^n is iid according to the distribution P_Z. The expressions (3.24) and (3.27) imply that

$$\delta_{1,n} = \Pr(W \notin \boldsymbol{\mathcal{L}}) \to 0$$

as $n \to \infty$.
The probability of the last event can be bounded as

$$\begin{aligned}&\Pr(\mathcal{E}_3|W = 1) \\ &\le \Pr\left\{\exists(w', l_{w'}),\ w' \ne 1, (Y^n, U^n(J_{w'}), V^n(J_{w'}, l_{w'})) \in \mathcal{T}_{\tilde{\epsilon}}^n,\ l'_w \in \mathfrak{B}(K_{w'})|W = 1\right\} \\ &\le \sum_{i=2}^{2^{n\hat{R}}} \Pr\{\exists l_i \in \mathfrak{B}(K_i), (Y^n, U^n(J_i), V^n(J_i, l_i)) \in \mathcal{T}_{\tilde{\epsilon}}^n|W = 1\}. \end{aligned}$$

[3.5] This can be explained in more details as follows. Conditioning on $W = 1$, Z^n is independent of $X^n(i)$ for $i \in [2 : M]$ and the codebook, while $U^n(J_i)$ depends only on $X^n(i)$ and the codebook. Hence, we can use a single codebook for all users.

$$\leq \sum_{i=2}^{2^{n\hat{R}}} \sum_{j_i,k_i} \sum_{l_i \in \mathfrak{B}(k_i)} P_{J_iK_i}(j_i,k_i)$$
$$\times \Pr\{(Y^n, U^n(j_i), V^n(j_i,l_i)) \in \mathcal{T}_{\bar{\epsilon}}^n | W=1, J_i=j_i, K_i=k_i\}.$$

Since for $i=2,\ldots,M$

$$\begin{aligned}
&\Pr\{(Y^n, U^n(j_i), V^n(j_i,l_i)) \in \mathcal{T}_{\bar{\epsilon}}^n | W=1, J_i=j_i, K_i=k_i\} \\
&= \sum_{u^n,v^n} \sum_{y^n \in \mathcal{T}_{\bar{\epsilon}}^n(Y|u^n,v^n)} P_{U^n(j_i)V^n(j_i,l_i)|J_iK_i}(u^n,v^n|j_i,k_i) P(Y^n=y^n|W=1) \\
&\overset{(a)}{\leq} 2^{-n(H(Y)-H(Y|U,V)-\gamma_n)} = 2^{-n(I(Y;U,V)-\gamma_n)},
\end{aligned} \tag{3.29}$$

$\Pr(\mathcal{E}_3|W=1) \to 0$ if $\hat{R} + \hat{R}'_V < I(Y;U,V) - \gamma_n$, where (a) is valid due to the independence of Y^n and W. Since $\Pr(\mathcal{E}_1) \to 0$ and $\Pr(\mathcal{E}_3) \to 0$,

$$\delta_{2,n} = \Pr(\hat{W} \neq W) \to 0.$$

Moreover, due to the union bound

$$\Pr\{\mathcal{E}\} \to 0, \text{ as } n \to \infty. \tag{3.30}$$

Given $\mathcal{E}^c$, we obtain

$$(1-\bar{\epsilon})\mathbb{E}[d(X,f(U,V,Y))] < d(X^n(W),\hat{X}^n) < (1+\bar{\epsilon})\mathbb{E}[d(X,f(U,V,Y))], \tag{3.31}$$

by the typical average lemma [EK11, p.26], which implies that $\big|d(X^n(W),\hat{X}^n) - D\big| < \bar{\epsilon}D$. Hence, choosing $\bar{\epsilon} \to 0$ as $n \to \infty$

$$\begin{aligned}
\mathbb{E}\big[\big|d(X^n(W),\hat{X}^n) - D\big|\big] &< \mathbb{E}\big[|d(X^n(W),\hat{X}^n) - D|\big|\mathcal{E}^c\big] \\
&\quad + \Pr(\mathcal{E})(d_{\max} + D) = \delta_{3,n}.
\end{aligned} \tag{3.32}$$

Since $\Pr(\hat{W} \neq W) = \mathbb{E}[\chi_{\{\hat{W}\neq W\}}]$ and $\Pr(W \notin \boldsymbol{\mathcal{L}}) = \mathbb{E}[\chi_{\{W\notin\boldsymbol{\mathcal{L}}\}}]$, by the Selection Lemma [BB11, Lemma 2.2] there exists a codebook $\mathcal{H}_n$ such that

$$\begin{aligned}
&\Pr(\hat{W} \neq W|\mathcal{H}_n) < \delta_n, \qquad \Pr(W \notin \boldsymbol{\mathcal{L}}|\mathcal{H}_n) < \delta_n, \\
&\mathbb{E}\big[\big|d(X^n(W),\hat{X}^n) - D\big|\big|\mathcal{H}_n\big] < \delta_n,
\end{aligned} \tag{3.33}$$

where $\delta_n = 4\max(\{\delta_{i,n}\}_{i=1}^3)$. Since the space of codebooks is discrete,

$$\big|\mathbb{E}[d(X^n(W),\hat{X}^n)|\mathcal{H}_n] - D\big| \leq \mathbb{E}\big[|d(X^n(W),\hat{X}^n) - D|\big|\mathcal{H}_n\big],$$

which implies that

$$\mathbb{E}[d(X^n(W),\hat{X}^n)|\mathcal{H}_n] < D + \delta_n. \tag{3.34}$$

In summary, given an $\epsilon > 0$ if the following conditions

$$\begin{aligned} R_1 > I(X;U), \quad R_2 + R'_V > I(X;V|U), \\ R'_V < I(Y;V|U), \\ R < R_L + I(Z;U), \\ R + R'_V < I(Y;U,V), \end{aligned} \tag{3.35}$$

hold, then there exists a data processing scheme that satisfies all the requirements of Definition 3.2 for sufficiently large n. By using Fourier-Motzkin elimination [EK11, Appendix D] to eliminate R'_V we obtain[3.6]

$$\begin{aligned} R_1 > I(X;U), \quad R_2 > I(X;V|U,Y), \\ R_2 - R > I(X;V|U) - I(Y;U,V) \\ R < \min\{R_L + I(Z;U), I(Y;U,V)\}. \end{aligned} \tag{3.36}$$

In the next step we simplify the above region by a rate transfer argument. Assume that R'_1, R'_2, and Θ are positive numbers such that

$$\begin{aligned} &R'_1 - \Theta > I(X;U), \\ &R'_2 + \Theta > \max\Big\{I(X;V|U,Y), R + I(X;V|U) - I(Y;U,V)\Big\}. \end{aligned} \tag{3.37}$$

Herein, Θ is the rate transferred from storage Node 2 to storage Node 1. Since $I(X;U) \geq 0$, by (3.36) there exists an identification scheme such that $(R'_1 - \Theta, R'_2 + \Theta)$ is achievable for the given R. This implies the achievability of (R'_1, R'_2) for the given R. Applying the Fourier-Motzkin approach for a second time to eliminate Θ, the achievable rate region is enlarged to[3.7]

$$\begin{aligned} R'_1 \geq I(X;U), \\ R'_1 + R'_2 \geq I(X;U) + I(X;V|U,Y), \\ R'_1 + R'_2 - R \geq I(X;U,V|Y), \\ R \leq \min\{R_L + I(Z;U), I(Y;U,V)\}, \end{aligned} \tag{3.38}$$

since by definition, the achievable region is closed.

3.3.B Cardinality bounding of $\mathcal{U}$ and $\mathcal{V}$

It is sufficient to preserve the following quantities $H(X|U)$, $H(X|U,Y)$, $H(X|U,V,Y)$, $H(Z|U)$, $H(Y|U,V)$, the distortion constraint, and $p(x)$ for all but one $x \in \mathcal{X}$. By

[3.6] This can be seen from the following constraints: $R'_V > I(X;V|U) - R_2$, $R'_V > 0$, $R'_V < I(Y;V|U)$ and $R'_V < I(Y;U,V) - R$.

[3.7] More specifically, the enlarged region can be obtained from the following constraints $\Theta > 0$, $\Theta > I(X;V|U,Y) - R'_2$, $\Theta > R + I(X;V|U) - I(Y;U,V) - R'_2$ and $\Theta < R'_1 - I(X;U)$.

the support lemma [EK11, Appendix C] the cardinality of $\mathcal{U}$ and $\mathcal{V}$ can be bounded by

$$\begin{aligned} |\mathcal{U}| &\leq |\mathcal{X}| + 6, \\ |\mathcal{V}| &\leq (|\mathcal{X}| + 6)(|\mathcal{X}| + 2). \end{aligned} \tag{3.39}$$

This implies that $\mathcal{R}^\star$ is a closed region.

3.3.C Converse

Given $\epsilon > 0$ small enough, assume that there exist mappings such that all the conditions are fulfilled for *all* sufficiently large n. Furthermore by taking n large enough, the condition $\frac{1}{n} < \epsilon$ is valid. For notation brevity we abbreviate $(J_i)_{i=1}^M$ as $\boldsymbol{J}$ and $(K_i)_{i=1}^M$ as $\boldsymbol{K}$. The corresponding realizations are denoted by $\boldsymbol{j}$, and $\boldsymbol{k}$. Since $\Pr(\hat{W} \neq W) < \epsilon$, Fano's inequality for the second stage implies

$$H(W|Y^n,(J_{\boldsymbol{\mathcal{L}}}, K_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{L}})) < 1 + \Pr(\hat{W} \neq W) \log_2 M < 1 + \epsilon \log_2 M. \tag{3.40}$$

We also establish a variant of Fano's inequality for the first stage. Define an auxiliary random variable

$$E = \chi_{\left\{W \in g_1(Z^n, \boldsymbol{J})\right\}}. \tag{3.41}$$

Since W is in the list when $E = 1$, the error probability $P_e = \Pr(E = 0)$ is bounded by ϵ. We obtain the following inequality

$$\begin{aligned} H(E, W|Z^n, \boldsymbol{J}) &= \mathbb{E}[-\log_2 \Pr(E, W|Z^n, \boldsymbol{J})] \\ &= \mathbb{E}[-\log_2 \Pr(W|Z^n, \boldsymbol{J})] + \mathbb{E}[-\log_2 \Pr(E|W, Z^n, \boldsymbol{J})] \\ &= H(W|Z^n, \boldsymbol{J}) \\ &= \mathbb{E}[-\log_2 \Pr(E|Z^n, \boldsymbol{J})] + \mathbb{E}[-\log_2 \Pr(W|E, Z^n, \boldsymbol{J})] \\ &\overset{(**)}{\leq} H(E) + \mathbb{E}\big[\Pr(E = 0|Z^n, \boldsymbol{J}) H(W|E = 0, Z^n, \boldsymbol{J}) \\ &\qquad + \Pr(E = 1|Z^n, \boldsymbol{J}) H(W|E = 1, Z^n, \boldsymbol{J})\big] \\ &\leq h_b(P_e) + P_e \log_2 M + n(R_L + \epsilon) \leq n(R_L + \epsilon_n), \end{aligned} \tag{3.42}$$

where $\epsilon_n = 2\epsilon + \frac{1}{n}\epsilon \log_2 M$ and $h_b(\cdot)$ is the binary entropy function. $(**)$ follows from the disintegration, i.e., the computing order is $\mathbb{E}_{Z^n \boldsymbol{J}}[\mathbb{E}_{E|Z^n \boldsymbol{J}}[\mathbb{E}_{W|E,Z^n,\boldsymbol{J}}(\cdot)]]$ Define random variables

$$U_i = (W, J_W, Y^{i-1}), \text{ and } V_i = (U_i, K_W, Y_{i+1}^n), \quad i \in [1 : n]. \tag{3.43}$$

Observe that $U_i - V_i - X_i(W) - Y_i - Z_i$ for all $i \in [1 : n]$, due to the memoryless property of the observation and pre-processing channels and the source. The identification rate can be bounded firstly as

$$n(R - \epsilon) \leq \log_2 M = H(W) = I(W; Z^n, \boldsymbol{J}) + H(W|Z^n, \boldsymbol{J})$$

$$
\begin{aligned}
&\overset{(\star)}{\leq} I(W; Z^n|\boldsymbol{J}) + n(R_L + \epsilon_n) \\
&\leq I(W, \boldsymbol{J}; Z^n) + n(R_L + \epsilon_n) = I(W, J_W; Z^n) + n(R_L + \epsilon_n) \\
&= \sum_{i=1}^{n} I(W, J_W, Z^{i-1}; Z_i) + n(R_L + \epsilon_n), \\
&\overset{(a)}{\leq} \sum_{i=1}^{n} I(W, J_W, Y^{i-1}; Z_i) + n(R_L + \epsilon_n),
\end{aligned} \tag{3.44}
$$

where $(\star)$ holds due to (3.42) and since W is independent of $\boldsymbol{J}$. (a) holds due to the Markov chain $Z^{i-1} - Y^{i-1} - (Z_i, W, J_W)$ for all $i \in [1:n]$, due to the memoryless property of the pre-processing. This implies that

$$
(R - \epsilon)(1 - \epsilon) \leq \frac{1}{n} \sum_{i=1}^{n} I(U_i; Z_i) + R_L + 2\epsilon. \tag{3.45}
$$

Secondly,

$$
\begin{aligned}
n(R - \epsilon) \leq \log_2 M = H(W) &= I(W; Y^n, (J_{\boldsymbol{\mathcal{L}}}, K_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{L}})) + H(W|Y^n, (J_{\boldsymbol{\mathcal{L}}}, K_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{L}})) \\
&\overset{(b)}{\leq} I(W; Y^n, Z^n, \boldsymbol{J}, \boldsymbol{K}) + 1 + \epsilon \log_2 M \\
&\overset{(c)}{=} I(W; Y^n, \boldsymbol{J}, \boldsymbol{K}) + 1 + \epsilon \log_2 M \\
&\overset{(\star\star)}{\leq} I(W, \boldsymbol{J}, \boldsymbol{K}; Y^n) + 1 + \epsilon \log_2 M \\
&= I(W, J_W, K_W; Y^n) + 1 + \epsilon \log_2 M \\
&\leq \sum_{i=1}^{n} I(W, J_W, K_W, Y^{n\setminus i}; Y_i) + 1 + \epsilon \log_2 M,
\end{aligned} \tag{3.46}
$$

where (b) holds since by eqs. (3.2) and (3.3), $\boldsymbol{\mathcal{L}} = g_1(Z^n, \boldsymbol{J})$, and

$$
(J_{\boldsymbol{\mathcal{L}}}, K_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{L}}) = ((J_i)_{i \in \boldsymbol{\mathcal{L}}}, (K_i)_{i \in \boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{L}}) = \pi(\boldsymbol{J}, \boldsymbol{K}, \boldsymbol{\mathcal{L}}),
$$

hold. We also use the inequality (3.40) in (b). (c) is valid due to the Markov chain $Z^n - Y^n - (W, \boldsymbol{J}, \boldsymbol{K})$. $(\star\star)$ holds since W is independent of both $\boldsymbol{J}$ and $\boldsymbol{K}$. Using (3.43) and (3.46) gives us

$$
(R - \epsilon)(1 - \epsilon) \leq \frac{1}{n} \sum_{i=1}^{n} I(U_i, V_i; Y_i) + \epsilon. \tag{3.47}
$$

Furthermore, the sum of the compressed rates can be bounded as

$$
\begin{aligned}
n(R_1 + R_2 + \epsilon) &\geq H(J_W, K_W|W) \geq I(X^n(W), Y^n; J_W, K_W|W) \\
&\geq I(Y^n; J_W|W) + I(X^n(W); J_W, K_W|W, Y^n)
\end{aligned}
$$

$$\begin{aligned}
&\overset{(d)}{=} I(Y^n; J_W, W) + I(X^n(W); J_W, K_W, W|Y^n) \\
&\overset{(e)}{=} \sum_{i=1}^{n} \Big(I(Y_i; W, J_W, Y^{i-1}) + I(X_i(W); W, J_W, K_W, Y^{n\setminus i}, X^{i-1}(W)|Y_i) \Big) \\
&\overset{(f)}{\geq} \sum_{i=1}^{n} \Big(I(X_i(W), Y_i; W, J_W, Y^{i-1}) + I(X_i(W); K_W, Y_{i+1}^n|Y_i, W, J_W, Y^{i-1}) \Big) \\
&\overset{(g)}{=} \sum_{i=1}^{n} I(X_i(W); U_i) + I(X_i(W); V_i|U_i, Y_i), \qquad (3.48)
\end{aligned}$$

where (d) is valid since W is independent of both $X^n(W)$ and Y^n. (e) is true due to the memoryless property of the observational channel. (f) holds since we drop the term $X^{i-1}(W)$ in the second term. (g) follows from the Markov chain $Y_i - X_i(W) - (W, J_W, Y^{i-1})$ for all $i \in [1:n]$. Similarly, we can show that

$$\begin{aligned}
n(R_1 + \epsilon) &\geq H(J_W|W) \geq I(X^n(W); J_W|W) = I(X^n(W); J_W, W) \\
&= \sum_{i=1}^{n} I(X_i(W); W, J_W, X^{i-1}(W)) \geq \sum_{i=1}^{n} I(X_i(W); U_i). \qquad (3.49)
\end{aligned}$$

In addition, using the first line of (3.48) and the second last line in (3.46) we obtain

$$\begin{aligned}
&n(R_1 + R_2 + \epsilon) - \log_2 M \\
&\geq H(J_W, K_W|W) - I(W, J_W, K_W; Y^n) - (1 + \epsilon \log_2 M) \\
&\geq I(X^n(W); J_W, K_W, W) - I(Y^n; J_W, K_W, W) - (1 + \epsilon \log_2 M) \\
&\overset{(*)}{=} I(X^n(W); J_W, K_W, W|Y^n) - (1 + \epsilon \log_2 M) \\
&= \sum_{i=1}^{n} I(X_i(W); J_W, K_W, W|Y^n, X^{i-1}(W)) - (1 + \epsilon \log_2 M) \\
&\overset{(e)}{=} \sum_{i=1}^{n} I(X_i(W); J_W, K_W, W, Y^{n\setminus i}, X^{i-1}(W)|Y_i) - (1 + \epsilon \log_2 M) \\
&\geq \sum_{i=1}^{n} I(X_i(W); U_i, V_i|Y_i) - (1 + \epsilon \log_2 M), \qquad (3.50)
\end{aligned}$$

where $(*)$ holds due to the memoryless property of the observation channel, i.e., $Y^n - X^n(W) - (W, J_W, K_W)$ and (e) holds as before. This leads to

$$R_1 + R_2 + 2\epsilon - (R - \epsilon)(1 - \epsilon) \geq \frac{1}{n} \sum_{i=1}^{n} I(X_i(W); U_i, V_i|Y_i). \qquad (3.51)$$

Since

$$D + \epsilon > \mathbb{E}[d(X^n(W), g_3(J_{\hat{W}}, K_{\hat{W}}, \hat{W}, Y^n))]$$

$$> \Pr(\hat{W} = W) \times \mathbb{E}[d(X^n(W), g_3(J_W, K_W, W, Y^n))|\hat{W} = W], \tag{3.52}$$

the following chain of expressions holds

$$\begin{aligned}
&\mathbb{E}[d(X^n(W), g_3(W, J_W, K_W, Y^n))] \\
&\quad \leq \mathbb{E}[d(X^n(W), g_3(J_W, K_W, W, Y^n))|\hat{W} = W]\Pr(\hat{W} = W) \\
&\quad + \Pr(\hat{W} \neq W)d_{\max} < D + (1 + d_{\max})\epsilon.
\end{aligned} \tag{3.53}$$

Let Q be a random variable uniformly distributed on $[1 : n]$ and independent of everything else. Define

$$U = (U_Q, Q), \quad V = (V_Q, Q), \text{ and } f(U, V, Y_Q) = g_{3Q}(J_W, K_W, W, Y^n). \tag{3.54}$$

Note that $U - V - X_Q(W) - Y_Q - Z_Q$ still holds. Then the above constraints can be rewritten as

$$\begin{aligned}
(R - \epsilon)(1 - \epsilon) &\leq I(U_Q; Z_Q|Q) + R_L + 2\epsilon = I(U; Z_Q) + R_L + 2\epsilon \\
(R - \epsilon)(1 - \epsilon) &\leq I(U, V; Y_Q) + \epsilon \\
R_1 + R_2 + \epsilon &\geq I(X_Q(W); U) + I(X_Q(W); V|U, Y_Q) \\
R_1 + \epsilon &\geq I(X_Q(W); U) \\
R_1 + R_2 + 2\epsilon - (R - \epsilon)(1 - \epsilon) &\geq I(X_Q(W); U, V|Y_Q) \\
D + (1 + d_{\max})\epsilon &> \mathbb{E}[d(X_Q(W), f(U, V, Y_Q))].
\end{aligned} \tag{3.55}$$

Since $(X_Q(W), Y_Q, Z_Q)$ has the joint distribution as $P_{XY} \times P_{Z|Y}$,

$$\Big((R - \epsilon)(1 - \epsilon) - \epsilon, R_1 + \epsilon, R_2 + \epsilon, R_L + \epsilon, D + (1 + d_{\max})\epsilon\Big) \in \mathcal{R}^\star$$

by the cardinality bounding arguments presented in Subsection 3.3.B. Taking $\epsilon \to 0$ completes the backward direction since $\mathcal{R}^\star$ is closed. ■

3.A Proof of Corollary 3.1

Direct part: Rate tuples that fulfill the conditions given in (3.11) also satisfy the conditions given in (3.9) with $D = d_{\max}$ and an arbitrary deterministic mapping f. Hence they are achievable.
Converse part: Similarly, we define auxiliary random variables

$$U_i = (W, J_W, Z^{i-1}), \text{ and } V_i = (U_i, K_W, Y^{i-1}), \quad i \in [1 : n]. \tag{3.56}$$

Then $U_i - V_i - X_i(W) - Y_i - Z_i$ for all $i \in [1 : n]$. The two first constraints on the compression rates can be derived shortly as

$$n(R_1 + \epsilon) \geq \sum_{i=1}^{n} I(X_i(W); W, J_W, X(W)^{i-1}) \geq \sum_{i=1}^{n} I(X_i(W); U_i), \tag{3.57}$$

and

$$\begin{aligned} n(R_1+R_2+\epsilon) &\geq \sum_{i=1}^{n} I(X_i(W); W, J_W, K_W, X^{i-1}(W)) \\ &= \sum_{i=1}^{n} I(X_i(W); W, J_W, K_W, X^{i-1}(W), Y^{i-1}, Z^{i-1}) \\ &\geq \sum_{i=1}^{n} I(X_i(W); U_i, V_i). \end{aligned} \tag{3.58}$$

Following the same steps which lead to (3.46), we obtain

$$\begin{aligned} n(R-\epsilon) \leq H(W) &= I(W; Y^n, (J_{\boldsymbol{\mathcal{L}}}, K_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{L}})) + H(W|Y^n, (J_{\boldsymbol{\mathcal{L}}}, K_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{L}})) \\ &\overset{(*)}{\leq} I(W, J_W, K_W; Y^n) + 1 + \epsilon \log_2 M \\ &\leq \sum_{i=1}^{n} I(U_i, V_i; Y_i) + 1 + \epsilon \log_2 M, \end{aligned} \tag{3.59}$$

where $(*)$ holds due to the Markov chain $Z^n - Y^n - (W, \boldsymbol{J}, \boldsymbol{K})$. In addition, from (3.44) we obtain

$$\begin{aligned} n(R-\epsilon) &\leq \sum_{i=1}^{n} I(W, J_W, Z^{i-1}; Z_i) + n(R_L + \epsilon_n) \\ &= \sum_{i=1}^{n} I(Z_i; U_i) + n(R_L + \epsilon_n). \end{aligned} \tag{3.60}$$

The rest follows by defining a uniform random variable Q on the set $[1:n]$ and taking $\epsilon \to 0$ as in Theorem 3.1. The cardinality of $\mathcal{U}$ and $\mathcal{V}$ can be bounded similarly using the support lemma [EK11, Appendix C].

3.B Proof of Proposition 3.1

The proof follows closely the one of Theorem 3.1 with some modifications.
Achievability: $2^{n\hat{R}_U}$ codewords $u^n(j)$ are generated as before. For each m we draw $2^{n\hat{R}_V}$ codewords $v^n(j,k)$ iid via the marginal $p_{V|U}$, i.e., no binning is used. The enrollment process follows accordingly. The identification process corresponding to Observer 2 works identically as the first stage in (3.18) while for Observer 1 the processing unit searches through *all* users to find the unique $\hat{w}$ such that

$$(y^n, u^n(j_{\hat{w}}), v^n(j_{\hat{w}}, k_{\hat{w}})) \in \mathcal{T}_\epsilon^n, \tag{3.61}$$

which leads to the following event in the analysis

$$\mathcal{E}_3' = \left\{ \exists w', w' \neq W, (Y^n, U^n(J_{w'}), V^n(J_{w'}, K_{w'})) \in \mathcal{T}_\epsilon^n \right\}.$$

Similarly, we have $\Pr(\mathcal{E}_3'|W=1) \to 0$ if $\hat{R} < I(Y;U,V) - \gamma_n$.
One might notice that the condition $X - Y - Z$ is not used in the achievability proof of Theorem 3.1. Hence, it can be concluded that the two stage processing in the achievability of Theorem 3.1 achieves the rate region of Proposition 3.1.
Converse: Define the random variables U_i and V_i as in (3.56). We also obtain the constraints as in (3.57), (3.58), and (3.60). To arrive at (3.59) we need the following modification

$$\begin{aligned} n(R-\epsilon) \leq H(W) &= I(W;Y^n,\boldsymbol{J},\boldsymbol{K}) + H(W|Y^n,\boldsymbol{J},\boldsymbol{K}) \\ &\overset{(\star)}{\leq} I(W,J_W,K_W;Y^n) + 1 + \epsilon \log_2 M, \end{aligned} \tag{3.62}$$

where $(\star)$ follows from the Fano's inequality and the requirement in (3.13).

4

Gaussian Hierarchical ID

In this chapter we extend the setting in Chapter 3 to a Gaussian setting. Namely, we assume that the users' data are independently Gaussian distributed, i.e., $X_i(w) \sim \mathcal{N}(0, \sigma_X^2)$, $\forall i \in [1:n], w \in [1:M]$. The observation sequence Y^n is assumed to be related to $X^n(W)$ via the following relation

$$Y_i = X_i(W) + N_{1i},\ \forall i \in [1:n],$$

while the pre-processed sequence Z^n is assumed to be given by

$$Z_i = Y_i + N_{2i},\ \forall i \in [1:n],$$

where $N_{1i} \sim \mathcal{N}(0, \sigma_{N_1}^2)$ and $N_{2i} \sim \mathcal{N}(0, \sigma_{N_2}^2)$, $\forall i$, are iid random variables, which are also independent of the users' data and each other. In other words, the observation and pre-processing channels are iid Gaussian. The AWGN assumptions facilitate the detailed analysis[4.1].

We take the reconstruction set to be the set of real numbers, i.e., $\hat{\mathcal{X}} = \mathbb{R}$. The distortion measure is the squared error distance

$$d(x^n, \hat{x}^n) = \frac{1}{n}\|x^n - \hat{x}^n\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2. \tag{4.1}$$

4.1 Statement of Results

The definition of an identification scheme and achievability follows similarly as the ones given in Definitions 3.1 and 3.2 in which the processing mappings $\{g_i\}_{i=1}^3$ are

[4.1] For a zero-mean unit variance jointly Gaussian pair (X, Y) we can express, from the structure of the conditional distribution $P_{Y|X}$, Y as $Y = \rho X + W$ where ρ is the Pearson's correlation coefficient and $W \sim \mathcal{N}(0, (1-\rho^2))$ is independent of X. Since the presence of the correlation coefficient ρ could make the analysis and the choice of auxiliary random variables even more complicated than our model, we do not pursue the details herein.

measurable. The enrollment mappings $\{\phi_{in}\}_{i=1}^{2}$ given by

$$\phi_{in}\colon \mathbb{R}^n \to \mathcal{M}_i,\ i=1,2,$$

are also required to be measurable. The reader is referred to Section 2.5 in Chapter 2 for the detailed convention. In the Section 4.2 we provide a proof of the following observation.

Theorem 4.1 *Let (R, R_1, R_2, R_L, D) be a rate-distortion tuple such that there exist random variables U and V with a joint conditional probability density*[4.2] *$p_{UV|X}$ and a measurable reconstruction mapping $g\colon \mathbb{R}^3 \to \mathbb{R}$ such that the following conditions are fulfilled.*

$$R_1 \geq I(X;U), \tag{4.2a}$$
$$R_1 + R_2 \geq I(X;U) + I(X;V|U,Y), \tag{4.2b}$$
$$R_1 + R_2 - R \geq I(X;U,V|Y), \tag{4.2c}$$
$$R \leq \min\{R_L + I(Z;U), I(Y;U,V)\}, \tag{4.2d}$$
$$D \geq \mathbb{E}[d(X, g(U,V,Y))]. \tag{4.2e}$$

Then (R, R_1, R_2, R_L, D) is achievable in the sense of Definition 3.2.

It will be clear from Section 4.2 that our proof for Theorem 4.1 can be transferred directly to the discrete case as the pmfs in the discrete case can be viewed as density functions w.r.t. the counting measure. Due to the formal analytical complexity of the Gaussian case, where we have a mixture of discrete and continuous random variables, we choose to present its proof separately for the sake of clarity. Theorem 4.1 allows us to derive the rate-distortion region for the Gaussian setting, denoted by $\mathcal{R}_{GS}$, which is given by the following theorem. For notation simplicity we define the following three auxiliary functions of R in which all other parameters are fixed.

$$h_0(R) = \log_2 \frac{\sigma_X^2 \sigma_{N_1}^2}{\sigma_Y^2 D}, \quad h_1(R) = \log_2 \frac{\sigma_X^2 2^{-2R}}{\sigma_Y^2 2^{-2R} - \sigma_{N_1}^2},$$
$$h_2(R) = \log_2 \frac{\sigma_X^2}{\sigma_Y^2} \frac{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2}{\sigma_Z^2 2^{-2(R-R_L)} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)}. \tag{4.3}$$

Theorem 4.2 *Assume that $0 \leq R_L \leq R$ and $0 < D \leq \frac{\sigma_X^2 \sigma_{N_1}^2}{\sigma_Y^2}$. Then the corresponding rate-distortion region $\mathcal{R}_{GS}$ is given by*

$$R < R_\gamma, \tag{4.4a}$$
$$R_1 \geq \frac{1}{2} \log_2 \left(\frac{\sigma_X^2}{\sigma_Z^2 2^{-2(R-R_L)} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)} \right), \tag{4.4b}$$

[4.2] The ranges of U and V are $\mathbb{R}$ and the joint density p_{XUV} is with respect to the (product) Lebesgue measure in $\mathbb{R}^3$.

$$R_1 + R_2 \geq \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2} + \Gamma, \tag{4.4c}$$

$$R_1 + R_2 - R \geq \Gamma, \tag{4.4d}$$

where

$$R_\gamma = \min\left\{\frac{1}{2}\log_2\left(\frac{\sigma_Z^2}{\sigma_{N_1}^2 + \sigma_{N_2}^2}\right) + R_L, \frac{1}{2}\log_2\left(\frac{\sigma_Y^2}{\sigma_{N_1}^2}\right)\right\},$$

$$\Gamma = \frac{1}{2}\max\big\{h_0(R), h_1(R), h_2(R)\big\}. \tag{4.5}$$

Remark 4.1 The constraint (4.4a) corresponds to the constraint (4.2d) where R_γ can be seen as the supremum of the right-hand side of (4.2d) w.r.t *any* pair of auxiliary random variables U and V such that $YZ - X - UV$ holds and the mutual information terms are well-defined.
The constraint $R_L \leq R$ is motivated from the fact that the first layer cannot reasonably output a list with size larger than the number of users in the system. As for the second restriction on the distortion level D, if we consider for each $i \in [1:n]$ estimating $X_i(W)$ using Y_i and the MMSE estimator, then the distortion level is exactly $\frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2}$. With additional information, the system in general can do better than this bound. If for some unknown reason, the target list size or the target distortion level is set above the corresponding thresholds, then the corresponding terms, related to D or R_L, in (4.4) are omitted. For instance, the rate-distortion trade-off when $R_L > R$ and $0 < D \leq \sigma_X^2\sigma_{N_1}^2/\sigma_Y^2$ is given by

$$R_1 + R_2 \geq R + \frac{1}{2}\max\left\{\log_2\frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2 D}, \log_2\frac{\sigma_X^2 2^{-2R}}{\sigma_Y^2 2^{-2R} - \sigma_{N_1}^2}\right\}$$

$$0 \leq R < \frac{1}{2}\log_2\left(\frac{\sigma_Y^2}{\sigma_{N_1}^2}\right). \tag{4.6}$$

By definition the rate-distortion region $\mathcal{R}$ is closed in the finite dimensional metric space induced by the ℓ_1 distance. However, the constraint $R < R_L$ and $D > 0$ may lead to the impression that the region is not necessary closed. In Appendix 4.C we show that $\mathcal{R}_{GS}$ is indeed closed.

The proof of Theorem 4.2 is divided into the following parts. We first establish an outer bound on the achievable rate-distortion region. It can be seen from the Markov structure in Theorem 4.1 that the second constraint (4.2b) can be reformulated as

$$R_1 + R_2 \geq I(Y;U) + I(X;U,V|Y).$$

The terms corresponding to the mutual information quantities $I(X;U)$ and $I(Y;U)$ can be lower bounded without difficulty by applying the entropy power inequalities. The crucial idea for deriving the outer bound is to minimize the term "related to" $I(X;U,V|Y)$ while all other parameters are fixed which results in the term Γ. The

Cases		Subcases	Dominating functions			Distributions of U and V
			$h_0(R)$	$h_1(R)$	$h_2(R)$	
$R_{cr_{12}} < R_\gamma$	$R_{cr_{01}} \leq R_{cr_{12}}$	I. $R_{cr_{01}} \leq R < R_{cr_{12}}$		✓		$U \sim P_U$, $V \sim P_V$
		II. $0 \leq R < R_{cr_{01}}$	✓			$U \sim P_U$, $V \sim P_V(R_{cr_{01}})$
		III. $R_{cr_{12}} \leq R < R_\gamma$			✓	$U \sim P_U$, V degenerate
	$R_{cr_{01}} > R_{cr_{12}}$	IV. $R_{cr_{02}} \leq R < R_\gamma$			✓	As in Case III
		V. $R_L \leq R < R_{cr_{02}}$	✓			$U \sim P_U$, $V \sim P_V(R_{cr_{01}})$
$R_{cr_{12}} \geq R_\gamma$	$R_{cr_{01}} > R_L$	VI. $R_{cr_{01}} \leq R < R_\gamma$		✓		As in Case I
		VII. $R_L \leq R < R_{cr_{01}}$	✓			As in Case II
	$R_{cr_{01}} \leq R_L$	VIII. $\forall R$		✓		As in Case I

Table 4.1: Summary of optimal (marginal) distributions of the auxiliary random variables U and V for all possible cases specified by the relation among $R_{cr_{12}}$, $R_{cr_{01}}$, $R_{cr_{02}}$, R_L, R_γ and R where $P_U = \mathcal{N}\big(0, \sigma_Z^2(1-2^{-2(R-R_L)})\big)$ and $P_V = \mathcal{N}\big(0, \sigma_Y^2(1-2^{-2R})\big)$. Note that the marginal distribution of the auxiliary random variable U does not change. Additionally, due to the relation (4.134) the distribution of V in Case V is identical to the one in Case II.

approach is particularly helpful in our scenario, since it does not create additional parameters for describing the region.

Then, we discuss in Subsection 4.3.C how to resolve the complicated outer bound into small subregions that can be achieved by different parameterized coding schemes. This is done by studying transition points in the characterization of Γ. Finally, we show that each region can be achieved by an appropriate choice of auxiliary random variables U and V in Theorem 4.1, hence implying that the complete outer bound is achievable. For an overview we summarize In Table 4.1 marginal distributions of the chosen U and V for all cases.

4.2 Proof of Theorem 4.1

We provide a justification for Theorem 4.1 in several steps. In the first step we establish a supporting covering lemma, which bypasses the need of a Markov lemma for weak typicality[4.3]. In the next step we provide a coding scheme which is based on the adapted covering lemma. The analysis only highlights the important parts. Our approach resembles the one given in [WZ76; Wyn78] with a tweak in the "error" analysis.

As reviewed in Chapter 2, since the setup is a continuous scenario, the idea is to look for a suitable mapping $\tilde{\psi}_n$ which is used to control "error" events. To facilitate the designing process, which is quite mechanical, we highlight the main steps in the proof in the following remark. Assume that we use a codebook in which $\boldsymbol{u}^n$ is used to represent the first layer, and $\boldsymbol{v}^n$ represents the second layer. As in the proof of the discrete case in Chapter 3 we would like to have that $(Y^n, Z^n, U^n(J_W), V^n(J_W, L_W))$ is jointly weakly typical with high probability. Since we do not have the conditional typicality lemma, we include this constraint in the definition of $\tilde{\psi}_n$. Once the joint typicality is guaranteed, the analysis of the list and identification constraints can be proceeded similarly as in Chapter 3.

Additionally since the distortion measure is not bounded, the average distortion level is not necessarily bounded even if the excess distortion probability goes to zero. To achieve the desired average distortion level we also include a failsafe mechanism in the definition of $\tilde{\psi}_n$ which consists of two steps. Firstly we quantize the reconstruction mapping g so that the resulting mapping called $\hat{g}$ takes only a finite number of output values. Secondly, we require that with high probability the combination of the chosen codewords and the side information sequence produces a reconstructed sequence with a typical normalized distortion level, less that $D+\epsilon$.

It is interesting to note that we do not quantize the auxiliary random variables as in [Wyn78]. This keeps us safe from having to define a jointly weak typical set with a mixture of continuous, notably X, and discrete random variables. The complete proof is given in the following.

4.2.A Prelude

To differentiate between weak and strong typicality, given $0 < \delta < 1$ we denote the weakly typical set by $\mathcal{A}_\delta^n$ whose definition for a tuple of random variables $(X_1, \ldots, X_k)$ with a joint probability density function $p_{X_1 X_2 \ldots X_k}$ is given by [CT12, p. 521], [Ooh98, Lemma 3]

$$\mathcal{A}_\delta^n(X_1 \ldots X_k) = \left\{ (x_1^n, \ldots, x_k^n) \middle| \left| -\frac{1}{n} \log p_{X_\mathcal{S}}^n(x_\mathcal{S}^n) - h(X_\mathcal{S}) \right| < \delta, \ \forall \mathcal{S} \subseteq [1:k] \right\}$$

[4.3] Markov lemmas for continuous alphabets can be found in the works [Ooh98, Lemma 5] in the context of weak typicality and [Mit15] in the sense of weak*-typicality. However, it is not obvious to extend Lemma 5 in [Ooh98] to multiple layers of auxiliary random variables used in our superposition coding scheme. Bounding the distortion level using the approach in [Mit15] is difficult since the distortion measure is unbounded.

where $h(\cdot)$ denotes the differential entropy[4.4] and $X_{\mathcal{S}} = (X_i)_{i\in\mathcal{S}}$. Some important properties of weakly typical sequences are given in the following:

- If $x_{\mathcal{S}}^n \in \mathcal{A}_\delta^n(X_{\mathcal{S}})$ then

$$2^{-n(h(X_{\mathcal{S}})+\delta)} \leq p_{X_{\mathcal{S}}}^n(x_{\mathcal{S}}^n) \leq 2^{-n(h(X_{\mathcal{S}})-\delta)}. \tag{4.7}$$

- If $\mathcal{S}_1 \cap \mathcal{S}_2 = \varnothing$ and $(x_{\mathcal{S}_1}^n, x_{\mathcal{S}_2}^n) \in \mathcal{A}_\delta^n(X_{\mathcal{S}_1\cup\mathcal{S}_2})$ where $\mathcal{S}_1, \mathcal{S}_2 \subset [1:k]$ then

$$2^{-n(h(X_{\mathcal{S}_1}|X_{\mathcal{S}_2})+2\delta)} \leq p_{X_{\mathcal{S}_1}|X_{\mathcal{S}_2}}^n(x_{\mathcal{S}_1}^n|x_{\mathcal{S}_2}^n) \leq 2^{-n(h(X_{\mathcal{S}_1}|X_{\mathcal{S}_2})-2\delta)}. \tag{4.8}$$

- For $x_{\mathcal{S}_1}^n \in \mathcal{A}_\delta^n(X_{\mathcal{S}_1})$ then

$$\mathrm{Vol}(\mathcal{A}_\delta^n(X_{\mathcal{S}_2}|x_{\mathcal{S}_1}^n)) \leq 2^{n(h(X_{\mathcal{S}_2}|X_{\mathcal{S}_1})+2\delta)}, \tag{4.9}$$

  where $\mathcal{A}_\delta^n(X_{\mathcal{S}_2}|x_{\mathcal{S}_1}^n)$ is the conditional typical set. Note that the left-hand side is zero if $x_{\mathcal{S}_1}^n \notin \mathcal{A}_\delta^n(X_{\mathcal{S}_1})$ as the set $\mathcal{A}_\delta^n(X_{\mathcal{S}_2}|x_{\mathcal{S}_1}^n)$ is empty in this case.

Assume that the tuple $(X^n, Y^n, Z^n, U^n, V^n)$ is generated iid from the joint density p_{XYZUV}. For brevity, we denote herein by $\tilde{Y}^n$ the pair (Y^n, Z^n). Then due to the weak law of large numbers we have the following properties

$$\Pr\{(\tilde{Y}^n, U^n, V^n) \notin \mathcal{A}_\delta^n(\tilde{Y}UV)\} \to 0, \tag{4.10}$$

as well as

$$\Pr\big\{\big|d(X^n, g(U^n, V^n, Y^n)) - D\big| > \delta\big\} \to 0, \text{ as } n \to \infty, \tag{4.11}$$

when we assume that $D = \mathbb{E}[d(X, g(U, V, Y))] < \infty$. As in [WZ76] we define the following indicator function

$$\psi_n(x^n, \tilde{y}^n, u^n, v^n) = \begin{cases} 1 & \text{if } \big|d(x^n, g(u^n, v^n, y^n)) - D\big| > \delta, \\ & \text{or } (\tilde{y}^n, v^n, u^n) \notin \mathcal{A}_\delta^n(\tilde{Y}UV) \\ 0 & \text{otherwise} \end{cases}. \tag{4.12}$$

Let $\delta_n = \mathbb{E}[\psi_n(X^n, \tilde{Y}^n, U^n, V^n)]$, then due to the union bound, (4.10) and (4.11) we have

$$\delta_n \to 0, \quad \text{as } n \to \infty. \tag{4.13}$$

For brevity define[4.5]

$$\mathcal{S}_n^\delta = \{(x^n, u^n, v^n) \colon \eta_{XUV}(x^n, u^n, v^n) \leq \delta_n^{1/2}\}, \tag{4.14}$$

[4.4] Note that $\mathcal{A}_\delta^n(X_1 \dots X_k)$ is a Borel-measurable set.

[4.5] Since ψ_n is a Borel measurable and integrable mapping as $\psi_n \in \{0, 1\}$, η_{XUV} is also Borel measurable and non-negative almost everywhere. Hence, the set $\mathcal{S}_n^\delta$ is Borel measurable.

where

$$\begin{aligned}\eta_{XUV}(x^n,u^n,v^n) &= \mathbb{E}[\psi_n(x^n,\tilde{Y}^n,u^n,v^n)|X^n=x^n,V^n=v^n,U^n=u^n]\\ &= \mathbb{E}[\psi_n(x^n,\tilde{Y}^n,u^n,v^n)|X^n=x^n]. \end{aligned} \tag{4.15}$$

Due to the Markov inequality we have

$$\Pr\{(X^n,U^n,V^n)\notin\mathcal{S}_n^\delta\} \le \frac{\mathbb{E}[\psi_n(X^n,\tilde{Y}^n,U^n,V^n)]}{\delta_n^{1/2}} = \delta_n^{1/2}. \tag{4.16}$$

Finally, define[4.6]

$$\mathcal{B}_n^\delta = \mathcal{A}_\delta^n(UVX)\cap\mathcal{S}_n^\delta, \tag{4.17}$$

and $\mathcal{B}_n^\delta(x^n)$, $\mathcal{B}_n^\delta(x^n,u^n)$ as sections of $\mathcal{B}_n^\delta$ corresponding to the sequence x^n and on the pair (x^n,u^n), respectively. Note that $\mathcal{B}_n^\delta(x^n)$ can be the empty set. A similar statement can be made about $\mathcal{B}_n^\delta(x^n,u^n)$. The following lemma is useful for analyzing the coding scheme that is presented in the next subsection.

Lemma 4.1 *Assume that $X^n\sim p_X^n$. Generate M codewords $u^n(j)$ iid according to the marginal density p_U, where $M\ge 2^{nR_U}$. For each $u^n(j)$ draw L codewords $v^n(j,l)$ via the conditional density $p_{V|U}$, where $L\ge 2^{nR_V}$. Then for a given δ, where $0<\delta<1$,*

$$\Pr\Big\{(X^n,U^n(j),V^n(j,l))\notin\mathcal{B}_n^\delta,\ \forall j,l\Big\}\to 0 \tag{4.18}$$

as $n\to\infty$ if $R_U\ge I(X;U)+4\delta$ and $R_V\ge I(X;V|U)+5\delta$.

The proof of Lemma 4.1 is deferred to Appendix 4.A.

4.2.B A coding scheme

As in the discrete case we begin with the codebook construction. Given a $\delta\in(0,1)$, whose value is determined later, fix a conditional density $p_{UV|X}$ and a measurable mapping $g\colon\mathbb{R}^3\to\mathbb{R}$ such that

$$\mathbb{E}[d(X,g(U,V,Y))]=D<\infty,\ \text{and}\ I(Y;V|U)>0. \tag{4.19}$$

We will discuss the degenerate case where $I(Y;V|U)=0$ at the end of this subsection. We note that since the distortion measure is the squared error distance, there exists a measurable quantization mapping $f\colon\hat{\mathcal{X}}\to\{\hat{x}_i\}_{i=1}^N\subset\hat{\mathcal{X}}$, with N sufficiently large and $\hat{\mathcal{X}}=\mathbb{R}$ such that [Wyn78, Eq. 2.11]

$$\hat{D}=\mathbb{E}[d(X,f(g(U,V,Y)))]\le(1+\delta)D. \tag{4.20}$$

[4.6] Hence $\mathcal{B}_n^\delta$ is a Borel measurable set as it is the intersection of two measurable sets, cf. Footnotes 4.4 and 4.5.

