UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

FINITE-LENGTH PERFORMANCE ANALYSIS AND DATA-DRIVEN DESIGN FOR HYPOTHESIS TESTING: AN INFORMATION THEORETIC PERSPECTIVE

TESIS PARA OPTAR AL GRADO DE DOCTOR EN INGENIERÍA ELÉCTRICA

SEBASTIÁN ANDRÉS ESPINOSA TRUJILLO

PROFESOR GUÍA:
JORGE SILVA SÁNCHEZ

MIEMBROS DE LA COMISIÓN:
MARCOS ORCHARD CONCHA
RONNY VALLEJOS ARRIAGADA
JUAN YUZ EISSMANN

SANTIAGO DE CHILE
2023

Resumen

RESUMEN DE TESIS PARA OPTAR AL
GRADO DE DOCTOR
EN INGENIERÍA ELÉCTRICA
POR: SEBASTIÁN ANDRÉS ESPINOSA TRUJILLO
FECHA: 2023
PROFESOR GUÍA: JORGE SILVA SÁNCHEZ

ANÁLISIS DE DESEMPEÑO NO ASINTÓTICO Y DISEÑO BASADO EN DATOS EN TEST DE HIPÓTESIS: UNA MIRADA DESDE LA TEORÍA DE LA INFORMACIÓN

Inspirado en el aporte de Claude E. Shannon en comunicaciones, esta tesis aborda tres problemas relevantes de toma de decisiones que tratan con incertidumbre, restricciones de comunicación y un número finito de observaciones en test de hipótesis binarios (HT).

En la primera parte de este trabajo, se entregan nuevas cotas de límites superiores e inferiores de desempeño para el test óptimo (Neyman-Pearson) en el problema clásico de HT binario. Nuestras cotas de régimen finito ofrecen la capacidad de cuantificar la relación entre el tamaño de la muestra y las probabilidades de error.

En la segunda parte de este trabajo, derivamos un nuevo límite de desempeño teórico para un HT bivariado distribuido. Derivamos una expresión analítica para el exponente de error del Tipo II dado una restricción de error de Tipo I y una restricción de tasa. También medimos la discrepancia entre nuestras expresiones prácticas de desempeño con un número finito de muestras y sus límites asintóticos.

Finalmente, estudiamos la colaboración en la inferencia distribuida en el caso del test de independencia con restricción de comunicación. Analizamos si la colaboración ofrece una ventaja en cuanto al compromiso óptimo entre los errores de Tipo I y Tipo II.

Summary

THESIS SUMMARY TO OBTAIN THE
DEGREE OF DOCTOR OF PHILOSOPHY
IN ELECTRICAL ENGINEERING
BY: SEBASTIÁN ANDRÉS ESPINOSA TRUJILLO
DATE: 2023
ADVISOR: JORGE SILVA SÁNCHEZ

FINITE-LENGTH PERFORMANCE ANALYSIS AND DATA-DRIVEN DESIGN FOR HYPOTHESIS TESTING: AN INFORMATION THEORETIC PERSPECTIVE

Inspired by the work of Claude E. Shannon in communication, this Thesis tackle three relevant decision-making problems that deal with uncertainty, communication constraints, and a finite number of observations (finite-length analysis) in binary hypothesis testing (HT).

In the first part of this work, we offer new upper and lower performance bounds for the optimal (Neyman-Pearson) test in the classical binary HT problem. Our finite-length bounds offer the ability to quantify the relationship between sample size and error probabilities.

In the second part of this work, we derive a new information-theoretic performance limit (error exponent) for a distributed bivariate HT. In distributed HT, observations are transmitted to the decision agent with a rate constraint (in bits per sample). We derive analytical expressions for the error exponent of the Type II error given a Type I error restriction and a rate constraint. We also extend the non-asymptotic finite-length performance bounds from the classical centralized setting to this distributed one.

Finally, we study collaboration in distributed inference for testing independence with a fixed-rate communication constraint. We analyze the benefits of collaboration and evaluate in theory and practice if collaboration offers an advantage regarding the optimal trade-off between Type I and Type II errors.

Esta tesis se la dedico a mi familia y toda persona que fue parte de este camino.

Bonds are stars that light up the entire universe.
May your heart be your guiding key.

Acknowledgements

Finalmente he llegado a la mejor parte de esta Tesis. Escribir los agradecimentos es un momento muy emocionante. Sin duda que el doctorado ha sido uno de los procesos más desafiantes y complejos que me ha tocado vivir como estudiante. Es muy curioso que siendo que estas páginas vayan al comienzo, en realidad son las últimas que decidí escribir. La razón principal es que tuve que esperar a madurar todo lo vivido para entender el valor del esfuerzo propio como del apoyo de los demás.

Los siguientes párrafos están dedicados a las personas que de una u otra forma fueron una influencia y un apoyo importante para poder llevar a cabo este trabajo y proceso. Lo más justo es mencionar a cada uno de ustedes por orden de aparición (tienen un $*$ para que se busquen rápidamente), posterior a los agradecimientos personales se encontrarán con unas palabras de cierre (lo notarán porque habrá una doble línea como la que hay más abajo). De antemano me disculpo si alguno cree que merece más palabras, pero créanme que cada frase estará escrita con el mayor cariño que me representa. Partamos!

$*$ Primero agradezco a mi familia. Papá, Mamá y mi hermano, son lo mejor que hay, así de simple. A ustedes ya les había agradecido por su apoyo desde que tengo uso de razón, gracias por permitirme tomar este camino no tan convencional pero a la vez tan llenador. Si estoy donde estoy, es porque ustedes confiaron en mi y me entregaron un apoyo incondicional. Ustedes siempre han estado para mi y, sinceramente, aquí las palabras sobran.

$*$ Ballero y Manuek, a ustedes dos les agradezco por tantos años de amistad. Los mejores momentos de risa y diversión son siempre al verlos. Aunque los tiempos a veces no nos acompañen para poder reunirnos, el hecho que nos podamos dar unas horas para poder jugar o simplemente reir es muy llenador.

Ballero, gracias por soportarme tantos fines de semana, sobretodo varios que fueron muy críticos, siempre dispuesto a recibirme aunque no tenga mucho que decir. Y siempre con tus comentarios que a veces me dejan sin respiración de tanto que me río.

Manuek, tu voluntad por mostrar disposición y cariño es algo que todo el mundo debería valorar, me alegro mucho y me emociono por la familia que formaste. Si bien eso hace dificil verte, se te nota feliz y ya con eso yo también estaré contento.

Tengo que también mencionar las partidas "colaborativas" de lfd2 y esperar al momento final donde no existe equipo, al contrario hay pura traición y además que se salve quien pueda (GGNTTY). Nada le gana a la buena jeringa y dejarlos botados por mancos. Un momento de diversión con ustedes ya me hacen el día. Gracias parcito de ..., los quiero. No olvidemos nuestro saludo AIAIAIAIAIAIAIAIAI.

* Mario, Franco y Nano, el team DIE-DCC. Me alegro harto que hayamos podido reencontrarnos aunque sea una vez al año. Las anécdotas que tenemos para contarnos son siempre una instancia para volver a reir y pasarla muy bien. Podría estar recordando todo el pregrado con ustedes, pero ya les agradecí oportunamente en mi magister, ahora hay que dejarle espacio al resto de privilegiados. Igual me alegra que nos ríamos siempre de lo mismo, incluso si, lamentablemente, soy el protagonista del chisme.

* Selmi y Caro, muchas gracias por el reencuentro, con ustedes se pasa muy bien. Me gusta mucho ver la evolución desde que las conocí como alumnas y ahora ya son todas unas profesionales. Lo mejor es que podemos divertirnos en temas bien afines (bueno la Caro malvada a veces arruga pero se le perdona), sólo espero que nos podamos seguir reuniendo y pasarla bien, esos panoramas con un toque infantil pero divertido lo agradezco mucho. Recuerdo con mucho cariño las anécdotas que hemos vivido, quisiera en particular destacar los globos y cómo se les cayó la cara cuando supe que eran ustedes (tengo buenos dotes detectivescos).

* Mario (dos), gracias por ayudarme con toda la parte algorítmica, sin duda que formamos un gran equipo de investigación y más aún con las publicaciones logradas. Lo emocionante es que también fuiste mi alumno por allá hace varios años atrás.

* Profesor Jorge, gracias por tu constante apoyo, agradezco todas las herramientas que me entregó para potenciarme en el área que más me gusta. Nunca olvidaré que el secreto para un buen entendimiento está en los detalles. Es algo que hasta el día de hoy aplico en mi vida diaria. Eso marca la diferencia y si he estado tantos años es porque era necesario entregar lo mejor de mi y eso es lo que siempre, de alguna u otra forma, trató de mostrarme.

* Don Rodri, te consideré un mentor para mi, tus consejos sobre cómo ser un buen investigador los sigo guardando y cada recomendación tuya es una fuente de conocimiento y sabiduría dificil de describir, gracias por guiarme en mis inicios como investigador. Hasta el día de hoy tus recomendaciones son claves para mi y los agradezco profundamente.

* Felipito, gracias por tener esa disposición que siempre te caracteriza, eres una persona muy admirable y querible. Me da harta pena que nuestros tiempos sean tan incompatibles, pero sabes, te llevo en mis recuerdos como esa persona apoyadora e inteligente que daba la sabiduría al laboratorio. Sólo te diré que te portes bien! Coqueto!

* Ceci, Toño y Vale, también les quiero agradecer su apoyo. Encontrarme con ustedes y ver en qué andaban era muy entretenido para mi, ahora ya están prontos a egresar (edit: egresados) y sólo quiero darles las gracias por esos momentos de copuchas y las conversaciones de pasillo que más de alguna vez me sacaron una sonrisa.

∗ Jo, te conocí ese día que entraste a la U por allá el 2017, desde ese entonces vi ese entusiasmo y esfuerzo tuyo por ser la mejor estudiante. Me alegra saber que hasta el día de hoy pueda seguir juntándome contigo. Aún recuerdo con entusiasmo las veces que ibas a realizarme tus preguntas DIM y trataba de darte la mayor orientación posible. Desde que me contaste que tu interés era la ingeniería eléctrica, supe que de alguna manera iba a seguir sabiendo de ti. Lo que quizás no me esperaba era que nos seguiríamos apoyando y formar una bonita amistad hasta el día de hoy, algo de lo cual estoy demasiado contento.

Agradezco de corazón que puedas entregarme parte de tu tiempo para poder conversar de todo, eres de las pocas personas que me da la confianza para hablarte mis cosas, gustos ñoños, que no los encuentres raros y tú los apoyes. Por culpa de eso pareciese que vomito información inservible, pero aprecio mucho que te des el tiempo de escuchar y reírte de todo lo que sale de mi boca. Podría estar recitándote todo el día los diálogos de Drake & Josh porque son muy buenos, pero el tiempo es limitado.

Por otra parte, me ayudaste a entender mi propio proceso de investigación y esa misma experiencia espero habértela transmitido. Ahora con lo que hemos vivido, tenemos claro la importancia de la superación, pero también vemos el poder del apoyo que te entregan tus seres queridos. No olvides que siempre puedes contar conmigo, con todo lo vivido ya podemos hacer los mejores trabajos de investigación y responder como se debe a todas las inquietudes de los supervisores (Ángela aprueba esto).

Me gustan mucho y aprecio las conversaciones tan abiertas, ese lado un tanto extrovertido tuyo, sumado al hecho que comprendemos el perfil del beauchefiano promedio (y lo pelamos), permite juntas muy fluidas y entretenidas. Me llama la atención que puedas hacer tantas cosas y reunirte con varios amigos tuyos, podrías compartirme tu secreto. Y también quisiera agradecerte que hables bien de mi tan abiertamente con tu familia, así como también que te acuerdes de vez en cuando de mi como un apoyo que he sido durante estos años ya que esto último me ha hecho sentir más importante.

Quisiera agradecerte todos esos momentos en que me has mostrado tu disposición para poder pasarla bien y divertirse, esto es algo que me llena de mucha alegría y aprecio que puedas entregarme ese tiempo cuando estás disponible. En especial me emociona ver todo tu cariño y bonitos gestos hacia mi durante estos años. Nunca había visto tanta dedicación en una celebración. Junto con el Freire la pasamos muy bien ese día y se debe repetir. Probablemente no había visto es faceta competitiva tuya pero la verdad es muy graciosa. Estuve muy contento esa vez, ya viste que el overcooked es un súper buen juego!

Finalmente te quiero decir que así como sientes que he formado parte de tu red de apoyo, quizás con el sólo hecho de escucharte, tú también formas parte de mi red. A veces son los pequeños gestos los que marcan las diferencia y esto lo sabemos muy bien. Tener ese cariño tuyo así como poder compartir momentos para saber de ti, hizo que este proceso se haya podido llevar a cabo de manera más amena. Serás la mejor eléctrica, de eso estoy seguro. Te quiero mucho Jo!

∗ Mauricio, tu llegada al laboratorio fue sin duda un cambio de aire para el laboratorio,

aunque ya te fuiste, no olvidaré cómo me entregaste el título inapelable de "generalísimo", hasta el día de hoy me río de todas las anécdotas que fuimos viviendo con "mis vientos" y el Miguel que no cuidaba las formas. Fueron momentos de mucha risa y se generaba un ambiente muy grato en el laboratorio, gracias por permitirme a vivir estas instancias.

∗ Valeria, o mejor dicho, VALERIA, cuando me acuerdo de ti solamente se me vienen a la cabeza recuerdos muy bonitos. Nunca había visto una estudiante que le diera tanta dedicación y esfuerzo en su paso por la U. Creo que cualquier profesor debería sentirse afortunado de tenerte como alumna, tal como me pasó a mi cuando te hice clases. Ese día que me dijiste que ibas a estar visitándome semanalmente por las dudas, sabía que hablabas en serio y sólo me llevo el mayor orgullo al ver todo lo que has logrado.

Este es por lejos uno de los agradecimientos más fáciles de escribir dado que siempre tengo algo que decirte, asimismo, muchas de las cosas que podría colocar la verdad es que ya las sabes y, muy probablemente, todo lo que leerás aquí alguna vez lo hemos conversado. La diferencia es que en esta oportunidad quedará por escrito, y cómo son tantas cosas lo mejor es que destaque algunas de ellas y será tu tarea completar esos bonitos momentos.

Lo primero que destaco son las preguntas VALERIA, con eso ya te dije todo y no hace falta hablar más de esto. Lo segundo que destaco es la mejor copucha que consiste en la historia de tu subida de puntaje en la corrección, no daré más detalles aquí porque ambos sabemos lo que eso implicaría, pero siempre hay que contarlo al reunirnos, ya que eso significa reirse hasta no poder respirar (además que siempre le agrego un detalle adicional). Lo tercero que destaco (esto es más reciente) son las reuniones con la auxiliar "tras bambalinas", es obvio que toda información relevante debe primero pasar por ti para su posterior aprobación.

Valoro también tu apoyo durante la pandemia y la EDV, sobretodo esos momentos que llamo "Full Synchro", donde se lograba un nivel de entendimiento telepático imposible de replicar con otra persona, era impresionante como no debíamos ni comunicarnos y ambos sabíamos lo que teníamos que hacer y, para rematar, nos anticipábamos a lo que nos íbamos a sugerir.

Ahora bien, lo que definitivamente más te quiero agradecer es que debido a ti pude lograr una mayor comprensión de mi personalidad, algo que siempre había querido saber. Diste en el clavo inmediatamente (como siempre) cuando hablamos sobre nuestro *don* y gracias a eso he llegado a ese punto en que entendemos como somos.

Tengo que mencionar obviamente las reuniones/sesiones contigo; esa tradición de principio de semestre de más de 3 horas para ponernos al día. Allí vemos y analizamos las cosas con un nivel de detalle sorprendente e inigualable lo que sin duda es algo que nos caracteriza y nos potencia. Identificamos las fortalezas (y debilidades) que podemos llegar a tener la gente como nosotros. Siempre vemos todo desde un punto de vista constructivo, con interés en mejorar y de manera muy sincera.

Aunque todo lo anterior son momentos a destacar, la verdad que lo que más me alegra es que podamos seguir teniendo este contacto, todo lo que he ido conociendo sobre ti y tú sobre mi nos ha ayudado a fortalecer esta conexión que tenemos, la cual espero siga siendo

cada día más fuerte. Cada vez que te recibo con la alegría que me caracteriza es una felicidad muy sincera y, sin mentirte, me emociona mucho verte. La razón es porque puedo anticipar que cada junta contigo será lo mejor, el hecho que vayas por tu propia voluntad lo hace aún más significativo. Tu apoyo va mucho más allá de querer escuchar, es más bien de *querer estar*, algo que genuinamente demuestras pero que es muy difícil de percibir en la gente.

Para qué hablar de tu futuro prometedor si eso ya está más que claro, lo mas chistoso es que tenemos que aconsejarnos las mismas cosas sobre confianza pese a que nos cueste de igual manera de colocar en práctica.

De verdad que es emocionante escribirte estas palabras y espero lleguen a ti de la misma manera que nuestros momentos "Full Synchro". Lamento si no me explayé más, pero lo que sí trato de destacar es nuestro *sello*, un sello donde existe entendimiento, respeto y comunicación. Si no estás cerca físicamente no me preocupa ya que si un día quisiera verte, verías tu apretado calendario, luego llegarías lo antes posible y eso ya es muy bonito. Te quiero mucho VALERIA, gracias por todo, siempre en mi corazón.

∗ Mati, siempre es un agrado y me da mucha alegría cada vez que te veo. Esa alma tan DIM tuya y llena de sabiduría me asegura que el departamento tiene al mejor. Gracias por realizar conmigo la colaboración maratónica que nos mandamos. Fue la mejor manera de poder dar cierre a mi proceso como auxiliar. Así como cuando fui auxiliar por primera vez, mi mentor también decidió hacer una clase en conjunto. Sentí que era necesario hacer el cierre de la misma manera. Ver estos ciclos son súper emocionantes y más si se comparten con las personitas que corresponden. Me alegro mucho que sigas el camino de las auxiliares, puedes notar lo llenadoras que son.

Ahora me entiendes cuando te preguntabas cómo era posible que yo siempre estuviese disponible para ayudarte. Agradezco que hayas escuchado mis recomendaciones y veo que te han ayudado mucho, pero también la verdad es que ya sé lo bien que te irá en todo, a estas alturas mi aporte será más bien de cariño y desearte lo mejor como siempre. Si requieres alguna orientación adicional, pues no se diga más, te apoyo. Te quiero crack.

∗ Carlitos, si bien ya no hablamos como antes quiero hacerte una mención honrosa ya que aún recuerdo nuestras partidas de Rocket League del 2020 y te agradezco cada sonrisa que me sacaste porque pucha que me ayudaron a sentirme mejor, te irá super bien en todo!

∗ Clared (o flaca para los confianzudos), de verdad es que estoy muy agradecido contigo. Eres una persona maravillosa, a pesar que nos separen como 2000km y tengamos unas diferencias bien marcadas en nuestros estilos de vida, nunca pensé que podría tener una amiga como tú. La hemos pasado súper bien y gracias por reírte de mis chistes, incluso si me salen fomes, por alguna razón igual te causan gracia.

También hemos tenido todo tipo de reuniones, ya sean emotivas, de copuchas, tragicómicas y, obviamente, de yuyines que nos caracteriza. Lo más destacable tiene que ser la china petrificada, estoy seguro que debe estar aún ahi. Ojalá puedas darte el tiempo de venir a Santiago, por más que no te guste la ciudad, los mejores artículos deportivos están aquí.

Realmente espero que sigas logrando todas esas metas deportísticas que te propones, sigas avanzando en tus proyectos tan importantes para el país y gracias de nuevo por darme de ese tiempo para escucharme en esos momentos de mayor fragilidad. Tu alma periodística no te lo quita nadie. Esa fuerza y coraje que nada te detiene es realmente admirable. A su vez, puedo notar esa pequeña ternura que sale de ti y siento que esta doble faceta tuya ha ayudado a fortalecer nuestra amistad. La clave de esto al final es el entendimiento y el respeto. Siempre me causan risa tus pinzas de cangrejo jaja. Te quiero flaquita!

∗ Thamar, o mejor dicho, la pana, eres la última de esta privilegiada lista. Así como el resto de las personas antes mencionadas me es fácil poder darte las gracias. Tengo que admitir que tu personalidad es bien especial, algunas veces aún me parece indescifrable el cómo te comportas conmigo, pero poco a poco voy sabiendo más de ti y de lo que estoy muy seguro es que siempre me da una alegría muy grande el verte (esto seguramente no lo habías notado).

Te empecé a conocer en un momento un poco complicado que tú sabes, mientras hacía pan. Siempre curiosa por los chismes y además te das el tiempo de escucharme y comprenderme, de la misma manera que lo he hecho contigo cuando tienes ganas de hablar (aunque con suerte me cuentas que "no te quejas").

Lo primero que debo destacar son los grandiosos, únicos e inigualables momentos Team Nabla y cómo logramos sacar adelante la mismísima perfección en no uno sino que en tres cursos (háganse esa). Te agradezco por apoyarme en todas mis ideas raras que iban desde vengarse de pelao hasta posar magistralmente luego de una semana redonda. Me impresionaba que no solamente apoyabas las ideas sino que además las potenciabas con otras -aún no entiendo cómo es que no nos expulsaron-.

A veces pienso qué fue lo que hizo que esta instancia fuese tan bonita, y la verdad pana, es que el curso por si solo no basta para eso, también es importante para mi al menos una sensación de unidad, algo que logramos indudablemente y hace que estos momentos los atesore como ningún otro.

Ahora bien, una de las cosas por las cuales estaré siempre agradecido contigo -y esto es de lo poco que no te he contado- es que me diste ese impulso para volver a sentirme *vivo*. Ese día en donde logramos la hazaña a 17 segundos de perder, fue de las cosas que me dieron las energías y el entusiasmo para seguir adelante pese a las adversidades que vendrían más adelante. Aunque suene puntual, incluso exagerado y para ti pudo ser un día más, en mi caso fue algo que me dejó dichoso ya que fue un tiempo para pasarla bien y divertirse en algo que había querido hacer, por lo que se debe repetir.

Lo que vi en ti e hizo que me encariñara mucho contigo se deduce de todas las ideas anteriores. Al leerlas te das cuenta que comparten algo en común y eso corresponde a ese *apañe* que me muestras de vez en cuando. Valoro tu bonita voluntad que tienes y ojalá la sigas mostrando conmigo ya que me hace sentir privilegiado el tenerte cerca.

Lo que sí debo decir es que hay una sola cosa que extraño y es que antes había una pana más conversadora a mediados del año pasado, debería y te exijo que vuelva si total Dr. Polo ya nos funó. Tú también puedes hablarme para todo, al menos en mi caso soy feliz cuando

converso contigo aunque sean cosas triviales. Tengo apañe que entregarte, no muerdo (sólo hablo como loro en confianza) y no me parece correcto que sólo vaya en una dirección. Recuerda que tienes mi apoyo permanente y sincero sea en el ámbito que sea.

Se nota que eres una persona con un corazón muy puro, muy acorde a tus creencias y definitivamente algo que debería verse más seguido en la gente. No tengo muy claro en qué he sido un aporte para ti salvo en los momentos concretos en que soy tu esclavo intelectual, pero estoy muy seguro que te irá demasiado bien en la U y en la vida.

Finalmente estoy muy agradecido por los momentos en que me has ido a ver (ya sea obligatorio o no), tu visita siempre me alegra harto ya que es sinónimo de que será un bonito día. Me has apoyado en ciertos casos muy concretos, pero a la vez certeros; es curioso porque apareces en un momento muy preciso que hasta llega a ser sorprendente (quizás lo percibes?) y entregas tu apoyo con tu forma bien peculiar de hacerlo. Tu excusa es que justo estabas modo nanai, pero ojalá vaya más allá de eso. Me gusta sentir ese cariño tuyo, es como si estuvieras cerca, dándome una palmadita en mi hombro (o quizás un golpe conociéndote) entregando un apoyo real.

A veces uno puede ser feliz con cosas tan resimples pero a la vez difíciles de conseguir, como grata compañía, en particular, la tuya. Estos pequeños agradecimientos hacia ti pretenden sintetizar un cariño muy grande, lo que se convierte en una tarea difícil mas no imposible. Aunque quede como imaginación, suelo verte como si fuéramos ese dúo imbatible que siempre quise formar pero que nunca pude, de ahí que el término "pana" es uno bien adecuado y simbólico. No me pidas buscar otra pana menos polla, nunca más.

Panita, te tengo una muy buena noticia, ahora sí puedes pasar los límites pa'entro y darme ese anhelado wate, es decir, tu forma particular de mostrar cariño hacia mi. De verdad te quiero mucho, por siempre en mi corazón de bobo.

Para finalizar, quisiera decir una frase motivacional, pero lamentablemente no tengo. Siempre he estado en busca de alguna que me represente, sin embargo, hay una que sí puede resumir muy bien los agradecimientos mencionados anteriormente. Varias veces para sacar la fuerza para seguir adelante no salieron de mis conocimientos, sino que de mis lazos. Creo firmemente que el apoyo mutuo es central para salir adelante.

Las relaciones son como estrellas fugaces, te guían en la oscuridad. Además, a partir de este apoyo derivan recuerdos y al menos en mi caso es relevante recordar estos hitos, ya que incluso si no están presentes lo que permanece es lo que queda en mi memoria. Por lo tanto, *"Even if we're apart, we're not alone anymore"*.

Bonita frase cierto? Es de un juego, no diré cual es :). Mentira, es de "Kingdom Hearts", juegazo, me tiene viciao como el color azul y el rojo, se los recomiendo 10/10. Y la analogía con las estrellas fugaces definitivamente no diré de dónde lo saqué, eso ya es muy personal.

Table of Content

List of Tables

List of Figures

Chapter 1

Introduction

Information theory was iniciated by Claude E. Shannon in 1948, in a landmark paper titled "A Mathematical Theory of Communication" [1]. Shannon, widely regarded as the father of information theory, made groundbreaking contributions that revolutionized the field of communication. Shannon's significant contribution to information theory laid the foundation for modern communication systems and significantly impacted many other areas.

In a nutshell, one of Shannon's key contributions was the development of the concept of entropy [2]. Entropy measures the uncertainty or randomness in a random variable or information's source. Shannon showed that entropy provides a fundamental limit on the achievable compression and transmission of information. He showed that any source of information can be encoded with arbitrary accuracy by exploiting the source's statistical properties and achieving compression rates close to its entropy [2]. Shannon's entropy concept also plays a crucial role in error control coding and channel capacity [2]. The basic goal of communication is to send a message over a noisy channel, and then to reconstruct it with low probability of error, in spite of the channel noise. The Shannon capacity theorem states that for a given communication channel with a certain level of noise, there exists a maximum data rate at which information can be reliably transmitted. This capacity is determined by the channel's noise characteristics and can be approached but not exceeded with error control coding techniques. Shannon's work highlighted the importance of channel capacity as a fundamental limit and provided guidelines for designing efficient error correction codes [3]. Overall by quantifying the amount of information in a probabilistic setting, Shannon provided a theoretical framework for understanding the limits and efficiencies of communication systems [4].

The importance of Shannon's contributions extends beyond the field of communication engineering. His work on information theory has had a profound impact on diverse disciplines, including computer science, cryptography, statistics, and data compression [5–14].

By establishing fundamental limits and introducing mathematical formalisms to quantify and manipulate information, Shannon's insights have shaped our understanding of communication problems and provided a framework for optimal system design and analysis [2]. In this context, information theory applied to statistics provides a probabilistic formalization

of this problem and, more importantly, fundamental performance limits for general decision problems [2, 4, 5]. This view inspires the work presented in this thesis where statistical tools are applied in three relevant problems in the area of HT.

1.1 Hypothesis Testing

In a nutshell, hypothesis testing (HT) is a statistical method used to make inferences from observations. The goal of HT is to determine whether a particular hypothesis about a random's object distribution is supported by the empirical evidence (data) or whether it should be rejected. The importance of HT lies in its ability to provide a framework for making objective decisions about data. For example, in the case of digital communications HT is used to distinguish between the presence or absence of a signal in noisy channel conditions [5]. In this context, the null hypothesis assumes that there is only noise, while the alternative hypothesis indicates the presence of a signal. Various detection techniques, such as energy-based detection or matched filtering [5], are employed to make decisions based on the received signal's characteristics. This allows decision makers to draw optimal decisions about the observed data with a known level of performance or accuracy.

HT is particularly useful for event detection in sensor networks [15–17]. Data correlation often occurs among observations of distributed devices in the presence of a relevant signal of interest [18–21].In particular, researchers within the field of statistical signal processing have been involved in a wide range of research initiatives studying decision and inference in the presence of measurement noise or corruptions driven by various types of perturbations [22]. In real-world applications, these sources of degradation come from factors such as noise at the sensors, communication restrictions between sensors and decision agents, or by the presence of external sources of perturbations [23].

Formally speaking, one of the key aspects of HT is selecting over a space of hypotheses a probability distribution that best fits the observations. This model selection task is essential and it is used to assess the evidence with respect to a collection of candidates (hypotheses).

1.1.1 Binary Hypothesis Testing

In this context, binary HT involves formulating a null hypothesis (H_0) and an alternative hypothesis (H_1). The null hypothesis represents the nominal (or normal) condition being tested, and the alternative hypothesis models the deviation scenario. The null hypothesis is assumed to be true unless there is sufficient evidence to reject it [24]. The null and alternative hypotheses are often specified in terms of a probability distribution. For example, if the data is assumed to be normally distributed, then the null hypothesis might be that the mean of the distribution is equal to a specific value [24]. Alternatively, if the data is assumed to be binomially distributed, then the null hypothesis might be that the probability of success is equal to a specific value [24].

Once the null hypothesis has been specified, the next step is to calculate a test statistic based on the observed data. The test statistic is a measurable function of the data that measures how far the evidence deviates from what is expected under the null hypothesis (H_0). The distribution of the test statistic, which is a random variable, is then compared to

a probability distribution that is consistent with H_0. Importantly, when the two distributions are known, the celebrated *Neyman-Pearson lemma* provides the optimal decision scheme for this testing task [25].

Selecting the appropriate probability distributions is a critical step in the mathemtical formulation of HT. The choice of the distribution depends on the type of data being analyzed and the assumptions being made about the problem. By selecting the correct distributions, researchers can calculate test statistics (scores) and evaluate the evidence for or against a specific scenario. For example, if the data is continuous and normally distributed, then the test statistic might follow a t-distribution. If the data is discrete and follows a binomial distribution, then the test statistic might follow a chi-squared distribution or a normal distribution [26].

In many applications of binary HT, the goal is to identify from measurements whether a particular signal or event is present or absent. The challenge is that there is always uncertainty in the measurements, and, consequently, there is the posibility of making errors. Deciding the null or the alternative hypotheses determines the probability of making errors. The two types of errors that can occur are the TYPE I and TYPE II errors. The TYPE I error occurs when the null hypothesis is rejected even though it is true. This is also known as the false positive [26]. The TYPE II error occurs, on the other hand, when the null hypothesis is not rejected even though it is false. This is also known as the false negative. Then, probability of errors provides a way to quantify the complexity of the task and helps designers make the best informed decisions about the cost associated with their conclusions.

1.2 Objectives

1.2.1 General Objectives of this Thesis

The general objective of this Thesis is to determine the optimal tradeoff between the Type I error and Type II error as a function of the number of observations. We study the finite sample-size regime in three interesting HT regimes. More specifically, in the topic of unconstrained HT, in the area of distributed HT, and in the new area of collaborative detection.

1.2.2 Specific Objectives of this Thesis

The specific objectives of this thesis are the following:

1. We study the optimal tradeoff between the Type I error and Type II error using the exponential rate of convergence of the optimal miss error probability — as the sample size tends to infinity — given some (positive) restrictions on the false alarm probabilities.
2. We study the gap between the minimum TYPE II error and its exponential approximation under different setups, including restrictions imposed on the vanishing TYPE I error probability under communication (information bits) constraints.
3. We analyze collaboration in distributed inference for testing independence with a fixed-rate communication constraint. We look at collaboration as a strategy to improve performance between the Type I error and Type II error.

1.3 General Contribution of this Thesis

This Thesis advances state-of-the-art in the context of statistical HT in three different areas. The main technical goal in this thesis is to develop asymptotic information limits and non asymptotic performance bounds by extending results on the literature of distributed HT [27], [28] and [29]. We have made significant contributions for the understanding and application of HT in three concrete areas: in the topic of finite length hypothesis testing (deriving performance bounds for the probabilities of error), in the area of distributed HT, and in the new area of collaborative detection.

1.3.1 Finite Length Result for Hypothesis Testing

Concerning the characterization of fundamental performance bounds in HT, a common approach is to determine the exponential rate of decay of the TYPE II error for a prescribed TYPE I error constraint when the number of observations tends to infinity [27]. In binary HT, this optimal error exponent is expressed by an information-theoretic quantity and given by the celebrated Stein's Lemma [30].

While many theoretical results in HT assume an infinite number of observations (asymptotic analysis) [4], a finite-length observation analysis acknowledges the finite nature of available data and provides valuable insights into the performance of practical testing schemes. One key importance of finite length analysis is its ability to quantify the relationship between sample size and error probabilities. In practical situations, it is often not feasible or efficient to collect an infinite number of observations. Then, finite length analysis allows designers to determine the minimum sample size required to achieve a desired level of statistical discrimination or to control error probabilities. Finite-length results could offer guidance on the optimal allocation of resources and enables informed decision-making in HT. This finite-length focus is particularly important in fields such as medicine, finance, and engineering, where decisions based on HT have real-world implications. For example, in genetics research or astronomical observation, the available data is often limited due to high experimental costs or rarity of certain genetic variants. Therefore, finite-length analysis could help evaluate the significance of observed associations in genetic markers and diseases [31, 32].

Adressing this challenge, in the first part of this thesis we obtain new upper and lower bounds to the optimal TYPE II error probability for the important case when we have finite observations. To illustrate the use of these new bounds, the derived expressions are evaluated and interpreted numerically for some vanishing TYPE I error restrictions and as a function of the number samples. The specific contributions are presented in Section 1.4.1.

1.3.2 Distributed Hypothesis Testing

It is commonly assumed that the observations to make decision (samples) are available for decision-making with no perturbation. However, in some practical settings, the data cannot be received directly due to some geographical or communication limitations. Then, extending the fundamental results of HT from centralized to decentralized scenarios is a very important problem. On this, Ahlswede and Csiszar [28] characterized the asymptotic behaviour of the error exponent with communication constraint modeling two agents located in different

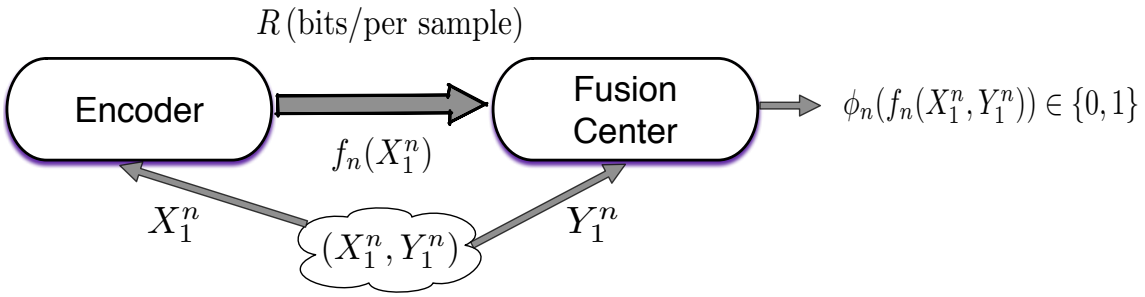


Figure 1.1: A schematic of a test with communication constraints. X_1^n and Y_1^n are random vectors. f_n and ϕ_n are the encoder and decoder, respectively.

locations (see an illustration in Fig. 1.1). More generally, Han [33] extended the derivation of performance fundamental bounds for the case where both sources are limited by a rate constraint. Several other constributions on this topic consider extension of this kind of problem (such as asymptotic decay of the TYPE I error or universal setting) and they can be found in [34–37].

In the second part of this thesis we address another important dimension of the problem of HT by assuming that the decision agent does not have direct access to the observations; rather, he/she has a lossy representation, more precisely, a finite rate version of it. In the classical HT setting (i.e. the centralized problem), the observations are collected at a single location with no perturbations. To begin the study of this decentralized inference problem, we consider the simplest scenario, namely, bivariate HT when one of the observation is measured remotely, and its information is transmitted over a noiseless channel of finite rate constraint [28] (see Fig. 1.1). On the specifics we derive general conditions on the Type I error restriction under which the error exponent of the optimal Type II error has a closed-form expression for the task of testing against independence. This expression shows the effect of the rate-constraint restriction in the inference power of the test. Importantly, we show that the performance limit (error exponent) is preserved for a large family of decreasing Type I error restrictions.

We also derive finite length performance bounds and show that these bounds can be used to accurately describe the optimal performance that can be achieved. We also describe the finite-length regimes where the error exponent is an excellent proxy for finite sample-size performances. The specific contributions are presented in Section 1.4.2.

1.3.3 Collaborative Hypothesis Testing

In the third part of this thesis we look at collaboration as a strategy to improve performance in HT. It is well known that collaboration is crucial in human communication, scientific research and decision-making processes. Collaboration as a communication strategy is key in human interaction as it allows people share their findings, discuss ideas, and identify areas where more research is needed. Effective collaboration requires clear communication and the ability to work together towards a common goal. On the technical side, collaboration has been adopted in sensor networks or surveillance systems [38], where collaboration between multiple sensors enhances HT capabilities. Each sensor may capture partial and noisy information,

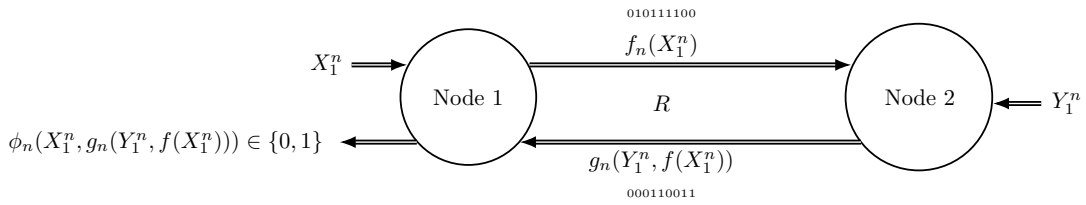


Figure 1.2: A schematic of collaborative distributed test.X_1^n and Y_1^n are random vectors. f_n and g_n are the encoders and ϕ_n is the decoder.

but by collaboratively fusing their evidence, a more comprehensive and accurate analysis can be performed. Collaboration helps in signal detection, pattern recognition, reducing error probabilities, and consequently, leading to improved decisions [39–42].

The problem of distributed HT in a collaboration context is interesting in theory and needed for a wide range of applications. Consider for example problems related with sensor networks and its practical applications (self-driving cars, array of sensors for measuring, internet of things). It is clear that the ability of automated systems to make minimum risk decisions in a timely manner is crucial in the 21st century. These systems will often operate under strict constraints over their resources. In some applications, e.g. automated systems, relatively short blocklengths are common both due to delay and complexity constraints imposed in the application. It is, therefore, of significance interest to assess the unavoidable penalty in performance (error exponents) required to sustain the desired fidelity at a given fixed blocklength.

In the context of our decentralized HT problem, collaboration means a decentralized framework where two agents (or nodes of the networks) interchange messages with an overall rate constraint to arrive at a final decision (see our setting in Fig 1.2). In this thesis, we derive an error exponent that expresses the benefit of collaboration in concrete terms — compared to the standard one-sided distributed strategy presented in Fig. 1.1. The bounds express the benefit of collaboration. Complementing this analysis, we also address the practical problem of designing encoders and decoders for this distributed task. For that problem, we propose an algorithm that uses ideas of machine learning for designing the encoder and decoder from data. The specific detail of these contributions are summarized in Section 1.4.3.

1.4 Technical Contributions of this Thesis

The technical contributions of this thesis are the following:

1.4.1 Finite Length Results for Hypothesis Testing

1. Building on the use of concentration inequalities [43], we offer new upper and lower bounds to the optimal Type II error probability for the case of finite observations.
2. The derived bounds are evaluated and interpreted numerically for some realistic models

considering different TYPE I error restrictions and number of samples.

1.4.2 Distributed Hypothesis Testing

1. We study a broader family of distributed HT problems (see Fig. 1.1) where the TYPE I error probability vanishes with the sample size. The objective here is to assess the impact of this more stringent set of restrictions on the asymptotic limit of TYPE II error probability given by the error exponent. Our main result here (cf. Theorem 3.1) gives new conditions on the admissible converge rate of the TYPE I error probability restriction under which the error exponent of the TYPE II error probability admits a closed-form expression.
2. For a family of sub-exponential decreasing TYPE I error probability restrictions, we show that the resulting error exponent matches the expression in [28, Theorem 3] while being consistent with the results obtained for the classical communication-free (centralized)problem [44].
3. Regarding the finite-length analysis, Theorem 3.2 offers new upper and lower bounds for the TYPE II error probabilities as a function of the number of samples, the underlying distributions, and the restriction on the TYPE I error probability.
4. Our bounds shed light on the speed at which the error exponent is achieved as the number of samples tends to infinity, and consequently, how well the performance limits represent the performances of practical decision schemes operating on a finite number of observations.
5. We evaluate our bounds numerically. We show that these expressions can be used to accurately describe the performance that can be achieved in practice with a scheme. We also analyze regimes where the error exponent is an excellent theoretical proxy for finite sample-size performances.

1.4.3 Collaborative Decentralized Hypothesis Testing

1. We introduce a one-round collaborative extension of the distributed setting introduced by Ahlswede and Csiszar in [28], as shown in Figure 1.2.
2. We derive an information limit (in the form of an error exponent) of the TYPE II error probability subject to a vanishing TYPE I error (Theorem 4.1).
3. We analyze the performance gain with respect to the one-sided (unidirectional) case introduced in [28]. We see that the overall quality of the test performance is governed by the bit assignment between the nodes and is affected, at the same time, by the distribution of the model.
4. On the practical side, we propose a data-driven design criterion for the two encoders and the decoder of the introduced one-round collaborative setting (see Figure 1.2). To design the encoder the problem is formulated as an info-max optimization task that learns the encoders from supervised data.
5. Empirical results based on simulations show that the proposed one-round collaboration strategy outperforms (in the ROC curve, i.e., TYPE I and TYPE II trade-off) the non-collaborative strategy and that the performance gain is a function of structural attributes in the model. Importantly, we show that the performance gain is proportional to a measure of the asymmetry of the underlying probability model.

