

A Functional Perspective on Information Measures

Présentée le 17 juin 2022

Faculté informatique et communications
Laboratoire d'information dans les systèmes en réseaux
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Amedeo Roberto ESPOSITO

Acceptée sur proposition du jury

Dr O. Lévêque, président du jury
Prof. M. C. Gastpar, directeur de thèse
Prof. M. Raginsky, rapporteur
Prof. I. Sason, rapporteur
Prof. E. Telatar, rapporteur

To the loving memory of my mother.
For you planted in me, amongst other things,
the fruitful seed of curiosity
and the unquenchable thirst for understanding.

The mathematician's patterns, like the painter's or the poet's must be beautiful;
the ideas, like the colours or the words must fit together in a harmonious way.
Beauty is the first test: there is no permanent place in this world for ugly mathematics.
— G.H. Hardy

Acknowledgements

This five-year journey has been particularly tough on me. I have (l)earned so much but, at the same time, lost so much. I have been very close to giving up more than once, but I persisted. Facing these challenges along the way rendered these acknowledgments ever more important. I am unsure where I would be now without every and each one of you.

First and foremost, I would like to express my gratitude to my supervisor: Michael Gastpar. The qualities that I could list here are far too many; I will thus focus on what impressed me the most and what I will carry with me in my future: your humility, work ethic, and constant presence. We met at least once every week, no matter how busy/tired/jet-lagged you were. You sat there, listened to whatever I had to say, and treated me as a respectable scientist from the first day. Even when I had no idea what I was talking about, you assumed there was something clever that *you* did not get from my words (I wish that were true!). You are someone not easy to impress, but trying to do so each and every week pushed me towards becoming a much better scientist.

I am grateful to the members of my jury committee: Emre Telatar, Olivier Levêque, Maxim Raginsky, and Igal Sason, for reading my thesis and for providing precious comments. While it is expected to acknowledge you, this is more than just duty. I chose each of you very carefully: you have all inspired me or taught me something meaningful in one way or another, and for this, I am grateful.

A big thank you goes to France Faille. Not only have you made my life at EPFL that much easier, but you always did more than “just your job”. I sincerely appreciated your kindness, reliability, and the fact that you always welcomed me in your (extremely relaxing and exotic) office with a big smile and helpful suggestions.

I owe an enormous debt of gratitude to my mentor, Ugo Vaccaro. Your passion for science is genuinely inspiring. You supported me from the beginning; you were always ready to provide advice and (continuously) tried to set me on the right path. A path I believe I have found: certainly not the easiest but undoubtedly the most meaningful to me. During my many mystical crises, the thought that I could someday influence students like you influenced me (and many others) has been one of the few sources of comfort. I would not be here if it weren't for you in many different ways, and, for that, I truly thank you.

And then there's Viktor Sanca, a fellow Ph.D. student who has become a close friend. Someone with whom I could freely rant, complain and moan without ever feeling judged but constantly feeling understood and supported. Your solution to my problems was cooking me overly caloric meals, a concept that I know all too well as an Italian. While “my dog days are over”

Acknowledgements

now, I hope I will be able to do the same for you.

I shared my path at EPFL with many people, shared many discussions, and blended into many different cultures. I have eaten many exotic treats (one too many pieces of baklava) and enjoyed hearing many stories. Thank you: Aditya, Bora, Erixhen, Ibrahim, Ido, Ignacio, Kirill, Marco, Reka, Su, Yanina and Yunus. A small special mention goes to Pierre: we have a similar approach to science and similar interests (which is rare in and of itself). I have deeply appreciated our conversations, even if sometimes they were some of the craziest I ever had. Talking with you, I did not feel so “out of place” as with some others. If anything, you are even more “out of place” than I am.

I am also very grateful for having had the opportunity to follow two students and accompany them while taking their first steps in research. Thank you, Adrien and Diyuan: the process of trying to teach you has taught me more than I would have ever expected. You came looking for guidance, but I truly feel like I am the one that gained something from our projects together. You are two bright minds whose passion and interests meaningfully nurtured mine.

And then there are those people who had little or nothing to do with Lausanne and EPFL: my oldest friends and family. Let me thank them in chronological order (so you can stop wondering whether first or last is more important :)).

Let us thus start with the guys from Unisa: Alessia, Enrica, Luigi, Maria Angela, Martina, Michel, and Simone. A group of (self-declared) socially awkward people with one too many Ph.D.'s. We have been facing similar challenges for the past ten years, and sharing them with you has made things much easier. Thank you for having listened and advised but most importantly, thank you for having been there despite the consistent physical distance that separates us.

Then there's “I servi del Naso”: Antonio, Federica, Nunzia, Simona and Veronica. We have known each other for a very long time. We have grown up over the years and lived in different countries but have always been there for each other. Sometimes, the only thing that kept me going was knowing that, eventually, I would fly home, and we would all have a pizza together, laugh for stupid reasons and jump back in time, even if just for an evening. We have all been through a lot, but thank you for the lightness and laughter that you kept bringing into my life. And then there is the person who stood by me the longest: Memela. It is probably impossible to put in writing how grateful I am for you. When I think of us, I think of two damaged souls that found each other at the right moment. We choose each day to face the difficulties together, some days one leaning on the other just a bit more, in the hope of one day strolling through life with our head held high, hand in hand. Thank you for being in my life.

Last but not least, my family. Thank you to my sisters: Michela, Angela, and Laura. You always believed so much in me and took for granted that I would succeed even when it was unclear: this gave me strength. If there is one thing I never lacked, was support from my family. You have always been there for me, ready to console me for every defeat and to celebrate with me every victory.

Thank you to my brother-in-law, Francesco: you always welcomed me into your house and made me feel like family. Thank you for standing by us when we needed it the most.

Thank you to my father, a man who fell on his knees one too many times, and yet he is still standing, ready to extend his hand in case I needed it.

Flying home for the holidays, knowing you all would be there, was yet another thing that kept me going. One of the very few certainties, even when my (or the) world turned upside-down. Thank you to my mother: you taught me how to walk, started my path with me, and then watched me run freely upon it. You infused me with your passion for math and teaching. I always had the freedom to make my own choices, but you were never too far away from me and were always ready to pick me up when I fell. I fell pretty often in these years of your absence. Getting back up has never felt (and, probably, never will) quite as safe. Yet, I did get up, hoping I would make you proud just a bit more.

All in all, I have been pretty lucky so far. Each person I mentioned above has enriched my life in many different ways and shaped my path, and I am fortunate to have met them. People like to say “we are the sum of our experiences” and that is true (scientifically and philosophically). However, I like to think that what truly leaves a mark on our beings are people. The acts of kindness, the mutual support, the selflessness, the closeness, and the moments of empathy: are the things that I will never ever forget. I am deeply grateful for each time you dedicated even just a minute to me and contributed to who I am today.

Lausanne, June 1, 2022

Abstract

Since the birth of Information Theory, researchers have defined and exploited various *information measures*, as well as endowed them with operational meanings. Some were born as a “solution to a problem”, like Shannon’s Entropy and Mutual Information. Others were the fruit of generalisation and the mathematical genius of bright minds like Rényi, Csizsár and Sibson, but started being used in practical problems only later on. These powerful objects allow us to manipulate probabilities intuitively and seem always to be somehow connected to concrete settings in communication, coding or estimation theory. A common theme is: take a problem in one of these areas, try to control (*i.e.*, upper or lower-bound) the expected value of some function of interest (often, probabilities of error) and, with enough work, an information measure appears as a fundamental limit of the problem.

The most striking example of this is in Shannon’s seminal paper in 1948: his purpose was to characterise the *smallest possible expected length* of a uniquely decodable encoding that compresses the realisations of a random variable. As he brilliantly proved, the smallest expected length one can hope for is the Entropy of the random variable. In establishing this connection, another quantity needed to be implicitly controlled: the Kraft’s sum of the code. Seemingly unrelated before, these three objects joined forces in harmony to provide a beautiful and fundamental result. But why are they even related? The answer seems to be: duality. Duality is an abstract notion commonly used in linear algebra and functional analysis. It has been expanded and generalised over the years and several incarnations have been discovered throughout mathematics. Still, its essence remained the same: “given a mathematical object one can associate to it a “dual” object that helps one to understand the properties of the object one started with” (Gowers et al., 2010). One particular instance of this involves vector spaces: given two vector spaces and a “duality pairing” (*e.g.*, a bilinear mapping defined on the Cartesian product of these two spaces), one can jump from one space to the other (its dual) through Legendre-Fenchel-like transforms. In the most common settings in Information Theory, the two spaces and the pairing are, respectively:

- 1) the space of (probability) measures defined on \mathcal{X} ;
- 2) the space of bounded functions defined on \mathcal{X} ;
- 3) the Lebesgue integral of the function (*e.g.*, the expected value of the function if the measure is a probability measure).

Once these are set, Legendre-Fenchel-like transforms allow us to connect

- a) a functional acting on the space described in Item 1);
- b) a functional acting on the space described in Item 2);

and the anchor point is

- c) the (expected) value described in Item 3).

These three pieces (Items a), b) and c)) represent the actors of many of the results provided in Information Theory. Once they are found, one usually bounds the functional described in Item b) and obtains a bound connecting the expected value (*i.e.*, the pairing, cf. Item c)) and the functional of measures (*e.g.*, an information measure, cf. Item a)).

Going back to Shannon's result, fixing a random variable (and thus, a probability measure) and selecting the function to be the length of a code:

- a1) the functional a) is the Shannon's Entropy of the source;
- b2) the functional b) is the Kraft's sum of the code, *i.e.*, the dual of Shannon's Entropy;
- c3) the pairing c) is the expected length of the code.

We explore this connection and this pattern throughout the thesis. We will see how it can be found in notable results like Coding Theorems for one-to-one (not necessarily uniquely decodable) codes, Campbell's Coding Theorem, Arikan's Guessing Theorem, Fano-like and Transportation-Cost Inequalities and so on. Moreover, unearthing the pattern allows us to generalise it to other information measures and apply the technique in a variety of fields, including:

- Learning Theory;
- Estimation Theory;
- Hypothesis Testing.

In particular, we provide results connecting expectations of the same function with two different measures to a divergence between these two measures. This allows us to generalise Transportation-Cost inequalities to divergences other than the Kullback-Leiber's. It will also enable us to retrieve many results in the literature connecting the expected generalisation error of learning algorithms to divergences and to generalise them to virtually any information measure.

As a particular case, we also consider probabilities of the same event under different probability measures. This allows us to generalise concentration of measure to functions of random variables that depend on the random variables themselves and, in turn, bound the probability of having a large generalisation error in supervised learning settings. Other applications can be found in lower-bounding the Bayesian risk in estimation procedures and analysing the error probability in hypothesis testing frameworks.

Abstract

Sin dalla nascita della Teoria dell'Informazione, i ricercatori hanno definito ed utilizzato varie *misure di informazione* dotandole inoltre di significati operativi. Alcune di esse sono nate come "soluzione ad un problema", come l'Entropia di Shannon e la Mutua Informazione. Altre sono invece state il frutto della generalizzazione e del genio matematico di menti brillanti come Rényi, Csizszár e Sibson ma sono state utilizzate in problemi pratici solo più tardi. Questi potenti strumenti ci permettono di manipolare le probabilità in maniera intuitiva e sembrano essere, in qualche modo, sempre connessi ad ambiti concreti come la teoria della comunicazione, della codifica o della stima di parametri. Un tema ricorrente è il seguente: considera un problema in una di queste aree, prova a controllare (*i.e.*, limitando superiormente o inferiormente) il valore di atteso di una specifica funzione di interesse (spesso, probabilità di errore) e, lavorando abbastanza, una misura di informazione sembra sempre apparire come limite fondamentale del problema.

L'esempio più lampante di ciò è nell'influente lavoro di Shannon del 1948: il suo obiettivo era quello di caratterizzare la *minima lunghezza attesa possibile* di un codice unicamente decifrabile che comprime le realizzazioni di una variabile casuale. Come ha brillantemente dimostrato, la minima lunghezza attesa a cui si può ambire è l'Entropia della variabile casuale. Nello stabilire tale connessione fu implicitamente necessario controllare un'altra quantità: la somma di Kraft del codice. Questi tre oggetti, alla prima apparenza non correlati tra loro, hanno unito le forze in armonia regalandoci un elegante e fondamentale risultato. Ma perchè tali oggetti sono correlati? La risposta sembra essere: dualità. La Dualità è una nozione astratta comunemente utilizzata in algebra lineare ed analisi funzionale. È stata generalizzata ed espansa negli anni e ne sono state individuate diverse istanze in vari ambiti della matematica. La sua essenza, però, è rimasta la stessa: "dato un oggetto matematico è possibile associare ad esso un oggetto "duale" che aiuta a comprendere le proprietà dell'oggetto di partenza" (Gowers et al., 2010). Una particolare istanza di questo fenomeno riguarda spazi vettoriali: dati due spazi vettoriali ed un "accoppiamento di dualità" (*e.g.*, una funzione bilineare definita sul prodotto Cartesiano di questi due spazi), è possibile saltare da uno spazio all'altro (il suo duale) tramite trasformate simili a quella di Legendre-Fenchel. Negli ambiti più comuni della Teoria dell'Informazione, questi due spazi ed il relativo accoppiamento sono, rispettivamente:

- 1) lo spazio delle misure (di probabilità) definite su \mathcal{X} ;
- 2) lo spazio di funzioni limitate definite su \mathcal{X} ;
- 3) l'integrale secondo Lebesgue della funzione (*e.g.*, il valore atteso della funzione se la misura è una misura di probabilità).

Una volta che questi elementi sono fissati, trasformate simili a quella di Legendre-Fenchel ci permettono di connettere:

- a) un funzionale che agisce sullo spazio descritto al punto 1);
- b) un funzionale che agisce sullo spazio descritto al punto 2);

ed il punto di connessione è

- c) il valore atteso descritto al punto 3).

Questi tre elementi (Punti a), b), and c)) rappresentano gli ingredienti di molti dei risultati dimostrati in Teoria dell'Informazione. Una volta individuati, si procede tipicamente limitando il funzionale descritto al punto b) e si può così ottenere un risultato che connette il valore atteso (*i.e.*, l'accoppiamento, cf. punto c)) ed il funzionale che agisce su misure (*e.g.*, una misura di informazione, cf. punto a)). Ritornando al risultato di Shannon, fissata una variabile casuale (e quindi, una misura di probabilità) e selezionando come funzione la lunghezza del codice:

- a1) il funzionale a) è l'Entropia di Shannon della sorgente;
- b2) il funzionale b) è la somma di Kraft del codice, *i.e.*, il duale dell'Entropia di Shannon;
- c3) l'accoppiamento c) è il valore atteso della lunghezza del codice.

Questo pattern e questa connessione verranno esplorati più volte in questa tesi. Vedremo come questo schema possa essere ritrovato in risultati noti come Teoremi di Codifica per codici uno-ad-uno (non necessariamente unicamente decifrabili), il Teorema di Codifica di Campbell, il Teorema sul Guessing di Arıkan, disuguaglianze simili a quella di Fano e del Trasporto-Costo e così via. Inoltre, portando alla luce questo pattern ci permette di generalizzarlo ad altre misure di informazione e di applicare la tecnica ad una varietà di ambiti, tra cui:

- Teoria dell'Apprendimento Automatico;
- Teoria delle Stime;
- Test d'Ipotesi.

In particolare, fornirò risultati che connettono i valori attesi della stessa funzione rispetto a due misure differenti ed una divergenza tra queste due misure. Questo ci permette di generalizzare le disuguaglianze del Trasporto-Costo a divergenze diverse da quella di Kullback-Leibler. Ci permetterà inoltre di recuperare molti dei risultati presenti in letteratura e che connettono il valore atteso dell'errore di generalizzazione di un algoritmo di apprendimento a divergenza e di generalizzarli a quasi tutte le misure di informazione.

Come caso particolare, considereremo anche le probabilità di uno stesso evento rispetto a misure di probabilità diverse. Questo ci permetterà di generalizzare risultati di concentrazione delle misure a funzioni di variabili casuali che dipendono dalle variabili casuali stesse e, di conseguenza, limitare la probabilità che l'errore di generalizzazione sia grande in ambiti di apprendimento supervisionato. Altre applicazioni possono essere trovate nel limitare inferiormente il rischio Bayesiano in procedure di stima dei parametri e nell'analisi della probabilità di errore in ambiti di test d'ipotesi.