Define[4.7]

$$\hat{g} = f \circ g. \tag{4.21}$$

With abuse of notation, we define $\mathcal{B}_n^\delta$ as before with $\hat{g}$ in place of g and $\hat{D}$ in place of D.

Additionally[4.8], we show in the following that there exist a deterministic mapping and an auxiliary random variable which produce the same effect as drawing an element from a set uniformly at random. We use the mapping and random variable in our formal coding scheme to show that the resulting mappings are measurable. Let $\mathcal{T}$ be the set of of all pairs (i, j) where $i \in [1 : 2^{nR_U}]$ and $j \in [1 : 2^{n\bar{R}_V}]$. The corresponding power set is $2^{\mathcal{T}}$. For each set $\mathcal{E} \in 2^{\mathcal{T}}$, we select one element of $\mathcal{E}$ uniformly at random if $\mathcal{E} \neq \varnothing$. Otherwise, we select one element of $\mathcal{T}$ uniformly at random. The corresponding conditional pmf is given by $\{P_{\mathcal{E}}(t) \mid t \in \mathcal{T}\}$. For each n by the functional representation lemma [EK11, Appendix B] there exists a discrete random variable $\hat{T}$, defined on the corresponding finite alphabet $\hat{\mathcal{T}}$, and a function $\hat{\psi} \colon 2^{\mathcal{T}} \times \hat{\mathcal{T}} \to \mathcal{T}$ such that

$$\hat{\psi}(\mathcal{E}, \hat{T}) \sim P_{\mathcal{E}}, \ \forall \mathcal{E} \in 2^{\mathcal{T}}. \tag{4.22}$$

Codebook generation: We generate a *single* codebook for all users which consists of 2^{nR_U} iid sequence $u^n(j)$ from the marginal pdf p_U. For each j, $2^{n\bar{R}_V}$ codewords $v^n(j, l)$ are drawn iid from the conditional pdf $p_{V|U}$. Each index l is parsed into a unique pair (k, k'), where $k \in [1 : 2^{nR_V}]$, $k' \in [1 : 2^{nR'_V}]$ and $\bar{R}_V = R_V + R'_V$, i.e., k is the corresponding bin index of l where the bin is given as in (3.16). We also fix two sequences u_e^n and v_e^n corresponding to the error message $\{e\}$.

Enrollment: Given $x^n(i)$ where $i \in \mathcal{M}$, we search for the set $\mathcal{I}_i$ which is determined as

$$\mathcal{I}_i = \left\{ (j_i, l_i) \mid (x^n(i), u^n(j_i), v^n(j_i, l_i)) \in \mathcal{B}_n^\delta, \ j_i \in [1 : 2^{nR_U}], \ l_i \in [1 : 2^{n\bar{R}_V}] \right\}. \tag{4.23}$$

If the set $\mathcal{I}_i$ is not empty then we select a tuple (j_i, l_i) uniformly at random from $\mathcal{I}_i$. Otherwise, (j_i, l_i) is selected uniformly from the set of all pairs $\mathcal{T}$. Formally the action is described by $\hat{\psi}(\mathcal{I}_i, \hat{t})$ as in (4.22) where $\hat{t}$ is the corresponding realization of $\hat{T}$. We store j_i in the first layer and the bin index k_i in the second layer[4.9].

Identification and Reconstruction: The two stage identification works similarly as

[4.7] Note that $\hat{g}$ is a measurable mapping since it is a composition of two measurable mappings.

[4.8] By our restrictions, all mappings are required to be deterministic and measurable. However, in our proof we use randomization in the encoding step to simplify the analysis. Hence, the existence of the mapping and the auxiliary random variable allow us to perform derandomization in the last step. Moreover, the output sequence is also a random vector since it is the output of the combination of deterministic transformations whose inputs are random vectors.

[4.9] We note that this encoding scheme is different from the one in Section 3.3 since the first layer message j_i is chosen after searching through codeword sequences in all layers. In contrast, in the discrete case the stored index in the first layer of the i-th user is chosen based only on the codewords in the first layer $\boldsymbol{u}^n$.

in the discrete case with the following modification. Condition (3.18) is replaced by

$$(z^n, u^n(j_i)) \in \mathcal{A}_\delta^n(ZU). \tag{4.24}$$

Condition (3.19) is replaced by searching for a unique $\hat{w}$ such that

$$(y^n, u^n(j_{\hat{w}}), v^n(j_{\hat{w}}, \tilde{l})) \in \mathcal{A}_\delta^n(YUV). \tag{4.25}$$

for some $\tilde{l} \in \mathfrak{B}(k_{\hat{w}})$. Condition (3.20) is changed to searching for a unique $\tilde{l} \in \mathfrak{B}(k_{\hat{w}})$ when $\hat{w} \neq e$ such that

$$(y^n, u^n(j_{\hat{w}}), v^n(j_{\hat{w}}, \tilde{l})) \in \mathcal{A}_\delta^n(YUV). \tag{4.26}$$

If $\hat{w} = e$ we set $\tilde{l} = 1$. When $\hat{w} = e$ or $\tilde{l} = e$, we set $u^n(j_{\hat{w}}) = u_e^n$ and $v^n(j_{\hat{w}}, \tilde{l}) = v_e^n$. Then the processing center outputs the corresponding sequence $\hat{x}_\tau = \hat{g}(u_\tau(j_{\hat{w}}), v_\tau(j_{\hat{w}}, \tilde{l}), y_\tau)$ for all $\tau = [1 : n]$ where $\hat{g}$ is defined in (4.21).

Properness of our coding scheme:

Roughly speaking, in each of the aforementioned steps the action consists of a combination of mappings whose pre-image of a Borel set is a finite intersections, or/and unions, of Borel sets. Hence the resulting mappings are measurable. The details are given in the following. We only need to show that mappings whose input arguments contain elements of $\mathbb{R}$ are measurable[4.10]. For notation brevity we define $\Xi = 1 + 2^{n(R_U + \bar{R}_V)}$, $\Upsilon = 1 + 2^{nR_U}$, $\boldsymbol{u}^n = (u^n(1), \ldots, u^n(2^{nR_U}))$ and $\boldsymbol{v}^n = (v^n(1,1), \ldots, v^n(2^{nR_U}, 2^{n\bar{R}_V}))$.

- We first show that the mappings from the users' data sequences and codebook to the stored indices are jointly measurable. For the sake of clarity, we focus on the first user. Consider the set of mappings $\{\psi_{i,j}\}$ where $i \in [1 : 2^{nR_U}]$ and $j \in [1 : 2^{n\bar{R}_V}]$ each is defined as

$$\psi_{i,j} \colon \mathbb{R}^{n \times \Xi} \to \{*, (i,j)\}$$
$$\psi_{i,j}(x^n(1), \boldsymbol{u}^n, \boldsymbol{v}^n) \mapsto \begin{cases} (i,j) & \text{if } (x^n(1), u^n(i), v^n(i,j)) \in \mathcal{B}_n^\delta \\ * & \text{otherwise} \end{cases}$$

  where $*$ is a dummy symbol. Then each $\psi_{i,j}$ is a measurable mapping since the pre-image

$$\psi_{i,j}^{-1}((i,j)) = \{(x^n(1), \boldsymbol{u}^n, \boldsymbol{v}^n) \mid (x^n(1), u^n(i), v^n(i,j)) \in \mathcal{B}_n^\delta, \text{ other codewords take values in } \mathbb{R}^n\}, \tag{4.27}$$

  is a Borel set. Hence the map

$$\psi = (\psi_{1,1}, \ldots, \psi_{2^{nR_U}, 2^{n\bar{R}_V}}) \colon \mathbb{R}^{n \times \Xi} \to \bigtimes_{i,j} \{*, (i,j)\}$$

[4.10] Mappings which map finite input alphabets to finite output alphabets are obviously measurable since the corresponding Borel σ-algebras are power sets.

$$\psi(x^n(1), \boldsymbol{u}^n, \boldsymbol{v}^n) \mapsto \hat{\mathcal{I}}_1, \tag{4.28}$$

is a measurable mapping. The one-to-one correspondence π_0 between the vector $\hat{\mathcal{I}}_1$ and the set of suitable pairs $\mathcal{I}_1$ given in (4.23) by eliminating all $*$, e.g., for $\hat{\mathcal{I}}_1 = \big(*, (1,2), *, (1,4), *, \ldots, *\big)$

$$\mathcal{I}_1 = \pi_0(\hat{\mathcal{I}}_1) = \{(1,2), (1,4)\}, \tag{4.29}$$

is obviously measurable. Let $\mathcal{T}$, $\hat{\psi}$ and $\hat{\mathcal{T}}$ be defined as in (4.22) then the map

$$\begin{aligned} \phi\colon \mathbb{R}^{n\times\Xi} \times \hat{\mathcal{T}} &\to \mathcal{T} \\ \phi(x^n(1), \boldsymbol{u}^n, \boldsymbol{v}^n, \hat{t}) = \hat{\psi}(\pi_0(\psi(x^n(1), \boldsymbol{u}^n, \boldsymbol{v}^n)), \hat{t}) &\mapsto (j_1, l_1), \end{aligned} \tag{4.30}$$

which is our selection map, is

$$\big(\mathbb{R}^{n\times\Xi} \times \hat{\mathcal{T}}, \mathcal{B}(\mathbb{R}^{n\times\Xi}) \times 2^{\hat{\mathcal{T}}}\big) \to (\mathcal{T}, 2^{\mathcal{T}})$$

measurable. We note that the mappings from the chosen pair to the stored pair are projections, hence measurable. In summary we show that the encoding mappings are measurable.

- To show that forming the list induces a measurable mapping, consider the following set of mappings $\{\hat{g}_{1i}\}_{i=1}^{M}$ where for each i, $\hat{g}_{1i}$ is defined as

$$\begin{aligned} \hat{g}_{1i}\colon \mathbb{R}^{n\times\Upsilon} \times \mathcal{M}_1^M &\to \{*, i\} \\ \hat{g}_{1i}(z^n, \boldsymbol{u}^n, \boldsymbol{j}) &\mapsto \begin{cases} i & \text{if } (z^n, u^n(j_i)) \in \mathcal{A}_\delta^n(ZU) \\ * & \text{otherwise} \end{cases}. \end{aligned} \tag{4.31}$$

Since $\mathcal{A}_n^\delta(ZU)$ is a Borel set, it can be seen that the map

$$\begin{aligned} \hat{g}_1 = (\hat{g}_{11}, \ldots, \hat{g}_{1M})\colon \mathbb{R}^{n\times\Upsilon} \times \mathcal{M}_1^M &\to \hat{\mathfrak{L}} = \bigtimes_{i=1}^{M} \{*, i\} \\ \hat{g}_1(z^n, \boldsymbol{u}^n, \boldsymbol{j}) &\mapsto \hat{\mathcal{L}}. \end{aligned} \tag{4.32}$$

is jointly measurable. Next, let π_1 be defined as

$$\begin{aligned} \pi_1\colon \hat{\mathfrak{L}} &\to \mathfrak{L} \\ \pi_1(\hat{\mathcal{L}}) &\mapsto \begin{cases} \mathcal{L} & \text{by eliminating all } * \text{ and if } 1 \le |\mathcal{L}| \le 2^{n\Delta} \\ \{e\} & \text{otherwise} \end{cases}. \end{aligned}$$

Since π_1 is a mapping from a discrete set to another discrete set, it is measurable w.r.t. the power set σ-algebras. Hence the map $\bar{g}_1 = \pi_1 \circ \hat{g}_1$ is a jointly measurable on

$$(\mathbb{R}^{n\times\Upsilon} \times \mathcal{M}_1^M, \mathcal{B}(\mathbb{R}^{n\times\Upsilon}) \times 2^{\mathcal{M}_1^M}) \to (\mathfrak{L}, 2^{\mathfrak{L}}).$$

Our first stage processing map g_1 can be obtain from $\bar{g}_1$ once a set of codewords is fixed.

- Similarly, for user identification in the second stage we look at the following set mappings $\{\hat{g}_{2i}\}_{i=1}^{M}$, whereas each is defined as

$$\hat{g}_{2i}\colon \mathbb{R}^{n\times\Xi}\times\mathfrak{M}_{12}\to\{*,i\}$$
$$\hat{g}_{2i}(y^n,\boldsymbol{u}^n,\boldsymbol{v}^n,(j_{\mathcal{L}},k_{\mathcal{L}},\mathcal{L}))\mapsto\begin{cases} i & \text{if } i\in\mathcal{L} \text{ and } (y^n,u^n(j_i),v^n(j_i,l))\in\mathcal{A}_\delta^n(YUV)\\ & \text{for some } l\in\mathfrak{B}(k_i)\\ * & \text{otherwise}\end{cases}.$$

We observe that for each i the mapping $\hat{g}_{2i}$ is jointly measurable. Next we need the mapping

$$\pi_2\colon \bigtimes_{i=1}^{M}\{*,i\}\to\mathcal{W}\cup\{e\}$$
$$\pi_2(\boldsymbol{\alpha})\mapsto\begin{cases}\hat{w} & \text{if it is the only non-}* \text{ element in } \boldsymbol{\alpha}\\ e & \text{otherwise}\end{cases}.$$

The second stage identification mapping g_2 can be obtained from $\bar{g}_2=\pi_2\circ\big((\hat{g}_{2i})_{i=1}^{M}\big)$ once a set of codewords is fixed.

- Finally, to describe the reconstruction mapping g_3 we need mappings $\hat{g}_3$ and π_3 which are defined in the following. Let $\hat{\mathfrak{M}}'_{12}=\mathcal{M}_1\times\big\{[1:2^{n\tilde{R}_V}]\cup\{e\}\big\}\times\big(\mathcal{W}\cup\{e\}\big)$. The mapping $\hat{g}_3$ searches for the unique second layer index $\tilde{l}$ of the chosen user $\hat{w}$, which has the bin index $k_{\hat{w}}$, and is defined formally as

$$\hat{g}_3\colon \mathbb{R}^{n\times\Xi}\times\hat{\mathfrak{M}}_{12}\to\hat{\mathfrak{M}}'_{12}$$
$$\hat{g}_3(y^n,\boldsymbol{u}^n,\boldsymbol{v}^n,(j_{\hat{w}},k_{\hat{w}},\hat{w}))\mapsto\begin{cases}(j_{\hat{w}},\tilde{l},\hat{w}) & \text{if } \hat{w}\neq e \text{ and } \tilde{l} \text{ is unique such that}\\ & (y^n,u^n(j_{\hat{w}}),v^n(j_{\hat{w}},\tilde{l}))\in\mathcal{A}_\delta^n(YUV)\\ & \text{as well as } \tilde{l}\in\mathfrak{B}(k_{\hat{w}})\\ (1,e,\hat{w}) & \text{if } \hat{w}\neq e\\ (1,1,e) & \text{if } \hat{w}=e\end{cases}$$

where $\hat{\mathfrak{M}}_{12}$ is defined in (3.5). The mapping π_3 outputs the corresponding codeword pair $(u^n(j_{\hat{w}}),v^n(j_{\hat{w}},\tilde{l}))$ given the input tuple $(j_{\hat{w}},\tilde{l},\hat{w})$ and the codebook. It is defined as

$$\pi_3\colon \mathbb{R}^{n\times(\Xi-1)}\times\hat{\mathfrak{M}}'_{12}\to\mathbb{R}^{2n}$$
$$\pi_3(\boldsymbol{u}^n,\boldsymbol{v}^n,(j_{\hat{w}},\tilde{l},\hat{w}))\mapsto\begin{cases}(u^n(j_{\hat{w}}),v^n(j_{\hat{w}},\tilde{l})) & \text{if } \hat{w}\neq e \text{ and } \tilde{l}\neq e\\ (u_e^n,v_e^n) & \text{otherwise}\end{cases}. \tag{4.33}$$

The measurable properties of $\hat{g}_3$ and π_3 can be shown similarly as the ones of $\hat{g}_1$ and $\hat{g}_2$. The reconstruction mapping g_3 can be obtained from $\bar{g}_3(\cdot,y^n)=\hat{g}(\pi_3(\cdot,\hat{g}_3(\cdot)),y^n)$, where $\hat{g}$, which has a finite output alphabet and is defined in (4.21), is applied symbolwisely.

Analysis: Let J_i and L_i, $i \in \mathcal{W}$, be the chosen indices for the i-th user. Furthermore, let $\boldsymbol{\mathcal{L}}_1$ be the list of indices $i \in \mathcal{W}$ that satisfy (4.24) in the first stage of the identification process.
Denoted by $\boldsymbol{\mathcal{H}}$ the random variable which represents the randomly generated codebook, i.e.,

$$\boldsymbol{\mathcal{H}} = \Big\{ (U^n(j), V^n(j,l)) \mid j \in [1 : 2^{nR_U}], l \in [1 : 2^{n\bar{R}_V}] \Big\}, \tag{4.34}$$

and its realization by $\mathcal{H}$. The Markov relation

$$(Y^n, Z^n) - X^n(W) - (W, J_W, L_W, \boldsymbol{\mathcal{H}})$$

follows by our coding scheme. However, for the error analysis we need the Markov relation in form of density terms.

Claim 1 *For each triple (w, j_w, l_w), the function*

$$\begin{aligned} &p_{X^n(W)Y^n\boldsymbol{\mathcal{H}}|J_W L_W W}(x^n, y^n, \mathcal{H}|j_w, l_w, w) \\ &= \frac{\Pr\{J_w = j_w, L_w = l_w, W = w | X^n(W) = x^n, \boldsymbol{\mathcal{H}} = \mathcal{H}\}}{P_{J_W L_W W}(j_w, l_w, w)} \\ &\times p_X^n(x^n) p_{Y|X}^n(y^n|x^n) p_{\boldsymbol{\mathcal{H}}}(\mathcal{H}) \end{aligned} \tag{4.35a}$$

is a conditional density function of the distribution

$$\mu(B, w, j_w, l_w) = \Pr\big\{(X^n(W), Y^n, \boldsymbol{\mathcal{H}}) \in B | J_w = j_w, L_w = l_w, W = w\big\}$$

w.r.t. the product of Lebesgue measures $\lambda^{\otimes n(1+\Xi)}$, where $B \in \mathcal{B}\big(\mathbb{R}^{n\times(1+\Xi)}\big)$ is a Borel set. It can also be argued that this function is jointly measurable in $(x^n, y^n, \mathcal{H}, j_w, l_w, w)$.

Proof. It is immediate from the definition of $p_{X^n(W)Y^n\boldsymbol{\mathcal{H}}|J_W L_W W}$ in (4.35a) that it is a jointly measurable function in $(x^n, y^n, \mathcal{H})$. Lemma 2.13 implies the following relation

$$\begin{aligned} &\Pr\{J_w = j_w, L_w = l_w, W = w | X^n(W) = x^n, \boldsymbol{\mathcal{H}} = \mathcal{H}\} \\ &= \Pr\{J_w = j_w, L_w = l_w, W = w | X^n(W) = x^n, Y^n = y^n, \boldsymbol{\mathcal{H}} = \mathcal{H}\} \\ &\quad P_{X^n(W)Y^n\boldsymbol{\mathcal{H}}} - \text{a.s.} \end{aligned} \tag{4.36}$$

Hence by integrating $p_{X^n(W)Y^n\boldsymbol{\mathcal{H}}|J_W L_W W}$, defined as in (4.35a), on each set Borel set B and using the relation (4.36) as well as the definition of conditional probability we obtain the conclusion. We further note that since our encoding procedure is identical among users and W is independent of users' data and the encoding process, we obtain

$$p_{X^n(W)Y^n\boldsymbol{\mathcal{H}}|J_W L_W W}(x^n, y^n, \mathcal{H}|j_w, l_w, w)$$

$$= \frac{\Pr\{J_1 = j_w, L_1 = l_w | X^n(1) = x^n, \boldsymbol{\mathcal{H}} = \mathcal{H}\}}{P_{J_1 L_1}(j_w, l_w)} \times p_X^n(x^n) p_{Y|X}^n(y^n|x^n) p_{\boldsymbol{\mathcal{H}}}(\mathcal{H}). \tag{4.37}$$

□

Note further that as $\Pr\{J_w = j, L_w = l | X^n(w) = x^n, \boldsymbol{\mathcal{H}} = \mathcal{H}\} = 0$ for some combinations of data sequence, observation and codebook for the w-th user, the corresponding density value is zero.

For brevity, we denote herein again by $\tilde{Y}^n$ the pair (Y^n, Z^n), by $\boldsymbol{\psi}_n$ the random variable $\psi_n(X^n(W), \tilde{Y}^n, U^n(J_W), V^n(J_W, L_W))$ and by P_{cp} the distribution $P_{X^n(W)U^n(J_W)V^n(J_W,L_W)WJ_WL_W}$. Additionally, we define

$$\begin{aligned} \boldsymbol{\chi}_{\mathcal{B}_n} &= \chi_{\mathcal{B}_n}(X^n(W), U^n(J_W), V^n(J_W, L_W)) \\ \boldsymbol{\chi}_{\mathcal{B}_n^c} &= 1 - \boldsymbol{\chi}_{\mathcal{B}_n}, \end{aligned} \tag{4.38}$$

where herein $\mathcal{B}_n$ is also a short notation for $\mathcal{B}_n^\delta$. We first notice that since $\psi_n(\cdot) \in \{0, 1\}$,

$$\begin{aligned} \mathbb{E}[\boldsymbol{\psi}_n] &= \mathbb{E}[\boldsymbol{\chi}_{\mathcal{B}_n^c} \boldsymbol{\psi}_n] + \mathbb{E}[\boldsymbol{\chi}_{\mathcal{B}_n} \boldsymbol{\psi}_n] \\ &\leq \Pr\{(X^n(W), U^n(J_W), V^n(J_W, L_W)) \notin \mathcal{B}_n\} + \mathbb{E}[\boldsymbol{\chi}_{\mathcal{B}_n} \boldsymbol{\psi}_n], \end{aligned} \tag{4.39}$$

With the help[4.11] of (4.35a) the second term can be bounded as

$$\begin{aligned} \mathbb{E}[\boldsymbol{\chi}_{\mathcal{B}_n} \boldsymbol{\psi}_n] &= \int \chi_{\mathcal{B}_n}(x^n, u^n, v^n) \times \mathbb{E}\big[\psi_n(x^n, u^n, v^n, \tilde{Y}^n) | X^n(w) = x^n, \\ &\qquad U^n(j_w) = u^n, V^n(j_w, l_w) = v^n, W = w, J_w = j_w, L_w = l_w\big] dP_{\text{cp}} \\ &= \int \chi_{\mathcal{B}_n}(x^n, u^n, v^n) \mathbb{E}[\psi_n(x^n, u^n, v^n, \tilde{Y}^n) | X^n(w) = x^n, W = w] dP_{\text{cp}} \\ &= \int \chi_{\mathcal{B}_n}(x^n, u^n, v^n) \eta_{XUV}(x^n, u^n, v^n) dP_{\text{cp}} \\ &\overset{(a)}{\leq} \delta_n^{1/2}, \end{aligned} \tag{4.40}$$

where (a) holds since given $(x^n, u^n, v^n) \in \mathcal{B}_n$, we have $\eta_{XUV}(x^n, u^n, v^n) \leq \delta_n^{1/2}$. Due to the symmetry of the problem we obtain

$$\begin{aligned} &\Pr\{(X^n(W), U^n(J_W), V^n(J_W, L_W)) \notin \mathcal{B}_n\} \\ &\qquad = \Pr\{(X^n(1), U^n(J_1), V^n(J_1, L_1)) \notin \mathcal{B}_n\} \end{aligned} \tag{4.41}$$

as W is independent of the enrollment process. Moreover, by our encoding rule we have

$$\{\omega \in \Omega \mid (X^n(1), U^n(J_1), V^n(J_1, L_1)) \notin \mathcal{B}_n\}$$

[4.11] See also the disintegration arguments in Corollary 2.2

$$= \{\omega \in \Omega \mid (X^n(1), U^n(j_1), V^n(j_1, l_1)) \notin \mathcal{B}_n, \ \forall j_1, l_1\} \tag{4.42}$$

Then by Lemma 4.1

$$\Pr\{(X^n(W), U^n(J_W), V^n(J_W, L_W)) \notin \mathcal{B}_n\} \to 0, \tag{4.43}$$

as $n \to \infty$ if

$$R_U \geq I(X;U) + 4\delta, \quad R_V + R'_V \geq I(X;V|U) + 5\delta. \tag{4.44}$$

Hence

$$\mathbb{E}[\boldsymbol{\psi}_n] \to 0, \text{ as } n \to \infty. \tag{4.45}$$

This implies that $(\tilde{Y}^n, U^n(J_W), V^n(J_W, L_W)) \in \mathcal{A}_\delta^n(YUV)$ with high probability, i.e.,

$$\Pr\{W \notin \boldsymbol{\mathcal{L}_1}\} \to 0, \text{ as } n \to \infty. \tag{4.46}$$

As in the discrete case we consider the following events

$$\begin{aligned}
\mathcal{E}_1 &= \{|\boldsymbol{\mathcal{L}}_1| > 2^{n\Delta}\}, \\
\mathcal{E}_2 &= \left\{(U^n(J_W), V^n(J_W, \tilde{l}), Y^n) \in \mathcal{A}_\delta^n(UVY), \text{ for some } \tilde{l} \neq L_W, \ \tilde{l} \in \mathfrak{B}(K_W)\right\}, \\
\mathcal{E}_3 &= \Big\{\exists (w', \tilde{l}), w' \neq W, w' \in \boldsymbol{\mathcal{L}}_1, \\
&\qquad (Y^n, U^n(J_{w'}), V^n(J_{w'}, \tilde{l})) \in \mathcal{A}_\delta^n(YUV), \ \tilde{l} \in \mathfrak{B}(K_{w'})\Big\}.
\end{aligned} \tag{4.47}$$

To bound the probability of the event $\mathcal{E}_1$ we only need to verify (3.28) for $i \geq 2$, which is expressed in our case as

$$\begin{aligned}
\Pr\{B_i | W = 1\} &= \int_{u^n, j_i} \int_{\mathcal{A}_\delta^n(Z|u^n)} p_{Z^n|W}(z^n|1) dz^n dP_{U^n(J_i)J_i}(u^n, j_i) \\
&\leq \int_{u^n, j_i} \int_{\mathcal{A}_\delta^n(Z|u^n)} 2^{-n(h(Z)-\delta)} dz^n dP_{U^n(J_i)J_i}(u^n, j_i) \\
&\leq 2^{-n(I(Z;U)-3\delta)}.
\end{aligned} \tag{4.48}$$

Therefore as in the discrete case $\Pr\{\mathcal{E}_1\} \to 0$ if $R - \Delta < I(Z;U) - 3\delta$. The analysis in (3.29) can be carried out similarly and we obtain the condition $R + R'_V < I(Y;U,V) - \delta$, which is needed for $\Pr\{\mathcal{E}_3\} \to 0$. This further leads to

$$\Pr\{\hat{W} \neq W\} \to 0. \tag{4.49}$$

Hence, we only need to bound the probability of the second event $\mathcal{E}_2$.

We use the same technique as the one in [MLK15, Lemma 1]. Due to the symmetry of the codebook construction and the encoding process, it is sufficient to condition on the following event[4.12]

$$\{J_1 = 1, L_1 = 1, W = 1\}. \tag{4.50}$$

We also assume that $l_1 = 1$ belongs to $\mathfrak{B}(1)$. Then due to the union bound and symmetry

$$\begin{aligned}
&\Pr\{\mathcal{E}_2|J_1 = 1, L_1 = 1, W = 1\} \\
&\leq \sum_{\tilde{l}\in\mathfrak{B}(1),\, \tilde{l}\neq 1} \Pr\{(U^n(1), V^n(1,\tilde{l}), Y^n) \in \mathcal{A}_\delta^n | J_1 = 1, L_1 = 1, W = 1\} \\
&\leq 2^{nR'_V} \Pr\{(U^n(1), V^n(1,2), Y^n) \in \mathcal{A}_\delta^n | J_1 = 1, L_1 = 1, W = 1\}.
\end{aligned} \tag{4.51}$$

The probability term in the right-hand side of (4.51) can be factorized[4.13]

$$\begin{aligned}
&\Pr\{(U^n(1), V^n(1,2), Y^n) \in \mathcal{A}_\delta^n | J_1 = 1, L_1 = 1, W = 1\} \\
&= \int_{\mathcal{A}_\delta^n(UVY)} p_{U^n(1)V^n(1,2)Y^n|J_1L_1W}(u^n, v^n, y^n|1,1,1)du^n dv^n dy^n \\
&= \int_{\mathcal{A}_\delta^n(UVY)} \Bigg(\int p(U^n(1) = u^n, V^n(1,2) = v^n, Y^n = y^n, \\
&\qquad X^n(1) = x^n, V^n(1,1) = \tilde{v}^n | J_1 = 1, L_1 = 1, W = 1) dx^n d\tilde{v}^n \Bigg) du^n dv^n dy^n \\
&\overset{(\star)}{=} \int_{\mathcal{A}_\delta^n(UVY)} \Bigg(\int p_{Y|X}^n(y^n|x^n) \\
&\qquad \times p_{U^n(1)V^n(1,2)X^n(1)V^n(1,1)|J_1L_1W}(u^n, v^n, x^n, \tilde{v}^n|1,1,1) dx^n d\tilde{v}^n \Bigg) du^n dv^n dy^n \\
&\overset{(\star\star)}{=} \int p_{Y|X}^n(y^n|x^n) p(X^n(1) = x^n, U^n(1) = u^n, V^n(1,1) = \tilde{v}^n | J_1 = 1, L_1 = 1) \\
&\times \Bigg(\int_{\mathcal{A}_\delta^n(V|u^n,y^n)} p(V^n(1,2) = v^n | U^n(1) = u^n, V^n(1,1) = \tilde{v}^n, X^n(1) = x^n, \\
&\qquad J_1 = 1, L_1 = 1) dv^n \Bigg) dx^n dy^n du^n d\tilde{v}^n.
\end{aligned} \tag{4.52}$$

The equality in $(\star)$ holds according to the relation (4.35a). Since densities are non-negative, $(\star\star)$ holds due to Fubini's theorem and (4.37). For brevity, we denote

[4.12] For simplicity we drop the subscript for the index of the first user, i.e., the notation (j_1, l_1) is simplified as (j, l).

[4.13] Since $\mathcal{A}_\delta^n(UVY)$ is a Borel measurable set, we do not need to consider the complete measure space. We also use the notation $p_{X|Y}(x|y)$ and $p(X = x|Y = y)$ for probability density function interchangeably where the latter is handy when a long tuple of random variables is present in the expression.

$\mathcal{F} = \{U^n(1) = u^n, V^n(1,1) = \tilde{v}^n, X^n(1) = x^n\}$ and

$$\boldsymbol{\mathcal{C}} = \{V^n(1,l) \mid l \geq 3\} \bigcup \left\{U^n(j), V^n(j,l)\right\}_{j,l\big|j\geq 2}, \tag{4.53}$$

as the rest of the codebook[4.14]. For given $\boldsymbol{\mathcal{C}} = \mathcal{C}$ define

$$\begin{aligned} n(\mathcal{C},\mathcal{F}) = &\left|\{l \mid v^n(1,l) \in \mathcal{C},\ (u^n, v^n(1,l), x^n) \in \mathcal{B}_n\}\right| \\ &+ \left|\{(j,l) \mid j \geq 2, (u^n(j), v^n(j,l)) \in \mathcal{C},\ (u^n(j), v^n(j,l), x^n) \in \mathcal{B}_n\}\right|, \end{aligned} \tag{4.54}$$

which is a Borel measurable function, and

$$i(\mathcal{C},\mathcal{F}) = \begin{cases} 1 & \text{if } (x^n, u^n, \tilde{v}^n) \notin \mathcal{B}_n \text{ and } n(\mathcal{C},\mathcal{F}) = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4.55}$$

As a standard step, we further define[4.15] $\mathfrak{G} = \{\mathcal{C} \colon p(\boldsymbol{\mathcal{C}} = \mathcal{C}|\mathcal{F}, J_1 = 1, L_1 = 1) = 0\}$. Then

$$\Pr\{\boldsymbol{\mathcal{C}} \in \mathfrak{G}|\mathcal{F}, J_1 = 1, L_1 = 1\} = 0, \tag{4.56}$$

which implies that

$$\begin{aligned} &\int_{\mathcal{A}_\delta^n(V|u^n,y^n)} p(V^n(1,2) = v^n|\mathcal{F}, J_1 = 1, L_1 = 1)dv^n \\ &= \int_{\mathcal{A}_\delta^n(V|u^n,y^n)} \int_{\mathfrak{G}} p(V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}|\mathcal{F}, J_1 = 1, L_1 = 1)d\mathcal{C}dv^n \\ &\quad + \int_{\mathcal{A}_\delta^n(V|u^n,y^n)} \int_{\mathfrak{G}^c} p(V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}|\mathcal{F}, J_1 = 1, L_1 = 1)d\mathcal{C}dv^n \\ &\overset{(b)}{=} \int_{\mathcal{A}_\delta^n(V|u^n,y^n)} \int_{\mathfrak{G}^c} p(V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}|\mathcal{F}, J_1 = 1, L_1 = 1)d\mathcal{C}dv^n, \end{aligned} \tag{4.57}$$

where (b) is valid since (4.56) can be seen as the integration of $p(V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}|\mathcal{F}, J_1 = 1, L_1 = 1)$ over $\mathbb{R}^n \times \mathfrak{G}$, which implies that the first term in the above sum is zero. A similar line of reasoning can be applied to resolve the case where

$$p(X^n(1) = x^n, U^n(1) = u^n, V^n(1,1) = \tilde{v}^n|J_1 = 1, L_1 = 1) = 0, \tag{4.58}$$

in (4.52).
Additionally, consider the case that $(x^n, u^n, \tilde{v}^n) \notin \mathcal{B}_n$. Define[4.16]

$$\mathfrak{D} = \{\mathcal{C} : n(\mathcal{C},\mathcal{F}) > 0\}, \tag{4.59}$$

[4.14] More precisely, $\boldsymbol{\mathcal{C}}$ is a random vector in which components are $V^n(1,l)$ where $l \geq 3$ and $(U^n(j), V^n(j,l))$ for $j \geq 2$ arranged in the presented order.

[4.15] From its definition $\mathfrak{G}$ is a Borel measurable set. In more details, due to the restriction (4.58) $\mathfrak{G}$ is the $(x^n, u^n, \tilde{v}^n)$-section of the measurable set $\bar{\mathfrak{G}} = \{(\mathcal{C}, x^n, u^n, \tilde{v}^n) \mid p(\mathcal{F}|J_1 = 1, L_1 = 1) > 0 \text{ and } p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}|J_1 = 1, L_1 = 1) = 0\}$. This implies that the inner integral over $\mathfrak{G}$ in (4.57) produces a measurable function in $(x^n, u^n, \tilde{v}^n, v^n)$.

[4.16] More specifically, $\mathfrak{D}$ is the $(x^n, u^n, \tilde{v}^n)$-section of the measurable set $\bar{\mathfrak{D}} = \{(\mathcal{C}, x^n, u^n, \tilde{v}^n) \mid (x^n, u^n, \tilde{v}^n) \notin \mathcal{B}_n \text{ and } n(\mathcal{C},\mathcal{F}) > 0\}$.

which is a Borel set. Then due to our encoding rule

$$\Pr\{\boldsymbol{\mathcal{C}} \in \mathfrak{D} | \mathcal{F}, J_1 = 1, L_1 = 1\} = 0, \tag{4.60}$$

which leads to

$$\begin{aligned}
&\int_{\mathcal{A}_\delta^n(V|u^n,y^n)} p(V^n(1,2) = v^n | \mathcal{F}, J_1 = 1, L_1 = 1) dv^n \\
&= \int_{\mathcal{A}_\delta^n(V|u^n,y^n)} \int_{(\mathfrak{D}\cup\mathfrak{G})^c} p(V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C} | \mathcal{F}, J_1 = 1, L_1 = 1) d\mathcal{C} dv^n. \quad (4.61)
\end{aligned}$$

Therefore, to upper bound (4.52), by combining the arguments in (4.57) and (4.61), it is sufficient to consider the following inner integral

$$\begin{aligned}
&\int_{\mathfrak{C}} p(V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C} | \mathcal{F}, J_1 = 1, L_1 = 1) d\mathcal{C} \\
&= \int_{\mathfrak{C}} p(\boldsymbol{\mathcal{C}} = \mathcal{C} | \mathcal{F}, J_1 = 1, L_1 = 1) \\
&\quad \times p(V^n(1,2) = v^n | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}, J_1 = 1, L_1 = 1) d\mathcal{C} \quad (4.62a) \\
&\overset{(c)}{=} \int_{\mathfrak{C}} p(\boldsymbol{\mathcal{C}} = \mathcal{C} | \mathcal{F}, J_1 = 1, L_1 = 1) p(V^n(1,2) = v^n | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) \\
&\qquad \times \frac{\Pr\{J_1 = 1, L_1 = 1 | V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\}}{\Pr\{J_1 = 1, L_1 = 1 | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\}} d\mathcal{C}, \quad (4.62b)
\end{aligned}$$

where[4.17]

$$\mathfrak{C} = \begin{cases} (\mathfrak{D} \cup \mathfrak{G})^c & \text{if } (x^n, u^n, \tilde{v}^n) \notin \mathcal{B}_n \\ \mathfrak{G}^c & \text{if } (x^n, u^n, \tilde{v}^n) \in \mathcal{B}_n \end{cases}. \tag{4.63}$$

Note that in both cases, $\Pr\{J_1 = 1, L_1 = 1 | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\} > 0$. In Appendix 4.2.C we provide an argument to verify (c) in (4.62) independently for interested readers. Next, we have

$$p(V^n(1,2) = v^n | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) = \prod_{i=1}^n p_{V|U}(v_i | u_i) \tag{4.64}$$

due to our codebook generation. In addition, we bound the numerator term in (4.62b) as follows:

$$\begin{aligned}
&\underbrace{\Pr\{J_1 = 1, L_1 = 1 | V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\}}_{\gamma} \\
&\qquad \leq \underbrace{\frac{1}{2^{nR}} i(\mathcal{C}, \mathcal{F}) + \frac{1}{n(\mathcal{C}, \mathcal{F}) + 1} (1 - i(\mathcal{C}, \mathcal{F}))}_{\gamma'}, \quad (4.65)
\end{aligned}$$

where $R = R_U + \bar{R}_V$. We verify the above inequality by the following cases:

[4.17] It can be seen that $\mathfrak{C}$ is the $(x^n, u^n, \tilde{v}^n)$-section of $(\bar{\mathfrak{G}} \cup \bar{\mathfrak{D}})^c$.

- $(x^n, u^n, \tilde{v}^n) \notin \mathcal{B}_n$ then $n(\mathcal{C}, \mathcal{F}) = 0$ by our restriction which implies that $i(\mathcal{C}, \mathcal{F}) = 1$. We have
$$\gamma \leq \gamma' = \frac{1}{2^{nR}}, \tag{4.66}$$
with the equality when $(x^n, u^n, v^n) \notin \mathcal{B}_n$.

- $(x^n, u^n, \tilde{v}^n) \in \mathcal{B}_n$, i.e., $i(\mathcal{C}, \mathcal{F}) = 0$, we always have
$$\gamma \leq \gamma' = \frac{1}{n(\mathcal{C}, \mathcal{F}) + 1}, \tag{4.67}$$
with the equality when $(x^n, u^n, v^n) \notin \mathcal{B}_n$. The "+1" term in the denominator is due to the event $(x^n, u^n, \tilde{v}^n) \in \mathcal{B}_n$.

Moreover, the denominator in (4.62b) can be lower bounded as

$$\begin{aligned}
&\Pr\{J_1 = 1, L_1 = 1 | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\} \\
&\quad \geq \int_{\mathcal{B}_n^c(x^n, u^n)} \Pr\{J_1 = 1, L_1 = 1 | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}, V^n(1,2) = v^n\} \\
&\qquad \times p(V^n(1,2) = v^n | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) dv^n \\
&\quad = \left(\frac{1}{2^{nR}} i(\mathcal{C}, \mathcal{F}) + \frac{1}{n(\mathcal{C}, \mathcal{F}) + 1} (1 - i(\mathcal{C}, \mathcal{F})) \right) \\
&\qquad \times \Pr\{V^n(1,2) \notin \mathcal{B}_n(x^n, u^n) | U^n(1) = u^n\} \\
&\quad \geq \left(\frac{1}{2^{nR}} i(\mathcal{C}, \mathcal{F}) + \frac{1}{n(\mathcal{C}, \mathcal{F}) + 1} (1 - i(\mathcal{C}, \mathcal{F})) \right) \\
&\qquad \times \Pr\{V^n(1,2) \notin \mathcal{A}_\delta^n(V | u^n, x^n) | U^n(1) = u^n\}.
\end{aligned} \tag{4.68}$$

Now for sufficiently large n,

$$\begin{aligned}
\Pr\{V^n(1,2) \notin \mathcal{A}_\delta^n(V|u^n, x^n) | U^n(1) = u^n\} &= 1 - \int_{\mathcal{A}_\delta^n(V|u^n, x^n)} p_{V|U}^n(v^n | u^n) dv^n \\
&\geq 1 - 2^{-n(h(V|U) - 2\delta)} \int_{\mathcal{A}_\delta^n(V|u^n, x^n)} dv^n \\
&\geq 1 - 2^{-n(h(V|U) - 2\delta)} 2^{n(h(V|U,X) + 2\delta)} \\
&= 1 - 2^{-n(I(X;V|U) - 4\delta)}.
\end{aligned} \tag{4.69}$$

This analysis implies that when $\delta < I(X; V|U)/4$ and for sufficiently large n

$$\begin{aligned}
p(V^n(1,2) = v^n | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) \times \frac{\Pr\{J_1 = 1, L_1 = 1 | V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\}}{\Pr\{J_1 = 1, L_1 = 1 | \boldsymbol{\mathcal{C}} = \mathcal{C}\}} \\
\leq \frac{1}{1 - 2^{-n(I(X;V|U) - 4\delta)}} \prod_{i=1}^n p_{V|U}(v_i | u_i)
\end{aligned}$$

$$\overset{(e)}{\leq} (1+\hat{\epsilon})2^{-n(h(V|U)-2\delta)}, \tag{4.70}$$

where $\hat{\epsilon}$ is a fixed positive number. (e) holds since $v^n \in \mathcal{A}_\delta^n(V|u^n, y^n)$. Combining (4.57), (4.62b) and (4.70) we obtain the following upper bound

$$\begin{aligned}
&\int_{\mathcal{A}_\delta^n(V|u^n,y^n)} p(V^n(1,2) = v^n|\mathcal{F}, J_1 = 1, L_1 = 1)dv^n \\
&\leq \int_{\mathcal{A}_\delta^n(V|u^n,y^n)} (1+\hat{\epsilon})2^{-n(h(V|U)-2\delta)}dv^n \int_{\mathfrak{C}} p(\boldsymbol{\mathcal{C}} = \mathcal{C}|\mathcal{F}, J_1 = 1, L_1 = 1)d\mathcal{C} \\
&\leq (1+\hat{\epsilon})2^{-n(h(V|U)-2\delta)}2^{n(h(V|U,Y)+2\delta)} \\
&= (1+\hat{\epsilon})2^{-n(I(V;Y|U)-4\delta)}.
\end{aligned} \tag{4.71}$$

Hence, inserting the above inequality in (4.52) we obtain

$$\Pr\{(U^n(1), V^n(1,2), Y^n) \in \mathcal{A}_\delta^n | J_1 = 1, L_1 = 1, W = 1\} \leq (1+\hat{\epsilon})2^{-n(I(V;Y|U)-4\delta)}$$

and

$$\Pr\{\mathcal{E}_2\} \to 0 \text{ as } n \to \infty, \tag{4.72}$$

if $R'_V < I(V;Y|U) - 4\delta$ and $\delta < I(V;Y|U)/4$.
Lastly, we bound now the distortion level of the reconstruction sequence. Define

$$\phi_n = (1-\boldsymbol{\psi}_n)(1-\chi_{\mathcal{E}_1})(1-\chi_{\mathcal{E}_2})(1-\chi_{\mathcal{E}_3}), \tag{4.73}$$

and $\bar{\phi}_n = (1-\phi_n)$. We have the following simple inequality, which is actually the union bound,

$$\bar{\phi}_n \leq 1-(1-\boldsymbol{\psi}_n)(1-\chi_{\mathcal{E}_1})(1-\chi_{\mathcal{E}_2})+\chi_{\mathcal{E}_3} \leq \cdots \leq \boldsymbol{\psi}_n + \chi_{\mathcal{E}_1} + \chi_{\mathcal{E}_2} + \chi_{\mathcal{E}_3}. \tag{4.74}$$

Then $\mathbb{E}[\bar{\phi}_n] \to 0$ as $n \to \infty$. We notice that

$$\phi_n = 1 \implies \big\{\big|d(X^n(W), \hat{X}^n) - \hat{D}\big| \leq \delta\big\}, \tag{4.75}$$

where $\hat{X}^n = \hat{g}(U^n(J_{\hat{W}}), V^n(J_{\hat{W}}, \tilde{L}), Y^n)$. Therefore the distortion level can be upperbounded as

$$\begin{aligned}
\mathbb{E}\big[\big|d(X^n(W), \hat{X}^n) - \hat{D}\big|\big] &= \mathbb{E}\big[\phi_n\big|d(X^n, \hat{X}^n) - \hat{D}\big|\big] + \mathbb{E}\big[\bar{\phi}_n\big|d(X^n, \hat{X}^n) - \hat{D}\big|\big] \\
&\leq \delta + \mathbb{E}[\bar{\phi}_n\hat{D}] + \mathbb{E}[\bar{\phi}_n d(X^n, \hat{X}^n)].
\end{aligned} \tag{4.76}$$

The last term in (4.76) can be bounded using similar techniques as in [Wyn78, Lemma 5.1]. First note that

$$\mathbb{E}[\bar{\phi}_n d(X^n(W), \hat{X}^n)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\bar{\phi}_n d(X_i(W), \hat{X}_i)] \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\bar{\phi}_n \zeta(X_i(W))], \tag{4.77}$$

where $\zeta(X_i(W)) = \max_{\{\hat{x}\}_{k=1}^N} d(X_i(W), \hat{x}_k)$. We further observe that $\{\zeta(X_i(W))\}_{i=1}^n$ are iid $\sim P_{\zeta(X)}$ and integrable random variables. The latter statement holds due to the property of the square distortion measure and P_X. Then for all $i \in [1:n]$ the following is valid for any $a > 0$

$$\mathbb{E}[\bar{\phi}_n \zeta(X_i(W))] \leq a\mathbb{E}[\bar{\phi}_n] + \mathbb{E}[\zeta(X_i(W))\chi_{\{\zeta(X_i(W))\geq a\}}]. \tag{4.78}$$

Due to the monotone convergence theorem and the iid property we have

$$\mathbb{E}[\zeta(X_i(W))\chi_{\{\zeta(X_i(W))\geq a\}}] = \mathbb{E}[\zeta(X)\chi_{\{\zeta(X)\geq a\}}] \leq \delta, \ \forall i \tag{4.79}$$

for sufficiently large $a \geq a_0$ where a_0 depends only on $(d, P_X, \{\hat{x}_k\}_{k=1}^N)$. This implies that when $a \geq a_0$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\bar{\phi}_n \zeta(X_i(W))] \leq a\mathbb{E}[\bar{\phi}_n] + \delta \leq 2\delta \tag{4.80}$$

when $n \to \infty$. In conclusion we obtain

$$\mathbb{E}\big[\big|d(X^n(W), \hat{X}^n) - \hat{D}\big|\big] \leq 4\delta. \tag{4.81}$$

for sufficiently large n. Recall that the discrete random variable $\hat{T}$ is used to select a pair of indices (j_i, l_i) randomly, cf. (4.22), and and $\boldsymbol{\mathcal{H}}$ is the random codebook. Put $\hat{\delta} = 4\delta$, by using Markov's inequality with the threshold $4\hat{\delta}$, as in the proof of [BB11, Lemma 2.2], we have for all sufficiently large n

$$\begin{aligned}\big|\mathbb{E}[d(X^n(W), \hat{X}^n)|\boldsymbol{\mathcal{H}}, \hat{T}] - \hat{D}\big| &< \mathbb{E}\big[\big|d(X^n(W), \hat{X}^n) - \hat{D}\big||\boldsymbol{\mathcal{H}}, \hat{T}\big], \ \mathbb{P}-\text{a.s}, \\ \Pr\Big\{\mathbb{E}[\chi_{\{W\neq\hat{W}\}}|\boldsymbol{\mathcal{H}}, \hat{T}] &< 4\hat{\delta}, \ \mathbb{E}[\chi_{\{W\notin\boldsymbol{\mathcal{L}}\}}|\boldsymbol{\mathcal{H}}, \hat{T}] < 4\hat{\delta}, \\ &\mathbb{E}\big[\big|d(X^n(W), \hat{X}^n) - \hat{D}\big||\boldsymbol{\mathcal{H}}, \hat{T}\big] < 4\hat{\delta}\Big\} > 1/4, \end{aligned} \tag{4.82}$$

which implies the existence of a codebook $\mathcal{H}$ and an instance of randomness $\hat{t}$. Choosing δ small enough, we therefore arrive at the conditions in (3.35). The rest follows immediately.