6. We evaluate how the number of samples (block-length) and the communication constraint (number of bits) affect the mentioned comparison.

1.5 Main Hypothesis

The main hypotheses of this work are the following:

1. For the finite length analysis, the hypothesis is that we can obtain new non-asymptotic result for the scenario with monotonic (sub-exponential decreasing) restriction on the Type I error probability. This hypothesis is supported by the fact that similar results have been established by Strassen [27].
2. For the problem of distributed HT, the hypothesis is that non-asymptotic expression for the error exponent can be obtained and from this being able to analize the rate of convergence of this expression to the theoretical asymptotic expression. Related results has been established in the classical problem and we conjecture that this type of analysis could be extended to the more challenging rate constrained case. The main technical challenge was the difficulties of dealing with the likelihood ratio in the rate constrained set-up. We propose to adress this technical issue by extending the approach of Zhang et al. [45] to the case of noisy rate distortion theory and obtain fundamental bounds via concentration inequalities (bounded difference inequality and the Berry-Esséen theorem) [43].
3. For the problem of collaborative HT, our hypothesis is that collaboration could play a crucial role and lead to better performance. The conjecture was that this collective effort can lead to better decision in binary HT. We claim that the concept of error exponent can be used as a metric to verify this hypothesis.
4. On the design of the encoder and the decoder , we conjecture that we can improve the decision peformance using a family of soft encoders based on the Bolztmann distributions [46]. The importance of the Boltzmann distribution for encoders lies in its ability to optimize encoding strategies by considering the statistical properties of the source and the communication channel. Using this approach, we conjecture that encoders have the capacity to allocate resources efficiently, prioritize important information, and maximize the overall performance of the coding decision system.

1.6 Structure of the Thesis

The structure of this Thesis will adopt a paper-based approach, wherein each published paper will be presented as a separate chapter. Chapter 2 will encompass the contents of the first paper, providing a comprehensive exploration of its objectives, methodology, findings, and conclusions. Similarly, Chapter 3 will be dedicated to the second paper, delving into its methodologies employed, results obtained, and the corresponding implications.

1.6.1 Specific Structure of the Thesis

This Thesis is organized in five Chapters. Chapter 2 introduces the binary HT. Here, we present a new non-asymptotic result for the scenario with monotonic (sub-exponential de-

creasing) restriction on the TYPE II error probability. In Chapter 3, we study distributed binary HT of statistical independence under communication (information bits) constraints. In Chapter 4 we study collaboration in distributed inference for testing independence with a fixed-rate communication constraint. Finally, Chapter 5 presents the conclusions and exhibits some future work.

1.7 Publications

2022 IEEE ICASSP 2022 (Corresponding Author) [Conference]
Title *A DATA-DRIVEN QUANTIZATION DESIGN FOR DISTRIBUTED TESTING AGAINST INDEPENDENCE WITH COMMUNICATION CONSTRAINTS*
Authors Sebastián Espinosa, Jorge F. Silva and Pablo Piantanida

2021 IEEE Transactions on Signal and Information Processing over Networks (Corresponding Author)
Título *ON THE EXPONENTIAL APPROXIMATION OF THE TYPE II ERROR PROBABILITY OF DISTRIBUTED TEST OF INDEPENDENCE*
Journal Impact Factor: 3.664
Authors Sebastián Espinosa, Jorge F. Silva and Pablo Piantanida

2021 IEEE Signal Processsing Letters (Corresponding Author)
Title *FINITE-LENGTH BOUNDS ON HYPOTHESIS TESTING SUBJECT TO VANISHING TYPE I ERROR RESTRICTIONS*
Journal Impact Factor: 3.109
Authors Sebastián Espinosa, Jorge F. Silva and Pablo Piantanida

2019 GlobalSIP (Corresponding Author) [Conference]
Title *NEW RESULTS ON TESTING AGAINST INDEPENDENCE WITH RATE-LIMITED CONSTRAINTS*
Authors Sebastián Espinosa, Jorge F. Silva and Pablo Piantanida

Chapter 2

Finite-Length Bounds on Hypothesis Testing Subject to Vanishing Type I Error Restrictions

A central problem in Binary Hypothesis Testing (BHT) is to determine the optimal tradeoff between the Type I error (referred to as *false alarm*) and Type II (referred to as *miss*) error. In this context, the exponential rate of convergence of the optimal miss error probability — as the sample size tends to infinity — given some (positive) restrictions on the false alarm probabilities is a fundamental question to address in theory. Considering the more realistic context of a BHT with a finite number of observations, this paper presents a new non-asymptotic result for the scenario with monotonic (sub-exponential decreasing) restriction on the Type I error probability, which extends the result presented by Strassen in 2009. Building on the use of concentration inequalities, we offer new upper and lower bounds to the optimal Type II error probability for the case of finite observations. Finally, the derived bounds are evaluated and interpreted numerically (as a function of the number samples) for some vanishing Type I error restrictions.

2.1 Introduction

Binary Hypothesis Testing (BHT) is a common problem in statistics and it has been richly used as a method to statistical signal detection [47,48]. In particular, the celebrated *Neyman-Pearson lemma* provides the optimal detection scheme for this testing task [25]. On the specifics, let us consider the classical n-length BHT setting given by

$$\begin{cases} H_0: & X_1^n \sim P^n, \\ H_1: & X_1^n \sim Q^n, \end{cases}$$

where $P, Q \in \mathcal{P}(\mathbb{X})$ with $D(P\|Q) > 0$ and $X_1^n = (X_1, ..., X_n)$ is a random vector with their length as a superscript. In this work, we restrict our attention to the case of a finite-alphabet $\mathbb{X}$, where $\mathcal{P}(\mathbb{X})$ denotes the family of probabilities on $\mathbb{X}$. A decision rule ϕ_n of length n is a

function $\phi_n : \mathbb{X}^n \to \Theta \triangleq \{0,1\}$, from which two types of errors are induced [49]:

$$\begin{aligned} P_0(\phi_n) &\equiv P^n(\{x_1^n \in \mathbb{X}^n : \phi_n(x_1^n) \neq 0\}) \triangleq P^n(\mathcal{A}^c(\phi_n)), \\ P_1(\phi_n) &\equiv Q^n(\{x_1^n \in \mathbb{X}^n : \phi_n(x_1^n) = 0\}) \triangleq Q^n(\mathcal{A}(\phi_n)), \end{aligned}$$

with decision region $\mathcal{A}(\phi_n) \triangleq \{x_1^n \in \mathbb{X}^n : \phi_n(x_1^n) = 0\}$.

For a given sequence $(\epsilon_n)_n$ of non-negative values such that $\lim_{n\to\infty} \epsilon_n = 0$, we study the solution to:

$$\beta_n(\epsilon_n) \equiv \min_{\phi_n \in \Phi_n} \{P_1(\phi_n) : \text{ s.t. } P_0(\phi_n) \leq \epsilon_n\}, \forall n \geq 1, \tag{2.1}$$

where $\Phi_n \equiv \{\phi_n : \mathbb{X}^n \to \Theta\}$ denotes the class of n-length detectors. Importantly, $(\beta_n(\epsilon_n))_{n\geq 1}$ represents the optimum TYPE II error sequence that satisfies a sequence of fixed TYPE I error constraints.

The *Neyman-Pearson* lemma [50] offers the optimal trade-off between the two type of errors[1]. In this context, the determination of the (exponential) rate of convergence of the TYPE II error, which is known as the error exponent, has been a central problem in HT's analysis. Indeed, the error exponent is seen as an indicator of the complexity of the decision task (function of P_0, P_1 and $(\epsilon_n)_n$) and has found numerous applications [15,16]. For the important case when $\epsilon_n = \epsilon > 0$ for all n, the celebrated *Stein's lemma* establishes that the error exponent of the TYPE II error is given by the KL divergence $D(P\|Q) \equiv \sum_{x\in\mathbb{X}} P(x) \log \frac{P(x)}{Q(x)}$ [2, 49].

Lemma 2.1.1 (*Stein's lemma* [2, 30]) For any fixed $\epsilon \in (0,1)$, $\lim_{n\to\infty} -\frac{1}{n}\log(\beta_n(\epsilon)) = D(P\|Q)$.

Importantly, the error exponent limit in Lemma 2.1.1 is independent of $\epsilon > 0$. However, this limit changes when we impose a setting with a monotonic decreasing TYPE I error restrictions. In particular, Han *et al.* [52] studied the case when the TYPE I error sequence has an exponential decreasing behaviour. Nagakawa *et al.* [44] extended this analysis for a family of decreasing sequence of TYPE I error restrictions:

Lemma 2.1.2 [44, Nakagawa] Let us assume that $\epsilon_n \leq e^{-rn}$ for some $r \in (0, D(P\|Q))$, then $\lim_{n\to\infty} -\frac{1}{n}\log(\beta_n(\epsilon_n)) = D(P_{t^*}\|Q)$, where $P_{t^*}(x) \equiv C_{t^*} P(x)^{1-t^*} Q(x)^{t^*}$ $\forall x \in \mathbb{X}$, and t^* is the solution of $D(P_{t^*}\|P) = r$.

A direct implication of Lemma 2.1.2 is the following result:

Corollary 1 [44] Let us assume that $(1/\epsilon_n)_n$ is $o(e^{rn})$ for any $r > 0$, then

$$\lim_{n\to\infty} -\frac{1}{n}\log(\beta_n(\epsilon_n)) = D(P\|Q). \tag{2.2}$$

Importantly, Corollary 1 shows that the same error exponent of the Stein's lemma is obtained for these stringent family of problems — where $(\epsilon_n)_n$ tends to zero at a sub-exponential

[1] See [51] for a new proof based on properties of exponential density function families.

rate. In contrast, when the TYPE I error restriction tends to zero exponentially fast (Lemma 2.1.2), the error exponent is strictly smaller than $D(P\|Q)$.

2.1.1 Finite-Length Context and Contribution

In many practical problems, the statistician has access only to a finite number of observations. Consequently, it is critical to obtain non-asymptotic bounds for the probability of error $\beta_n(\epsilon_n)$ for a finite n. Concerning the non-asymptotic analysis of this problem, the following result was derived by Strassen for the specific regime when $\epsilon_n = \epsilon > 0$ for all $n \geq 1$ [27].

Lemma 2.1.3 [27] Let us consider $\epsilon \in (0,1)$, then eventually with n, it follows that

$$-\frac{\log(\beta_n(\epsilon))}{n} = D(P\|Q) + \sqrt{\frac{V(P\|Q)}{n}}\Phi^{-1}(\epsilon) + \frac{\log n}{2n} + \mathcal{O}\left(\frac{1}{n}\right), \tag{2.3}$$

where $V(P\|Q) \equiv \sum_{x\in\mathbb{X}} P(\{x\}) \left[\log\left(\frac{P(\{x\})}{Q(\{x\})}\right) - D(P\|Q)\right]^2$.

Lemma 2.1.3 shows that $\left|D(P\|Q) - \left(-\frac{1}{n}\log(\beta_n(\epsilon))\right)\right|$ is $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, which expresses the velocity of convergence of $-\frac{1}{n}\log(\beta_n(\epsilon))$ to its limit $D(P\|Q)$. Given the practical importance of this type of finite length results, it is very relevant to derive new results that extend Lemma 2.1.3 to our general problem in (2.1), as a function of P, Q, $(\epsilon_n)_n$ and n. In addition, it is critical that these bounds can be evaluated for its practical use. This last aspect is not achieved in Lemma 2.1.3, which from that perspective is an asymptotic (convergence) result.

The main contribution of this chapter goes in this direction, where we derive new upper and lower bounds for the discrepancy between $-\frac{1}{n}\log(\beta_n(\epsilon_n))$ and its information limit $D(P\|Q)$ for any finite $n \geq 1$ when $(\epsilon_n)_n$ tends to zero at a sub-exponential rate. These expressions can be evaluated and interpreted numerically in any context where we know the models (P and Q) and the parameters of the problem (ϵ_n and n). In addition, these new bounds stipulate the velocity at which the error exponent is achieved as the sample size tends to infinity. From this, we could assess how realistic the information limits (asymptotic results) are in practice when facing a problem with a finite number of observations. To conclude our analysis, we numerically compute and evaluate the expressions obtained by our result to show the derived bounds' tightness for some specific scenarios.

2.1.2 Related Work

In a Bayesian setting, Sason [53] obtained an upper bound to the optimal Bayesian probability of error (non-asymptotic) by bounding the TYPE I and TYPE II errors simultaneously in such a way that they both decay to zero sub-exponentially with n. It is worth to mention that this work differs from the current setting in the sense that we are interested in bounding the discrepancy between $-\frac{1}{n}\log(\beta_n(\epsilon_n))$ and its information limit and how this analysis depends on the vanishing TYPE I error restrictions. In addition, we are interested in the velocity of convergence of $-\frac{1}{n}\log(\beta_n(\epsilon_n))$ to its information limit and the impact of considering stringent

restriction on TYPE I errors $(\epsilon_n)_n$. Complementing this chapter, [54] studies a distributed (two-terminal) version of the BHT problem subject to communication (rates) constraints. Our results here do not derive from [54] since the setups are very different from each other, and different tools are used to address them. Finally, a similar analysis of the TYPE I error has been addressed by Bahadur [55]. In contrast to this work's focus, this analysis considers a fixed restriction on the power of a test (1−TYPE II error) to determine the exponential rate of convergence of their sizes (TYPE I error) as n tends to infinity.

2.1.3 Notations and Organization

$(b_n)_n$ being $o(a_n)$ indicates that $\limsup_{n\to\infty}(b_n/a_n) = 0$ and $(b_n)_n$ being $\mathcal{O}(a_n)$ indicates that $\limsup_{n\to\infty}|b_n/a_n| < \infty$. We say that $(f(n))_n \approx (g(n))_n$ if there exists a constant $C > 0$ such that $f(n) = Cg(n)$ eventually in n. The rest of the chapter is organized as follows: Section 2.2 presents the main result of this work. Numerical analysis and discussions are presented in Section 2.3. The proof of is in Sect. 2.4.1.

2.2 Main Result

The main result of this Chapter extends Lemma 2.1.3 offering new non-asymptotic bounds for $\beta_n(\epsilon_n)$ in (2.1) under sub-exponential TYPE I error restrictions. In particular, the next result provides upper and lower bounds for the discrepancy between $-\frac{1}{n}\log(\beta_n(\epsilon_n))$ and $D(P\|Q)$.

Theorem 2.1 *Let us assume that $P \ll Q$ and that $(1/\epsilon_n)_n$ is $o(e^{rn})$ for any $r > 0$. Then, eventually in n, it follows that:*

$$-\tfrac{1}{n}\log(\beta_n(\epsilon_n)) \geq D(P\|Q) - C_X(P,Q)\sqrt{\frac{2\ln(1/\epsilon_n)}{n}}$$

$$-\tfrac{1}{n}\log(\beta_n(\epsilon_n)) \leq D(P\|Q) + \frac{\log\left(\frac{1}{1-\epsilon_n-\delta_n}\right)}{n} + \delta_n$$

where $C_X(P,Q) \equiv \sup_{x\in\mathbb{X}}\left|\log\left(\frac{P(\{x\})}{Q(\{x\})}\right)\right|$ *and* $\delta_n \equiv C_X(P,Q)\sqrt{\frac{2\ln(1/\epsilon_n)}{n}}$.

The proof is presented in Section 2.4.1.

2.2.1 Interpretation and Discussion of Theorem 2.1

1: This result establishes a non-asymptotic rate of convergence for the TYPE II error when we impose a vanishing condition on $(\epsilon_n)_n$ that is sub-exponential. Interestingly, the bounds for the discrepancy $-\frac{1}{n}\log(\beta_n(\epsilon_n))$ depend explicitly on the sequence $(\epsilon_n)_n$.
2: It is worth noting that the dependency on $(\epsilon_n)_n$ observed in our result is non-observed in the asymptotic limit in Corollary 1, which is $D(P\|Q)$ as long as $(1/\epsilon_n)_n$ is sub-exponential.
3: Adding on the previous point, the fact that the asymptotic error exponent is invariant from the simpler fixed TYPE I setup (in Lemma 2.1.1) to the more restrictive sub-exponential TYPE I error decay setting (in Corollary 1), it is however shown in our non-asymptotic result in term of the rate of convergence to the limit $D(P\|Q)$. In particular, there is a

concrete penalty $\mathcal{O}(\sqrt{\log(1/\epsilon_n)})$ on the velocity of convergence to zero of the discrepancy $(-\frac{1}{n}\log\beta_n(\epsilon_n) - D(P\|Q))$ in our result compared with what is obtained in Lemma 2.1.3.
4: The proof of the Theorem 2.1 has two parts: the constructive and unfeasibility arguments. Both arguments are constructed from concentration inequalities using the i.i.d. structure of the observations. For the constructive argument, we apply the bounded difference inequality [43]. On the unfeasibility argument, we use (concentration) results from typical sequences [2] to construct a lower bound on the minimum probability of TYPE II error.
5: If we impose a fixed value of $\epsilon_n = \epsilon \in (0,1)$, our result recovers the rate of convergence for the TYPE II error given by Lemma 2.1.3. However, we obtained explicit bounds.

2.3 Practical Implications of Theorem 2.1

In this section, we show how Theorem 2.1 may be adopted by a statistician to obtain bounds on $\beta_n(\epsilon_n)$ when n is finite. The resulting bounds provide an interval of feasibility for $\beta_n(\epsilon_n)$:

$$\mathrm{UB}(\epsilon_n) \equiv \exp\Big[-n\Big(D(P\|Q) - \sqrt{\frac{2\ln(1/\epsilon_n)}{n}}C_X(P,Q)\Big)\Big],$$
$$\mathrm{LB}(\epsilon_n) \equiv \exp\Big[-n\Big(D(P\|Q) - \frac{1}{n}\log\left(1-\epsilon_n-\delta_n(\epsilon_n)\right) + \delta_n(\epsilon_n)\Big)\Big].$$

The length of $[\mathrm{LB}(\epsilon_n), \mathrm{UB}(\epsilon_n)]$ indicates the precision of the result and, at the same time, the interval $[\mathrm{LB}(\epsilon_n), \mathrm{UB}(\epsilon_n)]$ can be used to measure how close $\beta_n(\epsilon_n)$ is to $e^{-nD(P\|Q)}$.

	Number of observations n					
ϵ_n	50	250	350	550	650	750
$1/\log(n)$	2.3587e-10	1.0595e-83	9.4592e-124	2.6103e-206	2.2862e-248	8.6970e-291
$1/n^{0.1}$	7.8229e-17	9.1096e-99	1.3994e-141	1.3117e-228	1.4272e-272	9.5440e-317
$1/n$	0.5571	7.4403e-56	2.7823e-89	1.4443e-160	1.2489e-197	2.3163e-235

Table 2.1: Magnitude of $\mathrm{UB}(\epsilon_n) - \mathrm{LB}(\epsilon_n)$ function of ϵ_n and n for the case when $D(P\|Q) = 1$.

Table 2.1 presents the length of $[\mathrm{LB}(\epsilon_n), \mathrm{UB}(\epsilon_n)]$ for three regimes of: $\epsilon_n \in \{n^{-1}, n^{-0.1}, 1/\log(n)\}$, and two models P, Q where $D(P\|Q) = 1$ with $|\mathbb{X}| = 15$. First, we observe that the length of $[\mathrm{LB}(\epsilon_n), \mathrm{UB}(\epsilon_n)]$ vanishes exponentially fast with the sample size. From this exponential decay, we observe that the centered value predicted by Theorem 2.1, i.e., the exponential behavior $\exp(-nD(P\|Q))$, is a good approximation for $\beta_n(\epsilon_n)$ provided that n is sufficiently large. This supports the idea that $\exp(-nD(P\|Q))$ is a useful proxy for $\beta_n(\epsilon_n)$ provided that a Critical Sample Size (CSS) is achieved (more details on this below). Table 2.1 also shows that the result's precision is affected by the velocity of convergence of the TYPE I error restriction $(\epsilon_n)_n$, which is consistent with the statement and the analysis of our main result. In particular, for a faster speed of convergence of $(\epsilon_n)_n$ to zero (i.e., a stringer problem), the gap between the bounds is more prominent, which means that the bounds of Theorem 2.1 are expected to be less informative about $\beta_n(\epsilon_n)$.

Regarding the implications of the above bounds to measure the gap between $\beta_n(\epsilon_n)$ and $e^{-nD(P\|Q)}$, we address the following question: given an arbitrary value of $\delta > 0$ of the

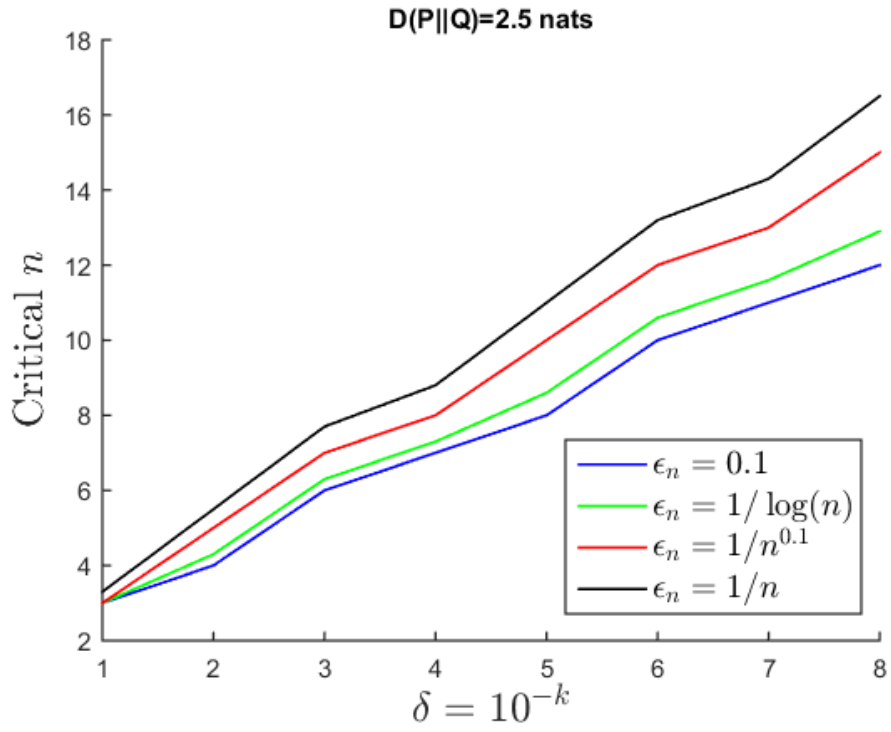


Figure 2.1: Critical number of samples (CSS) predicted by Th. 2.1 across different values of $\delta = 10^{-k}$. High divergence case with $D(P\|Q) = 2.5$ and $C_X(P,Q) = 2.04$.

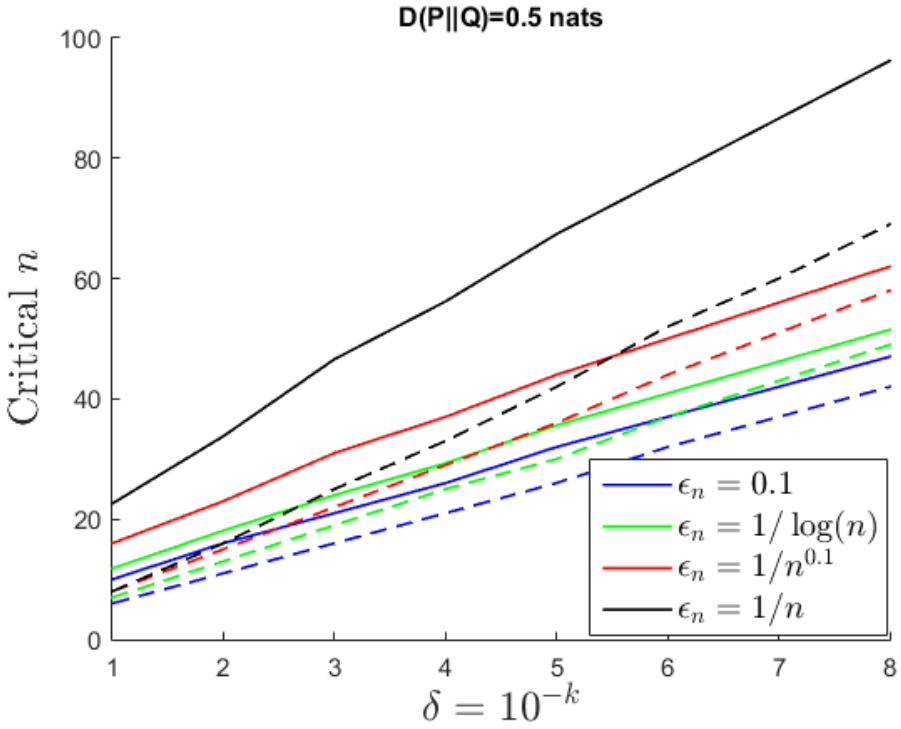


Figure 2.2: CSS predicted by Th. 2.1 across different values of $\delta = 10^{-k}$. Low divergence case with $D(P\|Q) = 0.5$ and $C_X(P,Q) = 1.03$. The dashed lines show an estimation of the exact CSS obtained from $\beta_n(\epsilon_n)$ directly.

form 10^{-k} with $k \in \{1, \ldots, 8\}$, and for two arbitrary models P and Q, we want to predict from Theorem 2.1 the minimum number of samples required to guarantee that $\beta_n(\epsilon_n) \in (e^{-nD(P\|Q)} - \delta, e^{-nD(P\|Q)} + \delta)$. The exponential decay of the length of $[\mathrm{LB}(\epsilon_n), \mathrm{UB}(\epsilon_n)]$, observed in Table 2.1, implies that this should happen eventually with n very quickly. Indeed, we can derive an upper bound for this critical number of samples (CSS) from the expressions we have for $\mathrm{LB}(\epsilon_n)$ and $\mathrm{UB}(\epsilon_n)$.[2] Figures 2.1 and 2.2 present the predicted CSS versus $\delta = 10^{-k}$ for different scenarios of P, Q (in terms of $D(P\|Q)$) and $(\epsilon_n)_n$. We consider two scenarios for P and Q (low divergence $D(P\|Q) = 0.5$ and high divergence $D(P\|Q) = 2.5$) and we explore $(\epsilon_n)_n \in \{n^{-1}, n^{-0.1}, 1/\log(n), 0.1\}$. Figures 2.1 and 2.2 show that even for really small precision $\delta = 10^{-8}$ the point at which $\beta_n(\epsilon_n)$ can be well approximated by $e^{-nD(P\|Q)}$ requires at most 16 samples and 60 samples for high and low divergence cases, respectively, and the majority of $(\epsilon_n)_n$. The dependency of these curves on the magnitude of $D(P\|Q)$ and $(\epsilon_n)_n$ is clearly expressed in these findings, which is consistent with our previous analyses.

Finally, to evaluate the tightness of our predictions, we simulate i.i.d. samples according to P and Q from which a precise empirical estimation of $\beta_n(\epsilon_n)$ is derived. In particular, given P, Q and $(\epsilon_n)_n$, we obtained empirical estimations of the error probabilities (Type I and Type II) from which we estimate $\beta_n(\epsilon_n)$. For this purpose, $2.5 \cdot 10^6$ realizations of P and Q were used to have good estimations of these probabilities. Using the estimated values of $\beta_n(\epsilon_n)$, we obtain the point where $\beta_n(\epsilon_n) \in (e^{-nD(P\|Q)} - \delta, e^{-nD(P\|Q)} + \delta)$ directly. Figure 2.2 contrasts our predictions and the true (estimated) values (the dashed lines) of the CSS. Consistent with our result's nature, our prediction of the CSS is more conservative than the true CSS estimated from simulations. Importantly, this discrepancy is not significant overall, expressing that our bounds are useful for this analysis and can be adopted in cases where it is impractical to estimate $\beta_n(\epsilon_n)$ from data. Indeed, in this analysis, we face this issue, and it is very difficult to obtain accurate estimates of $\beta_n(\epsilon_n)$ for high divergence regimes. Notice that $\beta_n(\epsilon_n)$ is of order: $O(e^{-nD(P\|Q)})$ for which around $e^{nD(P\|Q)}$ simulations (i.e., i.i.d. samples from P and Q) are needed. This becomes impractical even for n less than 30 when $D(P\|Q)$ is relatively large.

[2]The predicted CSS is the first $n \geq 1$ such that $\max\{\mathrm{UB}(\epsilon_n) - e^{-nD(P\|Q)}, e^{-nD(P\|Q)} - \mathrm{LB}(\epsilon_n)\} \leq \delta$, which is finite for any $\delta > 0$.

2.4 Appendix

2.4.1 Proof of Theorem 2.1

We divide the proof of Theorem 2.1 in two parts.

Lower Bound Analysis

Under the assumption of Theorem 2.1, let us verify that

$$D(P\|Q) - \left(-\frac{1}{n}\log\beta_n(\epsilon_n)\right) \leq \sqrt{\frac{2\ln(1/\epsilon_n)}{n}}C_X(P,Q).$$

Let us consider the corresponding optimal decision regions from the Neyman-Pearson Lemma parameterized in the following way: $\forall t > 0$,

$$\mathcal{B}_{n,t} = \left\{x_1^n \in \mathbb{X}^n : \frac{P^n(\{x_1^n\})}{Q^n(\{x_1^n\})} > e^{nt}\right\}. \tag{2.4}$$

Considering the induced test $\phi_{n,t}(\cdot) : \mathbb{X}^n \mapsto \{0,1\}$ such that $\phi_{n,t}^{-1}(\{0\}) = \mathcal{B}_{n,t}$. The TYPE I error probability is given by $P^n(\mathcal{B}_{n,t}^c)$. An upper bound for the TYPE II follows as:

$$Q^n\left(\mathcal{B}_{n,t}\right) \leq e^{-nt}. \tag{2.5}$$

Then, for any finite $n > 0$ and $\epsilon_n > 0$, finding an achievable TYPE II error exponent from this construction (and the bound in Eq.(2.5)) reduces to solve the following problem:

$$t_n^*(\epsilon_n) \triangleq \sup_t\{t : P^n(\mathcal{B}_{n,t}^c) \leq \epsilon_n\}. \tag{2.6}$$

It will be convenient to re-parameterize t with respect to the value $D(P\|Q)$. More precisely, let us define

$$t_\delta \triangleq D(P\|Q) - \delta,$$

for any $\delta > 0$. Then using the bounded difference inequality [43], we obtain

$$\begin{aligned} P^n\left(B_{n,t_\delta}^c\right) &= P^n\left(x_1^n \in \mathbb{X}^n : \left|\hat{D}(P\|Q) - D(P\|Q)\right| \geq \delta\right) \\ &\leq \exp\left(\frac{-n\delta^2}{2C_X(P,Q)^2}\right), \end{aligned} \tag{2.7}$$

where $\hat{D}(P\|Q) \triangleq \frac{1}{n}\sum_{i=1}^n \log\left(\frac{P(\{x_i\})}{Q(\{x_i\})}\right)$ is the empirical divergence. Finally, from Eq. (2.6) a lower bound for $t_n^*(\epsilon_n)$ can be determined from Eq. (2.7) by letting $\tilde{\delta}_n(\epsilon_n)$ to be the solution of the following equality:

$$\exp\left(\frac{-n\tilde{\delta}_n(\epsilon_n)^2}{2C_X(P,Q)^2}\right) = \epsilon_n. \tag{2.8}$$

Consequently, we have that

$$t_n^*(\epsilon_n) \geq t_{\tilde{\delta}_n(\epsilon_n)} \triangleq D(P\|Q) - \sqrt{\frac{2\log(1/\epsilon_n)}{n}}C_X(P,Q). \tag{2.9}$$

Finally, replacing the bound of (2.9) in (2.5) and taking logarithm we have that:

$$D(P\|Q) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n))\right) \leq \sqrt{\frac{2\ln(1/\epsilon_n)}{n}} C_X(P,Q), \tag{2.10}$$

which concludes this part.

Upper Bound Analysis

Let us consider the set

$$\mathcal{A}_{n,\delta}^c \triangleq \left\{ x_1^n \in \mathbb{X}^n : \left| \frac{1}{n}\log\left(\frac{P^n(\{x_1^n\})}{Q^n(\{x_1^n\})}\right) - D(P\|Q) \right| \geq \delta \right\}, \tag{2.11}$$

for any $\delta > 0$. We have the following result:

Lemma 2.4.1 [2, Sect 11.8] For any set $\mathcal{B}_n \subseteq \mathbb{X}^n$ and its induced test ϕ_n[3] such that operates at TYPE I error ϵ_n (i.e. $P^n(\mathcal{B}_n^c) \leq \epsilon_n$), then

$$Q^n(\mathcal{B}_n) \geq (1 - \epsilon_n - \delta)2^{-n(D(P\|Q)+\delta)}. \tag{2.12}$$

By construction, it is clear that there exists $\delta > 0$ such that $\mathcal{A}_{n,\delta}^c$ operates at TYPE I error ϵ_n. In fact, we consider

$$\delta_n^* \triangleq \sup\{\delta : P^n(\mathcal{A}_{n,\delta}^c) \leq \epsilon_n\}. \tag{2.13}$$

Using the bounded difference inequality [43], we get that

$$\begin{aligned} P^n\left(\mathcal{A}_{n,\delta}^c\right) &= P^n\left(x_1^n \in \mathbb{X}^n : \left|\hat{D}(P\|Q) - D(P\|Q)\right| \geq \delta\right) \\ &\leq \exp\left(\frac{-n\delta^2}{2C_X(P,Q)^2}\right). \end{aligned} \tag{2.14}$$

Using the same argument from the lower bound analysis, we obtain a lower bound for δ_n^* given by

$$\delta_n^* \geq \delta_n \triangleq \sqrt{\frac{2\log(1/\epsilon_n)}{n}} C_X(P,Q). \tag{2.15}$$

Finally, replacing δ_n in Eq. (2.12) and taking logarithm, we have that for any set $\mathcal{B}_n$ satisfying the assumptions of Lemma 2.4.1:

$$-\frac{1}{n}\log(Q^n(\mathcal{B}_n)) \leq D(P\|Q) + \frac{\log\left(\frac{1}{1-\epsilon_n-\delta_n}\right)}{n} + \delta_n. \tag{2.16}$$

Therefore, we can choose the optimum set which implies that

$$-\frac{1}{n}\log(\beta_n(\epsilon_n)) \leq D(P\|Q) + \frac{\log\left(\frac{1}{1-\epsilon_n-\delta_n}\right)}{n} + \delta_n. \tag{2.17}$$

This concludes the proof.

[3] Meaning that $\phi_n(x_1^n) = 0$ if $x_1^n \in \mathcal{B}_n$.

Chapter 3

On the Exponential Approximation of Type II Error Probability of Distributed Test of Independence

This chapter studies distributed binary test of statistical independence under communication (information bits) constraints. While testing independence is very relevant in various applications, distributed independence test is particularly useful for event detection in sensor networks where data correlation often occurs among observations of devices in the presence of a signal of interest. By focusing on the case of two devices because of their tractability, we begin by investigating conditions on Type I error probability restrictions under which the minimum Type II error admits an exponential behavior with the sample size. Then, we study the finite sample-size regime of this problem. We derive new upper and lower bounds for the gap between the minimum Type II error and its exponential approximation under different setups, including restrictions imposed on the vanishing Type I error probability. Our theoretical results shed light on the sample-size regimes at which approximations of the Type II error probability via error exponents became informative enough in the sense of predicting well the actual error probability. We finally discuss an application of our results where the gap is evaluated numerically, and we show that exponential approximations are not only tractable but also a valuable proxy for the Type II probability of error in the finite-length regime.

3.1 Introduction

Motivated by decision-making problems over networks, researchers within the field of statistical signal processing have been involved in a wide range of research initiatives studying decision and inference problems in the presence of quantization or measurement noise or data corruption by various types of perturbations. In real-world applications, these sources of degradation come from factors such as noise observations at the sensors, communication restrictions between sensors and decision agents, or by the presence of external sources of perturbations corrupting data [23]. The emerging field of Internet of Things (IoT) brings new dimensions and technical challenges to the classical problem as data is no longer centrally

available at the decision end. A related emerging domain is known as signal processing in the context of unlabeled or unordered data [15–21]. Another important domain, which is the general focus of this work, is distributed detection under data-compression [28, 52, 56]. The derivation of performance limits and characterization of statistical properties of optimal detectors have been active research areas over the past years.

Distributed detection, data fusion, and multisensor integration have a long history in statistics, signal, and information processing at large. Fundamental works can be traced back to [57], [25], and [58], among others. Applications of the decentralized decision framework arise in communications and sensor networks, for instance, in the context of Multiple Access Channels (MAC) [59] and wireless sensor networks [47]. These works do not only investigate practical solutions but, importantly, they study theoretical guarantees and performance bounds to understand the intrinsic complexity of these problems. In [60], the authors derived performances in the form of error exponents of the Type I and Type II error probabilities over Fading MACs. However, explicit communication restrictions between the sensors and the fusion center still remain a challenging problem [47, 59], which implies understanding how (detection) performances are affected by the introduction of non-trivial communication restrictions. Indeed, a crucial case of particular interest is when the fusion center receives quantized descriptions of the measurements taken by remote sensors [61, 62]. Some recent contributions have explored the asymptotic performance limits based on error exponents of distributed scenarios with multiple decision centers and rate constraints between sensors and detectors [63], [29], [64]. Communications constraints have also been studied within the framework of Bayesian detection in [65] and [66].

This chapter investigates the problem of distributed binary Hypothesis Testing (HT) of statistical independence under communication (information bits) constraints. In particular, we focus on non-asymptotic performance bounds. More specifically, we study the gap between the minimum Type II error probability and its exponential approximation under restrictions on the vanishing Type I error probability. To this end, we revisit the distributed scenario with communication constraints first introduced in [28]. This problem consists in testing against independence where the observations (e.g., sensor measurements) come from two modalities (e.g., two sensors), as shown in Fig. 3.1. One of the modalities is to be transmitted to the decision-maker (or detector) using an error-free communication channel that introduces a positive rate-constraint (in bits per sample). [28] derives the characterization of asymptotic performance bounds in terms of a closed-form expression for the error exponent of Type II error probability given a fixed restriction on the Type I error probability ($\epsilon > 0$) [28, Ths. 2 and 3]. Notably, the results show the effect of the communication constraints in asymptotic performance (error exponent), which is shown to be independent of ϵ. Later on, [52] derive an asymptotic lower bound for the error exponent when Type I restriction (as a sequence) tends to zero (with the sample size n) at an exponential rate given by $\mathcal{O}(e^{-nr})$.

3.1.1 Summary of contributions

Our work advances state-of-the-art in very different ways.

1. We study a broader family of problems (see Fig. 3.1) where the Type I error probability vanishes with the sample size. The objective here is to assess the impact of this stringer

set of restrictions on the asymptotic limit of TYPE II error probability given by the error exponent. Building on concentration inequalities and results from rate-distortion theory, our main result here (cf. Theorem 3.1) gives new conditions on the admissible converge rate of the TYPE I error probability restriction under which the error exponent of the TYPE II error probability admits a closed-form expression. Interestingly, for a family of sub-exponential decreasing TYPE I error probability restrictions, we show that the resulting error exponent matches the expression in [28, Theorem 3] while being consistent with the results obtained for the classical communication-free problem [44].

2. Regarding the non-asymptotic regime of this problem, Theorem 3.2 offers new upper and lower bounds for the TYPE II error probabilities as a function of the number of samples, the underlying distributions, and the restriction on the TYPE I error probability. As an important corollary, our bounds shed light on the speed at which the error exponent is achieved as the number of samples tends to infinity, and consequently, how well the performance limits represent the performances of practical decision schemes operating based on a finite number of samples.

3. Finally, we evaluate our bounds numerically and show that these can be used to accurately describe the optimal performance that can be achieved and, in particular, to devise the regimes where the error exponent is an accurate proxy for finite sample-size performances.

3.1.2 Related works

In terms of finite sample-size analysis within the centralized framework, [27] presented non-asymptotic results for the optimal TYPE II error probability under a constant TYPE I error restriction in the i.i.d case. Interestingly, the discrepancy between optimal finite-length and asymptotic performance was characterized, scaling as $\mathcal{O}(\sqrt{n})$ with the sample size n. In the same communication-free context, [53] borrows ideas from *moderate deviation analysis* [67] to obtain an interesting upper bound for the Bayesian error probability by bounding the TYPE I-TYPE II errors in a way that both decay to zero sub-exponentially with n. More recently, in [68], we obtained non-asymptotic upper and lower bounds for the TYPE II error probability for i.i.d samples draw according to two arbitrary distributions. We showed that the error exponent is a good approximation for the TYPE II error probability in the finite sample regime. Importantly, the distributed setting investigated in this work, with a non-trivial rate constraint in one of the modalities, induces a mathematical problem that is fundamentally different in terms of the requested tools. Communication restrictions subject to zero-rate (in bits-per sample) have been investigated in [56]. The error exponent and non-asymptotic bounds have been characterized. Extensions to interactive HT with zero-rate have been reported in [69].

A preliminary version of this work was presented in [54] with partial results and sketches of some of the arguments. In this chapter, we extend the results for a larger family of scenarios, provide complete proofs of the results and more systematic analysis of the practical implications of these results.

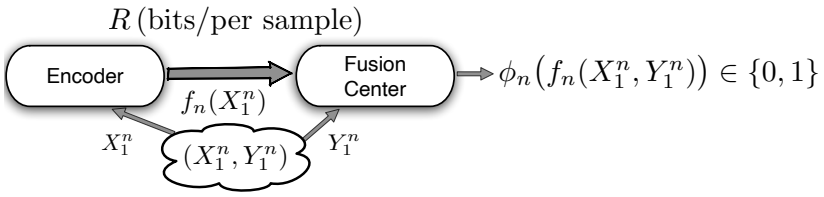


Figure 3.1: Illustration of the coding-decision problem with one-side communication constraint. f_n is the encoder of X_1^n (one of the modalities) and ϕ_n is the detector acting on the one-side compressed measurements $(f_n(X_1^n), Y_1^n)$.

3.1.3 Chapter Organization

The outline of the chapter is as follows. Section 3.2 introduces the main definitions and reviews some seminal results for the case of unconstrained communication. Sections 3.3 and 3.4 present our main theoretical results for the asymptotic and the non-asymptotic regimes, respectively. Numerical analysis and discussions are relegated to Section 3.5. Finally, the proofs are relegated to Appendix.