Contents

Acknowledgements	i
Abstract (English/Italian)	v
Introduction	1
Organisation of the Thesis	3
Related Work	4
I The Theory	7
1 Preliminaries	9
1.1 Banach Spaces	9
1.2 Duality	10
1.3 Orlicz Spaces	13
1.3.1 Hölder's inequality Duality	17
1.4 Divergences	19
1.4.1 Kullback-Leibler Divergence	19
1.4.2 φ -Divergences	20
1.4.3 Rényi's α -Divergences and Sibson's α -Mutual Information	21
2 Duality and Divergences	25
2.1 Introduction	25
2.1.1 Kullback-Leibler Divergence	26
2.1.2 Shannon's Entropy and Shannon's Coding Theorem	28
2.1.3 Beyond unique-decodability	32
2.2 Rényi's Entropy	34
2.2.1 A variational Representation for Rényi's Entropy	34
2.2.2 Campbell's Coding Theorem	36
2.2.3 Arikan's Guessing Theorem	38
2.2.4 Rényi's Divergence Variational Representation	40
2.3 A Variational Representation for φ -Divergences	43
3 Independence Vs Dependence	47
3.1 Ratio of Probabilities	48

Contents

3.1.1	General Results	48
3.1.2	Rényi's Information Measures	53
3.1.3	Other divergences	55
3.1.4	Tightness	57
3.1.5	Sibson's α -Mutual Information and its functional inequalities	58
3.2	Difference of Expectations	63
3.2.1	Generalising Transportation-cost Inequalities	65
3.2.2	Beyond Sub-Gaussianity	70
3.2.3	Transportation-cost inequalities with Rényi's α -Divergences	71
3.2.4	Orlicz spaces, tails of random-variables and the Kullback-Leibler Divergence	74
3.3	Application: Hypothesis Testing	75
3.3.1	Bounding the ratio of errors in Hypothesis Testing	78
Appendices		81
3.A	Hölder's inequality and information measures	81
3.B	Alternative proof of Theorem 13	83
II The Applications		85
4 Learning Theory		87
4.1	Background	87
4.2	The expected generalisation error	88
4.2.1	Recovering other known results	91
4.2.2	Generalising Individual Samples and Conditional Mutual-Information	93
4.3	The probability of having a large generalisation error	95
4.3.1	Sibson's α -Mutual Information	95
4.3.2	Maximal Leakage	97
4.3.3	Analysing Schemes via Maximal Leakage	98
4.3.4	Other Divergences	99
4.3.5	From Probability to Expected Value	101
4.3.6	Some considerations	103
4.4	Comparison with other bounds	104
4.4.1	Maximal Leakage and Mutual Information	104
4.4.2	Maximal Leakage and Differential Privacy	106
4.4.3	Sibson's Mutual Information, Maximal Leakage and Max Information	108
4.5	Adaptive Data Analysis	111
Appendices		115
4.A	Properties of Maximal Leakage	115
4.B	Examples	116
4.B.1	Proof of Lemma 8	116

5	Bayesian Risk in Estimation Procedures	119
5.1	Introduction	119
5.2	Problem Setting	120
5.3	The lower-bounds	120
5.4	Examples	125
5.4.1	Bernoulli Bias	125
5.4.2	Gaussian prior with Gaussian noise (and absolute error)	130
5.4.3	Hide-and-seek problem	132
5.5	Other approaches and generalisations	135
5.5.1	Inverting the roles	135
5.5.2	Conditioning	136
5.5.3	Leveraging SDPIs	137
5.5.4	Lower-Bounding the Risk Directly	139
5.5.5	Concave conjugates	142
 Bibliography		 145

Introduction

Duality is an important general theme that has manifestations in almost every area of mathematics.

(Gowers et al., 2010, Part III:
Mathematical Concepts - 19. Duality)

Duality is an abstract notion that comes in various shapes and forms. This thesis explores several notions of duality and their important roles in Information Theory. They will all stem from a duality pairing *i.e.*, a bilinear mapping between objects belonging to two (different) spaces. Once the pairing is set, it is possible to jump from one space to the other through Legendre-Fenchel-like transforms. This statement represents the central intuition behind this thesis and will allow us to interpret (old and new) results differently.

Since the birth of the field, a common theme in Information Theory has been the connection between expected values of functions and information measures. The most striking example of this can be found in the groundbreaking work of Shannon, in which he connected the smallest possible expected length of a uniquely-decodable code (*i.e.*, the expected value of a function) to what is now known as “Shannon’s Entropy” (*i.e.*, the information measure). One object (the function) lives in a space, while the other (the information measure) acts on the corresponding “dual” space. The thread connecting these two spaces is the “duality pairing”. Given a measure μ and a function f , a natural pairing is given by the Lebesgue integral:

$$\langle \mu, f \rangle = \mu(f) = \int f d\mu.$$

If the measure μ is a probability measure, the pairing is simply the expected value of f under μ . Thus, duality allows us to jump from the space of functions to the space of measures, and the connecting thread will be the *expected value*. It represents a simple yet essential object: probabilities of events themselves can be seen as expected values of indicator functions. The abundance of its presence in Information and Probability Theory implies that these fields have implicitly leaned on duality itself.

Other, almost omnipresent, objects in Information Theory are information measures. Most of their power comes from convexity and, in a way, from the connection between Duality

Introduction

and Convexity. Indeed, given a function $\psi : \mathcal{X} \rightarrow \mathbb{R}$ one can compute its Legendre-Fenchel transform. This leads us to a function ψ^* acting on the “dual” space of \mathcal{X} (which we will often denote as \mathcal{X}^*):

$$\psi^*(\mu) = \sup_{f \in \mathcal{X}} \langle \mu, f \rangle - \psi(f) \quad \text{with } \mu \in \mathcal{X}^*. \quad (1)$$

The object that we retrieve, ψ^* , is guaranteed to be convex (and lower semi-continuous with a suitable topology). If ψ itself is convex, then Equation (1) is guaranteed to be well-defined, and one has by the Fenchel-Moreau Theorem that $(\psi^*)^* = \psi$ (in reality, one must be a bit more careful. In some cases the domain of $(\psi^*)^*$ can be much larger than domain of ψ . The property holds nonetheless on the intersection of the spaces). This implies that:

$$\psi(f) = (\psi^*)^*(f) = \sup_{\nu} \langle \nu, f \rangle - \psi^*(\nu). \quad (2)$$

Hence, one can find the function ψ of which ψ^* is the Legendre-Fenchel dual (through Equation (2)) and vice versa.

Consider now divergences. Let us fix the second measure and see them as functionals acting on the first measure, *i.e.*,

$$D(\cdot \| \mu) : \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}. \quad (3)$$

One has that (for the objects of interest to us and in Information Theory) $D(\cdot \| \mu)$ is convex. The considerations just above essentially imply that they possess a representation of this form:

$$D(\nu \| \mu) = \psi_{\mu}^*(\nu) = \sup_f \langle \nu, f \rangle - \psi_{\mu}(f). \quad (4)$$

Many of these characterisations have been established in the literature at some point or another (Varadhan, 1984; Anantharam, 2017; Broniatowski and Keziou, 2010). As we will see throughout this thesis, they have been quite fruitful in providing different results in various fields.

Sometimes, however, this connection has gone unnoticed. Part of the contributions of this document consists in unearthing this connection. The journey will take us from Shannon’s Coding Theorem to Transportation-Cost Inequalities and Learning Theory. In the more classical results, the assumptions (or parts of the proof) often consisted in *implicitly* bounding the dual of the information measure of interest. This crucial step led then to a result connecting the duality pairing (*e.g.*, $\langle f, \nu \rangle$ and thus, the expected value of some function f) to the information measure (*e.g.*, $\psi^*(\nu)$, a divergence or entropy).

Other times, duality appears in disguise. For instance, some information measures are (transformations of) norms. For these divergences, the “dual” that we will consider (and bound, in some cases) will be the *dual norm*.

The essence, however, remains the same: the “dual norm” is defined on the “dual space” which is determined by the pairing. This represents a different way of jumping from functions to measures.

These abstract and intuitive statements will become more precise as the chapters evolve.

Multiple examples from the classical (and not so classical) Information Theory literature will be framed in this setting. Moreover, these very same tools will be used to derive new results, which will then be applied to a variety of settings, including:

- Learning Theory;
- Estimation Theory;
- Hypothesis Testing.

Organisation of the Thesis

This thesis is organised as follows:

- in Chapter 1 we will explore the basic objects and results that will be needed throughout the thesis:
 - Function Spaces (Banach and Orlicz Spaces, cf. Section 1.1 and Section 1.3);
 - Duality (Legendre-Fenchel Duality and Dual Norms, cf. Section 1.2);
 - Divergences (Kullback-Leibler's, Rényi's and φ -Divergences, cf. Section 1.4).
- in Chapter 2 the various notions of duality introduced in Chapter 1 will be instantiated to information measures. We will take a close look at the variational representations that stem from this duality for:
 - the Kullback-Leibler Divergence (cf. Section 2.1.1) (*i.e.*, the celebrated Donsker-Varadhan representation of the Kullback-Leibler Divergence);
 - Shannon's Entropy (cf. Section 2.1.2) and, as an immediate application, we will recover Shannon's Converse Coding Theorem as well as some results on one-to-one codes (cf. Section 2.1.3);
 - Rényi's entropy (cf. Section 2.2.1) with a variational representation that echoes the one provided for Rényi's divergences (cf. Section 2.2.4). As applications of this we will recover Campbell's coding theorem (cf. Section 2.2.2) and Arıkan's Guessing Theorem (cf. Section 2.2.3);
 - φ -Divergences for strongly convex function(al)s φ (cf. Section 2.3).

Most of these variational representations will then find further applications in the subsequent chapters;

- in Chapter 3 we will use the results presented in Chapter 2 to connect the expected value of a function f under a measure μ with the expected value of f but under a different measure ν . As a consequence, various information measures will appear in the analysis. When f is chosen to be the indicator function of an event E (cf. Section 3.1), we are essentially connecting $\mu(E)$ with $\nu(E)$ and some divergence $\hat{D}(\mu\|\nu)$.

The bounds provided will involve:

- very general objects, like Luxemburg and Amemiya norms of the Radon-Nikodym derivative and of the indicator function $\mathbb{1}_E$ (cf. Section 3.1.1);
- Rényi's information measures, *i.e.*, Rényi's Divergences, Sibson's α -Mutual Information and, in particular, Maximal Leakage (cf. Section 3.1.2);

Introduction

- other divergences, like the Hellinger– p divergences and, in particular, the χ^2 divergence (cf. Section 3.1.3).

When the function f is general (not set to be the indicator function), we will look at the problem of bounding the difference of the expectations of f with respect to the two measures μ and ν (cf. Section 3.2). Once again, duality will play an essential role in the results provided and will yield interesting applications in transportation-cost inequalities as well (cf. Section 3.2.1). Some immediate applications of these results in Hypothesis Testing frameworks are then considered in Section 3.3; The problems considered in Chapter 3 are of particular interest when μ is a joint measure and ν is the corresponding product of the marginals, as we will see in the subsequent chapters.

- They can be used in a Supervised Learning Theory framework (cf. Chapter 4) to provide bounds on:
 - the expected value of the generalisation error (cf. Section 4.2) via a variety of information measures;
 - the probability of having a large generalisation error (cf. Section 4.3) via a variety of information measures.

In particular, the problem of bounding the probability of having a large generalisation error can be seen as the problem of providing concentration bounds for a function of random variables when the function is **not** independent of the random variables;

- they can be used in a Bayesian Estimation framework (cf. Chapter 5) to provide lower bounds on the *risk* in estimation procedures that hold for general classes of estimators and that involve a variety of information measures.

Related Work

Following the same structure of the previous section, we will now provide the main references that are connected to (or inspired) this Thesis. The main inspiration for the employment of the duality between function spaces and measure spaces was drawn from the pedagogical proof of Kantorovich-Rubenstein's duality formula in (Villani, 2003, Chapter 1). Other significant sources of inspiration were also the book on Large Deviation by Varadhan (in particular, Section 10), the book on Measure-Theoretic probability (Pollard, 2001) and the monograph by Raginsky and Sason for the connection drawn among the Kullback-Leibler Divergence, log-Sobolev inequalities and the concentration of measure phenomenon (as well as Section 3.4 therein, where the link between Transportation-Cost inequalities and concentration is highlighted). Other meaningful resources were (Raginsky, 2016; Rassoul-Agha and Seppäläinen, 2015; Sason and Verdú, 2016). A document which is similar in spirit to this one is the thesis defended by Liu. Moreover,

- as for the variational representations presented in Chapter 2 and used in this document:
 - the variational representation for the Kullback-Leibler divergence dates back to (Varadhan, 1984, Section 10);
 - the variational representation for Rényi's divergences has appeared multiple times

- in the literature (Atar et al., 2013; Anantharam, 2017; Birrell et al., 2021);
- the variational representation for φ -Divergences in explicit closed-form has appeared in (Broniatowski and Keziou, 2010). Other meaningful references are (Nguyen et al., 2008; Ruderman et al., 2012; Sason, 2018a);
 - bounds of a similar flavour to the ones proposed in Chapter 3 have appeared in the literature in the past. For instance, the bounds regarding the Kullback-Leibler Divergence date back to (Arutjunjan, 1968). They have also often re-appeared due to the Data-Processing Inequality (Polyanskiy et al., 2010) or proven independently using the convexity of the Divergence (Bassily et al., 2018). For the bounds regarding Rényi-Divergences, the one instance we could find is in (Polyanskiy and Verdú, 2010). The problem of bounding the difference of expectations of the same functions with two different measures arises in several contexts: analysis of Bias in Models (Russo and Zou, 2016; Gourgoulis et al., 2020), uncertainty in Markov processes (Birrell and Rey-Bellet, 2020), generalisation error (cf. the next bullet-point), etc.
 - the connection between the expected generalisation error in learning and information measures started in (Russo and Zou, 2016) and evolved in (Xu and Raginsky, 2017c; Asadi et al., 2018; Bassily et al., 2018; Pensia et al., 2018; Bu et al., 2020; Steinke and Zakyntinou, 2020). A link has been drawn with other objects as well, like Wasserstein distances (Lopez and Jog, 2018; Wang et al., 2019; Rodríguez-Gálvez et al., 2021);
 - the connection between the Bayesian Risk in estimation procedures and information measures can be mainly found in the line of work started in (Duchi and Wainwright, 2013; Zhang et al., 2013) and then elaborated on in (Xu and Raginsky, 2017a).

The Theory **Part I**

1 Preliminaries

1.1 Banach Spaces

Definition 1 (L^p spaces, (Berberian, 1974, Def. 39.1)). Let $(\Omega, \mathcal{F}, \mu)$ be a σ -finite measure space, where Ω denotes the underlying space, \mathcal{F} represents the σ -field and μ a σ -finite measure. Given $1 \leq p < \infty$ one can consider the space $L^p(\Omega, \mathcal{F}, \mu)$ of the equivalence class of all functions $f : \Omega \rightarrow \mathbb{R}$ that are measurable with respect to (\mathcal{F}, μ) and such that

$$\int_{\Omega} |f(x)|^p d\mu(x) < \infty. \quad (1.1)$$

Two functions are equivalent if they are equal μ -a.e.

The space of functions $L^p(\Omega, \mathcal{F}, \mu)$ can be shown to be a vector space, i.e., it is closed with respect to addition and scalar multiplication. Moreover, one can endow these spaces with a natural norm that completes them, rendering them complete vector spaces or Banach spaces.

Definition 2 (L^p norm, (Berberian, 1974, Def. 39.1)). Let $1 \leq p < \infty$ and consider the space $L^p(\Omega, \mathcal{F}, \mu)$. Given $f \in L^p(\Omega, \mathcal{F}, \mu)$ one can define the following

$$\|f\|_{L^p(\Omega, \mathcal{F}, \mu)} = \left(\int_{\Omega} |f(x)|^p d\mu(x) \right)^{\frac{1}{p}}. \quad (1.2)$$

Conventionally, one would shorten the notation $\|f\|_{L^p(\Omega, \mathcal{F}, \mu)}$ to $\|f\|_{L^p(\Omega)}$, whenever the measure is clear from the context. However, given that we will mostly keep the space fixed but will constantly change measures, we will shorten $\|f\|_{L^p(\Omega, \mathcal{F}, \mu)}$ to $\|f\|_{L^p(\mu)}$. It is possible to show that given $p \geq 1$ the functional $\|\cdot\|_{L^p(\mu)}$ is actually a norm on $L^p(\Omega, \mathcal{F}, \mu)$. In particular, for every given $p \geq 1$, the L^p -norm completes the space (meaning that, given a Cauchy sequence (f_n) with respect to the metric induced by the L^p -norm in $L^p(\Omega, \mathcal{F}, \mu)$, then $f_n \xrightarrow[n \rightarrow \infty]{} f$ and $f \in L^p(\Omega, \mathcal{F}, \mu)$). More precisely, let us consider the following definition:

Definition 3 (Banach Spaces). A Banach Space is a vector space endowed with a norm, and that is complete with respect to that norm, i.e., complete in the metric induced by the norm.