Finally, we discuss about the case[4.18] when $I(Y;V|U) = 0$. We then have that

$$I(X;V|U,Y) = I(Y,X;V|U) = I(X;V|U), \tag{4.83}$$

as $Y - X - (U,V)$. The second constraint (4.2b) becomes

$$R_1 + R_2 \geq I(X;U,V).$$

[4.18] One needs to investigate the continuity problem of mutual information as function of distribution if one wants to use the previous result by taking the limit so that $I(Y;V|U)$ would go to 0. In our case a direct proof is far less complex than studying the continuity property.

The third constraint (4.2c) can be omitted as

$$R + I(X;U,V|Y) \le I(Y,X;U,V) = I(X;U,V), \tag{4.84}$$

where the first inequality holds since $R \le \min\{R_L + I(Z;U), I(Y;U,V)\}$. In summary we need to prove the following region is achievable

$$\begin{aligned} R_1 &\ge I(X;U) \\ R_1 + R_2 &\ge I(X;U,V) \\ R &\le \min\{R_L + I(Z;U), I(Y;U,V)\} \\ D &\ge \mathbb{E}[d(X, g(U,V,Y))]. \end{aligned} \tag{4.85}$$

The achievability of the region (4.85) can be proceeded in a similar manner as the one when $I(Y;V|U) > 0$. Namely, we need two layers of codewords $\boldsymbol{u}^n$ and $\boldsymbol{v}^n$. However binning is not used for the second layer. The reconstruction sequence is given by $\hat{g}(u^n(j_{\hat{w}}), v^n(j_{\hat{w}}, l_{\hat{w}}), y^n)$. In the analysis we simply omit the event $\mathcal{E}_2$ since no binning is used.
The following sub-region, which is useful in a later discussion, can be obtained by choosing V and g such that V is independent of everything else and $g\colon \mathbb{R}^2 \to \mathbb{R}$ such that

$$\begin{aligned} R_1 &\ge I(X;U) \\ R &\le \min\{R_L + I(Z;U), I(Y;U)\} \\ D &\ge \mathbb{E}[d(X, g(U,Y))]. \end{aligned} \tag{4.86}$$

4.2.C A detailed justification of (4.62)

The skeptic reader might be wary of the validity of (c) in (4.62) which is ensured by the following analysis. Let

$$\mathfrak{E} = \big\{(\mathcal{C}, x^n, u^n, \tilde{v}^n) \mid p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) > 0\big\}. \tag{4.87}$$

Then, we have $\Pr\big\{(\boldsymbol{\mathcal{C}}, X^n(1), U^n(1), V^n(1,1)) \in \mathfrak{E}^c\big\} = 0$. Therefore, we can modify the expression (4.52) as follows. The LHS of (4.52) is expanded to be an integral over $\boldsymbol{\mathcal{C}}$ and $\mathcal{A}_\delta^n(UVY)$ of the corresponding conditional density term. Then by restricting the integral on the set $\mathfrak{E}$ and following similar steps as in (4.52), the last integral in (4.52) is changed to $\int \int_{\mathcal{A}_\delta^n(V|u^n,y^n)} \int_{\mathfrak{E}(x^n,u^n,\tilde{v}^n)}$, where $\mathfrak{E}(x^n,u^n,\tilde{v}^n)$ is the corresponding section of $\mathfrak{E}$. The set $\mathfrak{C}$ in (4.62) can be modified to $\mathfrak{C}'$ which is the $(x^n,u^n,\tilde{v}^n)$-section of $(\bar{\mathfrak{G}} \cup \bar{\mathfrak{D}})^c \cap \mathfrak{E}$, cf. Footnote 4.15 and 4.16.

Claim 2 *Let $\mathbb{R}^\alpha$ be the product space of tuples $(\mathcal{C}, u^n, \tilde{v}^n, x^n, v^n)$ where $\alpha = n(4 + (2^{n\bar{R}_V} - 2 + (2^{nR_U} - 1)2^{n\bar{R}_V}))$ with the corresponding Borel σ-algebra $\mathcal{B}(\mathbb{R}^\alpha)$. Then*

$$p(V^n(1,2) = v^n | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}, J_1 = 1, L_1 = 1) \Pr\{J_1 = 1, L_1 = 1 | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\}$$

$$= p(V^n(1,2) = v^n | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) \times \Pr\{J_1 = 1, L_1 = 1 | V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\}$$

$\lambda^{\otimes\alpha}$-almost everywhere on $\{(\mathcal{C}, u^n, \tilde{v}^n, x^n, v^n) | p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) > 0\} \in \mathcal{B}(\mathbb{R}^\alpha)$. As a corollary, the conclusion also holds when we restrict to $v^n \in \mathcal{A}_\delta^n(V|u^n, y^n)$.

Proof. We first notice that $\Pr\{J_1 = 1, L_1 = 1 | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\} p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) = \Pr\{J_1 = 1, L_1 = 1\} p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F} | J_1 = 1, L_1 = 1)$, $\lambda^{\otimes(\alpha-n)}$-almost everywhere on $\mathbb{R}^{\alpha-n}$, hence also $\lambda^{\otimes\alpha}$-a.e. on $\mathbb{R}^\alpha$. This can be seen by integrating both sides w.r.t. $(\mathcal{C}, u^n, \tilde{v}^n, x^n)$ on any set $\hat{\mathcal{E}} \in \mathcal{B}(\mathbb{R}^{\alpha-n})$ and using the definition of conditional probability distribution. Then for any set $\mathcal{E} \in \mathcal{B}(\mathbb{R}^\alpha)$

$$\begin{aligned}
&\int_{\mathcal{E}} p(V^n(1,2) = v^n | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}, J_1 = 1, L_1 = 1) \\
&\quad \times \Pr\{J_1 = 1, L_1 = 1 | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\} p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) d\mathcal{C} du^n d\tilde{v}^n dx^n dv^n \\
&= P_{J_1 L_1}(1,1) \int_{\mathcal{E}} p(V^n(1,2) = v^n | \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}, J_1 = 1, L_1 = 1) \\
&\quad \times p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F} | J_1 = 1, L_1 = 1) d\mathcal{C} du^n d\tilde{v}^n dx^n dv^n \\
&\overset{(d)}{=} \Pr\{J_1 = 1, L_1 = 1, (\boldsymbol{\mathcal{C}}, U^n(1), V^n(1,1), X^n(1), V^n(1,2)) \in \mathcal{E}\} \\
&= \int_{\mathcal{E}} p(V^n(1,2) = v^n | \mathcal{C}, \mathcal{F}) \Pr\{J_1 = 1, L_1 = 1 | V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}\} \\
&\quad \times p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F}) d\mathcal{C} du^n d\tilde{v}^n dx^n dv^n.
\end{aligned}$$

In (d) we use the expression $p(V^n(1,2) = v^n, \boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F} | J_1 = 1, L_1 = 1) = p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F} | J_1 = 1, L_1 = 1) p(V^n(1,2) = v^n | \mathcal{F}, \boldsymbol{\mathcal{C}} = \mathcal{C}, J_1 = 1, L_1 = 1)$ which holds except on a zero probability set where $p(\boldsymbol{\mathcal{C}} = \mathcal{C}, \mathcal{F} | J_1 = 1, L_1 = 1) = 0$. The conclusion of the claim follows. □

Since, we are doing integration over $(u^n, y^n, v^n) \in \mathcal{A}_\delta^n(UVY)$ and $(\mathcal{C}, x^n, u^n, \tilde{v}^n) \in (\bar{\mathfrak{G}} \cup \bar{\mathfrak{D}})^c \cap \mathfrak{E}$, Claim 2 indicates that replacing (4.62a) by (4.62b) does not change the value of (4.51).

Remark 4.2 One might wonder if the overly complicated analysis of the probability of the event $\mathcal{E}_2$ could be avoidable by using the standard random binning trick. We explain in this remark that this passage is similarly complex. Assume that we did apply the random binning to generate K from L. Then we would have

$$\Pr\{\mathcal{E}_2 | W = 1\} \leq 2^{nR'_V} \Pr\{(Y^n, U^n(J_1), V^n(J_1, 1)) \in \mathcal{A}_\delta^n(YUV) | W = 1\}. \quad (4.88)$$

If J_1 would not depend on the second layer codewords, we would be easily done with the analysis as given $U^n(J_1)$, $V^n(J_1, 1)$ would be iid generated and independent of Y^n. However, it is not the case herein. The method mentioned in [EK11, Section 11.3] would have the same problem as it induces the same expression as the right-hand side of (4.88). We expect that a similar issue would arise in the analysis of the Heegard-Berger problem with side information at only one decoder.

4.3 Proof of Theorem 4.2

4.3.A Prelude

We first consider extreme cases which provide some points and hints about the whole rate-distortion region. The discussion is analogous to Remark 3.1 hence it might seem repetitive.

1. Our setup can be regarded as an extension of the Heegard-Berger [HB85] scheme and the Wyner-Ziv problem. Hence when additionally $M = 1$, the rate region collapses into

$$R_1 + R_2 \geq \frac{1}{2} \log_2 \frac{\sigma_X^2 \sigma_{N_1}^2}{\sigma_Y^2 D}. \tag{4.89}$$

2. For a given R_L the first term in the definition of R_γ is the first stage identification capacity $\frac{1}{2} \log_2 \frac{\sigma_Z^2}{\sigma_{N_1}^2 + \sigma_{N_2}^2} + R_L$ and the second term corresponds to the identification capacity $\frac{1}{2} \log_2 \frac{\sigma_Y^2}{\sigma_{N_1}^2}$ when the processing unit has the full access to both storage nodes.

Assume that we want to design an identification scheme such that a given tuple (R, R_1, R_2, R_L, D) is achievable in which the list size R_L is large enough such that $R_\gamma = 1/2 \log_2(\sigma_Y^2/\sigma_{N_1}^2)$. When the identification rate R is small, the distortion level D can be matched. However, when the identification rate R is close to the threshold R_γ, then the achieved distortion level by the identification scheme is likely to be lower than the requested distortion D. One can explain this observation as follows. In order for the identification rate to come close to the identification capacity R_γ, the compressed information must be *close* to the corresponding user's data, i.e., the distortion level for stored sequences will be extremely small and hence smaller than the requested level D. In other words, the distortion constraint in (4.2e) becomes inactive. This provides a hint that there will be a transition point from a region where the distortion constraint is active to a region where the distortion constraint is inactive when R increases. When the list size R_L is small or moderate, there exist additional transition points where the identification rate is limited at the first stage.

4.3.B An outerbound

Suppose that the rate-distortion tuple (R, R_1, R_2, R_L, D) is achievable, i.e., for a given $\epsilon > 0$ there exists an identification scheme such that all conditions in Definition 3.2 are satisfied for *all* sufficiently large n. In the following we consider the case where we have $R_L \leq R$ and $D \leq \frac{\sigma_X^2 \sigma_{N_1}^2}{\sigma_Y^2}$ and derive an outerbound for the achievable rate-distortion region. Similarly as in the converse direction for the

discrete case in Section 3.3.C, we denote by $\boldsymbol{J}$ and $\boldsymbol{K}$ the tuples $(J_i)_{i=1}^M$ and $(K_i)_{i=1}^M$. The distortion constraint implies that

$$\begin{aligned} D+\epsilon &> \mathbb{E}[d(X^n(W), g_3((J_{\hat{W}}, K_{\hat{W}}, \hat{W}), Y^n))] \\ &\geq \inf_g \mathbb{E}[d(X^n(W), g(W, \boldsymbol{J}, \boldsymbol{K}, \hat{W}, J_{\hat{W}}, K_{\hat{W}}, Y^n))] \\ &\geq \frac{1}{n}\sum_{i=1}^n \inf_{g_i} \mathbb{E}[d(X_i(W), g_i(W, \boldsymbol{J}, \boldsymbol{K}, \hat{W}, J_{\hat{W}}, K_{\hat{W}}, Y^n))] \end{aligned} \tag{4.90}$$

where the infimum is taken over all possible measurable functions g_i on $\mathcal{W}\times\mathcal{M}_1^M\times\mathcal{M}_2^M\times(\mathcal{W}\cup\{e\})\times\mathcal{M}_1\times\mathcal{M}_2\times\mathbb{R}^n$. In our identification scheme, $(\hat{W}, J_{\hat{W}}, K_{\hat{W}})$ are functions of $(\boldsymbol{J}, \boldsymbol{K}, Y^n, Z^n)$, which lead to the following relations

$$X^n(W) - (W, Y^n, Z^n, \boldsymbol{J}, \boldsymbol{K}) - (\hat{W}, J_{\hat{W}}, K_{\hat{W}}). \tag{4.91}$$

as well as

$$X^n(W) - (Y^n, W, J_W, K_W) - (Z^n, \boldsymbol{J}_{\setminus W}, \boldsymbol{K}_{\setminus W}), \tag{4.92}$$

where we use $\boldsymbol{J}_{\setminus W}$ as a shorthand notation of $(J_l)_{l=1,l\neq W}^M$ and similarly for $\boldsymbol{K}_{\setminus W}$. The Markov relation and the property of the squared error imply that

$$\begin{aligned} D+\epsilon &> \sum_{i=1}^n \frac{1}{n}\mathbb{E}[d(X_i(W), \mathbb{E}[X_i(W)|W, \boldsymbol{J}, \boldsymbol{K}, \hat{W}, J_{\hat{W}}, K_{\hat{W}}, Y^n])] \\ &\overset{(a)}{=} \sum_{i=1}^n \frac{1}{n}\mathbb{E}[d(X_i(W), \mathbb{E}[X_i(W)|W, J_W, K_W, Y^n])], \end{aligned} \tag{4.93}$$

where (a) holds due to Corollary 2.1 in Chapter 2. The constraint (4.93) can be interpreted in the following sense. A genie provides us the exact information (W, J_W, K_W). Then, we use the optimal estimator in the square error sense to reconstruct the original sequence using the aided information and the available information $(\hat{W}, J_{\hat{W}}, K_{\hat{W}}, Y^n)$. It turns out that the optimal estimator depends only on the exact information and the observation sequence.

As a standard step for a Gaussian setting, we next relate the distortion constraint (4.93) to a differential entropy term. Using[4.19] Lemma 2.11, and Lemma 2.12, (4.93) implies that

$$h(X^n(W)|J_W, K_W, W, Y^n) \leq \frac{n}{2}\log_2(2\pi e(D+\epsilon)). \tag{4.94}$$

[4.19] Since $h(X^n(W)|Y^n)$ as well as $h(X_i(W)|X^{i-1}(W), Y^n)$ are finite, conditional entropies $h(X_i(W)|W, J_W, K_W, X^{i-1}(W), Y^n)$ and $h(X^n(W)|W, J_W, K_W, Y^n)$ are well-defined due to Lemma 2.11. We further have $h(X^n(W)|W, J_W, K_W, Y^n) = h(X^n(W), Y^n|W, J_W, K_W) - h(Y^n|W, J_W, K_W) = h(X_n(W)|W, J_W, K_W, X^{n-1}(W), Y^n) + h(X^{n-1}(W), Y^n|W, J_W, K_W) - h(Y^n|W, J_W, K_W) = \cdots = \sum_i h(X_i(W)|W, J_W, K_W, X^{i-1}(W), Y^n)$ holds. Using the projection property $\mathbb{E}[(X_i(W) - \mathbb{E}[X_i(W)|W, J_W, K_W, Y^n])^2] \geq \mathbb{E}[(X_i(W) - \mathbb{E}[X_i(W)|W, J_W, K_W, X(W)^{i-1}, Y^n])^2]$, and Lemma 2.12, we obtain (4.94).

Furthermore

$$h(X^n(W)|J_W, K_W, W, Y^n) \overset{(*)}{\leq} h(X^n(W)|Y^n) = \frac{n}{2}\log_2\left(2\pi e\frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2}\right), \tag{4.95}$$

so that the assumption $D \leq \frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2}$ is to make the constraint (4.94) possibly active. $(*)$ is valid since conditioning reduces the entropy, cf. Lemma 2.11.

Next, using the variant of Fano's inequality (3.42), we arrive at the following expression, which corresponds to (3.44),

$$\begin{aligned} n(R-\epsilon) \leq \log_2 M &\leq I(W, J_W; Z^n) + n(R_L + \epsilon_n) \\ &= h(Z^n) - h(Z^n|W, J_W) + n(R_L + \epsilon_n) \\ &= \frac{n}{2}\log_2(2\pi e\sigma_Z^2) - h(Z^n|W, J_W) + n(R_L + \epsilon_n), \end{aligned} \tag{4.96a}$$

where $\epsilon_n = 2\epsilon + \frac{1}{n}\epsilon\log_2 M$. The second inequality follows from (3.42) and the first equality is valid due to Lemma 2.10. This leads to

$$n(R-\epsilon)(1-\epsilon) \leq \frac{n}{2}\log_2(2\pi e\sigma_Z^2) - h(Z^n|W, J_W) + n(R_L + 2\epsilon) \tag{4.97}$$

which implies that

$$h(Z^n|W, J_W) \leq \frac{n}{2}\log_2(2\pi e\sigma_Z^2 2^{-2((R-\epsilon)(1-\epsilon)-R_L)+4\epsilon}). \tag{4.98}$$

Lemma 2.10 also shows that for a given (w, j_w) the conditional pdfs $p_{X^n(W)|WJ_W}$ and $p_{Y^n|WJ_W}$ are well defined. Furthermore, N_1^n and N_2^n are independent of (W, J_W, K_W). Due to the entropy power inequality [EK11, p. 22], cf. also [Ooh98, Eq. (20)], we obtain

$$\begin{aligned} 2^{\frac{2}{n}h(Z^n|W,J_W)} &\geq 2^{\frac{2}{n}h(Y^n|W,J_W)} + 2^{\frac{2}{n}h(N_2^n|W,J_W)}, \\ 2^{\frac{2}{n}h(Z^n|W,J_W)} &\geq 2^{\frac{2}{n}h(X^n(W)|W,J_W)} + 2^{\frac{2}{n}h(N_1^n|W,J_W)} + 2^{\frac{2}{n}h(N_2^n|W,J_W)}, \end{aligned} \tag{4.99}$$

which leads to

$$\begin{aligned} 2\pi e(\sigma_Z^2 2^{-2((R-\epsilon)(1-\epsilon)-R_L)+4\epsilon} - \sigma_{N_2}^2) &\geq 2^{\frac{2}{n}h(Y^n|W,J_W)} \\ 2\pi e(\sigma_Z^2 2^{-2((R-\epsilon)(1-\epsilon)-R_L)+4\epsilon} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)) &\geq 2^{\frac{2}{n}h(X^n|W,J_W)}. \end{aligned} \tag{4.100}$$

Since $h(X^n|W, J_W) > -\infty$ by Lemma 2.10, we therefore have the following condition

$$(R-\epsilon)(1-\epsilon) - R_L - 2\epsilon < \frac{1}{2}\log_2\left(\frac{\sigma_Z^2}{\sigma_{N_1}^2 + \sigma_{N_2}^2}\right). \tag{4.101}$$

Inequality (4.101) further leads to, since by Definition 3.2 it must hold for every $\epsilon > 0$,

$$R \leq \frac{1}{2}\log_2\left(1 + \frac{\sigma_X^2}{\sigma_{N_1}^2 + \sigma_{N_2}^2}\right) + R_L, \tag{4.102}$$

as we take $\epsilon \to 0$.
Additionally, corresponding to (3.46) we obtain

$$\begin{aligned} n(R-\epsilon) &\le I(W, J_W, K_W; Y^n) + 1 + \epsilon \log_2 M \\ &= h(Y^n) - h(Y^n|W, J_W, K_W) + 1 + \epsilon \log_2 M \\ &= \frac{n}{2} \log_2(2\pi e \sigma_Y^2) - h(Y^n|W, J_W, K_W) + 1 + \epsilon \log_2 M, \end{aligned} \tag{4.103}$$

which leads to

$$h(Y^n|W, J_W, K_W) \le \frac{n}{2} \log_2(2\pi e \sigma_Y^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon}). \tag{4.104}$$

Similarly, using the entropy power inequality results in that

$$\begin{aligned} 2\pi e \sigma_Y^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon} &\ge 2^{\frac{2}{n} h(Y^n|W, J_W, K_W)} \\ &\ge 2^{\frac{2}{n} h(X^n(W)|W, J_W, K_W)} + 2^{\frac{2}{n} h(N_1^n|W, J_W, K_W)} > 2\pi e \sigma_{N_1}^2, \end{aligned} \tag{4.105}$$

since $h(X^n|W, J_W, K_W) > -\infty$ due to Lemma 2.10 as well. Thus, there exists an α_1 with $0 \le \alpha_1 < 1$, *which depends on other parameters*, such that

$$h(Y^n|W, J_W, K_W) = \frac{n}{2} \log_2(2\pi e((1-\alpha_1)\sigma_Y^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon} + \alpha_1 \sigma_{N_1}^2)), \tag{4.106}$$

and

$$h(X^n(W)|W, J_W, K_W) \le \frac{n}{2} \log_2(2\pi e(1-\alpha_1)(\sigma_Y^n 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon} - \sigma_{N_1}^2)). \tag{4.107}$$

From (4.105) we also obtain a constraint on the rate R, namely

$$R \le \frac{1}{2} \log_2 \left(\frac{\sigma_Y^2}{\sigma_{N_1}^2} \right). \tag{4.108}$$

Thus (4.102) and (4.108) imply that

$$0 \le R_L \le R \le R_\gamma. \tag{4.109}$$

Using the second inequality in (4.100) we have

$$\begin{aligned} &n(R_1 + \epsilon) \ge I(X^n(W); J_W, W) \\ &\ge \frac{n}{2} \left(\log_2(2\pi e \sigma_X^2) - \log_2(2\pi e(\sigma_Z^2 2^{-2((R-\epsilon)(1-\epsilon)-R_L)+4\epsilon} - (\sigma_{N_1}^2 + \sigma_{N_2}^2))) \right) \\ &= \frac{n}{2} \log_2 \left(\frac{\sigma_X^2}{\sigma_Z^2 2^{-2((R-\epsilon)(1-\epsilon)-R_L)+4\epsilon} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)} \right). \end{aligned} \tag{4.110}$$

Taking $\epsilon \to 0$ we obtain

$$R_1 \ge \frac{1}{2} \log_2 \left(\frac{\sigma_X^2}{\sigma_Z^2 2^{-2(R-R_L)} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)} \right). \tag{4.111}$$

Similarly, corresponding to (3.48) we obtain

$$\begin{aligned} n(R_1+R_2+\epsilon) &\geq I(Y^n;J_W,W)+I(X^n(W);J_W,K_W,W|Y^n) \\ &= \underbrace{h(Y^n)-h(Y^n|W,J_W)}_{\Delta_1}+\underbrace{h(X^n(W)|Y^n)-h(X^n(W)|J_W,K_W,W,Y^n)}_{\Delta_2}. \end{aligned} \tag{4.112}$$

The first term in (4.112) is bounded based on the first inequality in (4.100) as

$$\begin{aligned} \Delta_1 &\geq \frac{n}{2}\bigg(\log_2(2\pi e\sigma_Y^2)-\log_2(2\pi e(\sigma_Z^2 2^{-2((R-\epsilon)(1-\epsilon)-R_L)+4\epsilon}-\sigma_{N_2}^2)\bigg) \\ &= \frac{n}{2}\bigg(\log_2\frac{\sigma_Y^2}{\sigma_Z^2 2^{-2((R-\epsilon)(1-\epsilon)-R_L)+4\epsilon}-\sigma_{N_2}^2}\bigg). \end{aligned} \tag{4.113}$$

The second term is bounded in three different ways:

1. From (4.94) we obtain

$$\begin{aligned} \Delta_2 &\geq \frac{n}{2}\log_2\left(2\pi e\frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2}\right)-\frac{n}{2}\log_2 2\pi e(D+\epsilon) \\ &= \frac{n}{2}\log_2\frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2(D+\epsilon)}. \end{aligned} \tag{4.114}$$

   This implies in combination with (4.113) that

$$R_1+R_2 \geq \frac{1}{2}\log_2\frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)}-\sigma_{N_2}^2}+\frac{1}{2}h_0(R). \tag{4.115}$$

2. Secondly, the expressions in (4.106) and (4.107) lead to

$$\begin{aligned} \Delta_2 &= h(X^n(W)|Y^n)-h(X^n(W)|W,J_W,K_W) \\ &\quad -h(Y^n|X^n(W))+h(Y^n|W,J_W,K_W) \\ &\geq \frac{n}{2}\log_2\left(\frac{\sigma_X^2}{\sigma_Y^2}\frac{(1-\alpha_1)\sigma_Y^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon}+\alpha_1\sigma_{N_1}^2}{(1-\alpha_1)(\sigma_Y^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon}-\sigma_{N_1}^2)}\right) \\ &\overset{(a)}{\geq} \frac{n}{2}\log_2\bigg(\frac{\sigma_X^2}{\sigma_Y^2(\sigma_Y^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon}-\sigma_{N_1}^2)} \\ &\quad \times \inf_{0\leq\alpha_1<1}(\sigma_Y^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon}+\frac{\alpha_1}{1-\alpha_1}\sigma_{N_1}^2)\bigg). \end{aligned} \tag{4.116}$$

   We note that due to the inequality (4.105) the term $\frac{\sigma_X^2}{\sigma_Y^2(\sigma_Y^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon}-\sigma_{N_1}^2)}$ is positive hence (a) is valid. Note also that since α_1 might depend on ϵ and n, we should avoid taking the limit directly. Since $\frac{\alpha_1}{1-\alpha_1}$ is an increasing and positive function of α_1 on $[0,1)$, the infimum is attained at $\alpha_1=0$. Hence

$$\Delta_2 \geq \frac{n}{2}\log_2\left(\frac{\sigma_X^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon}}{\sigma_Y^2 2^{-2(R-\epsilon)(1-\epsilon)+2\epsilon}-\sigma_{N_1}^2}\right). \tag{4.117}$$

This implies by taking $\epsilon \to 0$ that

$$R_1 + R_2 \geq \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2} + \frac{1}{2}h_1(R), \tag{4.118}$$

3. Lastly, by applying a similar derivation we also observe that

$$R_1 + R_2 \geq \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2} + \frac{1}{2}h_2(R). \tag{4.119}$$

Combining these three bounds we obtain

$$R_1 + R_2 \geq \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2} + \frac{1}{2}\max\big\{h_0(R), h_1(R), h_2(R)\big\}. \tag{4.120}$$

Additionally, we have the following constraint which corresponds to (3.50)

$$\begin{aligned} n(R_1 + R_2 + \epsilon) - \log M &\geq I(X^n(W); W, J_W, K_W | Y^n) - (1 + \epsilon \log M) \\ &= \Delta_2 - (1 + \epsilon \log M). \end{aligned} \tag{4.121}$$

which implies that

$$R_1 + R_2 - R \geq \Gamma. \tag{4.122}$$

In summary, we obtain the following outerbound region

$$\begin{aligned} 0 < D \leq \frac{\sigma_X^2 \sigma_{N_1}^2}{\sigma_Y^2},\ 0 \leq R_L \leq R \leq R_\gamma, \\ R_1 \geq \frac{1}{2}\log_2 \left(\frac{\sigma_X^2}{\sigma_Z^2 2^{-2(R-R_L)} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)} \right), \\ R_1 + R_2 \geq \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2} + \Gamma, \\ R_1 + R_2 - R \geq \Gamma. \end{aligned} \tag{4.123}$$

As $R \to R_\gamma$ either $h_1(R)$ or $h_2(R)$ goes to ∞. However, since both R_1 and R_2 are finite we must have

$$0 \leq R_L \leq R < R_\gamma. \tag{4.124}$$

4.3.C Analyzing the outerbound

The above outerbound matches some properties such as the reduction to Wyner-Ziv's region which are mentioned in Subsection 4.3.A. We observe that for fixed D and R_L the three functions $h_0(R)$, $h_1(R)$ and $h_2(R)$ defined in (4.3) provide the key for the transition behavior from one extreme case to another since they are monotone in the identification rate R.

In Fig. 4.1 and Fig. 4.2 we plot different scenarios to indicate which of these functions is the dominant one in a given interval. Studying this behavior facilitates the process of selecting auxiliary random variables in the upcoming achievability part.

Phase transition points

We show in the following that there are three possible transition points $R_{cr_{12}}$, $R_{cr_{01}}$ and $R_{cr_{02}}$ as R varies, where the corresponding subscripts indicate which functions are involved. More specifically, we have

$$\begin{aligned} R_{cr_{12}} &= \frac{1}{2}\log_2 \frac{\sigma_Z^2 2^{2R_L} - \sigma_Y^2}{\sigma_{N_2}^2}, \quad R_{cr_{01}} = \frac{1}{2}\log_2 \frac{\sigma_Y^2(1-\frac{D}{\sigma_{N_1}^2})}{\sigma_{N_1}^2} \\ R_{cr_{02}} &= R_L + \frac{1}{2}\log_2 \frac{\sigma_Z^2}{\sigma_{N_2}^2 + \frac{\sigma_{N_1}^2}{1-\frac{D}{\sigma_{N_1}^2}}}. \end{aligned} \tag{4.125}$$

To derive $R_{cr_{12}}$ we first notice that the function

$$\frac{\sigma_Y^2 2^{-2R} - \sigma_{N_1}^2}{\sigma_X^2 2^{-2R}} = \frac{\sigma_Y^2}{\sigma_X^2} - \frac{\sigma_{N_1}^2}{\sigma_X^2 2^{-2R}} \tag{4.126}$$

is a decreasing function w.r.t. R, which implies that $h_1(R)$ is an increasing one. Similarly, $h_2(R)$ is also an increasing function w.r.t. R. Hence by solving the following equation

$$h_1(R) = h_2(R) \tag{4.127}$$

we can find the (possibly) unique intersection point $R_{cr_{12}}$ if the equation has a solution. The above expression implies that

$$\begin{aligned} \Rightarrow \sigma_Y^2 - \frac{\sigma_{N_1}^2}{2^{-2R}} &= \sigma_Y^2\left(1 - \frac{\sigma_{N_1}^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2}\right) \\ \Leftrightarrow R_{cr_{12}} &= \frac{1}{2}\log_2 \frac{\sigma_Z^2 2^{2R_L} - \sigma_Y^2}{\sigma_{N_2}^2}. \end{aligned} \tag{4.128}$$

Note that however $R_{cr_{12}}$ can lie outside the interval $[R_L, R_\gamma)$, i.e., $h_1(R) \neq h_2(R)$ for all $R \in [R_L, R_\gamma)$.

Next, note that $h_1(0) = 0$ and $h_1(R) \to \infty$ as $R \to \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_{N_1}^2}$. Since $h_1(R)$ is increasing, there is a unique point $R_{cr_{01}} \in [0, \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_{N_1}^2})$ such that $h_1(R) = h_0(R)$, i.e.,

$$\frac{\sigma_{N_1}^2}{\sigma_Y^2 D} = \frac{2^{-2R_{cr_{01}}}}{\sigma_Y^2 2^{-2R_{cr_{01}}} - \sigma_{N_1}^2}, \tag{4.129}$$

above which the $h_1(R)$ dominates $h_0(R)$. Solving for $R_{cr_{01}}$ we obtain

$$R_{cr_{01}} = \frac{1}{2}\log_2 \frac{\sigma_Y^2(1-\frac{D}{\sigma_{N_1}^2})}{\sigma_{N_1}^2}. \tag{4.130}$$

Analogously, $R_{cr_{02}} \in [R_L, R_L + \frac{1}{2}\log_2 \frac{\sigma_Z^2}{\sigma_{N_1}^2+\sigma_{N_2}^2})$ given as in (4.125) is the unique intersection point of $h_2(R)$ and $h_0(R)$, above which the $h_2(R)$ dominates $h_0(R)$, as $h_2(R_L) = 0$ while $h_2(R) \to \infty$ when $R \to R_L + \frac{1}{2}\log_2 \frac{\sigma_Z^2}{\sigma_{N_1}^2+\sigma_{N_2}^2}$. As the solution of the equation $h_2(R) = h_0(R)$, $R_{cr_{02}}$ also satisfies

$$D = \sigma_{N_1}^2\left(1 - \frac{\sigma_{N_1}^2}{\sigma_Z^2 2^{-2(R_{cr_{02}}-R_L)} - \sigma_{N_2}^2}\right). \tag{4.131}$$

Discussion
In this part we discuss some additional properties of the three functions and transitions points. We note that we have $R_{cr_{12}} \geq R_L$ because from $R_L \geq 0$ it follows that $\sigma_Z^2 2^{2R_L} - \sigma_Y^2 \geq 2^{2R_L}(\sigma_Z^2 - \sigma_Y^2) = 2^{2R_L}\sigma_{N_2}^2$. Additionally, again because we have $R_L \geq 0$ it follows that $h_1(0) = 0 \geq h_2(0)$ as

$$\sigma_Y^2(1 - \frac{\sigma_{N_1}^2}{\sigma_Z^2 2^{2R_L} - \sigma_{N_2}^2}) \geq \sigma_X^2. \tag{4.132}$$

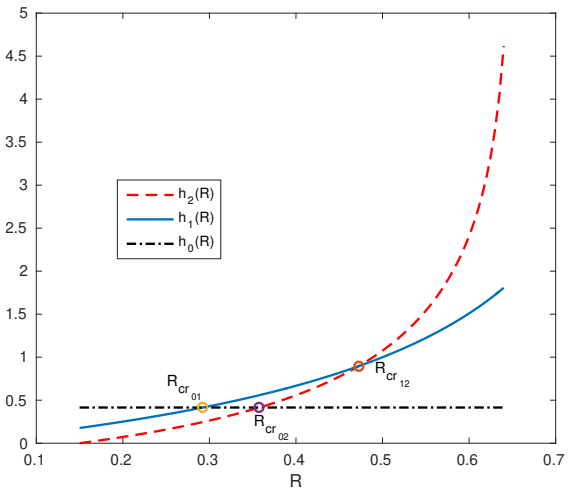
Thus for $0 \leq R \leq \min\{R_\gamma, R_{cr_{12}}\}$, $h_1(R) \geq h_2(R)$. Furthermore, we observe that when $R \leq \min\{R_{cr_{12}}, R_\gamma\}$ the following holds

$$R \geq \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2}. \tag{4.133}$$

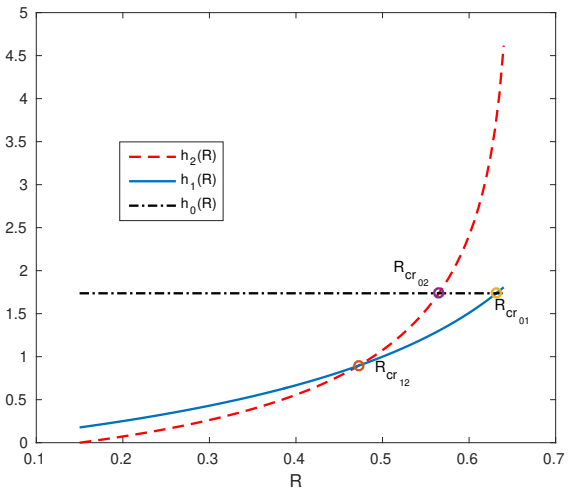
Therefore, the constraint (4.4c), can be omitted in this case. If the interval $(R_{cr_{12}}, R_\gamma)$ is not empty then the reverse inequality holds on it and the constraint (4.4d) can be omitted.
The following relation is helpful to relate Case II and Case V in the later paragraph.

$$\sigma_Y^2(1 - 2^{-2R_{cr_{01}}}) = \sigma_Z^2(1 - 2^{-2(R_{cr_{02}}-R_L)}) = \sigma_Y^2 - \frac{\sigma_{N_1}^2}{1-\frac{D}{\sigma_{N_1}^2}}. \tag{4.134}$$

Importantly, when $D \to 0$, we observe that as $R \to R_\gamma$ either $h_1(R)$ or $h_2(R)$ goes to ∞. Hence at least one of the point $R_{cr_{01}}$ or $R_{cr_{02}}$ lies in the interval $[R_L, R_\gamma)$. If $D \to \frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2}$, then $R_{cr_{01}}$ goes to 0 and hence might lie outside the interval $R_L \leq R < R_\gamma$. In this case $R_{cr_{02}}$ is always inside.



(a) Case 1: $R_{cr12} < R_\gamma$ and $R_L \leq R_{cr01} \leq R_{cr12}$. We can see that when $R_L \leq R \leq R_{cr01}$, $h_0(R)$ dominates over $h_1(R)$ and $h_2(R)$. $h_1(R)$ is the dominant component when $R_{cr01} < R \leq R_{cr12}$. When $R_{cr12} < R < R_\gamma$, then $h_2(R)$ dominates the other two functions.



(b) Case 2: $R_{cr12} < R_\gamma$ and $R_{cr01} \geq R_{cr12}$. In this case $h_0(R)$ dominates over the other two functions when $R_L \leq R \leq R_{cr02}$. For $R \in [R_{cr02}, R_\gamma)$, $h_2(R)$ is the dominant component

Figure 4.1: Two cases when D varies for fixed R_L.

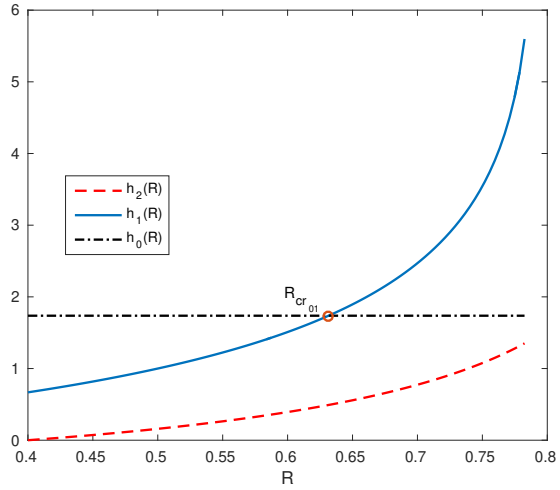


Figure 4.2: Case 3: $R_{cr_{12}} \geq R_\gamma$ and $R_{cr_{01}} \geq R_L$. In this case for $R_L \leq R \leq R_{cr_{01}}$, $h_0(R)$ is the dominant component, while for $R_{cr_{01}} < R < R_\gamma$, $h_1(R)$ is the dominant component.

4.3.D Achievability

From Fig. 4.1 and Fig. 4.2 we see that different constraints will be active in the outer bound depending on the identification rate R. In the achievability we will therefore distinguish between different cases and select the parameters accordingly. In Table 4.1 we provide an overview about the different cases as well as information about the marginal distributions of U and V that are encountered in the following. Fix a value of D and R_L where $0 \leq D \leq \frac{\sigma_X^2 \sigma_{N1}^2}{\sigma_Y^2}$ and $0 \leq R_L < R_\gamma(R_L)$.

The case $R_{cr_{12}} < R_\gamma$:
We consider first that $R_{cr_{12}} < R_\gamma$ which implies that both $h_1(R_{cr_{12}})$ and $h_2(R_{cr_{12}})$ are defined, i.e., $R_{cr_{12}}$ lies in both domains of $h_1(R)$ and $h_2(R)$. Note also that $R_{cr_{12}} \geq R_L$ holds, cf. (4.128).

a) $R_{cr_{01}} \leq R_{cr_{12}}$: In cases I and II we need to truncate the corresponding interval if necessary so that the condition $R \geq R_L$ holds.

- Case I: $R_{cr_{01}} \leq R < R_{cr_{12}}$, i.e., $h_1(R)$ is the dominant component in the outerbound since $h_1(R) \geq h_0(R)$ when $R \geq R_{cr_{01}}$ and $h_1(R) > h_2(R)$ when $R < R_{cr_{12}}$. Let $X = V + N_0$ where V and N_0 are independent Gaussian random variables with $\sigma_V^2 = \sigma_Y^2(1 - 2^{-2R})$. Note that $\sigma_V^2 < \sigma_X^2$ since $R < R_\gamma$. V should be understood as the output of the *test channel* $p_{V|X}$, cf. [CT12, p. 311]. Then, let $V = U + N_0'$ where U and N_0' are independent

Gaussian random variables such that $\sigma_U^2 = \sigma_Z^2(1 - 2^{-2(R-R_L)})$. We also observe that $\sigma_U^2 > 0$ if $R > R_L$. We note that

$$2^{-2R}(\sigma_Z^2 2^{2R_L} - \sigma_Y^2) > \sigma_{N_2}^2 \text{ or } \sigma_Y^2(1-2^{-2R}) > \sigma_Z^2(1-2^{-2(R-R_L)}), \quad (4.135)$$

since $R < R_{cr_{12}}$. This means that $\sigma_U^2 < \sigma_V^2$. Similarly, U is the output of the test channel $p_{U|V}$. By our choice of U and V the relation $U - V - X - Y - Z$ holds. We next examine whether the chosen random variables satisfy the constraints corresponding to the fixed parameters. The condition $I(Z;U) = R - R_L$ is satisfied by the chosen U. Furthermore, $I(Y;V) = R$ due to the choice of V. This means that the choice of U and V does not violate the constraint

$$R \leq \min\{I(Z;U) + R_L, I(Y;V)\}. \quad (4.136)$$

Next we calculate

$$\begin{aligned} h(X|V,Y) &= h(Y|X) + h(X|V) - h(Y|V) \\ &= \frac{1}{2}\log_2\left(2\pi e\frac{(\sigma_X^2 - \sigma_V^2)\sigma_{N_1}^2}{\sigma_Y^2 - \sigma_V^2}\right) = \frac{1}{2}\log_2\left(2\pi e\frac{\sigma_{N_1}^2}{\sigma_Y^2}\frac{\sigma_Y^2 2^{-2R} - \sigma_{N_1}^2}{2^{-2R}}\right) \\ &\overset{(\star)}{\leq} \frac{1}{2}\log_2(2\pi e D) \end{aligned} \quad (4.137)$$

where $(\star)$ is valid due to (4.129) as $R \geq R_{cr_{01}}$, i.e., the distortion level D is attainable using the MMSE decoder. Now, plugging the random variable U into the first expression in the achievable region we obtain

$$R_1 \geq I(X;U) = \frac{1}{2}\log_2\frac{\sigma_X^2}{\sigma_Z^2 2^{-2(R-R_L)} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)}. \quad (4.138)$$

Moreover, we have

$$\begin{aligned} I(X;V|Y) &= \frac{1}{2}\log_2\left(\frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2}\frac{\sigma_Y^2}{\sigma_{N_1}^2}\frac{2^{-2R}}{\sigma_Y^2 2^{-2R} - \sigma_{N_1}^2}\right) \\ &= \frac{1}{2}\log_2\frac{\sigma_X^2 2^{-2R}}{\sigma_Y^2 2^{-2R} - \sigma_{N_1}^2}. \end{aligned} \quad (4.139)$$

Since

$$\begin{aligned} I(X;U) + I(X;V|U,Y) &= I(Y;U) + I(X;U,V|Y) \\ &= I(Y;U) + I(X;V|Y), \end{aligned} \quad (4.140)$$

is valid where the first equality holds since $U - X - Y$, we obtain that

$$R_1 + R_2 \geq \frac{1}{2}\log_2\frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2} + \frac{1}{2}\log_2\frac{\sigma_X^2 2^{-2R}}{\sigma_Y^2 2^{-2R} - \sigma_{N_1}^2}$$

$$R_1 + R_2 - R \geq \frac{1}{2}\log_2 \frac{\sigma_X^2 2^{-2R}}{\sigma_Y^2 2^{-2R} - \sigma_{N_1}^2}, \tag{4.141}$$

which matches the outerbound. When $R = R_L$ we can simply choose U to be a Gaussian random variable which is independent of everything else. As discussed previously in (4.133), since $R \leq R_{cr_{12}}$ the first constraint in (4.141) can be omitted.

The other cases can be matched similarly using the same principle. Since this results in lengthy derivations with limited new insights, we provide the remaining proof in Appendix 4.B.