3.1.4 Notations and Conventions

Boldface letters x_1^n and upper-case letters X_1^n are used to denote vectors and random vectors of length n, respectively. Let X, Y and V be three random variables with joint probability $P_{X,Y,V}$. If $P_{X|Y,V}(x|y,v) = p_{X|Y}(x|y)$ for each x, y, v, then (X, Y, V) forms a Markov chain, which is denoted by $X \multimap Y \multimap V$. Let (b_n) and (a_n) be sequences, $(b_n) = o(a_n)$ indicates that $\limsup_{n\to\infty}(b_n/a_n) = 0$, and $(b_n) = \mathcal{O}(a_n)$ indicates that $\limsup_{n\to\infty}|b_n/a_n| < \infty$. We say that $(a_n) \approx (b_n)$ if for sufficiently large $N > 0$ there exists a constant $C > 0$ such that $a_n = Cb_n$, for all $n \geq N$.

3.2 Problem Setting and Preliminaries

Let us consider a finite alphabet product space $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$, where $\mathcal{P}(\mathbb{Z})$ denotes the family of probabilities on $\mathbb{Z}$. We have a joint random vector (X, Y) with values in $\mathbb{Z}$ and equipped with a joint probability $P \in \mathcal{P}(\mathbb{Z})$ where $P_X \in \mathcal{P}(\mathbb{X})$ and $P_Y \in \mathcal{P}(\mathbb{Y})$ denote the marginal of X and Y, respectively. $X_1^n = (X_1, ..., X_n)$ and $Y_1^n = (Y_1, ..., Y_n)$ denote the finite block vector with product (i.i.d.) distribution $P_{X,Y}^n \triangleq P_{X_1^n Y_1^n} \in \mathcal{P}(\mathbb{X}^n \times \mathbb{Y}^n)$. We consider two scenarios for the data generated distribution of (X_1^n, Y_1^n), i.e.,

$$\begin{aligned} H_0: &\quad (X_1^n, Y_1^n) \sim P_{XY}^n, \\ H_1: &\quad (X_1^n, Y_1^n) \sim Q_{XY}^n, \end{aligned} \tag{3.1}$$

where $Q_{XY}^n = P_X^n \cdot P_Y^n$ denote the product probability modeling the case where X_1^n and Y_1^n are independent. In order to make the problem non-trivial, we assume that [2]:

$$\begin{aligned}\mathcal{D}(P_{XY}\|Q_{XY}) &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{Q_{XY}(x,y)} \\ &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)\cdot P_Y(y)} \\ &= I(X;Y) > 0, \end{aligned} \tag{3.2}$$

where $\mathcal{D}(\cdot\|\cdot)$ denotes the KL divergence between two probabilities and $I(X;Y)$ is the mutual information [2] between X and Y. As presented in expression (3.2), the mutual information is the divergence between the joint distribution P_{XY} and the product of its marginals, i.e., $P_X \cdot P_Y$.

Without communication constraints, the fusion center needs to decide about the true underlying hypothesis (H_0 or H_1) based on an observation of the joint vector (X_1^n, Y_1^n). Here we introduce a decentralized version of this problem which is illustrated in Fig. 3.1. In this distributed context, the decision rule is composed by a pair of encoder and decoder (f_n, ϕ_n) of length n and rate R (in bits per sample), where:

$$\begin{aligned} &f_n : \mathbb{X}^n \to \{1,\dots,2^{nR}\},\ \text{(encoder)} \\ &\phi_n : \{1,\dots,2^{nR}\}\times\mathbb{Y}^n \to \Theta = \{0,1\},\ \text{(decoder)}. \end{aligned} \tag{3.3}$$

$f_n(\cdot)$ models a fixed-rate lossy encoder (or quantizer) of X_1^n and $\phi_n(\cdot)$ represents the detector (or classifier) acting on the one-sided compressed data $(f_n(X_1^n), Y_1^n) \in \{1,...,2^{nR}\}\times\mathbb{Y}^n$. The encoder represents a remote agent that senses X_1^n and transmit a finite description (using R bits per sample) of X_1^n to a fusion center (see Fig.3.1). The fusion center receives the quantization of X_1^n and at the same time senses locally a second modality Y_1^n to guess (using $\phi_n(\cdot)$) about the true distribution of the joint vector (X_1^n, Y_1^n). For any pair (f_n, ϕ_n) of length n and rate R, we introduce the corresponding Type I and Type II error probabilities [24], [49]:

$$P_0(f_n,\phi_n) \triangleq P_{XY}^n(\mathcal{A}^c(f_n,\phi_n)) \text{ and} \tag{3.4}$$

$$P_1(f_n,\phi_n) \triangleq Q_{XY}^n(\mathcal{A}(f_n,\phi_n)), \tag{3.5}$$

where $A(f_n,\phi_n) \triangleq \{(x_1^n, y_1^n) \in \mathbb{X}^n\times\mathbb{Y}^n : \phi_n(f_n(x_1^n), y_1^n) = 0\}$. Traditionally, for any $\epsilon > 0$, we are interested in the family of optimal encoder-decoder pairs satisfying:

$$\beta_n(\epsilon,R) \triangleq \min_{(f_n,\phi_n)} \{P_1(f_n,\phi_n) : P_0(f_n,\phi_n) \le \epsilon\}, \tag{3.6}$$

where the minimum is over all encoding-decoder pairs in (3.3). It is worth to mention that expression $\beta_n(\epsilon, R)$ (i.e., the optimization in (3.6)) is an explicit function of the underlying model P_{XY}. Consequently, in the analysis and results presented through this chapter, P_{XY} is assumed to be known.

We study the performance of the optimal scheme (3.6) by focusing on the case where a sequence of restrictions $(\epsilon_n)_{n\ge1}$ is required to tend to zero as the sample size grows. The

objective is to explore how this restriction is expressed in terms of $(\beta_n(\epsilon_n, R))_n$ with n in conjunction with other properties of the problem (e.g., the distribution P_{XY} and the rate R). In this work, we are primarily interested in deriving expressions to bound $\beta_n(\epsilon_n, R)$ in the large sample regime (non-asymptotic analysis). To this end, it would be essential to first characterize the asymptotic nature of the sequence $(\beta_n(\epsilon_n, R))_n$.

Before presenting the main contributions of this chapter, we review some essential asymptotic results for the classical communication-free (centralized) scenario.

3.2.1 Review of centralized HT results

For completeness, it is worth revisiting the centralized case where $f_n : \mathbb{X}^n \to \mathbb{X}^n$ is the identity and the solution of (3.6) is then denoted by $\beta_n(\epsilon_n)$. Furthermore, when $\epsilon_n = \epsilon > 0$ for all n, this is a classical HT setting where the celebrated *Stein's Lemma* implies the following result [2, 30]:

Lemma 3.2.1 (*Stein's Lemma*) For any $\epsilon \in (0, 1)$,

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_n(\epsilon) = \mathcal{D}(P\|Q).$$

This result establishes the asymptotic decayment of the TYPE II error subject to a fixed $\epsilon > 0$ implying that $\beta_n(\epsilon) \sim e^{-n\mathcal{D}(P\|Q)}$ as n tends to infinity (large sampling regime). Interestingly, in [68], we provided upper and lower bounds for $\beta_n(\epsilon)$ in the finite length regime showing that in practice the number of samples required to approximate the TYPE II error probability to $(e^{-n\mathcal{D}(P\|Q)})$ is not large. This observation supports the claim that the exponential approximation is a useful proxy for TYPE II error probability.

For the sub-exponential regime of $(\epsilon_n)_n$, the following result is known.

Lemma 3.2.2 ([44, Sect. IX]) if $(1/\epsilon_n)$ is $o(e^{rn})$ for any $r > 0$ then $\lim_{n\to\infty} -\frac{1}{n} \log \beta_n(\epsilon_n) = \mathcal{D}(P\|Q)$.

Therefore, the error exponent obtained with a fixed $\epsilon > 0$ in Lemma 3.2.1 is preserved for a family of stringer decision problems in (3.6) as long as $(\epsilon_n)_n$ tends to zero at a sub-exponential rate.

3.2.2 Review of distributed HT results

Returning to the main decentralized task with communication constraints in Fig.3.1, [28] determined the following result[1]:

[1]This result can be interpreted as the counterpart of the Stein's Lemma in the decentralized setting of Fig.3.1.

Lemma 3.2.3 [28, Theorem 3] For any $\epsilon > 0$, it follows that[2]

$$\xi(R) \triangleq \lim_{n\to\infty} -\frac{1}{n}\log\beta_n(\epsilon, R) = \max_{\substack{U: U \multimap X \multimap Y \\ I(U;X)\leq R \ |\mathbb{U}|\leq|\mathbb{X}|+1}} I(U;Y), \tag{3.7}$$

where $U \multimap X \multimap Y$ denotes the fact that (U, X, Y) forms a Markov chain (i.e., (U, Y) are independent conditioned to X).

The optimization presented in (3.7) is a trade-off between representation and regularization, in the sense that we seek to learn the best possible representation of X for predicting Y. As for the more challenging scenario where $(\epsilon_n)_n$ tends to zero with n, in [33] the author provided a lower bound for the error exponent of the TYPE II error probability in the case of exponentially decreasing TYPE I error restrictions:

Lemma 3.2.4 [52, Han and Kobayashi] Let us assume that $\epsilon_n \leq e^{-rn}$ for some $r > 0$, then: $\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n(\epsilon_n, R) \geq$

$$\max_{w\in\rho(R,r)} \min_{\substack{\tilde{P}_{UXY} \\ \mathcal{D}(\tilde{P}_{UXY}\|P_{UXY})\leq r \\ \tilde{P}_{U|X}=P_{U|X}=w \\ U \multimap X \multimap Y}} [\mathcal{D}(\tilde{P}_X\|P_X) + I(U;Y)], \tag{3.8}$$

$$\rho(R, r) \triangleq \{w \in \mathcal{P}(\mathbb{U}|\mathbb{X})| \max_{\substack{\tilde{P}_X: \mathcal{D}(\tilde{P}_X\|Q_X)\leq r \\ \tilde{P}_{U|X}=w \\ P_{UX}=w\cdot\tilde{P}_X}} I(U;X) \leq R\},$$

where $\mathcal{P}(\mathbb{U}|\mathbb{X})$ denotes all test channels (or conditional probabilities) from $\mathbb{X}$ to $\mathbb{U}$.

3.3 Asymptotic Result

Our first result complements the regime on $(\epsilon_n)_n$ presented in Lemma 3.2.4 to obtain an asymptotic characterization of $(\beta_n(\epsilon_n, R))_n$. In particular, we explore the important sub-exponential regime for the restriction sequence $(\epsilon_n)_n$ of TYPE I error probability.

Theorem 3.1 *Let us assume that $(1/\epsilon_n)_n = o(e^{rn})$ for any $r > 0$. Then,*

$$\lim_{n\to\infty} -\frac{1}{n}\log(\beta_n(\epsilon_n, R)) = \xi(R), \tag{3.9}$$

where $\xi(R)$ is defined in (3.7). (The proof is presented in Appendix 3.7.)

This result establishes an extensive regime on the speed at which $(\epsilon_n)_n$ tends to zero for which the error exponent of the problem is invariant and matches the expression obtained

[2]This result provides an interesting connection with the problem noisy lossy source coding with log-loss fidelity [70]. The performance limits in the right hand side (RHS) of (4.16) coincides precisely with the distortion-rate function of the information bottleneck problem [71].

for the less restrictive and classical setting ($\epsilon_n = \epsilon > 0$) presented in Lemma 3.2.3. This result is interesting because, as it was pointed out in [52], there was no guarantee that the asymptotic limit in (3.9) remains the same as the result in Lemma 3.2.3 when moving to stringer regimes on the speed at which $(\epsilon_n)_n$ vanishes with n. Besides this, the above result can be considered as being the counterpart of what is observed in the centralized setup when contrasting Lemmas 3.2.2 and 3.2.1.

The proof of Theorem 3.1, in Appendix 3.7.1, is divided into two parts. The direct part (i.e., constructive argument) is based on constructing an encoder-decision pair that guarantees that the error exponent of the optimal TYPE II is greater than $\xi(R)$. The second part of the argument (i.e., the infeasibility part) proves that no pair of encoder-decoder rule satisfying the restriction of the TYPE I error has an error exponent greater than $\xi(R)$. The proof argument used in both the achievable and infeasibility parts (see Appendix 3.7.1) is based on a refined use of concentration inequalities [43]. In particular, the achievable part is divided into two steps. The first step consists of reducing the problem to an i.i.d. structure over a block of X_1^n induced by the encoder, which will concentrate (in probability) to an error exponent that is different from $\xi(R)$ in (3.9). Importantly, the discrepancy between the concentration limit obtained from our approach (i.e., finite-block strategy) and $\xi(R)$ can be resolved analytically by connecting our problem with a noisy rate-distortion problem, where the discrepancy between its fundamental limit and a finite length version of this object is well understood [45]. The second step consists of optimizing our approach by giving concrete conditions to make the discrepancy between $\xi(R)$ and $-\frac{1}{n}\log(\beta_n(\epsilon_n, R))$ vanishes with n.

3.4 Finite-length Result

Our main result is concerned with the practically relevant task of offering a non-asymptotic characterization of the sequence $(\beta_n(\epsilon_n, R))_n$ for different scenarios of $(\epsilon_n)_n$, given the model P_{XY} and the rate constraint $R > 0$. To address this question, our methodology uses the asymptotic limit of $(\beta_n(\epsilon_n, R))_n$, stated in Theorem 3.1, and from this, analyzes the discrepancy between $-\frac{1}{n}\log\beta_n(\epsilon_n, R)$ and $\xi(R)$ as a function of n. In concrete, our main result (stated below) derives upper and lower bounds for $-\frac{1}{n}\log\beta_n(\epsilon_n, R)$ in different sub-exponential scenarios for the TYPE I restriction sequence $(\epsilon_n)_n$. As a corollary, we determine the speed at which $-\frac{1}{n}\log\beta_n(\epsilon_n, R)$ achieves its limit in (3.9). The proof of this result is presented in Appendix 3.7.2.

Theorem 3.2 *Let us assume that $R < H(X)$ and define*

$$C(P_{XY}) \triangleq \sup_{(x,y)\in\mathbb{X}\times\mathbb{Y}} \left| \log\left(\frac{P_{XY}(\{(x,y)\})}{Q_{XY}(\{(x,y)\})}\right)\right| < \infty. \tag{3.10}$$

Then, we have the following results

i) If $(\epsilon_n)_n = (1/\log(n))_n$ (logarithmic), it follows:

$$-\frac{1}{n}\log(\beta_n(\epsilon_n, R)) - \xi(R) \geq \left(\frac{dD(R)}{6dR} - \frac{\sqrt{2\ln(\log(n))}C(P_{XY})}{\log(n)} - o(1)\right)\frac{\log n}{n^{1/3}} \tag{3.11}$$

$$-\frac{1}{n}\log(\beta_n(\epsilon_n,R))-\xi(R)\leq\left(16C(P_{XY})+\frac{\log(\log(n))\sqrt{\log(n)}}{n}\right)\frac{1}{\sqrt{\log(n)}};\quad(3.12)$$

ii) If $(\epsilon_n)_n=(1/n^p)_n$ *(polynomial) with* $2>p>0$*, then*

$$-\frac{1}{n}\log(\beta_n(\epsilon_n,R))-\xi(R)\geq\left(\frac{1}{6}\frac{dD(R)}{dR}-\frac{\sqrt{2p\ln(n)}}{\log n}C(P_{XY})-o\left(1\right)\right)\frac{\log n}{n^{1/3}}\quad(3.13)$$

$$-\frac{1}{n}\log(\beta_n(\epsilon_n,R))-\xi(R)\leq\left(16C(P_{XY})+\frac{p\log(n)}{n^{1-p/2}}\right)\frac{1}{n^{p/2}};\quad(3.14)$$

iii) If $(\epsilon_n)_n=(1/n^p)_n$ *(polynomial) with* $p\geq 2$*, then*[3]

$$-\frac{1}{n}\log(\beta_n(\epsilon_n,R))-\xi(R)\geq\left(\frac{1}{6}\frac{dD(R)}{dR}-\frac{\sqrt{2p\ln(n)}}{\log n}C(P_{XY})-o\left(1\right)\right)\frac{\log n}{n^{1/3}}\quad(3.15)$$

$$-\frac{1}{n}\log(\beta_n(\epsilon_n,R))-\xi(R)\leq\left(8\sqrt{2}C(P_{XY})\frac{\sqrt{n^{2-p}+1}}{\log(n)}+2\right)\frac{\log(n)}{n};\quad(3.16)$$

iv) If $(\epsilon_n)_n=(1/e^{n^p})_n$ *(superpolynomial) with* $p\in(0,1)$*,*

$$-\frac{1}{n}\log(\beta_n(\epsilon_n,R))-\xi(R)\geq\left(\frac{(1-p)}{6}\frac{dD(R)}{dR}-\frac{\sqrt{2}C(P_{XY})}{\log(n)}-o\left(1\right)\right)\frac{\log n}{n^{(1-p)/3}}\quad(3.17)$$

$$-\frac{1}{n}\log(\beta_n(\epsilon_n,R))-\xi(R)\leq\left(8\sqrt{2}C(P_{XY})\frac{\sqrt{e^{-n^p}n^2+1}}{\log(n)}+2\right)\frac{\log(n)}{n}.\quad(3.18)$$

$D(R)$ *is the noisy distortion-rate function [2].*

3.4.1 Discussion of Theorem 3.2

(i) The results establish non-asymptotic bounds for the TYPE II error when we impose concrete scenarios for the monotonic behavior of $(\epsilon_n)_n$. We explore three main regimes for $(\epsilon_n)_n$: logarithmic, polynomial, super-polynomial. Each of these cases has its corresponding lower and upper bounds, which depends specifically on the considered $(\epsilon_n)_n$.

(ii) The proof of Theorem 3.2 involves an optimization problem of the upper and lower bounds presented in the proof of Theorem 3.1, for which the arguments used to prove Theorem 3.1 were instrumental for this analysis. Specifically, we refine the analysis introduced in (3.43), (3.45) and (3.55) by finding optimal values for l and s_n for a given ϵ_n. These choices of values for l and s_n give us non-asymptotic lower and upper bounds for $-\frac{1}{n}\log(\beta_n(\epsilon_n,R))$, for each scenario.

(iii) Regarding the upper bound of $-\frac{1}{n}\log(\beta_n(\epsilon_n,R))$ ((3.12), (3.14), (3.16) and (3.18)), obtained from the impossibility argument (converse part), as $(\epsilon_n)_n$ goes to zero faster (from

[3]It is worth to mention that there is a discrepancy in the constant $(\sqrt{2}C(P_{XY})$ used in the upper bounds in (3.16) and (3.18) with respect to the ones appearing in a preliminary version of this work in [54] which was obtained under stringer additional (implicit) assumptions.

case to case), the speed at which the bound tends to zero increases; from the slower rate $\mathcal{O}\left(1/\sqrt{\log(n)}\right)$ to the faster that is $\mathcal{O}\left(\log(n)/n\right)$. Therefore, by imposing a more restrictive $(\epsilon_n)_n$ there is an effect in the discrepancy between the fundamental limit $\xi(R)$ and the optimal TYPE II error $-\frac{1}{n}\log\beta_n(\epsilon_n, R)$ obtained from this upper bound analysis.

(iv) Regarding the lower bound of $-\frac{1}{n}\log\beta_n(\epsilon_n, R)$ ((3.11), (3.13), (3.15) and (3.17)), obtained from the direct argument (achievability part), as $(\epsilon_n)_n$ goes faster to zero (from case to case), the derived bound -for the super-polynomial case- decreases in the speed at which the discrepancy in error exponent (i.e., $-\frac{1}{n}\log(\beta_n(\epsilon_n, R))-\xi(R)$) tends to zero. For the other two cases (logarithmic and polynomial), the speed is not affected, but the constants change to slower magnitudes. These trends are consistent with the observation that by relaxing the speed of $(\epsilon_n)_n$ the decision problem is less restrictive and then, the result favors the possibility of obtaining a better TYPE II error (smaller) than the one predicted by the asymptotic limit, which is $e^{-n\xi(R)}$.

(v) Finally, it is worth noting that if we consider the relaxed restriction $\epsilon_n = \epsilon \in (0, 1)$ in Lemma 3.2.3, the achievability part of our argument still works and for $\xi(R) - \left(-\frac{1}{n}\log\beta_n(\epsilon, R)\right)$ it offers an upper bound that converges to zero as $\mathcal{O}\left(\frac{\log(n)}{n^{1/3}}\right)$.

This last speed of convergence is slower than the same result known for the unconstrained (centralized) problem presented in [27]. In fact, when X_1^n is fully observed at the detector (see Lemma 3.2.1), in [27] the author showed that the discrepancy $\left|\mathcal{D}(P\|Q) - \left(-\frac{1}{n}\log\beta_n(\epsilon)\right)\right|$ tends to zero as $\mathcal{O}\left(1/\sqrt{n}\right)$.[4]. We conjecture that our slower rate can be attributed to the non-trivial role of the communication constraint in our problem, which breaks the i.i.d. structure of X_1^n in a way that it is not possible to use the tools adopted to derive the unconstrained result in Lemma 3.7.3. It is a topic of further research to uncover if the upper bound $\mathcal{O}\left(\frac{\log(n)}{n^{1/3}}\right)$ for the discrepancy $\xi(R) - \left(-\frac{1}{n}\log\beta_n(\epsilon, R)\right)$ can be improved, or if it is possible to show (by a converse argument) that this rate is indeed optimal provided that $\epsilon_n = \epsilon > 0$.

3.4.2 Interpretation of Theorem 3.2

In general, Theorem 3.2 can be presented as two bounds:

$$\xi_o - f(n) \leq -\frac{1}{n}\log\beta_n \leq \xi_o + g(n), \tag{3.19}$$

where β_n is the optimal TYPE II error consistent with TYPE I error restriction (ϵ_n in the statement of Theorem 3.2), ξ_o is the performance limit (in Theorem 3.1), $f(n)$ is a positive sequences that goes to zero with n ($o(1)$) representing the penalization (in error exponent) for the use of finite simple-size, and $g(n)$ is a positive sequence that goes to zero representing a discrepancy with the limit but that can be seen as a gain in error exponent. Then, we have a feasibility range for β_n given by the interval:

$$[\exp[-n(\xi_o + g(n))], \exp[-n(\xi_o - f(n))]].$$

This interval contains the nominal value $e^{-n\xi_o}$, which is consistent with the error exponent limit in Theorem 3.1 but extrapolated to a finite length regime. If we consider $\exp(-n\xi_o)$ as

[4]For completeness, this is presented in Lemma 3.7.3 in Appendix 3.7.4.

our reference, we can study two feasible regions: the pessimistic interval

$$(\exp(-n\xi_o), \exp(-n(\xi_o - f(n)))]$$

where the error probability is greater than the nominal value $e^{-n\xi_o}$, and the optimistic interval

$$[\exp(-n(\xi_o + g(n))), \exp(-n\xi_o)]$$

where the appositive occurs. The length of the interval of the two regions is an indicator of the precision of our result (the worse case discrepancy with respect to $e^{-n\xi_o}$). For the pessimistic region, the length of that interval is $e^{-n\xi_o}(e^{nf(n)} - 1)$. From the fact that $f(n)$ is $o(1)$ (see the statement of Theorem 3.2), the length of this interval tends to zero strictly faster than $\mathcal{O}(e^{-n(\xi_o - \epsilon)})$ for any $\epsilon > 0$ and, consequently, the precision has an exponential rate of convergence that is asymptotically given by the nominal exponent $\xi_o > 0$. On the optimistic region, the length of this interval is $e^{-n\xi_o}(1 - e^{-ng(n)})$, which is $\mathcal{O}(e^{-n\xi_o})$. Overall, the length of the pessimistic interval dominates the analysis and, consequently, the precision of the result (i.e., the worse case discrepancy with respect to the nominal $e^{-n\xi_o}$) tends to zero as $\mathcal{O}(e^{-n(\xi_o - f(n))})$. This order is equivalent to the worse-case TYPE II error probability $(e^{-n(\xi_o - f(n))})$ predicted from Theorem 3.2.

In conclusion, the overall quality of the result is governed by ξ_o and affected in a smaller degree by how fast $f(n)$ goes to zero. Note that $g(n)$ plays no role from this perspective. We discussed on the previous section that $f(n)$ goes faster to zero when we relax the problem (i.e., passing from a scenario for $(\epsilon_n)_n$ to a scenario where this sequence tends to zero at a smaller speed). Then, the precision of Theorem 3.2 improves when simplifying the problem from one restriction $(\epsilon_n)_n$ to a relaxed restriction $(\tilde{\epsilon}_n)_n$ for the TYPE I error. This reinforces one of the points mentioned in Section 3.4.1, where we discussed that the speed at which $(\epsilon_n)_n$ goes to zero does not affect the limit ξ_o (Theorem 3.1) but it does affect our finite length result through $f(n)$.

3.5 Application Examples

In this section, we present some empirical evidences illustrating the possible implication of Theorem 3.2 to effectively bound $\beta_n(\epsilon_n, R)$ with finite-sample size n. Theorem 3.2 offers an interval of feasibility for $\beta_n(\epsilon_n, R)$ expressed by

$$\mathrm{UB}(\epsilon_n, R) = \exp\left[-n\left(\xi(R) + \frac{dD(R)}{dR}\frac{\log(l)}{2l} - \sqrt{\frac{2l\ln(1/\epsilon_n)}{n}}C(P_{XY})\right)\right], \tag{3.20}$$

$$\mathrm{LB}(\epsilon_n, R) = \exp\left[-n\left(\xi(R) + 4C(P_{XY}) \cdot \sqrt{2\ln\left(\frac{1}{1-\epsilon_n - h_n(s)}\right)} + \frac{\log(1/h_n(s))}{n}\right)\right], \tag{3.21}$$

where $\beta_n(\epsilon_n, R) \in [\mathrm{LB}(\epsilon_n, R), \mathrm{UB}(\epsilon_n, R)]$.[5] The length of $[\mathrm{LB}(\epsilon_n, R), \mathrm{UB}(\epsilon_n, R)]$ indicates the precision of our approximation and the interval itself can be used to evaluate how representative is $e^{-n\xi(R)}$ of $\beta_n(\epsilon_n, R)$ for a finite n.

[5] l and $h_n(s)$ are obtained according to the proof of Theorems 3.1 and 3.2 (see Appendix 3.7 and 3.7.2 for details).

We first evaluate the length of $[\mathrm{LB}(\epsilon_n, R), \mathrm{UB}(\epsilon_n, R)]$ by considering four cases $(\epsilon_n)_n \in \{0.01, 1/\log(n), n^{-0.01}, n^{-0.1}\}$ associated to a constant, a logarithmic and a polynomial TYPE I error restriction, respectively. We use a discretized version of a Gaussian pdf P_{XY} of $|\mathbb{X}|\times|\mathbb{Y}|$ where the mutual information between the two variables (X and Y) is 7 and 1.5 nats, respectively. To compute the expressions in (3.20) and (3.21), we need to evaluate $\xi(R)$. Obtaining $\xi(R)$ involves an optimization problem with respect to the encoder f_n and the rate R [71]. To this end, we use the algorithm in [72] which is a generalization of *Blahut-Arimoto algorithm* [73].[6]

	Number of observations n					
ϵ_n	50	250	350	550	650	750
$1/\log(n)$	1.2138e-12	3.3636e-62	3.0109e-87	1.4286e-137	8.2535e-163	4.3764e-188
$1/n^{0.01}$	6.4432e-10	4.2359e-52	4.5381e-74	8.5307e-119	1.9894e-141	3.4117e-164
$1/n^{0.1}$	0.0045	2.5598e-28	2.0497e-43	8.5949e-76	1.0006e-92	4.9903e-110

Table 3.1: Magnitude of $\mathrm{UB}(\epsilon_n)-\mathrm{LB}(\epsilon_n)$ function of ϵ_n and n for the case when $I(X;Y) = 1.5$ nats and $R = 2$ bits.

Table 3.1 shows the lengths of $[\mathrm{LB}(\epsilon_n, R), \mathrm{UB}(\epsilon_n, R)]$. We verify that $\mathrm{UB}(\epsilon_n, R)-\mathrm{LB}(\epsilon_n, R)$ tends to zero exponentially fast with the sample size as observed in Section 3.4.2. From this exponential decay, the nominal value predicted by Theorem 3.1, i.e., $\exp(-n\xi(R))$, is a very precise approximation of $\beta_n(\epsilon_n, R)$ provided that n is sufficiently large. This support the idea that $e^{-n\xi(R)}$ is an excellent proxy of $\beta_n(\epsilon_n, R)$ if a critical number of samples is achieved. Table 3.1 also shows that the precision of the result measured by $(\mathrm{UB}(\epsilon_n, R)-\mathrm{LB}(\epsilon_n, R))$ is affected by the speed at which the TYPE I error sequence tends to zero, which is consistent with our previous analysis in Section 3.4.2. In particular, we observe that for a faster convergence rate of $(\epsilon_n)_n$, i.e, a stringer distributed decision problem, the length of $[\mathrm{LB}(\epsilon_n, R), \mathrm{UB}(\epsilon_n, R)]$ is bigger, which means that our bounds are expected to be less informative on $\beta_n(\epsilon_n, R)$ when compared with a relaxed scenario.

The results presented in Table 3.1 support the claim that $\exp(-n\xi(R))$ can be adopted as practical proxy to $\beta_n(\epsilon_n, R)$. To formalize this, we address the following question: for a given arbitrary small $\delta > 0$ of the form 10^{-k} with $k \in \{1, .., 5\}$ and a joint model P_{XY}, we seek to find the lowest n such that $\beta_n(\epsilon_n, R) \in (e^{-n\xi(R)} - \delta, e^{-n\xi(R)} + \delta)$. The exponential decay of the length of $[\mathrm{LB}(\epsilon_n, R), \mathrm{UB}(\epsilon_n, R)]$, observed in Table 3.1, suggests that this condition happens eventually with n very quickly. Importantly, we can derive an upper bound for this Critical Number of Samples (CNS) from the closed-form expressions we have for $\mathrm{LB}(\epsilon_n, R)$ and $\mathrm{UB}(\epsilon_n, R)$.[7] Figs. 3.2 and 3.3 present the predicted CNS vs. $\delta = 10^{-k}$ for different scenarios of P_{XY} (in terms of the magnitude of $I(X;Y)$) and $(\epsilon_n)_n$. We consider two scenarios for P_{XY} ($I(X;Y) = 7$, $R = 4$ and $I(X;Y) = 1.5$ with $R = 2$) and we explore $(\epsilon_n)_n \in \{n^{-0.01}, n^{-0.1}, 1/\log(n), 0.1\}$. Figs. 3.2 and 3.3 show that even for a very small precision $\delta = 10^{-5}$, the point at which $\beta_n(\epsilon_n, R)$ is well approximated by $e^{-n\xi(R)}$ happens with less than 22 samples for the high-rate restriction case and in less than 80 samples for the low

[6] Importantly, under some mild conditions given in [72], this optimization (algorithm) converges to $\xi(R)$.

[7] The predicted CNS is the first $n \geq 1$ such that $\max\{\mathrm{UB}(\epsilon_n, R) - e^{-n\xi(R)}, e^{-n\xi(R)} - \mathrm{LB}(\epsilon_n, R)\} \leq \delta$, which is finite for any $\delta > 0$ and can be computed from our result.

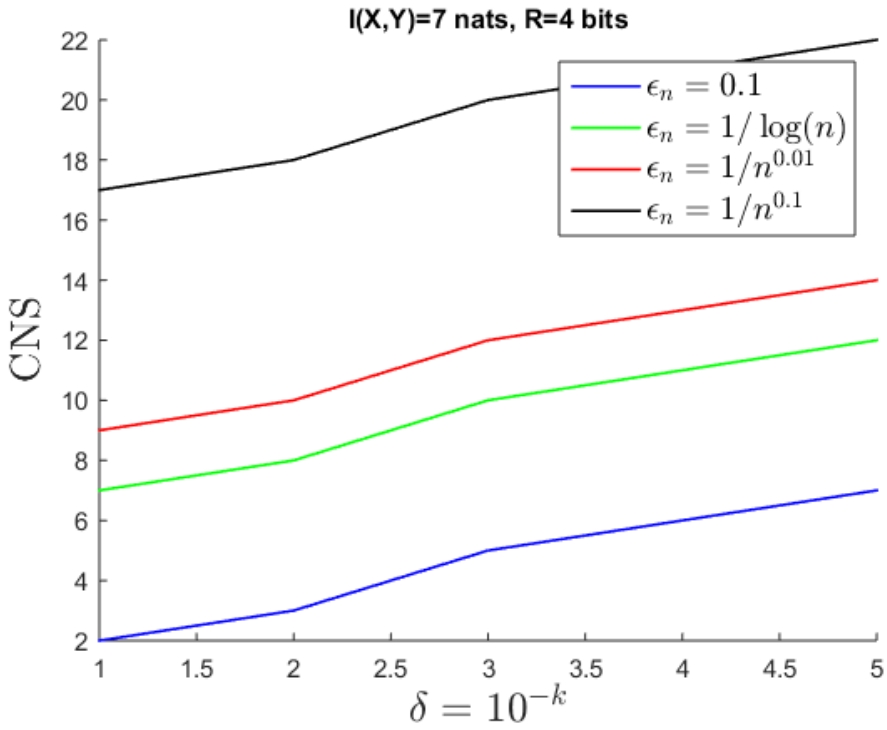


Figure 3.2: Critical Number of Samples (CNS) predicted by Theorem 3.2 across different values of $\delta = 10^{-k}$. The values used are $\xi(R) = 3$, $I(X;Y) = 7$, $R = 4$ and $C_X(P,Q) = 2.47$.

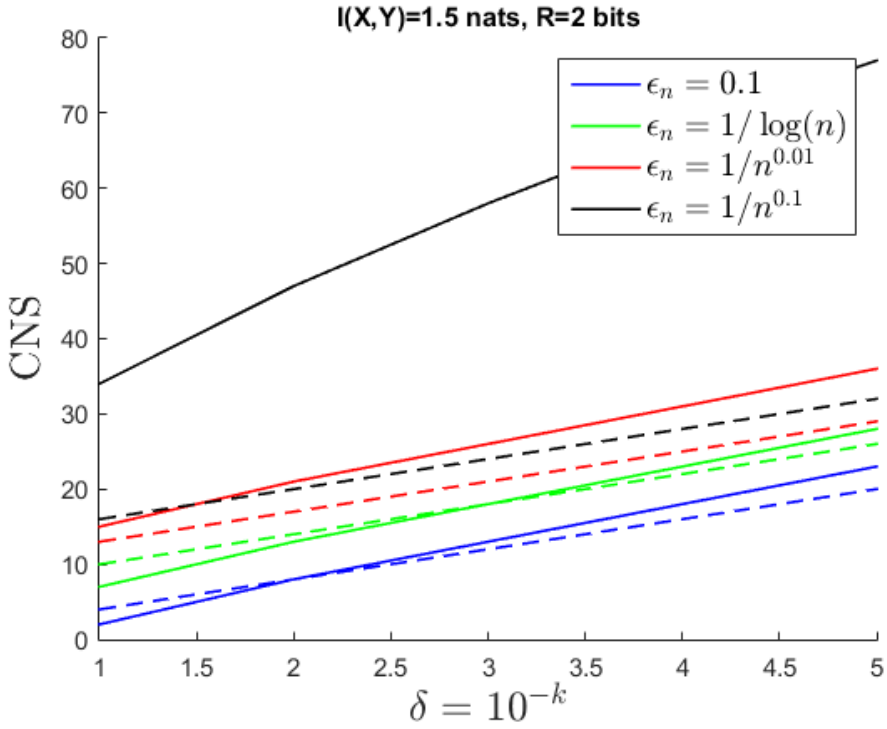


Figure 3.3: CNS predicted by Theorem 3.2 across different values of $\delta = 10^{-k}$. Low rate case with $\xi(R) = 0.7$, $I(X;Y) = 1.5$, $R = 2$ and $C_X(P,Q) = 1.92$. The dashed lines show an estimation of the exact CNS obtained from $\beta_n(\epsilon_n, R)$.

rate case for the majority of $(\epsilon_n)_n$.[8] The dependency of these values (predicted CNS from Theorem 3.2) on the magnitude of $I(X;Y)$ and $(\epsilon_n)_n$ is clearly expressed, which is consistent with our previous analyses in Section 3.4.2.

Finally, to evaluate the tightness of our theoretical bounds for the CNS, we simulate data from the true model P_{XY} (i.i.d. samples) to have a practical lower bound for $\beta_n(\epsilon_n)$. In particular, given P_{XY}, R and $(\epsilon_n)_n$, we obtained empirical estimations of the two error probabilities from which we estimate $\beta_n(\epsilon_n, R)$. $2.5 \cdot 10^6$ realizations of P_{XY} were used to obtain good estimations of these probabilities.[9] Using the estimated values of $\beta_n(\epsilon_n, R)$, we obtained for each $\delta > 0$ the corresponding CNS where the condition $\beta_n(\epsilon_n, R) \in (e^{-n\xi(R)} - \delta, e^{-n\xi(R)} + \delta)$ is meet directly (the empirical estimations of the CNS). Fig. 3.3 contrasts our predictions (theoretical upper bounds) with the empirical estimations (the dashed lines) of the CNS. Consistent with the nature of our result, the predicted CNS values are more conservative than the CNS estimated from simulations. This discrepancy is not significant overall, in particular for the regime when ϵ_n exhibits a relatively small speed of convergence to zero. Overall, we can conclude that the derived bounds are meaningful and can be adopted in cases where it is prohibitive to estimate $\beta_n(\epsilon_n, R)$ from data. Indeed, we face this issue in this analysis, as it was not possible to estimate $\beta_n(\epsilon_n, R)$ for the higher rate cases.[10]

3.6 Summary and Discussion

This thesis explores the problem of testing against independence with one-sided communication constraints. More specifically, the scenario of two memory-less sources is considered where one of the modalities is transmitted to the decision-maker (fusion center) over a rate-limited channel. In this context, we explored a general family of optimal tests (in the sense of *Neyman-Pearson*) where restrictions on the TYPE I error are imposed. We are interested in the speed at which the TYPE II error vanishes with the sample size. From a theoretical perspective, we obtained the performance limits for a rich family of problems with a decreasing sequence of TYPE I error probabilities (Theorem 3.1). This result establishes that the error exponent of the TYPE II error probability tends to an error exponent (fundamental limit) in the form of the classical *Stein's Lemma*. This error exponent is expressed in a closed-form, which is a function of the operational rate (in bits per sample) imposed on one of the information sources. Interestingly, this result implies that for a large family of TYPE I error restrictions (vanishing to zero with the sample size), the error exponent coincides with the result obtained in the (classical) scenario where the TYPE I error restriction is constant with n (Lemma 3.2.3).

Concerning the finite-sample size analysis, our main result (Theorem 3.2) provides bounds for the TYPE II error probability. Using results from rate-distortion theory and concentration inequalities, we obtained upper and lower bounds for this error as a function of n (the sample size), the sequence $(\epsilon_n)_n$ that models the restriction for the TYPE I error proba-

[8] The observed variations can be attributed to the value of $\frac{dD(R)}{dR}$, which tends to zero as long as $R > H(X)$.

[9] To achieve this, we use an scalar quantization based on the *Lloyd-max algorithm* [74] to obtain an induced quantized distribution $P_{f(X_1^n)Y}$.

[10] $\beta_n(\epsilon_n, R)$ is of order $O(e^{-n\xi(R)})$ so when R is relatively high, the value of $\xi(R)$ tends to $I(X;Y)$ for which $e^{nI(X;Y)}$ simulations are needed. This number becomes prohibitive, even for n of order of 30 when $I(X;Y) = 1.5$.

bility and the underlying distributions. We observed that the bounds offer an interval of feasibility for the optimal TYPE II error probability, which presents an accurate description. A closed-form expression for the worse-case TYPE II error probability was derived where a discrepancy in the error exponent (with respect to the asymptotic exponent) was identified. This discrepancy (overhead) can be attributed to using a finite number of observations in the decision. Furthermore, this penalization vanishes at a speed that is a function of $(\epsilon_n)_n$, and consequently, we observed the effect of the TYPE I error restriction in this non-asymptotic analysis.

We observed that the TYPE II error probability is arbitrary close (with n) to the nominal value predicted by the asymptotic result $e^{-n\xi(R)}$, where $\xi(R)$ is the limit in Theorem 3.1. Furthermore, the precision in Theorem 3.2, measured by the length of the feasible interval, tends to zero exponentially fast. Numerical analysis in some concrete scenarios confirms the predicted quality of the non-asymptotic results in Theorem 3.2.

3.6.1 Future Work

A relevant topic to be further explored is extending the results presented in this thesis to the problem of arbitrary binary hypothesis testing subject to communications constraints. However, a single-letter characterization of the error exponent of the TYPE II error is not available for the general setup and only a lower bound to it was derived in [33]. The characterization of this fundamental limit would be essential to be able to extend our results to the non-asymptotic analysis since a critical step was analyzing the discrepancy between the non-asymptotic and its corresponding asymptotic expression.

3.7 Appendix

3.7.1 Proof of Theorem 3.1:

The proof is divided in two parts: a lower and an upper bound result. We begin with the following bound that extends the result presented in [28, Theorem 3].