Then, one can show the following result:

Theorem 1 ((Berberian, 1974, Thm. 39.9)). *Let $1 \leq p < \infty$. The space $L^p(\Omega, \mathcal{F}, \mu)$ endowed with the L^p norm $\|\cdot\|_{L^p(\Omega, \mathcal{F}, \mu)}$ is a Banach Space, i.e. a complete normed vector space.*

1.2 Duality

As informally stated in the Introduction, our primary purpose will be to move from function spaces to measure spaces. In particular, what we will be seeking is a notion of duality between two vector spaces \mathcal{X}, \mathcal{Y} . In the general but classical sense this means that there exists a bilinear function $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$. Once the bilinear function is set (and, in the case of spaces of functions and measures, the bilinear mapping will typically be the Lebesgue integral of the function with respect to the measure), then the topology that one would seek on \mathcal{X} is the so-called weak topology $\sigma(\mathcal{X}, \mathcal{Y})$. I.e., the coarsest topology under which the mappings $\{x \rightarrow \langle x, y \rangle : y \in \mathcal{Y}\}$ are continuous. For the other direction, one would similarly consider the weak topology $\sigma(\mathcal{Y}, \mathcal{X})$ on \mathcal{Y} .

A common assumption is to require that, for each non-zero element $x \in \mathcal{X}$ there exists an element $y \in \mathcal{Y}$ such that $\langle x, y \rangle \neq 0$. Likewise, for each non-zero $y \in \mathcal{Y}$ one asks for the existence of an $x \in \mathcal{X}$ such that $\langle x, y \rangle \neq 0$. When this assumption is satisfied, the topologies $\sigma(\mathcal{X}, \mathcal{Y})$ and $\sigma(\mathcal{Y}, \mathcal{X})$ can be shown to be Hausdorff (Rassoul-Agha and Seppäläinen, 2015, Chapter 4).

Often, we will consider duality in the algebraic (topological) sense: given a vector space \mathcal{X} , the dual of \mathcal{X} , denoted with \mathcal{X}^* will be the space of (continuous/bounded) linear functionals on \mathcal{X} . In this case one has the natural duality pairing $\langle x, x^* \rangle = x^*(x)$ and the topologies $\sigma(\mathcal{X}, \mathcal{X}^*)$ and $\sigma(\mathcal{X}^*, \mathcal{X})$ are generally called, respectively, the weak and weak* topology (Rassoul-Agha and Seppäläinen, 2015, Chapter 4).

Given a pairing between two vector spaces \mathcal{X} and \mathcal{Y} and a norm on \mathcal{X} , one can also define a norm on \mathcal{Y} as follows:

Definition 4. *Given a norm $\|\cdot\| : \mathcal{X} \rightarrow \mathbb{R}^+$, the dual norm $\|\cdot\|_* : \mathcal{Y} \rightarrow \mathbb{R}^+$ can be defined as follows: let $y \in \mathcal{Y}$*

$$\|y\|_* = \sup_{x \in \mathcal{X} : \|x\| \leq 1} |\langle x, y \rangle|. \quad (1.3)$$

If $\mathcal{Y} = \mathcal{X}^*$ (the space of bounded linear functionals defined on \mathcal{X}) and $\langle x, y \rangle$ denotes the natural pairing $y(x)$, then \mathcal{X}^* equipped with $\|\cdot\|_*$ is a Banach space (regardless of whether \mathcal{X} itself is Banach) (Kreyszig, 2007, Section 2.10).

It is easy to see that various spaces of functions are vector spaces and they can be rendered Banach spaces with various norms. Moreover, there is a natural duality between spaces of functions and spaces of measures. Indeed, let $(\mathcal{X}, \mathcal{F})$ be a measurable space. Denote with $\mathcal{M}(\mathcal{X})$ the space of signed (Radon) measures on \mathcal{X} . Given a function $f : \mathcal{X} \rightarrow \mathbb{R}$ in some space

\mathcal{Y} of functions and a measure μ , one can consider the duality pairing:

$$\langle \cdot, \cdot \rangle : (\mathcal{M}(\mathcal{X}), \mathcal{Y}) \longrightarrow \mathbb{R} \quad (1.4)$$

$$(\mu, f) \longrightarrow \int f d\mu. \quad (1.5)$$

In particular, $\mathcal{M}(\mathcal{X})$ can be put in separating duality with: (Rassoul-Agha and Seppäläinen, 2015, Example 4.3)

- $B(\mathcal{X})$, the space of bounded measurable and real-valued functions on \mathcal{X} ;
- $C_b(\mathcal{X})$, the space of *continuous* and bounded real-valued functions on \mathcal{X} , if \mathcal{X} is metric.

These considerations can also be explored further considering topological vector spaces and the corresponding topological duals. For instance, one can leverage Riesz-Kakutani-Markov-like theorems to analyse the topological dual of a variety of spaces of functions (Dunford and Schwartz, 1988, Table IV.A). In general, the topological duals of vector spaces of functions (satisfying a variety of properties) are spaces of measure (again, satisfying a variety of properties).

Example 1. Assume that \mathcal{X} is a normal topological space. Consider the space $B(\mathcal{X})$ of bounded real-valued (or complex-valued) functions of \mathcal{X} with the sup-norm, i.e., given a function $f \in B(\mathcal{X})$ then

$$\|f\| = \sup_{x \in \mathcal{X}} |f(x)|. \quad (1.6)$$

There exists an isometric isomorphism between $B(\mathcal{X})^*$ and $ba(\mathcal{X})$ the space of bounded and finitely additive signed measures defined on \mathcal{X} when endowed with the total variation norm (Dunford and Schwartz, 1988, Theorem IV.5.1). Hence, given an element $h^* \in B(\mathcal{X})^*$, there exists a measure $\mu \in ba(\mathcal{X})$ such that

$$h^*(f) = \int f d\mu, \text{ with } f \in B(\mathcal{X}). \quad (1.7)$$

Example 2. When \mathcal{X} is a locally compact Hausdorff space, one can then consider more restricted sets of functions like $C_c(\mathcal{X})$ or $C_0(\mathcal{X})$ respectively, continuous functions with compact support and continuous functions vanishing at infinity. In this case there is an (isometric) isomorphism between the topological dual of said space and the space of signed Radon Measures defined on \mathcal{X} (Folland, 2013, Chapter 7).

Notice that $C_c(\mathcal{X}) \subset C_0(\mathcal{X}) \subset C_b(\mathcal{X}) \subset B(\mathcal{X})$.

In the spirit of the discussion just above and of Example 1, one can generally see measures as linear functionals acting on spaces of functions. We can thus denote any Lebesgue integral $\int f d\mu$ as the functional μ acting on f i.e., $\mu(f)$. While different from the usual notation, this approach allows us to the use of the common notation \mathbb{E} for the expectation operator. One of the upsides of using $\mu(f)$ rather than $\mathbb{E}_\mu[f]$ is that one does not have to denote constantly (e.g., with a subscript \mathbb{E}_μ) the measure with respect to which one is integrating. Moreover, it relieves us from using different notations when we integrate with respect to probability measures

($\mathbb{E}_\mu[\cdot]$) and other types of measures ($\int \cdot \mu$). We will thus adopt what is known in the literature as “de Finetti notation” (Pollard, 2001, Section 1.4) as the benefits outweigh the discomfort of facing a different-from-usual notation. With a slight abuse we will also denote with $\mu(E)$ the measure under μ of a (measurable) set E . Indeed, given the natural bijection between a set E and the corresponding indicator function $\mathbb{1}_E$, with $\mu(E)$ one can unambiguously denote $\mu(\mathbb{1}_E)$, maintaining a consistent notation that matches (to some extent) the one present in the literature.

Throughout this document we will mostly ignore the topological aspects. Thus, we will only consider the weak topology in both directions (whenever they are needed) unless otherwise specified. Now that we have characterised the spaces of interest, we will formally define one of the tools that allows us to relate functionals defined over one space to functionals defined over the dual: the Legendre-Fenchel transform (Fenchel, 1949). We follow the approach and notation in (Villani, 2003, Section 1.1.6).

Definition 5 (Legendre-Fenchel transform). *Let \mathcal{X} be a normed vector space and $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a functional over \mathcal{X} . The Legendre-Fenchel transform of ψ is the functional ψ^* defined on the topological dual \mathcal{X}^* of \mathcal{X} by the formula:*

$$\psi^*(x^*) = \sup_{x \in \mathcal{X}} \langle x^*, x \rangle - \psi(x), \quad (1.8)$$

where $\langle x^*, x \rangle$ denotes the natural pairing between a space and its topological dual i.e.,

$$\langle x^*, x \rangle = x^*(x).$$

Remark 1. *The definition of Legendre-Fenchel transform can be expressed more generally starting from a pairing between two spaces \mathcal{X} and \mathcal{Y} that puts them in separating duality. Essentially, given $\psi : \mathcal{X} \rightarrow \bar{\mathbb{R}}$ one can construct $\psi^* : \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ like in Equation (1.8) without \mathcal{Y} being necessarily the topological or algebraic dual of \mathcal{X} as long as $(\mathcal{X}, \mathcal{Y})$ are in duality (Rassoul-Agha and Seppäläinen, 2015, Definition 4.9). Most of the properties of Legendre-Fenchel transform (convexity, lower semi-continuity, etc. still hold (Rassoul-Agha and Seppäläinen, 2015, Chapter 4)). One can even consider more general notions of Legendre-Fenchel duality and of convexity itself, focusing on pairings that are not necessarily bilinear (Villani, 2003, Section 2.4.1).*

Given ψ , ψ^* is guaranteed to be convex and lower-semi continuous with respect to the weak* topology and $(\psi^*)^*$ will be the lower-semi continuous convex-hull of ψ (with respect to the corresponding weak topology) if $\psi \neq -\infty$ everywhere. Otherwise $(\psi^*)^* = -\infty$ (Gossez et al., 1976, page 187). The transform has an interesting property when considering convex and lower semi-continuous functions, as described in the following and important result:

Theorem 2 (Fenchel-Moreau Theorem, (Rassoul-Agha and Seppäläinen, 2015, Chapter 4)). *Let \mathcal{X} be a vector space and $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ a functional not identically equal to ∞ . Then, $(\psi^*)^* = \psi$ if and only if ψ is convex and lower semi-continuous.*

Another fundamental result in the area is the following:

Lemma 1 ((Dembo and Zeitouni, 2009, Lemma 4.5.8)). *Let \mathcal{X} be a locally convex Hausdorff topological vector space and let $\psi : \mathcal{X} \rightarrow (-\infty, \infty]$ be a convex lower semi-continuous function. Let*

$$\phi(\lambda) = \sup_{x \in \mathcal{X}} \langle \lambda, x \rangle - \psi(x). \quad (1.9)$$

One has that ψ is the Legendre-Fenchel transform of ϕ and thus

$$\psi(x) = \sup_{\lambda \in \mathcal{X}^*} \langle \lambda, x \rangle - \phi(\lambda). \quad (1.10)$$

Remark 2. *One of the purposes of Legendre-Fenchel duality (in this document) is to link a functional acting on a space to a functional acting on the corresponding dual space. Switching the focus slightly to optimisation problems, a similar connection was already established through Lagrange Duality. Indeed, following again the notation in (Villani, 2003, Theorem 1.9) one has that given two convex functions acting on a normed vector space E , under suitable assumptions the following holds:*

$$\inf_E (\Theta + \Xi) = \max_{z^* \in E^*} (-\Theta^*(-z^*) - \Xi^*(z^*)), \quad (1.11)$$

where E^ is the topological dual of E and Θ^*, Ξ^* represent the Legendre-Fenchel dual of, respectively, Θ and Ξ . The result is a consequence of the Hahn-Banach Theorem.*

One can see this result (typically known as Fenchel-Rockafellar or simply Fenchel Duality) as a generalisation of Lagrange Duality. Indeed, a specific instance of Equation (1.11) can be found in statements like (Luenberger, 1997, Theorem 8.6.1), connecting constrained minimisation problems to maximisation problems over a dual space. One way of obtaining (Luenberger, 1997, Theorem 8.6.1) from Equation (1.11) comes from incorporating the constraints (characterising a convex set C) into the objective function, and thus selecting Ξ to be the indicator function of the set of constraints C (a function which is 0 on C and $+\infty$ elsewhere). One can thus see Lagrange Duality (as well as statements on linear programming) as a consequence of Fenchel Duality with an appropriate choice of functions (Boş et al., 2006). This characterises a bit more explicitly the underlying geometry which might be hard to grasp from results as general as Equation (1.11).

1.3 Orlicz Spaces

The most commonly studied Banach function spaces are the L^p spaces with $p > 1$. However, it is possible to consider even more general function spaces. The family that we will often consider in this document is the family of Orlicz spaces. Orlicz spaces are defined starting from a family of convex functions known as Young functions (or, sometimes, Orlicz functions).

Definition 6 (Young Function, (M.M. Rao, 1991, Chapter 3)). *Let $\psi : \mathbb{R} \rightarrow \mathbb{R}^+$ be a function that satisfies the following list of properties:*

- *it is convex;*
- *it is even, i.e., $\psi(x) = \psi(-x)$;*
- *$\psi(0) = 0$;*

- $\psi(x) \xrightarrow{x \rightarrow \infty} \infty$,

then ψ is called a Young or Orlicz function.

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and denote with $L^0(\mu)$ the space of all the \mathcal{F} -measurable and real-valued functions on Ω . Given a Young/Orlicz function ψ , one can define the following functional

$$I_\psi : L^0(\mu) \rightarrow [0, +\infty] \quad (1.12)$$

$$f \rightarrow \int_{\Omega} \psi(|f|) d\mu. \quad (1.13)$$

Definition 7 ((M.M. Rao, 1991, Definition 5)). *Let ψ be a Young function. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and consider the functional I_ψ as defined in Equation (1.13). An Orlicz space can be defined to be the following set:*

$$L_\psi(\mu) = \{f \in L^0(\mu) : I_\psi(\lambda f) < +\infty \text{ for some } \lambda > 0\}. \quad (1.14)$$

Given a Young function ψ , one can define the corresponding complementary function to ψ in the sense of Young (M.M. Rao, 1991, Section 1.3):

Definition 8. *Given a Young function $\psi : [0, +\infty) \rightarrow \overline{\mathbb{R}}^+$, the complementary function to ψ in the sense of Young, denoted as $\psi_Y^* : \mathbb{R} \rightarrow \overline{\mathbb{R}}^+$ can be defined as follows:*

$$\psi_Y^*(x) = \sup \{\lambda|x| - \psi(\lambda) : \lambda \geq 0\}. \quad (1.15)$$

The complementary function ψ_Y^* is also a Young function and satisfies, by definition, the so-called Young's inequality:

$$xy \leq \psi(x) + \psi_Y^*(y), \quad x, y \in \mathbb{R}. \quad (1.16)$$

Example 3. *Let $\psi(x) = \frac{|x|^p}{p}$ with $p \geq 1$, ψ is a Young function and $\psi_Y^*(y) = \frac{|y|^q}{q}$ with $\frac{1}{p} + \frac{1}{q} = 1$.*

An Orlicz space can be endowed with several norms that render it a Banach Space (Hudzik and Maligranda, 2000): the Luxemburg norm

$$\|f\|_{L_\psi^L(\mu)} = \inf \left\{ \sigma > 0 : I_\psi(f/\sigma) \leq 1 \right\} = \inf \left\{ \sigma > 0 : \mu \left(\psi \left(\frac{|f|}{\sigma} \right) \right) \leq 1 \right\}, \quad (1.17)$$

the Orlicz norm

$$\|f\|_{L_\psi^O(\mu)} = \sup \left\{ \left| \int f g d\mu \right| : g \in L_{\psi_Y^*}(\mu), I_{\psi_Y^*}(g) \leq 1 \right\}, \quad (1.18)$$

and the Amemiya norm

$$\|f\|_{L_\psi^A(\mu)} = \inf_{t>0} \frac{I_\psi(tf) + 1}{t} = \inf_{t>0} \frac{\mu(\psi(t|f|)) + 1}{t}. \quad (1.19)$$

It can also be shown that the Amemiya norm is equivalent to the Orlicz norm in the following sense:

Theorem 3 ((Hudzik and Maligranda, 2000, Theorem 1)). *Let ψ be a Young function and $f \in L_\psi(\mu)$ then*

$$\|f\|_{L_\psi^A(\mu)} = \|f\|_{L_\psi^O(\mu)}. \quad (1.20)$$

We will thus mostly ignore Orlicz norms henceforth.