4.A Proof of Lemma 4.1

For notational brevity we suppress the superscript δ in $\mathcal{B}_n^\delta$ in the rest of this subsection. It is sufficient to prove the lemma for $R_U = I(X;U)+4\delta$, $R_V = I(X;V|U)+5\delta$ and $2^{nR_U} \leq M \leq \hat{M} = 2^{2nR_U}$. We first expand the left-hand side of (4.18) as[4.20]

$$\begin{aligned}&\Pr\Big\{(X^n, U^n(j), V^n(j,l)) \notin \mathcal{B}_n,\ \forall j,l\Big\}\\ &\quad= \int p_X(x^n)\Pr\Big\{(U^n(j), V^n(j,l)) \notin \mathcal{B}_n(x^n),\ \forall j,l\Big\}dx^n.\end{aligned} \tag{4.142}$$

The second term inside the integral can be decomposed as

$$\begin{aligned}\Pr\Big\{(U^n(j), V^n(j,l)) \notin \mathcal{B}_n(x^n),\ \forall j,l\Big\} &\overset{(a)}{=} \prod_{j=1}^{M} \Pr\{(U^n(j), V^n(j,l)) \notin \mathcal{B}_n(x^n),\ \forall l\}\\ &\overset{(b)}{=} \Big\{\Pr\{(U^n(1), V^n(1,l)) \notin \mathcal{B}_n(x^n),\ \forall l\}\Big\}^M,\end{aligned} \tag{4.143}$$

where (a) is valid due to the independence of tuples $\Big(U^n(j), \big(V^n(j,l)\big)_l\Big)_j$ for all j. (b) holds due to the iid of the codebook. Note that

$$\mathcal{B}_n(x^n) = \varnothing \implies \Pr\{(U^n(1), V^n(1,l)) \notin \mathcal{B}_n(x^n),\ \forall l\} = 1. \tag{4.144}$$

Otherwise, we define

$$\begin{aligned}&\mathcal{C}_n(x^n) = \{u^n \mid u^n \in \mathcal{A}_\delta^n(U|x^n), \text{ and } \{v^n \colon v^n \in \mathcal{B}_n(x^n, u^n)\} \neq \varnothing\}\\ &\mathcal{C}_n^c(x^n) = \mathcal{U}^n \backslash \mathcal{C}_n(x^n) = \mathbb{R}^n \backslash \mathcal{C}_n(x^n).\end{aligned} \tag{4.145}$$

[4.20] Note that herein dx^n is a friendly notation for $d\lambda^{\otimes n}$, i.e., we are considering the product of Lebesgue measures.

For brevity we also define $\tau_n = \int_{\mathcal{C}_n^c(x^n)} p_U^n(u^n)du^n$ when $\mathcal{B}_n(x^n) \neq \varnothing$. Then, for each x^n such that $\mathcal{B}_n(x^n) \neq \varnothing$ the following holds

$$\begin{aligned}
&\Pr\{(U^n(1), V^n(1,l)) \notin \mathcal{B}_n(x^n),\ \forall l\} \\
&= \tau_n + \int_{\mathcal{C}_n(x^n)} p_U^n(u^n)\Pr\{V^n(1,l) \notin \mathcal{B}_n(x^n,u^n),\ \forall l | U^n(1) = u^n\}du^n \\
&\overset{(b)}{=} \tau_n + \int_{\mathcal{C}_n(x^n)} p_U^n(u^n)\Big\{\Pr\{V^n(1,1) \notin \mathcal{B}_n(x^n,u^n)|U^n(1) = u^n\}\Big\}^L du^n, \quad (4.146)
\end{aligned}$$

where (b) holds due to the iid of the codebook. Moreover, for $u^n \in \mathcal{C}_n(x^n)$,

$$\begin{aligned}
&\Pr\{V^n(1,1) \notin \mathcal{B}_n(x^n,u^n)|U^n(1) = u^n\} \\
&\quad = 1 - \Pr\{V^n(1,1) \in \mathcal{B}_n(x^n,u^n)|U^n(1) = u^n\} \\
&\quad = 1 - \int_{\mathcal{B}_n(x^n,u^n)} p_{V|U}^n(v^n|u^n)dv^n. \quad (4.147)
\end{aligned}$$

From the definition of $\mathcal{B}_n$ for each $v^n \in \mathcal{B}_n(x^n,u^n)$ we have

$$\frac{p_{V|U}^n(v^n|u^n)}{p_{V|UX}^n(v^n|u^n,x^n)} \geq \frac{2^{-n(h(V|U)+2\delta)}}{2^{-n(h(V|U,X)-2\delta)}} = 2^{-n(I(X;V|U)+4\delta)}. \quad (4.148)$$

This implies that for $u^n \in \mathcal{C}_n(x^n)$ we have the following inequality

$$\begin{aligned}
&\Pr\{V^n(1,1) \notin \mathcal{B}_n(x^n,u^n)|U^n(1) = u^n\} \\
&\quad \leq 1 - 2^{-n(I(X;V|U)+4\delta)} \int_{\mathcal{B}_n(x^n,u^n)} p_{V|UX}^n(v^n|u^n,x^n)dv^n. \quad (4.149)
\end{aligned}$$

Therefore, for $u^n \in \mathcal{C}_n(x^n)$ the second integrand of the second integral in (4.146) is bounded as

$$\begin{aligned}
&\{\Pr\{V^n(1,1) \notin \mathcal{B}_n(x^n,u^n)|U^n(1) = u^n\}^L \\
&\quad \leq \left(1 - 2^{-n(I(X;V|U)+4\delta)} \int_{\mathcal{B}_n(x^n,u^n)} p_{V|UX}^n(v^n|u^n,x^n)dv^n\right)^L \\
&\quad \overset{(*)}{\leq} 1 - \int_{\mathcal{B}_n(x^n,u^n)} p_{V|UX}^n(v^n|u^n,x^n)dv^n + \exp\big(-L2^{-n(I(X;V|U)+4\delta)}\big) \\
&\quad \overset{(c)}{\leq} 1 - \int_{\mathcal{B}_n(x^n,u^n)} p_{V|UX}^n(v^n|u^n,x^n)dv^n + \exp(-2^{n\delta}), \quad (4.150)
\end{aligned}$$

where (c) follows from the definition of L. In $(*)$ we use the following inequality [CT12, Lemma 10.5.3]

$$(1-xy)^n \leq 1 - x + e^{-yn}, \quad (4.151)$$

where $0 \leq x, y \leq 1$, and $n > 0$. Thus, when $\mathcal{B}_n(x^n) \neq \varnothing$,

$$\begin{aligned}
&\Pr\{(U^n(1), V^n(1,l)) \notin \mathcal{B}_n(x^n),\ \forall l\} \\
&\quad \leq \tau_n + \int_{\mathcal{C}_n(x^n)} p_U^n(u^n)\left(1 - \int_{\mathcal{B}_n(x^n,u^n)} p_{V|UX}^n(v^n|u^n,x^n)dv^n + \exp(-2^{n\delta})\right)du^n \\
&\quad = 1 + \exp(-2^{n\delta}) \int_{\mathcal{C}_n(x^n)} p_U^n(u^n)du^n \\
&\quad - \int_{\mathcal{C}_n(x^n)} p_U^n(u^n) \int_{\mathcal{B}_n(x^n,u^n)} p_{V|UX}^n(v^n|u^n,x^n)dv^n du^n \\
&\quad \overset{(d)}{\leq} 1 + \exp(-2^{n\delta}) - 2^{-n(I(X;U)+3\delta)} \int_{\mathcal{B}_n(x^n)} p_{VU|X}^n(v^n,u^n|x^n)du^n dv^n \\
&\quad = 1 + \exp(-2^{n\delta}) - 2^{-n(I(X;U)+3\delta)} \Pr\{(U^n,V^n) \in \mathcal{B}_n(x^n)|X^n = x^n\}, \qquad (4.152)
\end{aligned}$$

where (d) follows since for $u^n \in \mathcal{C}_n(x^n)$ we have

$$\frac{p_U^n(u^n)}{p_{U|X}^n(u^n|x^n)} \geq \frac{2^{-n(h(U)+\delta)}}{2^{-n(h(U|X)-2\delta)}} = 2^{-n(I(X;U)+3\delta)}. \qquad (4.153)$$

Finally, putting the analysis together we obtain

$$\begin{aligned}
&\Pr\left\{(U^n(j), V^n(j,l)) \notin \mathcal{B}_n(x^n),\ \forall j,l\right\} \\
&\quad \leq \left(1 + \exp(-2^{n\delta}) - 2^{-n(I(X;U)+3\delta)} \Pr\{(U^n,V^n) \in \mathcal{B}_n(x^n)|X^n = x^n\}\right)^M \\
&\quad = (1+\exp(-2^{n\delta}))^M \left(1 - 2^{-n(I(X;U)+3\delta)} \frac{\Pr\{(U^n,V^n) \in \mathcal{B}_n(x^n)|X^n = x^n\}}{1+\exp(-2^{n\delta})}\right)^M \\
&\quad \overset{(*)}{\leq} (1+\exp(-2^{n\delta}))^M \\
&\quad \times \left(1 - \frac{\Pr\{(U^n,V^n) \in \mathcal{B}_n(x^n)|X^n = x^n\}}{1+\exp(-2^{n\delta})} + \exp(-M2^{-n(I(X;U)+3\delta)})\right) \\
&\quad \leq (1+\exp(-2^{n\delta}))^{\hat{M}} \left(1 - \frac{\Pr\{(U^n,V^n) \in \mathcal{B}_n(x^n)|X^n = x^n\}}{1+\exp(-2^{n\delta})} + \exp(-2^{n\delta})\right), \qquad (4.154)
\end{aligned}$$

where $(*)$ has the same explanation as before. From equation (4.144) we observe that the bound in (4.154) holds as well for the case $\mathcal{B}_n(x^n) = \varnothing$. Furthermore, note that

$$(1+\exp(-2^{n\delta}))^{\hat{M}} \to 1 \qquad (4.155)$$

as $n \to \infty$ which will be pointed out in the following. Define $\beta = 2^{n\delta}$ which implies that $\hat{M} = 2^{2n(I(X;U)+3\delta)} = \beta^\alpha$ where $\alpha = 2\frac{I(X;U)+3\delta}{\delta} > 0$. Also as $n \to \infty$, $\beta \to \infty$. It suffices to show that

$$\lim_{\beta\to\infty} \beta^\alpha \ln(1+e^{-\beta}) = 0, \qquad (4.156)$$

which can be concluded from L'Hospital's rule. Next, we average over x^n which gives us

$$\Pr\Big\{(X^n, U^n(j), V^n(j,l)) \notin \mathcal{B}_n,\ \forall j,l\Big\} \leq (1+\exp(-2^{n\delta}))^{\hat{M}}\left(1+\exp(-2^{n\delta}) - \frac{\Pr\{(U^n,V^n,X^n)\in\mathcal{B}_n\}}{1+\exp(-2^{n\delta})}\right). \quad (4.157)$$

The fact that

$$\Pr\{(U^n,V^n,X^n)\in\mathcal{B}_n\}\to 1, \quad (4.158)$$

follows from

$$\Pr\{(U^n,V^n,X^n)\in\mathcal{A}_\delta^n(UVX)\}\to 1, \text{ as } n\to\infty, \quad (4.159)$$

and (4.16). In conclusion we obtain

$$\Pr\Big\{(X^n, U^n(j), V^n(j,l)) \notin \mathcal{B}_n,\ \forall j,l\Big\}\to 0, \text{ as } n\to\infty. \quad (4.160)$$

4.B Achievability in Theorem 4.2

The case $R_{cr_{12}} < R_\gamma$:

a) $R_{cr_{01}} \leq R_{cr_{12}}$

- Case II: $0 \leq R < R_{cr_{01}}$, $h_0(R)$ dominates both $h_1(R)$ and $h_2(R)$. Let V and N_0 be independent Gaussian random variables such that $X = V + N_0$ where $\sigma_V^2 = \sigma_Y^2(1-2^{-2R_{cr_{01}}})$. Since $R_{cr_{12}} < R_\gamma$ we also have $\sigma_V^2 < \sigma_X^2$. Additionally, let U and N_0' be independent Gaussian random variables such that $V = U + N_0'$ where $\sigma_U^2 = \sigma_Z^2(1-2^{-2(R-R_L)})$. Note that $\sigma_U^2 > 0$, if $R > R_L$. Furthermore, we also observe that $\sigma_U^2 < \sigma_V^2$ since $\sigma_U^2(R)$ is a increasing function of R, and $\sigma_U^2(R_{cr_{01}}) \leq \sigma_V^2$ holds because $R_{cr_{01}} \leq R_{cr_{12}}$. We also have the Markov chain $U - V - X - Y - Z$. Moreover

$$R - R_L = I(Z;U),\ R < R_{cr_{01}} = I(Y;V). \quad (4.161)$$

Additionally

$$h(X|V,Y) = \frac{1}{2}\log_2\left(2\pi e\frac{\sigma_{N_1}^2}{\sigma_Y^2}\frac{\sigma_Y^2 2^{-2R_{cr_{01}}} - \sigma_{N_1}^2}{2^{-2R_{cr_{01}}}}\right) = \frac{1}{2}\log_2 2\pi eD, \quad (4.162)$$

which implies that the distortion level is matched. Hence the chosen random variables satisfy the constraints for fixed parameters. The rate constraint for R_1 is given as in (4.138). The other sum rate constraints can be calculated as

$$R_1 + R_2 \geq \frac{1}{2}\log_2\frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2} + \frac{1}{2}\log_2\frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2 D},$$

$$R_1 + R_2 - R \geq \frac{1}{2}\log_2 \frac{\sigma_X^2 \sigma_{N_1}^2}{\sigma_Y^2 D}. \tag{4.163}$$

which again match with the outerbound. If $R = R_L$ we can choose U to be a Gaussian and independent of everything else. Similarly, the first sum rate constraint in the above region can be removed, cf. (4.133).

- Case III: $R_\gamma > R \geq R_{cr_{12}} \geq R_L$ then $h_2(R)$ dominates the other functions since $h_2(R) \geq h_1(R) \geq h_0(R)$ on this interval. We also observe by the following contradiction that we have $R_{cr_{12}} \geq R_{cr_{02}}$, i.e., $R_{cr_{02}}$ lies inside the interval $[R_L, R_\gamma)$. Assume otherwise that $R_{cr_{02}} > R_{cr_{12}}$, then we have the following chain

$$h_0(R_{cr_{02}}) = h_2(R_{cr_{02}}) > h_2(R_{cr_{12}}) = h_1(R_{cr_{12}}) \geq h_1(R_{cr_{01}}) = h_0(R_{cr_{01}}),$$

which is a contradiction. Furthermore, note that as $\Gamma = \frac{1}{2}h_2(R)$ the third constraint (4.4c) becomes redundant due to (4.4b) as

$$\begin{aligned} R_1 + R_2 &\geq \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2} + \frac{1}{2}h_2(R) \\ &= \frac{1}{2}\log_2 \frac{\sigma_X^2}{\sigma_Z^2 2^{-2(R-R_L)} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)}. \end{aligned} \tag{4.164}$$

Additionally, since $R \geq R_{cr_{12}}$ we have

$$\begin{aligned} &2^{2R}(\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2) \leq \sigma_Y^2, \\ &\Rightarrow R + \frac{1}{2}h_2(R) \leq \frac{1}{2}\log_2 \frac{\sigma_X^2}{\sigma_Z^2 2^{-2(R-R_L)} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)}, \end{aligned} \tag{4.165}$$

which implies that the fourth constraint $R_1 + R_2 \geq R + \Gamma$ also becomes irrelevant, cf. also (4.133). Since this is a degenerate case, we use the region (4.86) for achieving the corresponding outer bound. Let $X = U + N_0$ where U and N_0 are independent Gaussian random variables, where $\sigma_U^2 = \sigma_Z^2(1 - 2^{-2(R-R_L)})$. We observe that $\sigma_U^2 < \sigma_X^2$ since $R < R_\gamma$. The Markov chain $U - X - Y - Z$ is satisfied. Additionally, the condition $I(Z;U) = R - R_L$ is satisfied by the chosen U. Since $R \geq R_{cr_{12}}$, we have $\sigma_U^2 \geq \sigma_Y^2(1 - 2^{-2R})$, cf. (4.135). This implies that $I(Y;U) \geq R$. Next,

$$\begin{aligned} h(X|U,Y) &= \frac{1}{2}\log_2(2\pi e \sigma_{N_1}^2) + \frac{1}{2}\log_2 2\pi e(\sigma_Z^2 2^{-2(R-R_L)} - (\sigma_{N_1}^2 + \sigma_{N_2}^2)) \\ &\quad - \frac{1}{2}\log_2 2\pi e(\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2) \\ &= \frac{1}{2}\log_2 2\pi e \sigma_{N_1}^2 \left(1 - \frac{\sigma_{N_1}^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2}\right) \\ &\overset{(4.131)}{\leq} \frac{1}{2}\log_2 2\pi e D, \end{aligned} \tag{4.166}$$

since $R \geq R_{cr_{12}} \geq R_{cr_{02}}$, i.e., the distortion level D is achievable. The last constraint (4.4b) follows by the choice of the auxiliary random variable U. We note again that in this case, the second layer is not necessary, hence *binning* can be omitted. A similar behavior is observed in [TG14, Section IV].

b) If $R_{cr_{01}} > R_{cr_{12}}$ then $R_{cr_{01}} \geq R_{cr_{02}} \geq R_{cr_{12}} \geq R_L$. Suppose that we have $R_{cr_{02}} < R_{cr_{12}}$ then the following meaningful chain of expressions, i.e., all involving terms are defined, shows the contradiction

$$h_0(R_{cr_{02}}) = h_2(R_{cr_{02}}) < h_2(R_{cr_{12}}) = h_1(R_{cr_{12}}) \leq h_1(R_{cr_{01}}) = h_0(R_{cr_{01}}).$$

Therefore the inequality $R_{cr_{02}} \geq R_{cr_{12}}$ holds.

Combining with our discussion in Subsection 4.3.C, we have $R_{cr_{02}} \in [R_{cr_{12}}, R_L + \frac{1}{2}\log_2 \frac{\sigma_Z^2}{\sigma_{N_1}^2+\sigma_{N_2}^2})$. For $R_\gamma > R \geq R_{cr_{12}}$, we have $h_2(R) \geq h_1(R)$. Additionally, as $R \to R_\gamma$ either $h_1(R)$ or $h_2(R)$ tend to ∞. This implies that $h_2(R)$ goes to ∞ and intersects $h_0(R)$ before $h_1(R)$. Hence, both relations $R_{cr_{01}} \geq R_{cr_{02}}$ and $R_{cr_{02}} \in [R_{cr_{12}}, R_\gamma)$ follow.

- Case IV: If $R_\gamma > R \geq R_{cr_{02}}$, then $h_2(R)$ dominates the outerbound. Since $R \geq R_{cr_{12}}$ the two constraints (4.4c) and (4.4d) are again redundant. U is selected as in Case III. We note that the requirements $I(Y;U) \geq R$ and $h(X|U,Y) \leq \frac{1}{2}\log_2 2\pi eD$ are still fulfilled since $R \geq R_{cr_{02}} \geq R_{cr_{12}}$.
- Case V: If $R_L \leq R < R_{cr_{02}}$, then $h_0(R)$ dominates the outerbound, since not only $h_0(R) = h_0(R_{cr_{02}}) = h_2(R_{cr_{02}}) \geq h_1(R_{cr_{02}})$, but also both $h_1(R_{cr_{02}}) \geq h_1(R)$ and $h_2(R_{cr_{02}}) \geq h_2(R)$ hold. Let V and N_0 be independent Gaussian random variables such that $X = V + N_0$ where $\sigma_V^2 = \sigma_Z^2(1 - 2^{-2(R_{cr_{02}} - R_L)}) = \sigma_Y^2(1 - 2^{-2R_{cr_{01}}})$, cf. (4.134). Additionally, let U and N_0' be independent Gaussian random variables such that $V = U + N_0'$ where $\sigma_U^2 = \sigma_Z^2(1-2^{-2(R-R_L)})$ and $\sigma_U^2 > 0$ if $R > R_L$. Note that $\sigma_U^2 < \sigma_V^2$ since $R < R_{cr_{02}}$. Again we have the relation $U - V - X - Y - Z$. Next,

$$\begin{aligned} h(X|V,Y) &= \frac{1}{2}\log_2 2\pi e\sigma_{N_1}^2 \left(1 - \frac{\sigma_{N_1}^2}{\sigma_Z^2 2^{-2(R_{cr_{02}} - R_L)} - \sigma_{N_2}^2}\right) \\ &\overset{(4.131)}{=} \frac{1}{2}\log_2 2\pi eD, \end{aligned} \tag{4.167}$$

which implies that the distortion level is matched. The choice of U and V also leads to $I(U;Z) = R - R_L$ and from (4.167), cf. also (4.134),

$$\begin{aligned} I(Y;V) &= \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R_{cr_{02}} - R_L)} - \sigma_{N_2}^2} \\ &= \frac{1}{2}\log_2 \sigma_Y^2 \frac{1 - \frac{D}{\sigma_{N_1}^2}}{\sigma_{N_1}^2} = R_{cr_{01}} \geq R. \end{aligned} \tag{4.168}$$

Lastly, the other constraints are calculated as

$$R_1 + R_2 \geq \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R-R_L)} - \sigma_{N_2}^2} + \frac{1}{2}\log_2 \frac{\sigma_X^2 \sigma_{N_1}^2}{\sigma_Y^2 D},$$
$$R_1 + R_2 - R \geq \frac{1}{2}\log_2 \frac{\sigma_X^2 \sigma_{N_1}^2}{\sigma_Y^2 D}, \tag{4.169}$$

which matches the outerbound. If $R = R_L$ we select U to be independent of the other random variables. When $R_L \leq R < R_{cr_{12}}$, ($R_{cr_{12}} \leq R \leq R_{cr_{02}}$), the first (second) constraint in the above region can be removed.

The case $R_{cr_{12}} \geq R_\gamma$:
Since $R_{cr_{12}} \geq R_\gamma$, $h_1(R) > h_2(R)$ holds for all $R_L \leq R < R_\gamma$. If $R_{cr_{01}} > R_L$ then the following argument shows that $R_{cr_{01}}$ lies in the interval $[R_L, R_\gamma)$. If $R_{cr_{01}}$ is outside the interval $[R_L, R_\gamma)$ then both $h_1(R)$ and $h_2(R)$ lie below $h_0(R)$ in $[R_L, R_\gamma)$. However this is not possible since $h_1(R)$ or both $h_1(R)$ and $h_2(R)$ tend to ∞ as $R \to R_\gamma$. Therefore, we need to consider the following subcases:

- Case VI: If $R_{cr_{01}} \leq R < R_\gamma$, then $h_1(R)$ dominates the outerbound. We select U and V as in Case I in the previous discussion. We note that (4.135) still holds since $R < R_\gamma \leq R_{cr_{12}}$.
- Case VII: If $R_L \leq R < R_{cr_{01}}$, then $h_0(R)$ is the dominating component in the outerbound. U and V are chosen identically as in Case II. $\sigma_U^2 < \sigma_V^2$ is valid since $R_{cr_{01}} < R_\gamma \leq R_{cr_{12}}$.

Case VIII: If $R_{cr_{01}} \leq R_L$, then $h_1(R)$ dominates the other functions on $[R_L, R_\gamma)$. So the construction can be done similarly as in Case I as $R < R_{cr_{12}}$ always holds.

4.C On the closedness of $\mathcal{R}_{GS}$

Assume that the sequence of tuples $(R_m, R_{1,m}, R_{2,m}, R_{L,m}, D_m)_{m\in\mathbb{N}} \in \mathcal{R}_{GS}$ tends to (R, R_1, R_2, R_L, D) as $m \to \infty$ w.r.t. ℓ_1-distance. This implicitly means that neither R_1 nor R_2 is ∞. From the definition of $\mathcal{R}_{GS}$ we only need to show that we always have $R < R_\gamma(R_L)$ and $D > 0$, where due to the definition $R_\gamma(R_L)$ depends on R_L. We show this by a proof by contradiction. Suppose that we have $R = R_\gamma(R_L)$ or $D = 0$. For an arbitrary but fixed $\epsilon > 0$, there exists $m_0(\epsilon) \in \mathbb{N}$ such that $\forall m > m_0(\epsilon)$

$$D_m < D + \epsilon,\ R_m > R - \epsilon, \text{ and } R_m - R_{L,m} > R - R_L - \epsilon. \tag{4.170}$$

This implies that

$$R_{1,m} + R_{2,m} \geq \frac{1}{2}\log_2 \frac{\sigma_Y^2}{\sigma_Z^2 2^{-2(R_m - R_{L,m})} - \sigma_{N_2}^2}$$

$$+\frac{1}{2}\max\left\{\log_2\frac{\sigma_X^2\sigma_{N_1}^2}{\sigma_Y^2(D+\epsilon)}, h_1(R-\epsilon), h_2(R-\epsilon)\right\}. \tag{4.171}$$

Taking $m\to\infty$, substituting $D=0$ or $R=R_\gamma(R_L)$ into the right-hand side of (4.171), and then taking $\epsilon\to 0$ we see the violation since if $D=0$ the first term in the maximization tends to ∞ whereas if $R=R_\gamma(R_L)$ one of the two latter terms in the maximization goes to ∞ which contradicts $R_1, R_2<\infty$. Therefore $(R,R_1,R_2,R_L,D)\in\mathcal{R}_{GS}$.

5

Uncertainty

CONSIDER a forensic system where biometric information such as facial features, height, fingerprint etc. are taken and stored into a database. As before compression of this information is desirable to reduce storage burdens. However, the system is required to work properly given some uncertainties. For example users' data distribution might be uncertain due to (ethnic) background information. Furthermore, the observation sequence might be observed through an uncertain channel.

We study several models for uncertainties in this chapter. Inspired by the classic work in the compound and arbitrary varying channels [BBT59; BBT60; Ahl78], we first study scenarios where users' data distribution and channel belong to some sets enumerated by states. Then assume that the state distributions are available, we study the identification system in mixture scenarios.

5.1 Identification problem: compound setting

In this section we consider the setting where users' data are generated iid from an unknown distribution that belongs to a set. Assume that P_X is a member of

$$\mathcal{P} = \{P_{X|S=s} \mid s \in \mathcal{S}\}, \tag{5.1}$$

which is a set of probability measures on the same measurable space $(\mathcal{X}, \mathcal{F})$. We can think that nature selects $s \in \mathcal{S}$ which we do not know. Given an underlying state $s \in \mathcal{S}$ the corresponding users' data sequences $(x^n(i))_{i=1}^M$ are generated from $P_{X|S=s}^{\otimes n}$. Furthermore, the observational channel is from the set

$$\mathcal{P}_c = \{P_{Y|X,\tau} \mid \tau \in \mathcal{T}\}, \tag{5.2}$$

where $P_{Y|X,\tau}$ is a shorthand notation for $P_{Y|X,T=\tau}$. Throughout this setting both $\mathcal{S}$ and $\mathcal{T}$ are finite[5.1]. Denote the corresponding set of "marginal" distributions on $\mathcal{Y}$ by, cf. notation convention,

$$\mathcal{P}_{\mathcal{Y}} = \{P_Y \mid P_Y = P_{Y|X,\tau}P_{X|S=s}, \text{ for some } (\tau, s)\}, \tag{5.3}$$

i.e., we allow different combinations of channels and input distributions resulting in the same (marginal) output distribution however no input distribution results in the same output distribution with different channels. We enumerate elements of $\mathcal{P}_{\mathcal{Y}}$ by $P_{Y,\kappa}$ where $\kappa \in [1 : |\mathcal{P}_{\mathcal{Y}}|]$.

Definition 5.1 An identification scheme consists of an enrollment mapping

$$\phi_n \colon \mathcal{X}^n \to \mathcal{M}_1 \tag{5.4}$$

which compresses the users' information and stores it in a database, and an identification mapping

$$\psi_n \colon \mathcal{Y}^n \times \mathcal{M}_1^M \to \mathcal{W} \cup \{e\}, \tag{5.5}$$

which identifies the true user from the observation and the stored information in the database. Herein e is used to indicate an error event. Note that both mappings are deterministic.

For a given pair (s, τ) the corresponding probability of error is given as

$$\begin{aligned}\Pr\{W \neq \hat{W}|S = s, T = \tau\} = \sum_w \frac{1}{M} \int dP_{Y|X,\tau}^{\otimes n}(y^n|x^n(w)) \\ \times \prod_{i=1}^{M} dP_{X|S=s}^{\otimes n}(X^n(i)) \mathbf{1}\{w \neq \psi_n(y^n, (\phi_n(x^n(i)))_{i=1}^M)\}.\end{aligned}$$

Definition 5.2 A compression-identification rate pair (R_c, R_i) is achievable if for every $\delta > 0$ there exists a pair of aforementioned mappings (ϕ_n, ψ_n) such that

$$\begin{aligned}\frac{1}{n}\log|\mathcal{M}_1| < R_c + \delta, \quad \frac{1}{n}\log|\mathcal{W}| > R_i - \delta \\ \sup_{\tau,s} \Pr\{W \neq \hat{W}|S = s, T = \tau\} \leq \delta.\end{aligned} \tag{5.6}$$

for all $n \geq n_0(\delta)$. The set of all achievable rate pair is denoted by $\mathcal{R}_{sc}$.

[5.1] Our model, which includes two types of uncertainties, source via $P_{X|S=s}$ and channel via $P_{Y|X,T=\tau}$, is a special case of the general source model $\{P_{XY}^s\}_{s\in\mathcal{S}}$. In the finite scenario, we can turn the general model into a source model (5.1) on a super alphabet $\hat{\mathcal{X}} = [1 : |\mathcal{S}||\mathcal{X}|]$ and a fixed channel $P_{Y|\hat{X}}$ as follows. For $\hat{x} \in \hat{\mathcal{X}}$, then $P_{Y|\hat{X}}(y|\hat{x}) = P_{Y|X}^i(y|j)$ where $\hat{x} = (i-1)|\mathcal{X}| + j$ with $j \in [1 : |\mathcal{X}|]$ and $i \in [1 : |\mathcal{S}|]$. Similarly, we define $P_{\hat{X}|S=i}(\hat{x}) = P_X^i(j)$ if $\hat{x} = (i-1)|\mathcal{X}| + j$, otherwise $P_{\hat{X}|S=i}(\hat{x}) = 0$.

Definition 5.3 If X and Y are discrete random variables which take values on finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, we define $\bar{\mathcal{R}}_{sc}$ to be the union over the set of distribution $\{P_{U|X,S=s}\}_s$ on $|\mathcal{U}| \leq |\mathcal{X}| + |\mathcal{T}|$ of the rate pairs (R_c, R_i) for which we have

$$\begin{aligned} R_c &\geq \max_s I(X;U|S=s) \\ R_i &\leq \min_{s,\tau} I(Y;U|S=s,T=\tau). \end{aligned} \tag{5.7}$$

Otherwise, if X and Y are continuous random variables where $\mathcal{P}$ can be represented by a set of density functions $p_{X|S=s}$ and $\mathcal{P}_c$ can be represented by the set of conditional density functions $p_{Y|X,\tau}$ then $\bar{\mathcal{R}}_{sc}$ is defined similarly as the closure of (5.7) over the set of test channels[5.2] $p_{U|X,S=s}$.

Theorem 5.1 *For given sets $\mathcal{P}$ and $\mathcal{P}_c$, the region $\bar{\mathcal{R}}_{sc}$ is achievable, i.e.,*

$$\bar{\mathcal{R}}_{sc} \subseteq \mathcal{R}_{sc}. \tag{5.8}$$

Furthermore, if further X and Y are finite then we have

$$\mathcal{R}_{sc} = \bar{\mathcal{R}}_{sc}. \tag{5.9}$$

The full proof of Theorem 5.1 is given in Appendix 5.B.
Due to the uncertainty caused by the underlying states of the users' data and the observation channel, it is intuitive to estimate these parameters. The estimation can be done due to the concentration effects at large block length. In the proof of Theorem 5.1, we use the following supporting lemma. Lemma 5.1 says that as long as elements in the set of the users' data distributions $\mathcal{P}$ are well-separated then with high probability we can estimate its underlying state correctly.

Lemma 5.1 *Assume that the alphabet $\mathcal{X}$ is a Polish space, specifically, a finite set with discrete metric or $\mathbb{R}$ with Euclidean distance, and $\mathcal{F}$ is the corresponding σ-algebra generated by open sets. Assume that the distributions $P_{X|S=s}$ are distinct for all $s \in \mathcal{S}$, i.e., $\forall s \in \mathcal{S}$ we have $d(P_{X|S=s}, P_{X|S=s'}) \neq 0$ where d is any metric on the space of probability measures. Then there exists a classifier $T : \mathcal{X}^n \to \mathcal{S} \cup \{e\}$, where e denotes an error, such that if $X^n \sim P_{X|S=s}^{\otimes n}$ then*

$$\Pr(T(X^n) = s|S=s) \to 1, \text{ as } n \to \infty. \tag{5.10}$$

The proof of Lemma 5.1 is given in Appendix 5.A.
Since the channel state is also unknown, we can also use Lemma 5.1 to estimate the unknown parameter κ. Then, if we can estimate the state s of the true user correctly then by using κ and s we know for sure the value of τ. Since all users in the system have the same underlying state we can use $S^\star = T_X(X^n(1))$ as an estimate of the true state of users' data sequence.

[5.2] We require that the corresponding entropy and mutual information are finite.

Remark 5.1 We note that Lemma 5.1 only mentions the existence of such a mapping. The Prohorov distance[5.3], which is used in the existence proof, is complex to calculate even in the discrete scenario. There are several alternative candidates for such a mapping. For the finite alphabet case we can use the total variational distance and define the corresponding decision region as

$$\hat{\mathcal{A}}_s^X = \big\{x^n \mid \|P_{x^n} - P_{X|S=s}\|_{TV} < \frac{\hat{t}_X}{2}\big\},\ \forall s \in \mathcal{S}, \tag{5.11}$$

where P_{x^n} is the n-type of the sequence x^n and $\hat{t}_X$ is the minimum total variation distance between any two distribution in $\mathcal{P}$. Note that

$$\|P_{\mathbf{x}^n} - P_{X|S=s}\|_{TV} = \frac{1}{2}\sum_{x\in\mathcal{X}}\left|\frac{N(x|x^n)}{n} - P_{X|S=s}(x)\right| \tag{5.12}$$

where $N(x|x^n)$ is the number of occurrences of x in x^n. From the weak law of large numbers we know that for any $\epsilon > 0$

$$\Pr\{|N(x|x^n)/n - P_{X|S=s}(x)| > \epsilon\} \to 0,\ \text{as } n \to \infty, \tag{5.13}$$

which implies that $\Pr(X^n \in \hat{\mathcal{A}}_s^X | S = s) \to 1$.

Alternatively, under regularity conditions such as for any two generic probability measures P and Q in $\mathcal{P}$ one has $D(P\|Q) < +\infty$ where $D(\cdot\|\cdot)$ is the relative entropy, we can form a classifier by using the binary hypothesis testing approach. This can be used as a workaround for the continuous case. Namely, fix an $\epsilon > 0$ and a $\delta > 0$ sufficiently small. Let P_1 be the first elements in $\mathcal{P}$ according to some enumerations, then for any other probability measure Q there exists a set $\mathcal{A}_Q \subset \mathcal{X}^n$ such that $P_1^{\otimes n}(\mathcal{A}_Q) \geq 1-\delta$ and $Q^{\otimes n}(\mathcal{A}_Q^c) > 1 - e^{-n(D(P_1\|Q)-\epsilon)} > 1-\delta$ for sufficiently large n by Stein's lemma, cf. Theorem 2.2. The decision region for P_1 is readily formed by the intersection of such $\mathcal{A}_Q$. The process continues with the second element and so on.

Example:
Let X be a zero mean Gaussian random variable with unknown variance which belongs to the set $\{\sigma_1^2, \sigma_2^2\}$. Without loss of generality we assume that $0 < \sigma_1^2 < \sigma_2^2$. The AWGN observation channel is modeled by

$$Y = X + N, \quad N \sim \mathcal{N}(0, \sigma_N^2). \tag{5.14}$$

It can be seen that the compression-identification rate region $\mathcal{R}_1$ is given by

$$R_c \geq \frac{1}{2}\log_2 \frac{\sigma_1^2}{\sigma_{Y_1}^2 2^{-2R_i} - \sigma_N^2}$$

[5.3]Given two probability measures P and Q on a Polish space $(\mathcal{X}, d)$ the Prohorov distance is defined as $d_{\mathrm{PH}}(P,Q) = \inf\{\epsilon \mid P(A) \leq Q(A^\epsilon)+\epsilon,\ Q(A) \leq Q(A^\epsilon)+\epsilon, \text{ for all Borel sets } A\}$, where A^ϵ is the ϵ-neighbor of A.

$$0 \le R_i < \frac{1}{2}\log_2(1+\sigma_1^2/\sigma_N^2), \tag{5.15}$$

where $\sigma_{Y_1}^2 = \sigma_1^2+\sigma_N^2$. The achievability follows from Theorem 5.1 with *test* channels $p_{U|X,1}$ and $p_{U|X,2}$ such that conditioning on each state we can write

$$X = U + N', \tag{5.16}$$

where U and N' are independent Gaussian random variables. More specifically, the distribution of U conditioning on the first state is given by $P_{U,1} = \mathcal{N}\big(0, \sigma_{Y_1}^2(1-2^{-2R_i})\big)$. Similarly we have $P_{U,2} = \mathcal{N}\big(0, \sigma_{Y_2}^2(1-2^{-2R_i})\big)$. The converse holds due to the entropy power inequality. The example illustrates that our scheme needs to adapt to the worst state.

5.2 Identification problem: arbitrarily varying settings

5.2.A Independent individual states

We now consider the scenario where each user $i \in \mathcal{W}$ has its own state $s_i \in \mathcal{S}$. Then the corresponding data sequence is generated independently from $P_{X|S=s_i}^{\otimes n}$. Compared to the setting of Theorem 5.1, the underlying state can be different from user to user. The observation channel is assumed to be given as $P_{Y|X}$ such that no two input distributions result in the same output distribution.

This scenario differs from the one given in Section 5.1. Herein the users' data are independent from each other, whereas in the setting in Section 5.1 they are independent only given S. Note that in both cases, without conditioning on $S = s$ (or $S_i = s$) each user's data are not iid generated. The definition of an identification scheme is identical as before. Similarly, we have the following definition of achievability.

Definition 5.4 A pair (R_c, R_i) is achievable if for every $\delta > 0$ there exists a compression-identification pair of mappings (ϕ_n, ψ_n) such that

$$\frac{1}{n}\log|\mathcal{M}_1| < R_c + \delta, \quad \frac{1}{n}\log|\mathcal{W}| > R_i - \delta,$$
$$\sup_{(s_i)_{i=1}^M \in \mathcal{S}^M} \Pr\{\hat{W} \neq W | (S_i)_{i=1}^M = (s_i)_{i=1}^M\} < \delta,$$

for all $n \ge n_0(\delta)$. The set of all achievable pairs is denoted by $\mathcal{R}_{iis}$.

Note that the number of constraints in the current setting grows exponentially with the block length n. With abuse of notation, when $|\mathcal{T}| = 1$ in the setting in Section 5.1, i.e., the observation channel is given, we also denote by $\bar{\mathcal{R}}_{sc}$ the set of rate pairs (R_c, R_i) defined as in Definition 5.3. The result for the current setting is summarized in the following theorem.

Theorem 5.2 *Given the channel $P_{Y|X}$ and the set of distributions $\mathcal{P}$, we have*

$$\mathcal{R}_{iis} = \bar{\mathcal{R}}_{sc}, \tag{5.17}$$

when both X and Y are finite. In case both X and Y are continuous, in which the observation channel is characterized by a conditional density function $p_{Y|X}$ and $\mathcal{P}$ is the set of density functions, then we have $\bar{\mathcal{R}}_{sc} \subseteq \mathcal{R}_{iis}$.

In the proof of Theorem 5.2 we use the same enrollment scheme as the one in Theorem 5.1 with a more careful analysis. The full proof is given in the Appendix 5.C.

Since the underlying state can be different from user to user, the processing unit can only estimate reliably the state of the true user based on the observation sequence y^n from the known observation channel $P_{Y|X}$.

5.2.B A connection to Wyner-Ahlswede-Körner network

It can be seen from the proofs of Theorem 5.1 and Theorem 5.2 that the same random coding arguments lead to identical optimal compression-identification trade-off in both settings when the observation channel is known. We show in this subsection that there exists a class of codes which leads to identical optimal performances in both settings. Moreover, this class includes those codes which were obtained by the random coding argument in the previous theorems.

We establish herein a connection between the *certain* identification problem and the lossless source coding with coded side information, which sheds some light on the reasons why the expression (5.17) holds. Recall that a code for Wyner-Ahlswede-Körner network, called a WAK-code, [AK75], [Wyn75] for the pair of discrete memoryless sources $(X^n, Y^n) \sim P_{XY}^{\otimes n}$ consists of three mappings,

$$\begin{aligned} \phi_{1n} \colon \mathcal{X}^n \to \mathcal{M}_1, \quad \phi_{2n} \colon \mathcal{Y}^n \to \mathcal{M}_2, \\ \psi_n \colon \mathcal{M}_1 \times \mathcal{M}_2 \to \mathcal{Y}^n. \end{aligned} \tag{5.18}$$

A pair (R_1, R_2) is achievable if it fulfills the following constraints for every $\delta > 0$

$$\begin{aligned} \frac{1}{n} \log |\mathcal{M}_i| < R_k + \delta, \ k = 1, 2, \\ \text{and } \ \Pr\{Y^n \neq \hat{Y}^n\} < \delta, \end{aligned} \tag{5.19}$$

for all $n \geq n_0(\delta)$. The set of all achievable rate pairs is denoted by $\mathcal{R}_{\text{WAK}}$. The region $\mathcal{R}_{\text{WAK}}$ is characterized by the following conditions

$$\begin{aligned} R_1 \geq I(X;U), \, R_2 \geq H(Y|U), \\ U - X - Y, \quad |\mathcal{U}| \leq |\mathcal{X}| + 1. \end{aligned} \tag{5.20}$$

If $|\mathcal{S}| = |\mathcal{T}| = 1$ then both settings in Theorem 5.1 and Theorem 5.2 collapse to the same setup. We denote the corresponding region by $\mathcal{R}_{\text{ID}}$. From the trade-off

characterized in Theorem 5.1 and the trade-off given in (5.20) it is immediate that $(R_c, R_i) \in \mathcal{R}_{\text{ID}}$ if and only if $(R_c, H(Y) - R_i) \in \mathcal{R}_{\text{WAK}}$. This observation was also noticed in [TC08, Section III.E]. In both problems R_i characterizes the number of confused (coded) codewords that the system can tolerate.

In fact, the two networks are closely related which is shown by the following proposition. It states that given an *arbitrarily* achievable scheme for the WAK-setting we can construct a corresponding achievable code for the identification setting with the corresponding rate pair.

Proposition 5.1 *Assume that* $(R_1, R_2) \in \mathcal{R}_{\text{WAK}}$*, where* $R_2 < H(Y)$ *then there exists an identification scheme based on the corresponding* WAK*-code such that* $(R_1, H(Y) - R_2)$ *is in* $\mathcal{R}_{\text{ID}}$*.*

A dual statement that given an achievable compression-identification rate pair (R_c, R_i) there exists a corresponding WAK-code such that $(R_c, H(Y)-R_i) \in \mathcal{R}_{\text{WAK}}$ is given in Proposition 5.2 in Section 5.2.D. In a recent work [ZVO19] we apply Proposition 5.1 to transform a polar code for the WAK-problem to a polar code for the identification system.

Proof. Fix $\delta > 0$ and $\gamma > 0$. Suppose that $(R_1, R_2) \in \mathcal{R}_{\text{WAK}}$ then there exists a code $(\phi_{1n}, \phi_{2n}, \psi_n)$ for the WAK network such that

$$\begin{aligned} &\frac{1}{n} \log |\mathcal{M}_k| < R_k + \delta, \quad k = 1, 2, \\ &\text{and } \Pr\{Y^n \neq \hat{Y}^n\} < \delta, \end{aligned} \tag{5.21}$$

for all $n \geq n_0(\delta)$. We define for each $m_1 \in \mathcal{M}_1$ the correctly decodable set

$$\mathcal{B}_{m_1} = \{y^n \mid y^n = \psi_n(m_1, \phi_{2,n}(y^n))\}. \tag{5.22}$$

We note that since the cardinality of the range of ϕ_{2n} is bounded by $|\mathcal{M}_2|$, then $|\mathcal{B}_{m_1}| \leq |\mathcal{M}_2|$ for all $m_1 \in \mathcal{M}_1$.

We take ϕ_{1n} as the enrollment mapping for the identification setting. The corresponding enrolled index is denoted by j_i for all $i \in \mathcal{W}$. The identification mapping is defined based on the sets $\{\mathcal{B}_i\}_{i \in \mathcal{M}_1}$ as follows. We look for a unique index $\hat{w}$ such that $y^n \in \mathcal{B}_{j_{\hat{w}}}$. Otherwise, if there is none or there is more than one such indices, we declare an error. For every $w \in \mathcal{W}$ we then have

$$\begin{aligned} &\Pr\{\hat{W} \neq w | W = w\} \\ &\quad \leq \Pr\{Y^n \notin \mathcal{B}_{J_w} | W = w\} + \Pr\{Y^n \notin \mathcal{A}_\gamma^n | W = w\} \\ &\quad + \Pr\{\exists w' \neq w,\ Y^n \in \mathcal{B}_{J_{w'}}, Y^n \in \mathcal{A}_\gamma^n | W = w\}, \end{aligned} \tag{5.23}$$

where $\mathcal{A}_\gamma^n$ is the weakly typical set based on P_Y. The first term in (5.23) corresponds to the error expression of the WAK problem and is independent of w. For each $w' \neq w$ we have

$$\Pr\{Y^n \in \mathcal{B}_{J_{w'}}, Y^n \in \mathcal{A}_\gamma^n | W = w\} = \sum_{j_{w'}} P_{j_{w'}} \sum_{y^n \in \mathcal{A}_\gamma^n \cap \mathcal{B}_{j_{w'}}} P_Y^{\otimes n}(y^n)$$

$$\leq e^{-n(H(Y)-\gamma)}|\mathcal{M}_2|. \tag{5.24}$$

Therefore we have

$$\begin{aligned}\Pr\{\hat{W} \neq w|W = w\} &\leq P_{\text{WAK}}(\text{error}) \\ &+ \Pr\{\tilde{Y}^n \notin \mathcal{A}_\gamma^n\} + M|\mathcal{M}_2|e^{-n(H(Y)-\gamma)},\end{aligned} \tag{5.25}$$

where $\tilde{Y}^n \sim P_Y^{\otimes n}$. Since $\frac{1}{n}\log|\mathcal{M}_2| \leq R_2 + \delta$ for all $n \geq n_0(\delta)$. Therefore if we take $M = e^{n(H(Y)-R_2-2\gamma-\delta)}$ then the last term in (5.25) goes to 0 as $n \to \infty$. Therefore

$$\Pr\{\hat{W} \neq W\} \leq 3\delta, \tag{5.26}$$

for all $n \geq n_1(\delta, \gamma)$. This shows that $(R_1, H(Y) - R_2)$ is an *achievable* rate pair for the identification setting. The conclusion follows. □

The bound in (5.25), where the last term is caused by the presence of multiple users in the system, can be extended to both settings in Theorem 5.1 and Theorem 5.2 as follows. For simplicity we consider that in both settings the observation channel is known. Let $\{(\phi_{1n,s}, \phi_{2n,s}, \psi_{n,s})\}_{s\in\mathcal{S}}$ be an *arbitrary* set of mappings for the WAK-problems where the corresponding joint distributions are $\{P_{Y|X} \times P_{X|S=s}\}_{s\in\mathcal{S}}$. A code for the setting of Theorem 5.2 is constructed as follows. In the enrollment phase we first estimate the state $\hat{s}_i$ and use the corresponding mapping $\phi_{1n,\hat{s}_i}$ to compress the data of the i-th user. In the identification phase the processing unit estimates the underlying state s' and uses the corresponding set $\mathcal{B}_{m_1,s'} = \{y^n \mid y^n = \psi_{n,s'}(m_1, \phi_{2n,s'}(y^n))\}$ as the decision region. Then we obtain the following upper bound for the setting in Theorem 5.2

$$\begin{aligned}&\Pr\{\hat{W} \neq w|W = w, (S_i)_{i=1}^M\} \\ &\leq P_{\text{WAK}}(\text{error}|S = s_w) + P(\text{estimation error}) \\ &+ \Pr\{\tilde{Y}_w^n \notin \mathcal{A}_{\gamma,w}^n\} + M|\mathcal{M}_{2,s}|e^{-n(H(Y|S=s_w)-\gamma_n)},\end{aligned} \tag{5.27}$$

where $\tilde{Y}_w^n \sim P_{Y|S=s_w}^{\otimes n}$ and $\mathcal{A}_{\gamma,w}^n$ is the corresponding weakly typical set. The last term in (5.27) is valid since conditioning on underlying states the independence still holds. A similar upper bound can be shown for the setting in Theorem 5.1 specialized to the known observation channel case. From the expression (5.27) we can conclude that two important reasons for the expression (5.17) are vanishing estimation error and the mutual independence of users' data conditioning on underlying states.