Theorem 3.3 *Let us assume that $\epsilon_n > 0$ for all n and $(1/\epsilon_n)_n = o(e^{rn})$ for any $r > 0$, then*

$$\liminf_{n\to\infty} -\frac{1}{n}\log(\beta_n(\epsilon_n, R)) \geq \xi(R). \tag{3.22}$$

PROOF. For an arbitrary encoder $f_n : \mathbb{X}^n \mapsto \{1, \ldots, 2^{nR}\}$ of rate $R > 0$, let us consider the corresponding optimal decision regions -according to Neyman-Pearson's Lemma- on the one-sided quantized space $\{1, \ldots, 2^{nR}\} \times \mathbb{Y}^n$ expressed by

$$\mathcal{B}_{n,t}(f_n) \triangleq \left\{(z, y_1^n) \in \{1, \ldots, 2^{nR}\} \times \mathbb{Y}^n : \frac{P_{f_n(X_1^n)Y_1^n}(z, y_1^n)}{Q_{f_n(X_1^n)Y_1^n}(z, y_1^n)} > e^{nt}\right\}. \tag{3.23}$$

$\mathcal{B}_{n,t}(f_n)$ is parametrized in terms of t, n and f_n. Let us denote by $\phi_{n,t}(\cdot) : \{1, \ldots, 2^{nR}\} \times \mathbb{Y}^n \mapsto \{0, 1\}$ the induced test (or decision rule) such that $\phi_{n,t}^{-1}(\{0\}) = \mathcal{B}_{n,t}(f_n)$. Then the TYPE I

error probability for the pair $(f_n, \phi_{n,t})$ is given by

$$P_0(f_n, \phi_{n,t}) = P_{f_n(X_1^n)Y_1^n}(\mathcal{B}_{n,t}^c(f_n)). \tag{3.24}$$

By construction of the pair $(f_n, \phi_{n,t})$, an upper bound for the TYPE II is obtained by

$$P_1(f_n, \phi_{n,t}) = Q_{f_n(X_1^n)Y_1^n}\left(\mathcal{B}_{n,t}(f_n)\right) \leq e^{-nt}. \tag{3.25}$$

Then, for any finite $n > 0$ and $\epsilon_n > 0$, finding an achievable TYPE II error exponent from this construction (and the bound in (3.25)) reduces to solve the following problem:

$$t_n^*(\epsilon_n) \triangleq \sup_{f_n \text{ encoder of rate } R} \sup_t \{t : P_{f_n(X_1^n)Y_1^n}(\mathcal{B}_{n,t}^c(f_n)) \leq \epsilon_n\}. \tag{3.26}$$

Note that f_n breaks the i.i.d. structure of the problem, then determining $t_n^*(\epsilon_n)$ is not a simple task. We will derive a lower bound for $t_n^*(\epsilon_n)$ using a finite block analysis approach. For this, let us consider a fixed $l \geq 1$ and let us consider $\tilde{f}_l$ an encoder of length l, i.e. $\tilde{f}_l : \mathbb{X}^l \to \{1, \ldots, 2^{lR}\}$. The idea is to decompose X_1^n in segments of finite length to use the induced block i.i.d. structure when n tends to infinity. More precisely, we construct an encoder that we denote by $\tilde{f}_{n,l}$ applying the function $\tilde{f}_l$ k-times to every sub-block of length l, assuming for the moment that $n = kl$, i.e.,

$$\begin{aligned}\tilde{f}_{n,l}(x_1, \ldots, x_l, x_{l+1}, \ldots, x_{2l}, \ldots, x_{l(k-1)+1}, \ldots, x_{kl}) \triangleq \\ (\tilde{f}_l(x_1, \ldots, x_l), \tilde{f}_l(x_{l+1}, \ldots, x_{2l}), \ldots, \tilde{f}_l(x_{l(k-1)+1}, \ldots, x_{kl})).\end{aligned} \tag{3.27}$$

In the use of the set $\mathcal{B}_{n,t}(\tilde{f}_{n,l})$ in (3.23), it will be convenient to parametrize t relative to the reference value $\frac{1}{l}\mathcal{D}(P_{\tilde{f}_l(X_1^l)Y_1^l} \| Q_{\tilde{f}_l(X_1^l)Y_1^l})$ that is a function of $\tilde{f}_l$. More precisely, let us define

$$t_\delta \triangleq \frac{1}{l}\mathcal{D}(P_{\tilde{f}_l(X_1^l)Y_1^l} \| Q_{\tilde{f}_l(X_1^l)Y_1^l}) - \delta,$$

for any $\delta > 0$. Using the l-block structure of $\tilde{f}_{n,l}$, the TYPE I error in (3.24) of the pair $(\tilde{f}_{n,l}, \phi_{n,t_\delta})$ can be expressed by:

$$P_{\tilde{f}_{n,l}(X_1^n)Y_1^n}\left(\mathcal{B}_{n,t_\delta}^c(\tilde{f}_{n,l})\right), \tag{3.28}$$

where $\mathcal{B}_{n,t_\delta}^c(\tilde{f}_{n,l}))$ has the elements $z_1^k, y_1^n \in \{1, \ldots, 2^{lR}\}^k \times \mathbb{Y}^n$ satisfying that

$$\left| \hat{\mathcal{D}}(P_{\tilde{f}_l(X_1^l)Y_1^l} \| Q_{\tilde{f}_l(X_1^l)Y_1^l}) - \mathcal{D}(P_{\tilde{f}_l(X_1^l)Y_1^l} \| Q_{\tilde{f}_l(X_1^l)Y_1^l}) \right| \geq l\delta, \tag{3.29}$$

where

$$\hat{\mathcal{D}}(P_{\tilde{f}_l(X_1^l)Y_1^l} \| Q_{\tilde{f}_l(X_1^l)Y_1^l}) \triangleq \frac{1}{k}\sum_{i=1}^k \log\left(\frac{P_{\tilde{f}_l(X_1^l)Y_1^l}(\{z_i, y_{k(i-1)+1}^{ki}\})}{Q_{\tilde{f}_l(X_1^l)Y_1^l}(\{z_i, y_{k(i-1)+1}^{ki}\})}\right)$$

denotes the empirical divergence. We will use a concentration inequality to bound the probability of the deviation event in (3.29). To this end, let us introduce the notation: $u_i = (z_i, y_{l(i-1)+1}, \ldots, y_{il}) \in \{1, \ldots, 2^{lR}\} \times \mathbb{Y}^l$ and

$$g(u_1, \ldots, u_i, \ldots, u_k) \triangleq \frac{1}{k}\sum_{j=1}^k \log\left(\frac{P_{\tilde{f}_l(X_1^l)Y_1^l}(\{u_j\})}{Q_{\tilde{f}_l(X_1^l)Y_1^l}(\{u_j\})}\right), \tag{3.30}$$

where it follows that for any $k > 0$ and $\forall i \in \{1, \ldots, k\}$:

$$\sup_{\substack{u_1,\ldots,u_i,\bar{u}_i,\ldots,u_k \\ \in \tilde{f}_l(\mathbb{X}^l)\times\mathbb{Y}^l}} \left| g(u_1, \ldots, u_i, \ldots, u_k) - g(u_1, \ldots, \bar{u}_i, \ldots, u_k) \right| \leq \frac{2}{k} C(\tilde{f}_l, P_{XY}), \tag{3.31}$$

where $C(\tilde{f}_l, P_{XY}) \triangleq \sup_{z, y_1^l \in \tilde{f}_l(\mathbb{X}^l)\times\mathbb{Y}^l} \left| \log \left(\frac{P_{\tilde{f}_l(X_1^l)Y_1^l}(\{z, y_1^l\})}{Q_{\tilde{f}_l(X_1^l)Y_1^l}(\{z, y_1^l\})} \right) \right|$. From the *bounded difference inequality* [75, Theorem 2.2], we have that

$$P_{\tilde{f}_{n,l}(X_1^n)Y_1^n} \left(\mathcal{B}^c_{n,t_\delta}(\tilde{f}_{n,l}) \right) \leq \exp \left(\frac{-k(l\delta)^2}{2C^2(\tilde{f}_l, P_{XY})} \right). \tag{3.32}$$

Finally, from (3.26), a lower bound for $t_n^*(\epsilon_n)$ can be obtained from (3.32) by making δ (that we denote by $\tilde{\delta}_{n,l}(\epsilon_n)$ in (3.33)) the solution of the following condition:

$$\exp \left(\frac{-k(l\tilde{\delta}_{n,l}(\epsilon_n))^2}{2C^2(\tilde{f}_l, P_{XY})} \right) = \epsilon_n. \tag{3.33}$$

Consequently, we have that

$$t_n^*(\epsilon_n) \geq \underbrace{\frac{1}{l} \mathcal{D}(P_{\tilde{f}_l(X_1^l)Y_1^l} \| Q_{\tilde{f}_l(X_1^l)Y_1^l}) - \tilde{\delta}_{n,l}(\epsilon_n)}_{t_{\tilde{\delta}_{n,l}(\epsilon_n)}=} \tag{3.34}$$

where from (3.33),

$$\tilde{\delta}_{n,l}(\epsilon_n) = \sqrt{\frac{2\ln(1/\epsilon_n)}{nl}} \cdot C(\tilde{f}_l, P_{XY}). \tag{3.35}$$

Finally, replacing the bound of $t_n^*(\epsilon_n)$ in (3.34) at the exponential term in (3.25) and taking logarithm, we have that:

$$\begin{aligned} &\xi(R) - \left(-\frac{1}{n} \log P_1(\tilde{f}_{n,l}, \phi_{n,t_{\tilde{\delta}_{n,l}(\epsilon_n)}}) \right) \\ &\leq \left[\xi(R) - \frac{1}{l} \mathcal{D}(P_{\tilde{f}_l(X_1^l)Y_1^l} \| Q_{\tilde{f}_l(X_1^l)Y_1^l}) \right] + \tilde{\delta}_{n,l}(\epsilon_n). \end{aligned} \tag{3.36}$$

Remark 1 Looking at (3.36) and using (2.6) and Theorem 3 in [28], $\forall \gamma > 0$, we can find a sufficient large l^* and f_l^* (function of γ) such that,

$$\xi(R) - \gamma < \frac{\mathcal{D}(P_{\tilde{f}_l^*(X_1^{l^*})Y_1^{l^*}} \| Q_{\tilde{f}_l^*(X_1^{l^*})Y_1^{l^*}})}{l^*} < \xi(R). \tag{3.37}$$

Returning to the proof, we have that $\forall l > 0$, $\forall n > 0$ and any $\epsilon_n > 0$

$$\begin{aligned}\xi(R) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n, R))\right) &\leq \xi(R) - \left(-\frac{1}{n}\log(P_1(\tilde{f}_{n,l}, \phi_{n,t_{\tilde{\delta}_{n,l}(\epsilon_n)}}))\right) \\ &\leq \xi(R) - \frac{1}{l}\mathcal{D}(P_{\tilde{f}_l(X_1^l)Y_1^l} \| Q_{\tilde{f}_l(X_1^l)Y_1^l}) + \tilde{\delta}_{n,l}(\epsilon_n) \\ &= \left(\max_{\substack{U: U \multimap X \multimap Y \\ I(U;X) \leq R \\ |\mathcal{U}| \leq |\mathcal{X}|+1}} I(U;Y) - \frac{1}{l} I(\tilde{f}_l(X_1^l); Y_1^l) \right) + \tilde{\delta}_{n,l}(\epsilon_n). \quad (3.38)\end{aligned}$$

The first inequality is from the fact that $\beta_n(\epsilon_n, R)) \leq P_1(\tilde{f}_{n,l}, \phi_{n,t_{\tilde{\delta}_{n,l}(\epsilon_n)}})$, the second from (3.36), and the last equality from the definition of $\xi(R)$ in Lemma 3.2.3, expressing the divergence as a mutual information [2].

It is worth noting that the bound in (3.38) is valid for an arbitrary $l > 0$. Considering that we know an expression for $\tilde{\delta}_{n,l}(\epsilon_n)$ from (3.35), we can address the problem of finding the best upper bound, i.e., the l that offers the best compromise between the two terms in the RHS of (3.38). For that, we need to focus on:

$$\max_{\substack{U: U \multimap X \multimap Y \\ I(U;X) \leq R \ |\mathcal{U}| \leq |\mathcal{X}|+1}} I(U;Y) - \max_{\tilde{f}_l : \mathcal{X}^l \to \{1,\ldots,2^{lR}\}} \frac{1}{l} I(\tilde{f}_l(X_1^l); Y_1^l), \quad (3.39)$$

which corresponds to the non-asymptotic analysis of the *information bottleneck (IB) problem* [71]. This coding problem can be viewed as a classical rate-distortion (fixed-rate) lossy source coding problem with the log-loss as the distortion function [76]. More precisely, (3.39) can be expressed by:

$$\min_{\tilde{f}_l : \mathcal{X}_1^l \to \{1,\ldots,2^{lR}\}} \frac{1}{l} H(Y_1^l | \tilde{f}_l(X_1^l)) - \min_{\substack{U: U \multimap X \multimap Y \\ I(U;X) \leq R \ |\mathcal{U}| \leq |\mathcal{X}|+1}} H(Y|U). \quad (3.40)$$

The following Lemma connects the expression in (3.40) with an instance of the classical rate distortion problem [77].

Lemma 3.7.1

$$\frac{1}{l} H(Y_1^l | \tilde{f}_l(X_1^l) \leq D(R) - \frac{d}{dR} D(R) \frac{\log(l)}{2l} + o\left(\frac{\log l}{l}\right), \quad (3.41)$$

where $D(R)$ is the noisy distortion-rate function given by

$$D(R) = \min_{\substack{U: U \multimap X \multimap Y \\ I(U;X) \leq R \ |\mathcal{U}| \leq |\mathcal{X}|+1}} H(Y|U). \quad (3.42)$$

The proof is presented in Appendix 3.7.3. Consequently, from (3.41) we have that the expression in (3.39) is upper bounded by $-\frac{d}{dR} D(R) \frac{\log(l)}{2l} + o\left(\frac{\log(l)}{l}\right)$. Applying this result to

(3.38), it follows that

$$\xi(R) - \left(-\frac{1}{n} \log(\beta_n(\epsilon_n, R)) \right) \leq -\frac{d}{dR} D(R) \frac{\log(l)}{2l} + \tilde{\delta}_{n,l}(\epsilon_n) + o\left(\frac{\log(l)}{l} \right). \tag{3.43}$$

To obtain a more explicit dependency of $\tilde{\delta}_{n,l}(\epsilon_n)$ on l we use the following result:

Proposition 3.4 *Let us consider two arbitrary probability distributions $\mu, \rho \in \mathbb{P}(\mathbb{X})$, an arbitrary encoder $f_n : \mathbb{X} \to \{1, \ldots, n\}$. and its induced partition of $\mathbb{X}$ given by $\pi_n = \{A_{i,n} \triangleq f_n^{-1}(\{i\}) : i \in \{1, \ldots, n\}\}$, then*

$$\sup_{A \in \pi_n} \frac{\mu(A)}{\rho(A)} \leq \sup_{x \in \mathbb{X}} \frac{\mu(\{x\})}{\rho(\{x\})}. \tag{3.44}$$

The proof is presented in Appendix 3.7.5.

From Proposition 3.4, we obtain that:

$$\begin{aligned} \tilde{\delta}_{n,l}(\epsilon_n) &= \sqrt{\frac{2 \ln(1/\epsilon_n)}{nl}} \cdot C(\tilde{f}_l, P_{XY}) \\ &\leq \sqrt{\frac{2l \ln(1/\epsilon_n)}{n}} \cdot C(P_{XY}). \end{aligned} \tag{3.45}$$

Using (3.45), the problem reduces to minimize the RHS of (3.43) as long as $(\epsilon_n)_n$ tends to zero at a sub-exponential rate, for which the assumption that $\left(\frac{1}{\epsilon_n}\right)_n$ is $o(e^{rn})$ for any $r > 0$ is central. In fact, it is sufficient to consider any sequence $(l_n)_n$ of integers such that $(1/l_n)_n$ is $o(1)$ and $(l_n)_n$ is $o\left(\frac{n}{\ln(1/\epsilon_n)}\right)$, from which we conclude that

$$\liminf_{n \to \infty} -\frac{1}{n} \log(\beta_n(\epsilon_n, R)) \geq \xi(R). \tag{3.46}$$

□

Conversely, we have the following result:

Theorem 3.5 *Let us assume that $\epsilon_n > 0$ for all n and that $(1/\epsilon_n)_n = o(e^{rn})$ for any $r > 0$, then*

$$\limsup_{n \to \infty} -\frac{1}{n} \log(\beta_n(\epsilon_n, R)) \leq \xi(R). \tag{3.47}$$

Proof. Let us consider a fixed-rate encoder $f_n : \mathbb{X}^n \to \{1, \ldots, 2^{nR}\}$ of rate R. We begin by using [78, Lemma 4.1.2], which states that for all $t > 0$ and $\forall \mathcal{A}_n \subset f_n(\mathbb{X}^n) \times \mathbb{Y}^n$

$$\begin{aligned} P_{f_n(X_1^n)Y_1^n}(\mathcal{A}_n^c) + e^{nt} Q_{f_n(X_1^n)Y_1^n}(\mathcal{A}_n) \\ \geq P_{f_n(X_1^n)Y_1^n}\left(\mathcal{B}_{n,t}^c(f_n) \right), \end{aligned} \tag{3.48}$$

where as before

$$\mathcal{B}_{n,t}(f_n) = \left\{(z, y_1^n) \in f_n(\mathbb{X}^n) \times \mathbb{Y}^n : \frac{P_{f_n(X_1^n)Y_1^n}(\{z, y_1^n\})}{Q_{f_n(X_1^n)Y_1^n}(\{z, y_1^n\})} > e^{nt}\right\}.$$

Eq. (3.48) is valid for any binary decision rule (represented by the set $\mathcal{A}_n$ in (3.48)) acting on $(f_n(X_1^n), Y_1^n)$. The rest of the argument focuses on finding a lower bound to the RHS of (3.48). The latter can be done by considering the following function $i(x_1^n, y_1^n) = \log\left(\frac{P_{Y_1^n|f_n(X_1^n)(y_1^n|f_n(x_1^n))}}{P_{Y_1^n}(y_1^n)}\right)$ and the fact that $\forall q \geq 1$

$$\mathbb{E}_{(X_1^n, Y_1^n) \sim P_{XY}^n}(i(X_1^n, Y_1^n)^q) \leq q!\,(4n^2 C(P_{XY})^2)^q. \tag{3.49}$$

Using $\mathcal{B}_{n,t}(f_n)$, it is useful to write $t = \frac{I(f_n(X_1^n);Y_1^n)}{n} + s$, then

$$P_{f_n(X_1^n)Y_1^n}\left(\mathcal{B}_{n,t}^c(f_n)\right) = P_{XY}^n\left(\{(x_1^n, y_1^n) : i(x_1^n, y_1^n) - \mathbb{E}(i(X_1^n, Y_1^n)) \leq ns\}\right), \tag{3.50}$$

where the expected values in (3.50) assumes that $(X_1^n, Y_1^n) \sim P_{XY}^n$. Using the bound on $i(X_1^n, Y_1^n)$ in (3.49), we can use the moment concentration inequality [43, Theorem 2.1] to obtain:

$$P_{XY}^n\left(\{i(x_1^n, y_1^n) - \mathbb{E}(i(X_1^n, Y_1^n)) \leq ns\}\right) \geq 1 - e^{-s^2/(32C(P_{XY})^2)}. \tag{3.51}$$

Combining (3.51) with (3.48), it follows that for any $s > 0$ and any set $\mathcal{A}_n \subset f_n(\mathbb{X}^n) \times \mathbb{Y}^n$

$$P_{f_n(X_1^n)Y_1^n}(\mathcal{A}_n^c) + e^{n\left(\frac{I(f_n(X_1^n);Y_1^n)}{n}+s\right)} Q_{f_n(X_1^n)Y_1^n}(\mathcal{A}_n) \geq 1 - e^{-s^2/(32C(P_{XY})^2)}. \tag{3.52}$$

At this point, we introduce the restriction on the TYPE I error in the analysis of $\mathcal{A}_n$. More preciasly, let us consider an arbitrary $\mathcal{A}_n$ such that $P_{f_n(X_1^n)Y_1^n}(\mathcal{A}_n^c) \leq \epsilon_n$. Then we have that:

$$e^{n\left(\frac{I(f_n(X_1^n);Y_1^n)}{n}+s\right)} Q_{f_n(X_1^n)Y_1^n}(\mathcal{A}_n) \geq 1 - e^{-s^2/(32C(P_{XY})^2)} - \epsilon_n. \tag{3.53}$$

Taking logarithm at both sides of (3.53) for any s satisfying the admisible condition $1 - \epsilon_n - e^{-\frac{s^2}{32C(P_{XY})^2}} > 0$, it follows that

$$\frac{I(f_n(X_1^n); Y_1^n)}{n} - \left(-\frac{1}{n}\log(Q_{f(X_1^n)Y_1^n}(\mathcal{A}_n))\right) \geq -s + \frac{\log\left(1 - \epsilon_n - e^{-\frac{s^2}{32C(P_{XY})^2}}\right)}{n}. \tag{3.54}$$

Since both f_n and the set $\mathcal{A}_n$ (the detector) are arbitrary in (3.54), the bound is valid for the optimal pair (f_n^*, ϕ_n^*) in (3.6) such that $Q_{f_n^*(X_1^n)Y_1^n}(\mathcal{A}_n^*) = \beta_n(\epsilon_n, R)$. In addition $\frac{I(f_n(X_1^n);Y_1^n)}{n} \leq \xi(R)$ by definition (see (2.5) in [28]), then for all $s > 4C(P_{XY})\sqrt{2\ln(1/1-\epsilon_n)}$ it follows that

$$\xi(R) + \frac{1}{n}\log(\beta_n(\epsilon_n, R)) \geq -s + \frac{\log\left(1 - \epsilon_n - e^{-\frac{s^2}{32C(P_{XY})^2}}\right)}{n}. \tag{3.55}$$

At this point, we use the assumption that $\lim_{n\to\infty} \epsilon_n = 0$, which implies that there is a sequence $(s_n)_n$ that is $\mathcal{O}(\sqrt{\log(n)/n})$ for which (3.55) evaluated at $s = s_n$ holds for any n, which implies that

$$\limsup_{n\to\infty} -\frac{1}{n}\log(\beta_n(\epsilon_n, R)) \leq \xi(R). \tag{3.56}$$

□

3.7.2 Proof of Theorem 3.2

PROOF. The proof can be divided in two independent parts from the analysis obtained in Theorems 3.3 and 3.5. On the one hand, we have an upper bound obtained by optimizing the RHS of (3.43) with respect to the blocklength l. More precisely, we have the following inequality:

$$\xi(R) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n, R))\right) \leq -\frac{d}{dR}D(R)\frac{\log l}{2l} + \sqrt{\frac{2l\ln(1/\epsilon_n)}{n}}C(P_{XY}) + o\left(\frac{\log l}{l}\right), \tag{3.57}$$

where $C(P_{XY}) \triangleq \sup_{(x,y)\in\mathbb{X}\times\mathbb{Y}}\left|\log\left(\frac{P_{XY}(\{(x,y)\})}{Q_{XY}(\{(x,y)\})}\right)\right|$. This expression depends on ϵ_n and it is valid for all $l \geq 1$. Then the tighest bound from (3.57), reduces to find l_n^* solution of:

$$\frac{\log l_n^*}{l_n^*} \approx \sqrt{\frac{l_n^*\ln(1/\epsilon_n)}{n}}. \tag{3.58}$$

To address this problem, we consider $l_n = n^\alpha$ to look for this optimal α (function of ϵ_n). This is the consequence of assuming that the condition in (3.58) holds, which reduces to:

$$\frac{\log n^\alpha}{n^\alpha} \approx \sqrt{\frac{n^\alpha\ln(1/\epsilon_n)}{n}}. \tag{3.59}$$

To solve (3.59), we move into the specific cases for (ϵ_n) stated in Theorem 3.2. We have three different scenarios:
a) $(\epsilon_n)_n = (1/n^p)_n$ with $p > 0$: The condition (3.59) reduces to

$$\frac{\alpha\log n}{n^\alpha} \approx \sqrt{\frac{n^\alpha p\ln(n)}{n}}, \tag{3.60}$$

where (non considering the logarithmic term) the equilibrium is obtained with $\alpha^* = 1/3$, which makes the upper bound in (3.57) of the form:

$$\begin{aligned}\xi(R) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n, R))\right) &\leq -\frac{dD(R)}{dR}\frac{\log n}{6n^{1/3}} + \sqrt{\frac{2p\ln(n)}{n^{2/3}}}C(P_{XY}) + o\left(\frac{\log n}{n^{1/3}}\right)\\ &= \left[-\frac{dD(R)}{dR}\cdot\frac{1}{6} + o(1)\right]\left(\frac{\log n}{n^{1/3}}\right).\end{aligned} \tag{3.61}$$

b) $(\epsilon_n)_n = (1/e^{n^p})_n$ with $p \in (0,1)$: Following the previous approach, we solve

$$\frac{\alpha\log n}{n^\alpha} \approx \sqrt{\frac{n^\alpha n^p}{n}}, \tag{3.62}$$

resulting in $\alpha^* = (1-p)/3$. This choice offers the bound

$$\begin{aligned}\xi(R) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n, R))\right) &\leq -\frac{dD(R)}{dR}\frac{(1-p)\log n}{6n^{(1-p)/3}} + \frac{\sqrt{2}C(P_{XY})}{n^{(1-p)/3}} + o\left(\frac{\log n}{n^{(1-p)/3}}\right)\\ &= \left[-\frac{dD(R)}{dR}\frac{(1-p)}{6} + o(1)\right]\left(\frac{\log n}{n^{(1-p)/3}}\right).\end{aligned} \tag{3.63}$$

c) $(\epsilon_n)_n = (1/\log(n))_n$: The matching condition reduces to find α such that

$$\frac{\alpha \log n}{n^\alpha} \approx \sqrt{\frac{n^\alpha \ln(\log(n))}{n}}. \tag{3.64}$$

It is simple to show that, as in the polynomial regime, the approximated solution is $\alpha^* = 1/3$, which offers the following upper bound:

$$\begin{aligned}\xi(R) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n, R))\right) &\leq -\frac{dD(R)}{dR}\frac{\log n}{6n^{1/3}} + \sqrt{\frac{2\ln(\log(n))}{n^{2/3}}}C(P_{XY}) + o\left(\frac{\log n}{n^{1/3}}\right) \\ &= \left[-\frac{dD(R)}{dR}\cdot\frac{1}{6} + o(1)\right]\left(\frac{\log n}{n^{1/3}}\right). \end{aligned} \tag{3.65}$$

For the lower bound, we use the following inequality from the proof of Theorem 3.5 (see (3.55)):

$$\xi(R) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n, R))\right) \geq -s + \frac{\log\left(1 - \epsilon_n - e^{-\frac{s^2}{32C(P_{XY})^2}}\right)}{n}. \tag{3.66}$$

This inequality is valid for any $s \in \mathbb{R}$ such that $1 - \epsilon_n - e^{-\frac{s^2}{32C(P_{XY})^2}} > 0$ or, equivalently, for all s such that $s > 4C(P_{XY})\sqrt{2\ln(1/1-\epsilon_n)}$. At this point, it is convenient to define $h_n(s) \triangleq 1 - \epsilon_n - e^{-\frac{s^2}{32C(P_{XY})^2}}$ in the domain $s > 4C(P_{XY})\sqrt{2\ln(1/1-\epsilon_n)}$. Then, (3.66) can be expressed in terms of $h_n(s)$ by

$$\begin{aligned}&\xi(R) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n, R))\right) \\ &\geq -4C(P_{XY})\sqrt{2\ln\left(\frac{1}{1-\epsilon_n-h_n(s)}\right)} - \frac{\log(1/h_n(s))}{n}, \end{aligned} \tag{3.67}$$

where $h_n(s) > 0$ if $s > 4C(P_{XY})\sqrt{2\ln(1/1-\epsilon_n)}$. We notice that as $(\epsilon_n)_n$ is $o(1)$ (function of n) the first term on the RHS of (3.67) tends to zero if, and only if, $(h_n(s))_n$ is $o(1)$. On the other hand, $(\log(1/h_n(s)))_n$ needs to be $o(n)$ to make the second terms on the RHS of (3.67) vanishing to zero with n. Then, there is a regime on the asymptotic behavior of $(h_n(s))_n$ where the bound in (3.67) is meaningful.

More precisely, for any finite n, we will address the problem of finding

$$s \in (4C(P_{XY})\sqrt{2\ln(1/1-\epsilon_n)}, \infty),$$

or equivalently finding $h_n(s) \in (0,1)$, that offers the best lower bound from (3.67). On the specifics, as $(\epsilon_n)_n$ and $(h_n(s))_n$ go to zero with n, for the first term $-4C(P_{XY})\sqrt{2\ln\left(\frac{1}{1-\epsilon_n-h_n(s)}\right)}$

a Taylor expansion around 1 is used to aproximate the function. In particular, it follows that:

$$
\begin{aligned}
-4\sqrt{2}C(P_{XY})\sqrt{\ln\left(\frac{1}{1-\epsilon_n-h_n(s)}\right)} &\geq -2\sqrt{2}C(P_{XY})\sqrt{\epsilon_n+h_n(s)}\frac{\sqrt{4-5(\epsilon_n+h_n(s))}}{1-\epsilon_n-h_n(s)} \\
&\geq -2\sqrt{2}C(P_{XY})\sqrt{\epsilon_n+h_n(s)}\frac{\sqrt{4}}{1/2} \\
&= -8\sqrt{2}C(P_{XY})\sqrt{\epsilon_n+h_n(s)}, \qquad (3.68)
\end{aligned}
$$

where the last inequality is obtained eventually as $(\epsilon_n+h_n(s))_n$ is $o(1)$. Then, from (3.67) and (3.68), the optimal lower bound reduces to find the optimal balance between $-8\sqrt{2}C(P_{XY})\sqrt{\epsilon_n+h_n(s)}$ and $\frac{\log(1/h_n(s))}{n}$. It is important to note that $-8\sqrt{2}C(P_{XY})\sqrt{\epsilon_n+h_n(s)}$ tends to zero at a speed that is proportional to how fast $(h_n(s))_n$ tends to zero, as long as, $(h_n(s))_n$ is $o(\epsilon_n)$, otherwise, the speed is dominated by $\mathcal{O}(\sqrt{\epsilon_n})$, which is independent of $(h_n(s))_n$. On the other hand, the second term $(\log(1/h_n(s)))_n$ tends to zero at a rate that is inversely proportional to the speed at which $(h_n(s))_n$ goes to zero. Therefore, the balance is function of $(\epsilon_n)_n$. We recognize two regimes for this optimization problem:

1- If for some $K>0$ we have that $\sqrt{2\epsilon_n}\geq K\frac{\log(1/\epsilon_n)}{n}$, eventually in n, then the solution of the optimization problem is achieved when $(h_n(s))_n\approx(\epsilon_n)_n$ (**Regime 1**);

2- Otherwise, if $(\sqrt{2\epsilon_n})_n$ is $o\left(\frac{\log(1/\epsilon_n)}{n}\right)$, then the solution of the optimization problem implies that $(\epsilon_n)_n$ is $o(h_n(s))$ (**Regime 2**).

Finally, to obtain the upper bound, we need to evaluate $(\epsilon_n)_n$ in the different scenarios stated in Theorem 3.2.

- $(\epsilon_n)_n=(1/\log(n))_n$: Regime 1 is met, then we choose $h_n(s)=\epsilon_n$. This implies that

$$
\begin{aligned}
\xi(R)-\left(-\frac{1}{n}\log(\beta_n(\epsilon_n,R))\right) &\geq \frac{-16C(P_{XY})}{\sqrt{\log(n)}}-\frac{\log(\log(n))}{n} \\
&= \left(-16C(P_{XY})-\frac{\log(\log(n))\sqrt{\log(n)}}{n}\right)\frac{1}{\sqrt{\log(n)}} \\
&= (-16C(P_{XY})-o(1))\left(\frac{1}{\sqrt{\log(n)}}\right). \qquad (3.69)
\end{aligned}
$$

- $(\epsilon_n)_n=(1/n^p)_n$ with $2>p>0$: Regime 1 is met, then we choose $h_n(s)=\epsilon_n$. This implies that

$$
\begin{aligned}
\xi(R)-\left(-\frac{1}{n}\log(\beta_n(\epsilon_n,R))\right) &\geq \frac{-16C(P_{XY})}{n^{p/2}}-\frac{p\log(n)}{n} \\
&= \left(-16C(P_{XY})-\frac{p\log(n)}{n^{1-p/2}}\right)\left(\frac{1}{n^{p/2}}\right) \\
&= (-16C(P_{XY})-o(1))\left(\frac{1}{n^{p/2}}\right). \qquad (3.70)
\end{aligned}
$$

- $(\epsilon_n)_n = (1/n^p)_n$ with $p \geq 2$: Regime 2 is met, then we have to solve the following matching condition

$$\sqrt{\epsilon_n + h_n(s)} \approx \frac{\log(1/h_n(s))}{n}. \tag{3.71}$$

Assuming $h_n(s) = 1/n^\alpha$, $\alpha \in (0, 2]$, the equilibrium is obtained with $\alpha^* = 2$. This implies that

$$\begin{aligned}\xi(R) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n, R))\right) &\geq -8\sqrt{2}C(P_{XY})\sqrt{n^{-p} + n^{-2}} - \frac{2\log(n)}{n} \\ &= \left(-8\sqrt{2}C(P_{XY})\frac{\sqrt{n^{2-p}+1}}{\log(n)} - 2\right)\left(\frac{\log(n)}{n}\right) \\ &= (-o(1) - 2)\left(\frac{\log(n)}{n}\right). \end{aligned} \tag{3.72}$$

- $(\epsilon_n)_n = (1/e^{n^p})_n$ with $p \in (0, 1)$: Regime 2 is met, then we follow the same condition in (3.71). The equilibrium is obtained with $\alpha^* = 2$. This implies that

$$\begin{aligned}\xi(R) - \left(-\frac{1}{n}\log(\beta_n(\epsilon_n, R))\right) &\geq -8\sqrt{2}C(P_{XY})\sqrt{e^{-n^p} + n^{-2}} - \frac{\log(n)}{n} \\ &= \left(-8\sqrt{2}C(P_{XY})\frac{\sqrt{e^{-n^p}n^2+1}}{\log(n)} - 2\right)\left(\frac{\log(n)}{n}\right) \\ &= (-o(1) - 2)\left(\frac{\log(n)}{n}\right). \end{aligned} \tag{3.73}$$

□

3.7.3 Proof of Lemma 3.7.1

Proof. Let us consider a family of probability distributions $P_\lambda \in \mathcal{P}(\mathbb{Y})$ indexed with a parameter $\lambda \in \Lambda$, where Λ is some parametric space. Given a vector of parameters $\lambda_1^n \in \Lambda^n$, the product probability distribution in $\mathcal{P}(\mathbb{Y}^n)$ is defined as

$$P_{\lambda^n}(\{y_1^n\}) \triangleq \prod_{i=1}^n P_{\lambda_i}(\{y_i\}). \tag{3.74}$$

Let $\rho(\lambda_1^n, Y_1^n) : \Lambda^n \times \mathbb{Y}_1^n \to \mathbb{R}^+ \cup \{0\}$ denote the logarithmic loss distortion defined by:

$$\rho(\lambda_1^n, y_1^n) \triangleq -\frac{1}{n}\log P_{\lambda_1^n}(\{y_1^n\}) = \sum_{i=1}^n -\frac{1}{n}\log P_{\lambda_i}(\{y_i\}).$$

By construction $\rho(\lambda_1^n, y_1^n)$ is additive ($\rho(\lambda_1^n, y_1^n) = \sum_{i=1}^n \rho(\lambda_i, y_i)$) and then the following result holds:

Lemma 3.7.2 [76, Lemma 1] Let X_1^l, Y_1^l be a random vector with known joint distribution. For any function $\tilde{f}_l : \mathbb{X}^l \to \{1, ..., 2^{lR}\}$ and function $g : \{1, ..., 2^{lR}\} \to \Lambda^n$ such that $g(\tilde{f}_l(X_1^l)) = \lambda_1^l$ it follows that

$$\mathbb{E}[\rho(g(u), Y_1^l)|\tilde{f}_l(X_1^l) = u] \geq \frac{1}{l} H(Y_1^l|\tilde{f}_l(X_1^l) = u). \tag{3.75}$$

Taking expectation on the two sides of (3.75) with repect to X_1^l, we get that

$$\mathbb{E}[\rho(g(\tilde{f}_l(X_1^l)), Y_1^l)] \geq \frac{1}{l} H(Y_1^l|\tilde{f}_l(X_1^l)). \tag{3.76}$$

Remark 2 We observe that, if we identify the $\tilde{f}_l$ as an encoder and g as the decoder, the term in the LHS of (3.76) corresponds to the noisy rate distortion function under the logarithmic loss. Then, for the purpose of the following result, it is convenient to redefine the distortion function $\tilde{\rho}(x_1^l, \lambda_1^l) : \mathbb{X}_1^l \times \Lambda_1^l \to \mathbb{R} \cup \{0\}$ as

$$\tilde{\rho}(x_1^l, \lambda_1^l) \triangleq \mathbb{E}[\rho(\lambda_1^l, Y_1^l)|X_1^l = x_1^l]. \tag{3.77}$$

Denoting $\lambda_i = g_i(\tilde{f}_l(x_1^l))$ and g_i is the ith component of g, we observe that $\tilde{\rho}(x_1^l, \lambda_1^l) = \sum_{i=1}^l \tilde{\rho}(x_i, \lambda_i)$ is additive.

Finally, using the previous observation, we can use $\tilde{f}_l$ as the encoder and g_i as the decoder, to recover an instance of the rate distortion problem [77]. Therefore, from [45, Theorem 3], we obtain that

$$\begin{aligned}\frac{1}{l} H(Y_1^l|\tilde{f}_l(X_1^l)) &\leq \mathbb{E}_{X \sim P_X^l}[\tilde{\rho}(X_1^l, \lambda_1^l)] \\ &\leq D(R) - \frac{d}{dR} D(R) \frac{\log(l)}{2l} + o\left(\frac{\log(l)}{l}\right),\end{aligned}$$

which concludes the proof. □

3.7.4 Finite-length Result for the Unconstrained Case

Lemma 3.7.3 [27] Let us consider $\epsilon \in (0, 1)$, then eventually in n it follows that $-\frac{\log(\beta_n(\epsilon))}{n} =$

$$\mathcal{D}(P\|Q) + \sqrt{\frac{V(P\|Q)}{n}} \Phi^{-1}(\epsilon) + \frac{\log n}{2n} + \mathcal{O}\left(\frac{1}{n}\right),$$

where $V(P\|Q) = \sum_{x \in \mathbb{X}} P(\{x\}) \left[\log\left(\frac{P(\{x\})}{Q(\{x\})}\right) - \mathcal{D}(P\|Q)\right]^2$.

A direct corollary of this result shows that $\left|\mathcal{D}(P\|Q) - \left(-\frac{1}{n}\log(\beta_n(\epsilon))\right)\right|$ is $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

3.7.5 Proof of Proposition 3.4

PROOF. Given $\mathcal{A} \in \pi_n$, we note that

$$\frac{\mu(\mathcal{A})}{\rho(\mathcal{A})} = \frac{\sum\limits_{j\in\mathcal{A}} \mu(\{j\})}{\sum\limits_{j\in\mathcal{A}} \rho(\{j\})}. \tag{3.78}$$

Then, given a collection of positive numbers $\{a_i : i \in \{1,\dots,n\}\}$ and $\{b_i : i \in \{1,\dots,n\}\}$, we use the following basic inequality

$$\sum_{i=1}^{n} a_i \leq \max_i \left\{\frac{a_i}{b_i}\right\} \sum_{i=1}^{n} b_i. \tag{3.79}$$

Finally, since $\mathcal{A}$ is arbitrary in (3.78) and using the positiveness of the probability, we conclude the desired result. □

Chapter 4

Collaboration in Decentralized Testing Against Independence: Performance Analysis and Data-Driven Design

In this chapter, we study collaboration in distributed inference for testing independence with a fixed-rate communication constraint. We look at collaboration as a strategy to improve performance. By collaboration, we mean a decentralized framework where two agents (or nodes of the networks) interchange messages with an overall rate constraint to arrive at a final decision. In this context, we derive a novel asymptotic performance result (an error exponent with a single-letter characterization) that expresses the benefit of collaboration in concrete terms — compared to the standard one-sided distributed strategy. We also address the practical problem of designing encoders and decoders for this distributed task. Using the knowledge derived about the information limit of this problem, we propose an info-max design principle where tools from machine learning are incorporated to facilitate practical solutions. Experiments using synthetic data show that our collaborative solution can substantially improve performance. Remarkably, we observe that a measure of the symmetrical structure of the underlying model of the test dictates this gain. Furthermore, our rate-constraint framework effectively preserves the discrimination information of the centralized (best scenario) test against independence.

4.1 Introduction

The study of distributed decision and inference, where measurements are collected remotely subject to communication constraints, is an important problem in signal processing over networks. In this regard, the emerging field of the Internet of Things (IoT) brings new dimensions and technical challenges to classic decision problems as data is no longer centrally available for inference, and we need to deal with technical challenges associated with data corruption, sensor noise, adversarial sources of perturbation [23] and unlabeled or unordered data [15–18]. In this context, an essential problem of signal processing over networks is the task of distributed decision, where observations are collected remotely and the decision is subject to communication constraints [19–21, 28, 79, 80]. In this chapter we focus on the im-

portant problem of detecting whether two remote measurements are statistically independent or not. This problem is fundamental from a theoretical standpoint [28,81–83] and relevant in many applications [47, 48, 84–88]. Examples of relevant applications include the problem of spectrum sensing in the area of cognitive radio networks where there is a primary agent and a group of secondary agents; the presence of the primary user's signal introduces dependence on the decentralized sensors, and then the secondary agents collaborate to detect whether the primary agent is present or not based on a test against independence [85, 86]. Other applications have been explored in the context of censoring and security where a node is interested in detecting independence of two remotely located sources with the presence of an eavesdropper, which has access to the encoded bits [87], [88].

The celebrated one-sided distributed setting introduced by Ahlswede and Csiszar in [28] consists of a test against independence where the observations (the evidence) are available at two remote nodes, as shown in Figure 4.1. In particular, one of the nodes transmits to another remote node (the detector) subject to a rate-constraint in bits per-sample. In this setting, an information-theoretic analysis of the problem has been addressed [28], where the asymptotic limit for the TYPE II error subject to a fixed TYPE I error constraint was derived in a closed form (the error exponent). More recently, finite-length performance bounds (non-asymptotic) have been presented for this one-sided distributed problem [54,68,82].

These bounds are important because they indicate what is possible to achieve with an optimal encoder and detector when a finite number of observations are available for decision making. In addition, these bounds shed light on the speed at which the error exponents are attained as the number of samples tends to infinity, and, consequently, how well the error exponents represent performances of practical decision schemes operating with finite samples [82].

Departing from previous works, in this chapter we explore the role that collaboration plays as a decentralized strategy to improve performance in distributed testing of independence. By collaboration, we mean a setting of decentralized detection where the two agents cooperate (interchange messages with an overall rate-constraint) to arrive at a final decision (see an illustration in Figure 4.1). Applications of this setting can be found, for example, in a cooperative communication system, where each wireless agent (user) transmits data and also acts as a cooperative agent for other users: each agent transmits both its own bits as well as some information for its partner to meet an operational requirement [89, 90]. Under this umbrella, we look for two relevant technical aspects associated with the role of collaboration. First, an important result is the derivation of performance limits (error exponent) that express the benefit of collaboration, if any, when the number of samples tends to infinity. From this information analysis, we could analyze the benefit of collaboration and how the (information) gains depend on specific attributes of the problem. Second, we propose concrete data-driven strategies to implement and deploy these distributed inference techniques. The idea here is to use the theoretical results (error exponent limits) as a guideline to design decentralized encoders-decoders that could offer competitive performances. To the best of our knowledge, these two aspects have not been studied systematically and remain open.