Orlicz spaces are quite general spaces of functions. Indeed:

Example 4. *Let $\psi(x) = |x|^p$ and let $(\Omega, \mathcal{F}, \mu)$ be a measure space. ψ is a Young function, $L_\psi(\mu)$ is an Orlicz space, $\|\cdot\|_{L_\psi(\mu)}$ is the L^p -norm and $L_\psi(\mu) = L^p(\mu)$.*

They also allow us to nicely characterise spaces of random variables according to the behaviour of their “tail”:

Example 5. *Let $\psi(x) = e^{x^2} - 1$ then, given a measure space $(\Omega, \mathcal{F}, \mu)$ one can define on (Ω, \mathcal{F}) the Orlicz space $L_\psi(\mu)$. Given a random variable X defined on this space then $\|X\|_{\psi, \mu}$ is the so-called sub-Gaussian norm i.e.,*

$$\|X\|_{L_\psi(\mu)} = \inf \left\{ \sigma > 0 : \mu \left(\exp \left(\frac{X^2}{\sigma^2} \right) \right) \leq 2 \right\}. \quad (1.21)$$

Thus, $L_\psi(\mu)$ represents the set of all sub-gaussian random variables with respect to μ (Vershynin, 2018, Definition 2.5.6).

Given a functional and the corresponding Young’s dual, one can prove the so-called generalised Hölder’s inequality.

Lemma 2. *Let ψ be an Orlicz function and ψ_Y^* denote its complementary function. For every couple of random variable U, V respectively defined over $(\Omega_U, \mathcal{F}_U, \mathcal{P}_U), (\Omega_V, \mathcal{F}_V, \mathcal{P}_V)$, given a coupling $\mathcal{P}_{UV}(UV)$ one has that:*

$$\mathcal{P}_{UV}(UV) \leq \|U\|_{L_\psi(\mathcal{P}_U)} \|V\|_{L_{\psi_Y^*}(\mathcal{P}_V)}. \quad (1.22)$$

Given that we could not find a proof of this result in the literature we will now provide one:

Proof. For every $\sigma, t > 0$ we have that:

$$\mathcal{P}(UV) = \mathcal{P}_{UV} \left(\sigma \frac{U}{\sigma} \frac{1}{t} V t \right) \quad (1.23)$$

$$\stackrel{(a)}{\leq} \frac{\sigma}{t} \mathcal{P}_{UV} \left(\psi \left(\frac{|U|}{\sigma} \right) + \psi_Y^*(|V|t) \right) \quad (1.24)$$

Where (a) follows from Young's inequality. Choosing $\sigma = \|U\|_{L_{\psi}^L(\mathcal{P}_U)}$ one retrieves

$$\mathcal{P}_{UV}(UV) \leq \frac{\|U\|_{L_{\psi}^L(\mathcal{P}_U)}}{t} \mathcal{P}_{UV} \left(\psi \left(\frac{|U|}{\|U\|_{L_{\psi}^L(\mathcal{P}_U)}} \right) + \psi_Y^*(|V|t) \right) \quad (1.25)$$

$$\stackrel{(b)}{\leq} \|U\|_{L_{\psi}^L(\mathcal{P}_U)} \frac{1 + \mathcal{P}_V(\psi_Y^*(|V|t))}{t}, \quad (1.26)$$

where (b) follows from the definition of Luxemburg norm, i.e., $\mathcal{P}_U \left(\psi \left(|U| / \|U\|_{L_{\psi}^L(\mathcal{P}_U)} \right) \right) \leq 1$. Taking the infimum with respect to t in Equation (1.26) gives us Equation (1.22) by definition of Amemiya norm. \square

Remark 3. With $\psi_1(t) = \frac{|t|^p}{p}$ as well as with $\psi_2(t) = |t|^p$, leading respectively to $\psi_{1Y}^*(t^*) = \frac{|t^*|^q}{q}$ and $\psi_{2Y}^*(t^*) = |t^*|^q \frac{(p-1)}{p^q}$ with $\frac{1}{p} + \frac{1}{q} = 1$, one recovers the classical Hölder's inequality:

$$\mathcal{P}_{UV}(UV) \leq \|U\|_{L^p(\mathcal{P}_U)} \cdot \|V\|_{L^q(\mathcal{P}_V)}. \quad (1.27)$$

Remark 4. An alternative approach to a generalised Hölder's inequality comes from Definition 4. Indeed, given a pairing $\langle x, y \rangle$, by definition of Dual norm one has that

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|_{\star}. \quad (1.28)$$

Indeed, for any $z \in \mathcal{X}$ such that $\|z\| \leq 1$

$$|\langle z, y \rangle| \leq \sup_{w \in \mathcal{X}: \|w\| \leq 1} |\langle w, y \rangle| = \|y\|_{\star}. \quad (1.29)$$

Selecting $z = \frac{x}{\|x\|}$ then leads to Equation (1.28). It is well-known that the dual norm with respect to the $L^p(\mu)$ -norm is the $L^q(\mu)$ -norm with $\frac{1}{p} + \frac{1}{q} = 1$, when the pairing is selected to be $\langle f, g \rangle = \int f \cdot g d\mu$, with $f \in L^p(\mu)$ and $g \in L^q(\mu)$ (Dunford and Schwartz, 1988, Theorem IV.8.1), thus recovering Hölder's inequality. More details on this are given in Section 1.3.1.

Similarly, fix an Orlicz function ψ and select the same pairing. Consider the ψ_Y^* -Luxemburg norm according to Equation (1.17). The ψ -Orlicz norm, by definition (cf. Equation (1.18)), can be seen as the dual norm (according to Definition 4) of the ψ_Y^* -Luxemburg norm. Indeed, let $g \in L_{\psi_Y^*}(\mu)$ such that $I_{\psi_Y^*}(g) \leq 1$ then, since $I_{\psi_Y^*}(g) = \mu(\psi_Y^*(|g|)) \leq 1$, it follows from Equation (1.17) that $\|f\|_{L_{\psi}^L(\mu)} \leq 1$. With some additional technical steps one can thus show that:

$$\|f\|_{L_{\psi}^L(\mu)} = \sup \left\{ \left| \int f g d\mu \right| : g \in L_{\psi_Y^*}(\mu), I_{\psi_Y^*}(g) \leq 1 \right\} \quad (1.30)$$

$$= \sup \left\{ \left| \int f g d\mu \right| : g \in L_{\psi_Y^*}(\mu), \|g\|_{L_{\psi_Y^*}^L(\mu)} \leq 1 \right\}. \quad (1.31)$$

Then, by Theorem 3, one has that the ψ -Amemiya norm is equivalent to the ψ -Orlicz norm and thus, represents the dual norm of the ψ_Y^* -Luxemburg norm in the sense of Definition 4. Lemma 2

would then follow from the reasoning made at the beginning of this remark i.e., from Equation (1.28).

1.3.1 Hölder's inequality Duality

A central tool throughout this document will also be the classical Hölder's¹ inequality. Given the theme of this thesis, it is only fair to dedicate a section to this result and show how it is connected to the rest of this document (other than from a technical standpoint).

Theorem 4. *Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space and let $\alpha, \beta \in [1, +\infty]$ with $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Let $f : \mathcal{X} \rightarrow \mathbb{R}$ and $g : \mathcal{X} \rightarrow \mathbb{R}$ be measurable functions, then*

$$\mu(|fg|) \leq \|f\|_{L^\alpha(\mu)} \|g\|_{L^\beta(\mu)}. \quad (1.32)$$

One interpretation of Hölder's inequality comes from the notion of dual norms as defined in Definition 4. In particular, given a measure μ , two spaces \mathcal{X}, \mathcal{Y} and a norm on \mathcal{X} then the functional

$$\langle \cdot, \cdot \rangle : (\mathcal{X} \times \mathcal{Y}) \rightarrow \mathbb{R} \quad (1.33)$$

$$(f, g) \rightarrow \mu(fg) \quad (1.34)$$

represents a bilinear mapping, i.e., a duality pairing. Once the duality pairing is set then, from the definition of dual norm (cf. Definition 4), one has that:

$$\mu(fg) \leq \|f\| \cdot \|g\|_*. \quad (1.35)$$

Setting $\mathcal{X} = L^\alpha(\mu)$ with $1 \leq \alpha < +\infty$ then one has that \mathcal{Y} can be set to be \mathcal{X}^* (the algebraic dual of \mathcal{X}) and it is possible to show that $\mathcal{Y} = \mathcal{X}^* = L^\beta(\mu)$ with $\beta = \frac{\alpha}{\alpha-1}$ (cf. (Dunford and Schwartz, 1988, Theorem IV.8.1)). The corresponding dual norm, in the sense of Definition 4, is the $L^\beta(\mu)$ -norm. This immediately provides that, given f, g then:

$$\mu(|fg|) \leq \|f\|_{L^\alpha(\mu)} \cdot \|g\|_{L^\beta(\mu)}. \quad (1.36)$$

These observations show a first connection between Hölder's inequality and "duality" when the duality pairing is fixed to be $\mu(fg)$ and the dual space is defined to be the algebraic (or topological) dual in the classical functional analysis sense.

Another approach that allows us to establish a more direct link between Hölder's inequality and (Legendre-Fenchel) duality, comes from Orlicz spaces and the corresponding norms. Indeed, we have seen that given $\psi_\alpha(x) = \frac{|x|^\alpha}{\alpha}$ one has that ψ_α is an Orlicz function and that $L_{\psi_\alpha}(\mu) = L^\alpha(\mu)$ is an Orlicz space. Moreover, the Luxemburg norm $L_{\psi_\alpha}^L(\mu)$ corresponds to the classical $L^\alpha(\mu)$ -norm. Given ψ_α , the corresponding Young's complementary function is given

¹The result is usually attributed to Hölder (Hölder, 1889). It has been, however, discovered one year before by Rogers. An historical treatise of the inequality is proposed in (Maligranda, 1998), where Maligranda advances the name Hölder-Roger's inequality.

by $\psi_{\alpha_Y}^*(x) = \psi_\beta(x) = \frac{|x|^\beta}{\beta}$ with $\beta = \frac{\alpha}{\alpha-1}$. The function ψ_β is also an Orlicz function and thus implicitly defines a function space L_{ψ_β} . Hence, one can see Hölder's inequality as a Corollary of Lemma 2 with the choice of functions just described (cf. Remark 3). It is, however, instructive to go through the steps. Please notice that in this case the Legendre-Fenchel transform of ψ_α i.e., ψ_α^* , matches with the Young's complementary function, $\psi_{\alpha_Y}^*$. Let us assume that both f and g are positive-valued functions. One has that, given the bilinearity of the duality pairing, for every $t, \sigma > 0$:

$$\mu(fg) = \frac{\sigma}{t} \mu\left(\frac{f}{\sigma} \cdot (tg)\right) \quad (1.37)$$

$$\leq \frac{\sigma}{t} \mu\left(\psi_\alpha\left(\frac{f}{\sigma}\right) + \psi_\alpha^*(tg)\right). \quad (1.38)$$

Then, setting $\sigma = \|f\|_{L_{\psi_\alpha}^L(\mu)}$ leads us to the following:

$$\mu(fg) \leq \|f\|_{L_{\psi_\alpha}^L(\mu)} \inf_{t>0} \frac{1 + \psi_\beta(tg)}{t} \quad (1.39)$$

$$= \|f\|_{L_{\psi_\alpha}^L(\mu)} \|g\|_{L_{\psi_\beta}^A(\mu)}. \quad (1.40)$$

Then one can see that Equation (1.32) follows from Equation (1.40) either by going through the computations and finding the two infima involved in the Luxemburg and Amemiya norms or by following the considerations in Remark 3 along with Theorem 3. Indeed, given the equality between the Amemiya Norm and the Orlicz Norm, one can see the definition of the Orlicz Norm as an instance of Definition 4 and, thus, see the Amemiya norm as the dual norm with respect to the Luxemburg norm induced by ψ_α . Given that $\|\cdot\|_{L_\alpha^L(\mu)}$ is actually equal to $\|\cdot\|_{L^\alpha(\mu)}$ (cf. (M.M. Rao, 1991, Page 5)), then it is known that the dual norm is the $L^\beta(\mu)$ with $\beta = \frac{\alpha}{\alpha-1}$ and the result follows from a contextualisation of Definition 4. What is worth mentioning in particular is the step depicted in Equation (1.38): it explicitly shows the role of duality in proving such an inequality. Surprisingly, Hölder's inequality can also be proved with an approach that mimics the one presented so far, although the equivalence between the approaches is not entirely clear. In particular, starting again from Legendre-Fenchel transforms, one can prove the following:

$$\mu(fg) \leq \inf_{\lambda>0} \frac{\lambda^\alpha}{\alpha} \mu(f^\alpha) + \frac{\lambda^{-\beta}}{\beta} \mu(g^\beta). \quad (1.41)$$

It is easy to see that Equation (1.41) is actually a convex function in λ , the infimum is attained at $\hat{\lambda} = \left(\frac{\mu(f^\alpha)}{\mu(g^\beta)}\right)^{\frac{1}{\alpha\beta}}$ and leads to Equation (1.32). Although slightly different from Equation (1.38), even in the formulation in Equation (1.41) the pattern remains similar: one starts from a duality-pairing, uses then Young's inequality (stemming from Legendre-Fenchel transforms or Young's complementary functions) and applies an additional minimisation step (reminiscent of the Minkowski Functional/Gauge Function (V.I. Bogachev, 2017, Definition 1.4.4)) to achieve norm-like objects.

1.4 Divergences

Ever since the birth of Information Theory, information measures and divergences have played a pivotal role in providing the fundamental results of the field. They represent an intuitive and powerful tool to manipulate, compare and understand probability measures. From the very beginning, Shannon described in 1948 the fundamental limits of lossless compression through the notion of “Entropy” of a source and the fundamental limits of communication through the “Mutual Information”. Throughout the years, more and more information measures have been defined, discovered and exploited to provide results of a similar flavour:

- The Kullback-Leibler divergence (also implicitly used in Shannon’s work) (Shannon, 1948; Kullback and Leibler, 1951);
- Rényi’s Entropy, used by Campbell to provide yet another fundamental result in lossless compression although with a different characterisation (Rényi, 1960; Campbell, 1965);
- Rényi’s Divergences (Rényi, 1960);
- φ -Divergences (Morimoto, 1963; Csiszár, 1963; Ali and Silvey, 1966; Csiszár, 1967, 1972a);
- etc.

Each of these objects has found applications in a variety of fields, ranging from Hypothesis Testing and Statistics (Shannon, 1951; Chernoff, 1952; Ben-Bassat and Raviv, 1978; Liese and Vajda, 2006; Zhang et al., 2013; van Erven and Harremoës, 2014; Xu and Raginsky, 2017a,b; Tomamichel and Hayashi, 2018; Lapidath and Pfister, 2018; Sason and Verdú, 2018; Esposito and Gastpar, 2021; Esposito et al., 2021b), Coding (Shannon, 1948; Campbell, 1965; Csiszar, 1995; Polyanskiy and Verdú, 2010), Concentration of Measure (Varadhan, 1984; Raginsky and Sason, 2014), Guessing (Arikan, 1996; Sason, 2018b; Graczyk and Sason, 2021) and so on and so forth.