5.2.C A general identification-compression trade-off

Consider the setting in Section 5.2.A with an additional assumption that the distribution of states P_S with $P_s = P_S(s) > 0$, $\forall s \in \mathcal{S}$ is known. Then the corresponding distribution of each user's data sequence is given by

$$P_{X^n}(x^n) = \sum_{s\in\mathcal{S}} P_s P_{X|S=s}^{\otimes n}(x^n). \tag{5.28}$$

In other words, the user's data distribution is a mixture of iid components. Similarly, users' data are mutually independent and the observation channel $P_{Y|X}$ is known. Additionally, we consider only finite alphabets $\mathcal{X}$ and $\mathcal{Y}$ in this subsection. Roughly speaking, we are interested in what the largest identification rate R_i given a compression rate R_c is such that

$$\lim_{n\to\infty} \Pr\{\hat{W} \neq W\} = 0. \tag{5.29}$$

Before providing a definite answer for the above question, we make a digression and consider a general problem where the underlying processes are not necessarily memoryless. Assume that users' data are generated independently from the same distribution P_{X^n} on a finite alphabet $\mathcal{X}_n$, which is not necessary a discrete memoryless source. The observation channel is given by $P_{Y^n|X^n}$, where $P_{Y^n|X^n=x^n}$ is a probability distribution on a finite alphabet $\mathcal{Y}_n$, which is not necessary a discrete memoryless channel. To study this general problem we use the following definition, which is slightly different from the previous ones[5.4].

Definition 5.5 A rate pair (R_c, R_i) is achievable if there exists a pair of identification-compression mappings (ϕ_n, ψ_n) such that

$$\limsup_{n\to\infty} \frac{1}{n}\log|\mathcal{M}_1| \leq R_c,\ \liminf_{n\to\infty} \frac{1}{n}\log M \geq R_i,$$
$$\lim_{n\to\infty} \Pr\{\hat{W} \neq W\} = 0. \tag{5.30}$$

Let $\mathcal{R}_{gen}$ denote the closure of all achievable rate pairs (R_c, R_i).

To characterize $\mathcal{R}_{gen}$ we need the following quantities. For a joint discrete process $(\mathbf{X}, \mathbf{Y})$ such that $(X^n, Y^n) \sim P_{X^nY^n}$, the spectral sup-mutual information (inf-mutual information) rate [Han03] is defined respectively as

$$\bar{I}(\mathbf{X};\mathbf{Y}) = \text{p-}\limsup_{n\to\infty} \frac{1}{n}\log\frac{P_{Y^n|X^n}(Y^n|X^n)}{P_{Y^n}(Y^n)}$$
$$= \inf\left\{\alpha \Big| \lim_{n\to\infty}\Pr\left[\frac{1}{n}\log\frac{P_{Y^n|X^n}(Y^n|X^n)}{P_{Y^n}(Y^n)} > \alpha\right] = 0\right\},$$
$$\underline{I}(\mathbf{X};\mathbf{Y}) = \text{p-}\liminf_{n\to\infty} \frac{1}{n}\log\frac{P_{Y^n|X^n}(Y^n|X^n)}{P_{Y^n}(Y^n)}$$
$$= \sup\left\{\alpha \Big| \lim_{n\to\infty}\Pr\left[\frac{1}{n}\log\frac{P_{Y^n|X^n}(Y^n|X^n)}{P_{Y^n}(Y^n)} < \alpha\right] = 0\right\}.$$

[5.4]The achievability in Definition 5.5 requires a fixed sequence of mappings (ϕ_n, ψ_n) for a rate pair (R_c, R_i) which is different from the one in Definition 5.2 that allows for each $\delta > 0$ a different set of mappings $(\phi_n(\delta), \psi_n(\delta))$. Since the notion of spectral sup-mutual information, inf-mutual information heavily depends on a fixed sequence of random variables, it is more convenient to use Definition 5.5 when dealing with these quantities.

Theorem 5.3 *$\mathcal{R}_{gen}$ is the closure of the set of all pairs (R_c, R_i) such that*

$$\begin{aligned} R_c &\geq \bar{I}(\mathbf{X};\mathbf{U}) \\ R_i &\leq \underline{I}(\mathbf{Y};\mathbf{U}), \end{aligned} \tag{5.31}$$

for some general process $\mathbf{U} = \{U^n\}_{n=1}^{\infty}$. U^n takes values in a finite alphabet $\mathcal{U}_n$ such that there exists $M > 0$ for which $\frac{1}{n}\log|\mathcal{U}_n| < M$, $\forall n$. Additionally, for each n, $U^n - X^n - Y^n$ holds.

Proof of Theorem 5.3

Achievability:
Choose a general source $\mathbf{U}$ where U^n takes values in a finite alphabet $\mathcal{U}_n$ such that $(X^n, Y^n, U^n) \sim P_{Y^n|X^n} \times P_{X^nU^n}$. The condition $\frac{1}{n}\log|\mathcal{U}_n| < M$ ensures that the right-hand sides in (5.31) are finite, cf. the analysis in [Han03, p.46-47] and [Han03, Theorem 3.5.2]. Given $\gamma > 0$ we first define the set

$$\mathcal{T}_n = \left\{(y^n, u^n)\middle|\frac{1}{n}\log\frac{P_{Y^n|U^n}(y^n|u^n)}{P_{Y^n}(y^n)} \geq \underline{I}(\mathbf{Y};\mathbf{U}) - \gamma\right\}.$$

Then, we define the set $\mathcal{B}_n$ as follows

$$\mathcal{B}_n = \left\{(x^n, u^n) \mid \Pr\{(Y^n, u^n) \notin \mathcal{T}_n | X^n = x^n\} \leq \delta_n^{1/2}\right\}$$

where $\delta_n = \Pr\{(Y^n, U^n) \notin \mathcal{T}_n\} \to 0$ as $n \to \infty$.
Generate a codebook for 2^{nR_c} sequence $U^n(m)$, $m \in [1 : 2^{nR_c}]$ where $U^n(m) \sim P_{U^n}$ and $R_c = \bar{I}(\mathbf{X};\mathbf{U}) + 2\gamma$. In the enrollment phase for the i-th user we look for an index m_i such that $(X^n, U^n(i)) \in \mathcal{B}_n$ and store it in the database. In the identification phase we look for a unique $\hat{w}$ such that $(Y^n, U^n(J_{\hat{w}})) \in \mathcal{T}_n$. The rest of the achievability part follows similarly as in the proof of Theorem 5.1. Hence we omit it.

Converse:
Assume that the identification-compression rate pair (R_i, R_c) is *achievable*. Then, there exists a pair of identification-compression mappings (ϕ_n, ψ_n) such that for every $\gamma > 0$ we have

$$|\mathcal{M}_1| \leq e^{n(R_c+\gamma)},\ M \geq e^{n(R_i-\gamma)}, \tag{5.32}$$

for all $n \geq n_0(\gamma)$. Define $U^n = \phi_n(X^n)$ which takes values on $\mathcal{U}_n$. From (5.32) it can be seen that for all n, $\frac{1}{n}\log|\mathcal{U}_n|$ is upper bounded by a (large enough) constant and $U^n - X^n - Y^n$ holds. It can be then shown that

$$R_c + 2\gamma \geq \bar{I}(\mathbf{X};\mathbf{U}) \tag{5.33}$$

along the lines of arguments in [Han03, p.342]. To show the other inequality we use the following lemma which appears in a general form in [ZTM17, Lemma 9]. The proof of Lemma 5.2 is given in Appendix 5.D for completeness.

Lemma 5.2 *Given an arbitrary $\gamma > 0$, for all sufficiently large n we have*

$$\begin{aligned}&\Pr\{\hat{W} \neq W\}\\ &\geq \Pr\left\{\frac{1}{n}\log\frac{dP_{\bar{Y}^n\phi_n(\bar{X}^n)}}{d(P_{\bar{Y}^n}\times P_{\phi_n(\bar{X}^n)})}(\bar{Y}^n,\phi_n(\bar{X}^n)) \leq R_i - 2\gamma\right\} - e^{-n\gamma},\end{aligned}$$

where $(\bar{Y}^n, \bar{X}^n) \sim P_{Y^nX^n}$.

Using Lemma 5.2 and taking n to ∞ we observe that

$$R_i \leq \underline{I}(\mathbf{Y};\mathbf{U}). \tag{5.34}$$

Finally taking[5.5] $\gamma \to 0$ we obtain the claim. □

We now turn back to the question posed at the beginning of this subsection. We use Definition 5.5 as the definition of achievability and denote $\mathcal{R}_{mix}$ as the closure of all achievable rate pairs in our mixture setting. Define

$$R_{\max}(R_c) = \sup\{R_i \mid (R_i, R_c) \in \mathcal{R}_{mix}\}. \tag{5.35}$$

The following theorem provides the characterization for the mixture model.

Theorem 5.4 *For a given compression rate R_c the supremum of the achievable identification rate, according to Definition 5.5, is given by*

$$\begin{aligned}R_{\max}(R_c) &= \min_{s\in\mathcal{S}} \max_{\substack{P_{U|X,S=s}:\ |\mathcal{U}|\leq|\mathcal{X}|+1\\ I(X;U|S=s)\leq R_c}} I(Y;U|S=s)\\ &= \min_{s\in\mathcal{S}} \theta^s(R_c),\end{aligned} \tag{5.36}$$

which is independent of P_S.

Theorem 5.4 rediscovers a familiar fact that for a mixture model, the optimal trade-off does not depend on the mixing probability distribution, here P_S, cf. for example [Han03, Theorem 1.4.2, Theorem 5.10.1, Theorem 3.3.1].

Proof of Theorem 5.4

We first show that

$$R_{\max}(R_c) \geq \min_{s\in\mathcal{S}} \theta^s(R_c).$$

By the achievability proof of Theorem 5.2, there exists a sequence of mappings $(\tilde{\phi}_n, \tilde{\psi}_n)$ such that when $R_c > \max_s I(X;U|S=s)$ and $R_i < \min_s I(Y;U|S=s)$ where $|\mathcal{U}| \leq |\mathcal{X}| + 1$ then for every $\epsilon > 0$ we have

$$\sup_{(s_i)_{i=1}^M} \Pr\{\hat{W} \neq W | (S_i)_{i=1}^M = (s_i)_{i=1}^M\} \leq \epsilon, \tag{5.37}$$

[5.5] Since the pair of mappings (ϕ_n, ψ_n) is fixed beforehand, taking the limit does not change the value of the right-hand side.

for all $n \geq n_0(\epsilon)$. This implies that for every $\epsilon > 0$ we have

$$\Pr\{\hat{W} \neq W\} \leq \epsilon, \ \forall n \geq n_0(\epsilon). \tag{5.38}$$

Take a sequence $\epsilon_k \to 0$. Let $(\phi_{k,n}, \psi_{k,n})$ be the corresponding sequences of mappings such that (5.38) is satisfied with ϵ_k for the aforementioned (R_c, R_i). We construct a sequence of mappings (ϕ_n, ψ_n) such that if $n_0(\epsilon_k) \leq n < n_0(\epsilon_{k+1})$ then $\phi_n = \phi_{k,n}$ and $\psi_n = \psi_{k,n}$. This leads to

$$\begin{aligned} \frac{1}{n}\log|\mathcal{M}_1| \leq R_c + \epsilon_k, \ \frac{1}{n}\log M \geq R_i - \epsilon_k \\ \Pr\{\hat{W} \neq W\} \leq \epsilon_k, \end{aligned} \tag{5.39}$$

which further implies that $(R_c, R_i) \in \mathcal{R}_{mix}$. Hence, since $\mathcal{R}_{mix}$ is closed,

$$\begin{aligned} R_{\max}(R_c) &\geq \sup_{\substack{\{P_{U|X,S=s}\}_{s\in\mathcal{S}}:\ |\mathcal{U}|\leq|\mathcal{X}|+1 \\ \max_s I(X;U|S=s)\leq R_c}} \min_s I(Y;U|S=s) \\ &\overset{(\star)}{=} \min_s \sup_{\substack{P_{U|X,S=s}:\ |\mathcal{U}|\leq|\mathcal{X}|+1 \\ I(X;U|S=s)\leq R_c}} I(Y;U|S=s) \\ &= \min_s \theta^s(R_c). \end{aligned} \tag{5.40}$$

$(\star)$ holds since in the optimization domain the constraint on $P_{U|X,S=s}$ does not depend on the others $\{P_{U|X,S=s'}\}_{s'\neq s}$[5.6].

Now we show the converse direction, i.e.,

$$R_{\max}(R_c) \leq \min_s \theta^s(R_c).$$

It suffices to consider all *achievable* pairs (R_c, R_i). Let (ϕ_n, ψ_n) be a pair of mappings such that (R_c, R_i) is achievable. Similarly, let $n_0(\gamma)$ be such that (5.32) is satisfied for all $n \geq n_0(\gamma)$. Since

$$P_{Y^n\phi_n(X^n)}(y^n, \phi_n(x^n)) = \sum_s P_s P^s_{Y^n\phi_n(X^n)}(y^n, \phi_n(x^n)), \tag{5.41}$$

we obtain that $\underline{I}(\mathbf{Y};\mathbf{U}) = \min_s \underline{I}^s(\mathbf{Y}, \phi_n(\mathbf{X}))$, cf. [Han03, Lemma 3.3.2], where the superscript s denotes the evaluation w.r.t. the distribution

$$P^s_{Y^n\phi_n(X^n)}(y^n, m_1) = \sum_{x^n:\ \phi_n(x^n)=m_1} P^{\otimes n}_{Y|X}(y^n|x^n) P^{\otimes n}_{X|S=s}(x^n).$$

[5.6] In more details, denote the corresponding values for the left-hand and right-hand side by α_{opt} and β_{opt}, respectively. Both quantities are finite. The inequality $\alpha_{\text{opt}} \leq \beta_{\text{opt}}$ is straightforward since the domain is relaxed. Given $\epsilon > 0$ there exists for each $s \in \mathcal{S}$ a conditional distribution $\hat{P}_{U|X,S=s}$ such that $I(X;U|S=s) \leq R_c$ and $I(Y;U|S=s) > \beta_{\text{opt}} - \epsilon$. Since the collection $\{\hat{P}_{U|X,S=s}\}_{s\in\mathcal{S}}$ is in the optimizing domain of the left-hand side we obtain $\alpha_{\text{opt}} > \beta_{\text{opt}} - \epsilon$. Since ϵ is arbitrary, the reverse direction holds.

We then have

$$R_i \leq \min_s \underline{I}^s(\mathbf{Y}; \phi_n(\mathbf{X})). \tag{5.42}$$

Next we use the following well-known relation [Han03, Theorem 3.5.2] that the spectral-inf mutual information rate is less than or equal to the inf-mutual information rate

$$\underline{I}^s(\mathbf{Y}; \phi_n(\mathbf{X})) \leq \liminf_{n\to\infty} \frac{1}{n} I^s(Y^n; \phi_n(X^n)),\ \forall s \in \mathcal{S}. \tag{5.43}$$

For a given $s \in \mathcal{S}$ let $(n_{s,k})$ be a subsequence of indices such that the subsequence $\big(\frac{1}{n_{s,k}} I^s(Y^{n_{s,k}}; \phi_{n_{s,k}}(X^{n_{s,k}}))\big)$ converges to the corresponding lim inf-limit. Then for all $n_{s,k} \geq n_{s,k}(\gamma) \geq n_0(\gamma)$ we have

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} I^s(Y^n; \phi_n(X^n)) &\leq \frac{1}{n_{s,k}} I^s(Y^{n_{s,k}}; \phi_{n_{s,k}}(X^{n_{s,k}})) + \gamma \\
&\overset{(a)}{\leq} \sup_{\phi_{n_{s,k}}:\, |\mathcal{M}_1| \leq e^{n_{s,k}(R_c+\gamma)}} \frac{1}{n_{s,k}} I^s(Y^{n_{s,k}}; \phi_{n_{s,k}}(X^{n_{s,k}})) + \gamma \\
&\leq \sup_n \sup_{\phi_n:\, |\mathcal{M}_1| \leq e^{n(R_c+\gamma)}} \frac{1}{n} I^s(Y^n; \phi_n(X^n)) + \gamma \\
&\overset{(b)}{=} \theta^s(R_c + \gamma) + \gamma
\end{aligned} \tag{5.44}$$

where

$$\theta^s(R_c) = \max_{\substack{P_{U|X,S=s}:\, |\mathcal{U}| \leq |\mathcal{X}|+1 \\ I(X;U|S=s) \leq R_c}} I(Y; U|S=s). \tag{5.45}$$

(a) is valid due to (5.32) as $n_{s,k} \geq n_0(\gamma)$. (b) holds due to the entropy characterization [AC86, Theorem 2]. Therefore we end up with

$$R_i \leq \min_s \theta^s(R_c + \gamma) + \gamma. \tag{5.46}$$

Taking $\gamma \to 0$ we obtain $R_i \leq \min_s \theta^s(R_c)$ as $\theta^s(R_c)$ is continuous for $R_c > 0$. In case that $R_c = 0$, it can be seen that $R_i = 0$ (for example along the line of using Fano's inequality[5.7]). The converse direction is hence shown.

Remark 5.2 If we select the conditional distribution $P_{U^n|X^n}$ such that

$$P_{X^nU^n}(x^n, u^n) = \sum_{s\in\mathcal{S}} P_s P_{XU|S=s}^{\otimes n}$$

[5.7] This can be seen as follows. We have $n(R-\gamma) \leq H(W) \leq I(Y^n, \mathbf{J}; W) + H(W|Y^n, \mathbf{J}) \leq I(Y^n; W|\mathbf{J}) + n\epsilon_n \leq I(Y^n; J_W, W) + n\epsilon_n = I(Y^n; J_W|W) + n\epsilon_n \leq n(R_c + \gamma + \epsilon_n)$. The third inequality holds due to Fano's inequality and the fact that $\mathbf{J}$ is independent of W. The equality is valid since Y^n is independent of W. Alternatively, due to the data processing inequality we have $\theta^s(R_c) \leq R_c$. This implies further that $\theta^s(R_c)$ is a right-continuous function at $R_c = 0$.

where U^n takes values on the Cartesian product set $\mathcal{U}^n$ with $|\mathcal{U}| \leq |\mathcal{X}| + 1$ then it can be seen that, cf. [Han03, Lemma 5.10.1] that $\bar{I}(\mathbf{X};\mathbf{U}) = \max_s I^s(X;U|S=s)$ and $\underline{I}(\mathbf{Y};\mathbf{U}) = \min_{s\in\mathcal{S}} I^s(Y;U|S=s)$. Therefore (5.40) also follows from Theorem 5.3. An advantage of this approach is that we do not need to assume that the distributions $P_{Y|S=s}$ are all different as in the proof of Theorem 5.2. However the latter approach does not use the underlying distribution $\{P_s\}_{s\in\mathcal{S}}$ in the construction.

Remark 5.3 Since $\theta^s(R_c)$ is a concave function of R_c [AC86, Lemma 1], $R_{\max}(R_c)$ is a concave in R_c. Therefore the region $\mathcal{R}_{mix}$, cf. $\mathcal{R}_{iis}$, is convex.

Generalized mixture models

In the following we generalize the mixture setting in two directions. We keep using Definition 5.5 for the achievability. In the first generalization, the distribution of users' data is kept as in (5.28). The observation channel is modeled as

$$P_{Y^n|X^n} = \sum_{\tau\in\mathcal{T}} P_\tau P_{Y|X,\tau}^{\otimes n}, \tag{5.47}$$

where $\mathcal{T}$ is a finite set.

Theorem 5.5 *$\bar{\mathcal{R}}_{sc}$ characterizes the compression-identification trade-off for the model given by* (5.28) *and* (5.47)

Proof. By choosing the conditional distribution $P_{U^n|X^n}$ as in Remark 5.2, Theorem 5.3 tells us that the set of all rate pairs (R_c, R_i) such that

$$R_c \geq \max_s I(X;U|S=s),\ R_i \leq \min_{s,\tau} I(Y;U|S=s,T=\tau), \tag{5.48}$$

where $P_{YXU|S=s,T=\tau} = P_{Y|X,\tau} \times P_{X|S=s} \times P_{U|X,S=s}$ with $|\mathcal{U}| \leq |\mathcal{X}| + |\mathcal{T}|$ is achievable.

Additionally, it can be seen that for all sufficiently large n we have

$$R_c + \gamma \geq \frac{1}{n} H(\phi_n(X^n)|S=s) \geq \frac{1}{n} I^s(X^n;\phi_n(X^n)). \tag{5.49}$$

Using similar steps as in the converse proof for Theorem 5.4 we obtain the corresponding inequality for each pair (s,τ)

$$R_i \leq \liminf_{n\to\infty} \frac{1}{n} I^{s,\tau}(Y^n;\phi_n(X^n)) \leq \frac{1}{n_{s,\tau,k}} I^{s,\tau}(Y^{n_{s,\tau,k}};\phi_{n_{s,\tau,k}}(X^{n_{s,\tau,k}})) + \gamma, \tag{5.50}$$

for all $n_{s,\tau,k} \geq n_{s,\tau,k}(\gamma)$. We can then apply the same single-letterizing steps as in the converse proof of Theorem 5.1 to show that $(R_c+\gamma, R_i-\gamma)$ is inside the region defined by (5.48), cf. $\bar{\mathcal{R}}_{sc}$. This implies $\bar{\mathcal{R}}_{sc}$ characterizes the compression-identification trade-off. Additionally, applying a similar line of reasoning as in Footnote 5.6, we obtain that

$$R_{\max}(R_c) = \max_{\substack{\{P_{U|X,S=s}\}_{s\in\mathcal{S}}:\ |\mathcal{U}|\leq|\mathcal{X}|+|\mathcal{T}| \\ \max_s I(X;U|S=s)\leq R_c}} \min_{s,\tau} I(Y;U|S=s,T=\tau)$$

$$= \min_{s\in\mathcal{S}} \max_{\substack{P_{U|X,S=s}:\, |\mathcal{U}|\le|\mathcal{X}|+|\mathcal{T}| \\ I(X;U|S=s)\le R_c}} \min_{\tau} I(Y;U|S=s,T=\tau) \tag{5.51}$$

as the multiple mutual information terms $\{I(Y;U|S=s,T=\tau)\}_{\tau\in\mathcal{T}}$ share the same $P_{U|X,S=s}$. □

In the second direction we generalize (5.28) by

$$P_{X^n} = \sum_{s\in\mathcal{S}} P_s P_{X|S=s}^{\otimes n} \tag{5.52}$$

to allow $\mathcal{S}$ to be a *countably infinite* alphabet and $P_s > 0$ for all $s \in \mathcal{S}$. The observation channel $P_{Y|X}$ is fixed. Denote the closure of all achievable rate pairs by $\mathcal{R}_{enum}$ and define $R_{\max}(R_c) = \sup\{R_i|(R_c,R_i)\in\mathcal{R}_{enum}\}$. The following theorem characterizes $R_{\max}(R_c)$ for the setting given by (5.52)

Theorem 5.6

$$R_{\max}(R_c) = \inf_{s\in\mathcal{S}} \theta^s(R_c), \tag{5.53}$$

where $\theta^s(R_c)$ *is defined in* (5.45).

Proof. Let $\mathbf{U}$ be a discrete process where U^n takes values on the Cartesian product set $\mathcal{U}^n$ such that $P_{X^nU^n} = \sum_{s\in\mathcal{S}} P_s P_{XU|S=s}^{\otimes n}$. Following the proof of [Han03, Lemma 3.3.2] we obtain that $\bar{I}(\mathbf{X};\mathbf{U}) = \sup_{s\in\mathcal{S}} I(X;U|S=s)$.

Applying the above calculation to the result of Theorem 5.3 we obtain that[5.6]

$$R_{\max}(R_c) \ge \sup_{\substack{\{P_{U|X,S=s}\}_{s\in\mathcal{S}}:\, |\mathcal{U}|\le|\mathcal{X}|+1 \\ \sup_s I(X;U|S=s)\le R_c}} \inf_s I(Y;U|S=s) \tag{5.54}$$

$$= \inf_{s\in\mathcal{S}} \theta^s(R_c). \tag{5.55}$$

Similarly, given an achievable rate pair (R_c,R_i) we obtain for the reverse direction the expression (5.44). Since γ is arbitrarily we see that $R_i \le \theta^s(R_c)$ for all $s\in\mathcal{S}$. Let $(R_c,R_i)\in\mathcal{R}_{enum}$ be any rate pair, then there exists a sequence of achievable rate pairs $(R_{c,k},R_{i,k})$ such that $R_{c,k}\to R_c$ and $R_{i,k}\to R_i$ as $k\to\infty$. Using the given bound we obtain

$$\begin{aligned} R_i &= \limsup_{k\to\infty} R_{i,k} \le \limsup_{k\to\infty} \inf_{s\in\mathcal{S}} \theta^s(R_{c,k}) \\ &\le \inf_{s\in\mathcal{S}} \limsup_{k\to\infty} \theta^s(R_{c,k}) = \inf_{s\in\mathcal{S}} \theta^s(R_c). \end{aligned} \tag{5.56}$$

Therefore $R_{\max}(R_c) \le \inf_{s\in\mathcal{S}} \theta^s(R_c)$. □

5.2.D A connection ϵ-achievable regions

We now show a dual statement to Proposition 5.1. For simplicity we cheat a bit by again using Definition 5.5 for $\mathcal{R}_{\mathrm{ID}}$.

Proposition 5.2 *Assume that $(R_i, R_c) \in \mathcal{R}_{\mathrm{ID}}$ is achievable, where $\mathcal{R}_{\mathrm{ID}}$ is now defined according to Definition 5.5. Then there exists a* WAK*-code which corresponds to the identification-compression scheme such that $(R_c, H(Y) - R_i)$ is in $\mathcal{R}_{\mathrm{WAK}}$.*

Proof. Given a pair of compression-identification mappings (ϕ_n, ψ_n) such that the rate pair (R_i, R_c) is achievable. We construct a code for the WAK-problem as follows. For notation clarity, we denote the sources in the WAK problem by $(\bar{X}^n, \bar{Y}^n) \sim P_{XY}^{\otimes n}$. Randomly assign each sequence y^n to a bin $\mathcal{B}(m_2)$ where $m_2 \in [1 : e^{nR_2}]$. Define the decoding set $\mathcal{D}_n(m_1)$ as

$$\mathcal{D}_n(m_1) = \left\{ y^n \middle| \frac{1}{n} \log \frac{P_{Y^n|\phi_n(X^n)}(y^n|m_1)}{P_{Y^n}(y^n)} \geq R_i - 2\gamma \right\}, \tag{5.57}$$

We define a decoder for the WAK-problem as follows. If $\hat{y}^n$ is a unique sequence such that $\hat{y}^n \in \mathcal{B}(m_2) \cap \mathcal{D}_n(m_1) \cap \mathcal{A}_\gamma^n$, where $\mathcal{A}_\gamma^n$ is the weakly typical set, then $\hat{y}^n$ is output as the reconstruction sequence. Let $\bar{M}_1$ and $\bar{M}_2$ be the encoded messages at Encoder 1 and 2. Applying Lemma 5.2 we obtain

$$\Pr\{\bar{Y}^n \notin \mathcal{D}_n(\bar{M}_1) \cap \mathcal{A}_\gamma^n\} \leq \Pr\{\hat{W} \neq W\} + e^{-n\gamma} + \Pr\{\bar{Y}^n \notin \mathcal{A}_\gamma^n\}, \tag{5.58}$$

for all $n \geq n^*(\gamma)$. Furthermore, we have

$$\begin{aligned}
&\Pr\{\exists \hat{y}^n \neq \bar{Y}^n, \hat{y}^n \in \mathcal{B}(\bar{M}_2) \cap \mathcal{D}_n(\bar{M}_1) \cap \mathcal{A}_\gamma^n\} \\
&\quad \leq \sum_{(m_1, y^n)} P_{Y^n \phi_n(X^n)}(y^n, m_1) \\
&\qquad \times \sum_{m_2} \Pr\{\bar{M}_2 = m_2 | Y^n = y^n\} \sum_{\hat{y}^n \in \mathcal{D}_n(m_1) \cap \mathcal{A}_\gamma^n} \Pr\{\hat{y}^n \in \mathcal{B}(m_2)\} \\
&\quad \leq e^{-nR_2} e^{n(H(Y) - R_i + 3\gamma)},
\end{aligned} \tag{5.59}$$

where the last inequality holds due to (5.138). Therefore the rate pair $(R_c, H(Y) - R_i)$ is achievable for the WAK-problem. □

Given $\epsilon \in [0, 1)$. We recall the definitions of ϵ-achievability of both WAK-problem and ID-problem in the table on the top of the next page.
We first have the following observation which is a strong converse w.r.t. the identification rate, see also [YY16] for another strong converse statement.

Proposition 5.3 *Given $\epsilon > 0$ if $(R_c, R_i) \in \mathcal{R}_{\mathrm{ID},\epsilon}$ then $R_i \leq I(Y; X)$.*

$\mathcal{R}_{\mathrm{WAK},\epsilon}$ is the closure of (R_1, R_2) s.t.	$\mathcal{R}_{\mathrm{ID},\epsilon}$ is the closure of (R_c, R_i) s.t.
$\limsup_{n\to\infty} \frac{1}{n} \log \lvert\mathcal{M}_k\rvert \le R_k,\ k = 1, 2$	$\limsup_{n\to\infty} \frac{1}{n} \log \lvert\mathcal{M}_1\rvert \le R_c,$
$\limsup_{n\to\infty} \Pr\{\hat{Y}^n \neq Y^n\} \le \epsilon,$	$\liminf_{n\to\infty} \frac{1}{n} \log \lvert\mathcal{W}\rvert \ge R_i,$
	$\limsup_{n\to\infty} \Pr\{\hat{W} \neq W\} \le \epsilon.$

Proof. Suppose that $(R_c, R_i) \in \mathcal{R}_{\mathrm{ID},\epsilon}$ with $R_i = I(X;Y) + 3\gamma$ for some $\gamma > 0$. It suffices to consider the uncompressed scenario. By Lemma 5.2 and the weak law of large numbers we obtain for any identification mapping

$$\Pr\{\hat{W} \neq W\} \to 1, \text{ as } n \to \infty, \tag{5.60}$$

a contradiction to $(R_c, R_i) \in \mathcal{R}_{\mathrm{ID},\epsilon}$. □

The following theorem generalizes the results from Propositions 5.1 and 5.2.

Theorem 5.7 *Given $(R_a, R_b) \in \mathbb{R}_+^2$ with $R_b \le H(Y)$,*

$$(R_a, R_b) \in \mathcal{R}_{\mathrm{WAK},\epsilon} \Leftrightarrow (R_a, H(Y) - R_b) \in \mathcal{R}_{\mathrm{ID},\epsilon}, \ \ \forall \epsilon \in [0, 1). \tag{5.61}$$

Theorem 5.7 and Proposition 5.3 imply that for each $\epsilon > 0$, $\mathcal{R}_{\mathrm{ID},\epsilon}$ corresponds to the sub-region[5.8] of $\mathcal{R}_{\mathrm{WAK},\epsilon}$ with $R_2 \le H(Y)$. Hence a strong converse for the WAK-problem is equivalent to a strong converse for the identification problem.

Proof. ($\Leftarrow$) follows from the proof of Proposition 5.2.
($\Rightarrow$): Consider an *achievable* pair of $(R_a, R_b) \in \mathcal{R}_{\mathrm{WAK},\epsilon}$ with the corresponding mapping $(\phi_{1n}, \phi_{2n}, \psi_n)$. Similarly as in the proof of Proposition 5.1, given $\gamma, \delta > 0$, there exists $n_0(\delta)$ such that for all $n \ge n_0(\delta)$

$$\lvert\mathcal{M}_2\rvert \le e^{n(R_b+\delta)}. \tag{5.62}$$

In inequality (5.25) if we take $M = e^{n(H(Y)-R_b-2\gamma-\delta)}$ then the last term in (5.25) goes to 0, as $n \to \infty$. Therefore

$$\limsup_{n\to\infty} \Pr\{\hat{W} \neq W\} \le \epsilon, \tag{5.63}$$

and hence $(R_a, H(Y) - R_b - 2\gamma - \delta) \in \mathcal{R}_{\mathrm{ID},\epsilon}$. Since γ, δ are arbitrary, the forward direction follows. □

Remark 5.4 In Appendix 5.E we present a strong converse proof for the ID-problem for both discrete and Gaussian settings. The arguments therein resemble the ones used in the strong converse proof of the WAK problem [AGK76, Theorem 3].

[5.8] The $\mathcal{R}_{\mathrm{WAK},\epsilon}$ also includes all tuples (R_1, R_2) with $R_2 > H(Y)$.

5.2.E Arbitrarily varying observation channel

Consider the following scenario where $\mathcal{X}$ and $\mathcal{Y}$ are also finite. We assume that the data is generated iid according to P_X where $P_X(x) > 0$ for all $x \in \mathcal{X}$ and enrolled into the database. An observer obtains the observation y^n from one user in the system which is arbitrarily corrupted and wants to search for the true user. Denote by

$$\mathcal{P}_c = \{P_{Y|X,s}(\cdot|\cdot, s) \mid s \in \mathcal{S}\} \tag{5.64}$$

the set of possible channel distributions. For simplicity[5.9] we also assume that $P_{Y|X,s} > \eta > 0$ for all x, y, s. Assume that the observation channel is given for a state sequence s^n by

$$P_{Y^n|X^n,s^n}(y^n|x^n) = \prod_{i=1}^{n} P(y_i|x_i, s_i), \quad \forall s^n. \tag{5.65}$$

Accordingly, the average error expression for a given enrollment-identification mapping pair (ϕ_n, ψ_n) and a state sequence s^n is given by

$$\begin{aligned} e(s^n, \phi_n, \psi_n) &= \Pr\{W \neq \hat{W}|s^n\} \\ &= \sum_w \frac{1}{M} \sum P((j_i)_{i=1, i\neq w}^{M}) P(j_w, x^n(w)) \\ &\qquad \times P_{Y^n|X^n,s^n}(y^n|x^n) \chi_{\{\psi_n(y^n, (j_i)_{i=1}^{M}) \neq w\}}. \end{aligned} \tag{5.66}$$

Definition 5.6 A pair (R_c, R_i) is achievable if for every $\delta > 0$ there exist mappings (ϕ_n, ψ_n) such that

$$\begin{aligned} \frac{1}{n} \log |\mathcal{M}_1| < R_c + \delta, \quad \frac{1}{n} \log |\mathcal{W}| > R_i - \delta, \\ \text{and} \sup_{s^n \in \mathcal{S}^n} e(s^n, \phi_n, \psi_n) < \delta, \end{aligned} \tag{5.67}$$

for all $n \geq n_0(\delta)$. The set of all achievable pairs is denoted by $\mathcal{R}_{avc}$.

Denote by $\bar{\mathcal{P}}_c$ the convex closure of $\mathcal{P}_c$, i.e., the set of all conditional distributions P on $\mathcal{Y}$ given $\mathcal{X}$ such that

$$P(y|x) = \sum_{s \in \mathcal{S}} P_s P_{Y|X,s}(y|x, s), \tag{5.68}$$

where $\{P_s, s \in \mathcal{S}\}$ is a distribution on $\mathcal{S}$. Furthermore, for each $\tilde{P} \in \bar{\mathcal{P}}_c$ the corresponding achievable compression-identification rate region is given by

$$\bar{\mathcal{R}}_{\tilde{P}} = \Big\{ (R_c, R_i) \mid R_c \geq I(\bar{X}; \bar{U}), R_i \leq I(\bar{Y}; \bar{U}),$$

[5.9] This technical assumption allows us to avoid terms of the form $a \log 0$ where $a > 0$ when taking limit as in [CK11, Eq. 12.19].

$$P_{\bar{Y}\bar{X}\bar{U}}(y,x,u) = \tilde{P}(y|x)P_X(x)P_{\bar{U}|\bar{X}}(u|x),\ |\bar{\mathcal{U}}| \leq |\mathcal{X}|+1\Big\}.$$

The following proposition provides an outer bound for the achievable rate region.

Proposition 5.4 $\mathcal{R}_{avc} \subseteq \bigcap_{p\in\bar{\mathcal{P}}_c} \bar{\mathcal{R}}_p$.

The proof of Proposition 5.4 is given in Appendix 5.F. An inner bound for the *achievable* rate region is given in the following proposition.

Proposition 5.5 *If users' data sequence are uncompressed in the database then*

$$R_i \leq \inf_{P_{Y|X}\in\bar{\mathcal{P}}_c} I(P_X, P_{Y|X}). \tag{5.69}$$

The proof of Proposition 5.5 is identical to the proof of [CK11, Lemma 12.10] hence omitted since the average error probability (over random codewords $X^n(m) \sim P_X^{\otimes n}$) therein can be also viewed as the identification error probability in our setting. Propositions 5.4 and 5.5 establish the following result.

Corollary 5.1 *The maximum achievable identification rate, identification capacity, for the given model is given by*

$$C = \min_{P_{Y|X}\in\bar{\mathcal{P}}_c} I(P_X, P_{Y|X}). \tag{5.70}$$

Remark 5.5 The expression (5.70) is similar to the (transmission) capacity of the arbitrarily varying channel $\mathcal{P}_c$ achieved using random codes. The only difference is that in our setting the channel input distribution P_X is fixed. It will be clear from the proof of Proposition 5.5 that there is no need for a randomness elimination procedure to achieve the identification capacity as in [Ahl78]. In other words, C can be achieved in the current setting using deterministic mappings. We believe that the randomness in the users' data and the deterministic mappings help to simulate as random code for the corresponding observation channel.

5.A Proof of Lemma 5.1

Let d be a metric on the set of probability measures $\mathcal{M}_1(\mathcal{X})$ that induces the weak* topology, for example the Prohorov metric, and define

$$t_X = \min_{s,s'} d(P_{X|S=s}, P_{X|S=s'}). \tag{5.71}$$

As $|\mathcal{S}|$ is finite, $t_X > 0$. For each x^n the empirical distribution is denoted by P_{x^n}, which is given by

$$P_{x^n}(\mathcal{A}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(\mathcal{A}), \tag{5.72}$$

where $\mathcal{A} \in \mathcal{F}$, and δ_x is the corresponding Dirac measure. Define

$$\mathcal{A}_s^X = \big\{x^n \mid d(P_{x^n}, P_{X|S=s}) < \frac{t_X}{2}\big\},\ \forall s \in \mathcal{S}. \tag{5.73}$$

By the triangular inequality we see that the sets $\mathcal{A}_s^X$ are disjoint. The corresponding classifier is given by

$$T(x^n) \mapsto \begin{cases} s & \text{if } x^n \in \mathcal{A}_s^X \\ e & \text{otherwise} \end{cases}. \tag{5.74}$$

Furthermore we also see that, if the elements of X^n are generated iid from the distribution $P_{X|S=s}$ then

$$\Pr(X^n \notin \mathcal{A}_s^X | S = s) \to 0, \text{ as } n \to \infty \tag{5.75}$$

due to [Mit15, Theorem 4.4].