4.1.1 Contributions

1. We formalize and present a one-round collaborative extension of the distributed setting introduced by Ahlswede and Csiszar in [28], as shown in Figure 4.1.
2. In the theoretical analysis of the one-round collaborative setting, we derive an information limit (in the form of an error exponent) of the TYPE II error probability subject to a vanishing TYPE I error (Theorem 4.1). This result has two technical parts. We first derive an achievable result using for that a collaborative encoder-decoder strategy with a single round of interaction. This feasibility approach offers a lower bound for the error exponent for our task (testing against independence). The second part of the result presents an unfeasible argument (i.e., a converse argument) that fully determines a single-letter error exponent expression for our problem. This result is in line with what is known as the one-sided distributed setting [28] with a distinctive information term attributed to the role played by collaboration in the problem.
3. Based on this asymptotic result, we analyze the performance gain with respect to the one-sided (unidirectional) case introduced in [28]. We see that the overall quality of the test performance is governed by the bit assignment between the nodes and is affected, at the same time, by the distribution of the model.
4. On the practical side, we propose a data-driven design criterion for the two encoders and the decoder of the introduced one-round collaborative setting (see Figure 4.1). To design the encoders, our criterion is informed by the single-letter optimization that determines the error exponent in Theorem 4.1. This problem reduces to an info-max optimization task that learns the encoders (quantizers) from supervised data of the problem. Borrowing ideas from machine learning (ML) [91–93], a collection of soft-quantizer, i.e. conditional probabilities, is used to formalize the info-max problem. In particular, we consider the rich collection of *Boltzmann distributions* to represent the space of soft-quantizers.[1]
5. Corroborating our previous analysis, empirical results based on simulations show that the proposed one-round collaboration strategy outperforms (in the ROC curve, i.e., TYPE I and TYPE II trade-off) the non-collaborative strategy in general and that the performance gain is a function of structural attributes in the model. Importantly, we observe that the performance gain is proportional to a measure of the asymmetry of the underlying probability model. In this analysis, we also evaluate how the number of samples (block-length) and the communication constraint (number of bits) affect the comparison.

4.1.2 Related Works

The contribution that is closest to our work was made by Kim *et al.* [29]. They explored an interactive communication scheme for testing against independence similar to the one-round collaborating strategy studied in this work. They presented a construction that provides a lower bound for the error exponent of the one-round collaboration problem. In our work, we present a new achievable construction that offers a strong feasibility result and a converse

[1] Interestingly, this design task connects with the so-called Variational *Information Bottleneck* (IB) problem in machine learning [71, 94, 95], as we are optimizing an empirical mutual information between class and latent variables (the output of the quantizer) subject to a compression constraint [96].

argument (impossibility result) that allows us to obtain a closed-form analytical expression (single letter) for the error exponent. Other related works addressed collaboration in the context of hypothesis testing with multiple detectors. Two interesting examples are [63, 64] in which they obtain achievable TYPE II error exponents for testing against conditional independence for two decision agents.

4.1.3 Chapter Organization

This chapter is organized as follows. Section 4.2 introduces the definitions and the problem setting for the unidirectional and collaborative case. Section 4.3 presents the main result (Theorem 4.1) for the collaborative regime. Sections 4.4 and 4.5 present our data-driven algorithms for the unidirectional and the collaborative regimes, respectively. Numerical analysis and discussions are presented in Section 4.6. To conclude, the proof of Theorem 4.1 and supporting results can be found in Appendix Sections.

4.1.4 Basic Notation and Conventions

We use upper-case letters to denote random variables (RVs) and lower-case letters to denote realizations of RVs. Vectors are denoted by $X_1^n = (X_1, ..., X_n)$ with their length as a superscript. Sets, including alphabets of RVs, are denoted by calligraphic letters. Throughout this chapter we assume all RVs are defined over finite alphabets. $P_X \in \mathcal{P}(\mathbb{X})$ denotes the distribution for a RV X defined on the set $\mathbb{X}$ and $\mathcal{P}(\mathbb{X})$ denotes the set of all possible distributions over $\mathbb{X}$. $X \multimap Y \multimap Z$ indicates that X, Y and Z form a Markov chain. For a RV $X \sim P_X$, the Shannon *entropy* is defined by $H(X) = H(P_X) = -\sum_{x \in \mathbb{X}} P_X(x) \log P_X(x)$. Similarly, the *conditional entropy* is

$$H(Y|X) = H(P_{Y|X}|P_X) = -\sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} P_X(x) P_{Y|X}(y|x) \log P_{Y|X}(y|x) \tag{4.1}$$

where $P_{Y|X} : \mathbb{X} \mapsto \mathcal{P}(\mathbb{Y})$ denotes the conditional distribution and

$$D(P_X \| Q_X) = \sum_{x \in \mathbb{X}} P_X(x) \log \frac{P_X(x)}{Q_X(x)}$$

is the *Kullback-Leiber (KL)-divergence* between P_X and Q_X in $\mathcal{P}(\mathbb{X})$. The *conditional KL divergence* between two stochastic mappings $P_{Y|X} : \mathbb{X} \mapsto \mathcal{P}(\mathbb{Y})$ and $Q_{Y|X} : \mathbb{X} \mapsto \mathcal{P}(\mathbb{Y})$ wrt $P_X \in \mathcal{P}(\mathbb{X})$ is

$$\mathcal{D}(P_{Y|X} \| Q_{Y|X} | P_X) = \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} P_X(x) P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)}. \tag{4.2}$$

For any two RVs, $X, Y \sim P_{XY}$, the *mutual information* is $I(X;Y) = \mathcal{D}(P_{XY} \| P_X \cdot P_Y)$. The logarithm is assumed to be of base 2.

4.2 Problem Setting

We introduce the general setting of the problem addressed in this work. We have a decentralized system conformed by two agents or nodes depicted in Fig. 4.1. These nodes sense the

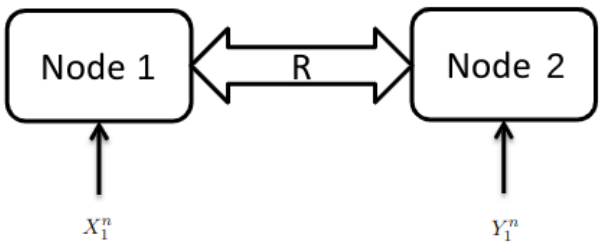


Figure 4.1: Schematics of the collaborative distributed hypothesis testing problem setting.

environment in a distributed way. Node 1 has lossless access to $X_1^n = (X_1, .., X_n)$ and Node 2 has lossless access to $Y_1^n = (Y_1, ..., Y_n)$. Then the two nodes communicate using a finite number of resources (in bits per sample) and make a decision about the joint underlying probabilistic structure of (X_1^n, Y_1^n). Below we introduce the formal elements of this problem and the main questions addressed in this work.

4.2.1 Testing Against Independence

Let us consider a finite observation space $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$. The joint random vector (X, Y) in $\mathbb{Z}$ represents two information sources sensed independently by two remote nodes illustrated in Fig. 4.1. In this context, we consider two scenarios: either (X, Y) follows a joint distribution $P_{X,Y} \in \mathcal{P}(\mathbb{Z})$, which is the main model of our problem, or (X, Y) follows an alternative model $Q_{X,Y} = P_X \cdot P_Y$ that consists of the product of the marginals of $P_{X,Y}$: i.e., the scenario where X and Y are independent[2]. The two nodes collect n i.i.d. samples of (X, Y). We denote by $X_1^n = (X_1, ..., X_n)$ and $Y_1^n = (Y_1, ..., Y_n)$ the n-samples where $P_{X,Y}^n$ denotes the n-product (i.i.d) distribution of (X_1^n, Y_1^n). Then, we can introduce the two hypotheses of the test against independence by

$$\begin{aligned} H_0 : &\quad (X_1^n, Y_1^n) \sim P_{XY}^n, \\ H_1 : &\quad (X_1^n, Y_1^n) \sim Q_{X,Y}^n. \end{aligned} \tag{4.3}$$

The observational setting in (4.3) is the standard centralized context where based on the joint evidence (X_1^n, Y_1^n) a decision about the underlying data-generation process should be made [97]. In this decision problem, we recognize two classic error events: the Type I error is when independence is assumed despite H_0 being true, while the Type II error is when non-independence is declared despite H_1 being true.

We will introduce two different decentralized settings to address the problem depicted in Fig. 4.1, where information of (X_1^n, Y_1^n) is acquired remotely by the nodes, and a total constraint in the number of messages (i.e., the number of bits) is imposed to make a decision about the two underlying hypotheses: H_0 and H_1. Here, we recognize the standard unidirectional setting introduced in [28], and an alternative strategy that involves one-round of collaboration between the nodes.

[2] $P_X \in \mathcal{P}(\mathbb{X})$ and $P_Y \in \mathcal{P}(\mathbb{Y})$ denote the marginals distributions of X and Y, respectively.

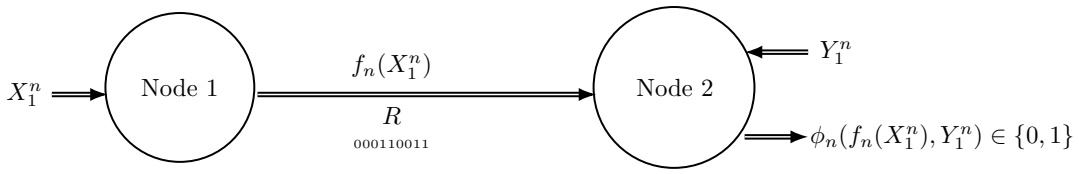


Figure 4.2: The one-directional distributed test in which $f_n(\cdot)$ is the encoder and $\phi_n(\cdot)$ is the detector acting on $(f_n(X_1^n), Y_1^n)$.

4.2.2 The One-directional (non-collaborative) Strategy

We begin with the distributed setting introduced by Ahlswede and Csiszar in [28] that is depicted in Figure 4.2. The problem is to test the two hypotheses in (4.3) where the observations are available at two remote locations and one of the observations X_1^n needs to be transmitted from Node 1 to Node 2 subject to a rate-constraint of $R > 0$ in bits per-sample. In this setting, Node 2 is acting as the detector, i.e., it observes Y_1^n (lossless) and receives a compressed (lossy) version of X_1^n to decide about H_0 and H_1. In this context, we recognize an encoder and a detector. More precisely, let us consider *the n-block pair of encoder-detector* (f_n, ϕ_n) that uses R bits per sample by the following structure:

$$\begin{aligned} &f_n : \mathbb{X}^n \to \{1, \ldots, 2^{nR}\}, \text{ (encoder)} \\ &\phi_n : \{1, \ldots, 2^{nR}\} \times \mathbb{Y}^n \to \Theta = \{0, 1\}, \text{ (detector).} \end{aligned} \tag{4.4}$$

The detector $\phi_n(\cdot)$ makes a decision from the one-sided compressed observations $(f_n(X_1^n), Y_1^n) \in \{1, ..., 2^{nR}\} \times \mathbb{Y}^n$ as illustrated in Fig.4.2. Then, we have the following TYPE I and TYPE II errors:

$$\alpha_n((f_n, \phi_n)) \equiv P_{XY}^n(\mathcal{A}^c(f_n, \phi_n)) \text{ and} \tag{4.5}$$

$$\beta_n((f_n, \phi_n)) \equiv Q_{XY}^n(\mathcal{A}(f_n, \phi_n)) \tag{4.6}$$

where $A(f_n, \phi_n) \equiv \{(x_1^n, y_1^n) : \phi_n(f_n(x_1^n), y_1^n) = 0\} \subset \mathbb{X}^n \times \mathbb{Y}^n$.

To express the optimal trade-off between the TYPE I and TYPE II errors of this problem, we consider the following: given $n \geq 1$ (the number of samples), $R > 0$ (the bits per sample) and $\epsilon > 0$ (the TYPE I error constraint), the best TYPE II is the solution of

$$\beta_n(\epsilon, R) \equiv \min_{(f_n, \phi_n)} \{\beta_n((f_n, \phi_n)) : \alpha_n((f_n, \phi_n)) \leq \epsilon\}. \tag{4.7}$$

The solution in (4.7) is optimized over all n-blocks pairs (f_n, ϕ_n) that uses R bits of information (see Eq.(4.4)). Therefore, $(\beta_n(\epsilon, R))_{\epsilon > 0}$ expresses the optimal operational performance for detecting H_0 vs. H_1 given a finite number of observations n and nR bits of communications resources.

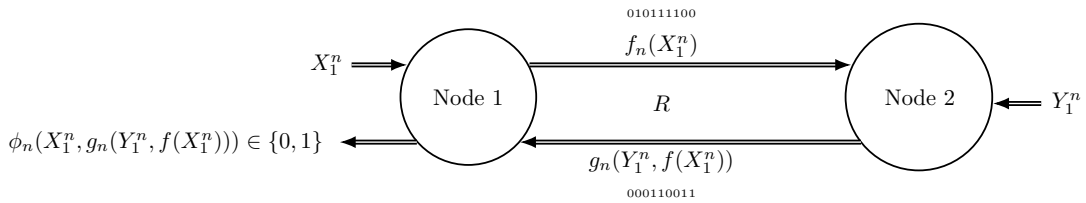


Figure 4.3: The one-round distributed test in which $f_n(\cdot)$, $g_n(\cdot)$ is the encoder and $\phi_n(\cdot)$ is the detector acting on $(X_1^n, g^1(f_n^1(X_1^n), Y_1^n))$.

4.2.3 The Collaborative Strategy

We consider a simple collaborative strategy where Node 1 and Node 2 interchange messages before making a decision. In particular, we work on what we call a one-round communication setting depicted in Fig. 4.3.

In this setting, we recognize two encoders $(f_n(\cdot), g_n(\cdot))$ and a detector ϕ_n given by

$$\begin{aligned} f_n &: \mathbb{X}^n \longrightarrow \mathbb{U}, \ (encoder\ 1) \\ g_n &: \mathbb{Y}^n \times \mathbb{U} \longrightarrow \mathbb{V}, \ (encoder\ 2) \\ \phi_n &: \mathbb{X}^n \times \mathbb{V} \longrightarrow \{0,1\}, \ (detector). \end{aligned} \tag{4.8}$$

The encoders $f_n(\cdot)$ and $g_n(\cdot)$ satisfy overall fixed-rate communication constraints in bits per sample given by

$$\log\left(|\mathbb{U}||\mathbb{V}|\right) \leq nR. \tag{4.9}$$

Given the encoders-decoder $(f_n(\cdot), g_n(\cdot), \phi_n(\cdot))$ in (4.8), there are two stages of data transmission. In the first stage, $f_n(\cdot)$ is used to transmit information from Node 1 to Node 2. In the second stage, $g_n(\cdot)$ is used to transmit information from Node 2 to Node 1 (see Figure 4.3). Then, the final decision is made by Node 1 after receiving the message from Node 2 with an overall rate constraint expressed in (4.9). The information flow goes from right-to-left and then from left-to-right as shown in Fig. 4.3. It is worth noting that the total bits budget used on these two data-compression stages is constrained by the same fixed-rate restriction $R > 0$ introduced in Section 4.2.2.

As for performance, the corresponding TYPE I and TYPE II errors are given by

$$\alpha_n((f_n, g_n, \phi_n)) \equiv P_{XY}^n(\mathcal{A}^c(f_n, g_n, \phi_n)) \text{ and} \tag{4.10}$$

$$\beta_n((f_n, g_n, \phi_n)) \equiv Q_{XY}^n(\mathcal{A}(f_n, g_n, \phi_n)), \tag{4.11}$$

where $\mathcal{A}(f_n, g_n, \phi_n) \equiv \{(x_1^n, y_1^n) \in \mathbb{X}^n \times \mathbb{Y}^n : \phi_n(x_1^n, g_n(y_1^n, f(x_1^n))) = 0\}$. As an analogy to what was presented in (4.7), we introduce an expression for the optimal trade-off between TYPE I and TYPE II errors given by

$$\beta_n^c(\epsilon, R) \equiv \min_{(f_n, g_n, \phi_n)} \{\beta_n((f_n, g_n, \phi_n)) : \alpha_n((f_n, g_n, \phi_n)) \leq \epsilon\}, \tag{4.12}$$

where the minimization in (4.12) is above all the rules that use n-samples of (X, Y) and satisfy the fixed-rate communication constraint of (4.9).

The main focus of this work is to analyze the benefit of collaboration and evaluate in theory and in practice if collaboration offers an advantage in terms of the optimal trade-off between the TYPE I and TYPE II errors i.e. on the analysis of $(\beta_n^c(\epsilon, R))_{\epsilon>0}$ in (4.12) vs. $(\beta_n(\epsilon, R))_{\epsilon>0}$ in (4.7) for the same model $P_{X,Y}$. An important aspect is to quantify the performance gain of collaboration, if any, and see how that depends on properties of the underlying model $P_{X,Y}$. To address these questions, Section 4.3 presents information-theoretic limits to contrast these two strategies from a theoretical perspective (when $n \longrightarrow \infty$), while Sections 4.4 and 4.5 explore a non-asymptotic comparison where practical solutions for the encoder and decoders of the two strategies are presented.

4.3 Asymptotic Analysis

In this section, we present the main theoretical contribution of this work: Theorem 4.1. This result offers a precise asymptotic characterization for $\beta_n^c(\epsilon, R)$ in (4.12) when $n \longrightarrow \infty$. We revisit the asymptotic result for the non-collaborative case studied in [28] and discuss the interpretation of these two information limits. This comparative analysis is presented in terms of an error exponent analysis [2]. Error exponents (EE) are fundamental performance limits that express the discrimination power of the test when $n \longrightarrow \infty$. Importantly, EE can be adopted as a very good approximation of the optimal performance, $\beta_n^c(\epsilon_n, R)$, when n is sufficiently large [3].

4.3.1 Collaborative Hypothesis Testing

In this subsection, we derive a closed-form expression for $-\frac{1}{n}\log\beta_n^c(R, \epsilon)$ when n tends to infinity. This expression determines the exponential velocity at which the TYPE II error tends to zero (with the number of samples) given $\epsilon > 0$ (a fixed TYPE I error restriction) and $R > 0$ (a fixed-rate constraint for the collaboration between Node 1 and Node 2). The result is the following:

Theorem 4.1 *Given $P_{X,Y}$ in (4.3) and $R > 0$, the best performance trade-off of the collaborative setting in (4.12) satisfies the following*

$$\lim_{\epsilon\to 0} \liminf_{n\to\infty} -\frac{1}{n}\log\beta_n^c(R, \epsilon) = E(R), \tag{4.13}$$

where

$$E(R) = \max_{\substack{P_{U|X} : \mathbb{X} \mapsto \mathcal{P}(\mathbb{U}) \\ P_{V|UY} : \mathbb{U}\times\mathbb{Y} \mapsto \mathcal{P}(\mathbb{V}) \\ s.t.\ I(U;X) + I(V;Y|U) \le R}} [I(U;Y) + I(V;X|U)]. \tag{4.14}$$

[3] [82, Th. 2] shows that $e^{-nE(R)}$ is an excellent proxy for $\beta_n(\epsilon_n, R)$, where $E(R)$ denotes the EE of the test within the limits of large samples. Importantly, [82, Sec.V] shows that this approximation happens very quickly when n increases.

In (4.14) $(X,Y) \sim P_{XY}$ and U and V are obtained from the joint vector $(X,Y,U,V) \sim P_{X,Y} \cdot P_{U|X} \cdot P_{V|U,Y}$. $P_{U|X}$ is a conditional probability from $\mathbb{X}$ to $\mathbb{U}$, and $P_{V|U,Y}$ denotes the conditional probabilities from $\mathbb{U} \times \mathbb{V}$ to $\mathbb{V}$, meaning that

$$X \multimap (Y,U) \multimap V. \tag{4.15}$$

The proof of this result is presented in Appendix 4.8.1.

The result in (4.13) states that $\beta_n^c(\epsilon, R)$ decreases exponentially with n with an equivalent exponent $E(R) > 0$ that is determined analytically by (4.14). $E(R)$ is the solution of an information-theoretic optimization problem (an info-max single-letter optimization task) that is a function of the model $P_{X,Y}$ (under H_0) and $R > 0$ (the operational constraint). Importantly, this asymptotic expression is obtained when considering an arbitrary small TYPE I error restriction (parametrized by $\epsilon > 0$). We will analyze the magnitude of $E(R)$ when compared with the equivalent result known for the non-collaborative setting presented next.

4.3.2 Non-Collaborative Hypothesis Testing

For the non-collaborative setting introduced in Section 4.2.2, we have the counterpart of Theorem 4.1, where its respective error exponent is also expressed as a single letter info-max optimization that is a function of $P_{X,Y}$ and R:

Theorem 4.2 *[28, Th.1]. Given the model $P_{X,Y}$ in (4.3) and $R > 0$[4]:*

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} -\frac{1}{n} \log \beta_n(\epsilon, R) = \xi(R) = \max_{\substack{P_{U|X}: \mathbb{X} \mapsto \mathcal{P}(\mathbb{U}) \\ (X,Y,U) \sim P_{Y|X} \cdot P_{U|X} \cdot P_X \\ I(U;X) \leq R \\ |\mathbb{U}| \leq |\mathbb{X}|+1}} I(U;Y). \tag{4.16}$$

4.3.3 Discussion of the Results

i) Our main result in Theorem 4.1 establishes an asymptotic limit for the TYPE II error when we consider an arbitrary small TYPE I error parameterized by $\epsilon > 0$ in (4.13). The proof of Theorem 4.1 has two parts: an achievable argument (i.e., the construction of a decision scheme) and an impossibility argument that shows the optimality of the proposed construction. It is worth noting that testing against independence in a cooperative scenario was first studied in [29] for the case of a single round of interaction. They show that the expression in (4.14) offers a lower bound for the limit of the TYPE II error in (4.13). Our proof completes this analysis adding an impossibility argument that proves that the obtained lower bound is optimal.

ii) The achievable argument mentioned implies the construction of a scheme, which is based on properties of typical sequences [2]. In a nutshell, given a sequence X_1^n, Node

[4] As a side comment, in [82, Th.1] we presented a non-trivial extension of Theorem 4.2 exploring vanishing Type I error restriction (function of n). We show that if $(1/\epsilon_n)_n = o(e^{rn})$ for any $r > 0$, then $\lim_{n \to \infty} -\frac{1}{n} \log \beta_n(\epsilon_n, R) = \xi(R)$, where $\xi(R)$ has the same expression presented in (4.16)

1 searches in its respective codebook: if the sequence belongs to their typical set, the sequence is re-transmitted to Node 2, if not, Node 1 declares H_1. Then, Node 2 repeats the same strategy; that is, Node 2 searches in its respective codebook and retransmits to Node 1 if the sequence belongs to the codebook of Node 2. The details are presented in Appendix 4.8.1. Importantly, this construction is used later for the task of designing a practical detector in Section 4.5.2.

iii) Comparing the obtained error exponents, i.e., $E(R)$ in (4.14) vs. its respective non-collaborative expression $\xi(R)$ in (4.16), we observe that both are the solution of a single letter optimization task function of the model $P_{X,Y}$ and $R > 0$. These optimizations are very similar and related. Both objectives use information measures with information restrictions. Importantly, in the single-letter task used to obtain $E(R)$, a non-zero additive term given by $I(V, X_1^n|U) > 0$ is observed with respect to the related expression used to determine $\xi(R)$. This extra information component offers a non-zero gain in the asymptotic error-exponent. These results can be used to argue that collaboration in theory and asymptotically offers better performance: i.e., $E(R) \geq \xi(R)$ (see the argument in Appendix 4.8.4). The inclusion of this additional term $I(V, X_1^n|U) > 0$ in (4.14) comes from the ability of re-transmission (see the proof and (4.45)), which is exclusive of the collaborative setting. This observation (a the new additive term) will be exploited in the next Section where we use these two single-letter optimizations as criteria for designing both (f_n, g_n, ϕ_n) and (f_n, ϕ_n).

In summary, we observe that in contrasting the non-collaborative versus the collaborative strategy, there is a discrepancy in the error exponents expressed analytically by the extra term $I(V, X|U) \geq 0$ in (4.13). This conditional MI behaves as an additive information gain that emerges exclusively when we have the ability of collaboration. To complement this theoretical observation, in the next two Sections we will evaluate the benefit of collaboration when n is finite (a non-asymptotic analysis). Importantly for a finite $n > 0$, selecting (f_n, ϕ_n) and (f_n, g_n, ϕ_n) as being the solutions to the operational problems in (4.7) and (4.12) respectively is not possible. To address this in practice, we look at the task of designing practical encoders and a decoder for each strategy using samples of the distributions presented in (4.14) and (4.16). For these designs, we will propose new data-driven approaches that are based on the adoption of the info-max principles presented in (4.14) and (4.16). For the optimizations, we adopt some ideas of machine learning to select a collection of expressive parametric distributions and the stochastic gradient decent (SGD) to make the problems tractable.

4.4 Data-Driven Design:Non-Collaborative

In this and the next section, we introduce an info-max learning principle for the design of the encoders and decoders of the two distributed inference strategies presented in Section 4.2. We begin with the non-collaborative approach.

Our basic conjecture is that designing the encoder from (4.16) would improve the detection performance at the decision stage. For that, we propose a concrete data-driven info-max criterion for selecting $f_n(\cdot)$ (and implicitly $\phi_n(\cdot)$) for testing independence.

4.4.1 The Multi-letter Info-Max Problem

From our analysis of the previous section, the fundamental limit of the TYPE II error in (4.16) is given by the mutual information (MI) maximization between a soft (lossy) representation of X (represented by U) and the class level Y. Using this optimization, we propose to maximize the MI between the representation $U = f_n(X_1^n)$ and Y_1^n given a size constraint on the range of $f_n(\cdot)$. This yields the following info-max problem

$$\max_{f_n:\mathbb{X}^n\to\mathbb{U}=\{1,\ldots,|\mathbb{U}|\}} I(U = f_n(X_1^n); Y_1^n). \tag{4.17}$$

The problem in (4.17) can be interpreted as a multi-letter version of (4.16), where we use deterministic mappings (quantizers) instead of the soft mappings (conditional probabilities) expressed in (4.16).

4.4.2 Approximations and Design Considerations

The problem in (4.17) is combinatorial and non-tractable for large n. Some approximations are needed to make it numerically tractable and to estimate MI. For this, we relax some assumptions making the problem data-driven.

Empirical Version of (4.17)

First, we note that the sequence $U = f(X_1^n) \multimap X_1^n \multimap Y_1^n$ (and the model $P_{U,X,Y}$) forms a *Markov chain*. From this, the mutual information $I(U; Y_1^n)$ in (4.17) can be conveniently expressed as

$$\sum_{\substack{x_1^n\in\mathbb{X}^n\\ u\in\mathbb{U}}} P_{U|X_1^n} P_{X_1^n} \log\left(\frac{1}{\sum\limits_{x_1^n\in\mathbb{X}^n} P_{U|X_1^n} P_{X_1^n}}\right) - \sum_{\substack{x_1^n\in\mathbb{X}^n\\ u\in\mathbb{U}\\ y_1^n\in\mathbb{Y}^n}} P_{Y_1^n|X_1^n} P_{U|X_1^n} P_{X_1^n} \log\left(\frac{\sum\limits_{x_1^n\in\mathbb{X}^n} P_{Y_1^n|X_1^n} P_{X_1^n}}{\sum\limits_{x_1^n\in\mathbb{X}^n} P_{Y_1^n|X_1^n} P_{U|X_1^n} P_{X_1^n}}\right) \tag{4.18}$$

where $P_{X_1^n}, P_{Y_1^n|X_1^n}$ and $P_{U|X_1^n}$ are short-hand notations for $P_{X_1^n}(x_1^n)$, $P_{Y_1^n|X_1^n}(y_1^n|x_1^n)$ and $P_{U|X_1^n}(u|x_1^n)$, respectively.

In practical communication problems, the distribution of the sources at the nodes is often unknown. Then, instead of assuming $P_{X_1^n}$ in (4.18), we assume a training (i.i.d) set $\{\overline{x}_1, \ldots, \overline{x}_m\}$, with $\overline{x}_i \in \mathbb{X}^n$ that will be used to approximate the expectations in (4.18) (w.r.t. $P_{X_1^n}$) by their respective empirical means. In addition for large n, the second expectation is impractical to compute. For that, we assume i.i.d. samples $S_{x_1^n} = \{\overline{y}_1, \ldots, \overline{y}_{m'}\}$ of Y_1^n given $X_1^n = x_1^n$. These conditional samples are used to approximate the expectation in (4.18) (w.r.t. $P_{X_1^n, Y_1^n}$) by their respective empirical average. Then, our empirical (and

numerically tractable) version of $I(U;Y_1^n)$ is written as

$$\hat{I}_\alpha(U;Y_1^n) \equiv \frac{1}{m}\sum_{i=1}^{m}\sum_{u\in\mathbb{U}} P_{U|X_1^n}(u|\overline{x}_i)\log\left(\frac{1}{\frac{1}{m}\sum_{l=1}^{m} P_{U|X_1^n}(u|\overline{x}_l)}\right)$$
$$-\frac{\alpha}{mm'}\sum_{i=1}^{m}\sum_{\overline{y}_j\in S_{\overline{x}_i}}\sum_{u\in\mathbb{U}} P_{U|X_1^n}(u|\overline{x}_i)\cdot\log\left(\frac{\sum_{l=1}^{m} P_{Y_1^n|X_1^n}(\overline{y}_j|\overline{x}_l)}{\sum_{l=1}^{m} P_{Y_1^n|X_1^n}(\overline{y}_j|\overline{x}_l)P_{U|X_1^n}(u|\overline{x}_l)}\right). \quad (4.19)$$

The derivation of this equation is presented in Appendix 4.8.6.

Soft-quantizers based on Boltzmann distributions

We need to determine the collection of models $P_{U|X_1^n}$ in (4.19). Every $P_{U|X_1^n}$ is directly linked to the encoder $f_n(\cdot)$. Instead of using a deterministic mapping to index $P_{U|X_1^n}$, we consider more general soft-quantizers (or conditional distributions from X_1^n to U). We relax this deterministic assumption to enrich the space of hypotheses used to solve (4.17). In the process, our problem connects naturally with the type of info-max optimization addressed in representation learning for selecting the encoder [98, 99]. Following [91], we consider the family of *Boltzmann distribution* [46] as it is a rich and expressive collection of parametric distributions for $P_{U|X_1^n}$. More precisely, we consider the following family of conditional models (soft-encoders):

$$p^W(u|x_1^n) \equiv P^{W,\tau}_{U|X_1^n}(u|x_1^n) = \frac{e^{-\frac{\tau\|w_u - x_1^n\|^2}{2}}}{\sum_{l\in\mathbb{U}} e^{-\frac{\tau\|w_l - x_1^n\|^2}{2}}}, \quad (4.20)$$

where W is a weight matrix $W\in\mathbb{R}^{n\times|\mathbb{U}|}$, given by $W=[w_1;\ldots;w_{|\mathbb{U}|}]$ and $w_u\in\mathbb{X}^n$, $\forall u\in\mathbb{U}$. This family of conditional distributions has been widely used in machine learning because of its expressiveness and learning properties [100, 101].

Info-Max Learning

At the end, we address our main design problem as follows:

$$W^* = \arg\max_{W\in\mathbb{R}^{n\times|\mathbb{U}|}} \hat{I}_\alpha(U;Y_1^n). \quad (4.21)$$

Importantly, the expression in (4.21) is smooth and differentiable with respect to W, i.e., our collection $(p^W(u|x_1^n))$. Consequently, the solution of (4.21) can be approximated based on the Stochastic Gradient Descent (SGD) algorithm [102]. A pseudo-code for this mutual information maximization algorithm is provided here:

The encoder $f_n(\cdot)$

Finally, the solution in (4.21) is a weight matrix that produces a soft quantizer (or conditional distribution). To obtain a hard-quantizer or encoder (denoted by $f_n^W(\cdot)$), we use the MAP (or soft-max) rule:

$$f_n^W(x_1^n) = \arg\max_{u\in\mathbb{U}} \frac{e^{-\frac{\tau\|w_u - x_1^n\|^2}{2}}}{\sum_{l\in\mathbb{U}} e^{-\frac{\tau\|w_l - x_1^n\|^2}{2}}}^{5}. \quad (4.22)$$

[5] For this analysis we set $\tau = 1$

Algorithm 1 Unidirectional Mutual Information Maximization

Initialize: τ, α, λ, m (number of iterations), W
for $i \leftarrow 0$ **to** m **do**
  $p^W(u|\overline{x}_i) \leftarrow e^{-\frac{\tau \|w_u - \overline{x}_i\|^2}{2}} \; u \in \mathbb{U}$
  $p^W(u|\overline{x}_i) \leftarrow \frac{p^W(u|\overline{x}_i)}{\sum_{l \in \mathbb{U}} p^W(u|\overline{x}_l)}$
  $w_u \leftarrow w_u - \lambda \frac{\partial \hat{I}_\alpha(Y_1^n;U)}{\partial w_u}$
end for
Result: Prediction $f(\overline{x}_i) = \arg\max_{u \in \mathbb{U}} p^W(u|\overline{x}_i)$

Decision Stage: $\phi_n(\cdot)$

Given $U = f_n^W(X_1^n)$ and Y_1^n, the decision rule $\phi_n(\cdot)$ is given by the (optimal) *Neyman-Pearson (NP) test* acting on U, which is the optimal decision rule given (U, Y_1^n), and it offers the optimal trade-off between the two types of errors. Therefore, the decision (decoder) is given by the family $\phi_n^\tau(u, y_1^n) = 0$ if $\frac{P_{U,Y_1^n}(u,y_1^n)}{P_U(u)P_{Y_1^n}(y_1^n)} > \tau$ and $\phi_n^\tau(u, y_1^n) = 1$, otherwise.

4.5 Data-Driven Design: Collaborative

To design the collaborative setting in Fig. 4.3, we need to construct two encoders $f_n(X_1^n)$ and $g_n(f(X_1^n), Y_1^n)$ and the detector $\phi_n(\cdot)$. As in the previous section, we use the information limit, in this case stated in Theorem 4.1, to inform the objective function (or loss) needed to select from data $f_n(\cdot)$ and $g_n(\cdot)$. In particular, the fundamental limit of the TYPE II error in (4.13) is a specific single letter optimization that expresses the role played by the two encoders of the problem: maximizing the MI between a soft lossy version of X (represented by U) and Y plus a conditional MI between a lossy representation of (U, Y) (represented by V) and X. The two compressed terms U and V (or latent variables) represent the role played by $f_n(\cdot)$ and $g_n(\cdot)$, respectively. Consequently, using the single letter optimization in (4.14) as a proxy, we maximize the multi-letter version of this problem

$$\max_{f_n(\cdot):\mathbb{X}^n \to \mathbb{U}, g_n(\cdot):\mathbb{U}\times\mathbb{Y}^n \to \mathbb{V}} I(U; Y_1^n) + I(V; X_1^n|U) \tag{4.23}$$

where $U = f_n(X_1^n)$, $V = g(f(X_1^n), Y_1^n)$ for a given a cardinality constraint on $\mathbb{U}$ and $\mathbb{V}$.

The expression in Eq. (4.23) has two information components $I(U; Y_1^n)$ and $I(V; X_1^n|U)$. A key observation on this is that the first term $I(U; Y_1^n)$ depends only on $f_n(\cdot)$. Consequently — and for numerical simplicity — we decided to address the optimization in (4.23) sequentially: First we optimize $f_n(\cdot)$ and with that solution solve the problem for $g_n(\cdot)$. More precisely, we use (4.17) to solve $f_n(\cdot)$ (the first optimization). Then, fixing $f_n(\cdot)$, we solve $g_n(\cdot)$ by maximizing $I(V; X_1^n|U)$ from (4.23). The focus of the next subsections is to present a numerically effective way to address the second optimization task: $\max_{g(\cdot)} I(V; X_1^n|U)$. This reduces to find a practical way to estimate $I(V; X_1^n|U)$.

4.5.1 Approximations and Design Considerations

Empirical Version of $I(V;X_1^n|U)$

Using the fact that $V \multimap (U, Y_1^n) \multimap X_1^n$ and the definition of the conditional MI [97], we have that $I(V;X_1^n|U)$ can be conveniently expressed as

$$I(V;X_1^n|U) = \sum_{\substack{x_1^n\in\mathbb{X}^n \\ u\in\mathbb{U} \\ v\in\mathbb{V} \\ y_1^n\in\mathbb{Y}^n}} P_{V|Y_1^n,U} P_{U|X_1^n} P_{Y_1^n|X_1^n} P_{X_1^n} \log\left(\frac{\sum\limits_{x_1^n\in\mathbb{X}^n} P_{U|X_1^n} P_{X_1^n}}{\sum\limits_{\substack{x_1^n\in\mathbb{X}^n \\ y_1^n\in\mathbb{Y}^n}} P_{V|Y_1^n,U} P_{U|X_1^n} P_{Y_1^n|X_1^n} P_{X_1^n}}\right)$$

$$- \sum_{\substack{x_1^n\in\mathbb{X}^n \\ u\in\mathbb{U} \\ v\in\mathbb{V} \\ y_1^n\in\mathbb{Y}^n}} P_{V|Y_1^n,U} P_{U|X_1^n} P_{Y_1^n|X_1^n} P_{X_1^n} \log\left(\frac{1}{\sum\limits_{y_1^n\in\mathbb{Y}^n} P_{V|Y_1^n,U} P_{Y_1^n|X_1^n}}\right) \tag{4.24}$$

where $P_{X_1^n}, P_{Y_1^n|X_1^n}, P_{V|Y_1^n,U}$ and $P_{U|X_1^n}$ are short-hand notations for $P_{X_1^n}(x_1^n)$, $P_{Y_1^n|X_1^n}(y_1^n|x_1^n)$, $P_{V|Y_1^n,U}(v|y_1^n,u)$ and $P_{U|X_1^n}(u|x_1^n)$, respectively. Instead of assuming $P_{X_1^n}$ in (4.24), we assume a training set (i.i.d. samples of $P_{X_1^n}$) $\{\overline{x}_1,\ldots,\overline{x}_m\} \subset \mathbb{X}^n$. The samples of $P_{X_1^n}$ are used to approximate the expectations in (4.24) (w.r.t. $P_{X_1^n}$) by their respective empirical means. In addition, we generate i.i.d. of the conditional distribution of Y_1^n given $X_1^n = x_1^n$. We denote these conditional sets by $S_{x_1^n} = \{\overline{y}_1,\ldots,\overline{y}_{m'}\} \subset \mathbb{Y}^n$ for any $x_1^n \in \mathbb{X}^n$. Consequently, a semi-empirical version of $I(V;X_1^n|U)$ is given/denoted by $\hat{I}_\alpha(V;X_1^n|U) \equiv$

$$= \frac{1}{mm'} \sum_{i=1}^{m} \sum_{\overline{y}_j\in S_{\overline{x}_i}} \sum_{\substack{u\in\mathbb{U} \\ v\in\mathbb{V}}} P_{V|Y_1^n,U}(v|\overline{y}_j,u) P_{U|X_1^n}(u|\overline{x}_i) \log\left(\frac{\sum\limits_{l=1}^{m} P_{U|X_1^n}(u|\overline{x}_l)}{\frac{1}{m'}\sum\limits_{l=1}^{m}\sum\limits_{\overline{y}_j\in S_{\overline{x}_l}} P_{V|Y_1^n,U}(v|\overline{y}_j,u) P_{U|X_1^n}(u|\overline{x}_l)}\right)$$

$$- \frac{\alpha}{mm'} \sum_{i=1}^{m} \sum_{\overline{y}_j\in S_{\overline{x}_i}} \sum_{\substack{u\in\mathbb{U} \\ v\in\mathbb{V}}} P_{V|Y_1^n,U}(v|\overline{y}_j,u) P_{U|X_1^n}(u|\overline{x}_i) \log\left(\frac{1}{\sum\limits_{\overline{y}_j\in S_{\overline{x}_i}} P_{V|Y_1^n,U}(v|\overline{y}_j,u)}\right). \tag{4.25}$$

The derivation of this equation is presented in Appendix 4.8.7.