1.4.1 Kullback-Leibler Divergence

Arguably the largest building-block of Information Theory, the Kullback-Leibler Divergence can be defined as follows:

Definition 9. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and μ, ν two probability measures in $\mathcal{P}(\mathcal{X})$. The Kullback-Leibler divergence between ν and μ is defined as follows

$$D(\nu \parallel \mu) = \begin{cases} \mu \left(\frac{d\nu}{d\mu} \log \left(\frac{d\nu}{d\mu} \right) \right) = \nu \left(\log \left(\frac{d\nu}{d\mu} \right) \right) & \text{if } \nu \ll \mu \\ +\infty & \text{else.} \end{cases} \quad (1.42)$$

Remark 5. The Kullback-Leibler divergence is finite if and only if $\nu \ll \mu$ **and** $f \log(f) \in L^1(\mu)$, with $f = \frac{d\nu}{d\mu}$ (Varadhan, 1984, Lemma 10.1). If the absolute continuity constraint does not hold, then it is set to be $+\infty$. However, it can also be equal to $+\infty$ if the probability measures are absolutely continuous with respect to each other e.g., if ν is the Cauchy measure and μ is the Gaussian measure, then $\nu \ll \mu$ but $D(\nu \parallel \mu) = +\infty$.

Whenever one has that ν is a joint probability measure between two random variables X, Y (\mathcal{P}_{XY}) and μ represents the corresponding product of the marginals ($\mathcal{P}_X \mathcal{P}_Y$) it goes by the name of “Mutual Information” (between X and Y):

$$I(X; Y) = D(\mathcal{P}_{XY} \parallel \mathcal{P}_X \mathcal{P}_Y). \quad (1.43)$$

The Kullback-Leibler divergence/Mutual Information is strongly connected to the concentration of measure phenomenon (Varadhan, 1984; Dembo and Zeitouni, 2009; Raginsky and Sason, 2014), hypothesis testing problems (Cover and Thomas, 2006), and is routinely applied when characterising the communication capabilities of a stochastic channel (Shannon, 1948), the generalisation error of a learning algorithm (Xu and Raginsky, 2017c; Bassily et al., 2018), lower-bounding the risk in a Bayesian Setting (Xu and Raginsky, 2017a) and so on and so forth.

1.4.2 φ -Divergences

A straightforward generalisation of the Kullback Leibler-Divergence can be obtained by considering a generic convex function $\varphi : \mathbb{R}^+ \rightarrow \mathbb{R}$, usually with the additional constraint that $\varphi(1) = 0$.

Definition 10. *Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probability spaces. Let $\varphi : \mathbb{R}^+ \rightarrow \mathbb{R}$ be a convex function such that $\varphi(1) = 0$. Consider a measure μ such that $\mathcal{P} \ll \mu$ and $\mathcal{Q} \ll \mu$. Denoting with p, q the densities of the measures with respect to μ , the φ -Divergence of \mathcal{P} from \mathcal{Q} is defined as follows:*

$$D_\varphi(\mathcal{P} \parallel \mathcal{Q}) = \int q \varphi\left(\frac{p}{q}\right) d\mu. \quad (1.44)$$

Despite the fact that the definition uses μ and the densities with respect to this measure, it is possible to show that φ -divergences are actually independent of the dominating measure (Liese and Vajda, 2006). Indeed, when absolute continuity between \mathcal{P}, \mathcal{Q} holds, i.e. $\mathcal{P} \ll \mathcal{Q}$ we retrieve the following (Liese and Vajda, 2006, Equation (26)):

$$D_\varphi(\mathcal{P} \parallel \mathcal{Q}) = \mathcal{Q}\left(\varphi\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right) = \int \varphi\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q}. \quad (1.45)$$

Denoting with \mathcal{F}_X the Sigma-field generated from the random variable X the φ -Mutual Information, a generalisation of Shannon’s Mutual Information, can be defined as follows:

Definition 11. *Let X and Y be two random variables jointly distributed according to \mathcal{P}_{XY} over the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{XY})$. Let $(\mathcal{X}, \mathcal{F}_X, \mathcal{P}_X), (\mathcal{Y}, \mathcal{F}_Y, \mathcal{P}_Y)$ be the corresponding probability spaces induced by the marginals. Let $\varphi : \mathbb{R}^+ \rightarrow \mathbb{R}$ be a convex function such that $\varphi(1) = 0$. The φ -Mutual Information between X and Y is defined as:*

$$I_\varphi(X, Y) = D_\varphi(\mathcal{P}_{XY} \parallel \mathcal{P}_X \mathcal{P}_Y). \quad (1.46)$$

If $\mathcal{P}_{XY} \ll \mathcal{P}_X \mathcal{P}_Y$, one has that:

$$I_\varphi(X, Y) = \mathcal{P}_X \mathcal{P}_Y \left(\varphi \left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right) \right) = \int \varphi \left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right) d\mathcal{P}_X \mathcal{P}_Y. \quad (1.47)$$

It is possible to see that if φ satisfies $\varphi(1) = 0$ and it is strictly convex at 1, then $I_\varphi(X, Y) = 0$ if and only if X and Y are independent (Liese and Vajda, 2006, Equation (34)). This generalization includes a large family of divergences, including:

- Kullback-Leibler divergence, with $\varphi(t) = t \log(t)$;
- Total Variation distance, with $\varphi(t) = \frac{1}{2}|t - 1|$;
- Hellinger distance, with $\varphi(t) = (\sqrt{t} - 1)^2$;
- Pearson χ^2 -divergence, with $\varphi(t) = (t - 1)^2$;
- E_γ -divergence, with $\varphi(t) = \max\{0, t - \gamma\}, \gamma \geq 1$.

1.4.3 Rényi's α -Divergences and Sibson's α -Mutual Information

Introduced by Rényi as a generalisation of Shannon Entropy and the Kullback Leibler-divergence (cf. (Rényi, 1960)), Rényi's α -entropy and divergence have found many applications ranging from hypothesis testing to guessing and several other statistical inference and coding problems (van Erven and Harremoës, 2014; Verdú, 2015). Indeed, it has several operational interpretations (e.g., in hypothesis testing, and as the cut-off rate in block coding (Csiszar, 1995; van Erven and Harremoës, 2014)). It can be defined as follows (van Erven and Harremoës, 2014):

Definition 12. Let $(\Omega, \mathcal{F}, \mathcal{P}), (\Omega, \mathcal{F}, \mathcal{Q})$ be two probability spaces. Let $\alpha > 0$ be a positive real number different from 1. Consider a measure μ such that $\mathcal{P} \ll \mu$ and $\mathcal{Q} \ll \mu$ and denote with p, q the densities of \mathcal{P}, \mathcal{Q} with respect to μ . The α -Divergence of \mathcal{P} from \mathcal{Q} is defined as follows:

$$D_\alpha(\mathcal{P} \parallel \mathcal{Q}) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} d\mu. \quad (1.48)$$

Remark 6. The definition is independent of the chosen measure μ . It is indeed possible to show that $\int p^\alpha q^{1-\alpha} d\mu = \int \left(\frac{q}{p}\right)^{1-\alpha} d\mathcal{P}$, and that whenever $\mathcal{P} \ll \mathcal{Q}$ or $0 < \alpha < 1$, we have $\int p^\alpha q^{1-\alpha} d\mu = \int \left(\frac{p}{q}\right)^\alpha d\mathcal{Q}$ (van Erven and Harremoës, 2014).

It can be proved that if $\alpha > 1$ and $\mathcal{P} \not\ll \mathcal{Q}$ then $D_\alpha(\mathcal{P} \parallel \mathcal{Q}) = \infty$. The behaviour of the measure for $\alpha \in \{0, 1, \infty\}$ can be defined by continuity. In general, one has that $D_1(\mathcal{P} \parallel \mathcal{Q}) = D(\mathcal{P} \parallel \mathcal{Q})$ but if $D(\mathcal{P} \parallel \mathcal{Q}) = \infty$ or there exists $\beta > 1$ such that $D_\beta(\mathcal{P} \parallel \mathcal{Q}) < \infty$ then $\lim_{\alpha \downarrow 1} D_\alpha(\mathcal{P} \parallel \mathcal{Q}) = D(\mathcal{P} \parallel \mathcal{Q})$ (van Erven and Harremoës, 2014, Theorem 5). It is important to notice that Rényi's α -Divergences do not belong to the family of φ -Divergences, however they are connected

them through a one-to-one mapping linking them to Hellinger Integrals ²:

$$D_\alpha(\mathcal{P}\|\mathcal{Q}) = \frac{1}{\alpha-1} \log(H_\alpha(\mathcal{P}\|\mathcal{Q})). \quad (1.49)$$

For an extensive treatment of Rényi's α -divergences and their properties we refer the reader to (van Erven and Harremoës, 2014). Starting from Rényi's Divergence and the geometric averaging that it involves, Sibson built the notion of Information Radius (Sibson, 1969):

Definition 13. Let (μ_1, \dots, μ_n) be a family of probability measures and (w_1, \dots, w_n) be a set of weights s.t. $w_i \geq 0$ for $i = 1, \dots, n$ and such that $\sum_{i=1}^n w_i > 0$. Let $\alpha \geq 1$, the information radius of order α is defined as:

$$\min_{\nu \ll \sum_i w_i \mu_i} \frac{1}{\alpha-1} \log \left(\sum_i w_i \exp((\alpha-1)D_\alpha(\mu_i\|\nu)) \right).$$

Suppose now that we have two random variables X, Y jointly distributed according to \mathcal{P}_{XY} . It is possible to generalise Definition 13 and see that the information radius is a special case of the following quantity (Verdú, 2015):

$$I_\alpha(X, Y) = \min_{\mathcal{Q}_Y} D_\alpha(\mathcal{P}_{XY}\|\mathcal{P}_X\mathcal{Q}_Y). \quad (1.50)$$

$I_\alpha(X, Y)$ represents a generalisation of Shannon's Mutual Information and possesses many interesting properties (Verdú, 2015). Indeed, $\lim_{\alpha \rightarrow 1} I_\alpha(X, Y) = I(X; Y)$. On the other hand when $\alpha \rightarrow \infty$, we get:

$$I_\infty(X, Y) = \log \mathcal{P}_Y \left(\sup_{x: \mathcal{P}_X(\{x\}) > 0} \frac{\mathcal{P}_{XY}(\{x, Y\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{Y\})} \right) = \mathcal{L}(X \rightarrow Y),$$

where $\mathcal{L}(X \rightarrow Y)$ denotes the Maximal Leakage from X to Y , a recently defined information measure with an operational meaning in the context of privacy and security (Issa et al., 2020). More details on Maximal Leakage will be provided in a subsequent section. To conclude, let us list some of the properties of I_α :

Proposition 1 (Verdú, 2015, Theorem 4).

1. **Data Processing Inequality:** given $\alpha > 0$, $I_\alpha(X, Z) \leq \min\{I_\alpha(X, Y), I_\alpha(Y, Z)\}$ if the Markov Chain $X - Y - Z$ holds;
2. $I_\alpha(X, Y) \geq 0$ with equality iff X and Y are independent;
3. Let $\alpha_1 \leq \alpha_2$ then $I_{\alpha_1}(X, Y) \leq I_{\alpha_2}(X, Y)$;
4. Let $\alpha \in (0, 1) \cup (1, \infty)$, for a given \mathcal{P}_X , $\frac{1}{\alpha-1} \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X, Y)\right)$ is convex in $\mathcal{P}_{Y|X}$;
5. $I_\alpha(X, Y) \leq \min\{\log|X|, \log|Y|\}$;

²Please notice that with Hellinger integral between two absolutely continuous measure $\nu \ll \mu$ we denote $\int \left(\frac{d\nu}{d\mu}\right)^\alpha d\mu$ (cf. Definition 19). Another similar family of divergences which are linked to Rényi's Divergences are the Hellinger Divergences (cf. (Sason and Verdú, 2016, page 5976, item 5)) which are slightly different from what we consider in this document.

For a more extensive treatment of Sibson's α -MI we refer the reader to (Verdú, 2015).

Maximal Leakage

A dependence measure of particular interest to us, belonging to the family of Sibson's Mutual Information, is the Maximal Leakage, denoted by $\mathcal{L}(X \rightarrow Y)$. It was introduced as a way of measuring the leakage of information from X to Y and defined as follows:

Definition 14 ((Issa et al., 2020, Definition 1)). *Given a joint distribution \mathcal{P}_{XY} on finite alphabets \mathcal{X} and \mathcal{Y} , the maximal leakage from X to Y is defined as:*

$$\mathcal{L}(X \rightarrow Y) = \sup_{U-X-Y-\hat{U}} \log \frac{\mathcal{P}_{U\hat{U}}(\{U = \hat{U}\})}{\max_{u \in \mathcal{U}} \mathcal{P}_U(\{u\})}, \quad (1.51)$$

where U and \hat{U} take values in the same finite, but arbitrary, alphabet.

It is shown in (Issa et al., 2020, Theorem 1) that, for finite alphabets:

$$\mathcal{L}(X \rightarrow Y) = \log \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}: \mathcal{P}_X(x) > 0} P_{Y|X}(y|x). \quad (1.52)$$

If X and Y have a jointly continuous pdf $f(x, y)$, one gets (Issa et al., 2020, Corollary 4):

$$\mathcal{L}(X \rightarrow Y) = \log \int_{\mathbb{R}} \sup_{x: f_X(x) > 0} f_{Y|X}(y|x) dy. \quad (1.53)$$

One can also show that $\mathcal{L}(X \rightarrow Y) = I_\infty(X; Y)$, *i.e.*, Maximal Leakage corresponds to the Sibson's Mutual Information of order infinity. This allows the measure to retain the properties listed in Proposition Proposition 1. A conditional version of this information measure has also been presented in (Issa et al., 2020):

Definition 15 (Conditional Maximal Leakage). *Given a joint distribution P_{XYZ} on alphabets \mathcal{X}, \mathcal{Y} , and \mathcal{Z} , define:*

$$\mathcal{L}(X \rightarrow Y|Z) = \sup_{U: U-X-Y|Z} \log \frac{\mathbb{P}(\{U = \hat{U}(Y, Z)\})}{\mathbb{P}(\{U = \tilde{U}(Z)\})}, \quad (1.54)$$

where U takes value in an arbitrary finite alphabet and we consider \hat{U}, \tilde{U} to be the optimal estimators of U given (Y, Z) and Z , respectively.

It is shown in (Issa et al., 2020) that for discrete random variables X, Y, Z :

$$\mathcal{L}(X \rightarrow Y|Z) = \log \max_{z: P_Z(z) > 0} \sum_y \max_{x: P_{X|Z}(x|z) > 0} P_{Y|XZ}(y|xz),$$

and that

$$\mathcal{L}(X \rightarrow (Y, Z)) \leq \mathcal{L}(X \rightarrow Y) + \mathcal{L}(X \rightarrow Z|Y). \quad (1.55)$$

2 Duality and Divergences

2.1 Introduction

Throughout this document we will look at information measures as functionals acting over the space of signed (or probability) measures. One can look at divergences (Kullback-Leibler, Rényi or φ -Divergences) from various angles. In this document, we will generally fix the second measure and look at them as functionals acting on the first measure. *I.e.*, take any convex function φ and any probability measure μ , one can consider the φ -Divergence as a functional

$$\begin{aligned}\psi_\mu^\varphi(\cdot) : \mathcal{M}(\mathcal{X}) &\longrightarrow \bar{\mathbb{R}} \\ v &\longrightarrow \mu\left(\varphi\left(\frac{dv}{d\mu}\right)\right).\end{aligned}$$

Once this perspective is undertaken, it is clear that one can try and understand whether ψ_μ^φ has a variational representation on a space of functions

$$\psi_\mu^\varphi(v) = D_\varphi(v\|\mu) = \sup_f v(f) - (\psi_\mu^\varphi)^\star(f).$$

Under suitable additional assumptions, one can then use Lemma 1 and see, for instance, $(\psi_\mu^\varphi)^\star$ as the Legendre-Fenchel dual of D_φ and thus say that for $f \in \mathcal{M}(\mathcal{X})^\star$ (for some space of measures $\mathcal{M}(\mathcal{X})$):

$$(\psi_\mu^\varphi)^\star(f) = \sup_{v \in \mathcal{M}(\mathcal{X})} v(f) - D_\varphi(v\|\mu).$$

A similar route can be undertaken whenever one restricts themselves to measures that are absolutely continuous with respect to each other. In particular, fix a measure μ and consider the corresponding space $\mathcal{M}(\mathcal{X}, \mu) = \{v \in \mathcal{M}(\mathcal{X}) : v \ll \mu\}$. One can then identify (the quotient space of) $\mathcal{M}(\mathcal{X}, \mu)$ with $L^1(\mu)$ (the space of measurable functions with $L^1(\mu)$ -norm) via the Radon-Nikodym theorem. This is because given μ and $f \in L^1(\mu)$ one can always construct a new measure ν through the relationship $d\nu = fd\mu$ and vice versa. Thus, given μ , to each such function f corresponds a measure ν_f and to each measure $\nu \ll \mu$ a function $f_\nu \in L^1(\mu)$ can be

Chapter 2. Duality and Divergences

associated. One can then see divergences, with the second measure fixed to be μ , as (convex) functionals acting on $L^1(\mu)$, *i.e.*,

$$D_\varphi(v\|\mu) = \begin{cases} \mu\left(\varphi\left(\frac{dv}{d\mu}\right)\right) & \text{whenever } v \ll \mu \\ +\infty & \text{else.} \end{cases}$$

In this case one can show that, for a given convex functional φ , $\mu \in \mathcal{M}(\mathcal{X})$ and $v \in \mathcal{M}(\mathcal{X}, \mu)$, denoting with $h = \frac{dv}{d\mu} \in L^1(\mu)$,

$$\psi_\mu^\varphi(h) = \sup_{f \in L^\infty(\mu)} v(f) - (\psi_\mu^\varphi)^*(f),$$

and

$$(\psi_\mu^\varphi)^*(f) = \sup_{g \in L^1(\mu)} \mu(fg) - \psi_\mu^\varphi(g).$$

These relationships hold (cf. (Ruderman et al., 2012)), as it is well-known that the dual space of $L^1(\mu)$ is equal to $L^\infty(\mu)$. The dual of $L^\infty(\mu)$ however, can be much larger than $\mathcal{M}(\mathcal{X}, \mu)$ and involve finitely additive measures (Dunford and Schwartz, 1988, Theorem IV.8.16).