5.B Proof of Theorem 5.1

Achievability

Let T_X be a classifier for s and T_Y be a classifier for κ from Lemma 5.1.
Let U be a random variable[5.10] such that given $S = s$ and $X = x$ it is distributed according to the law $P_{U|X=x,S=s}$. Also fix a $\delta > 0$. For each $s \in \mathcal{S}$ generate 2^{nR_c} sequences $u^n(m_s) \sim P_{U|S=s}$, $m_s \in [1 : e^{nR_c}]$, where $P_{U|S=s}$ is the marginal corresponding to $P_{UX|S=s}$. Hence, we have a total $|\mathcal{S}|e^{nR_c}$ codeword sequences which are used to enroll all users' data.
For each $s \in \mathcal{S}$ and $\tau \in \mathcal{T}$ assume that the tuple

$$(\hat{X}^n, \hat{Y}^n, \hat{U}^n) \sim (P_{UX|S=s} \times P_{Y|X,\tau})^{\otimes n}.$$

Denote $\mathcal{A}_s = \mathcal{A}_\delta^n(P_{XU|S=s})$, where the latter is the weakly typical set[5.11]. Note that

$$\Pr\{(\hat{X}^n, \hat{U}^n) \notin \mathcal{A}_s | S = s\} \to 0, \text{ as } n \to \infty. \tag{5.76}$$

Moreover, define

$$\phi_{s,\tau}(x^n, y^n, u^n) = \chi_{\left\{(y^n,u^n) \notin \mathcal{A}_\delta^n(P_{YU|\tau s})\right\}}, \tag{5.77}$$

then $\delta_{n,s\tau} = \mathbb{E}[\phi_{s,\tau}(\hat{X}^n, \hat{Y}^n, \hat{U}^n)|S = s, T = \tau] \to 0$, as $n \to \infty$, due to the weak law of large numbers. Define, hence,

$$\mathcal{B}_{s,\tau} = \big\{(x^n, u^n) \mid \mathbb{E}[\phi_{s,\tau}(x^n, \hat{Y}^n, u^n)|\hat{X}^n = x^n, S = s, T = \tau] \le \delta_{n,s\tau}^{1/2}\big\}, \tag{5.78}$$

[5.10] We can choose for each s a different random variable U_s. Herein we abuse the notation for brevity.
[5.11] $\{P_{U|X=x,S=s}\}_{s\in\mathcal{S}}$ must be chosen such that the corresponding entropy terms (discrete or continuous) exist.

and accordingly, let $\hat{\mathcal{A}}_s = \mathcal{A}_s \cap \bigcap_\tau \mathcal{B}_{s,\tau}$. By Markov's inequality we have

$$\Pr\{\mathcal{B}_{s,\tau}^c | S = s\} \leq \frac{\mathbb{E}[\phi_{s,\tau}(\hat{X}^n, \hat{Y}^n, \hat{U}^n)|S = s, T = \tau]}{\delta_{n,s\tau}^{1/2}} = \delta_{n,s\tau}^{1/2} \to 0, \ \forall \tau \in \mathcal{T}, \quad (5.79)$$

so that

$$\Pr\{\hat{\mathcal{A}}_s | S = s\} \to 1, \text{ as } n \to \infty. \quad (5.80)$$

Enrollment: For each user i, we first let $\hat{s}_i = T_X(x^n(i))$. Assume that $\hat{s}_i \in \mathcal{S}$. Then we search for a codeword such that

$$(x^n(i), u^n(m_{\hat{s}_i,i})) \in \hat{\mathcal{A}}_{\hat{s}_i}. \quad (5.81)$$

We then store the corresponding pair of $(\hat{s}_i, m_{\hat{s}_i,i})$ as j_i in the database. Note that when either $\hat{s}_i = e$ or $m_{\hat{s}_i,i} = e$, i.e., there does not exist a codeword such that the above condition is satisfied, the corresponding stored index is e.
Identification: The processing unit first assigns the observation y^n to one of $|\mathcal{P}_\mathcal{Y}|$ states or e. We denote the corresponding state by $\kappa^\star = T_Y(y^n)$. We use $s^\star = \hat{s}_1$ as an estimate of the state of users' data. If $\kappa^\star$ is not compatible with $s^\star$, i.e., if $\kappa^\star = e$ or $s^\star = e$ or

$$\nexists \tau \colon P_{Y,\kappa^\star} = P_{Y|X,\tau} P_{X|S=s^\star}, \quad (5.82)$$

the processing center aborts the operations and declares an error. Otherwise, denote the corresponding channel index by $\hat{\tau}$. We then search for a unique $\hat{w}$ such that

$$(y^n, u^n(m_{\hat{s}_{\hat{w}},\hat{w}})) \in \mathcal{A}_\delta^n(P_{YU|\hat{\tau}s^\star}), \quad (5.83)$$

where $P_{YU|\hat{\tau}s^\star}$ is the marginal of $P_{Y|X,\hat{\tau}} \times P_{X|S=s^\star} \times P_{U|X,S=s^\star}$. An error is declared if there does not exist any such index or there is more than one.
Analysis: Let $\hat{S}_W$, $M_{\hat{S}_W,W}$, $S^\star$ and $\hat{T}$ be the corresponding random variables induced by the enrollment and identification processes. Without loss of generality we condition on the event $W = 1$, $S = s$ and $T = \tau$. If $\hat{W} \neq 1$ then at least one of the following events occurs

$$\begin{aligned}
\mathcal{E}_{es} &= \{T_X(X^n(1)) \neq s\} \cup \{T_Y(Y^n) \neq \kappa\} \\
\mathcal{E}_{no,sc} &= \{(Y^n, U^n(M_{\hat{S}_1,1})) \notin \mathcal{A}_\delta^n(P_{YU|\hat{T}S^\star})\} \\
\mathcal{E}_{\geq 2,sc} &= \{\exists w' \neq 1 \mid (Y^n, U^n(M_{\hat{S}_{w'},w'})) \in \mathcal{A}_\delta^n(P_{YU|\hat{T}S^\star})\}.
\end{aligned}$$

We define the following events

$$\begin{aligned}
\mathcal{E}_{xu} &= \{(X^n(1), U^n(m_{s,1})) \notin \hat{\mathcal{A}}_s, \forall m_{s,1}\} \\
\mathcal{E}_1 &= \{(Y^n, U^n(M_{s,1})) \notin \mathcal{A}_\delta^n(P_{YU|T=\tau,S=s})\}.
\end{aligned} \quad (5.84)$$

Then

$$\Pr\{W \neq \hat{W} | W = 1, S = s, T = \tau\}$$

$$
\begin{aligned}
&\leq \Pr\{\mathcal{E}_{es} \cup \mathcal{E}_{no,sc} \cup \mathcal{E}_{\geq 2,sc} | W=1, S=s, T=\tau\} \\
&\leq \Pr\{\mathcal{E}_{es} | W=1, S=s, T=\tau\} \\
&\quad + \Pr\{\mathcal{E}_{es}^c \cap \mathcal{E}_{no,sc} | W=1, S=s, T=\tau\} \\
&\quad + \Pr\{\mathcal{E}_{es}^c \cap \mathcal{E}_{\geq 2,sc} | W=1, S=s, T=\tau\}. \qquad (5.85)
\end{aligned}
$$

From the previous analysis we know that

$$
\begin{aligned}
&\Pr\{T_X(X^n(1)) \neq s | S=s\} \to 0, \\
&\quad \text{and, } \Pr\{T_Y(Y^n) \neq \kappa | W=1, T=\tau, S=s\} \to 0, \qquad (5.86)
\end{aligned}
$$

as $n \to \infty$. Hence $\Pr\{\mathcal{E}_{es} | W=1, S=s, T=\tau\} \to 0$ as $n \to \infty$. The second term in (5.85) is upper bounded further as in (5.87). The last inequality holds since W is independent of users' data sequences, the codebook and the states.

$$
\begin{aligned}
&\Pr\{\mathcal{E}_{es}^c \cap \mathcal{E}_{no,sc} | W=1, S=s, T=\tau\} \\
&\leq \Pr\{\mathcal{E}_{es}^c \cap \{(X^n(1), U^n(M_{\hat{S}_1,1})) \notin \hat{\mathcal{A}}_{\hat{S}_1}\} | W=1, S=s, T=\tau\} \\
&+ \Pr\{\mathcal{E}_{es}^c \cap \{(X^n(1), U^n(M_{\hat{S}_1,1})) \in \hat{\mathcal{A}}_{\hat{S}_1}\} \cap \mathcal{E}_{no,sc} | W=1, S=s, T=\tau\} \\
&\leq \Pr\{\mathcal{E}_{xu} | S=s\} + \underbrace{\Pr\{\{(X^n(1), U^n(M_{s,1})) \in \hat{\mathcal{A}}_s\} \cap \mathcal{E}_1\} | S=s, W=1, T=\tau\}}_{t_2(s,\tau)}. \qquad (5.87)
\end{aligned}
$$

For a given state $s \in \mathcal{S}$

$$
\begin{aligned}
&\Pr\{\mathcal{E}_{xu} | S=s\} \leq \Pr\big\{(\hat{X}^n, \hat{U}^n) \notin \hat{\mathcal{A}}_s | S=s\big\} \\
&+ \Pr[\iota(\hat{X}^n; \hat{U}^n | S=s) \geq nR_c - \gamma | S=s] + e^{-\exp(\gamma)} \to 0,
\end{aligned}
$$

as $n \to \infty$ if $\gamma = n\delta/2$ and $R_c > I(X;U|S=s) + \delta$. The inequality follows from the non-asymptotic covering lemma [Ver12], cf. also Lemma 2.1. This implies that

$$
\Pr\big\{\mathcal{E}_{xu} | S=s\big\} \to 0, \ \forall s \in \mathcal{S}, \qquad (5.88)
$$

if

$$
R_c > \max_s I(X;U|S=s) + \delta. \qquad (5.89)
$$

The constraint (5.89) can be viewed as a sufficient condition such that the encoding succeeds with high probability regardless of the unknown state.
Now we look at the term $t_2(s,\tau)$ which can be bounded as

$$
\begin{aligned}
t_2(s,\tau) = \int_{\hat{\mathcal{A}}_s} \int_{\mathcal{A}_\delta^n(P_{Y|U,\tau s}|u^n)^c} & dP_{Y|X,\tau}^{\otimes n}(y^n|x^n) \\
& \times dP_{X^n(1)U^n(M_{s,1})|SW}(x^n, u^n | s, 1) \leq \delta_{n,\tau s}^{1/2}, \qquad (5.90)
\end{aligned}
$$

since $\hat{\mathcal{A}}_s \subset \mathcal{B}_{s,\tau}$. Thus, $t_2(s,\tau) \to 0$ as $n \to \infty$.
Define

$$\mathcal{E}_2 = \{\exists w' \neq 1 \mid (Y^n, U^n(M_{\hat{S}_{w',w'}})) \in \mathcal{A}_\delta^n(P_{YU|\tau s})\}.$$

Then the (conditional) probability of the event $\mathcal{E}_{es}^c \cap \mathcal{E}_{\geq 2,sc}$ can be bounded as

$$\Pr\{\mathcal{E}_{es}^c \cap \mathcal{E}_{\geq 2,sc} | S = s, W = 1, T = \tau\} \leq \Pr\{\mathcal{E}_2 | S = s, T = \tau, W = 1\}. \tag{5.91}$$

We have

$$\begin{aligned}
&\Pr\{\mathcal{E}_2 | S = s, T = \tau, W = 1\} \\
&\leq e^{nR_i} \times \Pr\big\{(Y^n, U^n(M_{\hat{S}_2,2})) \in \mathcal{A}_\delta^n(P_{YU|\tau s}) \big| T = \tau, S = s, W = 1\big\} \\
&= e^{nR_i} \int \int_{\mathcal{A}_\delta^n(P_{Y|U,\tau s}|u^n)} dP_{Y|T=\tau,S=s}^{\otimes n}(y^n) dP_{U^n(M_{\hat{S}_2,2})|S}(u^n|s) \\
&\leq e^{nR_i} e^{-n(I(Y;U|T=\tau,S=s)-3\delta)}.
\end{aligned} \tag{5.92}$$

Therefore

$$\Pr\{\mathcal{E}_2 | T = \tau, S = s, W = 1\} \to 0, \ \forall s \in \mathcal{S} \tag{5.93}$$

as $n \to \infty$ if $R_i < \min_s I(Y;U|T = \tau, S = s) - 3\delta$.
In summary, if $n \geq n_0(\delta)$ then

$$\sup_{s,\tau} \Pr\{W \neq \hat{W} | T = \tau, S = s\} < \delta,$$

which implies that the rate pair

$$\big(\max_s I(X;U|S = s), \min_{s,\tau} I(Y;U|S = s, T = \tau)\big)$$

is in $\mathcal{R}_{sc}$.

Converse for the discrete case

Assume that the pair (R_c, R_i) is *achievable*, i.e., for every $\delta > 0$ there exists an identification scheme such that

$$\begin{aligned}
&\frac{1}{n} \log |\mathcal{M}_1| < R_c + \delta, \ \frac{1}{n} \log |\mathcal{W}| > R_i - \delta, \\
&\qquad \Pr\{\hat{W} \neq W | S = s, T = \tau\} < \delta, \ \forall s, \tau,
\end{aligned} \tag{5.94}$$

for all $n \geq n_0(\delta)$. Similarly as in Chapter 2 we abbreviate $(J_i)_{i=1}^M$ as $\boldsymbol{J}$ where $J_i = \phi_n(X^n(i))$ for all $i \in [1 : M]$. Then by Fano's inequality we obtain

$$H(W|Y^n, \boldsymbol{J}, S = s, T = \tau) < \log 2 + \delta \log |\mathcal{W}|, \ \forall (s,\tau) \in \mathcal{S} \times \mathcal{T}. \tag{5.95}$$

Define

$$U_i = (J_W, W, X^{i-1}(W)), \ \forall i \in [1 : n]. \tag{5.96}$$

Due to the memoryless property of the observation channel we have

$$\begin{aligned} &P_{Y_i X_i(W) U_i|S=s,T=\tau} = P_{Y_i|X_i(W),T=\tau} \times P_{X_i(W)|S=s} \times P_{U_i|X_i(W),S=s} \\ &= P_{Y|X,\tau} \times P_{X|S=s} \times P_{U_i|X_i(W),S=s},\ \forall i \in [1:n]. \end{aligned} \tag{5.97}$$

We start bounding the compression rate. For each pair $(s,\tau) \in \mathcal{S} \times \mathcal{T}$ we have

$$\begin{aligned} n(R_c+\delta) &\overset{(a)}{\geq} H(J_W|W,S=s) \geq I(X^n(W);J_W|W,S=s) \\ &= I(X^n(W);J_W,W|S=s) \\ &\overset{(b)}{=} \sum_{i=1}^{n} I(X_i(W);J_W,W,X^{i-1}(W)|S=s) \\ &= \sum_{i=1}^{n} I(X_i(W);U_i|S=s). \end{aligned} \tag{5.98}$$

(a) is valid since $\frac{1}{n}\log|\mathcal{M}_1| < R_c+\delta$ holds. In (b) we use the following property,

$$\Pr\{X^n(W) \in \mathcal{B}|S=s\} = P_{X|S=s}^{\otimes n}(\mathcal{B}),$$

in particular if $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_n$ then the independence follows. Also for each pair $(s,\tau) \in \mathcal{S} \times \mathcal{T}$

$$\begin{aligned} &n(R_i-\delta) \leq H(W|S=s,T=\tau) \\ &= I(Y^n,\boldsymbol{J};W|S=s,T=\tau) + H(W|Y^n,\boldsymbol{J},S=s,T=\tau) \\ &\overset{(c)}{\leq} I(Y^n;W|S=s,T=\tau,\boldsymbol{J}) + \delta_n \leq I(Y^n;W,\boldsymbol{J}|S=s,T=\tau) + \delta_n \\ &\overset{(d)}{=} I(Y^n;W,J_W|S=s,T=\tau) + \delta_n \overset{(e)}{=} \sum_{i=1}^{n} I(Y_i;W,J_W,Y^{i-1}|S=s,T=\tau) + \delta_n \\ &\overset{(f)}{\leq} \sum_{i=1}^{n} I(Y_i;W,J_W,X^{i-1}(W)|S=s,T=\tau) + \delta_n \\ &= \sum_{i=1}^{n} I(Y_i;U_i|S=s,T=\tau) + \delta_n, \end{aligned} \tag{5.99}$$

where $\delta_n = \log 2 + \delta \log|\mathcal{W}|$. (c) is true since W is independent of $(T,S,\boldsymbol{J})$. (d) is valid since

$$(Y^n,J_W) - (W,S,T) - (J_i)_{i=1,i\neq W}^{M}. \tag{5.100}$$

(e) is valid since conditioning on $T=\tau$ and $S=s$, Y_i are independent. Finally (f) holds since conditioning on $T=\tau$ we have the Markov chain

$$Y^{i-1} - X^{i-1}(W) - (Y_i,W,J_W,S),\ \forall i \in [1:M]. \tag{5.101}$$

Let Q be a uniform random variable on $[1:n]$ which is independent of everything. Define further $U = (U_Q, Q)$. Then we obtain, $\forall (s,\tau) \in \mathcal{S} \times \mathcal{T}$

$$\begin{aligned} R_c + \delta &\geq I(X_Q(W); U|S=s) \\ (R_i - \delta)(1-\delta) &\leq I(Y_Q(W); U|S=s, T=\tau) + \delta. \end{aligned} \tag{5.102}$$

Note that

$$P_{Y_Q X_Q(W) U|S=s,T=\tau} = P_{Y|X,\tau} \times P_{X|S=s} \times P_{U|X_Q(W),S=s}. \tag{5.103}$$

Since each conditional distribution $P_{U|X_Q(W),S=s}$ acts independently, we can upper bound the cardinality[5.12] of $\mathcal{U}$ by $|\mathcal{X}| + |\mathcal{T}|$ by following [EK11, Appendix C] as each $P_{U|X_Q(W),S=s}$ affects $|\mathcal{T}|$ terms $H(Y_Q|U, S=s, T=\tau)$. This implies that $(R_c + \delta, (R_i - \delta)(1-\delta) - \delta) \in \bar{\mathcal{R}}_{sc}$. Hence taking $\delta \to 0$ we obtain the desired conclusion.

5.C Proof of Theorem 5.2

Achievability
As in the proof of Theorem 5.1 we generate a codebook for all users which consists of $|S|e^{nR_c}$ codewords. The definitions of the sets $\mathcal{A}_s$, $\mathcal{B}_s$ and $\hat{\mathcal{A}}_s$ are similar as in Section 5.1, i.e.,

$$\mathcal{B}_s = \{(x^n, u^n) \mid \mathbb{E}[\phi_s(x^n, \hat{Y}^n, u^n)|\hat{X}^n = x^n, S=s] \leq \delta_{n,s}^{1/2}\}.$$

where $\phi_s(x^n, y^n, u^n) = \chi_{\{(y^n,u^n) \notin \mathcal{A}_\delta^n(P_{YU|S=s})\}}$ and $\hat{\mathcal{A}}_s = \mathcal{A}_s \cap \mathcal{B}_s$.
Enrollment: For each user $i \in [1:M]$ we first assign $x^n(i)$ to one of the label using the classifier T_X if it is possible. The resulted label is denoted by $\hat{s}_i = T_X(x^n(i))$. Assuming that there is no error, then we proceed as before to look for an index $m_{\hat{s}_i,i}$ such that

$$(x^n(i), u^n(m_{\hat{s}_i,i})) \in \hat{\mathcal{A}}_{\hat{s}_i}. \tag{5.104}$$

Then $m_{\hat{s}_i,i}$ is stored in the database in the corresponding position. Note that in this case storing $\hat{s}_i$ does not help.
Identification: Given the observation sequence y^n, the processing unit first searches for a suitable label by using the classifier T_Y if it is possible. We denote this label by s', i.e., $s' = T_Y(y^n)$. If there is no error, then the processing unit looks for a unique index $\hat{w}$ such that

$$(y^n, u^n(m_{\hat{s}_{\hat{w}},\hat{w}})) \in \mathcal{A}_\delta^n(P_{YU|S=s'}). \tag{5.105}$$

[5.12] The bounding procedure may result in different random variables U_s and test channels $P_{U|X_Q(W),S=s}$ for different s.

If there is more than one of such $\hat{w}$ or there is none the system declares an error.
Analysis: For a given sequence of states $(s_i)_{i=1}^M$, we define $\mathcal{E}_{no}(w)$, $\mathcal{E}_{\geq 2}(w)$, $\mathcal{E}_X(w)$, $\mathcal{E}_Y(w)$, $\mathcal{E}_{xu}(w)$, $\mathcal{E}_1(w)$ and $\mathcal{E}_2(w)$ similarly as

$$\begin{aligned}
\mathcal{E}_{es}(w) &= \{T_X(X^n(w)) \neq s_w\} \cup \{T_Y(Y^n) \neq s_w\},\\
\mathcal{E}_{xu}(w) &= \{(X^n(w), U^n(m_{s_w,w})) \notin \hat{\mathcal{A}}_{s_w},\ \forall m_{s_w,w}\}\\
\mathcal{E}_1(w) &= \{(Y^n, U^n(M_{s_w,w})) \notin \mathcal{A}_\delta^n(P_{YU|S=s_w})\},\\
\mathcal{E}_2(w) &= \{\exists w' \neq w \mid (Y^n, U^n(M_{\hat{S}_{w'},w'})) \notin \mathcal{A}_\delta^n(P_{YU|S=s_w})\},\\
\mathcal{E}_{no}(w) &= \{(Y^n, U^n(M_{\hat{S}_w,w})) \notin \mathcal{A}_\delta^n(P_{YU|S=S'})\},\\
\mathcal{E}_{\geq 2}(w) &= \{\exists w' \neq w \mid (Y^n, U^n(M_{\hat{S}_{w'},w'})) \in \mathcal{A}_\delta^n(P_{YU|S=S'})\}.
\end{aligned} \tag{5.106}$$

We need to be careful in this scenario since the symmetry w.r.t. W when conditioning on a state sequence does not hold. Assume that $(S_i)_{i=1}^M = (s_i)_{i=1}^M$ and $W = w$, if $\hat{W} \neq w$ then the following event occurs

$$\mathcal{E}_{es}(w) \cup \mathcal{E}_{no}(w) \cup \mathcal{E}_{\geq 2}(w).$$

First, we observe that

$$\Pr\{\mathcal{E}_{es}(w) | (S_i)_{i=1}^M = (s_i)_{i=1}^M, W = w\} = \Pr\{\mathcal{E}_{es}(w) | S_w = s_w, W = w\}. \tag{5.107}$$

From Lemma 5.1, there exists an $n_1(\epsilon, s_w)$ such that if $n \geq n_1(\epsilon, s_w)$ then

$$\Pr\{T_X(X^n(w)) \neq s_w | S_w = s_w, W = w\} \leq \epsilon. \tag{5.108}$$

Since s_w can be any element of the finite set $\mathcal{S}$, there are only $|S|$ possible values for the left-hand side as w varies, due to the assumption that $X^n(i) \sim P_{X|S=s_i}^{\otimes n}$. Therefore, if we take $n_1(\epsilon) = \max_{s \in \mathcal{S}} n_1(\epsilon, s)$ then for $n \geq n_1(\epsilon)$

$$\Pr\{T_X(X^n(w)) \neq s_w | S_w = s_w, W = w\} \leq \epsilon,\ \forall s_w \in \mathcal{S},\ \forall w \in [1:M]. \tag{5.109}$$

Similarly the exists an $n_2(\epsilon)$ such that if $n \geq n_2(\epsilon)$ holds then

$$\Pr\{T_Y(Y^n) \neq s_w | S_w = s_w, W = w\} \leq \epsilon,\ \forall s_w \in \mathcal{S},\ \forall w \in [1:M]. \tag{5.110}$$

Next, we bound the conditional probability of the event $\mathcal{E}_{es}^c(w) \cap \mathcal{E}_{no}(w)$ as in (5.111).

$$\begin{aligned}
&\Pr\{\mathcal{E}_{es}^c(w) \cap \mathcal{E}_{no}(w) | (S_i)_{i=1}^M = (s_i)_{i=1}^M, W = w\}\\
&\leq \Pr\{\mathcal{E}_{xu}(w) | (S_i)_{i=1}^M = (s_i)_{i=1}^M, W = w\}\\
&+ \Pr\{\{(X^n(w), U^n(M_{s_w,w})) \in \hat{\mathcal{A}}_{s_w}\} \cap \mathcal{E}_1(w) | (S_i)_{i=1}^M = (s_i)_{i=1}^M, W = w\}\\
&= \Pr\{\mathcal{E}_{xu}(w) | S_w = s_w, W = w\} + \bar{t}_2(w, (s_i)_{i=1}^M).
\end{aligned} \tag{5.111}$$

The first term can be bounded as, given $W = w$ and $S_w = s_w$

$$\begin{aligned}
&\Pr\{\mathcal{E}_{xu}(w)|S_w = s_w, W = w\} \\
&\le \Pr\big\{(\hat{X}^n, \hat{U}^n) \notin \hat{\mathcal{A}}_{s_w}|S = s_w\big\} \\
&+ \Pr\big[\iota(\hat{X}^n; \hat{U}^n|S = s_w) \ge nR_c - \gamma\big|S = s_w\big] + e^{-\exp(\gamma)} \le \epsilon,
\end{aligned} \tag{5.112}$$

for $n \ge n_3(\epsilon, \delta, s_w)$ if we take $\gamma = n\delta/2$ and

$$R_c \ge I(X; U|S = s_w) + \delta. \tag{5.113}$$

Consequently, if we take $n \ge n_3(\epsilon, \delta) = \max_{s\in\mathcal{S}} n_3(\epsilon, \delta, s)$ and

$$R_c > \max_{s\in\mathcal{S}} I(X; U|S = s) + \delta$$

then

$$\Pr\{\mathcal{E}_{xu}(w)|S_w = s_w, W = w\} \le \epsilon,\ \forall s_w \in \mathcal{S},\ \forall w \in [1 : M]. \tag{5.114}$$

Lastly, we have

$$\begin{aligned}
\bar{t}_2(w, (s_i)_{i=1}^M) &= \int_{\hat{\mathcal{A}}_{s_w}} \int_{\mathcal{A}_\delta^n(P_{Y|U,S=s_w}|u^n)^c} dP_{Y|X}^{\otimes n}(y^n|x^n) \\
&\qquad \times dP_{X^n(1)U^n(M_{s_w,w})|SW}(x^n, u^n|s_w, w) \\
&\le \delta_{n,s_w}^{1/2} \le \epsilon,
\end{aligned} \tag{5.115}$$

if $n \ge n_4(\epsilon, s_w)$. Hence taking $n \ge n_4(\epsilon) = \max_{s\in\mathcal{S}} n_4(\epsilon, s)$ we obtain

$$\bar{t}_2(w, (s_i)_{i=1}^M) \le \epsilon,\ \forall s_w \in \mathcal{S},\ \forall w \in [1 : M]. \tag{5.116}$$

Next, we have

$$\begin{aligned}
&\Pr\{\mathcal{E}_{es}^c(w) \cap \mathcal{E}_{\ge 2}(w)|(S_i)_{i=1}^M = (s_i)_{i=1}^M, W = w\} \\
&\le \Pr\{\mathcal{E}_2(w)|W = w, (S_i)_{i=1}^M = (s_i)_{i=1}^M\}.
\end{aligned} \tag{5.117}$$

The right-hand side of (5.117) can be bounded further as follows

$$\begin{aligned}
&\Pr\{\mathcal{E}_2(w)|W = w, (S_i)_{i=1}^M = (s_i)_{i=1}^M\} \\
&\quad \le \sum_{w'\ne w} \Pr\big\{(Y^n, U^n(M_{\hat{S}_{w'},w'})) \in \mathcal{A}_\delta^n(P_{YU|S=s_w})\big|S_{w'} = s_{w'}, S_w = s_w, W = w\big\}. \\
&\quad = \sum_{w'\ne w} \int \int_{\mathcal{A}_\delta^n(P_{Y|U,S=s_w}|u^n)} dP_{Y|S=s_w}^{\otimes n}(y^n) \times dP_{U^n(M_{\hat{S}_{w'},w'})|S_{w'}}(u^n|s_{w'}) \\
&\quad \le \sum_{w'\ne w} e^{-n(I(Y;U|S=s_w)-3\delta)} \le e^{nR_i} e^{-n(I(Y;U|S=s_w)-3\delta)} < \epsilon,
\end{aligned} \tag{5.118}$$

if $R_i < I(Y;U|S=s_w) - 3\delta$ for $n \geq n_5(\epsilon, \delta, s_w)$. The second inequality holds due to the property of weak typicality, which is independent of n. Hence if we take $R_i < \min_{s\in\mathcal{S}} I(Y;U|S=s) - 3\delta$ and $n \geq n_5(\epsilon,\delta) = \max_{s\in\mathcal{S}} n_5(\epsilon,\delta,s)$ then

$$\Pr\{\mathcal{E}_2(w)|W=w,(S_i)_{i=1}^M=(s_i)_{i=1}^M\} < \epsilon,\ \forall s_w \in \mathcal{S},\ w\in[1:M]. \tag{5.119}$$

In summary, by taking $n > \max\{n_i(\epsilon,\delta)\}_{i=1}^5$ and

$$\begin{aligned} R_c &> \max_{s\in\mathcal{S}} I(X;U|S=s) + \delta, \\ R_i &< \min_{s\in\mathcal{S}} I(Y;U|S=s) - 3\delta, \end{aligned} \tag{5.120}$$

then

$$\Pr\{W \neq \hat{W}|(S_i)_{i=1}^M=(s_i)_{i=1}^M\} < 6\epsilon, \tag{5.121}$$

for every tuple $(s_i)_{i=1}^M$. This implies the achievable conclusion of Theorem 5.2.

Converse for the discrete case

We show in the following that: Any achievable compression-identification rate pair for the setting in Theorem 5.2 is also achievable for the one in Theorem 5.1, i.e.,

$$\mathcal{R}_{iis} \subseteq \mathcal{R}_{sc}. \tag{5.122}$$

The converse follows since

$$\bar{\mathcal{R}}_{sc} \subseteq \mathcal{R}_{iis} \subseteq \mathcal{R}_{sc} = \bar{\mathcal{R}}_{sc}. \tag{5.123}$$

Let (ϕ_n,ψ_n) be an identification scheme for the setting of Theorem 5.2. Consider the case that $s_i = s$ for all $i \in [1:M]$. Then expanding the error expression conditioning on $(S_i)_{i=1}^M = s^M$ we obtain

$$\begin{aligned} &\Pr\{W \neq \hat{W}|(S_i)_{i=1}^M = s^M\} \\ &= \frac{1}{M}\sum_w \int_{\mathcal{D}_w} P_{Y|X}^{\otimes n}(y^n|x^n(w)) \times \prod_i P_{\phi_n}(j_i|x^n(i)) d\Big(\prod_i P_{X|S}^{\otimes n}(x^n(i)|s)\Big) \\ &= \Pr\{W \neq \hat{W}|S=s\}, \end{aligned} \tag{5.124}$$

where the last term in (5.124) is the error expression of the setting in Theorem 5.1 and

$$\mathcal{D}_w = \{(y^n,(j_i)_{i=1}^M) \mid \psi_n(y^n,(j_i)_{i=1}^M) \neq w\}. \tag{5.125}$$

Hence

$$\begin{aligned} &\sup_{s\in\mathcal{S}} \Pr\{W\neq\hat{W}|S=s\} \overset{(5.124)}{=} \sup_{s\in\mathcal{S}} \Pr\{W\neq\hat{W}|(S_i)_{i=1}^M=s^M\} \\ &\leq \sup_{(s_i)_{i=1}^M\in\mathcal{S}^M} \Pr\{W\neq\hat{W}|(S_i)_{i=1}^M=(s_i)_{i=1}^M\} < \delta, \end{aligned} \tag{5.126}$$

for all $n \geq n_0(\delta)$ if (ϕ_n,ψ_n) is an achievable identification scheme for the current setting. This implies that $\mathcal{R}_{iis} \subseteq \mathcal{R}_{sc}$.

5.D Proof of Lemma 5.2

Proof. Define the set

$$\mathcal{A} = \left\{ (y^n, j_w, w) \middle| \frac{1}{n} \log \frac{dP_{Y^n|J_W,W}(y^n|j_w, w)}{dP_{Y^n}(y^n)} > R_i - 2\gamma \right\}. \tag{5.127}$$

Due to the symmetry as all users' data follows the same underlying distribution P_{X^n} the Radon-Nikodym derivative in the definition of $\mathcal{A}$ does not depend on w and equals to $\frac{dP_{\bar{Y}^n|\phi_n(\bar{X}^n)}}{dP_{\bar{Y}^n}}$. Hence, the first term in the statement of Lemma 5.2 is given by $\Pr\{(Y^n, J_W, W) \in \mathcal{A}^c\}$. Additionally, as $\hat{W} = \psi_n(Y^n, (J_i)_{i=1}^M)$ we have

$$\Pr\{\hat{W} = W\} \leq \Pr\{\hat{W} = W, (Y^n, J_W, W) \in \mathcal{A}^c\} + \Pr\{(Y^n, J_W, W) \in \mathcal{A}\}. \tag{5.128}$$

Next,

$$\begin{aligned}
&\Pr\{\hat{W} = W, (Y^n, J_W, W) \in \mathcal{A}^c\} \\
&= \sum_{w=1}^{M} \sum_{\substack{(y^n, (j_i)_{i=1}^M):\\ \psi_n(y^n, (j_i)_{i=1}^M) = w \\ (y^n, j_w, w) \in \mathcal{A}^c}} dP_{Y^n|J_W W}(y^n|j_w, w) dP((j_i)_{i=1}^M) \frac{1}{M} \\
&\overset{(\star)}{\leq} \sum_{w=1}^{M} \frac{e^{n(R_i - 2\gamma)}}{M} \sum_{\substack{(y^n, (j_i)_{i=1}^M):\\ \psi_n(y^n, (j_i)_{i=1}^M) = w}} dP_{Y^n}(y^n) dP((j_i)_{i=1}^M) \\
&\overset{(\star\star)}{\leq} e^{-n\gamma} \sum_{(j_i)_{i=1}^M} P((j_i)_{i=1}^M) P_{Y^n}\left[\bigcup_w \{y^n : \psi_n(y^n, (j_i)_{i=1}^M) = w\} \right] \\
&\leq e^{-n\gamma},
\end{aligned} \tag{5.129}$$

where $(\star)$ is valid due to the definition of $\mathcal{A}$ and $(\star\star)$ holds for all $n \geq n_0(\gamma)$, cf. (5.32). The conclusion of the lemma follows. □

5.E A Strong Converse Proof

We present herein a strong converse proof for the discrete setting in the case that both the source distribution P_X and the channel $P_{Y|X}$ are known. The arguments are transferred immediately to the Gaussian case. In the discrete setting our proof is weaker than in [ZTM17] where the authors show the exponential strong converse. We begin with some definitions and important tools. The definition of ϵ-achievability is already given in Section 5.2.C, we restate it here for convenience. Given an $\epsilon \in [0, 1)$.

Definition 5.7 A rate pair (R_c, R_i) is ϵ-achievable if there exists a pair of identification-compression mappings (ϕ_n, ψ_n) such that

$$\limsup_{n\to\infty} \frac{1}{n}\log|\mathcal{M}_1| \le R_c,\ \liminf_{n\to\infty}\frac{1}{n}\log M \ge R_i,$$
$$\text{and } \limsup_{n\to\infty}\Pr\{\hat{W}\neq W\}\le\epsilon. \tag{5.130}$$

Let $\mathcal{R}_{\mathrm{ID},\epsilon}$ be the closure of all ϵ-achievable rate pairs (R_c, R_i).

Our proof follows essentially the same arguments as in [Liu18, Theorem 4.4.1] and [AGK76, Theorem 3], where the former uses the following theorem:

Theorem 5.8 *[LHV17, Theorem 9], [Liu18, Theorem 4.3.1] Consider P_X a probability measure on a finite set $\mathcal{X}$, ν a probability measure on $\mathcal{Y}$ and $P_{Y|X}$. Let $\beta_X = 1/\min_x P_X(x) \in [1,\infty)$, $\alpha = \sup_x \|\frac{dP_{Y|X=x}}{d\nu}\|_\infty \in [1,\infty)$. Let $c\in(0,\infty)$, $\eta,\delta\in(0,1)$, and $n > 3\beta_X\log\frac{|\mathcal{X}|}{\delta}$. We can choose some set $\mathcal{C}_n$ with $P_X^{\otimes n}[\mathcal{C}_n]\ge 1-\delta$, such that for $\mu_n = P_X^{\otimes n}\big|_{\mathcal{C}_n}$ we have*

$$\log\mu_n[x^n : P_{Y^n|X^n=x^n}(f)\ge\eta] - c\log\nu^{\otimes n}(f)$$
$$\le nd^\star(P_X, P_{Y|X}, \nu, c) + A\sqrt{n} + c\log\frac{1}{\eta} \tag{5.131}$$

for any[5.13] $f\in\mathcal{H}_{[0,1]}(\mathcal{Y}^n)$ *where*

$$A = \log(\alpha^c\beta_X^{c+1})\sqrt{3\beta_X\log\frac{|\mathcal{X}|}{\delta}} + 2c\sqrt{(\alpha-1)\log\frac{1}{\eta}}. \tag{5.132}$$

$d^\star$ is defined as[5.14]

$$d^\star(P_X, P_{Y|X}, \nu, c)$$
$$= \sup_{Q_{UX}: Q_X = P_X}\{cD(Q_{Y|U}||\nu|Q_U) - D(Q_{X|U}||P_X|Q_U)\}.$$

We first examine the discrete case. Let (R_c, R_i) be an ϵ-achievable pair. Then, there exists a pair of mappings (ϕ_n, ψ_n) such that (5.130) are satisfied. Let $\gamma > 0$ be such that $\epsilon + 3\gamma < 1$. From Lemma 5.2 we know that for all $n \ge n_0(\gamma)$ we have

$$\epsilon + 2\gamma \ge \Pr\left\{\frac{1}{n}\log\frac{dP_{Y^n\phi_n(X^n)}}{d(P_{Y^n}\times P_{\phi_n(X^n)})}(Y^n,\phi_n(X^n)) \le R_i - 2\gamma\right\}$$
$$= \Pr\{(Y^n,\phi_n(X^n))\in\mathcal{A}_{1,n}^c\}, \tag{5.133}$$

[5.13] $\mathcal{H}_{[0,1]}(\mathcal{Y}^n)$ is the set of measurable (hence integrable) mappings $f:\mathcal{Y}^n\to[0,1]$.
[5.14] The convention is that $\infty-\infty=-\infty$.

where for notational clarity $\mathcal{A}_{1,n}$ corresponds to the set $\mathcal{A}$ in (6.98). Moreover, we denote by $\mathcal{A}^n_{2,\gamma}(Y)$ the weakly typical set w.r.t P_Y and threshold γ. Then for $n \geq n_1(\gamma)$ we have

$$\Pr\{(Y^n, \phi_n(X^n)) \in \mathcal{A}_{1,n},\ Y^n \in \mathcal{A}^n_{2,\gamma}\} \geq 1 - \epsilon - 3\gamma. \tag{5.134}$$

For simplicity we define $\hat{\epsilon} = \epsilon + 3\gamma$. For each $m_1 \in \mathcal{M}_1$ we define the following set

$$\mathcal{B}_{m_1} = \{y^n \mid (y^n, m_1) \in \mathcal{A}_{1,n},\ y^n \in \mathcal{A}^n_{2,\gamma}\}. \tag{5.135}$$

The inequality (5.134) can be rewritten as

$$P_X^{\otimes n}[P_{Y^n|X^n}[\mathcal{B}_{\phi_n(X^n)}]] \geq 1 - \hat{\epsilon}. \tag{5.136}$$

Choose $\epsilon' \in (\hat{\epsilon}, 1)$ and $\delta = \frac{\epsilon' - \hat{\epsilon}}{2\epsilon'}$. Then by Markov's inequality we have

$$P_X^{\otimes n}\big\{x^n \mid P_{Y^n|X^n=x^n}[\mathcal{B}_{\phi_n(x^n)}] \geq 1 - \epsilon'\big\} \geq 1 - \frac{\hat{\epsilon}}{\epsilon'}. \tag{5.137}$$

Additionally, we obtain

$$\begin{aligned} 1 &\geq \Pr\{Y^n \in \mathcal{B}_{m_1} | \phi_n(X^n) = m_1\} = \int_{\mathcal{B}_{m_1}} dP_{Y^n|\phi_n(X^n)=m_1}(y^n) \\ &\geq e^{n(R_i - 2\gamma)} \int_{\mathcal{B}_{m_1}} dP_Y^{\otimes n}(y^n) \geq e^{n(R_i - 3\gamma - h(Y))} \mathrm{vol}(\mathcal{B}_{m_1}),\ \forall m_1 \in \mathcal{M}_1, \end{aligned} \tag{5.138}$$

where $\mathrm{vol}(\cdot)$ denotes the number of sequences in the discrete case[5.15], which implies that $\mathrm{vol}(\mathcal{B}_{m_1}) \leq 2^{n(h(Y)+3\gamma-R_i)}$, for all $m_1 \in \mathcal{M}_1$. We choose ν as the uniform distribution on $\mathcal{Y}$. Let $c \in (0, \infty)$, $\eta = 1 - \epsilon'$. By Theorem 5.8 we then can find a measure μ_n such that

$$\mu_n\big\{x^n \mid P_{Y^n|X^n=x^n}[\mathcal{B}_{\phi_n(x^n)}] \geq \eta\big\} \geq 1 - \frac{\hat{\epsilon}}{\epsilon'} - \delta = \delta, \tag{5.139}$$

and

$$\begin{aligned} &\log \mu_n\{x^n \mid P_{Y^n|X^n=x^n}(f) \geq \eta\} - c \log \nu^{\otimes n}(f) \\ &\quad \leq n d^\star(P_X, P_{Y|X}, \nu, c) + \mathcal{O}\big(\sqrt{n}\big), \end{aligned} \tag{5.140}$$

for any integrable f with range $[0, 1]$. Further calculation indicates that

$$d^\star(P_X, P_{Y|X}, \nu, c) = \sup_{\substack{U: U - X - Y \\ I(X;U) < \infty}} \{-ch(Y|U) - I(X;U)\} + c \log |\mathcal{Y}|$$

[5.15] With abuse of notation, we use $h(Y)$ to denote both the discrete entropy and the differential entropy in this section.

$$= \max_{\substack{U:U-X-Y \\ |\mathcal{U}|\leq|\mathcal{X}|+1}} \{-ch(Y|U) - I(X;U)\} + c\log|\mathcal{Y}| \tag{5.141}$$

due to the support lemma [CK11, Lemma 15.4]. Since there are $|\mathcal{M}_1|$ possible values of m_1 it follows that there must exists $m^\star$ such that

$$\mu_n\{x^n \mid P_{Y^n|X^n=x^n}[\mathcal{B}_{m^\star}] \geq \eta\} \geq \frac{\delta}{|\mathcal{M}_1|}. \tag{5.142}$$

Moreover we also have

$$\nu^{\otimes n}(\mathcal{B}_{m^\star}) \leq |\mathcal{Y}|^{-n} e^{n(h(Y)+3\gamma-R_i)}. \tag{5.143}$$

Combining (5.141), (5.142) and (5.143) with $f = \chi_{\mathcal{B}_{m^\star}}$ we obtain that

$$\begin{aligned} &-\log|\mathcal{M}_1| - cn(-\log|\mathcal{Y}| + h(Y) + 3\gamma - R_i) \\ &\leq n\left(\max_{\substack{U:U-X-Y \\ |\mathcal{U}|\leq|\mathcal{X}|+1}} \{-ch(Y|U) - I(X;U)\} + c\log|\mathcal{Y}|\right) + \mathcal{O}(\sqrt{n}). \end{aligned} \tag{5.144}$$

which implies that

$$\begin{aligned} R_c - cR_i + (1+3c)\gamma &\geq \min_{\substack{U:U-X-Y \\ |\mathcal{U}|\leq|\mathcal{X}|+1}} \{I(X;U) - cI(Y;U)\} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \\ &= f(c) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{5.145}$$

Taking $n \to \infty$ we see that

$$(R_c + \gamma, R_i - 3\gamma) \in \bigcap_{c>0} \{R_a - cR_b \geq f(c)\} \overset{(*)}{=} \mathcal{R}_{\mathrm{ID}}, \tag{5.146}$$

where $(*)$ follows from the hyper plane characterization of a closed convex set, which leads to $\mathcal{R}_{\mathrm{ID},\epsilon} \subseteq \mathcal{R}_{\mathrm{ID}}$.

In the Gaussian case, assume that $P_X = \mathcal{N}(0, \sigma_X^2)$ and $P_{Y|X=x} = \mathcal{N}(x, 1)$, then by doing the same steps, with ν in this case the Lebesgue measure, we obtain (5.140) using [Liu18, Theorem 4.3.2]. Since ν is the Lebesgue measure, we have $\nu(\mathcal{B}_{m^\star}) = \mathrm{vol}(B_{m^\star})$. Additionally,

$$\begin{aligned} d^\star(P_X, P_{Y|X}, \nu, c) &= \sup_{\substack{U:U-X-Y \\ I(X;U)<\infty}} \{-ch(Y|U) - I(X;U)\} \\ &\overset{(**)}{\leq} \sup_{1<\beta\leq 1+\sigma_X^2} \left\{ -\frac{c}{2}\log 2\pi e\beta - \frac{1}{2}\log\frac{\sigma_X^2}{\beta-1} \right\}, \end{aligned} \tag{5.147}$$

where $(**)$ follows by first putting $h(Y|U) = \frac{1}{2}\log 2\pi e\beta$ and then using the entropy power inequality. Therefore we obtain,

$$\begin{aligned} & R_c - cR_i + (1+3c)\gamma \\ & \geq \inf_{0\leq\beta<\log(1+\sigma_X^2)} \frac{1}{2}\left\{ \log \frac{\sigma_X^2}{(\sigma_X^2+1)e^{-\beta}-1} - c\beta \right\} \\ & = \begin{cases} 0 & \text{if } 0 \leq c \leq 1 \text{ at } \beta = 0 \\ \frac{1}{2}\big\{ \log \sigma_X^2(c-1) & \text{if } c > 1 \\ \quad -c\log(\sigma_X^2+1)(1-1/c)\big\} & \text{at } \beta = \log(\sigma_X^2+1)(1-1/c) \end{cases} . \end{aligned} \tag{5.148}$$

Compared with the characterization in (5.15), we observe that

$$(R_c + \gamma, R_i - 3\gamma) \in \mathcal{R}_{\mathrm{ID}}. \tag{5.149}$$

Therefore, we have $\mathcal{R}_{\mathrm{ID},\epsilon} \subseteq \mathcal{R}_{\mathrm{ID}}$.

On $(*)$: The inclusion $\bigcap_{c>0}\{R_a - cR_b \geq f(c)\} \supseteq \mathcal{R}_{\mathrm{ID}}$ is quite straightforward. Since $\mathcal{R}_{\mathrm{ID}}$ is a closed convex subset of $\mathbb{R}_+^2$ if $(x,y) \in \mathbb{R}_+^2$ and $(x,y) \notin \mathcal{R}_{\mathrm{ID}}$ then there exists a vector $(a,b) \in \mathbb{R}^2$ such that

$$ax + by < aR_1 + bR_2, \ \forall (R_1, R_2) \in \mathcal{R}_{\mathrm{ID}}. \tag{5.150}$$

Since $(0,0) \in \mathcal{R}_{\mathrm{ID}}$, we see that either a or b must be negative. If $a < 0$, then plugging $(R_1, 0)$ in the above inequality we obtain the violation for sufficiently large R_1. Hence, we have $a > 0$ and $b < 0$. We can normalize further to obtain

$$x - cy < R_1 - cR_2, \ \forall (R_1, R_2) \in \mathcal{R}_{\mathrm{ID}}, \tag{5.151}$$

where $c = -b/a > 0$. The minimum of the right-hand side, which is attained by a point $(R_1^*, R_2^*) \in \mathcal{R}_{\mathrm{ID}}$, is $f(c)$. Therefore if $(x,y) \notin \mathcal{R}_{\mathrm{ID}}$ then $(x,y) \notin \bigcap_{c>0}\{R_a - cR_b \geq f(c)\}$, which implies that $\bigcap_{c>0}\{R_a - cR_b \geq f(c)\} \subseteq \mathcal{R}_{\mathrm{ID}}$.

5.F Proof of Proposition 5.4

Let S be any random variable on $\mathcal{S}$ and let $P_S \in \bar{\mathcal{P}}_c$ be the corresponding induced distribution. The corresponding multi-letter extension is denoted by S^n. The corresponding channel is given by

$$P_S^n(y^n|x^n) = \mathbb{E}P(y^n|x^n, S^n). \tag{5.152}$$

Assume that the rate pair (R_c, R_i) is achievable in the setting with an arbitrarily varying observation channel with the corresponding pair of mappings (ϕ_n, ψ_n). If we use the same enrollment mapping ϕ_n and identification mapping ψ_n for the

discrete memoryless observation channel P_S with the source distribution P_X then we have

$$\Pr_{P_S}\{W \neq \hat{W}\} = \mathbb{E}[e(S^n, \phi_n, \psi_n)] \leq \sup_{s^n \in \mathcal{S}^n} e(s^n, \phi_n, \psi_n) < \delta, \tag{5.153}$$

for all $n \geq n_0(\delta)$. This implies that for any $p \in \bar{\mathcal{P}}_{\mathcal{S}}$

$$\mathcal{R}_{avc} \subseteq \bar{\mathcal{R}}_p, \tag{5.154}$$

which leads to the conclusion of Proposition 5.4.

6

Equivalence

MEMBERSHIP testing has not been actively considered in existing works on identification systems. It is often assumed that the observation sequence is related to the data inside the system. In this work we put our attention to this important problem. Assume a database that stores *compressed* versions of data sequences of M users $(x^n(m))_{m=1}^M$. An observation sequence y^n is provided to a processing center which has access to these compressed data sequences. The processing center performs a screening step and returns *Yes/No* when y^n is *related to one of the user/ independent of all users* in the system. We call the first case hypothesis H_0 and the second case hypothesis H_1.

Other hypothesis testing problems related to the identification problem include [Vol+10; Sch02; Mou10]. In [Vol+10] the hypothesis H_1 was tested against M other hypotheses in the binary setting where the focus was to minimize the overall identification error under a specific decision rule. In [Sch02] the author considered the M-ary hypothesis testing problem with fixed M and studied the large deviation regime. In [Mou10] the decision rule was based on a decoding metric using the hashed data and observation sequences at different lengths. The exponents of the probability of miss and the expected number of incorrect items on the list were provided for a fixed hashed function. Error exponent aspects of the probability of estimating the correct user in the identification systems have been studied in [YY16; DD11; Mer17].

In this chapter we first study the exponent of the probability of the second kind of error $E^\star(R, R_c)$ provided that the probability of the first kind is vanishing. Next, we show that the lower bound is tight in the strong converse sense if R is less than the threshold value $R_{\max}(R_c)$. In-between we show the equivalence between the single-user hypothesis testing problem studied by Ahlswede and Csiszár, the Wyner-Ahlswede-Körner problem and the identification problem.

6.1 A Lower Bound of $E^\star(R, R_c)$

We begin with some notational conventions. We use $I_{H_0}(\cdot;\cdot)$ to indicate that the mutual information is evaluated w.r.t. the distribution related to H_0. Additionally, we use the $\bar{(\cdot)}$ notation, e.g. $\bar{\alpha}_n$, $\bar{\mathcal{A}}_n$, to emphasize that the single-user scenario, i.e. $M = 1$, is considered.