Soft quantizers based on Boltzmann distributions

We need to determine the collection of models $P_{V|Y_1^n,U}$ used in (4.25). $P_{V|Y_1^n,U}$ is directly linked to the encoder $g_n(\cdot)$. As in the case of $P_{U|X_1^n}$ (see Section 4.4.2), instead of using a deterministic mapping, we consider more general soft-quantizers (or conditional distributions from $\mathbb{U}\times\mathbb{Y}^n$ to $\mathbb{V}$) given by

$$p^{W_2}(v|\overline{y}_j,u) \equiv P_{V|U,Y_1^n}^{W_2,\tau}(v|\overline{y}_j,u) = \frac{e^{-\frac{\tau||w_v^2-(\overline{y}_j,u)||^2}{2}}}{\sum_{\tilde{v}\in\mathbb{V}} e^{-\frac{\tau||w_{\tilde{v}}^2-(\overline{y}_j,u)||^2}{2}}}, \tag{4.26}$$

where $W_2 \in \mathbb{R}^{(n+1)\times|\mathbb{V}|}$ is a weight matrix such that $W_2 = [w_1^2;...;w_{|\mathbb{V}|}^2]$ and $w_v^2 \in \mathbb{U}\times\mathbb{Y}^n$, $\forall v \in \mathbb{V}$ and $U = f(X_1^n)$ is the solution of (4.22). With this parametric selection, our main

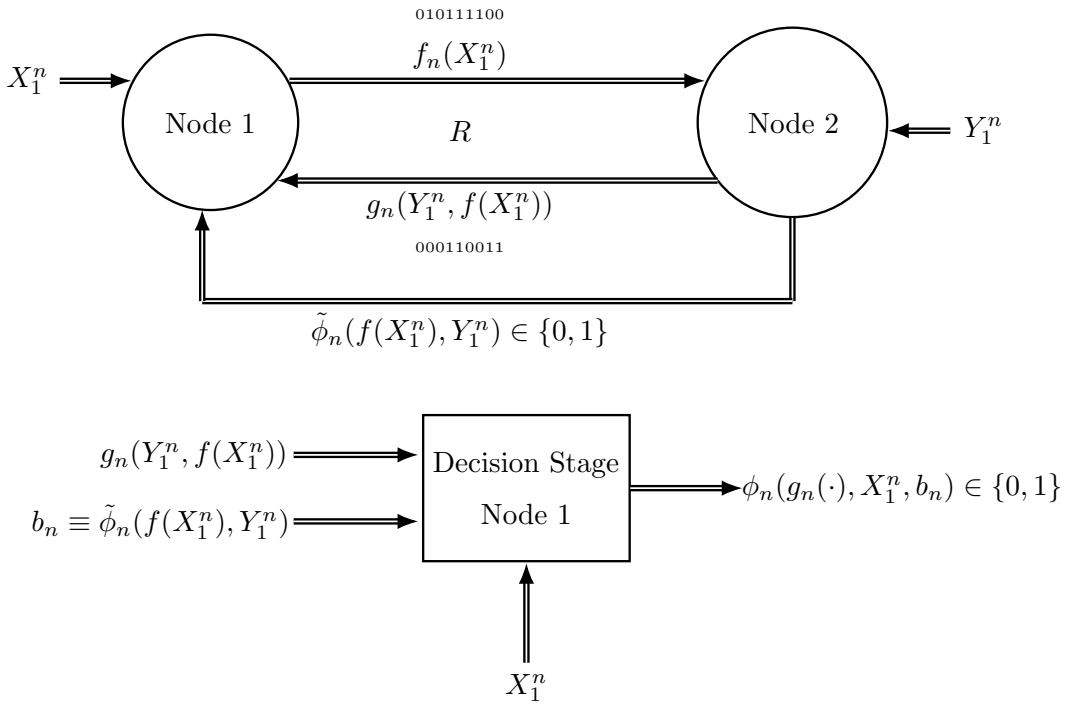


Figure 4.4: Collaborative strategy to detect H_0 and H_1 given an overall rate-communication constraint.

design problem is

$$W_2^* = \arg \max_{W_2 \in \mathbb{R}^{(n+1)\times|\mathbb{V}|}} \hat{I}_\alpha(V; X_1^n|U). \tag{4.27}$$

Finally, the solution in (4.27) is a weight matrix that produces a conditional probability (or soft quantizer). The hard-quantizer or encoder (denoted by $g_n^{W_2}(\cdot)$) is obtained with the MAP (soft-max) rule:

$$g_n^{W_2}(Y_1^n, U = f_n^W(X_1^n)) = \arg\max_{v\in\mathbb{V}} \frac{e^{-\frac{\tau||w_v^2-(Y_1^n,U)||^2}{2}}}{\sum_{\tilde{v}\in\mathbb{V}} e^{-\frac{\tau||w_{\tilde{v}}^2-(Y_1^n,U)||^2}{2}}}. \tag{4.28}$$

Algorithm 2 Collaborative Mutual Information Maximization

1: Initialize: τ, α, λ, m (number of iterations), W (pretrained), W_2
2: **for** $i \leftarrow 0$ **to** m **do**
3: $\quad p^{W_2}(v|y_1^n, u) \leftarrow e^{-\frac{\tau\|w_v^2-(y_1^n,u)\|^2}{2}} \quad v \in \mathbb{V}$
4: $\quad p^{W_2}(v|y_1^n, u) \leftarrow \frac{p^{W_2}(v|y_1^n,u)}{\sum_{v\in\mathbb{V}} p^{W_2}(v|y_1^n,u)}$
5: $\quad w_v \leftarrow w_v - \lambda \frac{\partial \hat{I}_\alpha(V;X_1^n|U)}{\partial w_v}$
6: **end for**
7: **Result:** Prediction $g(f(\overline{x}_i), y_1^n) = \arg\max_{v\in\mathbb{V}} p^{W_2}(v|y_1^n, u)$

4.5.2 Decision Stage:

Assuming $f_n(\cdot)$ and $g_n(\cdot)$, we need to design the decision rule $\phi_n(\cdot)$. Because of the non-trivial interaction between the nodes in this setting (see Fig. 4.3), the design of $\phi_n(\cdot)$ does not follow from a simple adoption of the NP test (used in Section 4.4.2 for the non-collaborative setting). Instead, our solution is inspired by the achievable construction used in the proof of Theorem 4.1[6]. In this construction, there are two decision rules: one rule taken by Node 2 after receiving the information from Node 1, i.e., acting on $(f_n(x_1^n), y_1^n)$ as evidence, and a second rule located at Node 1 that uses the preliminary decision of Node 2 (1 bit of information) and the information (bits) from Node 2, i.e., $v = g_n(f(x_1^n, y_1^n))$, as evidence. This process is illustrated in Fig. 4.4. More precisely, we propose the following collaborative two-stage detection scheme.

i) For the first decision, Node 2 runs an optimal NP test based on the evidence $(f_n(x_1^n), y_1^n)$ at this stage of the process (first-round).
Let us define accordingly the following decision region for the alternative hypothesis (H_1):

$$A_{f_n,n}^{\tau} = \left\{ (x_1^n, y_1^n) \in \mathbb{X}^n \times \mathbb{Y}^n : \log \left(\frac{P_{f(X_1^n)Y_1^n}(f(x_1^n), y_1^n)}{P_{f(X_1^n)}(f(x_1^n)) P_Y^n(y_1^n)} \right) > \log(\tau) \right\} \tag{4.29}$$

function of the threshold $\tau > 0$. Note that $(A_{f_n,n}^{\tau}, (A_{f_n,n}^{\tau})^c)$ corresponds to a partition of the observation space induced by the classical NP test. This is the optimal strategy at this stage.

ii) For the second and final decision stage, Node 1 receives the evidence $(v = g_n(f(x_1^n, y_1^n)), x_1^n)$ and one extra bit indicating if $(x_1^n, y_1^n) \in A_{f_n,n}^{\tau}$ (from stage 1). This final stage aggregates information in the following way:

   A: Under the condition that $(x_1^n, y_1^n) \in A_{f_n,n}^{\tau}$, Node 2 runs an optimal NP test based on (v, x_1^n) and a threshold $\tilde{\tau} > 0$. More precisely, it decides H_0 if

$$\log \left(\frac{P_{g(Y_1^n, f(X_1^n))X_1^n}(g_n(y_1^n, f(x_1^n)), x_1^n)}{P_{g(Y_1^n, f(X_1^n))}(g_n(y_1^n, f(x_1^n))) P_X^n(x_1^n)} \right) > \log(\tilde{\tau})$$

   or H_1 otherwise.

   B: Under the condition that $(x_1^n, y_1^n) \notin A_{f_n,n}^{\tau}$, Node 1 trust Node 2's judgment and decides the alternative H_1.

4.5.3 Error Computation

Finally, it is worth expressing the two types of errors of this joint collaboration scheme and comparing them with their respective expression for the one-side approach. The unidirectional setting has the following TYPE I and TYPE II errors given, for a fixed $\tau > 0$, by

$$\alpha_n((f_n, \phi_n)) \equiv P_{XY}^n((A_{f_n,n}^{\tau})^c) \text{ and} \tag{4.30}$$

$$\beta_n((f_n, \phi_n)) \equiv Q_{XY}^n(A_{f_n,n}^{\tau}). \tag{4.31}$$

[6] In particular, the construction used to prove Lemma 4.8.1 in Appendix 4.8.1.

For the collaborative case, for a fixed $\tau, \tilde{\tau} > 0$ let us consider, additionally to $A^{\tau}_{f_n,n}$, the following set

$$B^{\tilde{\tau}}_{f_n,g_n,n} = \left\{ (x_1^n, y_1^n) \in \mathbb{X}^n \times \mathbb{Y}^n : \log \left(\frac{P_{g(Y_1^n, f(X_1^n))X_1^n}(g_n(y_1^n, f(x_1^n)), x_1^n)}{P_{g(Y_1^n, f(X_1^n))}(g_n(y_1^n, f(x_1^n)))P_X^n(x_1^n)} \right) > \log(\tilde{\tau}) \right\}. \tag{4.32}$$

Then, we divide the decision stage into two options:

A: under the condition that $(x_1^n, y_1^n) \in A^{\tau}_{f_n,n}$, Node 1 runs an optimal NP test based on the partition induced by $B^{\tilde{\tau}}_{f_n,g_n,n}$ for $\tilde{\tau} > 0$.

B: otherwise, if $(x_1^n, y_1^n) \notin A^{\tau}_{f_n,n}$, Node 1 trusts Node 2's judgment and decides the alternative H_1.

Then, from these conditions, we have that the TYPE I error occurs when Node 2 declares H_1 given H_0 (using $(A^{\tau}_{f_n,n})^c$) or when Node 1 declares H_1 given H_0 (using $(B^{\tau}_{f_n,n})^c$ and $A^{\tau}_{f_n,n}$). On the other hand, the TYPE II error occurs when Node 1 declares H_0 given H_1 (using $B^{\tau}_{f_n,n}$ and $A^{\tau}_{f_n,n}$). Formally. the expressions are given by

$$\alpha_n((f_n, g_n, \phi_n)) \equiv P^n_{XY}((A^{\tau}_{f_n,n})^c) + P^n_{XY}((B^{\tilde{\tau}}_{f_n,g_n,n})^c \cap A^{\tau}_{f_n,n}) \text{ and} \tag{4.33}$$

$$\beta_n((f_n, g_n, \phi_n)) \equiv Q^n_{XY}(B^{\tilde{\tau}}_{f_n,g_n,n} \cap A^{\tau}_{f_n,n}). \tag{4.34}$$

Comparing these expressions with Eq.(4.30) and (4.31), we note that there is no evident performance relationship between the two schemes. The main reason is that for a fixed rate $R > 0$, Node 1 in the non-collaborative setting has more bits assigned to it than its respective counterpart in the collaborative setting. This justifies the numerical analysis presented in Section 4.6, where under different symmetric conditions for $P_{X,Y}$, we compare (numerically) the trade-off derived from the expressions in Eqs.(4.33)-(4.34).

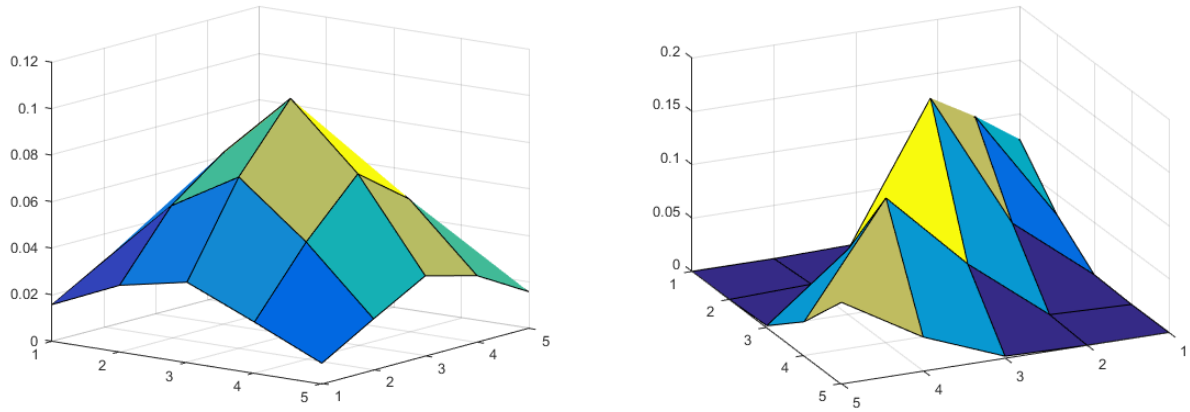
Finally, it is worth mentioning that the two errors in (4.33) and (4.34) are functions of two parameters $(\tau, \tilde{\tau})$ of this scheme. Then, we have two degrees of freedoms, i.e., a bi-dimensional space of plausible solutions, that produces different trade-offs between TYPE I and TYPE II errors. To simplify this exploration, we consider a functional one-to-one relationship between $(\tau, \tilde{\tau})$ using a monotone mapping between τ and $\tilde{\tau}$ given by $\tilde{\tau} = r(\tau) = \tau^{\alpha}$ with $\alpha \in (0, 100]$. These mappings offer the possibility of exploring a rich collection of performance trade-offs without compromising the expressiveness of the original 2D space $(\tau, \tilde{\tau})$. In the next section, we show this strategy to explore the 2D parameter domain.

4.6 Numerical Analyses

We evaluate the performance of the collaborative and non-collaborative strategies proposed in this work to see if we can ratify the performance discrepancy predicted by the error exponent results in Section 4.3. We present two empirical analyses. First, we evaluate the performance of the info-max encoder $f_n^W(\cdot)$ in (4.22) and compare it with some other quantization design principle across different quantization sizes $|\mathbb{U}|$ (associated to the rate) and sample-length n. Secondly and more importantly, we compare the collaborative scheme and the non-collaborative scheme presented in this work under different rates and sample lengths scenarios. In this last part, we evaluate the effect of collaboration between the nodes in terms of performances as a function of some structural attributes of the model (i.e., $P_{X,Y}$).

4.6.1 Preliminary Analysis

For the experimental setting, accordingly to Section 4.4, we consider a joint space of size $|\mathbb{X}|\times|\mathbb{Y}|= 5\times 5$. We derive a discrete probability by partitioning $\mathbb{R}^2$ with a Gaussian density in $\mathbb{R}^2$ of parameters $(\mu_X, \mu_Y, \sigma_X, \sigma_Y.\rho)$. To this end, we assume that $\mu_X = \mu_Y$ and $\sigma_X^2 = \sigma_Y^2$ where ρ is the cross-covariance. Using this construction, we control the statistical dependency induced in the vector (X, Y) with the parameter $\rho > 0$ of the continuous model. Figures 4.5 (a) and (b) show the joint distribution P_{XY} for $\rho = 0.5$ and $\rho = 0.8$, respectively.



(a) (b)

Figure 4.5: Distribution of P_{XY} for the non-collaborative experiment. Figures (a) and (b) show a correlation coefficient of $\rho = 0.5$ and $\rho = 0.8$, respectively.

		Effective Cardinality (EC)			
		0.000512	0.00256	0.4	0.8
ρ	0.5	0.0614	0.0018	0.0015	0.0009
	0.7	0.0820	0.0029	0.0012	0.001

Table 4.1: ROC curve performance discrepancy for different level statistical dependency or discrimination (indexed by ρ) and for different effective cardinalities (measured by $\frac{|\mathbb{U}|}{|\mathbb{X}|^n}$).

Figures 4.6 presents the ROC curve for different level statistical dependency or discrimination (indexed by ρ), for different quantization levels $|\mathbb{U}|$, and for different values of $n \in \{2, 4, 6\}$. In Figure 4.7, we also contrast our strategy (continuous line) with an unsupervised method that is agnostic to the task and only tries to preserve the information of X_1^n (dashed line). More precisely, the unsupervised method is the solution of $\max_{f_n:\mathbb{X}^n\to\mathbb{U}=\{1,\dots,|\mathbb{U}|\}} I(U; X_1^n)$. In all the settings (indexed by n, ρ), as $|\mathbb{U}|$ increases, the performance of both strategies improves: a large $|\mathbb{U}|$ implies that more bits $\log_2(|\mathbb{U}|)$ are transferred to the decision stage. In comparing the results (ROC curves) between the one-side approach and the non-supervise approach, we clearly see in Fig. 4.7 (a) and (b) the advantage of optimizing the encoder $f_n(\cdot)$ to maximize the MI between the representation and Y_1^n (4.17). Importantly, this advantage is more prominent when the level of dependency between X and

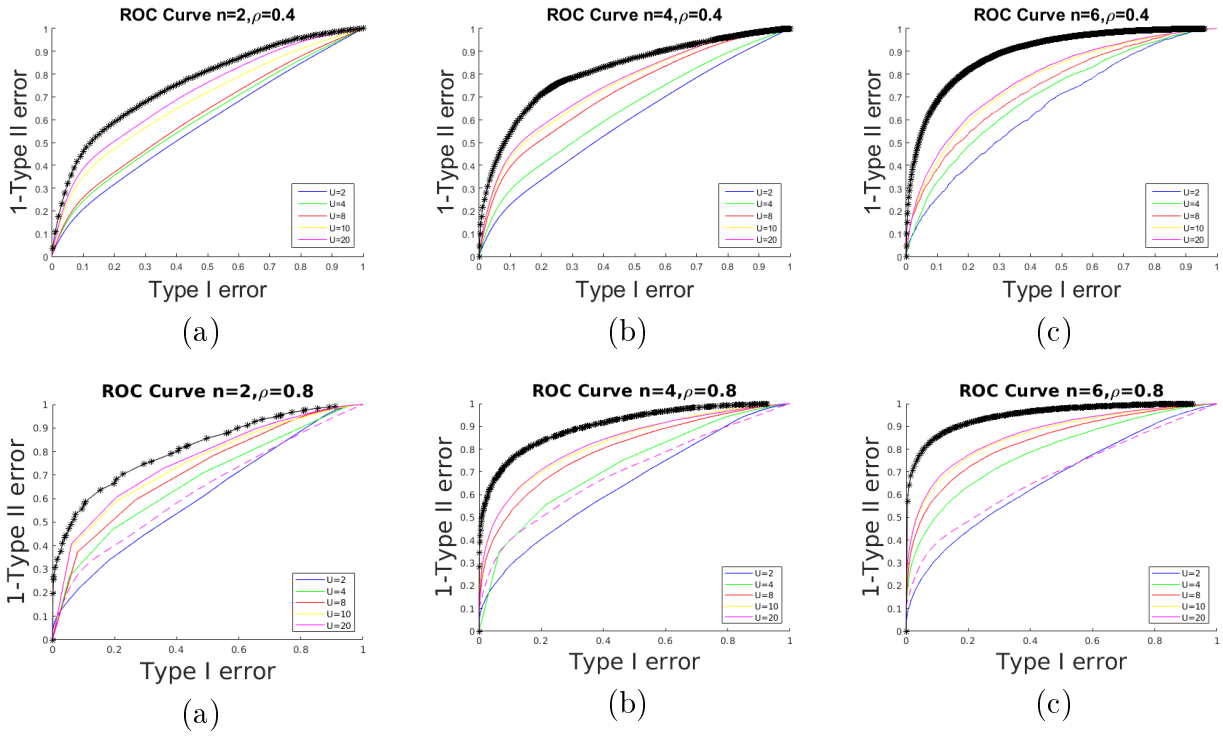


Figure 4.6: ROC curve for different levels of statistical dependency or discrimination (indexed by ρ), for different quantization levels $|\mathbb{U}|$, and for different values of n.

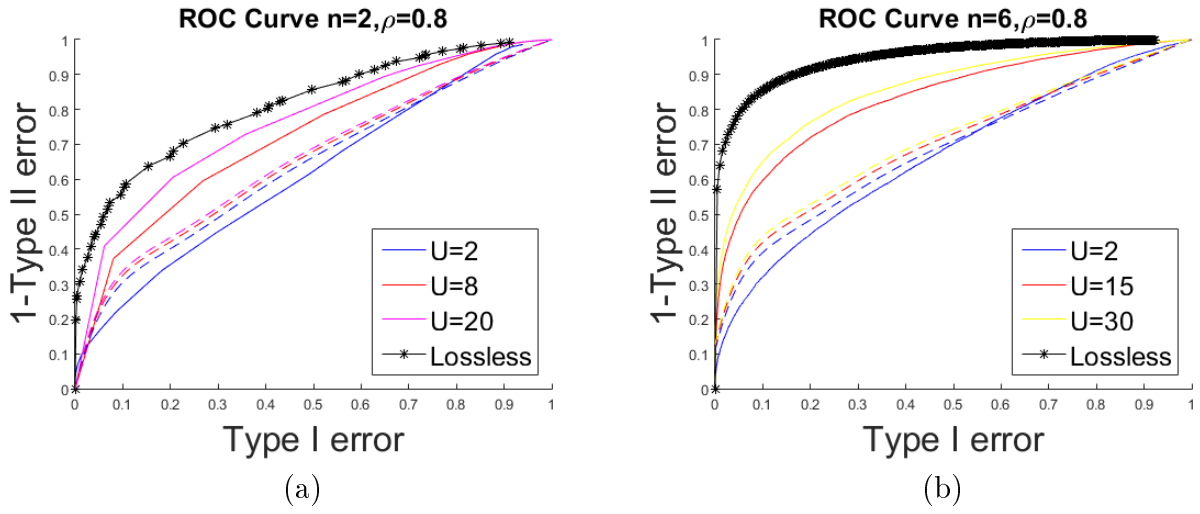


Figure 4.7: ROC curve for different levels of statistical dependency in which the algorithm is compared with respect to the unsupervised method using the same quantization level $|\mathbb{U}|$ for each color.

Y increases (by increasing ρ). In addition, there is a concrete effect of the correlation coefficient, a greater ρ implies a higher mutual information between (X, Y) and, as a consequence, a better trade-off between the errors [97].

Focusing in the expressiveness of our strategy, our info-max data-driven quantizer $f_n(\cdot)$

shows to be effective in representing with few bits the information that the lossless sample offers to discriminate H_0 from H_1. This is observed when comparing the performances of our info-max lossy strategy with the oracle (lossless) NP test acting on (X_1^n, Y_1^n), for the different regimes of mutual information. It is worth noting that $|\mathbb{U}|$ is a small (almost zero) fraction of the cardinality of $\mathbb{X}^n$, however $f_n^{W^*}(X_1^n)$ produces modest degradation in performance with respect to the use of X_1^n. To further illustrate this last point, Table 4.1 shows the reduction of the power of the test with respect to the centralized (lossless) decision case. For that, we compute the area below the ROC curve of the lossless case for different effective cardinality levels defined as $\frac{|\mathbb{U}|}{|\mathbb{X}|^n} \in (0, 1)$. Interestingly, as the effective capacity increases (relative to the alphabet size 5^n) the performance of our test increases significantly. This observation implies that although we can not achieve the optimal NP lossless results, our info-max compression scheme can achieve close to optimal NP results by using a negligible fraction of the size of $\mathbb{X}^n$.

4.6.2 Collaborative vs Non-Collaborative Analysis

In this subsection, we evaluate the effect of collaboration by comparing the performance of the two strategies presented in Sections 4.5.1 and 4.5.2, respectively. In theory, we noticed that the re-transmission from Node 2 to 1 produces additional information (expressed by the term $I(V; X_1^n|U)$ in Eq. (4.23)). Our conjecture is that this error exponent gain (or information gain) could translate into a non-asymptotic ROC performance gain: in terms of the trade-off between the two errors.

Analyzing the error exponent expressions, we observe that the information gain $E(R) - \xi(R) \geq 0$ is influenced by the similarity between the conditional distribution $P_{Y|X}$ from Node 1 to 2 (forward direction) and the conditional distribution $P_{X|Y}$ from Node 2 to 1 (backward direction). To quantify this discrepancy, we define a similarity indicator between two arbitrary conditional distributions $(\mu_{Y|X}(\cdot|x))_{x\in\mathcal{X}} \subset \mathbb{P}(\mathcal{Y})$ and $(\nu_{Y|X}(\cdot|x))_{x\in\mathcal{X}} \subset \mathbb{P}(\mathcal{Y})$ as

$$\mathcal{D}(\mu_{Y|X}\|\nu_{Y|X}|P_X) \equiv \sum_{x\in\mathbb{X}} P_X(x) \sum_{y\in\mathbb{Y}} \mu_{Y|X}(y|x) \log \frac{\mu_{Y|X}(y|x)}{\nu_{Y|X}(y|x)} \geq 0, \tag{4.35}$$

with respect to $P_X \in \mathbb{P}(\mathcal{X})$. For the symmetric case where $|\mathbb{Y}| = |\mathbb{X}|$, we propose to measure the level of symmetry or asymmetry of a joint model $P_{X,Y}$ by

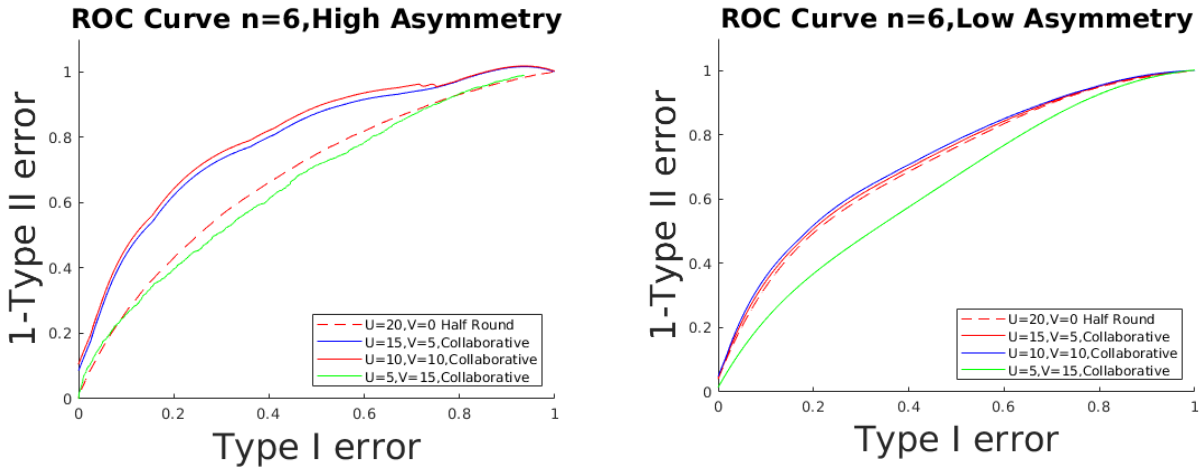
$$\Lambda(P_{X,Y}) = \mathcal{D}(P_{Y|X}\|P_{X|Y}|P_X) \geq 0, \tag{4.36}$$

where $P_{Y|X}$ and $P_{X|Y}$ denote the two conditional distributions that can be obtained from $P_{X,Y}$.

Bit Allocation

For the collaborative scheme, the allocation of quantization levels between the nodes needs to be addressed. In particular, for a fixed number of transmission symbols we want to have a good bit assignment between the nodes. Figures 4.8 (a) and (b) show the performance of different collaborative assignments of quantization levels between Node 1 and 2. The curves are also compared with its non-collaborative (half-round) case with $|\mathbb{U}| = 20$. Interestingly, we see that the overall quality of the ROC curve is governed by the assignment between

the nodes and affected at the same time by the asymmetry of the distribution. From this analysis, we observe that a good strategy is to distribute in a balanced way the bits assigned to each node. The performance of the ROC curve decreases when Node 2 has more bits available with respect to Node 1. This is consequence of the design of the quantizers that mostly uses the information contained in the first half round given by $P_{Y|X}$. This is reinforced by comparing the main results in (4.16) and (4.14), where the term $I(V, X_1^n|U)$ behaves as an additive information gain that emerges exclusively when there is re-transmission.

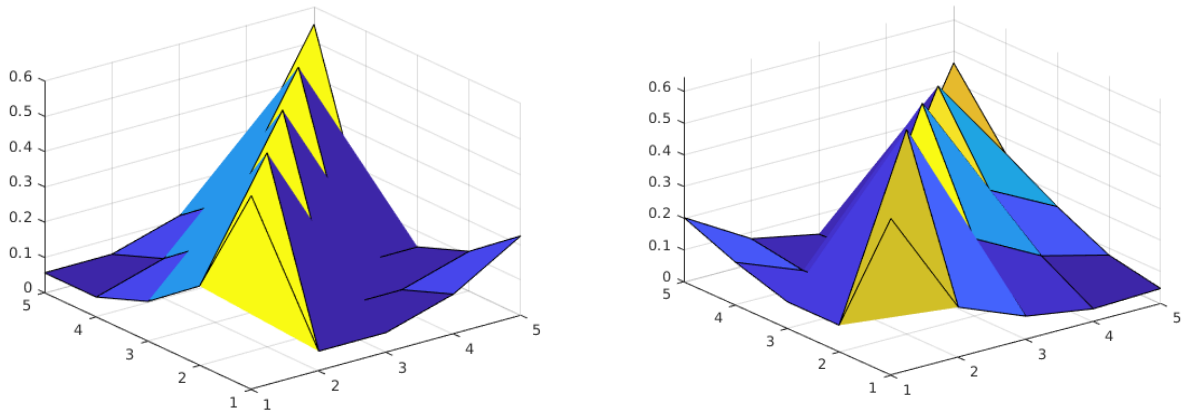


(a) (b)

Figure 4.8: ROC curves for different collaboration schemes between Node 1 and 2, using different quantization levels for $|\mathbb{U}|$ and $|\mathbb{V}|$. All these curves are compared with their corresponding half round performance using $|\mathbb{U}|= 20$.

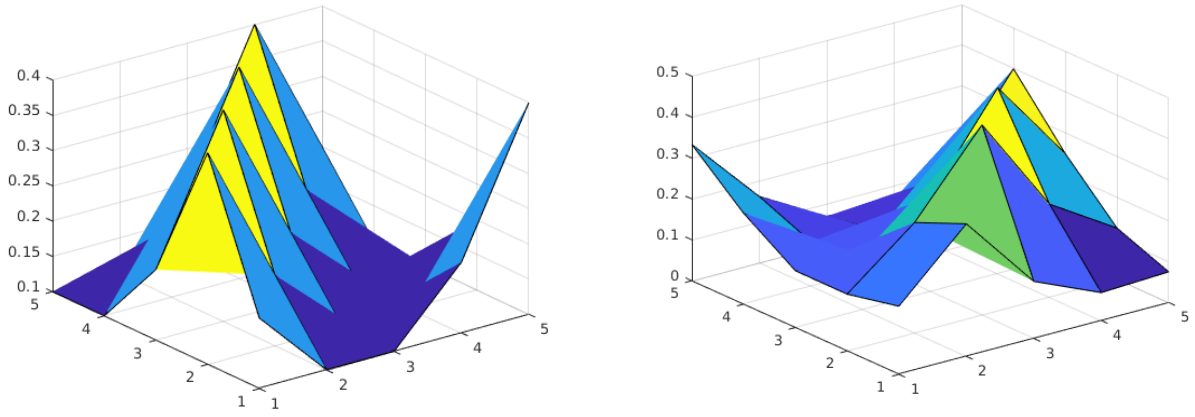
Comparison

Returning to the comparison, Figures 4.9 and 4.10 illustrate two models with low and high symmetry in terms of $\Lambda(P_{X,Y})$. Here, we want to analyze the contribution of collaboration (performance gain) when there is a discrepancy between the forward ($P_{Y|X}$) and backward channel ($P_{X|Y}$) measured by $\Lambda(P_{X,Y})$. To explore the collaborative scheme proposed in 4.5.2, we explore the ROC curves over the 2D surface parameterized by $(\tau, \tilde{\tau})$. More precisely, given a fixed τ (described in (4.29)), we generate a ROC curve exploring different values of $\tilde{\tau} \in (0, \infty)$. To simplify this exhaustive search, we consider a monotone relationship between τ and $\tilde{\tau}$ expressed by $\tilde{\tau}(\tau) = \tau^{\alpha}$ with $\alpha \in (0, 100]$. Using this selection, Figures 4.11 (a) and (b) show the performance of the ROC curve for the two cases associated with high and low symmetry, respectively. For the collaborative setting, we use $|\mathbb{U}|= 5$ and $|\mathbb{V}|= 4$ and we compared with the non-collaborative case with the same rate, i.e., $|\mathbb{U}|= 20$. In the collaborative setting, we have many solutions (indexed by α). Consequently, we obtain the ROC by the superposition of all obtained curves selecting the α value that offers the best performance trade-off in different regimes. The color zones separated with the vertical dashed black lines indicate the best α chosen for different areas of the ROC curve. Thus, from Figs 4.11(a) and 4.11(b), our collaborative scheme shows improvement with respect to the non-collaborative case. The improvement is a function of $P_{X,Y}$ and dictated by its symmetry $\Lambda(P_{X,Y})$ (illustrated in Figs. 4.10 (a) and (b)). This confirms our intuition that model symmetry plays an important role in this performance analysis.



(a) (b)

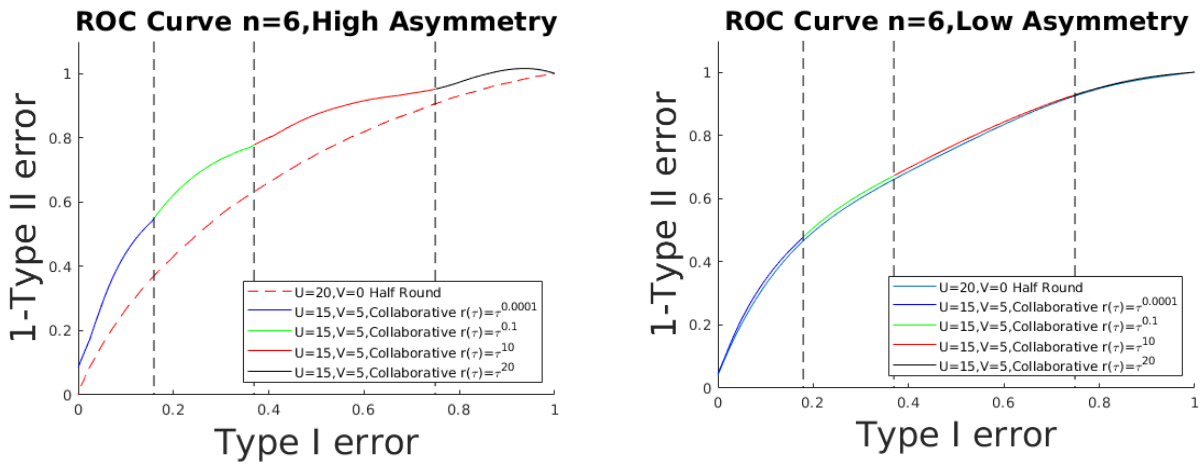
Figure 4.9: Illustrative example of the two channels (conditional probabilities) with $\Lambda = 0.34$ (low asymmetry). Figure 4.9 (a) corresponds to the graphical illustration of the frontward channel ($P_{Y|X}$) for Model 1, and Figure 4.9 (b) corresponds to the graphical illustration of the backward channel ($P_{X|Y}$) for Model 1.



(a) (b)

Figure 4.10: Illustrative example of the two channels (conditional probabilities) with $\Lambda = 0.65$ (high asymmetry). Figure 4.10(a) corresponds to the graphical illustration of the frontward channel ($P_{Y|X}$) for Model 2, and Figure 4.10 (b) corresponds to the graphical illustration of the backward channel ($P_{X|Y}$) for Model 2.

Finally, to complement the results illustrated in Figs 4.11(a) and 4.11(b), Table 4.2 shows the relative reduction of the power of the collaborative test (area under the ROC curve) with respect to the area of its non-collaborative (half-round) for different different models (organized by $\Lambda(P_{X,Y})$). Additionally, we present the relative performance gain of the TYPE II error for the operational points of 3 different TYPE I errors in the ROC curve (0.15, 0.35 and 0.75). For the collaborative scheme, we chose the symmetric assignment $|\mathbb{U}|= 5$ and $|\mathbb{V}|= 4$, and $|\mathbb{U}|= 20$ for the non-collaborative scheme. As the asymmetry of the model



(a) (b)

Figure 4.11: ROC curves for the case of one round collaboration, using $|\mathbb{U}|= 10$, $|\mathbb{V}|= 10$. Different color zones indicate the optimal relationship between $\tilde{t} = f(t)$. All these curves are compared with their corresponding half-round performance using $|\mathbb{U}|= 20$.

increases, the relative gain in performance of our collaborative strategy increases significantly. These results support the idea that there is an important improvement when we introduce collaboration between the nodes, and this improvement is dictated by a specific structural property of $P_{X,Y}$.

$\Lambda(P_{X,Y})$	Accumulative performance	TYPE II Relative Performance		
		0.15	0.35	0.75
0.72	0.1237	42.45%	25.39%	6.71%
0.52	0.0832	28.68 %	18.89%	3.93%
0.41	0.0385	19.21 %	8.74%	1.77%
0.25	0.0087	8.49%	3.71%	0.21%
0	0.0000001	0.00014%	0.00017%	0.00014%

Table 4.2: Accumulative and relative performance gain in terms of the ROC curve with respect to the asymmetry coefficient $\Lambda(P_{X,Y})$. The accumulative gain is calculated using the difference of the area below the ROC curve. For the relative gain, we fixed different TYPE I errors and calculated the relative gain of the power of the test. For both cases, we compare the collaborative case (using $|\mathbb{U}|= 5$ and $|\mathbb{V}|= 4$) with respect to the unidirectional case using $|\mathbb{U}|= 20$.

4.7 Discussion and Concluding Remarks

This work investigated and proposed new data-driven schemes for testing against independence with communication constraints. The main conceptual focus was understanding collaboration's role in this task. We obtained analytical expressions to measure the effect of node cooperation by deriving and exploring asymptotic performance limits. From this theoretical understanding, we proposed an info-max design method to learn some practical strategies (encoders-decoder).

In particular, an algorithm is derived to tackle a multi-letter info-max learning task reminiscent of the type of representation for learning algorithms used in modern ML algorithms. Importantly, our solution does not need a description of the model $P_{X,Y}$ (data-driven) as it builds upon information obtained from i.i.d. samples (empirical observations).[7]

When analyzing our collaborative strategy, we observe that the performance of the test is governed by the assignment between the nodes and affected at the same time by the asymmetry of the data-generating distribution. In all the cases explored, our collaborative scheme shows improvement with respect to its non-collaborative counterpart where this gain was a function of a specific attribute of the model $P_{X,Y}$: its symmetry.

Finally, our results support the adoption of collaboration as a way of effectively using resources (bits) in distributed inference settings. We also show the importance of deriving fundamental information limits for distributed inference problems. These expressions admit analytical interpretations and also have the power to inform the design of practical schemes operating in non-asymptotic conditions.

The promising results presented in this work illuminate many areas of further research. One relevant topic is extending the results presented in this thesis to multiple rounds of node interactions, which is a challenging problem as the derivation of an error exponent for the TYPE II error for multiple rounds is not a simple extension of the argument presented in this work. Another relevant topic is the problem of arbitrary binary hypothesis testing subject to communications constraints. In this area, a single-letter characterization of the TYPE II error exponent remains an open problem where only a lower bound was derived in [33]. Characterizing this fundamental limit would be essential in extending the type of design algorithms proposed in this thesis.

[7] Although the proposed scheme uses partial information of the underlying model $P_{Y|X}$ (the channel), it can be directly extended to a scenario where $P_{Y|X}$ is also estimated from data.

4.8 Appendix

4.8.1 Proof of Theorem 4.1

PROOF. The proof is divided in two distinctive parts associated to a feasibility (or constructive) and a unfeasibility (or converse) argument.

Feasibility

We first introduce a key Lemma that offers a lower bound (a constructive argument) for the error exponent of the Type II error in the regime when $\epsilon > 0$ (the Type I error restriction) is arbitrary small.

Lemma 4.8.1 Given $P_{XY} \in \mathcal{P}(\mathbb{X} \times \mathbb{Y})$ and $R > 0$, let $\mathscr{S}(R) \subset \mathcal{P}(\mathbb{U} \times \mathbb{V} \times \mathbb{X} \times \mathbb{Y})$ and $\mathscr{L}(P_{UVXY}) \subset \mathcal{P}(\mathbb{U} \times \mathbb{V} \times \mathbb{X} \times \mathbb{Y})$ denote the sets of probability defined by

$$\begin{aligned}\mathscr{S}(R) = \{&P_{UVXY} \in \mathcal{P}(\mathbb{U} \times \mathbb{V} \times \mathbb{X} \times \mathbb{Y}) : (X,Y,U) \sim P_{XY} \cdot P_{U|X}, \\ &(X,Y,U,V) \sim P_{X,Y} \cdot P_{U|X} \cdot P_{V|U,Y},\ s.t.\ \ I(U;X) + I(V;Y|U) \leq R\ , \\ &|\mathbb{U}|, |\mathbb{V}| < +\infty\}\ , \end{aligned} \tag{4.37}$$

$$\mathscr{L}(P_{UVXY}) = \{\mu_{UVXY} \in \mathcal{P}(\mathbb{U} \times \mathbb{V} \times \mathbb{X} \times \mathbb{Y})\ :\ \mu_{UVX} = P_{UVX}\ ,\ \mu_{UVY} = P_{UVY}\}\ . \tag{4.38}$$

Then, it follows that

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} -\frac{1}{n} \log \beta_n^c(R,\epsilon) \geq \max_{P_{UVXY} \in \mathscr{S}(R)} \min_{\mu_{UVXY} \in \mathscr{L}(P_{UVXY})} \mathcal{D}(\mu_{UVXY} || Q_{UVXY})\ , \tag{4.39}$$

where $Q_{UVXY} \equiv Q_{XY} \cdot P_{U|X} \cdot P_{V|UY}$ derives from $P_{UVXY} = P_{XY} \cdot P_{U|X} \cdot P_{V|UY}$ being $P_{U|X}$ and $P_{V|UY}$ the two channels used to construct $P_{UVXY} \in \mathscr{S}(R)$ from $P_{X,Y}$ (see the definition in Eq.(4.37)).

The proof of Lemma 4.8.1 is presented in Appendix 4.8.2.