More generally, one could try and find the largest dual pair of spaces $(\mathcal{A}, \mathcal{B})$ such that \mathcal{B} includes a class of functions of interest or, alternatively, such that \mathcal{A} includes a class of measures of interest and then characterise a divergence acting on \mathcal{A} through a variational representation acting on \mathcal{B} and vice versa (cf. (Broniatowski and Keziou, 2010, Section 2)).

2.1.1 Kullback-Leibler Divergence

As stated in Section 1.4.1, the classical approach would be to see the Kullback-Leibler Divergence as a mapping

$$D: P(\mathcal{X}) \times P(\mathcal{X}) \rightarrow \overline{\mathbb{R}}^+ \\ (v, \mu) \rightarrow v\left(\log\left(\frac{dv}{d\mu}\right)\right)$$

i.e., as a mapping that takes as an input two (probability) measures and provides as an output a positive real number (possibly $+\infty$). This number should somehow quantify how “distant” the two measures are. Following the framework described in Section 2.1 we will fix the second measure μ and see the Kullback-Leibler divergence as a (convex) functional of the first measure. *I.e.*, assume that we have a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ defined on a measurable space $(\mathcal{X}, \mathcal{F})$, one can see the Kullback-Leibler divergence as follows:

$$D: \mathcal{M}(\mathcal{X}) \rightarrow \overline{\mathbb{R}}^+ \tag{2.1}$$

$$v \rightarrow \mu\left(\frac{dv}{d\mu} \log \frac{dv}{d\mu}\right), \tag{2.2}$$

where one assumes (in congruence with the literature) that given ν, μ , $D(\nu\|\mu) = +\infty$ if $\nu \not\ll \mu$ or if $\nu \in \mathcal{M}(\mathcal{X}) \setminus \mathcal{M}_1(\mathcal{X})$. We can now leverage the underlying duality principle between measures and functionals. In particular, we will consider the well-known variational representation for the Kullback-Leibler Divergence, known as the Donsker-Varadhan Representation, originally stated¹ in (Varadhan, 1984, Section 10):

Theorem 5 ((Rassoul-Agha and Seppäläinen, 2015, Theorem 5.6)). *Let \mathcal{X} be a metric space and let $\mu \in \mathcal{M}_1(\mathcal{X})$. For a given $\nu \in \mathcal{M}(\mathcal{X})$ one has that*

$$D(\nu\|\mu) = \sup_{f \in C_b(\mathcal{X})} \nu(f) - \log \int_{\mathcal{X}} e^f d\mu = \sup_{f \in C_b(\mathcal{X})} \nu(f) - \log \mu(\exp(f)). \quad (2.3)$$

More generally, one can also state the following:

Theorem 6 ((Rassoul-Agha and Seppäläinen, 2015, Theorem 5.4)). *Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and let $\mu \in \mathcal{M}_1(\mathcal{X})$. For a given $\nu \in \mathcal{M}(\mathcal{X})$ one has that*

$$D(\nu\|\mu) = \sup_{f \in B(\mathcal{X})} \nu(f) - \log \mu(\exp(f)), \quad (2.4)$$

where $B(\mathcal{X})$ denotes the space of bounded measurable functions $f : \mathcal{X} \rightarrow \mathbb{R}$.

One thus has that the Kullback-Leibler Divergence $D(\cdot\|\mu)$ is the Legendre-Fenchel dual of the log-moment generating function and in particular, one can also have that given $f \in B(\mathcal{X})$

$$\log(\mu(\exp(f))) = \sup_{\nu \in B(\mathcal{X})^*} \nu(f) - D(\nu\|\mu), \quad (2.5)$$

with the assumption that $D(\nu\|\mu) = +\infty$ if $\nu \notin \mathcal{M}_1(\mathcal{X})$ (Dembo and Zeitouni, 2009, Lemma 6.2.13). These considerations, although seemingly trivial, represent the tools that allow us to jump from divergences to functions and connect these quantities via inequalities. As an example, let us immediately recover a well-known result from the literature, starting from Equation (2.3). Let $f = \alpha \mathbb{1}_E$ with $\alpha > 0$, for some measurable set E . Given any probability

¹The variational formulation given in (Varadhan, 1984, Equation 10.1) takes an equivalent but different and interesting perspective: it sees the Kullback-Leibler Divergence $D(\nu\|\mu)$ as the smallest constant defined as follows:

$$D(\nu\|\mu) = \inf \{c : \nu(f) \leq c + \log(\mu(\exp(f))), f \in B(\mathcal{X})\}.$$

Hence, the smallest functional $\psi_\mu^*(\nu)$, $\nu \in \mathcal{P}(\mathcal{X})$ such that $\nu(f) = \langle \nu, f \rangle = \mu\left(f \frac{d\nu}{d\mu}\right)$ satisfies Young's inequality:

$$\nu(f) = \langle \nu, f \rangle \leq \psi_\mu(f) + \psi_\mu^*(\nu), \text{ with } \psi_\mu(f) = \log(\mu(\exp(f))).$$

measure ν (such that $\nu \ll \mu$ and that E is measurable with respect to ν) we have that:

$$D(\nu \parallel \mu) \geq \nu(f) - \log(\mu(\exp(f))) \quad (2.6)$$

$$= \alpha \nu(E) - \log(\mu(\exp(\alpha \mathbb{1}_E))) \quad (2.7)$$

$$= \alpha \nu(E) - \log(e^\alpha \mu(E) + 1 - \mu(E)) \quad (2.8)$$

$$= \nu(E) \log(1/\mu(E)) - \log(2 - \mu(E)), \quad (2.9)$$

where the last steps follows from setting $\alpha = \log(1/\mu(E))$. Re-arranging we recover the following result:

$$\nu(E) \leq \frac{D(\nu \parallel \mu) + \log(2 - \mu(E))}{\log(1/\mu(E))} \leq \frac{D(\nu \parallel \mu) + 1}{\log(1/\mu(E))}. \quad (2.10)$$

Remark 7. Clearly, one can try to optimise the inequality with respect to α , providing the best bound one can obtain through this family of f 's and this technique, on $\nu(E)$. The corresponding minimisation problem is, however, not trivial. One could also naturally consider general linear transformations of the form $f = \lambda \mathbb{1}_E + \beta$ with $\lambda > 0$. However, it is easy to see that this does not provide any extra leverage in this specific setting. Indeed, the corresponding inequality would be

$$D(\nu \parallel \mu) \geq \lambda \nu(E) + \beta - \log(e^{(\lambda+\beta)} \mu(E) + (1 - \mu(E)) e^\beta) \quad (2.11)$$

$$= \lambda \nu(E) + \beta - \beta - \log(e^\lambda \mu(E) + (1 - \mu(E))). \quad (2.12)$$

This simple Fano-like inequality allows us to immediately relate probabilities of an event measured through different measures and the divergence between these measures. Instances of this have appeared repeatedly in the literature and have been proven in a variety of ways (e.g., (Nishiyama and Sason, 2020, Section 3.4)). They found application in learning theory (Bassily et al., 2018), hypothesis testing and (“weak”) converse bounds in coding settings (Polyanskiy et al., 2010). Moreover, it captures the essence of this document and allows us to connect through an inequality a divergence (in this case, the Kullback-Leibler divergence, ψ_μ) the expectation of a function (here, $\nu(\lambda \mathbb{1}_E)$) and the Legendre-Fenchel dual of the divergence evaluated at said function ($\psi_\mu^*(\lambda \mathbb{1}_E) = \log(\mu(\exp(\lambda \mathbb{1}_E)))$). In this particular case one can characterise $\psi_\mu^*(\lambda \mathbb{1}_E)$ completely. In other settings that we will encounter, such an explicit characterisation will not be accessible and an upper-bound on the ψ_μ^* will be assumed. Throughout the exposition we will use these very same tools with different functionals of measures and re-derive some fundamental results of information theory as well as derive new results.

2.1.2 Shannon's Entropy and Shannon's Coding Theorem

The approach mentioned in the previous section is powerful and quite general. Moreover, it does not simply apply to divergences (with a measure fixed) but also to other convex functionals of measures, e.g., entropies. Let us consider, for instance, Shannon's Entropy. Let us step aside from the standard notation here as well, in order to highlight the fact that we

consider entropies as functionals acting over measures rather than functionals over random variables. Moreover, in this and the following section we will always consider the existence of a dominating measure ξ^2 .

Definition 16. Let \mathcal{P} be a probability measure defined over a measurable space $(\mathcal{X}, \mathcal{F})$. Assume that there exists a measure ξ such that $\mathcal{P} \ll \xi$ and denote the corresponding density with p . The Shannon Entropy of \mathcal{P} (in bits) is given by³:

$$H(\mathcal{P}) = - \int p \log_2(p) d\xi. \quad (2.13)$$

Given the Shannon Entropy, one can compute its “dual”. For simplicity let us restrict ourselves to discrete probability measures (and, thus, select ξ to be the counting measure which, in turn, implies that p is the probability mass function associated to \mathcal{P}). One has that:

Lemma 3. Let H be the Shannon Entropy as defined in Definition 16 and let \mathcal{P} be a probability measure that is absolutely continuous with respect to the counting measure. One has that:

$$H(\mathcal{P}) = \inf_{f \in B(\mathcal{X})} \log_2 \sum_x 2^{f(x)} - \mathcal{P}(f). \quad (2.14)$$

Proof. There are multiple ways to recover Equation (2.14). If one is dealing with discrete objects and assuming that $|\mathcal{X}| = n$, one possibility is to go through the fact that $H(\mathcal{P}) = \log(n) - D(\mathcal{P} \parallel \mathcal{U})$, where \mathcal{U} is the uniform measure over a set of n points. Using then Equation (2.3) it is possible to recover Equation (2.14). Indeed, given a discrete probability measure \mathcal{P} :

$$-H(\mathcal{P}) + \log n = D(\mathcal{P} \parallel \mathcal{U}) \quad (2.15)$$

$$= \sup_f \mathcal{P}(f) - \log \mathcal{U}(\exp(f)) \quad (2.16)$$

$$= \sup_f \mathcal{P}(f) - \log \left(\frac{1}{n} \sum_x \exp(f(x)) \right). \quad (2.17)$$

Thus, one has that

$$H(\mathcal{P}) = - \sup_f \mathcal{P}(f) - \log \left(\frac{1}{n} \sum_x \exp(f(x)) \right) - \log n \quad (2.18)$$

$$= \inf_f \log \left(\sum_x \exp(f(x)) \right) - \mathcal{P}(f). \quad (2.19)$$

Yet another technique that makes use of the variational characterisation of the Kullback-Leibler Divergence follows from interpreting the Shannon entropy $H(\mathcal{P})$ as $-D(\mathcal{P} \parallel \xi)$, where

²Defining Shannon's and Rényi's entropy for general spaces is not a trivial task and often requires elaborated definitions leveraging partitions or the notion of “information dimension” (Rényi, 1959; Wu and Verdú, 2010; Śmieja and Tabor, 2014).

³In order to be precise, one should bring the dominating measure ξ in the notation used for the Shannon entropy, since the quantity depends on ξ . However, given that the dominating measure will always be clear from the context or stated explicitly, we will adopt a simpler notation that resembles the original notation more.

Chapter 2. Duality and Divergences

ξ is the dominating measure (cf. Definition 16). This approach clearly works well with both discrete and continuous objects, but not necessarily with more general measures. \square

Remark 8. *The infimal representation in Equation (2.19) can be seen as the “concave conjugate” of Shannon’s Entropy. Indeed, given a function ψ , the concave conjugate of ψ , denoted with ψ_\star , can be defined as follows (Rockafellar, 1970, Page 308):*

$$\psi_\star(f) = \inf_{\mu} \mu(f) - \psi(\mu). \quad (2.20)$$

Moreover, if ψ is concave one also has that

$$\psi(\mu) = \inf_f \mu(f) - \psi_\star(f). \quad (2.21)$$

One can also easily see that (Rockafellar, 1970, Page 308)

$$\psi_\star(f) = -(-\psi)^\star(-f). \quad (2.22)$$

In the case of Shannon entropy one has that setting $\psi(\mu) = H(\mu)$ then $\psi_\star(f) = H_\star(f) = -\log \sum_x 2^{-f(x)}$ which in turn implies that $(-\psi)^\star(-f) = (-H)^\star(-f) = \log \sum_x 2^{-f(x)}$. Hence,

$$-H(\mu) = \sup_f \mu(f) - (-H)^\star(f) \quad (2.23)$$

$$= \sup_f \mu(f) - \log \sum_x 2^{f(x)} \quad (2.24)$$

which then leads us to the following variational representation:

$$H(\mu) = \inf_f \log \sum_x 2^{f(x)} - \mu(f). \quad (2.25)$$

Substituting f with $-f$ in Equation (2.25) one obtains

$$H(\mu) = \inf_f \mu(f) - (-\log \sum_x 2^{-f(x)}) = \inf_f \mu(f) - H_\star(f). \quad (2.26)$$

In particular, let us assume that \mathcal{P} is defined over the finite set \mathcal{X} . Simply using Equation (2.14) one can easily recover Shannon’s converse of the source coding theorem. Let us set the stage for such a result. Denote with \mathcal{D} the target alphabet. Our purpose is to “compress” the realisations of a random variable taking values in \mathcal{X} (and whose probability measure is \mathcal{P}) using only symbols from \mathcal{D} (with the assumption that $|\mathcal{D}| < |\mathcal{X}|$ hence, “compressing” the realisations of the random variable).

Definition 17. *A source code is a mapping $C : \mathcal{X} \rightarrow \mathcal{D}^{K_\star}$ from the source alphabet \mathcal{X} to the Kleene-star operator over \mathcal{D} , \mathcal{D}^{K_\star} i.e., the set of all finite-length strings that can be generated by concatenating arbitrary elements of \mathcal{D} , allowing the use of the same element multiple times as well as the empty string.*

Given $x \in \mathcal{X}$, $C(x)$ represents the code-word corresponding to x and we will denote with $l(x) = |C(x)|$ the length of $C(x)$. Moreover, for a given probability measure \mathcal{P} we can define the expected length of a code C as $\mathcal{P}(l) = \sum_{x \in \mathcal{X}} p(x)l(x)$, with p denoting the point mass function associated to \mathcal{P} . Let $|\mathcal{D}| = 2$. Shannon's source coding theorem states the following:

Theorem 7. *Let $C : \mathcal{X} \rightarrow \mathcal{D}^{K^*}$ be a binary code (i.e., such that $|\mathcal{D}| = 2$). Let $l : \mathcal{X} \rightarrow \mathbb{N}$ be the function representing the length of the code, i.e., $l(x) = |C(x)|$.*

If the code satisfies Kraft's Inequality, i.e.,

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1, \quad (2.27)$$

then

$$\mathcal{P}(l) \geq H(\mathcal{P}). \quad (2.28)$$

Clearly, Shannon's source-coding theorem is a statement connecting a concave functional $H(\mathcal{P})$ to the expected value of a function l under \mathcal{P} with certain assumptions on said function (that relate to desirable properties of the code). Consequently, following our discussion from the previous section, one should be able to retrieve Equation (2.28) from Equation (2.14) and, indeed, that is the case.