We assume that the database consists of M users with $\lim_{n\to\infty} \frac{1}{n} \log M = R$. i.e., the number of users grows with the block length n at rate R. We consider mainly in this work the discrete scenario where alphabets $\mathcal{X}$ and $\mathcal{Y}$ are finite. For each i the corresponding data sequence $x^n(i)$ is generated iid from the distribution P_X. Under H_0 the joint distribution of the sequence y^n and sequences $(x^n(m))_{m=1}^M$ is given by

$$P_{H_0} = \sum_{i=1}^{M} \frac{1}{M} P_{Y^n X^n(i)} \times \prod_{k=1, k\neq i}^{M} P_{X^n(k)}, \tag{6.1}$$

i.e., the sequence y^n is related to one randomly chosen user in the system, where $P_{Y^n X^n(i)} = P_{Y|X}^{\otimes n} \times P_{X^n(i)}$. The joint distribution under H_1 is given by

$$P_{H_1} = P_Y^{\otimes n} \times \prod_{i=1}^{M} P_{X^n(i)}, \tag{6.2}$$

i.e., the sequence y^n is not related to the information in the database. Note that under both hypotheses the users' data sequences are mutually independent. Additionally, the joint distributions P_{H_0} and P_{H_1} are not product distributions of iid random variables since the number of components M grows with n. We can also view P_{H_0} as the result of mixing M general random processes uniformly where the distributions at instance n are given by $P_{Y^n X^n(i)} \times \prod_{k=1,k\neq i}^{M} P_{X^n(k)}$, $i \in [1:M]$.

Definition 6.1 A *testing scheme* consists of a compression mapping which enrolls the users' data sequences into the database according to

$$\phi_n \colon \mathcal{X}^n \to \mathcal{M}_1, \tag{6.3}$$

and a decision mapping which outputs whether H_0 or H_1 is deemed true

$$\psi_n \colon \mathcal{Y}^n \times \mathcal{M}_1^M \to \{0, 1\}. \tag{6.4}$$

The acceptance region is defined accordingly as

$$\mathcal{A}_n = \big\{(y^n, (x^n(i))_{i=1}^M) \mid \psi_n(y^n, (j_i)_{i=1}^M) = 0\big\}, \tag{6.5}$$

where $j_i = \phi_n(x^n(i)) \in \mathcal{M}_1$, for all $i \in [1:M]$. An error of the first (second) type occurs when y^n is related to (independent from) one unknown user (all users) in the system but the testing scheme declares otherwise. The probability of first and second type of error, also called false alarm and miss detection probabilities, are given respectively as

$$\alpha_n = P_{H_0}(\mathcal{A}_n^c),\ \beta_n = P_{H_1}(\mathcal{A}_n). \tag{6.6}$$

Definition 6.2 An error exponent E of type II is *achievable* given (R, R_c) if there exist enrollment and decision mappings (ϕ_n, ψ_n) such that

$$\lim_{n\to\infty} \alpha_n = 0,\ \limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_1| \leq R_c,$$
$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_n} \geq E. \tag{6.7}$$

We define

$$E^\star(R, R_c) = \sup\{E \mid E \text{ is achievable given } (R, R_c)\}. \tag{6.8}$$

$E^\star(R, R_c)$ serves as an upper bound on the performance of testing schemes at the given compression rate R_c, i.e., the best scheme with (R, R_c) will not achieve a better error exponent.
The single-user testing against independence setting studied by [AC86] corresponds to the case $M = 1$ in our scenario. Although both settings involve compression at rate R_c, the motivating application contexts are, however, different. In [AC86] R_c represents the communication rate from a remote sensor to aid the decision making at a processing center. Herein, R_c represents stored information of each user in the database.

Remark 6.1 A straightforward observation when the compression is at zero-rate, i.e., the compression alphabet is of sub-exponential size, is

$$E^\star(R, 0) = 0,\ \forall R. \tag{6.9}$$

To see this result we look at $I(Y^n; (J_i)_{i=1}^M)$ under H_0 where $J_i = \phi_n(X^n(i))$, $\forall i \in [1 : M]$. Let W be the hidden random variable that characterizes the uniformly chosen user in the hypothesis H_0 which is independent of users' sequences, i.e.,

$$P_{H_0}(y^n, (x^n(i))_{i=1}^M, W = w)$$
$$= \frac{1}{M} P_{YX}^{\otimes n}(y^n, x^n(w)) \prod_{i\neq w} P_X^{\otimes n}(x^n(i)). \tag{6.10}$$

From (6.10) it can be inferred that

$$I_{H_0}(Y^n; (J_i)_{i=1}^M) \leq I_{H_0}(Y^n; W, (J_i)_{i=1}^M) = I_{H_0}(Y^n; W, J_W)$$
$$= I_{H_0}(Y^n; J_W|W) \leq \log |\mathcal{M}_1|.$$

When the compression is done at zero-rate then we have that $nH(Y) \approx n(H(Y) - 1/n \log |\mathcal{M}_1|) \leq H(Y^n|(J_i)_{i=1}^M) \leq nH(Y)$ since $1/n \log |\mathcal{M}_1|$ tends to 0 as $n \to \infty$. This implies roughly that Y^n is asymptotically independent of $(J_i)_{i=1}^M$. As the (marginal) distributions of Y^n and $(J_i)_{i=1}^M$ from both P_{H_0} and P_{H_1} are identical, the likelihood that we can distinguish both hypotheses from each other is low asymptotically. Since the probability of the first kind of error $\alpha_n \to 0$, the probability of the second kind of error is bounded away from 0 due to the indistinguishability, cf. Theorem 6.6 for a rigorous discussion.

Given a compression rate R_c we define the following functions

$$R_{\max}(R_c) = \max_{\substack{U-X-Y,\ |\mathcal{U}|\leq|\mathcal{X}|+1,\\ I(X;U)\leq R_c}} I(Y;U)$$
$$\theta(R,R_c) = R_{\max}(R_c) - R,\ \text{on } 0 \leq R < R_{\max}(R_c). \tag{6.11}$$

Different interpretations of $R_{\max}(R_c)$ appear in previous works. In [AC86] $R_{\max}(R_c)$ is the maximum error exponent of type II for the single-user testing against independence problem. In [Tun09] $R_{\max}(R_c)$ characterizes the number of users that can be identified with vanishing probability of error given the compression rate R_c. As mentioned in the problem description, the distribution of each user's data sequence is $P_X^{\otimes n}$ in both hypotheses. We also observe further symmetric properties in both distributions. Namely, in both distributions all users' data have the same marginal distribution. Furthermore, the channels are also identical across all users: $P_{Y|X}$ in H_0 and P_Y in H_1. The symmetry suggests us to build a testing scheme for our multi-user setting from a *generic* single-user testing scheme in which the same compression mapping is used to enroll each user's sequence into the database. We then have the following lower bound on $E^\star(R,R_c)$.

Theorem 6.1 *Given a single-user hypothesis testing scheme (ϕ_n,ψ_n) with probabilities of errors $\bar{\alpha}_n$ and $\bar{\beta}_n$, there exists a testing scheme (ϕ_n,ψ_n'), using the same compression mapping ϕ_n, for the multi-user case that achieves the following multi-user probabilities of error*

$$\alpha_n \leq \bar{\alpha}_n,\ \beta_n \leq \bar{\beta}_n M. \tag{6.12}$$

Consequently, given that the condition $0 \leq R < R_{\max}(R_c)$ is satisfied, then the multi-user error exponent is lowered bounded by

$$E^\star(R,R_c) \geq \theta(R,R_c). \tag{6.13}$$

Given a compression rate R_c, if the achieved error exponent by the single-user scheme, E, is less than R, then the bound for the probability of error of type II in our multi-user setting in Theorem 6.1 is loose. In other words, we need to design the decision mapping ψ_n for the single-user hypothesis testing scheme such that E is larger than R if it is *possible*[6.1] so that the corresponding multi-user error exponent would be positive.
It is difficult to apply the same trick using Stein's Lemma as in the proof of [AC86, Theorem 1] mainly as the number of components M in both distributions grows with n. For notational brevity, in the sequel we abbreviate $R_{\max}(R_c)$ as $R_{\max}$, $(X^n(i))_{i=1}^M$ as $\mathbf{X}^n$ and $(\phi_n(X^n(i)))_{i=1}^M$ as $\phi_n(\mathbf{X}^n)$.

[6.1] Even in the case that $R_{\max}(R_c) > R$ holds we can not use any single-user testing scheme with an error exponent $E \leq R$ to produce a positive error exponent in the multi-user case.

Proof. Let $\bar{\mathcal{A}}_n$ be the acceptance region for the single-user scenario corresponding to (ϕ_n, ψ_n). For each user $i \in [1 : M]$ the compression mapping ϕ_n maps the data sequence $x^n(i)$ into an index that is stored in the database. We define the acceptance region for the multi-user case as follows

$$\mathcal{A}_n = \{(y^n, \mathbf{x}^n) \mid (y^n, x^n(i)) \in \bar{\mathcal{A}}_n \text{ for some } i \in [1 : M]\}. \tag{6.14}$$

The probability of the first type of error in the multi-user scenario is upper bounded as

$$\begin{aligned}\alpha_n = P_{Y^n\mathbf{X}^n}\{\mathcal{A}_n^c\} &= \sum_w P_{Y^n\mathbf{X}^nW}\{\mathcal{A}_n^c, w\} \\ &\overset{(\star)}{\leq} \sum_w P_{Y^nX^n(w)W}\{\bar{\mathcal{A}}_n^c, w\} \overset{(6.10)}{=} \bar{\alpha}_n.\end{aligned}$$

The inequality $(\star)$ holds since $\{(y^n, \mathbf{x}^n) | (y^n, x^n(w)) \in \bar{\mathcal{A}}_n\} \subset \mathcal{A}_n$. Moreover, the probability of the second type of error is upper bounded as

$$\begin{aligned}\beta_n &= P_{Y^n} \times P_{\mathbf{X}^n}\{\mathcal{A}_n\} \\ &\leq \sum_i P_{Y^n} \times P_{X^n(i)}\{\bar{\mathcal{A}}_n\} = M\bar{\beta}_n,\end{aligned} \tag{6.15}$$

where the inequality is valid since $\mathcal{A}_n \subset \bigcup_{i=1}^M \{(y^n, \mathbf{x}^n) | (y^n, x^n(i)) \in \bar{\mathcal{A}}_n\}$ holds. The existence of testing schemes that achieve the exponent $R_{\max}(R_c) - \gamma$ for the single-user scenario where $\gamma > 0$ is well-known, cf. [AC86; Han87]. Hence, this proves the theorem. □

In the following, we investigate some cases where the lower bound is actually tight.

Proposition 6.1 *Assume that the compression is lossless, i.e., $R_c \geq H(X)$, then the lower bound $\theta(R, R_c)$ is tight, i.e. the maximum exponent for the second type of error is given by*

$$E_{ll}^\star = I(Y; X) - R, \text{ for } R < I(Y; X). \tag{6.16}$$

Proof. Denote the losslessly compressed sequences by $(\hat{X}_i^n)_{i=1}^M$ abbreviated in the following as $\hat{\mathbf{X}}^n$. For a given n we have

$$\begin{aligned}D_n = D(P_{Y^n\mathbf{X}^n} \| P_{Y^n} \times P_{\mathbf{X}^n}) &\overset{(a)}{\geq} D(P_{Y^n\hat{\mathbf{X}}^n} \| P_{Y^n} \times P_{\hat{\mathbf{X}}^n}) \\ &\overset{(b)}{\geq} P_{H_0}(\mathcal{A}_n) \log \frac{P_{H_0}(\mathcal{A}_n)}{P_{H_1}(\mathcal{A}_n)} + P_{H_0}(\mathcal{A}_n^c) \log \frac{P_{H_0}(\mathcal{A}_n^c)}{P_{H_1}(\mathcal{A}_n^c)} \\ &\geq -(1 - \alpha_n) \log \beta_n - h_2(\alpha_n)\end{aligned} \tag{6.17}$$

where $h_2(\cdot)$ is the binary entropy function. (a) follows due to the data-processing inequality with the channel $P_{\hat{\mathbf{X}}^n|\mathbf{X}^n} \times \mathrm{id}_{Y^n}$. (b) holds due to the log-sum inequality [CT12, Theorem 2.7.1]. Note that we actually have

$$\begin{aligned} D_n = I_{H_0}(Y^n; \mathbf{X}^n) &= nH(Y) - H(Y^n|\mathbf{X}^n) \\ &\overset{(c)}{=} nH(Y) - H(Y^n|W, \mathbf{X}^n) - H(W) + H(W|Y^n, \mathbf{X}^n) \\ &\overset{(6.10)}{=} nI(Y;X) - \log M + H(W|Y^n, \mathbf{X}^n). \end{aligned} \tag{6.18}$$

The equality in (c) holds since W is independent of the users' sequences $\mathbf{X}^n$. Define the deterministic mapping $f_n \colon \mathcal{Y}^n \times \mathcal{X}^{nM} \to \mathcal{W} \cup \{e\}$ as follows. If there exists a unique index $\hat{w}$ such that $(y^n, x^n(\hat{w})) \in \mathcal{T}_\epsilon^n(P_{XY})$ then we set $f_n(y^n, (x(i))_{i=1}^M)$ to $\hat{w}$. If there is more than one such index or there is none then the corresponding value of f_n is e. Then as $R < I(Y;X)$ we have [WKL03]

$$\Pr_{H_0}\{W \neq \hat{W}\} \to 0 \text{ as } n \to \infty. \tag{6.19}$$

Therefore, by Fano's inequality we obtain

$$H(W|Y^n, \mathbf{X}^n) \leq n\epsilon_n, \tag{6.20}$$

where $\epsilon_n \to 0$ as $n \to \infty$. Hence combining both expressions (6.17) and (6.18), the following bound holds for all n

$$-(1-\alpha_n)\log\beta_n - h_2(\alpha_n) \leq nI(X;Y) - \log M + n\epsilon_n. \tag{6.21}$$

Hence, with $\alpha_n \to 0$ as $n \to \infty$ we obtain

$$E \leq \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_n} \leq I(X;Y) - R. \tag{6.22}$$

The proof of Proposition 6.1 is complete. □

Remark 6.2 Note that in the case of no compression our setting is an instance of hypothesis testing for the mixed source problem. When M is a constant, we obtain a similar result as in [Han03, Example 4.1.1]. Therefore, Proposition 1 states that allowing the mixing coefficients, herein $1/M$ in P_{H_0}, to depend on n can lead to a non-trivial reduction in the error exponent of type II.

Another scenario where the lower bound is tight is given in the following proposition, which is a slight generalization of the Ahlswede-Csiszár result on testing against independence [AC86, Theorem 2].

Proposition 6.2 *Assume that $R = 0$, i.e., the number of users in the database grows sub-exponentially, then we obtain*

$$E^\star(0, R_c) = \theta(0, R_c) = R_{\max}(R_c). \tag{6.23}$$

Proof. Similarly as in the proof of Proposition 6.1, we use the log-sum inequality to obtain the following relation

$$D(P_{Y^n\phi_n(\mathbf{X}^n)}\|P_{Y^n} \times P_{\phi_n(\mathbf{X}^n)}) \geq -(1-\alpha_n)\log\beta_n - h_2(\alpha_n).$$

We can upper bound the left-hand side using the monotonicity of divergence so that we obtain

$$\begin{aligned} D(P_{Y^n\phi_n(\mathbf{X}^n)W}\|P_{Y^n} \times P_{\phi_n(\mathbf{X}^n)} \times P_W) \\ \geq -(1-\alpha_n)\log\beta_n - h_2(\alpha_n). \end{aligned}$$

The left-hand side is equal to

$$\begin{aligned} H(Y^n) - H(Y^n|\phi_n(\mathbf{X}^n), W) &= nH(Y) - H(Y^n|J_W, W) \\ &= n(H(Y) - H(Y_Q|U)) \end{aligned} \tag{6.24}$$

where we first define $U_i = (J_W, Y^{i-1})$ and then $U = (U_Q, Q)$ in which Q is a uniform random variable on $[1:n]$ that is independent of everything else. For any $\delta > 0$ the compression rate can be readily bounded (for all sufficiently large n) as

$$R_c + \delta > I(X_Q; U). \tag{6.25}$$

Using the standard cardinality bounding technique [EK11, Appendix C] we can find an "equivalent" $\bar{U}$, which takes values on $\bar{\mathcal{U}}$ such that $|\bar{\mathcal{U}}| \leq |\mathcal{X}| + 1$, and a probability kernel $P_{\bar{X}|\bar{U}}$, preserving $H(Y_Q|U) = H(\bar{Y}|\bar{U})$ and $H(X_Q|U) = H(\bar{X}|\bar{U})$ where $P_{\bar{X}\bar{Y}} = P_{XY}$. Hence

$$-(1-\alpha_n)\log\beta_n - h_2(\alpha_n) \leq nR_{\max}(R_c + \delta). \tag{6.26}$$

As $\alpha_n \to 0$ when $n \to \infty$, we get

$$E \leq R_{\max}(R_c + \delta),\ \forall\delta > 0. \tag{6.27}$$

Taking $\delta \to 0$, we obtain that $E \leq \theta(0, R_c)$ for $R = 0$ since $R_{\max}(R_c)$ is continuous in R_c. □

Remark 6.3 Note that the term $D(P_{Y^n\phi_n(\mathbf{X}^n)}||P_{Y^n} \times P_{\phi_n(\mathbf{X}^n)})$ in the proof of Proposition 6.2 can be expanded as

$$\begin{aligned} D(P_{Y^n\phi_n(\mathbf{X}^n)}||P_{Y^n} \times P_{\phi_n(\mathbf{X}^n)}) = nH(Y) - H(Y^n|\phi_n(\mathbf{X}^n), W) \\ - H(W) + H(W|Y^n, \phi_n(\mathbf{X}^n)). \end{aligned}$$

Using the same arguments from (6.24) onwards, we also obtain the conclusion of Proposition 6.2. Hence, a central theme in the proofs for the two aforementioned propositions is that

$$H(W|Y^n, \phi_n(\mathbf{X}^n)) \approx n\epsilon_n, \text{ where } \epsilon_n \to 0 \text{ as } n \to \infty. \tag{6.28}$$

In the first case, this holds due to the fact that we can construct a decoding function by the identification problem. In the second case, the approximation is valid since the alphabet $\mathcal{W}$ has sub-exponential size. If we restrict the compression mappings to the following set

$$\begin{aligned}\mathcal{B}(R,R_c)=\big\{(\phi_n)\mid{}&\text{there exists a sequence of } (\tilde{\psi}_n) \text{ such that}\\ &\limsup_{n\to\infty}\frac{1}{n}\log|\phi_n|\le R_c,\\ &\lim_{n\to\infty}\Pr{}_{H_0}\{\tilde{\psi}_n(Y^n,\phi_n(\mathbf{X}^n))\neq W\}=0\big\},\end{aligned}\tag{6.29}$$

then by following essentially the same steps as in the proofs of Proposition 6.1 and 6.2, we can show, cf. Appendix 6.A, that the restricted error exponent is characterized by $E_{\mathcal{B}}^{\star}(R,R_c)=\theta(R,R_c)$ when $R<R_{\max}(R_c)$. In other words, (6.29) is a sufficient condition for the tightness of the lower bound. It will be clear later that (6.29) is also the necessary condition, cf. Remark 6.8.

6.2 Characterization via Strong Converse

We first establish the strong relation between the single-user HT and the WAK problem. Then we close the lower bound on the optimal error exponent $E^{\star}(R,R_c)$ in the previous section via strong converse proofs. Finally, we show the equivalence of the single-user HT and the identification problem.

6.2.A Equivalence between single-user HT and WAK problems

Assume that in both problems the source is given by $\bar{X}^n\bar{Y}^n\sim P_{XY}^{\otimes n}$. We briefly recap the definition of a WAK-code. It consists of two encoding mappings $\phi_{1n}\colon\mathcal{X}^n\to\mathcal{M}_1$, $\phi_{2n}\colon\mathcal{Y}^n\to\mathcal{M}_2$ and a decoding function $\psi_n\colon\mathcal{M}_1\times\mathcal{M}_2\to\mathcal{Y}^n$. The WAK problem aims to control $\Pr\{\bar{Y}^n\neq\hat{\bar{Y}}^n\}$, where $\hat{\bar{Y}}^n=\psi_n(\phi_{1n}(\bar{X}^n),\phi_{2n}(\bar{Y}^n))$. We present in the following an equivalent relation between a WAK-code and a single-user HT scheme.

Theorem 6.2 *Fix an arbitrary $\gamma>0$. Given a WAK-code $(\phi_{1n},\phi_{2n},\psi_n)$, we can construct a single-user testing scheme (ϕ_{1n},ψ_n') such that the corresponding error probabilities of type I and II are given by*

$$\begin{aligned}\bar{\alpha}_n&\le P_{\mathrm{WAK}}\{\text{error}\}+\Pr\{\bar{Y}^n\notin\mathcal{A}_\gamma^n\}\\ \bar{\beta}_n&\le e^{-n(H(\bar{Y})-\gamma)}|\mathcal{M}_2|,\end{aligned}\tag{6.30}$$

where $\mathcal{A}_\gamma^n$ is the weakly typical set w.r.t P_Y. Conversely, given a testing scheme (ϕ_n,ψ_n) for the single-user hypothesis testing problem there exists a WAK-code $(\phi_n,\phi_{2n}',\psi_n')$ such that

$$\Pr\{\hat{\bar{Y}}^n\neq\bar{Y}^n\}\le\bar{\alpha}_n+e^{n(\hat{E}-\gamma)}\bar{\beta}_n+\Pr\{\bar{Y}^n\notin\mathcal{A}_\gamma^n\}+\frac{e^{n(H(\bar{Y})+2\gamma-\hat{E})}}{|\mathcal{M}_2|},\tag{6.31}$$

where $\hat{E}$ is a free parameter that satisfies $H(\bar{Y}) > \hat{E} - 2\gamma$.

The proof of Theorem 6.2 is similar to the ones of Proposition 5.1 and Proposition 5.2 in Chapter 5 and is given in the following.

Proof. WAK $\Rightarrow$ Single-user HT: For a given $m_1 \in \mathcal{M}_1$, define the following correctly decodable set of the WAK-code

$$\mathcal{D}_{m_1} = \{y^n \mid y^n = \psi_n(m_1, \phi_{2n}(y^n)),\ y^n \in \mathcal{A}_\gamma^n\}. \tag{6.32}$$

Then, it is clear that for all $m_1 \in \mathcal{M}_1$ we have $|\mathcal{D}_{m_1}| \leq |\mathcal{M}_2|$ as ϕ_{2n} can only take at most $|\mathcal{M}_2|$ values. A decision region for the single-user HT, based on $\mathcal{Y}^n \times \mathcal{M}_1$, is defined with a slight abuse of notation as

$$\bar{\mathcal{A}}_n = \bigcup_{m_1} \big(\mathcal{D}_{m_1} \times \{m_1\}\big) \subset \mathcal{Y}^n \times \mathcal{M}_1. \tag{6.33}$$

The validity of $\bar{\mathcal{A}}_n$, i.e., the existence of a decision mapping ψ_n in Definition 6.1, follows from the fact that we have full access to the sequence y^n when making a decision. We use the mapping ϕ_{1n} as the compression mapping for $\mathcal{X}^n$ in the single-user HT problem. From (6.33) the probability of type I of error is bounded by

$$\begin{aligned}\bar{\alpha}_n &= P_{\bar{Y}^n \phi_{1n}(\bar{X}^n)}(\bar{\mathcal{A}}_n^c) \\ &\leq P_{\mathrm{WAK}}\{\text{error}\} + \Pr\{\bar{Y}^n \notin \mathcal{A}_\gamma^n\},\end{aligned} \tag{6.34}$$

and the probability of type II of error is bounded above by

$$\begin{aligned}\bar{\beta}_n &= P_{\bar{Y}^n} \times P_{\phi_{1n}(\bar{X}^n)}(\bar{\mathcal{A}}_n) = \sum_{m_1} P_{\phi_{1n}(\bar{X}^n)}(m_1) P_{\bar{Y}^n}(\mathcal{D}_{m_1}) \\ &\leq \sum_{m_1} P_{\phi_{1n}(\bar{X}^n)}(m_1) |\mathcal{D}_{m_1}| e^{-n(H(Y)-\gamma)} \\ &\leq e^{-n(H(\bar{Y})-\gamma)} |\mathcal{M}_2|.\end{aligned} \tag{6.35}$$

WAK $\Leftarrow$ Single-user HT: Let $\hat{E}$ be arbitrary such that $\hat{E} - \gamma < H(\bar{Y})$. Given a single-user HT testing scheme (ϕ_n, ψ_n), we define the set

$$\mathcal{D}_n(m_1) = \left\{ y^n \middle| \frac{1}{n} \log \frac{P_{\bar{Y}^n|\phi_n(\bar{X}^n)}(y^n|m_1)}{P_{\bar{Y}^n}(y^n)} > \hat{E} - \gamma \right\} \cap \mathcal{A}_\gamma^n. \tag{6.36}$$

$\mathcal{D}_n(m_1)$ plays the role of the conditional typical set in the standard proof of the WAK setting, cf. [CT12],[EK11]. We use the mapping ϕ_n as the compression mapping for $\mathcal{X}^n$ in the WAK-problem. From the definition of $\mathcal{D}_n(m_1)$ we obtain

$$\Pr\big\{\bar{Y}^n \notin \mathcal{D}_n(\phi_n(\bar{X}^n))\big\}$$

$$\overset{(*)}{\leq} \bar{\alpha}_n + e^{n(\hat{E}-\gamma)}\bar{\beta}_n + \Pr\{\bar{Y}^n \notin \mathcal{A}_\gamma^n\}, \tag{6.37}$$

where $(*)$ follows from [Han03, Lemma 4.1.2]. Furthermore, we have $|\mathcal{D}_n(m_1)| \leq e^{n(H(\bar{Y})+2\gamma-\hat{E})}$ for all m_1 since

$$\begin{aligned} 1 &\geq P_{\bar{Y}|\phi_n(\bar{X}^n)}(\mathcal{D}_n(m_1)|m_1) \\ &= \sum_{y^n \in \mathcal{D}_n(m_1)} P_{\bar{Y}^n|\phi_n(\bar{X}^n)}(y^n|m_1) \geq \sum_{y^n \in \mathcal{D}_n(m_1)} P_{\bar{Y}^n}(y^n) e^{n(\hat{E}-\gamma)} \\ &\geq |\mathcal{D}_n(m_1)| e^{-n(H(Y)+\gamma)} e^{n(\hat{E}-\gamma)}. \end{aligned} \tag{6.38}$$

Let m_2 be a uniformly random bin index of y^n and $\mathcal{B}(m_2)$ be the set of all such y^n. The decoder decides that $\hat{y}^n$ is the reconstructed sequence if it is the unique sequence such that $\hat{y}^n \in \mathcal{B}(m_2) \cap \mathcal{D}_n(m_1)$, where m_1 and m_2 are sent messages from Encoder 1 and 2. It then follows that

$$\begin{aligned} &\Pr\{\hat{\bar{Y}}^n \neq \bar{Y}^n\} \leq \Pr\{\bar{Y}^n \notin \mathcal{B}(M_2) \cap \mathcal{D}_n(\phi_n(\bar{X}^n))\} \\ &+ \Pr\{\exists \tilde{y}^n \neq \bar{Y}^n, \tilde{y}^n \in \mathcal{D}_n(\phi_n(\bar{X}^n)) \cap \mathcal{B}(M_2)\} \\ &\overset{(a)}{\leq} \Pr\{\bar{Y}^n \notin \mathcal{D}_n(\phi_n(\bar{X}^n))\} + \Pr\{\exists \tilde{y}^n \neq \bar{Y}^n, \tilde{y}^n \in \mathcal{D}_n(\phi_n(\bar{X}^n)) \cap \mathcal{B}(M_2)\} \\ &\overset{(b)}{\leq} \alpha_n + e^{n(\hat{E}-\gamma)}\beta_n + \Pr\{\bar{Y}^n \notin \mathcal{A}_\gamma^n\} + \frac{e^{n(H(\bar{Y})+2\gamma-\hat{E})}}{|\mathcal{M}_2|}. \end{aligned} \tag{6.39}$$

(a) is valid since $\bar{Y}^n \in \mathcal{B}(M_2)$. The inequality (b) holds since each $\tilde{y}^n$ is assigned independently to a bin with probability $1/|\mathcal{M}_2|$ and the number of such $\tilde{y}^n$ is bounded by $e^{n(H(\bar{Y})+2\gamma-\hat{E})}$, cf. (6.38). The existence of deterministic mappings ϕ'_{2n} and ψ'_n follows immediately by the random coding argument. □

Remark 6.4 We discuss herein briefly the effect of a randomized test for the single-user HT. Let $T\colon \mathcal{M}_1 \times \mathcal{Y}^n \to [0,1]$ represent the probability that H_0 is chosen given (m_1, y^n). The corresponding probabilities of errors of type I and II are

$$\begin{aligned} \bar{\alpha}_n &= \sum_{y^n,x^n} P_{\bar{Y}^n\phi_n(\bar{X}^n)}(y^n, \phi_n(x^n))(1 - T(\phi_n(x^n), y^n)), \\ \bar{\beta}_n &= \sum_{y^n,x^n} P_{\bar{Y}^n}(y^n) P_{\phi_n(\bar{X}^n)}(\phi_n(x^n)) T(\phi_n(x^n), y^n). \end{aligned} \tag{6.40}$$

Then Theorem 6.2 is still valid since $(*)$ in (6.37) holds due to [PW17, Lemma 12.2].

As a consequence of Theorem 6.2 we show in the following a relation between ϵ-achievable rate regions. We recap the relevant definitions again in the following. Fix an $\epsilon \in [0,1)$.

Definition 6.3 Let the ϵ-achievable rate region for the WAK problem $\mathcal{R}_{\mathrm{WAK},\epsilon}$ be the closure of all (R_1, R_2) such that there exists a WAK-code $(\phi_{1n}, \phi_{2n}, \psi_n)$ which satisfies

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_i| \leq R_i,\ i = 1, 2, \text{ and } \limsup_{n\to\infty} \Pr\{\hat{\bar{Y}}^n \neq \bar{Y}^n\} \leq \epsilon. \tag{6.41}$$

Define the *minimum ϵ-achievable compression rate of Y^n* for a given compression rate R_1 of X^n as $R^\star_{2,\epsilon}(R_1) = \inf\{R_2 \mid (R_1, R_2) \in \mathcal{R}_{\mathrm{WAK},\epsilon}\}$.

Note that since $\mathcal{R}_{\mathrm{WAK},\epsilon}$ is a closed set, the $\inf\{\cdot\}$ operation in the definition of $R^\star_{2,\epsilon}(R_1)$ can be replaced by the $\min\{\cdot\}$ operation. As $(R_1, H(Y)) \in \mathcal{R}_{\mathrm{WAK},\epsilon}$, $\forall \epsilon \in [0, 1)$, $R^\star_{2,\epsilon}(R_1)$ is finite. Additionally, let us define the following:

Definition 6.4 Let the ϵ-achievable rate region for the single-user HT problem $\mathcal{R}_{\mathrm{HT},\epsilon}$ be the closure of all (R_c, E) such that there exists a single-user testing scheme (ϕ_n, ψ_n) which satisfies

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_1| \leq R_c,\ \limsup_{n\to\infty} \bar{\alpha}_n \leq \epsilon, \text{ and } \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\bar{\beta}_n} \geq E. \tag{6.42}$$

Define the *maximum ϵ-achievable error exponent* for a given compression rate R_c of X^n as $E^\star_\epsilon(R_c) = \sup\{E \mid (R_c, E) \in \mathcal{R}_{\mathrm{HT},\epsilon}\}$.

We observe further that $E^\star_\epsilon(R_c) \leq I(X;Y)$ since the later is the error exponent of uncompressed data. Additionally $E^\star_\epsilon(R_c) \geq 0$ also holds since $(R_c, 0) \in \mathcal{R}_{\mathrm{HT},\epsilon}$, $\forall \epsilon \in [0, 1)$.
The following result provides an alternative view on the proof of the strong converse for the single-user testing against independence problem.

Theorem 6.3 *For all $R_c > 0$ and for all $\epsilon \in [0, 1)$ we have*

$$E^\star_\epsilon(R_c) + R^\star_{2,\epsilon}(R_c) = H(\bar{Y}). \tag{6.43}$$

Proof. Given a $\gamma > 0$ and $\epsilon > 0$, there exists a WAK-code $(\phi_{1n}, \phi_{2n}, \psi_n)$ which satisfies all the conditions in Definition 6.3 for the rate pair $(R_c + \gamma, R^\star_{2,\epsilon}(R_c) + \gamma)$. This implies that for all sufficiently large n we have $|\mathcal{M}_2| \leq e^{n(R^\star_{2,\epsilon}(R_c)+2\gamma)}$. By the first part of Theorem 6.2, the corresponding testing scheme satisfies

$$\limsup_{n\to\infty} \bar{\alpha}_n \leq \epsilon,\ \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\bar{\beta}_n} \geq H(\bar{Y}) - 3\gamma - R^\star_{2,\epsilon}(R_c).$$

This implies that $(R_c + \gamma, H(\bar{Y}) - 3\gamma - R^\star_{2,\epsilon}(R_c)) \in \mathcal{R}_{\mathrm{HT},\epsilon}$. Since $\gamma > 0$ is arbitrary and $\mathcal{R}_{\mathrm{HT},\epsilon}$ is closed by definition, we obtain $E^\star_\epsilon(R_c) \geq H(\bar{Y}) - R^\star_{2,\epsilon}(R_c)$.
Conversely, given $\gamma > 0$ and $\epsilon > 0$ there exists a testing scheme (ϕ_n, ψ_n) such that $(R_c + \gamma, E^\star_\epsilon(R_c) - \gamma)$ is ϵ-achievable according to Definition 6.4, i.e., for all sufficiently large n we have

$$\bar{\beta}_n \leq e^{-n(E^\star_\epsilon(R_c)-2\gamma)}.$$

Choosing $\hat{E} = E_\epsilon^\star(R_c) - 2\gamma$ and $|\mathcal{M}_2| = e^{n(H(\bar{Y})+5\gamma-E_\epsilon^\star(R_c))}$ we obtain by the second part of Theorem 6.2 that

$$\limsup_{n\to\infty} \Pr\{\hat{\bar{Y}}^n \neq \bar{Y}^n\} \leq \epsilon, \tag{6.44}$$

which implies that $(R_c + \gamma, H(\bar{Y}) + 5\gamma - E_\epsilon^\star(R_c)) \in \mathcal{R}_{\mathrm{WAK},\epsilon}$. Since γ is arbitrary and $\mathcal{R}_{\mathrm{WAK},\epsilon}$ is also closed by definition, therefore we obtain $R_{2,\epsilon}^\star(R_c) \leq H(\bar{Y}) - E_\epsilon^\star(R_c)$. □

Another important consequence of Theorem 6.2, which states an equivalence of exponentially strong converse statements, is given as follows.

Theorem 6.4 *The following statements are equivalent:*

1. *For any code which satisfies*

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_1| \leq R_c, \textit{ and, } \limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_2| \leq R_2, \tag{6.45}$$

   in the WAK problem, if $(R_c, R_2) \notin \mathcal{R}_{\mathrm{WAK},0}$*, then* $\Pr\{Y^n \neq \hat{Y}^n\} \to 1$ *exponentially fast at a positive convergence rate.*

2. *For any single-user HT scheme with*

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_1| \leq R_c, \textit{ and, } \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_n} \geq E, \tag{6.46}$$

   if $(R_c, E) \notin \mathcal{R}_{\mathrm{HT},0}$*, then* $\alpha_n \to 1$ *exponentially fast at a positive convergence rate.*

Proof. Assume that the first statement holds. It suffices to show the second statement when $E < H(Y)$. Let (ϕ_n, ψ_n) be a hypothesis testing scheme such that $\limsup_{n\to\infty} \frac{1}{n} \log M \leq R_c$ and $\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_n} \geq E$ where $(R_c, E) \notin \mathcal{R}_{\mathrm{HT},0}$. Let $\gamma > 0$ be small enough such that $(R_c, H(Y) + 4\gamma - E) \notin \mathcal{R}_{\mathrm{WAK},0}$. By the second part of Theorem 6.2, there exists a WAK-code $(\phi_n, \phi'_{2n}, \psi'_n)$ such that with $\hat{E} = E - \gamma$ and $|\mathcal{M}_2| = e^{n(H(Y)+4\gamma-E)}$ we have

$$\Pr\{\hat{Y}^n \neq Y^n\} \leq \alpha_n + 2e^{-n\gamma} + \Pr\{Y^n \notin \mathcal{A}_\gamma^n\}. \tag{6.47}$$

Since the weakly typical set $\mathcal{A}_\gamma^n$ includes the strongly typical set $\mathcal{T}_\epsilon^n$ for a fixed, positive, and small enough ϵ, the last term goes to 0 exponentially with a convergence rate of at least $2\epsilon^2$. By the assumption $\Pr\{Y^n = \hat{Y}^n\}$ goes to 0 exponentially at a rate of $\eta > 0$, we then have $\alpha_n \to 1$ exponentially at a rate of $\min\{\eta, \gamma, 2\epsilon^2\}$.

Conversely, assume that the second statement holds. Let $(\phi_{1n}, \phi_{2n}, \psi_n)$ be a WAK-code such that

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_1| \leq R_c, \ \limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_2| \leq R_2, \tag{6.48}$$

where $(R_c, R_2) \notin \mathcal{R}_{\mathrm{WAK},0}$. Let $\gamma > 0$ be small enough such that $(R_c, H(Y) - 2\gamma - R_2) \notin \mathcal{R}_{\mathrm{HT},0}$. By the first part of Theorem 6.2, the constructed testing scheme satisfies $\liminf \frac{1}{n} \log \frac{1}{\beta_n} \geq H(Y) - 2\gamma - R_2$. The corresponding false alarm probability α_n hence goes to 1 exponentially at a rate of $\xi > 0$, or $P_{\mathrm{WAK}}\{\text{error}\} \to 1$ exponentially at a rate of $\min\{\xi, 2\epsilon^2\}$. □

Remark 6.5 Note that the following two conditions are needed for proof of Theorems 6.2 and 6.3: $P_{Y^n} = P_Y^{\otimes n}$ to ensure that $2^{-n(H(Y)+\gamma)} \leq P_{Y^n}(y^n) \leq 2^{-n(H(Y)-\gamma)}$, $(R_1, H(Y)) \in \mathcal{R}_{\mathrm{WAK},\epsilon}$ as well as $\Pr\{Y^n \notin \mathcal{A}_\gamma^n\} \to 0$, and $E^\star(R_c)$ is upper bounded by $H(Y)$. We present now a simple example on finite alphabets without the iid assumption of the joint distribution where these two conditions are still active. Hence, Theorem 6.3 is still valid. Let $P_{1,XY}$ and $P_{2,XY}$ be distributions which belong to the set $\{Q_{XY} \mid Q_X = P_X,\ Q_Y = P_Y\}$ such that $D(P_{1,XY} \| P_X \times P_Y) < D(P_{2,XY} \| P_X \times P_Y)$. Let the source distribution of the WAK problem and the distribution under H_0 in the HT problem be $P_{X^nY^n} = \alpha P_{1,XY}^{\otimes n} + (1-\alpha) P_{2,XY}^{\otimes n}$ where $\alpha \in (0,1)$. The distribution under H_1 is still $P_X^{\otimes n} \times P_Y^{\otimes n}$. The process governed by $P_{X^nY^n}$ is stationary but nonergodic, cf. [Han03, Fig. 1.5] for an illustration of its information-spectrum. The maximum error exponent of type II in the uncompressed case is given by [Han03, Example 4.2.1]

$$E_\epsilon^\star(\log|\mathcal{X}|) = \begin{cases} D(P_{1,XY} \| P_X \times P_Y) & \text{if } 0 \leq \epsilon < \alpha \\ D(P_{2,XY} \| P_X \times P_Y) & \text{when } \alpha \leq \epsilon < 1 \end{cases}.$$

Note that $E_\epsilon^\star(\log|\mathcal{X}|) \leq H(Y)$ holds. The strong converse however does not hold.

6.2.B Soft-covering implies strong converse

We now get back to our multi-user setting. Applying the blowing up approach from [AC86] to study the ϵ-achievable error exponent is challenging, since both distributions are not product distributions of iid random variables. We instead employ the information spectrum approach on top of the result by Ahlswede and Csiszár to show the converse. Similar to Definition 6.2 we have the following definition.

Definition 6.5 Let $\epsilon \in [0,1)$ be an arbitrarily given constant. An error exponent E of type II is *ϵ-achievable* given (R, R_c) if there exist compression and decision mappings (ϕ_n, ψ_n) such that

$$\limsup_{n\to\infty} \alpha_n \leq \epsilon,\ \limsup_{n\to\infty} \frac{1}{n} \log|\mathcal{M}_1| \leq R_c,\ \text{and}\ \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_n} \geq E. \tag{6.49}$$

We define $E_\epsilon^\star(R, R_c) = \sup\{E \mid E \text{ is } \epsilon\text{-achievable given } (R, R_c)\}$.

We first present a lemma that relates our multi-user setting to the single-user setting. It says that when the likelihood ratio test is considered the probability of type I of error in our multi-user setting is greater than or equal to the one of the single-user setting due to the presence of multiple users.

Lemma 6.1 *For any compression sequence (ϕ_n), and $\hat{E}, \gamma > 0$, we have for all sufficiently large n*

$$\begin{aligned}&\Pr\left\{\frac{1}{n}\log\frac{P_{Y^n\phi_n(\mathbf{X}^n)}(Y^n,\phi_n(\mathbf{X}^n))}{P_{Y^n}\times P_{\phi_n(\mathbf{X}^n)}(Y^n,\phi_n(\mathbf{X}^n))} > \hat{E}\right\}\\ &\le \Pr\left\{\iota(\bar{Y}^n;\phi_n(\bar{X}^n)) > \log M + n(\hat{E}-\gamma)\right\} + \mathcal{O}(\exp(-n\hat{E})),\end{aligned} \tag{6.50}$$

where the left-hand side of (6.50) *is evaluated with $(Y^n, \mathbf{X}^n) \sim P_{H_0}$ as given in* (6.1)*, $(\bar{Y}^n, \bar{X}^n) \sim P_{YX}^{\otimes n}$ and*

$$\iota(y^n;\phi_n(x^n)) = \log\frac{P_{\bar{Y}^n|\phi_n(\bar{X}^n)}(y^n|\phi_n(x^n))}{P_{\bar{Y}^n}(y^n)}.$$

Proof. We denote the LHS of (6.50) by $L_n(\hat{E},\gamma)$ and suppress the dependency on $(\hat{E},\gamma)$ in the proof for notation brevity. Given a compressed tuple $\phi_n(\mathbf{x}^n)$ of users' data sequences, we define the following (conditional) distribution on $\mathcal{Y}^n$

$$\hat{P}_{H_0,\phi_n(\mathbf{x}^n)}(y^n) = \frac{1}{M}\sum_{i=1}^{M} P_{Y^n|\phi_n(X^n)}(y^n|\phi_n(x^n(i))).$$

We observe that under hypothesis H_0 the induced joint distribution of $(Y^n, \phi_n(\mathbf{X}^n))$ can be reformulated as

$$P_{H_0,\phi_n}(y^n,\phi_n(\mathbf{x}^n)) = \hat{P}_{H_0,\phi_n(\mathbf{x}^n)}(y^n)\times\prod_{i=1}^{M} P_{\phi_n(X^n)}(\phi_n(x^n(i))). \tag{6.51}$$

The corresponding induced joint distribution under hypothesis H_1, P_{H_1,ϕ_n}, is equal to

$$P_{Y^n}(y^n)\times\prod_{k=1}^{M} P_{\phi_n(X^n)}(\phi_n(x^n(k))). \tag{6.52}$$

Therefore, since $(Y^n, \mathbf{X}^n) \sim P_{H_0}$ we can rewrite L_n as

$$\begin{aligned}L_n &= \mathbb{E}_{\phi_n(\mathbf{X}^n)}\left[\Pr\left\{\frac{\hat{P}_{H_0,\phi_n(\mathbf{X}^n)}}{P_{Y^n}}(Y^n) > \eta\middle|\phi_n(\mathbf{X}^n)\right\}\right]\\ &= \mathbb{E}_{\phi_n(\mathbf{X}^n)}F_\eta(\hat{P}_{H_0,\phi_n(\mathbf{X}^n)}||P_{Y^n}).\end{aligned} \tag{6.53}$$

Herein we have

$$F_\eta(P||Q) = \Pr\left\{\frac{dP}{dQ}(X) > \eta\right\}$$

where $X \sim P$ is the excess relative information metric with threshold η, $\eta = e^{n\hat{E}}$, as defined in[6.2] [LCV17]. For each tuple $\phi_n(\mathbf{x}^n)$, which is a realization of $\phi_n(\mathbf{X}^n)$, we

[6.2] The metric is denoted therein by $\bar{F}_\eta(P||Q)$. We use a slightly different notation herein, since $\bar{(\cdot)}$ has been employed to denote the single-user case.

can view $\hat{P}_{H_0,\phi_n(\mathbf{x}^n)}$ as the output distribution induced by selecting one sequence in the tuple uniformly at random and feeding it into the input of the channel $P_{Y^n|\phi_n(X^n)}$. The soft-covering lemma for the F_η metric in [LCV17, Theorem 24] states that

$$\begin{aligned}
&\mathbb{E}_{\phi_n(\mathbf{X}^n)} F_\eta(\hat{P}_{H_0,\phi_n(\mathbf{X}^n)} || P_{Y^n}) \\
&\leq \Pr\big[\iota(\bar{Y}^n; \phi_n(\bar{X}^n)) > \log(M\sigma)\big] + \frac{1}{\nu}\Pr\big[\iota(\bar{Y}^n; \phi_n(\bar{X}^n)) > \log M - \tau\big] \\
&+ \frac{\exp(-\tau)}{(\eta - 1 - \nu - \sigma)^2},
\end{aligned} \tag{6.54}$$

where herein $\sigma, \nu > 0$ are arbitrarily satisfying $\eta - 1 > \nu + \sigma$, $\tau \in \mathbb{R}$ and $(\bar{Y}^n, \bar{X}^n) \sim P_{YX}^{\otimes n}$. If we take, $\tau = -n\hat{E}$, $\sigma = \eta/4 - 1$ and $\nu = \eta/4$, then we obtain

$$\begin{aligned}
&\mathbb{E}_{\phi_n(\mathbf{X}^n)} F_\eta(\hat{P}_{H_0,\phi_n(\mathbf{X}^n)} || P_{Y^n}) \\
&\leq \Pr\Big\{\iota(\bar{Y}^n; \phi_n(\bar{X}^n)) > \log M + n\hat{E} + \log(1/4 - 1/\eta))\Big\} \\
&+ 8\exp(-n\hat{E}).
\end{aligned} \tag{6.55}$$

The conclusion of the lemma follows. □

Roughly speaking, to provide a strong converse statement we aim to drive L_n to 0 as $n \to \infty$. It can be seen that if $\hat{E}$ is greater than the spectral-sup mutual information of $(Y^n, \phi_n(\mathbf{X}^n))$ then L_n goes to 0. However, the bound is hard to characterize in a single letter form. The following corollary shows that there exists a sequence (L_{n_k}) which goes to 0. It will be shown later that the conclusion is sufficient for proving a strong converse statement.