Using Lemma 4.8.1, the lower bound in Eq.(4.39) can be alternatively expressed as[8]:

$$\begin{aligned}\lim_{\epsilon \to 0} \liminf_{n \to \infty} &-\frac{1}{n} \log \beta_n^c(R,\epsilon) \geq \\ &\max_{P_{UVXY} \in \mathscr{S}(R)} \min_{\mu_{UVXY} \in \mathscr{L}(P_{UVXY})} \left[\mathcal{D}(\mu_{UXY} || Q_{UXY}) + \mathcal{D}(\mu_{XV|UY} || \mu_{X|UY} \cdot \mu_{V|UY} | \mu_{UY})\right]\ . \end{aligned} \tag{4.40}$$

For our test against independence problem, we analyze each information term in (4.40) separately. Considering first $\mathcal{D}(\mu_{UXY} \| Q_{UXY})$, we have that

$$\begin{aligned}\mathcal{D}(\mu_{UXY} \| Q_{UXY}) &\overset{(a)}{=} \mathcal{D}(\mu_{UY} || Q_{UY}) + \mathcal{D}(\mu_{X|UY} || Q_{X|UY} | \mu_{UY}) && (4.41)\\ &\overset{(b)}{=} I(U;Y) + \mathcal{D}(\mu_{X|UY} || Q_{X|U} | \mu_{UY}) && (4.42)\\ &= I(U;Y) + \mathcal{D}(\mu_{X|UY} || \mu_{X|U} | \mu_{UY}) + \mathcal{D}(\mu_{X|U} || Q_{X|U} | \mu_U) && (4.43)\\ &\overset{(c)}{\geq} I(U;Y) + \mathcal{D}(\mu_{X|UY} || \mu_{X|U} | \mu_{UY})\ , && (4.44)\end{aligned}$$

[8] The details are presented in (4.108), Appendix 4.8.2

where (a) is due to the chain rule of the divergence [97] and $\mathcal{D}(\mu_{X|UY}||Q_{X|UY}|\mu_{UY})$ denotes the conditional KL-divergence; (b) derives from the independence assumption on $Q_{X,Y}$ (the alternative hypothesis), the Markov chain structure $U \multimap X \multimap Y$ and the fact (from the definition of $\mathscr{L}(P_{UVXY})$) that $\mu_{UY} = P_{UY}$; and (c) follows from the fact that the KL-divergence is non-negative. Continuing the analysis of $\mathcal{D}(\mu_{UXY}\|Q_{UXY})$, we obtain from (4.44) that

$$\mathcal{D}(\mu_{UXY}\|Q_{UXY}) \geq I(U;Y) + \sum_{(u,x,y)\in\mathbb{U}\times\mathbb{X}\times\mathbb{Y}} \mu_{UXY}(u,x,y)\log\left(\frac{\mu_{X|UY}(x|u,y)}{\mu_{X|U}(x|u)}\right) \tag{4.45}$$

$$= I(U;Y) + \sum_{(u,x,y)\in\mathbb{U}\times\mathbb{X}\times\mathbb{Y}} \mu_{UXY}(u,x,y)\log\left(\frac{\mu_{XY|U}(x,y|u)}{\mu_{X|U}(x|u)\mu_{Y|U}(y|u)}\right) \tag{4.46}$$

$$= I(U;Y) + \mathcal{D}(\mu_{XY|U}||\mu_{X|U}\cdot\mu_{Y|U}|\mu_U)\ . \tag{4.47}$$

For the second information term in (4.40), $\mathcal{D}(\mu_{XV|UY}||\mu_{X|UY}\cdot\mu_{V|UY}|\mu_{UY}) =$

$$\begin{aligned}
&\mathcal{D}(\mu_{VXY|U}||\mu_{VY|U}\cdot\mu_{X|U}|\mu_U) - \mathcal{D}(\mu_{XY|U}||\mu_{X|U}\cdot\mu_{Y|U}|\mu_U)\\
&= \mathcal{D}(\mu_{XY|UV}||\mu_{Y|UV}\cdot\mu_{X|UV}|\mu_{UV}) + \mathcal{D}(\mu_{VX|U}||\mu_{V|U}\cdot\mu_{X|U}|\mu_U) - \mathcal{D}(\mu_{XY|U}||\mu_{X|U}\cdot\mu_{Y|U}|\mu_U)\\
&\geq \mathcal{D}(\mu_{VX|U}||\mu_{V|U}\cdot\mu_{X|U}|\mu_U) - \mathcal{D}(\mu_{XY|U}||\mu_{X|U}\cdot\mu_{Y|U}|\mu_U)\ .
\end{aligned} \tag{4.48}$$

Integrating these derivations in (4.40), we obtain that

$$\lim_{\epsilon\to 0}\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n^c(R,\epsilon) \geq \max_{P_{UVXY}\in\mathscr{S}(R)} \min_{\mu_{UVXY}\in\mathscr{L}(P_{UVXY})} \left[I(U;Y) + \mathcal{D}(\mu_{VX|U}||\mu_{V|U}\cdot\mu_{X|U}|\mu_U)\right] \tag{4.49}$$

$$= \max_{P_{UVXY}\in\mathscr{S}(R)} \left[I(U;Y) + I(V;X|U)\right]\ , \tag{4.50}$$

which completes the constructive part of the proof.

Weak unfeasibility

We use the following key Lemma:

Lemma 4.8.2 (Theorem 2, [29]) Let us consider $R > 0$. For any pair of mappings $(\tilde{f}_n(\cdot), \tilde{g}_n(\cdot))$ satisfying the information bound

$$R \geq \frac{1}{n}[I(U_{N_1};X_1^n) + I(V_{N_2};Y_1^n|U_{N_1})]$$

where $U_{N_1} \equiv \tilde{f}_n(X_1^n)$ and $V_{N_2} \equiv \tilde{g}_n(\tilde{f}_n(X_1^n), Y_1^n)$, the Type II error exponent of the testing against independence with one round is upper bounded by:

$$\lim_{\epsilon\to 0}\liminf_{m\to\infty} -\frac{1}{m}\log\beta_m^c(R,\epsilon) \leq \frac{1}{n}[I(U_{N_1};Y_1^n) + I(V_{N_2};X_1^n|U_{N_1})]\ , \tag{4.51}$$

for all $n \geq 1$.

The proof of Lemma 4.8.2 follows from many known results in [28, 29].

Using Lemma 4.8.2, it follows that:

$$\lim_{\epsilon\to 0}\liminf_{m\to\infty} -\frac{1}{m}\log\beta_m^c(R,\epsilon) \leq \limsup_{n\to\infty} \underbrace{\frac{1}{n}[I(U_{N_1};Y_1^n)+I(V_{N_2};X_1^n|U_{N_1})]}_{\Delta_n\equiv} \tag{4.52}$$

where U_{N_1} represents the message sent from Node 1 while V_{N_2} represents the reply from Node 2. To derive a single-letter expression of this upper bound similar to the result in (4.49), we expand Δ_n in (4.52) as follows.

$$\begin{aligned}
\Delta_n &\overset{(d)}{=} \frac{1}{n}\sum_{i=1}^{n}\left[I(U_{N_1};Y_i|Y_{i+1}^n)+I(V_{N_2};X_i|U_{N_1},X_1^{i-1})\right] && (4.53)\\
&\overset{(e)}{=} \frac{1}{n}\sum_{i=1}^{n}\left[I(U_{N_1},Y_{i+1}^n;Y_i)+I(V_{N_2},Y_{i+1}^n;X_i|U_{N_1},X_1^{i-1})-I(Y_{i+1}^n;X_i|U_{N_1},V_{N_2},X_1^{i-1})\right] && (4.54)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[I(U_{N_1},X_1^{i-1},Y_{i+1}^n;Y_i)-I(X_1^{i-1};Y_i|U_{N_1},Y_{i+1}^n)+I(Y_{i+1}^n;X_i|U_{N_1},X_1^{i-1})\right. && (4.55)\\
&+ \left. I(V_{N_2};X_i|U_{N_1},X_1^{i-1},Y_{i+1}^n)-I(Y_{i+1}^n;X_i|U_{N_1},V_{N_2},X_1^{i-1})\right] && (4.56)\\
&\overset{(f)}{=} \frac{1}{n}\sum_{i=1}^{n}\left[I(\hat{U}_i;Y_i)+I(V_i;X_i|\hat{U}_i)-I(Y_{i+1}^n;X_i|U_{N_1},V_{N_2},X_1^{i-1})\right]\ , && (4.57)
\end{aligned}$$

where X_1^i denotes the first i samples and $X_1^n=(X_1,\ldots,X_n)$. (d) derives from the chain rule of the MI [97] and (e) from the assumed i.i.d. nature of the joint sources (X_1^n,Y_1^n). Finally to obtain the equality presented in (f), the identity

$$\sum_{i=1}^{n} I(X_1^{i-1};Y_i|U_{N_1},Y_{i+1}^n)=\sum_{i=1}^{n} I(Y_{i+1}^n;X_i|U_{N_1},X_1^{i-1})\ , \tag{4.58}$$

presented in [3, chapter 15] is used where we also introduce the following auxiliary RVs on the measurable spaces $(\mathbb{U}_i\times\mathbb{V}_i,\mathcal{B}_{\mathbb{U}_i\times\mathbb{V}_i})$:

$$\hat{U}_i\equiv(U_{N_1},X_1^{i-1},Y_{i+1}^n)\quad\text{and}\quad V_i\equiv V_{N_2}\ ,\ \forall\, i=\{1,...,n\}\ . \tag{4.59}$$

It is important to emphasize that our choice in (4.59) satisfies the required Markov chains $X \multimap (\hat{U}_i,Y) \multimap V_i$ for each $i=\{1,...,n\}$ (the argument is presented in Appendix 4.8.3).

If Q denotes a RV uniformly distributed over $\{1,...,n\}$, then Eq. (4.57) can be expressed as:

$$\begin{aligned}
\Delta_n &= I(\hat{U}_Q;Y_Q|Q)+I(V_Q;X_Q|\hat{U}_Q,Q)-\frac{1}{n}\sum_{i=1}^{n} I(Y_{i+1}^n;X_i|U_{N_1}V_{N_2}X_1^{i-1}) && (4.60)\\
&= I(U;Y)+I(V;X|U)-T\ , && (4.61)
\end{aligned}$$

where $U\equiv(\hat{U}_Q,Q)$, $V\equiv V_Q$ and $T\equiv\frac{1}{n}\sum_{i=1}^{n} I(Y_{i+1}^n;X_i|U_{N_1}V_{N_2}X_1^{i-1})$. Since X_1^n and Y_1^n are i.i.d we have that $X=X_Q$ and $Y=Y_Q$, over any value in the support of Q.

Let us return to our operational problem. For that let us consider an arbitrary encoder-decoder $f_n(\cdot), g_n(\cdot)$ meeting the fixed-rate constraint $\log(|\mathbb{U}||\mathbb{V}|) \leq nR$. If we denote by $U_{N_1} \equiv f_n(X_1^n)$ and $V_{N_2} \equiv g_n(f_n(X_1^n), Y_1^n)$ then it is simple to verify that:

$$nR \geq H(U_{N_1}) + H(V_{N_2}) \geq I(U_{N_1}; X_1^n) + I(V_{N_2}; Y_1^n, U_{N_1}) \geq I(U_{N_1}; X_1^n) + I(V_{N_2}; Y_1^n | U_{N_1}) \ . \tag{4.62}$$

We analyze each of the two last information terms in (4.62) separately.

$$\begin{aligned} I(U_{N_1}; X_1^n) &= \sum_{i=1}^{n} I(U_{N_1}, X_1^{i-1}; X_i) && (4.63) \\ &= \sum_{i=1}^{n} \left[I(U_{N_1}, X_1^{i-1}, Y_{i+1}^n; X_i) - I(Y_{i+1}^n; X_i | U_{N_1}, X_1^{i-1}) \right] \ , && (4.64) \end{aligned}$$

where (4.63) is due to the i.i.d nature of joint samples $(X_1, Y_1)....(X_n, Y_n)$.

The second term $I(V_{N_2}; Y_1^n | U_{N_1})$ writes as

$$\begin{aligned} I(V_{N_2}; Y_1^n | U_{N_1}) &= \sum_{i=1}^{n} \left[I(V_{N_2}, X_1^{i-1}; Y_i | U_{N_1}, Y_{i+1}^n) - I(X_1^{i-1}; Y_i | U_{N_1}, V_{N_2}, Y_{i+1}^n) \right] && (4.65) \\ &= \sum_{i=1}^{n} \big[I(X_1^{i-1}; Y_i | U_{N_1}, Y_{i+1}^n) \\ &\qquad + I(V_{N_2}; Y_i | U_{N_1}, X_1^{i-1} Y_{i+1}^n) - I(X_1^{i-1}; Y_i | U_{N_1}, V_{N_2}, Y_{i+1}^n) \big] && (4.66) \\ &= \sum_{i=1}^{n} \big[I(V_{N_2}; Y_i | U_{N_1}, X_1^{i-1}, Y_{i+1}^n) \\ &\qquad + I(X_i; Y_{i+1}^n | U_{N_1}, X_1^{i-1}) - I(X_1^{i-1}; Y_i | U_{N_1}, V_{N_2}, Y_{i+1}^n) \big] \ , && (4.67) \end{aligned}$$

where for the final step we use the identity in Eq. (4.58). Integrating these derivations in (4.62), we obtain that

$$nR \geq \sum_{i=1}^{n} \left[I(U_{N_1}, X_1^{i-1}, Y_{i+1}^n; X_i) + I(V_{N_2}; Y_i | U_{N_1}, X_1^{i-1} Y_{i+1}^n) - I(X_1^{i-1}; Y_i | U_{N_1}, V_{N_2}, Y_{i+1}^n) \right] \ . \tag{4.68}$$

Introducing $\hat{U}_i \equiv (U_{N_1}, X_1^{i-1}, Y_{i+1}^n)$ and $V_i \equiv V_{N_2}$, $\forall i \in \{1, ..., n\}$, we have that

$$\begin{aligned} R &\geq I(\hat{U}_Q; X_Q | Q) + I(V_Q; Y_Q | \hat{U}_Q, Q) - I(X_1^{i-1}; Y_i | U_{N_1}, V_{N_2}, Y_{i+1}^n) && (4.69) \\ &= I(U; X) + I(V; Y | U) - \frac{1}{n} \sum_{i=1}^{n} I(Y_{i+1}^n; X_i | U_{N_1} V_{N_2} X_1^{i-1}) && (4.70) \\ &= I(U; X) + I(V; Y | U) - T, && (4.71) \end{aligned}$$

where $U \equiv (\hat{U}_Q, Q)$, $V \equiv V_Q$, $T \equiv \frac{1}{n} \sum_{i=1}^{n} I(Y_{i+1}^n; X_i | U_{N_1} V_{N_2} X_1^{i-1})$ and $Y = Y_Q$, $X = X_Q$ for a. In (4.71) we use the identity of (4.58) using the pair $(U_{N_1} V_{N_2})$ instead of U_{N_1}.

Finally adopting (4.61) and (4.71), we have the following related bounds

$$\begin{cases} \Delta_n \leq I(U; Y) + I(V; X | U) - T \ , \\ R \geq I(U; X) + I(V; Y | U) - T \ , \end{cases} \tag{4.72}$$

where (U, V) are auxiliary RVs that respect the required Markov chains in (4.15).

Next we show that the region in (4.72) is equivalent to

$$\begin{cases} \Delta_n \leq I(U;Y) + I(V;X|U) \ , \\ R \geq I(U;X) + I(V;Y|U) \ , \end{cases} \tag{4.73}$$

that means that all the pairs (R, Δ_n) that are forbidden in the region in (4.72) are also forbidden in (4.73). First of all, we use the *Fourier-Motzkin* elimination [chapter 12, [103]]. This allow us to remove $T \geq 0$ by using $T = I(U;Y) + I(V;X|U) - \Delta_n$ from (4.61) in (4.72), from this we get:

$$\begin{cases} \Delta_n \leq I(U;Y) + I(V;X|U) \ , \\ R \geq I(U;X) + I(V;Y|U) - I(U;Y) - I(V;X|U) + \Delta_n \ , \end{cases} \tag{4.74}$$

and using the Markovian relations $U \multimap X \multimap Y$ and $X \multimap (U,Y) \multimap V$ we obtain that

$$\begin{cases} \Delta_n \leq I(U;Y) + I(V;X|U) \ , \\ R \geq I(U;X|Y) + I(V;Y|UX) + \Delta_n \ . \end{cases} \tag{4.75}$$

Finally, to obtain that the two regions in (4.73) and (4.75) are equivalent, we only need to check the extremal points of Δ_n, i.e., the scenarios $\Delta_n = 0$ and $\Delta_n = I(U;Y) + I(V;X|U)$. If $\Delta_n = 0$ the result is trivial because $R = 0$ is optimal under both regions. On the other hand, if $\Delta_n = I(U;Y) + I(V;X|U)$, we have that

$$\begin{aligned} R &\geq I(U;X|Y) + I(V;Y|UX) + I(U;Y) + I(V;X|U) && (4.76) \\ &= I(U;X) + I(V;Y|U) \, . && (4.77) \end{aligned}$$

Therefore, the two regions in (4.72) and (4.73) are equivalent.

Since the upper bound for Δ_n in (4.73) is independent of n and for any of this induced pair of variable (U, V) they meet the condition $R \geq I(U;X) + I(V;Y|U)$, we obtain from the definition of $\mathscr{S}(R)$ in (4.37) that:

$$\limsup_{n\to\infty} \Delta_n \leq \max_{P_{UVXY} \in \mathscr{S}(R)} I(U;Y) + I(V;X|U). \tag{4.78}$$

Finally from (4.50), (4.52) and (4.78), we have that

$$\lim_{\epsilon\to 0} \liminf_{m\to\infty} -\frac{1}{m} \log \beta_m^c(R,\epsilon) = \max_{P_{UVXY} \in \mathscr{S}(R)} I(U;Y) + I(V;X|U) \ . \tag{4.79}$$

This concludes the proof of Theorem 4.1. □

4.8.2 Proof of Lemma 4.8.1

For the proof of this result, the notation and well-known information-theoretic results presented in Appendix 4.8.5 will be used.

PROOF. We start by describing the random construction of codebooks, as well as encoding and decision functions. By analyzing the asymptotic properties of such decision systems, we aim at implying a *feasibility (existence) result* of interactive functions and decision regions that satisfy, for any given $\epsilon, \varepsilon > 0$, the following inequalities:

$$\frac{1}{n}\log\left(|f_n||g_n|\right) \leq I(U;X) + I(V;Y|U) + \varepsilon \ , \ \alpha_n^c(R) \leq \epsilon \ , \tag{4.80}$$

$$-\frac{1}{n}\log\beta_n^c(R,\epsilon) \geq \min_{\mu_{UVXY}\in\mathscr{L}(P_{UVXY})} \mathcal{D}(\mu_{UVXY}||P_{\bar{U}\bar{V}\bar{X}\bar{Y}}) - \varepsilon \ , \tag{4.81}$$

provided that n is large enough and for any given distribution $P_{UV} \in \mathscr{S}(R)$, where $|f_n|$ and $|g_n|$ denote the number of codewords in the codebooks used for interaction (note that *feasibility* is defined in the information-theoretic sense which implies the *random existence* of interactive and decision functions with desired properties).

Codebook generation. Without loss of generality, we assume that Node 1 is the first to communicate. Fix a conditional probability $P_{UV|XY}(u,v|x,y) = P_{U|X}(u|x)P_{V|UY}(v|u,y)$ that attains the maximum in (4.39). Let

$$P_U(u) = \sum_{x\in\mathbb{X}} P_{U|X}(u|x)P_X(x) \ , \tag{4.82}$$

$$P_{V|U}(v|u) = \sum_{y\in\mathbb{Y}} P_{V|UY}(v|u,y)P_Y(y). \tag{4.83}$$

For this choice of RVs, set the rates (R_U, R_V) to be

$$I(U;X) + \epsilon(\delta) = R_U \ , \tag{4.84}$$

$$I(V;Y|U) + \epsilon(\delta') = R_V \tag{4.85}$$

with $\epsilon(\delta) \to 0$ as $\delta \to 0$. By the definition of the set $\mathscr{S}(R)$ in (4.37), it is clear that $R_U + R_V \leq R + \epsilon(\delta) + \epsilon(\delta')$. Randomly and independently draw 2^{nR_U} sequences $u_1^n = (u_1, \ldots, u_n)$, each according to $\prod_{i=1}^n P_U(u_i)$. Index these sequences by $m_U \in [1 : M_U \equiv 2^{nR_U}]$ to form the random codebook $\mathcal{C}_{u_1^n} \equiv \{u_1^n(m_U) : m_U \in [1 : M_U]\}$. As a second step, for each word $u_1^n \in \mathcal{C}_{u_1^n}$, build a codebook $\mathcal{C}_{v_1^n}(m_U)$ by randomly and independently drawing 2^{nR_V} sequences v_1^n, each according to $\prod_{i=1}^n P_{V|U}(v_i|u_i(m_U))$. Index these sequences by $m_V \in [1 : M_V \equiv 2^{nR_V}]$ to form the collection of codebooks $\mathcal{C}_{v_1^n}(m_U) \equiv \{v_1^n(m_U, m_V) : m_V \in [1 : M_V]\}$ for $m_U \in [1 : M_U]$.

Encoding and decision mappings. Given a sequence x_1^n, Node 1 searches in the codebook $\mathcal{C}_{u_1^n}$ for an index m_U such that $(u_1^n(m_U), x_1^n) \in \mathcal{T}^n_{[UX]_\delta}$ (note that this notation denotes the δ-typical set with relation to the probability measure implied by H_0). If no such index is found, Node 1 declares H_1. If more than one sequence is found, Node 1 chooses one at random. Node 1 then communicates the chosen index m_U to Node 2, using a portion R_U bits of the available exchange rate. Upon receiving the index m_U, Node 2 checks if $(u_1^n(m_U), y_1^n) \in \mathcal{T}^n_{[UY]_{\delta'}}$. If not, Node 2 declares H_1. If the received sequence u_1^n and y_1^n (the observed sequence at Node 2) are jointly typical, Node 2 looks in the specific codebook $\mathcal{C}_{v_1^n}(m_U)$, for an index m_V such that $(u_1^n(m_U), v_1^n(m_U, m_V), y_1^n) \in \mathcal{T}^n_{[UVY]_{\delta'}}$. If such an index is not found, Node 2 declares H_1. If Node 2 finds more than one such index, it chooses one of them at random. Node 2 then

transmits the chosen index m_V to Node 1. Upon reception of the index m_V, Node 1 checks if $(u_1^n(m_U), v_1^n(m_U, m_V), x_1^n) \in \mathcal{T}^n_{[UVX]_{\delta''}}$. If so, it declares H_0, otherwise it declares H_1. The relation between δ, δ' and δ'' can be deducted from Lemma 4.8.7 in Appendix 4.8.5. It is, however, important to emphasize that $\delta'(\delta) \to 0$ as $\delta \to 0$, and $\delta''(\delta') \to 0$ as $\delta' \to 0$ with $n \to \infty$.

Analysis of α_n (Type I). The analysis of α_n is identical to the one proposed in [29], for the case of testing against independence. We give here a short summary of the analysis available in [29]. Assuming that the measure that controls X and Y is P_{XY}, and denoting the chosen indices at nodes 1 and 2 by m_U and m_V respectively, the error probability of the Type I can be expressed as follows

$$\alpha_n \equiv \mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3) \leq \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_1^c \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_1^c \cap \mathcal{E}_2^c \cap \mathcal{E}_3) \ , \tag{4.86}$$

where $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$ represent the following error events:

$$\mathcal{E}_1 \equiv \{(u_1^n(m_U), x_1^n) \notin \mathcal{T}^n_{[UX]\delta} \, \forall \, m_U \in [1 : M_U]\} \ , \tag{4.87}$$

$$\mathcal{E}_2 \equiv \{(v_1^n(m_U, m_V), u_1^n(m_U), y_1^n) \notin \mathcal{T}^n_{[VUY]_{\delta'}} \, \forall \, m_V \in [1 : M_V] \tag{4.88}$$
$$\text{and the specific } m_U \text{ selected at Node 1}\} \ ,$$

$$\mathcal{E}_3 \equiv \{(v_1^n(m_U, m_V), u_1^n(m_U), x_1^n) \notin \mathcal{T}^n_{[VUX]_{\delta''}}, \tag{4.89}$$
$$\text{for the specific } m_U \text{ and } m_V \text{ previously chosen}\} \ .$$

Analyzing each of the probabilities in (4.86) separately, $\mathbb{P}(\mathcal{E}_1) \to 0$ as $n \to \infty$ by the *covering lemma* [104], provided that $R_U \geq I(U;X) + \epsilon(\delta)$, with $\epsilon(\delta) \to 0$ as $\delta \to 0$. $\mathbb{P}(\mathcal{E}_1^c \cap \mathcal{E}_2) \to 0$ when $n \to \infty$ by the *conditional typicality lemma* [104], in addition to the covering lemma, provided that $R_V \geq I(V;Y|U) + \epsilon(\delta')$. Finally, the third term in (4.86) can be shown to tend to zero through the use of the Markov lemma (see Lemma 4.8.8), as well as Lemma 4.8.6 and Lemma 4.8.7 in Appendix 4.8.5. Thus, as all three components tend to zero with large n, we may conclude that $\alpha_n \leq \epsilon$ for any constraint $0 < \epsilon < 1$ and n large enough.

Analysis of β_n (Type II). The error probability of Type II is defined by

$$\beta_n^c(R, \epsilon) \equiv \mathbb{P}(\text{decide } H_0 | XY \sim Q_{XY}) \ . \tag{4.90}$$

Thus, we assume that $P_{\bar{X}\bar{Y}}$ controls the measure of the observed RVs throughout this analysis. We use similar methods to what was done in [33], although we choose to work with random codebooks. The influence of this choice is on the analysis of α_n only, as seen above, and not on β_n.

For a given pair of sequences (x_1^n, y_1^n) with type variables $\hat{P}_{XY} \in \mathcal{P}_n(\mathbb{X} \times \mathbb{Y})$, we count all possible events that lead to an error. We notice first, that given a pair of vectors $(x_1^n, y_1^n) \in \mathbb{X}^n \times \mathbb{Y}^n$ the probability that these vectors will be the result of n i.i.d. draws, according to the measure implied by H_1, is given by Lemma 4.8.6 in Appendix 4.8.5 to be:

$$Q((X_1^n, Y_1^n) = (x_1^n, y_1^n)) = \exp\left[-n\left(H(\hat{P}_{XY}) + \mathcal{D}(\hat{P}_{XY} || Q_{XY})\right)\right] \ , \tag{4.91}$$

where $\hat{P}_{XY} \in \mathcal{P}_n(\mathbb{X} \times \mathbb{Y})$ are the type variables of (x_1^n, y_1^n) (see Appendix 4.8.5). For each pair of codewords $(u_1^n)_i \in \mathcal{C}_{u_1^n}$ and $(v_1^n)_{ij} \in \mathcal{C}_{v_1^n}(i)$, we define the set:

$$\mathcal{S}_{ij}(x_1^n) \equiv \{(u_1^n)_i\} \times \{(v_1^n)_{ij}\} \times \mathcal{G}_{ij} \times \{x_1^n\} \ , \tag{4.92}$$

where $\mathcal{G}_{ij} \subseteq \mathbb{Y}^n$ is the set of all vectors y_1^n that, given the received message $(u_1^n)_i$, will result in the message $(v_1^n)_{ij}$ being transmitted back to Node 1. Denoting by $K_{ij}(x_1^n)$ the number of elements $((u_1^n)_i, (v_1^n)_{ij}, y_1^n, x_1^n) \in \mathcal{S}_{ij}(x_1^n)$ whose type variables coincide with $U^{(n)}V^{(n)}X^{(n)}Y^{(n)}$, we have by Lemma 4.8.5 in Appendix 4.8.5 that:

$$K_{ij}(x_1^n) \leq \exp\left[nH(Y^{(n)}|U^{(n)}V^{(n)}X^{(n)})\right] \ . \tag{4.93}$$

Let $K(U^{(n)}V^{(n)}X^{(n)}Y^{(n)})$ denote the number of all elements:

$$\mathscr{S}_n \equiv \bigcup_{i=1}^{M_U}\bigcup_{j=1}^{M_V} \bigcup_{x_1^n \in \mathcal{T}^n_{[X|(u_1^n)_i(v_1^n)_{ij}]_{\delta''}}} \mathcal{S}_{ij}(x_1^n) \tag{4.94}$$

that have type variable $U^{(n)}V^{(n)}X^{(n)}Y^{(n)} \in \mathcal{P}_n(\mathbb{U} \times \mathbb{V} \times \mathbb{X} \times \mathbb{Y})$, then

$$\begin{aligned} K(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) &\leq \sum_{i=1}^{M_U}\sum_{j=1}^{M_V} \exp\left[nH(Y^{(n)}|U^{(n)}V^{(n)}X^{(n)})\right] |\mathcal{T}^n_{[X|(u_1^n)_i(v_1^n)_{ij}]_{\delta''}}| & (4.95)\\ &\leq \exp\Big[n\Big(H(Y^{(n)}|U^{(n)}V^{(n)}X^{(n)}) & (4.96)\\ &+ I(U;X) + I(V;Y|U) + H(X|UV) + \mu_n\Big)\Big] \ , & (4.97) \end{aligned}$$

where M_U and M_V are the sizes of the codebooks $\mathcal{C}_{u_1^n}$ and $\mathcal{C}_{v_1^n}(\cdot)$. The first and second additional terms in the final expression come from the size of the codebooks and the third is a bound over the size of the delta-typical set (see Lemma 4.8.9 in Appendix 4.8.5). The resulting sequence μ_n is a function of $\delta, \delta', \delta''$ that complies with $\mu_n \to 0$ as $n \to \infty$. The error probability of Type II satisfies:

$$\beta_n^c(R,\epsilon) \leq \sum_{U^{(n)}V^{(n)}X^{(n)}Y^{(n)} \in \mathscr{S}_n} \exp\left[-n\left(k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) - \mu_n\right)\right] \ , \tag{4.98}$$

where the function $k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)})$ is defined by

$$\begin{aligned} k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) &\equiv H(X^{(n)}Y^{(n)}) + \mathcal{D}(\hat{P}_{XY}||Q_{XY}) & (4.99)\\ &- H(Y^{(n)}|U^{(n)}V^{(n)}X^{(n)}) - H(X|UV) & (4.100)\\ &- I(U;X) - I(V;Y|U) \ . & (4.101) \end{aligned}$$

We deliberately made an abuse of notation in (4.98) to indicate that the sum is taken over all possible type-variables $U^{(n)}V^{(n)}X^{(n)}Y^{(n)} \in \mathcal{P}_n(\mathbb{U} \times \mathbb{V} \times \mathbb{X} \times \mathbb{Y})$ formed by empirical probability measures from elements $(u_1^n, v_1^n, x_1^n, y_1^n) \in \mathscr{S}_n$.

From the construction of $\mathscr{S}_n$, it is clear that if $(u_1^n, v_1^n, x_1^n, y_1^n) \in \mathscr{S}_n$, then at least $(u_1^n, v_1^n, x_1^n) \in \mathcal{T}^n_{[UVX]_{\delta''}}$ and $(u_1^n, v_1^n, y_1^n) \in \mathcal{T}^n_{[UVY]_{\delta'}}$. Thus, the summation in (4.98) is only over all types satisfying:

$$\begin{aligned} |P_{U^{(n)}V^{(n)}X^{(n)}}(u,v,x) - P_{UVX}(u,v,x)| &\leq \delta'' \ ,\\ |P_{U^{(n)}V^{(n)}Y^{(n)}}(u,v,y) - P_{UVY}(u,v,y)| &\leq \delta' \ , \end{aligned} \tag{4.102}$$

for all $(u,v,x) \in \text{supp}(P_{UVX})$ and $(u,v,y) \in \text{supp}(P_{UVY})$. In addition, it follows by Lemma 4.8.4 (from the total number of types of length n) that:

$$\begin{aligned}\beta_n^c(R,\epsilon) &\leq (n+1)^{|\mathbb{U}||\mathbb{V}||\mathbb{X}||\mathbb{Y}|} \\ &\times \max_{U^{(n)}V^{(n)}X^{(n)}Y^{(n)} \in \mathscr{S}_n} \exp\left[-n\left(k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) - \mu_n\right)\right] \ .\end{aligned} \tag{4.103}$$

By (4.102) and the continuity of the entropy as well as the KL divergence [2], we can conclude that

$$\begin{aligned}k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) &= H(\mu_{XY}) + \mathcal{D}(\mu_{XY}||Q_{XY}) - H(\mu_{UVXY}) \\ &+ H(\mu_{UV}) - \mathcal{D}(\mu_{UX}||\mu_U \cdot \mu_X) - \mathcal{D}(\mu_{VY|U}||\mu_{V|U} \cdot \mu_{Y|U}|\mu_U) + \mu_n' \qquad (4.104) \\ &= \mathcal{D}(\mu_{UVXY}||\mu_{UV} \cdot \mu_{XY}) + \mathcal{D}(\mu_{XY}||Q_{XY}) \\ &- \mathcal{D}(\mu_{UX}||\mu_U \cdot \mu_X) - \mathcal{D}(\mu_{VY|U}||\mu_{V|U} \cdot \mu_{Y|U}|\mu_U) + \mu_n' \qquad (4.105) \\ &= \mathcal{D}(\mu_{UXY}||\mu_{XY} \cdot \mu_U) + \mathcal{D}(\mu_{VXY|U}||\mu_{XY|U} \cdot \mu_{V|U}|\mu_U) + \mathcal{D}(\mu_{XY}||Q_{XY}) \\ &- \mathcal{D}(\mu_{UX}||\mu_U \cdot \mu_X) - \mathcal{D}(\mu_{VY|U}||\mu_{V|U} \cdot \mu_{Y|U}|\mu_U) + \mu_n' \qquad (4.106) \\ &= \mathcal{D}(\mu_{UXY}||\mu_{XY} \cdot \mu_U) + \mathcal{D}(\mu_{VX|UY}||\mu_{V|UY} \cdot \mu_{X|UY}|\mu_{UY}) + \mathcal{D}(\mu_{XY}||Q_{XY}) \\ &- \mathcal{D}(\mu_{UX}||\mu_U \cdot \mu_X) + \mu_n' \qquad (4.107) \\ &= \mathcal{D}(\mu_{U|X}||\mu_{Y|X}|\mu_X) + \mathcal{D}(\mu_{VX|UY}||\mu_{V|UY} \cdot \mu_{X|UY}|\mu_{UY}) + \mathcal{D}(\mu_{XY}||Q_{XY}) + \mu_n' \\ &= \mathcal{D}(\mu_{UXY}||Q_{UXY}) + \mathcal{D}(\mu_{VX|UY}||\mu_{V|UY} \cdot \mu_{X|UY}|\mu_{UY}) + \mu_n' \qquad (4.108)\end{aligned}$$

with $\mu_{UVXY} \in \mathscr{L}(P_{UVXY})$ and $\mu_n' \to 0$ when $n \to \infty$.

Finally, the following Markov chain: $X \multimap (U,Y) \multimap V$ holds under both hypotheses.

Then, from (4.108) the development of $k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)})$ goes as follows:

$$\begin{aligned}k(U^{(n)}V^{(n)}X^{(n)}Y^{(n)}) &= \mathcal{D}(\mu_{UXY}||Q_{UXY}) + \mathcal{D}(\mu_{X|UY}||\mu_{V|UY}|\mu_{UY}) + \mu_n' \qquad (4.109) \\ &= \sum_{\forall(u,v,x,y)} \mu_{UVXY}(u,v,x,y) \times \\ &\quad \times \log\left(\frac{\mu_{UXY}(u,x,y)}{Q_{UXY}(u,x,y)} \frac{\mu_{XV|UY}(x,v|u,y)}{\mu_{X|UY}(x|u,y)\mu_{V|UY}(v|u,y)}\right) + \mu_n' \qquad (4.110) \\ &\overset{(g)}{=} \sum_{\forall(u,v,x,y)} \mu_{UVXY}(u,v,x,y) \log\left(\frac{\mu_{UVXY}(u,v,x,y)}{Q_{UXY}(u,x,y)Q_{V|UY}(v|u,y)}\right) + \mu_n' \qquad (4.111) \\ &= \sum_{\forall(u,v,x,y)} \mu_{UVXY}(u,v,x,y) \log\left(\frac{\mu_{UVXY}(u,v,x,y)}{Q_{UVXY}(u,v,x,y)}\right) + \mu_n' \qquad (4.112) \\ &= \mathcal{D}(\mu_{UVXY}||Q_{UVXY}) + \mu_n' \ , \qquad (4.113)\end{aligned}$$

where the sums are over the $\text{supp}(\mu_{UVXY})$; and (g) is due to the definition of the set $\mathscr{L}(P_{UVXY})$ that implies that $\mu_{V|UY}(v|u,y) = P_{V|UY}(v|u,y)$. In addition, as coding (at each side) is performed before a decision is made, it is clear that it must be done in the same way under both hypotheses. Thus, while $P_{UVY}(u,v,y) \neq Q_{UVY}(u,v,y)$, it is true that $Q_{V|UY}(v|u,y) = P_{V|UY}(v|u,y) = \mu_{V|UY}(v|u,y)$. As μ_n, μ_n' are arbitrarily small (as a function of the choices of δ and δ' provided that n is large enough) this concludes the proof of Lemma 4.8.1. □

4.8.3 Proof of Markov Chain Structure

As a part of the weak unfeasibility part of the proof of Theorem 4.1, two Markov chains are necessary:

$$\begin{cases} \hat{U}_i \multimap X_i \multimap Y_i \ , \ \forall\, i = \{1, ..., n\} \\ V_i \multimap (\hat{U}_i, Y_i) \multimap X_i \ , \ \forall\, i = \{1, ..., n\}. \end{cases} \tag{4.114}$$

Using the chosen RVs from (4.59), these Markov chains are represented by

$$\begin{cases} (U_{N_1}, X_1^{i-1}, Y_{i+1}^n) \multimap X_i \multimap Y_i \ , \ \forall\, i = \{1, ..., n\} \\ V_{N_2} \multimap (U_{N_1}, X_1^{i-1}, Y_i^n) \multimap X_i \ , \ \forall\, i = \{1, ..., n\}. \end{cases} \tag{4.115}$$

In order to check this, we use the following result.

Lemma 4.8.3 Let A_1, A_2, B_1, B_2 be RVs with joint probability measure $P_{A_1A_2B_1B_2} = P_{A_1B_1}P_{A_2B_2}$ and assume that $\{f^i\}_{i=1}^k, \{g^i\}_{i=1}^k$ are any collection of P-measurable mappings with domain structure given by:

$$f^1(A_1, A_2); f^2(A_1, A_2, g^1); \ldots; f^k(A_1, A_2, g^1, \ldots, g^{k-1}) \ , \tag{4.116}$$

$$g^1(B_1, B_2, f^1); g^2(B_1, B_2, f^1, f^2); \ldots; g^k(B_1, B_2, f^1, \ldots, f^k) \ . \tag{4.117}$$

Then,

$$I(A_2; B_1 | f^1, f^2, \ldots, f^k, g^1, g^2, \ldots, g^k, A_1, B_2) = 0 \ . \tag{4.118}$$

Proof. Refer to reference [105, Lemma 1]. □

In order to prove the first Markov chain, we simply let:

$$\begin{cases} A_1 = X_i, \quad B_1 = Y_i \ , \\ A_2 = (X_1^{i-1}, X_{i+1}^n, Y_{i+1}^n) \ , \quad B_2 = Y_1^{i-1} \ . \end{cases} \tag{4.119}$$

It can be easily verified that $P_{A_1A_2B_1B_2} = P_{A_1B_1}P_{A_2B_2}$, which stems directly from the i.i.d. nature of the samples. Thus, according to Lemma 4.8.3:

$$\begin{aligned} 0 &= I(X_1^{i-1}, X_{i+1}^n, Y_{i+1}^n; Y_i | X_i, Y_1^{i-1}) \\ &= I(X_1^{i-1}, X_{i+1}^n, Y_1^{i-1}, Y_{i+1}^n; Y_i | X_i) - I(Y_1^{i-1}; Y_i | X_i) \ , \end{aligned} \tag{4.120}$$

which shows the Markov chain:

$$(X_1^{i-1}, X_{i+1}^n, Y_1^{i-1}, Y_{i+1}^n) \multimap X_i \multimap Y_i \ , \ \forall\, i = \{1, ..., n\}. \tag{4.121}$$

As $U_{N_1} = f_n(X_1^n)$, the following Markov chain is also true:

$$(U_{N_1}, X_1^{i-1}, Y_{i+1}^n) \multimap X_i \multimap Y_i \ , \ \forall\, i = \{1, ..., n\}, \tag{4.122}$$

which proves the first Markov chain in (4.115). As for the second one, we let:

$$\begin{cases} A_1 = X_1^{i-1} \ , \quad B_1 = Y_1^{i-1} \ , \\ A_2 = (X_i, X_{i+1}^n) \ , \quad B_2 = (Y_i, Y_{i+1}^n) \ . \end{cases} \tag{4.123}$$

Under this choice, $U_{N_1} = f_n(A_1, A_2)$ and thus,

$$I(X_i, X_{i+1}^n; Y_1^{i-1} | U_{N_1}, X_1^{i-1}, Y_i, Y_{i+1}^n) = 0 \ , \ \forall\, i = \{1, ..., n\}. \tag{4.124}$$

The later identity proves the following Markov chain:

$$(X_i, X_{i+1}^n) \multimap (U_{N_1}, X_1^{i-1}, Y_i, Y_{i+1}^n) \multimap Y_1^{i-1} \ , \ \forall\, i = \{1, ..., n\}. \tag{4.125}$$

As $V_{N_2} = g_n(U_{N_1}, Y_1^n)$, it also holds that:

$$X_i \multimap (U_{N_1}, X_1^{i-1}, Y_i^n) \multimap V_{N_2} \ , \ \forall\, i = \{1, ..., n\}, \tag{4.126}$$

which yields the desired Markov chain.

4.8.4 Proof that $E(R) \geq \xi(R)$

Suppose that the solution of $\xi(R)$ is $P_{U|X}^*$ such that $I(U;X) = R$. This implies that $|\mathbb{V}| = 1$ and then $I(V;Y|U) = 0$. This implies that $E(R) = \xi(R)$. Now assume that there exists $P_{U|X}^*$ such that $I(U;X) < R$. Then, we can always find $P_{V|UY}^*$ such that $I(U;X) + I(V;Y|U) \leq R$ and plugin $P_{U|X}^*$ and $P_{V|UY}^*$ into $I(U;Y) + I(V;X|U)$ we have by definition that $I(U;Y) + I(V;X|U) = \xi(R) + I(V;X|U)$ and, therefore, $E(R) \geq \xi(R)$.

4.8.5 Technical Definitions and Lemmas

In this appendix, we revise fundamental notions and properties of *method of types* [81], which are extensively used through this chapter. Given a vector $x_1^n = (x_1, \ldots, x_n) \in \mathbb{X}^n$, let $N(a|x_1^n)$ be the *counting measure*, i.e., the number of times the letter $a \in \mathbb{X}$ appears in the vector x_1^n. The *type* of the vector x_1^n, denoted by $Q_{x_1^n}$, is defined through its *empirical measure*: $Q_{x_1^n}(a) = n^{-1} N(a|x_1^n)$ with $a \in \mathbb{X}$. $\mathcal{P}_n(\mathbb{X})$ denotes the set of all possible types (or empirical measures) of length n over $\mathbb{X}$. We use type variables of the form $X^{(n)} \in \mathcal{P}_n(\mathbb{X})$ to denote a RV with a probability measure identical to the empirical measure induced by x_1^n. The set of all vectors x_1^n that share this type is denoted by $\mathcal{T}(Q_{x_1^n}) = \mathcal{T}_{[Q_{x_1^n}]}$.