Proof. Setting then $f = -l$ in Lemma 3 and re-arranging, one retrieves

$$\sum_x l(x)p(x) \geq H(\mathcal{P}) - \log_2 \sum_x 2^{-l(x)}. \quad (2.29)$$

Using now the assumption that $\sum_x 2^{-l(x)} \leq 1$ one has that $-\log_2 \sum_x 2^{-l(x)} \geq 0$ and the statement of the theorem follows:

$$\mathcal{P}(l) \geq H(\mathcal{P}) - \log_2 \sum_x 2^{-l(x)} \geq H(\mathcal{P}). \quad (2.30)$$

□

Remark 9. *A simple pattern to be extrapolated here comes from noticing that most of these results are in the following form: let ψ be a convex (concave) functional over measures and let ψ^* denote its convex (concave) dual.*

If

$$\psi^*(f) \stackrel{(\geq)}{\leq} k$$

then

$$\mu(f) \stackrel{(\geq)}{\leq} \psi(\mu) - k.$$

In the case just above, abstracting from technical details, $\psi(\mu) = H(\mu)$ represents the Shannon entropy (a concave functional), and bounding its dual $\psi^(l)$ (i.e., asking for the code to be uniquely-decodable), one can lower-bound $\mu(l)$ using $H(\mu)$. We will see this pattern appear over and over in this document.*

Remark 10. *An intermediate approach comes from Lagrange Duality. This approach is under-*

taken, for instance, in (Cover and Thomas, 2006, Thm 5.3.1). Indeed, one can see the Shannon's Entropy of the probability mass function p as the solution to the problem of minimising the expected length of the code (under p , i.e. $\sum_x p(x)l(x)$) subject to having $\sum_x 2^{-l(x)} \leq 1$, cf. Remark 2.

The proof is clearly not simpler than the standard approaches one can find in introductory books on Information Theory. Moreover, it requires relatively advanced tools like Legendre-Fenchel transforms and Duality. However, it is instructive as it shows that even basic and fundamental results like Shannon's source coding theorem (connecting the expected value of a function and a concave functional) do indeed follow from Duality. Once again, the purpose of this work is to show that most of the results connecting a convex functional of measures (Divergences, Entropies, etc.) to expectations of functions, follow from some form of duality.

2.1.3 Beyond unique-decodability

With an approach identical to Theorem 7, one can also retrieve known statements about one-to-one (not necessarily uniquely decodable) codes. The codes considered usually assign a distinct binary code-word to each outcome of the random variable X without necessarily constraining the concatenations of the code-words to be uniquely decodable (Leung-Yan-Cheong and Cover, 1978; Dunham, 1980). In this case, one typically has that if \mathcal{X} is finite and of size n then one can construct a bijection between \mathcal{X} and the set $\{1, \dots, n\}$. The optimal code, assuming (without loss of generality) that the probabilities $\mathcal{P}(X = i)$ are non-increasing, would be such that $l(i) = \lfloor \log(i) \rfloor$ (Alon and Orlicsky, 1994). In this case, using Lemma 3, then one can prove the result in (Dunham, 1980) i.e.,

$$\mathcal{P}(l) \geq H(\mathcal{P}) - \log \log(n + 1). \quad (2.31)$$

Indeed, one has that:

$$\sum_{i=1}^n 2^{-l(i)} = \sum_{i=1}^n 2^{-\lfloor \log(i) \rfloor} \leq \log(n + 1). \quad (2.32)$$

Plugging $-l$ and Equation (2.32) in Theorem 10 one recovers Equation (2.31). With a more convoluted argument it is also possible to consider codes defined on a domain in bijection with \mathbb{N} . In particular, consider codes such that the corresponding lengths satisfy the following:

$$l(i) = \lfloor \log(i + 1) \rfloor. \quad (2.33)$$

These codes are optimal one-to-one mappings (Leung-Yan-Cheong and Cover, 1978). This choice of lengths is similar to the previous one with the sole exception that here we do not assign the code-word of length 0 to $i = 1$. In both cases one has that, for every $k \in \mathbb{N}$ there are 2^k code-words of length k . Thus,

$$\sum_i 2^{-l(i)} = \sum_{k \in \mathbb{N}} 2^k 2^{-k} = +\infty, \quad (2.34)$$

i.e., the Kraft's sum does **not** converge.

It is nonetheless possible to provide a lower-bound in this case considering a transformation of the length of the code. Consider, for instance, a bijection between l and m where $m(i) = a \cdot l(i)$ with $a > 1$. In this case, using Lemma 3, one has that:

$$\mathcal{P}(m) = a\mathcal{P}(l) \geq H(\mathcal{P}) - \log_2 \sum_i 2^{-m(i)} \quad (2.35)$$

$$= H(\mathcal{P}) - \log_2 \sum_{k \in \mathbb{N}} 2^k 2^{-ak} \quad (2.36)$$

$$= H(\mathcal{P}) - \log_2 \sum_{k \in \mathbb{N}} 2^{-(a-1)k} \quad (2.37)$$

$$= H(\mathcal{P}) - \log_2 \left(\frac{1}{2^{(a-1)} - 1} \right). \quad (2.38)$$

Re-writing Equation (2.38), this means that

$$\mathcal{P}(l) \geq \sup_{a>1} \frac{1}{a} \left(H(\mathcal{P}) - \log_2 \left(\frac{1}{2^{(a-1)} - 1} \right) \right). \quad (2.39)$$

Equation (2.39) is essentially a re-interpretation of (Verriest, 1986, Theorem 1 and Remark 1). As a last example, consider the following bijection:

$$n(i) = l(i) + a \log_2(l(i)). \quad (2.40)$$

In this particular case one has

$$\log_2 \sum_i 2^{-n(i)} = \log_2 \sum_i 2^{-l(i) - a \log_2(l(i))} \quad (2.41)$$

$$= \log_2 \sum_{k \in \mathbb{N}} 2^k 2^{-k - a \log_2(k)} \quad (2.42)$$

$$= \log_2 \sum_{k \in \mathbb{N}} 2^{-a \log_2(k)} \quad (2.43)$$

$$= \log_2 \sum_{k \in \mathbb{N}} k^{-a} \quad (2.44)$$

$$= \log_2 \zeta(a), \quad (2.45)$$

where $\zeta(a)$ denotes the Riemann's Zeta Function. Plugging n and Equation (2.45) in Lemma 3, one recovers the following for every $a > 1$:

$$\mathcal{P}(l) \geq H(\mathcal{P}) - a\mathcal{P}(\log_2(l)) - \log_2(\zeta(a)) \quad (2.46)$$

$$\geq H(\mathcal{P}) - a(\log_2(\mathcal{P}(l))) - \log_2(\zeta(a)) \quad (2.47)$$

$$\geq H(\mathcal{P}) - a(\log_2(H(\mathcal{P}) + 1)) - \log_2(\zeta(a)), \quad (2.48)$$

where Equation (2.47) follows from Jensen's inequality and Equation (2.48) follows from the fact that, since l correspond to the length of the best one-to-one code then it is such that $\mathcal{P}(l) \leq H(\mathcal{P}) + 1$ (*i.e.*, it has a lower expected value than the best uniquely-decodable code

which is guaranteed to be smaller than $H(\mathcal{P}) + 1$. In particular, one has the following lower-bound on the expected length

$$\mathcal{P}(l) \geq \sup_{a>1} H(\mathcal{P}) - a(\log_2(H(\mathcal{P}) + 1)) - \log_2(\zeta(a)). \quad (2.49)$$

Selecting $a = \frac{3}{2}$ in Equation (2.49) one has that $\log_2(\zeta(3/2)) < \log_2(e)$ leading us to the following lower-bound

$$\mathcal{P}(l) \geq H(\mathcal{P}) - \frac{3}{2}(\log_2(H(\mathcal{P}) + 1)) - \log_2(e), \quad (2.50)$$

which is close to the Theorem in (Alon and Orlitsky, 1994). Moreover, with the technique described in this section and the right transformation of $l(i)$ one can recover all the lower-bounds presented in (Leung-Yan-Cheong and Cover, 1978, Theorem 3). Notice also that while in (Leung-Yan-Cheong and Cover, 1978) the authors transform their codes so that Kraft's inequality is satisfied in order to provide the lower-bound, with the approach described here this is not necessary: having a converging Kraft's sum is enough to reach a lower-bound on $\mathcal{P}(l)$. Moreover, Lemma 3 does not require the function to necessarily be the length of a code, hence one can include, for instance, a log (like in Equation (2.40)) without having to constrain the function to be integer-valued.

2.2 Rényi's Entropy

The same approach as before can be used to recover another source coding result, this time due to Campbell and involving Rényi's entropy and a different type of expected value.

2.2.1 A variational Representation for Rényi's Entropy

Mimicking the previous section, in order to provide Campbell's Coding Theorem (which is a result connecting Source Coding and Rényi's entropy) it is necessary to provide a variational characterisation for H_α .

Definition 18 (Rényi's Entropy). *Let \mathcal{P} be a probability measure defined over a measurable space $(\mathcal{X}, \mathcal{F})$ and let $\alpha > 0$. Assume that there exists a measure ξ such that $\mathcal{P} \ll \xi$ and denote the corresponding density with p . The Rényi's Entropy (in bits) of \mathcal{P} is given by*

$$H_\alpha(\mathcal{P}) = -\frac{1}{\alpha-1} \log_2 \int p^\alpha d\xi. \quad (2.51)$$

Remark 11. *Clearly this definition of Rényi's Entropy will depend on the dominating measure that one chooses. Selecting ξ to be the counting measure and assuming $\mathcal{P} \ll \xi$ one recovers the usual definition of Rényi's entropy for discrete random variables (Rényi, 1960).*

We will focus, like before, on discrete measures. Hence, we can re-write the Rényi entropy as follows

$$H_\alpha(\mathcal{P}) = -\frac{1}{\alpha-1} \log_2 \sum_{x \in \mathcal{X}} p(x)^\alpha$$

with p the probability mass function associated to \mathcal{P} . Focusing on the part inside the \log_2 we can see the Rényi's entropy of \mathcal{P} as the ℓ_α -norm of the probability mass function p with respect to the counting measure, i.e., $H_\alpha(\mathcal{P}) = -\frac{\alpha}{\alpha-1} \log_2 \|p\|_{\ell^\alpha}$. Leveraging this, one can thus prove the following variational representation for Rényi's entropy.

Lemma 4. *Let p be a probability mass function defined on \mathcal{X} . Let $\alpha > 0$, and denote with $\beta = \frac{\alpha}{\alpha-1}$ then*

$$H_\alpha(\mathcal{P}) = \inf_{g \in L^0(\mathcal{P}), g: \mathcal{X} \rightarrow \mathbb{R}^+} \log \|g\|_{\ell^\beta}^\beta - \beta \log(\mathcal{P}(g)). \quad (2.52)$$

Proof. Given $\alpha > 1$, for every positive function g defined on \mathcal{X} one has that

$$\mathcal{P}(g) = \int g d\mathcal{P} = \int g p d\xi \quad (2.53)$$

$$\stackrel{(c)}{\leq} \|p\|_{L^\alpha(\xi)} \|g\|_{L^\beta(\xi)} \quad (2.54)$$

$$= \|p\|_{\ell^\alpha} \|g\|_{\ell^\beta} \quad (2.55)$$

$$= 2^{\left(\frac{1-\alpha}{\alpha} H_\alpha(\mathcal{P})\right)} \cdot \|g\|_{\ell^\beta}, \quad (2.56)$$

where (c) follows from Hölder's inequality. Equation (2.56) can be re-written as follows

$$\left(\frac{1-\alpha}{\alpha} H_\alpha(\mathcal{P})\right) \geq \log_2 \mathcal{P}(g) - \log_2 \|g\|_{\ell^\beta}. \quad (2.57)$$

Multiplying both sides of Equation (2.57) by $\frac{1-\alpha}{\alpha}$ leads to:

$$H_\alpha(\mathcal{P}) \leq \frac{\alpha}{1-\alpha} \left(\log_2 \mathcal{P}(g) - \log_2 \|g\|_{\ell^\beta} \right). \quad (2.58)$$

Consider now $0 < \alpha < 1$. One can achieve Equation (2.58) starting from Equation (2.53) but using the reverse Hölder's inequality⁴. Given a probability mass (density) function p then $g = p^{\alpha-1}$ achieves equality in Equation (2.52). Indeed,

$$\log \|g\|_{\ell^\beta}^\beta - \beta \log(\mathcal{P}(g)) = \log \sum_x g(x)^\beta - \beta \log \sum_x p(x) g(x) \quad (2.59)$$

$$= \log \sum_x (p(x)^{\alpha-1})^\beta - \beta \log \sum_x p(x) p(x)^{\alpha-1} \quad (2.60)$$

$$= \log \sum_x p(x)^{\beta(\alpha-1)} - \beta \log \sum_x p(x)^\alpha \quad (2.61)$$

$$= (1-\beta) \log \sum_x p(x)^\alpha \quad (2.62)$$

$$= \left(1 - \frac{\alpha}{\alpha-1}\right) \log \sum_x p(x)^\alpha = \frac{1}{1-\alpha} \log \sum_x p(x)^\alpha = H_\alpha(\mathcal{P}). \quad (2.63)$$

□

⁴The objects $\|p\|_{L^\alpha(\xi)}$ and $\|g\|_{L^\beta(\xi)}$ will not denote norms but they simply represent a compact notation for the corresponding functionals $(\int p^\alpha d\xi)^{\frac{1}{\alpha}}$ and $(\int g^\beta d\xi)^{\frac{1}{\beta}}$

Remark 12. Similarly to Shannon's entropy, one can see the Rényi's entropy (cf. Equation (2.51)) as $(-D_\alpha(\mathcal{P} \parallel \xi))$. One can then use the variational representation for D_α to reach Equation (2.52). More details on this will be given in Section 2.2.4.

As the sketch of proof in Section 2.2.4 will show, both variational representations are a consequence of Hölder's inequality which is connected to duality in a more implicit way. A discussion on the connection between Hölder's inequality and duality is given in Section 1.3.1 while a discussion on Hölder's inequality and (some) information measures can be found in Section 3.A. Once the link between these variational representations and (reverse) Hölder's inequality is established, the proofs provided here for both Campbell's Coding Theorem and Arikan's Guessing Theorem are essentially identical to their own proofs (cf. (Campbell, 1965; Arikan, 1996)). The contribution provided in these sections has more of a conceptual nature: the idea is to provide a framework for thinking about Information Measures, Duality and the potential applications of such a framework.

One can now apply Lemma 4 to the setting described in (Campbell, 1965).

2.2.2 Campbell's Coding Theorem

The setting for this Coding Theorem is identical to the one proposed by Shannon with the sole exception that the purpose is not to provide a fundamental lower-bound on $\mathcal{P}(l)$ but rather on $\frac{1}{t} \log \mathcal{P}(2^{tl})$ with $0 < t < \infty$. Under this framework, the price in which we incur when we compress \mathcal{X} using the code C is no longer measured as the expected length $\mathcal{P}(l)$ but through a parametrised "loss" function. The result that Campbell provided was thus the following.

Theorem 8. Let $C : \mathcal{X} \rightarrow \mathcal{D}^{K^*}$ be a binary code (i.e., such that $|\mathcal{D}| = 2$) such that $\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1$, then

$$\frac{1}{t} \log_2 \mathcal{P}(2^{tl}) \geq H_{\frac{1}{t+1}}(\mathcal{P}). \quad (2.64)$$

If $t \rightarrow 0^+$ one essentially recovers Theorem 7. Let us now prove Theorem 8 using Lemma 4.