Corollary 6.1 *Given a compression sequence (ϕ_n) such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_1| \leq R_c.$$

If $\hat{E} = R_{\max} - R + 3\gamma$ where $R \leq R_{\max}$ and $\gamma > 0$ is arbitrary, then there exists a subsequence $(n_k)_{k=1}^{\infty}$ such that

$$\lim_{k\to\infty} L_{n_k}(\hat{E}, \gamma) = 0. \tag{6.56}$$

Proof. It suffices to show that there exists a subsequence (n_k) such that the first term in the RHS of (6.50) converges to 0.
Define the following acceptance region for the single-user hypothesis testing problem

$$\bar{\mathcal{A}}_n = \Big\{(y^n, \phi_n(x^n)) \mid \iota(y^n; \phi_n(x^n)) > n\tilde{E}\Big\}, \tag{6.57}$$

where $\tilde{E} = R_{\max} + \gamma$. Then it can be seen that

$$\begin{aligned}\bar{\beta}_n &= P_{\bar{Y}^n} \times P_{\phi_n(\bar{X}^n)}(\bar{\mathcal{A}}_n) \\ &= \sum_{(y^n,\phi_n(x^n))\in\bar{\mathcal{A}}_n} P_{\bar{Y}^n}(y^n) P_{\phi_n(\bar{X}^n)}(\phi_n(x^n)) \\ &\leq e^{-n\tilde{E}} \sum_{(y^n,\phi_n(x^n))\in\bar{\mathcal{A}}_n} P_{\bar{Y}^n\phi_n(\bar{X}^n)}(y^n,\phi_n(x^n)) \\ &\leq e^{-n\tilde{E}}, \end{aligned} \tag{6.58}$$

thus $\liminf_{n\to\infty} \frac{1}{n}\log\frac{1}{\bar{\beta}_n} \geq \tilde{E}$. Since $\tilde{E} > R_{\max}$ by the strong converse result of Ahlswede and Csiszár [AC86, Theorem 3] we have

$$\limsup_{n\to\infty} \bar{\alpha}_n = 1, \text{ where } \bar{\alpha}_n = P_{\bar{Y}^n\phi_n(\bar{X}^n)}(\bar{\mathcal{A}}_n^c). \tag{6.59}$$

Then there exists a subsequence $(n_k)_{k=1}^{\infty}$ such that

$$\lim_{k\to\infty} \bar{\alpha}_{n_k} = 1 \Leftrightarrow \lim_{k\to\infty} P_{\bar{Y}^{n_k}\phi_{n_k}(\bar{X}^{n_k})}(\bar{\mathcal{A}}_{n_k}) = 0. \tag{6.60}$$

Additionally, for all sufficiently large n we have

$$\hat{E} - \gamma + \frac{1}{n}\log M > \tilde{E}$$

which implies further that

$$P_{\bar{Y}^n\phi_n(\bar{X}^n)}(\bar{\mathcal{A}}_n) \geq \Pr\{\iota(\bar{Y}^n;\phi_n(\bar{X}^n)) > \log M + n(\hat{E}-\gamma)\}.$$

By Lemma 6.1 and (6.60) we then obtain

$$\lim_{k\to\infty} L_{n_k}(\hat{E},\gamma) = 0. \tag{6.61}$$

□

We now summarize the above analysis in the following theorem, which is the strong converse statement.

Theorem 6.5 *The strong converse holds, i.e., for all* ϵ, $0 \leq \epsilon < 1$,

$$E_\epsilon^\star(R,R_c) = \theta(R,R_c) = R_{\max} - R, \text{ if } R \leq R_{\max}. \tag{6.62}$$

It is interesting to note that Theorem 6.5 shows the tightness of $E^\star(R,R_c)$ for $R \leq R_{\max}$. It is not clear whether the same statement can be reached with the methods in Section 6.1. Although, the linear dependency of $E_\epsilon^\star(R,R_c)$ on R when $R \leq R_{\max}$ is interesting, we do not have a clear explanation for this behaviour. However, we believe that it holds due to the specific symmetry of the setting where the processes are mixed uniformly in P_{H_0}. The proof of Theorem 6.5 uses a similar idea as the strong converse proof of Stein's Lemma 2.2 in Chapter 2.

Proof. Suppose that there exists a sequence of compression mappings (ϕ_n) and a sequence of decision mappings (ψ_n) such that

$$\limsup_{n\to\infty} \alpha_n \le \epsilon,\ \limsup_{n\to\infty} \frac{1}{n}\log|\mathcal{M}_1| \le R_c,\ \text{and}\ \liminf_{n\to\infty} \frac{1}{n}\log\frac{1}{\beta_n} \ge E. \tag{6.63}$$

Given $\gamma > 0$ small enough such that $\epsilon + \gamma < 1$, then for all sufficiently large n we have $\alpha_n \le \epsilon + \gamma$ and $\beta_n \le e^{-n(E-\gamma)}$. Select $\hat{E} = R_{\max} - R + 3\gamma$ and let (n_k) be the corresponding subsequence such that (6.56) holds. Then by applying Lemma 2.8, c.f. also [PW17, Section 13.1], we have

$$1 - \alpha_n - e^{n\hat{E}}\beta_n \le \Pr\left\{\frac{1}{n}\log\frac{P_{Y^n\phi_n(\mathbf{X}^n)}}{P_{Y^n}\times P_{\phi_n(\mathbf{X}^n)}}(Y^n,\phi_n(\mathbf{X}^n)) > \hat{E}\right\},$$

holds for any n. Since, the RHS is actually $L_n(\hat{E},\gamma)$ as defined in proof of Lemma 6.1, we obtain for all sufficiently large n_k the following

$$\begin{aligned} 1-\epsilon-\gamma-e^{n_k\hat{E}}e^{-n_k(E-\gamma)} &\le L_{n_k} \\ &\implies \hat{E} - E + \gamma \ge \frac{1}{n_k}\log(1-\epsilon-\gamma-L_{n_k}) \\ &\implies R_{\max} - R + 4\gamma \ge E \overset{\gamma\to 0}{\implies} R_{\max} - R \ge E \\ &\implies R_{\max} - R \ge E_\epsilon^\star(R,R_c). \end{aligned} \tag{6.64}$$

The last inequality holds since E is an arbitrarily ϵ-achievable exponent. The conclusion follows since $R_{\max} - R \le E_\epsilon^\star(R,R_c)$ when $R < R_{\max}$ by Theorem 6.1. Note that the upper bound still holds even if we define $E_\epsilon^\star(R,R_c)$ on the closure of ϵ-achievable region of (R_c,E). □

Due to the presence of multiple users in our setting, it is natural to ask for the behavior of errors of type I and II when the number of users exceeds the identifiable threshold, i.e., $R > R_{\max}$. We provide some partial information for their behaviors in the following theorem.

Theorem 6.6 *Given a sequence of compression mappings (ϕ_n) such that*

$$\limsup_{n\to\infty} \frac{1}{n}\log|\mathcal{M}_1| \le R_c.$$

Consider the case $R = R_{\max}(R_c)+\gamma$ where $\gamma > 0$ is arbitrary, then for any sequence of decision mappings (ψ_n), the following holds

$$\limsup_{n\to\infty}(\alpha_n + \beta_n) \ge 1. \tag{6.65}$$

Moreover, in case of no compression, i.e., correspondingly $R = I(X;Y) + \gamma$, we obtain $\lim_{n\to\infty}(\alpha_n + \beta_n) = 1$.

We interpret the result of Theorem 6.6 via the receiver operating characteristic curve as follows. In the design process, one aims to attain the highest detection probability for a given false alarm level. Theorem 6.5 states that the detection probability can be driven to 1 as long as the number of users is below $R_{\max}(R_c)$ and the false alarm level is strictly below 1. However, Theorem 6.6 says that when the number of users in the system exceeds $R_{\max}(R_c)$, the performance of *any* decision rule is not significantly better than a random guess.

Proof. From (6.51) and (6.52), the variational distance between P_{H_0,ϕ_n} and P_{H_1,ϕ_n} is given by

$$\|P_{H_0,\phi_n} - P_{H_1,\phi_n}\|_{TV} = \mathbb{E}_{\phi_n(\mathbf{X}^n)}\big\|\hat{P}_{H_0,\phi_n(\mathbf{X}^n)} - P_{Y^n}\big\|_{TV}.$$

By the soft-covering lemma [Cuf13, Corollary VII.2], [Hay06, Lemma 2], we obtain that

$$\begin{aligned}&\|P_{H_0,\phi_n} - P_{H_1,\phi_n}\|_{TV}\\ &\le \Pr\left\{\frac{1}{n}\iota(\bar{Y}^n;\phi_n(\bar{X}^n)) > R_{\max} + \frac{\gamma}{2}\right\} + \frac{1}{2}\sqrt{\frac{e^{n(R_{\max}+\gamma/2)}}{M}}.\end{aligned} \tag{6.66}$$

From the definition of the total variational distance we obtain

$$\begin{aligned}|1-\alpha_n-\beta_n| &\le \sup_{\mathcal{A}} |P_{H_0,\phi_n}(\mathcal{A}) - P_{H_1,\phi_n}(\mathcal{A})|\\ &= \|P_{H_0,\phi_n} - P_{H_1,\phi_n}\|_{TV}.\end{aligned} \tag{6.67}$$

Let $\bar{\mathcal{A}}_n$ be defined as in (6.57) with $\tilde{E} = R_{\max} + \gamma/2$ instead. Then by using (6.66) we obtain

$$\begin{aligned}\liminf_{n\to\infty}|1-(\alpha_n+\beta_n)| &\le \liminf_{n\to\infty}\|P_{H_0,\phi_n} - P_{H_1,\phi_n}\|_{TV}\\ &\le \liminf_{n\to\infty} P_{\bar{Y}^n\phi_n(\bar{X}^n)}(\bar{\mathcal{A}}_n) \overset{(6.60)}{=} 0,\end{aligned} \tag{6.68}$$

which implies $\limsup_{n\to\infty}(\alpha_n+\beta_n) \ge 1$. In case of no compression we replace the $\liminf_{n\to\infty}(\cdot)$ operation by the $\lim_{n\to\infty}(\cdot)$ operation and use the weak law of large numbers in the last step. □

Combining the results of Theorem 6.5 and Theorem 6.6 we obtain $E_\epsilon^\star(R,R_c) = \max\{R_{\max}(R_c) - R, 0\}$. We observe an information loss in the sense that when $R > R_{\max}(R_c)$ a compression scheme ϕ_n can achieve a positive error exponent in the single-user case while our multi-user error exponent is zero. Along the flow of Section 6.2.A we establish in the following a reverse statement of Theorem 6.1.

Proposition 6.3 *Fix $\hat{E}$ and $\gamma > 0$. Given a multi-user testing scheme (ϕ_n,ψ_n) with probabilities of errors (α_n,β_n) there exists a single-user testing scheme (ϕ_n,ψ_n') such that the corresponding probabilities of errors are given by*

$$\bar{\alpha}_n \le \alpha_n + e^{n\hat{E}}\beta_n + \mathcal{O}(\exp(-n\hat{E}))$$

$$\bar{\beta}_n \le e^{-n(\hat{E}-\gamma)} \frac{1}{M}, \tag{6.69}$$

for all sufficiently large n.

Similarly to the discussion after Theorem 6.1 the above bounds are loose when $R > R_{\max}(R_c)$.

Proof. Define a single-user decision region as

$$\bar{\mathcal{A}}_n = \big\{(y^n, \phi_n(x^n)) \mid \iota(y^n; \phi(x^n)) > \log M + n(\hat{E} - \gamma)\big\}. \tag{6.70}$$

By Lemma 6.1 we obtain

$$\begin{aligned} \alpha_n + e^{n\hat{E}}\beta_n &\ge 1 - L_n(\hat{E}, \gamma) \\ &\ge P_{\bar{Y}^n \phi_n(\bar{X}^n)}(\bar{\mathcal{A}}_n^c) - 8\exp(-n\hat{E}) = \bar{\alpha}_n - 8\exp(-n\hat{E}). \end{aligned} \tag{6.71}$$

Similar to (6.58), by the change of measure we also have

$$\bar{\beta}_n \le \frac{e^{-n(\hat{E}-\gamma)}}{M}. \tag{6.72}$$

□

Theorem 6.1 and Proposition 6.3 indicate that when $R < R_{\max}(R_c)$ the same, close to optimal, compression mapping ϕ_n can be used for the single-user HT and multi-user HT settings to achieve the ϵ-achievable performances. This mapping can also be used for the WAK setting according to Theorem 6.2. The relation about ϵ-achievable performance for the WAK and the identification problem is given in Theorem 5.7. We summarize these relations in Fig. 6.1 where the solid arrow indicates a constructive transformation while the dashed arrow indicates a non-constructive transformation. The constructive transformations between the single-user HT against independence and the identification problem is shown in the next subsection.

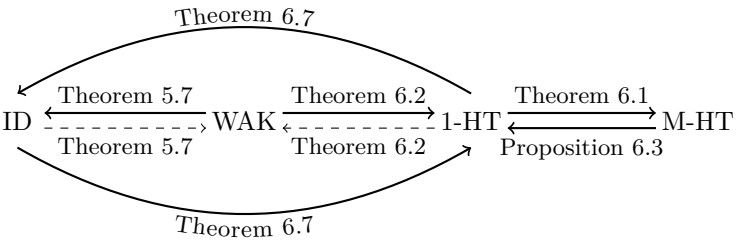


Figure 6.1: Reusability of the same compression mapping across settings.

6.2.C Equivalence between single-user HT and Identification

We establish herein the equivalence between the single-user hypothesis testing against independence and the identification problem. In both settings we allow the compression of the sequence y^n. Two hypotheses in the HT problem are $H_0\colon P_{X^nY^n}$, $H_1\colon P_{X^n}\times P_{Y^n}$. Note that we do not assume either $P_{X^nY^n}=P_{XY}^{\otimes n}$, or $\mathcal{X}$ and $\mathcal{Y}$ are finite. However, it should be assumed that $\mathcal{X}$ and $\mathcal{Y}$ are *nice* enough, for example Polish spaces. With abuse of terminology and notation we will redefine some terms and notations in the subsequent development.

A testing scheme consists of two compression mappings (ϕ_{1n},ϕ_{2n}) and a decision mapping ψ_n where

$$\begin{aligned}\phi_{1n}\colon \mathcal{X}^n\to\mathcal{M}_1,\ \phi_{2n}\colon \mathcal{Y}^n\to\mathcal{M}_2\\ \psi_n\colon \mathcal{M}_1\times\mathcal{M}_2\to\{0,1\}.\end{aligned}\tag{6.73}$$

The acceptance region can be defined similarly as in (6.5). The probabilities of type I and II errors $\bar{\alpha}_n$ and $\bar{\beta}_n$ can also be determined accordingly. For the iid case, this setup was discussed briefly in [AC86], for which a single-letter characterization for the optimal achievable error exponent of type II of error is still challenging.

Similarly, the joint distribution of the users' information, the observation and the randomly selected index in the ID problem is given as

$$P_{Y^n\mathbf{X}^nW}(y^n,\mathbf{x}^n,w)=\frac{1}{M}P_{Y^n|X^n}(y^n|x^n(w))\times\prod_k P_{X^n}(x^n(k)).\tag{6.74}$$

An identification scheme consists of two compression mappings (ϕ_{1n},ϕ_{2n}) and a decoding mapping ψ_n where

$$\phi_{1n}\colon \mathcal{X}^n\to\mathcal{M}_1,\ \phi_{2n}\colon \mathcal{Y}^n\to\mathcal{M}_2,\ \text{and}\ \psi_n\colon \mathcal{M}_1^M\times\mathcal{M}_2\to\mathcal{W}\cup\{e\}.\tag{6.75}$$

In the identification problem one wants to control the probability of incorrect identification $\Pr\{\hat{W}\neq W\}$, where $\hat{W}=\psi_n(\phi_{1n}(\mathbf{X}^n),\phi_{2n}(Y^n))$. This setting was studied in [WO08] when $P_{X^nY^n}=P_{XY}^{\otimes n}$, where inner bounds and outer bounds on the achievable rate region were derived. A connection between the achievable regions of these two problems has been drawn recently in [PPM16] via the entropy characterization.

We first establish the following useful lemma, which is a generalization of Lemma 5.2.

Lemma 6.2 *For a given $\gamma>0$ and a given identification scheme $(\phi_{1n},\phi_{2n},\psi_n)$, we have*

$$\Pr\{\hat{W}\neq W\}\geq\Pr\Big\{\iota(\phi_{2n}(\bar{Y}^n);\phi_{1n}(\bar{X}^n))\leq\log M-n\gamma\Big\}-e^{-n\gamma},$$

where again $(\bar{Y}^n,\bar{X}^n)\sim P_{Y^nX^n}$ and

$$\iota(\phi_{2n}(y^n);\phi_{1n}(x^n))=\log\frac{P_{\phi_{2n}(\bar{Y}^n)|\phi_{1n}(\bar{X}^n)}}{P_{\phi_{2n}(\bar{Y}^n)}}(\phi_{2n}(y^n),\phi_{1n}(x^n)).\tag{6.76}$$

It can be seen that the first term on the right-hand side of Lemma 6.2 is the corresponding false alarm probability when testing $H_0 : P_{X^nY^n}$ versus $H_1 : P_{X^n} \times P_{Y^n}$ using the log-likelihood ratio test with the corresponding threshold $\log M - n\gamma$. The proof of Lemma 6.2 is presented in Appendix 6.B. Using Lemma 6.2 we can establish an analogue of Theorem 6.2 as follows.

Theorem 6.7 *Fix an arbitrary $\gamma > 0$. Given an identification scheme $(\phi_{1n}, \phi_{2n}, \psi_n)$, we can construct a single-user HT scheme $(\phi_{1n}, \phi_{2n}, \psi'_n)$ such that the corresponding error probabilities of type I and II are given by*

$$\bar{\alpha}_n \le \Pr\{\hat{W} \ne W\} + e^{-n\gamma}, \text{ and } \bar{\beta}_n \le \frac{e^{n\gamma}}{M}. \tag{6.77}$$

Conversely, given a testing scheme $(\phi_{1n}, \phi_{2n}, \psi_n)$ for the single-user HT problem, we can construct an identification scheme $(\phi_{1n}, \phi_{2n}, \psi'_n)$ such that

$$\Pr\{\hat{W} \ne W\} \le \bar{\alpha}_n + M\bar{\beta}_n. \tag{6.78}$$

Proof. ID $\Rightarrow$ Single-user HT: From a given identification scheme $(\phi_{1n}, \phi_{2n}, \psi_n)$ we use the same pair of mappings (ϕ_{1n}, ϕ_{2n}) to compress information in the single-user HT. From the interpretation of the Lemma 6.2 given above, it is natural to define an acceptance region for the single-user HT setup as

$$\bar{\mathcal{A}}_n = \{(\phi_{1n}(x^n), \phi_{2n}(y^n)) \mid \iota(\phi_{2n}(y^n); \phi_{1n}(x^n)) > \log M - n\gamma\}. \tag{6.79}$$

The probability of type I of error is then given by

$$\bar{\alpha}_n = P_{\phi_{1n}(\bar{X}^n)\phi_{2n}(\bar{Y}^n)}(\bar{\mathcal{A}}_n^c) \le \Pr\{\hat{W} \ne W\} + e^{-n\gamma}, \tag{6.80}$$

where the inequality follows from Lemma 6.2. By the change of measure we also obtain

$$\bar{\beta}_n = P_{\phi_{1n}(\bar{X}^n)} \times P_{\phi_{2n}(\bar{Y}^n)}(\bar{\mathcal{A}}_n) \le \frac{e^{n\gamma}}{M}. \tag{6.81}$$

ID $\Leftarrow$ Single-user HT: Given a testing scheme $(\phi_{1n}, \phi_{2n}, \psi_n)$ of the single-user HT, let $\bar{\mathcal{A}}_n$ be the acceptance region. We use the mapping ϕ_{1n} to compress each user's information and store it into a database and the mapping ϕ_{2n} to compress the observation sequence y^n in the identification setting. We define the decoding rule as follows. We search for a unique $\hat{w}$ such that

$$(\phi_{2n}(y^n), \phi_{1n}(x^n(\hat{w}))) \in \bar{\mathcal{A}}_n. \tag{6.82}$$

If there exists none or there is more than one of such index, we output e. Define the following error events

$$\begin{aligned} \mathcal{E}_1 &= \{(\phi_{2n}(Y^n), \phi_{1n}(X^n(W))) \notin \bar{\mathcal{A}}_n\} \\ \mathcal{E}_2 &= \{\exists \tilde{w} \ne W \mid (\phi_{2n}(Y^n), \phi_{1n}(X^n(\tilde{w}))) \in \bar{\mathcal{A}}_n\}. \end{aligned} \tag{6.83}$$

Then

$$\Pr\{\hat{W} \neq W\} \leq \Pr\{\mathcal{E}_1\} + \Pr\{\mathcal{E}_2\}. \tag{6.84}$$

The probability of the first event is given as

$$\Pr\{\mathcal{E}_1\} = P_{\phi_{2n}(\bar{Y}^n)\phi_{1n}(\bar{X}^n)}(\bar{\mathcal{A}}_n^c) = \bar{\alpha}_n. \tag{6.85}$$

The probability of the second event is upper bounded as

$$\begin{aligned}
\Pr\{\mathcal{E}_2\} &\leq \sum_w \frac{1}{M} \sum_{\tilde{w}\neq w} P_{\phi_{2n}(Y^n)|W=w} \times P_{\phi_{1n}(X^n(\tilde{w}))}(\bar{\mathcal{A}}_n) \\
&= \sum_w \frac{1}{M} \sum_{\tilde{w}\neq w} P_{\phi_{2n}(\bar{Y}^n)} \times P_{\phi_{1n}(\bar{X}^n)}(\bar{\mathcal{A}}_n) \\
&\leq M\bar{\beta}_n.
\end{aligned} \tag{6.86}$$

Hence the reverse direction follows. □

We are now ready to present a connection between ϵ-achievable rate regions of the two settings. For that purpose we need some additional definitions, which we briefly state in the following. For an arbitrary but fixed $\epsilon \in [0,1)$, define $\mathcal{R}_{\mathrm{ID},\epsilon}$ to be the closure of all tuples (R_1, R_2, R) such that there exists an identification scheme $(\phi_{1n}, \phi_{2n}, \psi_n)$ which satisfies

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_i| \leq R_i,\ i = 1,2, \\
&\liminf_{n\to\infty} \frac{1}{n} \log M \geq R,\ \limsup_{n\to\infty} \Pr\{\hat{W} \neq W\} \leq \epsilon.
\end{aligned} \tag{6.87}$$

Then, we can define the *ϵ-identification capacity* for a given compression rate pair (R_1, R_2) as $R_\epsilon^\star(R_1, R_2) = \sup\{R \mid (R_1, R_2, R) \in \mathcal{R}_{\mathrm{ID},\epsilon}\}$.
Further, define $\mathcal{R}_{\mathrm{HT},\epsilon}$ to be the closure of all tuples (R_1, R_2, E) such that there exists a single-user HT scheme $(\phi_{1n}, \phi_{2n}, \psi_n)$ such that

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_i| &\leq R_i,\ i = 1,2, \\
\limsup_{n\to\infty} \bar{\alpha}_n \leq \epsilon,\ \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\bar{\beta}_n} &\geq E.
\end{aligned} \tag{6.88}$$

Similarly we define the *maximum ϵ- achievable error exponent* for a given compression rate pair (R_1, R_2) as $E_\epsilon^\star(R_1, R_2) = \sup\{E \mid (R_1, R_2, E) \in \mathcal{R}_{\mathrm{HT},\epsilon}\}$.

Theorem 6.8 *For all $\epsilon \in [0,1)$ and for all $(R_1, R_2) \in \mathbb{R}_+^2$, the following equality holds $E_\epsilon^\star(R_1, R_2) = R_\epsilon^\star(R_1, R_2)$.*

Proof. Assume that both quantities are finite. Given $\gamma > 0$ there exists an identification scheme $(\phi_{1n}, \phi_{2n}, \psi_n)$ such that all conditions in (6.87) are satisfied for $(R_1 + \gamma, R_2 + \gamma, R_\epsilon^\star(R_1, R_2) - \gamma)$. This implies that for all sufficiently large n we have $M \geq e^{n(R_\epsilon^\star(R_1,R_2)-2\gamma)}$. Then by the first part of Theorem 6.7 the probabilities of error of the corresponding single-user HT scheme are bounded by

$$\limsup_{n\to\infty} \bar{\alpha}_n \leq \epsilon, \ \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\bar{\beta}_n} \geq R_\epsilon^\star(R_1, R_2) - 3\gamma. \tag{6.89}$$

This implies that $E_\epsilon^\star(R_1, R_2) \geq R_\epsilon^\star(R_1, R_2)$, by taking $\gamma \to 0$.
Conversely, there exists a single-user testing scheme $(\phi_{1n}, \phi_{2n}, \psi_n)$ such that all conditions in (6.88) are satisfied for $(R_1 + \gamma, R_2 + \gamma, E_\epsilon^\star(R_1, R_2) - \gamma)$. This implies that for all sufficiently large n we have $\bar{\beta}_n \leq e^{-n(E_\epsilon^\star(R_1,R_2)-2\gamma)}$. By choosing $M = e^{n(E_\epsilon^\star(R_1,R_2)-3\gamma)}$ we obtain

$$\limsup_{n\to\infty} \Pr\{\hat{W} \neq W\} \leq \epsilon, \tag{6.90}$$

which implies that $(R_1 + \gamma, R_2 + \gamma, E_\epsilon^\star(R_1, R_2) - 3\gamma) \in \mathcal{R}_{\mathrm{ID},\epsilon}$. Hence, we have $E_\epsilon^\star(R_1, R_2) \leq R_\epsilon^\star(R_1, R_2)$.

Next, if $E_\epsilon^\star(R_1, R_2) = \infty$ for some pair $(R_1, R_2) \in \mathbb{R}_+^2$ and $\epsilon > 0$, then we can modify the proof as follows: Let $\{E_m\}_{m=1}^\infty$ be a sequence such that $E_m < \infty, \forall m$, and $E_m \to \infty$ as $m \to \infty$. Then we replace $E_\epsilon^\star(R_1, R_2)$ with E_m in the last paragraph to get $(R_1, R_2, E_m) \in \mathcal{R}_{\mathrm{ID},\epsilon}$. This holds for any m, hence $R_\epsilon^\star(R_1, R_2) = \infty$ as well. The case $R_\epsilon^\star(R_1, R_2) = \infty$ can be handled similarly. □

Remark 6.6 It follows from [Han03, Theorem 3.5.2] that $\bar{I}(\phi_{1n}(\bar{\mathbf{X}}), \phi_{2n}(\bar{\mathbf{Y}})) \leq \min\{R_1, R_2\} + \gamma$ holds where $\gamma > 0$ is arbitrary. Therefore, Lemma 6.2 implies that $R_\epsilon^\star(R_1, R_2) \leq \min\{R_1, R_2\} + \gamma$. Hence both $R_\epsilon^\star(R_1, R_2)$ and $E_\epsilon^\star(R_1, R_2)$ are finite and equal each other.

Remark 6.7 As the consequence of Theorem 6.8, the strong converse for the Gaussian ID setting presented in Appendix 5.E implies a strong converse proof of the HT against independence for the iid Gaussian with one side compression scenario. Note further that we can replace $n\gamma$ in Lemma 6.2 with $\sqrt{n}\gamma$. Hence, we could establish a second-order relation between these two settings as follows. If in the first part of Theorem 6.7 we have $M \geq e^{n(R+S/\sqrt{n})}$ then we obtain $\beta_n \leq e^{-n(R+(S-\gamma)/\sqrt{n})}$. If in the second part of Theorem 6.7 we have $\beta_n \leq e^{-n(E+S/\sqrt{n})}$ then we can choose $M = e^{n(E+(S-\gamma)/\sqrt{n})}$.

Remark 6.8 Let (ϕ_n, ψ_n) be a multi-user testing scheme which achieves a positive error exponent $E > 0$ when $R < R_{\max}(R_c)$. The scheme exists since $E(R, R_c) = R_{\max}(R_c) - R$ is positive. Then by choosing $\gamma < \hat{E} < E$ in Proposition 6.3 we obtain a single-user testing scheme (ϕ_n, ψ_n') such that $\bar{\alpha}_n \to 0$ and $\bar{\beta}_n$ goes to 0 at a rate $\hat{E} + R$. Applying Theorem 6.7 we obtain an identification mapping $\tilde{\psi}_n$ such that

$$\Pr\{\hat{W} \neq W\} \leq \bar{\alpha}_n + e^{-n(\hat{E}-\gamma)} \to 0. \tag{6.91}$$

Hence the relation $\phi_n \in \mathcal{B}(R, R_c)$ holds, i.e. (6.29) is also the necessary condition.

6.A Supplementary arguments for Remark 6.3

Since $R < R_{\max}$ there exists for any $\delta > 0$ a conditional distribution $P_{U|X}$ such that $R_c > I(X;U)$ and $\max\{R, R_{\max} - \delta\} < I(Y;U) < R_{\max}$ hold. We are going to show that $I(Y;U) - R$ is an achievable exponent of the second type of error and the probability of identifying W erroneously goes to zero as n goes to ∞.
Codebook: Generate e^{nR_c} codewords $u^n(m)$ iid according to the marginal distribution P_U.
Enrollment: For each user i, we look for a codeword $u^n(m_i)$ such that

$$(x^n(i), u^n(m_i)) \in \mathcal{T}_\epsilon^n(P_{XU}),$$

where $\mathcal{T}_\epsilon^n$ denotes the strongly typical set, and store m_i as j_i into the database.
Acceptance region & Decoding mapping: $\mathcal{A}_n$ is defined (for a given codebook) as the set

$$\mathcal{A}_n = \{(y^n, (x^n(i))_{i=1}^M) \mid (y^n, u^n(j_i)) \in \mathcal{T}_\epsilon^n(P_{YU}) \text{ for some } i\}. \tag{6.92}$$

To obtain $\tilde{\psi}_n$ we define the decoding rule as follows: We declare $\hat{w}$ to be the output of $\tilde{\psi}_n$ if it is the unique index such that $(y^n, u^n(j_{\hat{w}})) \in \mathcal{T}_\epsilon^n(P_{YU})$. If there is none or more than one such index then the output of $\tilde{\psi}_n$ is a fixed error indicator symbol e. The analysis of the probabilities of errors of type I, II and decoding error can be proceeded similarly as in the proof of Theorem 6.1 and as in [Tun09]. We obtain

$$E_{\mathcal{B}}^\star(R, R_c) \geq \theta(R, R_c). \tag{6.93}$$

To show the other direction, i.e., $E_{\mathcal{B}}^\star(R, R_c) \leq \theta(R, R_c)$, we note that from the weak converse

$$I(Y^n; \phi_n(\mathbf{X}^n)) \geq -(1 - \alpha_n) \log \beta_n - h_2(\alpha_n). \tag{6.94}$$

The left-hand side can be expressed further as

$$nH(Y) - H(Y^n|W, \phi_n(\mathbf{X}^n)) - H(W|\phi_n(\mathbf{X}^n)) + H(W|Y^n, \phi_n(\mathbf{X}^n)), \tag{6.95}$$

where the underlying probability distribution is induced by the hypothesis H_0. Since (ϕ_n) belong to $\mathcal{B}(R, R_c)$, we obtain that

$$H(W|Y^n, \phi_n(\mathbf{X}^n)) \leq n\epsilon_n, \tag{6.96}$$

due to Fano's inequality where $\epsilon_n \to 0$ as $n \to \infty$. Therefore, we obtain

$$-(1 - \alpha_n) \log \beta_n - h_2(\alpha_n) \leq nH(Y) - H(Y^n|W, J_W) - \log M + n\epsilon_n. \tag{6.97}$$

The rest follows from the proof of Proposition 6.2.

6.B Proof of Lemma 6.2

Define the set

$$\mathcal{A} = \big\{(\phi_{2n}(y^n), \phi_{1n}(x^n(w)), w) \mid \iota(\phi_{2n}(y^n); \phi_{1n}(x^n(w))) > \log M - n\gamma\big\}. \quad (6.98)$$

Note that the first term on the right-hand side in the statement of Lemma 6.2 is given by

$$\Pr\{(\phi_{2n}(Y^n), \phi_{1n}(X^n(W)), W) \in \mathcal{A}^c\}, \quad (6.99)$$

cf. the joint distribution in (6.74). Additionally, as $\hat{W} = \psi_n(\phi_{2n}(Y^n), \phi_{1n}(\mathbf{X}^n))$ we have

$$\begin{aligned}\Pr\{\hat{W} = W\} \le{}& \Pr\{\hat{W} = W, (\phi_{2n}(Y^n), \phi_{1n}(X^n(W)), W) \in \mathcal{A}^c\} \\ &+ \Pr\{(\phi_{2n}(Y^n), \phi_{1n}(X^n(W)), W) \in \mathcal{A}\}. \end{aligned} \quad (6.100)$$

We can bound the first term in the above inequality as

$$\begin{aligned}&\Pr\{\hat{W} = W, (\phi_{2n}(Y^n), \phi_{1n}(X^n(W)), W) \in \mathcal{A}^c\} \\ &= \sum_{w=1}^{M} \sum_{\substack{(\phi_{2n}(y^n), \phi_{1n}(\mathbf{x}^n)):\\ \psi_n((\phi_{2n}(y^n), \phi_{1n}(\mathbf{x}^n)) = w \\ (\phi_{2n}(y^n), \phi_{1n}(x^n(w)), w) \in \mathcal{A}^c}} P_{\phi_{2n}(\bar{Y}^n)|\phi_{1n}(\bar{X}^n)}(\phi_{2n}(y^n)|\phi_{1n}(x^n(w))) \\ &\qquad \times P_{\phi_{1n}(\mathbf{X}^n)}(\phi_{1n}(\mathbf{x}^n)) \frac{1}{M} \\ &\overset{(\star)}{\le} \sum_{w=1}^{M} \frac{M e^{-n\gamma}}{M} \sum_{\substack{(\phi_{2n}(y^n), \phi_{1n}(\mathbf{x}^n)):\\ \psi_n((\phi_{2n}(y^n), \phi_{1n}(\mathbf{x}^n)) = w}} P_{\phi_{2n}(\bar{Y}^n)}(\phi_{2n}(y^n)) P_{\phi_{1n}(\mathbf{X}^n)}(\phi_{1n}(\mathbf{x}^n)) \\ &= e^{-n\gamma} \sum_{\phi_{1n}(\mathbf{x}^n)} P_{\phi_{1n}(\mathbf{X}^n)}(\phi_{1n}(\mathbf{x}^n)) \\ &\qquad \times P_{\phi_{2n}(\bar{Y}^n)}\Big[\bigcup_w \big\{\phi_{2n}(y^n) \colon \psi_n(\phi_{2n}(y^n), \phi_{1n}(\mathbf{x}^n)) = w\big\}\Big] \\ &\le e^{-n\gamma}, \end{aligned} \quad (6.101)$$

where $(\star)$ is valid due to the definition of $\mathcal{A}$. The conclusion of the lemma follows.

7

Conclusion

In this thesis we investigated several aspects of identification systems: reducing processing complexity with a hierarchical architecture; relaxing the assumption on the joint distribution; and examining whether the observation is related to the system or not. We studied and provided fundamental characterizations on the performances in each case. Our study filled some gaps between the initial models and practice. Still, some additional work could be done to refine the theory.

The results presented in this thesis are mostly in terms of mutual information, entropy of auxiliary random variables. Computing these results for practical purposes are generally difficult since we still face the challenge to solve the numerical (non-convex) optimization problem. Therefore we need some approximation procedures, algorithms for such quantities.

The relation between the single-user HT problem and the ID problem is quite interesting. To go further we could try to establish the (preferably closed) distributed hypothesis testing regions when compression of both x^n and y^n is considered. The task is very challenging for the discrete case as it has been open for a long time. We however think that something could be done for the Gaussian case since the problem can be perhaps related to the Gaussian distributed source coding problem.

Finally, we can change some more assumptions to make the models more practical. Take for a simple example that our sequence $x^n(i)$ are images of *small* size 100×100 pixels. The block length is then 10^4. Even with a *very small* identification rate of 0.01 bits the number of supported users would be $\approx 10^{30}$, a huge number. This implies that in reality we possibly operate far below the identification capacity given a compression rate. Furthermore, discussing search complexity in a system with such huge number of users, is also problematic. An interesting question therefore could be: Given a number of users M, a block length n and a joint distribution P_{XY} what is the optimal error probability? Herein M is not necessary an exponential function of n.

"I gotta stop somewhere. I'll leave you something to imagine."
Richard Feynman

Bibliography

[PBJ00] S. Pankanti, R. M. Bolle, and A. Jain. "Biometrics: The Future of Identification". In: *Computer* 33.2 (2000), pp. 46–49.

[WKL03] F. M. J. Willems, T. Kalker, and J.-P. Linnartz. "On the capacity of a biometrical identification system". In: *2003 IEEE International Symposium on Information Theory*. IEEE. 2003, p. 82.

[Tun09] E. Tuncel. "Capacity/storage tradeoff in high-dimensional identification systems". In: *IEEE Transactions on Information Theory* 55.5 (2009), pp. 2097–2106.

[TG14] E. Tuncel and D. Gündüz. "Identification and lossy reconstruction in noisy databases". In: *IEEE Transactions on Information Theory* 60.2 (2014), pp. 822–831.

[WO08] M. B. Westover and J. A. O'Sullivan. "Achievable rates for pattern recognition". In: *IEEE Transactions on Information Theory* 54.1 (2008), pp. 299–320.

[Wil09] F. M. J. Willems. "Searching methods for biometric identification systems: Fundamental limits". In: *2009 IEEE International Symposium on Information Theory*. IEEE. 2009, pp. 2241–2245.

[Tun12] E. Tuncel. "Recognition capacity versus search speed in noisy databases". In: *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE. 2012, pp. 2566–2570.

[FW16] F. Farhadzadeh and F. M. J. Willems. "Identification Rate, Search and Memory Complexity Tradeoff: Fundamental Limits". In: *IEEE Transactions on Information Theory* 62.11 (2016), pp. 6173–6188.

[VOS17] M. T. Vu, T. J. Oechtering, and M. Skoglund. "Hierarchical identification with pre-processing". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2017, pp. 2746–2750.

[VOS18a] M. T. Vu, T. J. Oechtering, and M. Skoglund. "Gaussian Hierarchical Identification with Pre-processing". In: *2018 Data Compression Conference*. IEEE. 2018, pp. 277–286.

[VOSB18] M. T. Vu et al. "Uncertainty in identification systems". In: *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2018, pp. 2386–2390.

[VOS18b] M. T. Vu, T. J. Oechtering, and M. Skoglund. "Testing in Identification Systems". In: *2018 IEEE Information Theory Workshop (ITW)*. IEEE. 2018, pp. 1–5.

[VOS19] M. T. Vu, T. J. Oechtering, and M. Skoglund. "Operational Equivalence of Distributed Hypothesis Testing and Identification Systems". In: *Accepted at ISIT*. 2019.

[Gal68] R. G. Gallager. *Information theory and reliable communication*. Wiley, 1968.

[EK11] A. El Gamal and Y.-H. Kim. *Network information theory*. Cambridge university press, 2011.

[Hay06] M. Hayashi. "General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel". In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1562–1575.

[Cuf13] P. Cuff. "Distributed channel synthesis". In: *IEEE Transactions on Information Theory* 59.11 (2013), pp. 7071–7096.

[Cuf16] P. Cuff. "Soft covering with high probability". In: *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2016, pp. 2963–2967.

[AK75] R. Ahlswede and J. Körner. "Source coding with side information and a converse for degraded broadcast channels". In: *IEEE Transactions on Information Theory* 21.6 (1975), pp. 629–637.

[Wyn75] A. D. Wyner. "On source coding with side information at the decoder". In: *IEEE Transactions on Information Theory* 21.3 (1975), pp. 294–300.

[SW73] D. Slepian and J. Wolf. "Noiseless coding of correlated information sources". In: *IEEE Transactions on information Theory* 19.4 (1973), pp. 471–480.

[Cov75] T. Cover. "A proof of the data compression theorem of Slepian and Wolf for ergodic sources (Corresp.)" In: *IEEE Transactions on Information Theory* 21.2 (1975), pp. 226–228.

[CT12] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[AGK76] R. Ahlswede, P. Gács, and J. Körner. "Bounds on conditional probabilities with applications in multi-user communication". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 34.2 (1976), pp. 157–177.

[Han03] T. S. Han. *Information-Spectrum Methods in Information Theory*. Springer-Verlag Berlin Heidelberg, 2003.

[AC86] R. Ahlswede and I. Csiszár. "Hypothesis testing with communication constraints". In: *IEEE Transactions on Information Theory* 32.4 (1986), pp. 533–542.

[WW75] H. Witsenhausen and A. Wyner. "A conditional entropy bound for a pair of discrete random variables". In: *IEEE Transactions on Information Theory* 21.5 (1975), pp. 493–501.

[TC08] C. Tian and J. Chen. "Successive refinement for hypothesis testing and lossless one-helper problem". In: *IEEE Transactions on Information Theory* 54.10 (2008), pp. 4666–4681.

[Dur10] R. Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

[Kal06] O. Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

[Tao15] T. Tao. *275A, Notes 2: Product measures and independence*. `https://terrytao.wordpress.com/2015/10/12/275a-notes-2-product-measures-and-independence/`. 2015.

[Gra13] R. Gray. *Entropy and Information theory*. 2013. URL: `https://ee.stanford.edu/~gray/it.pdf`.

[PW17] Y. Polyanskiy and Y. Wu. "Lecture notes on information theory". In: *MIT (6.441), UIUC (ECE 563)* (2017).

[Goo] Google. *Google Lens*. URL: `https://lens.google.com/`.

[Ama] Amazon. *Amazon Flow*. URL: `https://www.amazon.com/A9-Innovations-LLC-Powered-Amazon/dp/B008G318PE`.

[TKR04] E. Tuncel, P. Koulgi, and K. Rose. "Rate-distortion approach to databases: Storage and content-based retrieval". In: *IEEE Transactions on Information Theory* 50.6 (2004), pp. 953–967.

[BB11] M. Bloch and J. Barros. *Physical-layer security: from information theory to security engineering*. Cambridge University Press, 2011.

[Ooh98] Y. Oohama. "The rate-distortion function for the quadratic Gaussian CEO problem". In: *IEEE Transactions on Information Theory* 44.3 (1998), pp. 1057–1070.

[Mit15] P. Mitran. "On a Markov Lemma and Typical Sequences for Polish Alphabets". In: *IEEE Transactions on Information Theory* 61.10 (2015), pp. 5342–5356.

[WZ76] A. D. Wyner and J. Ziv. "The rate-distortion function for source coding with side information at the decoder". In: *IEEE Transactions on Information Theory* 22.1 (1976), pp. 1–10.

[Wyn78] A. D. Wyner. "The rate-distortion function for source coding with side information at the decoder-II: General sources". In: *Information and control* 38.1 (1978), pp. 60–80.

[MLK15] P. Minero, S. H. Lim, and Y.-H. Kim. "A unified approach to hybrid coding". In: *IEEE Transactions on Information Theory* 61.4 (2015), pp. 1509–1523.

[HB85] C. Heegard and T. Berger. "Rate distortion when side information may be absent". In: *IEEE Transactions on Information Theory* 31.6 (1985), pp. 727–734.

[BBT59] D. Blackwell, L. Breiman, and A. Thomasian. "The capacity of a class of channels". In: *The Annals of Mathematical Statistics* (1959), pp. 1229–1241.

[BBT60] D. Blackwell, L. Breiman, and A. Thomasian. "The capacities of certain channel classes under random coding". In: *The Annals of Mathematical Statistics* 31.3 (1960), pp. 558–567.

[Ahl78] R. Ahlswede. "Elimination of correlation in random codes for arbitrarily varying channels". In: *Probability Theory and Related Fields* 44.2 (1978), pp. 159–175.

[ZVO19] L. Zhou, M. T. Vu, and T. J. Oechtering. "Polar Codes for Identification Systems". In: *12th International ITG Conference on Systems, Communications and Coding*. 2019.

[ZTM17] L. Zhou, V. Y. F. Tan, and M. Motani. "Exponential Strong Converse for Content Identification with Lossy Recovery". In: *ArXiv e-prints* (Feb. 2017). URL: `https://arxiv.org/abs/1702.06649`.

[YY16] V. Yachongka and H. Yagi. "Reliability function and strong converse of biomedical identification systems". In: *2016 International Symposium on Information Theory and Its Applications (ISITA)*. IEEE. 2016, pp. 547–551.

[CK11] I. Csiszar and J. Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[Ver12] S. Verdú. "Non-asymptotic achievability bounds in multiuser information theory". In: *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE. 2012, pp. 1–8.

[Liu18] J. Liu. "Information Theory from a Functional Viewpoint". PhD thesis. Princeton University, 2018.

[LHV17] J. Liu, R. van Handel, and S. Verdú. "Beyond the blowing-up lemma: Sharp converses via reverse hypercontractivity". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2017, pp. 943–947.

[Vol+10] S. Voloshynovskiy et al. "Information-theoretical analysis of private content identification". In: *IEEE Information Theory Workshop (ITW)*. IEEE. 2010, pp. 1–5.

[Sch02] N. A. Schmid. "Large deviations performance analysis for biometrics recognition". In: *40th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2002.

[Mou10] P. Moulin. "Statistical modeling and analysis of content identification". In: *Information Theory and Applications Workshop (ITA)*. IEEE. 2010, pp. 1–5.

[DD11] G. Dasarathy and S. C. Draper. "On reliability of content identification from databases based on noisy queries". In: *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE. 2011, pp. 1066–1070.

[Mer17] N. Merhav. "Reliability of universal decoding based on vector-quantized codewords". In: *IEEE Transactions on Information Theory* 63.5 (2017), pp. 2696–2709.

[Han87] T. Han. "Hypothesis testing with multiterminal data compression". In: *IEEE Transactions on Information Theory* 33.6 (1987), pp. 759–772.

[LCV17] J. Liu, P. Cuff, and S. Verdú. "E_γ-Resolvability". In: *IEEE Transactions on Information Theory* 63.5 (2017), pp. 2629–2658.

[PPM16] G. Pichler, P. Piantanida, and G. Matz. "Distributed Information-Theoretic Clustering". In: *arXiv preprint arXiv:1602.04605* (2016).