Definition 4.3 (Types [3]) *The type of a sequence $x_1^n \in \mathbb{X}^n$ is the measure $\hat{P}_X$ on $\mathbb{X}$ defined by $\hat{P}_X(a) = \frac{1}{n} N(a|x_1^n)$, $\forall a \in \mathbb{X}$, where $N(a|x_1^n)$ is the counting measure of the letter a in x_1^n. The* joint type *of a pair $(x_1^n, y_1^n) \in \mathbb{X}^n \times \mathbb{Y}^n$ is the empirical measure $\hat{P}_{XY}$ on $\mathbb{X} \times \mathbb{Y}$ such that*

$$\hat{P}_{XY}(a,b) = \frac{1}{n} N(a,b|x_1^n, y_1^n) \ , \quad \forall (a,b) \in \mathbb{X} \times \mathbb{Y} \ , \tag{4.127}$$

where $N(a,b|x_1^n, y_1^n)$ is the joint counting measure of the pair (a,b) in (x_1^n, y_1^n).

Definition 4.4 (Conditional Types [3]) *The vector $y_1^n \in \mathbb{Y}_1^n$ is said to have* conditional type *$V : \mathbb{X} \mapsto \mathcal{P}_n(\mathbb{Y})$ given $x_1^n \in \mathbb{X}^n$ if*

$$N(a,b|x_1^n, y_1^n) = N(a|x_1^n) V(b|a) \ , \quad \forall (a,b) \in \mathbb{X} \times \mathbb{Y} \ , \tag{4.128}$$

where V is a stochastic mapping.

Lemma 4.8.4 (Type Counting) Let $\mathcal{P}_n(\mathbb{X})$ be the set of all possible types of sequences in $\mathbb{X}^n$. Then, $|\mathcal{P}_n(\mathbb{X})| \leq (n+1)^{|\mathbb{X}|}$.

PROOF. Refer to reference [3, Lemma 2.2]. □

Lemma 4.8.5 For any type $\hat{P} \in \mathcal{P}_n(\mathbb{X})$ of sequences in $\mathbb{X}^n$, denote by $\mathcal{T}_{[\hat{P}]}$ the set of all sequences with this type. Then,

$$(n+1)^{-|\mathbb{X}|} \exp\left[nH(\hat{P})\right] \leq |\mathcal{T}_{[\hat{P}]}| \leq \exp\left[nH(\hat{P})\right] . \tag{4.129}$$

In a similar fashion, for every $x_1^n \in \mathbb{X}^n$ and stochastic mapping $V : \mathbb{X} \mapsto \mathcal{P}_n(\mathbb{Y})$, let $\mathcal{T}_{[V]}(x_1^n)$ be the set of all sequences $y_1^n \in \mathbb{Y}_1^n$ with the conditional type V given x_1^n. Then,

$$(n+1)^{-|\mathbb{X}||\mathbb{Y}|} \exp\left[nH(V|\hat{P})\right] \leq |\mathcal{T}_{[V]}(x_1^n)| \leq \exp\left[nH(V|\hat{P})\right] , \tag{4.130}$$

where $H(V|\hat{P})$ is the conditional entropy function,

$$H(V|\hat{P}) = \sum_{x \in \mathbb{X}} \hat{P}(x) H(V(\cdot|x)) . \tag{4.131}$$

PROOF. Refer to reference [3, Lemma 2.3, Lemma 2.5]. □

Lemma 4.8.6 (Inaccuracy) Let $\hat{P} \in \mathcal{P}_n(\mathbb{X})$ be the type of $x_1^n \in \mathbb{X}^n$ ($X^{(n)} \sim \hat{P}$ is referred to as the *type variable*). Then, for any RV X on $(\mathbb{X}, \mathcal{B}_{\mathbb{X}}, P_X)$,

$$P_X^n(X_1^n = x_1^n) = \exp\Big\{-n\left[H(\hat{P}) + \mathcal{D}(\hat{P}\|P_X)\right]\Big\} . \tag{4.132}$$

PROOF. Refer to reference [33, Lemma 3], [3, Lemma 2.6]. □

Definition 4.5 (δ-Typicality [33]) *Let $\delta > 0$, an n-sequence x_1^n is called δ-typical, denoted by $\mathcal{T}_{[X]_\delta}$, if $|N(a|x_1^n) - nP_X(a)| \leq \mathcal{O}(\delta), \quad \forall a \in \mathbb{X}$, and $\hat{P}_X \ll P_X$. Jointly δ-typical $\mathcal{T}_{[XY]_\delta}$ and conditionally δ-typical sequences $\mathcal{T}_{[Y|X]_\delta}(x_1^n)$ are defined in a similar manner.*

Lemma 4.8.7 Let $\mathcal{T}_{[X]_\delta}$, $\mathcal{T}_{[XY]_\delta}$ and $\mathcal{T}_{[Y|X]_\delta}$ denote the sets of typical, jointly typical and conditionally typical sequences, respectively. For any $x_1^n \in \mathcal{T}_{[X]_\delta}$ and $y_1^n \in \mathcal{T}_{[Y|X]_{\delta'}}$, then $(x_1^n, y_1^n) \in \mathcal{T}_{[XY]_{\delta+\delta'}}$. Moreover, $y_1^n \in \mathcal{T}_{[Y]_{\delta''}}$, with $\delta'' = (\delta + \delta')|\mathbb{X}|$.

PROOF. Refer to reference [3]. □

Lemma 4.8.8 (Generalized Markov Lemma) Let $P_{UXY} \in \mathcal{P}(\mathbb{U} \times \mathbb{X} \times \mathbb{Y})$ be a probability measure that satisfies: $U \multimap X \multimap Y$. Consider $(x_1^n, y_1^n) \in \mathcal{T}^n_{[XY]_{\epsilon'}}$ and random vectors U^n generated according to:

$$\mathbb{P}\left(U_1^n = u_1^n | U_1^n \in \mathcal{T}^n_{[U|X]_{\epsilon''}}(x_1^n), x_1^n, y_1^n\right) = \frac{\mathbb{1}\left\{u_1^n \in \mathcal{T}^n_{[U|X]_{\epsilon''}}(x_1^n)\right\}}{|\mathcal{T}^n_{[U|X]_{\epsilon''}}(x_1^n)|} . \tag{4.133}$$

For sufficiently small $\epsilon, \epsilon', \epsilon'' > 0$,

$$\mathbb{P}\left(U^n \notin \mathcal{T}^n_{[U|XY]_\epsilon}(x_1^n, y_1^n) \middle| U^n \in \mathcal{T}^n_{[U|X]_{\epsilon''}}(x_1^n), x_1^n, y_1^n\right) \equiv \mathcal{O}\left(c^{-n}\right) \tag{4.134}$$

holds uniformly on $(x_1^n, y_1^n) \in \mathcal{T}^n_{[XY]_{\epsilon'}}$ where $c > 1$.

Proof. Refer to reference [106]. □

Lemma 4.8.9 For every probability measure $P_X \in \mathcal{P}(\mathbb{X})$ and stochastic mapping $W : \mathbb{X} \mapsto \mathcal{P}(\mathbb{Y})$, there exist sequences $(\varepsilon_n)_{n\in\mathbb{N}_+}, (\varepsilon'_n)_{n\in\mathbb{N}_+} \to 0$ as $n \to \infty$ satisfying:

$$\left|\frac{1}{n}\log|\mathcal{T}_{[X]_\delta}| - H(X)\right| \leq \varepsilon_n \ , \tag{4.135}$$

$$\left|\frac{1}{n}\log|\mathcal{T}_{[Y|X]_\delta}(x_1^n)| - H(Y|X)\right| \leq \varepsilon_n \ , \tag{4.136}$$

for each $x_1^n \in \mathcal{T}_{[X]_\delta}$ where $\varepsilon_n \equiv \mathcal{O}(n^{-1}\log n)$, and

$$P_{X_1^n}(\mathcal{T}_{[X]_\delta}) \geq 1 - \varepsilon'_n \ , \tag{4.137}$$

$$W^n(\mathcal{T}_{[Y|X]_\delta}(x_1^n)|X_1^n = x_1^n) \geq 1 - \varepsilon'_n \ , \tag{4.138}$$

for all $x_1^n \in \mathbb{X}^n$ where $\varepsilon'_n \equiv \mathcal{O}\left(\frac{1}{n\delta^2}\right)$, provided that n is sufficiently large.

Proof. Refer to reference [3, Lemma 2.13]. □

4.8.6 Derivation of Half-Round Algorithm

Using the fact that $f(X_1^n) \multimap X_1^n \multimap Y_1^n$ (Markov chain) we can rewrite (4.17) as

$$\begin{aligned} I(U;Y_1^n) &= H(U) - H(U|Y_1^n) \\ &= \sum_{u\in\mathbb{U}} P_U(u)\log\left(\frac{1}{P(u)}\right) - \sum_{y_1^n\in\mathbb{Y}^n}\sum_{u\in\mathbb{U}} P_{U,Y_1^n}(u, y_1^n)\log\left(\frac{P_{Y_1^n}(y_1^n)}{P_{U,Y_1^n}(u, y_1^n)}\right) \end{aligned} \tag{4.139}$$

The first term $H(U)$ can be computed using the following identity

$$\begin{aligned} H(U) &= \sum_{u\in\mathbb{U}} P_U(u)\log\left(\frac{1}{P_U(u)}\right) \\ &= \sum_{x_1^n\in\mathbb{X}^n}\sum_{u\in\mathbb{U}} P_{U|X_1^n}(u|x_1^n)P_{X_1^n}(x_1^n)\log\left(\frac{1}{\sum_{x_1^n\in\mathbb{X}^n} P_{U|X_1^n}(u|x_1^n)P_{X_1^n}(x_1^n)}\right) \end{aligned} \tag{4.140}$$

The second term $H(U|Y_1^n)$ can be computed using the folowing Markov chain property:

$$
\begin{aligned}
&H(U|Y_1^n) \\
&= \sum_{y_1^n \in \mathbb{Y}^n} \sum_{u \in \mathbb{U}} P_{U,Y_1^n}(u, y_1^n) \log\left(\frac{P_{Y_1^n}(y_1^n)}{P_{U,Y_1^n}(u, y_1^n)}\right) \\
&= \sum_{y_1^n \in \mathbb{Y}^n} \sum_{x_1^n \in \mathbb{X}^n} \sum_{u \in \mathbb{U}} P_{Y_1^n,U,X_1^n}(y_1^n, u, x_1^n) \log\left(\frac{\sum_{x_1^n \in \mathbb{X}^n} P_{Y_1^n,X_1^n}(y_1^n, x_1^n)}{\sum_{x_1^n \in \mathbb{X}^n} P_{Y_1^n,U,X_1^n}(y_1^n, u, x_1^n)}\right) \\
&= \sum_{\substack{x_1^n \in \mathbb{X}^n \\ u \in \mathbb{U} \\ y_1^n \in \mathbb{Y}^n}} P_{Y_1^n|X_1^n}(y_1^n|x_1^n) P_{U|X_1^n}(u|x_1^n) P_{X_1^n}(x_1^n) \log\left(\frac{\sum\limits_{x_1^n \in \mathbb{X}^n} P_{Y_1^n|X_1^n}(y_1^n|x_1^n) P_{X_1^n}(x_1^n)}{\sum\limits_{x_1^n \in \mathbb{X}^n} P_{Y_1^n|X_1^n}(y_1^n|x_1^n) P_{U|X_1^n}(u|x_1^n) P_{X_1^n}(x_1^n)}\right)
\end{aligned}
\tag{4.141}
$$

In practice, we do not have access to the marginal distribution of X but we have a training set $\{\overline{x}_1, ..., \overline{x}_m\}$ i.i.d, with $\overline{x}_i \in \mathbb{X}^n$. Then, we can introduce the empirical version of the marginal of $P_{X_1^n}(x_1^n)$, by:

$$
P_{X_1^n}(x_1^n) \approx \hat{P}_{X_1^n}(x_1^n) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\overline{x}_i = x_1^n}. \tag{4.142}
$$

We also observe that (4.141) is computationally costly because it depends on the alphabet of $\mathbb{Y}^n$. We can approximate this expression by substituting the conditional distribution by its empirical mean. More precisely, given $X_1^n = x_1^n$ we have a set $S_{x_1^n} = \{\overline{y}_1, ..., \overline{y}_{m'}\}$ i.i.d, with $\overline{y}_i \in \mathbb{Y}^n$ and

$$
P_{Y_1^n|X_1^n}(y_1^n|x_1^n) \approx \hat{P}(Y_1^n = y_1^n | X_1^n = x_1^n) = \frac{1}{m'} \sum_{\overline{y}_j \in S_{x_1^n}} \mathbb{1}_{\overline{y}_j = y_1^n}. \tag{4.143}
$$

Plugin these two expressions in (4.139) we have the empirical version of the mutual information $I(U; Y_1^n)$.

$$
\hat{I}_\alpha(Y_1^n; U) \equiv \hat{H}(U) - \alpha \hat{H}(U|Y_1^n) \tag{4.144}
$$

with

$$
\begin{aligned}
\hat{H}(U) &= \sum_{x_1^n \in \mathbb{X}^n} \sum_{u \in \mathbb{U}} P_{U|X_1^n}(u|x_1^n) \hat{P}_{X_1^n}(x_1^n) \log\left(\frac{1}{\sum_{x_1^n \in \mathbb{X}^n} P_{U|X_1^n}(u|x_1^n) \hat{P}_{X_1^n}(x_1^n)}\right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{u \in \mathbb{U}} P_{U|X_1^n}(u|\overline{x}_i) \log\left(\frac{1}{\frac{1}{m}\sum_{i=1}^{m} P_{U|X_1^n}(u|\overline{x}_i)}\right)
\end{aligned}
\tag{4.145}
$$

and

$$
\begin{aligned}
&\hat{H}(U|Y_1^n) \\
&= \sum_{\substack{x_1^n \in \mathbb{X}^n \\ u \in \mathbb{U} \\ y_1^n \in \mathbb{Y}^n}} \hat{P}_{Y_1^n|X_1^n}(y_1^n|x_1^n) P_{U|X_1^n}(u|x_1^n) \hat{P}_{X_1^n}(x_1^n) \log \left(\frac{\sum\limits_{x_1^n \in \mathbb{X}^n} P_{Y_1^n|X_1^n}(y_1^n|x_1^n) \hat{P}_{X_1^n}(x_1^n)}{\sum\limits_{x_1^n \in \mathbb{X}^n} P_{Y_1^n|X_1^n}(y_1^n|x_1^n) P_{U|X_1^n}(u|x_1^n) \hat{P}_{X_1^n}(x_1^n)} \right) \\
&= \frac{1}{mm'} \sum_{i=1}^{m} \sum_{\overline{y}_j \in S_{\overline{x}_i}} \sum_{u \in \mathbb{U}} P_{U|X_1^n}(u|\overline{x}_i) \log \left(\frac{\sum\limits_{i=1}^{m} P_{Y_1^n|X_1^n}(\overline{y}_j|\overline{x}_i)}{\sum\limits_{i=1}^{m} P_{Y_1^n|X_1^n}(\overline{y}_j|\overline{x}_i) P_{U|X_1^n}(u|\overline{x}_i)} \right).
\end{aligned} \tag{4.146}
$$

4.8.7 Derivation of One-Round Algorithm

Using the fact that $V \multimap (U, Y_1^n) \multimap X_1^n$ and the definition of the conditional MI, we have that $I(V; X_1^n|U)$ can be expressed as $H(V|U) - H(V|X_1^n, U)$. The first term $H(V|U)$ can be computed using the following Markov chain property

$$
\begin{aligned}
&= H(V|U) \\
&= \sum_{u \in \mathbb{U}} \sum_{v \in \mathbb{V}} P_{U,V}(u,v) \log \left(\frac{P_U(u)}{P_{U,V}(u,v)} \right) \\
&= \sum_{u \in \mathbb{U}} \sum_{v \in \mathbb{V}} \sum_{y_1^n \in \mathbb{Y}^n} P_{U,V,Y_1^n}(u,v,y_1^n) \log \left(\frac{P_U(u)}{\sum_{y_1^n \in \mathbb{Y}^n} P_{U,V,Y_1^n}(u,v,y_1^n)} \right) \\
&= \sum_{u \in \mathbb{U}} \sum_{v \in \mathbb{V}} \sum_{y_1^n \in \mathbb{Y}^n} P_{V|Y_1^n,U}(v|y_1^n,u) P_{U,Y_1^n}(u,y_1^n) \log \left(\frac{P_U(u)}{\sum_{y_1^n \in \mathbb{Y}^n} P_{V|Y_1^n,U}(v|y_1^n,u) P_{U,Y_1^n}(u,y_1^n)} \right) \\
&= \sum_{\substack{u \in \mathbb{U} \\ v \in \mathbb{V} \\ x_1^n \in \mathbb{X}^n \\ y_1^n \in \mathbb{Y}^n}} P_{V|Y_1^n,U} P_{Y_1^n|X_1^n} P_{U|X_1^n} P_{X_1^n} \log \left(\frac{\sum\limits_{x_1^n \in \mathbb{X}^n} P_{U|X_1^n} P_{X_1^n}}{\sum\limits_{y_1^n \in \mathbb{Y}^n} \sum\limits_{x_1^n \in \mathbb{X}^n} P_{V|Y_1^n,U} P_{Y_1^n|X_1^n} P_{U|X_1^n} P_{X_1^n}} \right).
\end{aligned} \tag{4.147}
$$

The second term $H(V|X_1^n, U)$ can be computed using the folowing Markov chain property

$$
\begin{aligned}
&= H(V|X_1^n, U) \\
&= \sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}\\ x_1^n\in\mathbb{X}^n}} P_{U,V,X_1^n}(u,v,x_1^n)\log\left(\frac{P_{U|X_1^n}(u|x_1^n)}{P_{U,V,X_1^n}(u,v,x_1^n)}\right) \\
&= \sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}\\ x_1^n\in\mathbb{X}^n}} P_{U,V,X_1^n}(u,v,x_1^n)\log\left(\frac{P_{U|X_1^n}(u|x_1^n)}{\sum_{y_1^n\in\mathbb{Y}^n} P_{U,V,X_1^n,Y_1^n}(u,v,x_1^n,y_1^n)}\right) \\
&= \sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}\\ x_1^n\in\mathbb{X}^n}} P_{U,V,X_1^n}(u,v,x_1^n)\log\left(\frac{P_{U|X_1^n}(u|x_1^n)}{\sum_{y_1^n\in\mathbb{Y}^n} P_{V|U,Y_1^n}(v|u,y_1^n)P_{X_1^n|U,Y_1^n}(x_1^n|u,y_1^n)P_{U,Y_1^n}(u,y_1^n)}\right) \\
&= \sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}\\ x_1^n\in\mathbb{X}^n}} P_{U,V,X_1^n}(u,v,x_1^n)\log\left(\frac{P_{U|X_1^n}(u|x_1^n)}{\sum_{y_1^n\in\mathbb{Y}^n} P_{V|U,Y_1^n}(v|u,y_1^n)P_{U,X_1^n,Y_1^n}(u,x_1^n,y_1^n)}\right) \\
&= \sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}\\ x_1^n\in\mathbb{X}^n}} P_{U,V,X_1^n}(u,v,x_1^n)\log\left(\frac{P_{U|X_1^n}(u|x_1^n)}{\sum_{y_1^n\in\mathbb{Y}^n} P_{V|U,Y_1^n}(v|u,y_1^n)P_{Y_1^n|X_1^n}(y_1^n|x_1^n)P_{U|X_1^n}(u|x_1^n)}\right) \\
&= \sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}\\ x_1^n\in\mathbb{X}^n}} P_{U,V,X_1^n}(u,v,x_1^n)\log\left(\frac{1}{\sum_{y_1^n\in\mathbb{Y}^n} P_{V|U,Y_1^n}(v|u,y_1^n)P_{Y_1^n|X_1^n}(y_1^n|x_1^n)}\right) \\
&= \sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}\\ x_1^n\in\mathbb{X}^n\\ y_1^n\in\mathbb{Y}^n}} P_{V|Y_1^n,U}P_{Y_1^n|X_1^n}P_{U|X_1^n}P_{X_1^n}\log\left(\frac{1}{\sum\limits_{y_1^n\in\mathbb{Y}^n} P_{V|U,Y_1^n}P_{Y_1^n|X_1^n}}\right).
\end{aligned}
\tag{4.148}
$$

Again, we do not have access to the marginal distribution of X but we have a training set $\{\overline{x}_1, ..., \overline{x}_m\}$ i.i.d, with $\overline{x}_i \in \mathbb{X}^n$, to introduce the empirical version of the marginal of $P_{X_1^n}(x_1^n)$, by:

$$
\hat{P}_{X_1^n}(x_1^n) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\overline{x}_i = x_1^n}. \tag{4.149}
$$

We also approximate the conditional distribution by its empirical mean. More precisely, given $X_1^n = x_1^n$ we have a set $S_{x_1^n} = \{\overline{y}_1, ..., \overline{y}_{m'}\}$ i.i.d, with $\overline{y}_i \in \mathbb{Y}^n$ and

$$
P_{Y_1^n|X_1^n}(y_1^n|x_1^n) \approx \hat{P}(Y_1^n = y_1^n | X_1^n = x_1^n) = \frac{1}{m'}\sum_{\overline{y}_j \in S_{x_1^n}} \mathbb{1}_{\overline{y}_j = y_1^n}. \tag{4.150}
$$

With this we have that the empirical version of the mutual information $I(V; X_1^n|U)$, by:

$$
\hat{I}_\alpha(V; X_1^n|U) \equiv \hat{H}(V|U) - \alpha\hat{H}(V|X_1^n, U) \tag{4.151}
$$

with

$$\hat{H}(V|U)$$

$$= \sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}\\ x_1^n\in\mathbb{X}^n\\ y_1^n\in\mathbb{Y}^n}} P_{V|Y_1^n,U}\hat{P}_{Y_1^n|X_1^n}P_{U|X_1^n}\hat{P}_{X_1^n}\log\left(\frac{\sum\limits_{x_1^n\in\mathbb{X}^n}P_{U|X_1^n}\hat{P}_{X_1^n}}{\sum\limits_{y_1^n\in\mathbb{Y}^n}\sum\limits_{x_1^n\in\mathbb{X}^n}P_{V|Y_1^n,U}\hat{P}_{Y_1^n|X_1^n}P_{U|X_1^n}\hat{P}_{X_1^n}}\right)$$

$$= \frac{1}{mm'}\sum_{i=1}^{m}\sum_{\overline{y}_j\in S_{\overline{x}_i}}\sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}}}P_{V|Y_1^n,U}(v|\overline{y}_j,u)P_{U|X_1^n}(u|\overline{x}_i)\log\left(\frac{\sum\limits_{l=1}^{m}P_{U|X_1^n}(u|\overline{x}_l)}{\frac{1}{m'}\sum\limits_{l=1}^{m}\sum\limits_{\overline{y}_j\in S_{\overline{x}_l}}P_{V|Y_1^n,U}(v|\overline{y}_j,u)P_{U|X_1^n}(u|\overline{x}_l)}\right) \tag{4.152}$$

and

$$\hat{H}(V|X_1^n,U) = \sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}\\ x_1^n\in\mathbb{X}^n\\ y_1^n\in\mathbb{Y}^n}} P_{V|Y_1^n,U}\hat{P}_{Y_1^n|X_1^n}P_{U|X_1^n}\hat{P}_{X_1^n}\log\left(\frac{1}{\sum\limits_{y_1^n\in\mathbb{Y}^n}P_{V|U,Y_1^n}\hat{P}_{Y_1^n|X_1^n}}\right)$$

$$= \frac{1}{mm'}\sum_{i=1}^{m}\sum_{\overline{y}_j\in S_{\overline{x}_i}}\sum_{\substack{u\in\mathbb{U}\\ v\in\mathbb{V}}}P_{V|Y_1^n,U}(v|\overline{y}_j,u)P_{U|X_1^n}(u|\overline{x}_i)\log\left(\frac{1}{\sum\limits_{\overline{y}_j\in S_{\overline{x}_i}}P_{V|Y_1^n,U}(v|\overline{y}_j,u)}\right). \tag{4.153}$$

Chapter 5

Conclusion

5.1 Concluding Remarks

This thesis presents significant contributions to the area of finite-length analysis, distributed inference, and collaborative HT. By exploring the impact of finite observations and practical communication restrictions on distributed inference, we provide novel results in the form of concrete performance bounds. These results help the understanding of practical limitations and provide the means to design concrete encoder-decoder strategies.

Our contribution on decentralized HT, specifically Theorem 3.2 offers achievable performance bounds that establish non-asymptotic bounds for the TYPE II error when we impose concrete scenarios for the monotonic behavior of $(\epsilon_n)_n$. These results link the gap between practical finite length results and fundamental asymptotic limits offering a deeper understanding of the detection problem with real-world constraints and the presence of degradation sources such as noisy observations, communication restrictions between sensors and decision agents, and the presence of external sources of perturbations. Furthermore, Theorem 4.1 in collaborative HT highlights the importance of collaboration as an inference strategy. The complementary information provided by multiple interaction between sensors and decision agents enhances accuracy. This underscores the significance of collaboration as a strategy for improving hypothesis testing outcomes.

While our theoretical frameworks provide a foundation for understanding the principles and limitations of detection schemes, developing practical methods is crucial for real-world applications. On this dimension, the contributions of this thesis shed light on the design and implementation of algorithms that facilitate effective collaboration between sensors or agents in binary HT scenarios. Our data-driven solutions offer an efficient information exchange and a coordination strategy (encoder design), and an optimal decision-making framework (decoder design). Furthermore, the algorithm proposed in Section 4.5 brings a novel perspective to the area of collaborative HT. Our algorithm establishes conditions to obtain a performance improvements. We see an improvement in the detection performance in scenarios where there is the presence of channel asymmetry. This groundbreaking insight opens up exciting possibilities for optimizing collaborative HT processes, shedding light on how

asymmetry can be leveraged to achieve heightened performance levels.

5.2 Future Work

A relevant topic to be further explored is extending the results presented in this Thesis to the problem of arbitrary binary HT subject to communications constraints. Our contribution that focuses on testing independence can be extended to two arbitrary distributions. These considerations become crucial as they reflect a broader conditions under which decisions could be made. A crucial step in this direction is the characterization of a fundamental limit for arbitrary binary HT setting. Obtaining this result is essential to extend our non-asymptotic study since a critical step was analyzing the discrepancy between non-asymptotic bounds and its corresponding asymptotic limit. Existing contributions that go in this direction are found in [33, 56], where they derived a lower bound for the general bivariate HT.

Another topic to investigate is the extension of collaborative HT with multiple rounds of interactions. In many domains, such as social networks, collaborative environments, or distributed systems, interactions occur between multiple participants simultaneously or over extended periods. Multiple interactions offer potential advantages. Firstly, they enable robustness and reliability in communication [74]. By allowing numerous opportunities for information exchange, errors and disruptions in individual interactions can be mitigated or corrected. This could enhance the overall accuracy of communication, reducing the impact of noise, channel impairments, or transmission failures [23]. Finally, this iterative nature is particularly relevant in contexts such as negotiations, iterative algorithms, or interactive protocols, where multiple rounds of interaction are necessary to converge towards optimal outcomes [39–42]. A possible way to address this problem is by borrowing ideas from [29] combined with the method of types [81].

Finally, the data driven encoder-decoder algorithm proposed in this Thesis has many applications beyond binary HT where digitalization is necessary to make decisions. For example, quantization algorithms with multiple sensors have applications in cryptographic systems. The detection of spoofing attacks against cryptographically-secured signals takes the form of an hypothesis test that accounts for the statistical profile of a replay-type spoofing attack [107]. Discretizing data also provides a means of transforming information into a form that can be securely processed and transmitted. Quantization-based encryption schemes and cryptographic protocols utilize quantization algorithms to ensure the confidentiality and integrity of digital data. In [108], for example, they offer an improved security mechanism, considering the security level for wireless sensor networks. This is an area where our distributed framework could be adopted.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[3] I. Csiszar and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[4] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.

[5] R. M. Gray and L. D. Davisson, *An introduction to statistical signal processing*. Cambridge University Press, 2004.

[6] Y. Wu, Y. Zhou, G. Saveriades, S. Agaian, J. P. Noonan, and P. Natarajan, "Local shannon entropy measure with statistical tests for image randomness," *Information Sciences*, vol. 222, pp. 323–342, 2013.

[7] A. Delgado, "Social conflict analysis on a mining project using shannon entropy," in *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pp. 1–4, IEEE, 2017.

[8] J. Gao, J. Hu, and W.-w. Tung, "Entropy measures for biological signal analyses," *Nonlinear Dynamics*, vol. 68, pp. 431–444, 2012.

[9] R. Zhou, R. Cai, and G. Tong, "Applications of entropy in finance: A review," *Entropy*, vol. 15, no. 11, pp. 4909–4931, 2013.

[10] C. Cachin, *Entropy measures and unconditional security in cryptography*. PhD thesis, ETH Zurich, 1997.

[11] U. M. Maurer, "The role of information theory in cryptography," in *Fourth IMA Conference on Cryptography and Coding*, pp. 49–71, Citeseer, 1993.

[12] P. Bromiley, N. Thacker, and E. Bouhova-Thacker, "Shannon entropy, renyi entropy, and information," *Statistics and Inf. Series (2004-004)*, vol. 9, pp. 2–8, 2004.

[13] A. Monaco, N. Amoroso, L. Bellantuono, E. Lella, A. Lombardi, A. Monda, A. Tateo,

R. Bellotti, and S. Tangaro, "Shannon entropy approach reveals relevant genes in alzheimer's disease," *PloS One*, vol. 14, no. 12, p. e0226190, 2019.

[14] S. Mishra and B. M. Ayyub, "Shannon entropy for quantifying uncertainty and risk in economic disparity," *Risk Analysis*, vol. 39, no. 10, pp. 2160–2181, 2019.

[15] S. Marano and P. K. Willet, "Algorithm and fundamental limits for unlabeled detection using types," *IEEE Transactions on Signal Processing*, vol. 67, no. 8, pp. 2022–2035, 2019.

[16] G. Wang, J. Zhu, R. Blum, P. K. Willet, S. Marano, V. Matta, and P. Braca, "Signal amplitude estimation and detection from unlabeled binary quantized samples," *IEEE Transactions on Signal Processing*, vol. 66, pp. 4291–4303, August 2018.

[17] J. Zhu, H. Cao, C. Song, and Z. Xu, "Parameter estimation via unlabelled sensing using distributed sensors," *IEEE Commun. Letter*, vol. 21, no. 10, pp. 2130–2133, 2017.

[18] S. Marano and P. K. Willet, "The importance of being earnest: Social network with unknown agent quality," *IEEE Transactions on Signal Processing*, vol. 2, pp. 306–320, September 2016.

[19] J. Unnikrishnam, S. Haghighasthoar, and M. Vetterli, "Unlabeled sensing with random linear measurements," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3237–3253, 2018.

[20] Z. Liu and J. Zhu, "Signal detection from unlabeled ordered samples," *IEEE Commun. Letter*, vol. 22, pp. 2431–2434, December 2018.

[21] S. Haghighasthoar and G. Caire, "Signal recovery from unlabeled samples," *IEEE Transactions on Signal Processing*, vol. 66, pp. 1242–1257, March 2018.

[22] S. Bayram, S. Gezici, and H. V. Poor, "Noise enhanced hypothesis-testing in the restricted bayesian framework," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 3972–3989, 2010.

[23] R. Mahler, *Statistical Multisources-Multitarget Information Fusion*. Norwood, MA, USA, 2007.

[24] M. Kendall, A. Stuart, K. J. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics: Volume 2A–Classical Inference and and the Linear Model*. 1999.

[25] R. R. Tenney and N. R. Sandell, "Detection with distributed sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-17, no. 4, pp. 501–510, 1981.

[26] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

[27] V. Strassen, "Asymptotic estimates in Shannon's information theory," in *Proc. 3rd Trans. Prague Conf. Inf. Theory*, pp. 689–723, 2009.

[28] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Transactions on Information Theory*, vol. 32, no. 4, pp. 533–542, 1986.

[29] Y. Xiang and Y.-H. Kim, "Interactive hypothesis testing with communication constraints," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1065–1072, IEEE, 2012.

[30] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, pp. 493–507, 1952.

[31] M. C. Neale, A. Heath, J. Hewitt, L. Eaves, and D. Fulker, "Fitting genetic models with lisrel: Hypothesis testing," *Behavior genetics*, vol. 19, no. 1, pp. 37–49, 1989.

[32] C.-Y. Liao, T. E. Johnson, and J. F. Nelson, "Genetic variation in responses to dietary restriction—an unbiased tool for hypothesis testing," *Experimental gerontology*, vol. 48, no. 10, pp. 1025–1029, 2013.

[33] T. Han, "Hypothesis testing with multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 33, no. 6, pp. 759–772, 1987.

[34] V. Y. Tan *et al.*, "Asymptotic estimates in information theory with non-vanishing error probabilities," *Foundations and Trends® in Communications and Information Theory*, vol. 11, no. 1-2, pp. 1–184, 2014.

[35] V. Kostina, *Lossy data compression: nonasymptotic fundamental limits*. PhD thesis, Princeton University, 2013.

[36] I. Kontoyiannis and S. Verdú, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 777–795, 2014.

[37] M. Hayashi, "Second-order asymptotics in fixed-length source coding and intrinsic randomness," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4619–4637, 2008.

[38] E. Soltanmohammadi, M. Orooji, and M. Naraghi-Pour, "Decentralized hypothesis testing in wireless sensor networks in the presence of misbehaving nodes," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 205–215, 2012.

[39] M. Alanyali, S. Venkatesh, O. Savas, and S. Aeron, "Distributed bayesian hypothesis testing in sensor networks," in *Proceedings of the 2004 American control conference*, vol. 6, pp. 5369–5374, IEEE, 2004.

[40] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Muller, E. Sackinger, P. Simard, *et al.*, "Comparison of classifier methods: a case study in handwritten digit recognition," in *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, vol. 2, pp. 77–82, IEEE, 1994.

[41] S. Theodoridis and K. Koutroumbas, *Pattern recognition*. Elsevier, 2006.

[42] R. Cogranne and J. Fridrich, "Modeling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2627–2642, 2015.

[43] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[44] K. Nakagawa and F. Kanaya, "On the converse theorem in statistical hypothesis testing," *IEEE Transactions on Information Theory*, vol. 39, no. 2, pp. 623–628, 1993.

[45] Z. Zhang, E.-H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion-part one: Known statistics," *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 71–91, 1997.

[46] M. Boudiaf, I. Ziko, J. Rony, J. Dolz, P. Piantanida, and I. Ben Ayed, "Information maximization for few-shot learning," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 2445–2457, Curran Associates, Inc., 2020.

[47] J. Chamberland and V. V. Veeravalli, "Wireless sensors in distributed detection applications," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 16–25, 2007.

[48] Dan Li, K. D. Wong, Yu Hen Hu, and A. M. Sayeed, "Detection, classification, and tracking of targets," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 17–29, 2002.

[49] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[50] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. R. Soc. Lond. A*, vol. 231, no. 694-706, pp. 289–337, 1933.

[51] S. Kay, "A new proof of the neyman–pearson theorem using the eef and the vindication of sir r. fisher," *IEEE Signal Processing Letters*, vol. 19, no. 8, pp. 451–454, 2012.

[52] T. S. Han and K. Kobayashi, "Exponential-type error probabilities for multiterminal hypothesis testing," *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 2–14, 1989.

[53] I. Sason, "Moderate deviations analysis of binary hypothesis testing," in *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 821–825, IEEE, 2012.

[54] S. Espinosa, J. F. Silva, and P. Piantanida, "New results on testing against independence with rate-limited constraints," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, IEEE, 2019.

[55] Y. Nikitin, *Asymptotic efficiency of nonparametric tests*. Cambridge University Press, 1995.

[56] S. Watanabe, "Neyman–pearson test for zero-rate multiterminal hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 4923–4939, 2018.

[57] P. K. Varshney, *Distributed Detection and Data Fusion*. Berlin, Heidelberg: Springer-Verlag, 1st ed., 1996.

[58] W. Baek and S. Bommareddy, "Optimal m-ary data fusion with distributed sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 31, no. 3, pp. 1150–1152, 1995.

[59] C. Tepedelenlioglu and S. Dasarathan, "Distributed detection over gaussian multiple access channels with constant modulus signaling," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2875–2886, 2011.

[60] G. Mergen, V. Naware, and L. Tong, "Asymptotic detection performance of type-based multiple access over multiaccess fading channels," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1081–1092, 2007.

[61] V. S. S. Nadendla and P. K. Varshney, "Design of binary quantizers for distributed detection under secrecy constraints," *IEEE Transactions on Signal Processing*, vol. 64, no. 10, pp. 2636–2648, 2016.

[62] R. Chandramouli and N. Ranganathan, "Quantization for robust sequential m-ary signal detection," in *ISCAS '98. Proceedings of the 1998 IEEE International Symposium on Circuits and Systems (Cat. No.98CH36187)*, vol. 4, pp. 317–320 vol.4, 1998.

[63] M. Wigger and R. Timo, "Testing against independence with multiple decision centers," in *2016 International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5, IEEE, 2016.

[64] S. Salehkalaibar, M. Wigger, and R. Timo, "On hypothesis testing against conditional independence with multiple decision centers," *IEEE Transactions on Communications*, vol. 66, no. 6, pp. 2409–2420, 2018.

[65] K. R. Varshney and L. R. Varshney, "Quantization of prior probabilities for hypothesis testing," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4553–4562, 2008.

[66] J. B. Rhim, L. R. Varshney, and V. K. Goyal, "Quantization of prior probabilities for collaborative distributed hypothesis testing," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4537–4550, 2012.

[67] A. Dembo and O. Zeitouni, "Large deviations techniques and applications. 1998," *Applications of Mathematics*, vol. 38, 2011.

[68] S. Espinosa, J. F. Silva, and P. Piantanida, "Finite-length bounds on hypothesis testing subject to vanishing type i error restrictions," *IEEE Signal Processing Letters*, January 2021.

[69] G. Katz, P. Piantanida, and M. Debbah, "Collaborative distributed hypothesis testing,"

CoRR, vol. abs/1604.01292, 2016.

[70] Y. Shkel, M. Raginsky, and S. Verdú, "Universal lossy compression under logarithmic loss," in *Information Theory (ISIT), 2017 IEEE International Symposium on*, pp. 1157–1161, IEEE, 2017.

[71] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[72] M. Vera, L. R. Vega, and P. Piantanida, "Compression-based regularization with an application to multitask learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 1063–1076, 2018.

[73] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[74] R. Gallager, *Principles of digital communication*. Technical Publications, 2008.

[75] L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[76] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[77] T. Berger, "Rate-distortion theory," *Encyclopedia of Telecommunications*, 1971.

[78] T. S. Han, *Information-spectrum methods in information theory*. Springer Science & Business Media, 2013.

[79] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[80] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 326–339, 1948.

[81] I. Csiszár, "The method of types," *Information Theory, IEEE Transactions on*, vol. 44, pp. 2505–2523, Oct 1998.

[82] S. Espinosa, J. F. Silva, and P. Piantanida, "On the exponential approximation of type ii error probability of distributed test of independence," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 7, pp. 777–790, 2021.

[83] G. Pichler, P. Piantanida, and G. Matz, "Distributed information-theoretic biclustering of two memoryless sources," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 426–433, IEEE, 2015.

[84] Feng Zhao, Jaewon Shin, and J. Reich, "Information-driven dynamic sensor collaboration," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 61–72, 2002.

[85] M. Mhanna, P. Piantanida, and P. Duhamel, "Privacy-preserving quantization learning

with applications to smart meters," in *IEEE International Conference on Communications, ICC 2017, Paris, France, May 21-25, 2017*, pp. 1–6, IEEE, 2017.

[86] M. Chen, W. Liu, B. Chen, and J. Matyjas, "Quantization for distributed testing of independence," in *2010 13th International Conference on Information Fusion*, pp. 1–5, IEEE, 2010.

[87] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication-rate scheme for distributed detection," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 2, pp. 554–568, 1996.

[88] M. Mhanna and P. Piantanida, "On secure distributed hypothesis testing," in *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1605–1609, IEEE, 2015.

[89] A. Nosratinia, T. E. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE communications Magazine*, vol. 42, no. 10, pp. 74–80, 2004.

[90] K. B. Letaief and W. Zhang, "Cooperative communications for cognitive radio networks," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 878–893, 2009.

[91] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 7, pp. 1294–1309, 2008.

[92] Z. Peng, L. Zhang, and T. Luo, "Learning to communicate via supervised attentional message processing," in *Proceedings of the 31st International Conference on Computer Animation and Social Agents*, pp. 11–16, 2018.

[93] A. v. d. Oorod, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *arXiv preprint arXiv:1711.00937*, 2017.

[94] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[95] B. Dai, C. Zhu, B. Guo, and D. Wipf, "Compressing neural networks using the variational information bottleneck," in *International Conference on Machine Learning*, pp. 1135–1144, PMLR, 2018.

[96] M. Vera, L. R. Vega, and P. Piantanida, "Compression-based regularization with an application to multitask learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 1063–1076, 2018.

[97] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[98] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[99] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new

perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[100] H. Schulz, A. Müller, S. Behnke, *et al.*, "Investigating convergence of restricted boltzmann machine learning," in *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, vol. 1, pp. 6–1, 2010.

[101] F. X. Albizuri, A. d'Anjou, M. Graña, and J. A. Lozano, "Convergence properties of high-order boltzmann machines," *Neural networks*, vol. 9, no. 9, pp. 1561–1567, 1996.

[102] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[103] A. Schrijver, *Theory of linear and integer programming*. John Wiley & Sons, 1998.

[104] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge University Press, 2011.

[105] A. Kaspi, "Two-way source coding with a fidelity criterion," *Information Theory, IEEE Transactions on*, vol. 31, pp. 735–740, Nov 1985.

[106] P. Piantanida, L. Rey Vega, and A. Hero, "A proof of the generalized markov lemma with countable infinite sources," in *Information Theory Proceedings (ISIT), 2014 IEEE International Symposium on*, July 2014.

[107] T. E. Humphreys, "Detection strategy for cryptographic gnss anti-spoofing," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 49, no. 2, pp. 1073–1090, 2013.

[108] U. Panahi and C. Bayılmış, "Enabling secure data transmission for wireless sensor networks based iot applications," *Ain Shams Engineering Journal*, vol. 14, no. 2, pp. 1–11, 2023.