Proof. Let $g = 2^{tl}$ in Equation (2.52), one has that for every $0 < \alpha < 1$

$$\frac{1}{t} \log_2 \sum_x 2^{tl(x)} p(x) \geq \frac{(1-\alpha)}{t\alpha} H_\alpha(\mathcal{P}) + \frac{1}{t} \log_2 \left(\sum_x 2^{\beta tl(x)} \right)^{\frac{1}{\beta}}. \quad (2.65)$$

Set $\alpha = \frac{1}{1+t}$ (which, in turn, implies that $\beta = -\frac{1}{t}$), one has that

$$\frac{1}{t} \log_2 \sum_x 2^{tl(x)} p(x) \geq \frac{(1-\alpha)}{t\alpha} H_\alpha(\mathcal{P}) + \frac{1}{t} \log_2 \left(\sum_x 2^{-l(x)} \right)^{\frac{1}{\beta}} \quad (2.66)$$

$$= \frac{(1-\alpha)}{t\alpha} H_\alpha(\mathcal{P}) - \log_2 \sum_x 2^{-l(x)} \quad (2.67)$$

$$\geq \frac{(1-\alpha)}{t\alpha} H_\alpha(\mathcal{P}) \quad (2.68)$$

$$= H_{\frac{1}{1+t}}(\mathcal{P}), \quad (2.69)$$

where Equation (2.69) holds since $(1+t) = \frac{1}{\alpha}$ and consequently $\frac{1-\alpha}{t\alpha} = \frac{t}{t} = 1$. \square

Remark 13. The parameter t allows us to interpolate between the average length and the maximum length of our code-words. Indeed, we are trying to lower-bound a norm of 2^l , i.e. the quantity $\log \|2^l\|_{L^t(\mathcal{P})}$.

Shannon's Coding Theorem follows from Theorem 8, as one has the following:

$$\lim_{t \rightarrow 0^+} \frac{1}{t} \log \mathcal{P}(2^{tl}) = \lim_{t \rightarrow 0^+} \log \|2^l\|_{L^t(\mathcal{P})} = \log_2(\exp(\mathcal{P}(\log(2^l)))) = \mathcal{P}(l), \quad (2.70)$$

while, on the right-hand side of Equation (2.64),

$$\lim_{t \rightarrow 0^+} H_{\frac{1}{1+t}}(\mathcal{P}) = H(\mathcal{P}). \quad (2.71)$$

Considering the other extreme:

$$\lim_{t \rightarrow +\infty} \frac{1}{t} \log \mathcal{P}(2^{tl}) = \lim_{t \rightarrow +\infty} \log \|2^l\|_{L^t(\mathcal{P})} = \log_2 \operatorname{ess\,sup}_{\mathcal{P}} 2^l = \max_x l(x), \quad (2.72)$$

assuming that $l(X)$ has full support of with respect to \mathcal{P} and is positive. The larger t is, the more weight we give to the larger code-words. The smaller t is, the more evenly we treat each code-word.

Remark 14. The pattern analysed in Remark 9 appears here as well, although in a slightly altered form. The difference lies in the fact that we are not computing the Legendre-Fenchel dual of $H_\alpha(\cdot)$ but rather interpreting it as “norm” (ℓ_α -norms of the pmf) and computing its “dual norm”. In this description we are abusing the notation as some of these objects are, in fact, **not** norms (if one chooses $0 < \alpha < 1$, which is necessary in order to have $\beta < 0$ and, consequently, be able of lower-bounding $\log(\mathcal{P}(g))$ in Equation (2.52)). A re-writing of the result would be:

$$\mathcal{P}(2^{tl}) \geq \|p\|_{\ell_\alpha} \|2^{tl}\|_{\ell_\beta}, \quad (2.73)$$

with α and β being Hölder's conjugates. Taking the log on both sides of Equation (2.73) one reaches a Legendre-Fenchel-like transformation, where the role of the Legendre-Fenchel dual is played by $\log \|2^{tl}\|_{\ell_\beta}$. Once this is done, the underlying idea is essentially the same: if we can lower-bound the “dual”, i.e. $\log \|2^{tl}\|_{\ell_\beta}$ (again, assuming unique-decodability and with a

proper choice of β in terms of t), then we can lower-bound the expected value of 2^{tl} with the ℓ_α -“norm” of p . This last object is, essentially, a logarithm away from Rényi’s entropy.

2.2.3 Arikan’s Guessing Theorem

Another immediate result that stems from Equation (2.52) is Arikan’s result connecting moments of the number of guesses and Rényi’s entropy. The setting is the following: let X be a discrete random variable taking values in \mathcal{X} and such that $|\mathcal{X}| = M$. The idea is that we want to guess the value of X asking questions of the type “Is X equal to x ?” (Massey, 1994). Let $G(x)$ represent the number of guesses required by a particular guessing strategy when $X = x$. In (Arikan, 1996) the following result is proved:

Theorem 9. *Let $G : \mathcal{X} \rightarrow \{1, \dots, M\}$ be an arbitrary guessing function and let X be distributed according to \mathcal{P} . Given $\rho > 0$, then:*

$$\mathcal{P}(G^\rho) \geq (1 + \log M)^{-\rho} 2^{\left(\rho H_{\frac{1}{1+\rho}}(\mathcal{P})\right)}. \quad (2.74)$$

Proof. This result is a simple application of Lemma 4 with $g = G$ and $\alpha = \frac{1}{1+\rho}$. Indeed one has that, given α , $\frac{1}{\alpha} = 1 + \rho$ and consequently $\frac{1}{\beta} = -\rho$ and $\beta = -\frac{1}{\rho}$. Plugging these quantities in Equation (2.52) one has:

$$\log \sum_x G(x)^\rho p(x) \geq \frac{1-\alpha}{\alpha} H_\alpha(\mathcal{P}) + \log \|G^\rho\|_\beta \quad (2.75)$$

$$= \frac{1-\alpha}{\alpha} H_\alpha(\mathcal{P}) + \frac{1}{\beta} \log \sum_{x=x_1}^{x_M} G(x)^{\beta \cdot \rho} \quad (2.76)$$

$$= \frac{1-\alpha}{\alpha} H_\alpha(\mathcal{P}) + \log \left(\sum_{i=1}^M i^{-1} \right)^{-\rho}. \quad (2.77)$$

Substituting α and applying $2^{(\cdot)}$ on both sides one recovers

$$\sum_x G(x)^\rho p(x) \geq 2^{\left(\rho H_{\frac{1}{1+\rho}}(\mathcal{P})\right)} \cdot \left(\sum_{i=1}^M i^{-1} \right)^{-\rho}. \quad (2.78)$$

The result then follows from noticing that $\sum_{i=1}^M i^{-1} \leq (1 + \log M)$ and the positivity of ρ . \square

Remark 15. *Once again, the idea is: given ρ and $\alpha = \frac{1}{1+\rho}$, if we can lower-bound the dual norm of G^ρ we can lower-bound the expected-value of G^ρ under \mathcal{P} with the Rényi’s entropy of \mathcal{P} . In this case, assuming $\rho > 0$, the lower-bound on the dual norm was:*

$$\left(\sum_{i=1}^M i^{-1} \right)^{-\rho} \geq (1 + \log M)^{-\rho}. \quad (2.79)$$

Remark 16. *One can also utilise the same approach used for Shannon’s Coding Theorem. Hence,*

setting $f = -G$ in Lemma 3 one recovers the following bound:

$$\sum_x G(x) p(x) \geq H(\mathcal{P}) - \log \sum_i 2^{-i} = H(\mathcal{P}) - \log(1 - 2^{-M}). \quad (2.80)$$

Notice that trying to retrieve Shannon's entropy in Equation (2.74) requires us to consider the limit of $\rho \rightarrow 0$ which leads us to the 0-th moment of G : an uninteresting quantity. On the other hand, to provide a lower-bound on $\mathcal{P}(G)$ one is required to consider $\rho = 1$ in Equation (2.74), leading us to the following result:

$$\sum_x G(x) p(x) \geq (1 + \log M)^{-1} 2^{\left(\frac{H_1}{2}(p)\right)}. \quad (2.81)$$

Consider the following setting: $p = (1/2, 1/2)$, then $M = 2$ and $H(p) = H_1(p) = 1$. One has then

$$\sum_x G(x) p(x) \geq \max \left\{ \frac{2}{(1 + \log_2 2)}, 1 - \log_2(1 - 2^{-2}) \right\} \quad (2.82)$$

$$= \max \left\{ 1, 1 - \log_2 \left(\frac{3}{4} \right) \right\} = 1 - \log_2 \left(\frac{3}{4} \right) \approx 1.42. \quad (2.83)$$

Hence, the lower-bound provided by Shannon's Entropy (cf. Equation (2.80)) is larger than the one provided by Equation (2.74). Or, more generally, for any given $\rho \geq 1$ setting $f = -G^\rho$ in Lemma 3 one recovers the following.

$$\sum_x G(x)^\rho p(x) \geq H(\mathcal{P}) - \log \sum_{i=1}^M 2^{-i^\rho}. \quad (2.84)$$

If p is uniform over a set of size M then $H(p) = \log M$. Similarly, one has that:

$$H_{\frac{1}{1+\rho}}(\mathcal{P}) = \frac{1}{1 - \frac{1}{1+\rho}} \log_2 \sum_{i=1}^M \left(\frac{1}{M} \right)^{\frac{1}{1+\rho}} \quad (2.85)$$

$$= \frac{\rho + 1}{\rho} \log_2 M^{1 - \frac{1}{1+\rho}} \quad (2.86)$$

$$= \frac{\rho + 1}{\rho} \log_2 M^{\frac{\rho}{1+\rho}} = \log_2 M. \quad (2.87)$$

Hence,

$$2^{\left(\rho H_{\frac{1}{1+\rho}}(\mathcal{P})\right)} = 2^{(\rho \log_2 M)} = M^\rho. \quad (2.88)$$

The corresponding lower-bound would then be

$$\sum_x G(x)^\rho p(x) \geq \left(\frac{M}{\sum_{i=1}^M i^{-1}} \right)^\rho \geq \left(\frac{M}{1 + \log_2 M} \right)^\rho \quad (2.89)$$

while with Shannon Entropy one retrieves the following

$$\sum_x G(x)^\rho p(x) \geq \log \left(\frac{M}{\sum_{i=1}^M 2^{-i^\rho}} \right). \quad (2.90)$$

Plotting the behaviour of Equation (2.90) and Equation (2.89), one can see that the lower-bound provided by Equation (2.90) is larger than the one provided by Equation (2.89) only if $\rho = 1$ and $M \leq 9$ or if $\rho = 2$ and $M = 2$. With $\rho \geq 3$ then the lower-bound given by Rényi's Entropy is always larger (cf. Figures 2.1a, 2.1b, 2.1c). Considering other probability mass functions p could lead to more interesting results stemming from Lemma 3.

2.2.4 Rényi's Divergence Variational Representation

Using the tools developed so far, one can also provide an alternative proof of Rényi's Variational Representation that stems from Hölder's inequality (and thus, as argued in Appendix 1.3.1, from Legendre-Fenchel Duality). When trying to prove such a characterisation for Rényi's Divergences the literature points to different directions: Birrell et al. as well as Atar et al. use an ad-hoc approach to show the expression, while Anantharam uses a variational representation connecting the Kullback-Leibler and Rényi's Divergence. However, none of them uses explicitly the same type of duality we have considered so far. In order to prove this result via Hölder's inequality one has to look at a seemingly different but strongly related object: Hellinger integrals. Formally,

Definition 19. Let $(\Omega, \mathcal{F}, \mu), (\Omega, \mathcal{F}, \nu)$ be two probability spaces and assume that $\nu \ll \mu$. Let $\alpha > 0$, the Hellinger Integral of order α between ν and μ is defined as follows

$$H_\alpha(\mu \parallel \nu) = \mu \left(\left(\frac{d\nu}{d\mu} \right)^\alpha \right). \quad (2.91)$$

Remark 17. $H_\alpha(\cdot \parallel \mu)$ is clearly a convex functional for $\alpha \geq 1$ while it is concave for $0 < \alpha < 1$. One could say that H_α is a φ -divergence with $\varphi(x) = x^\alpha$, relaxing the assumption that $\varphi(1) = 0$. It is easy to see that the Hellinger Integral of order α is nothing but the $(\alpha$ -th power of the) $L^\alpha(\mu)$ -norm of the Radon-Nikodym derivative between ν and μ , i.e., $H_\alpha(\mu \parallel \nu) = \left\| \frac{d\nu}{d\mu} \right\|_{L^\alpha(\mu)}^\alpha$.

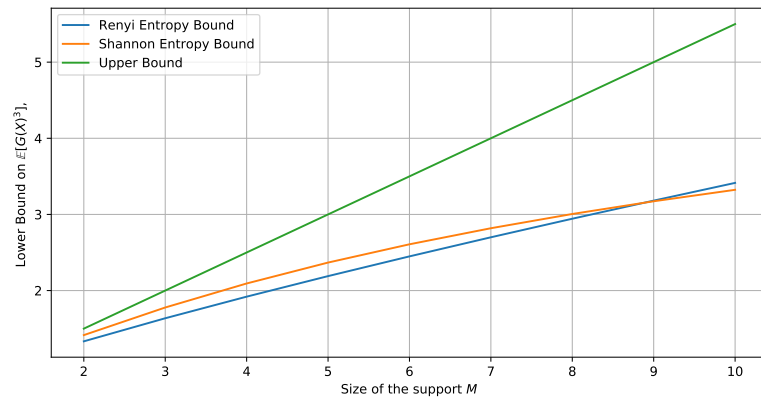
One can show the following Hölder-like inequality for every $f : \mathcal{X} \rightarrow \mathbb{R}^+$ and every $\nu \ll \mu$ (cf. Section 3.A):

$$\int f d\nu \leq (H_\alpha(\nu \parallel \mu))^{\frac{1}{\alpha}} \left(\int f^\beta d\mu \right)^{\frac{1}{\beta}}. \quad (2.92)$$

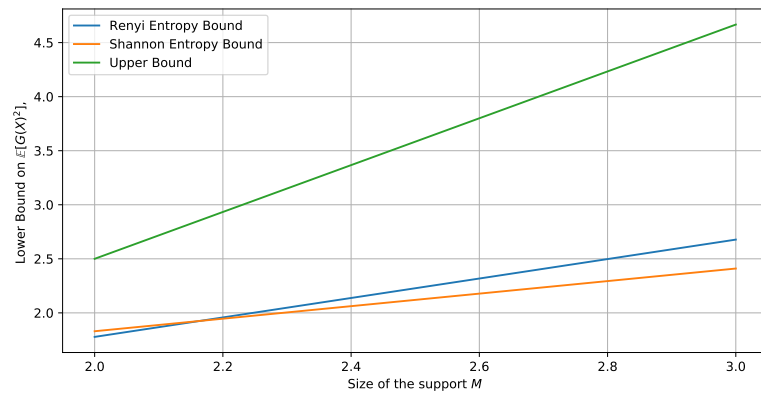
While, if $\alpha < 1$ one retrieves the following:

$$\int f d\nu \geq (H_\alpha(\nu \parallel \mu))^{\frac{1}{\alpha}} \left(\int f^\beta d\mu \right)^{\frac{1}{\beta}}. \quad (2.93)$$

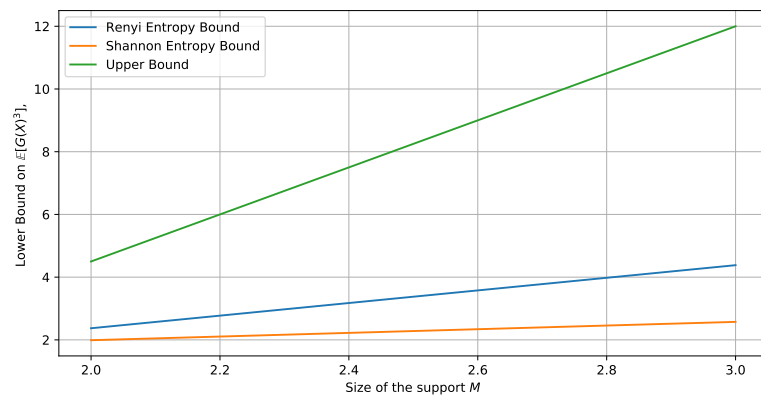
In both equations $\beta = \frac{\alpha}{\alpha-1}$ denotes the Hölder's conjugate of α . Starting from Equation (2.92) and Equation (2.93) one can then retrieve the so-called Rényi's variational representation:



(a) The picture shows the behaviour of Equation (2.89) and Equation (2.90) with $\rho = 1$ and as a function of M .



(b) The picture shows the behaviour of Equation (2.89) and Equation (2.90) with $\rho = 2$ and as a function of M .



(c) The picture shows the behaviour of Equation (2.89) and Equation (2.90) with $\rho = 3$ and as a function of M .

Figure 2.1: Comparisons of the lower-bound on the ρ -th moment of the Guessing strategy G via Equations Equation (2.89) and and Equation (2.90). The values of the sums $\sum_{i=1}^M i^{-1}$ and $\sum_{i=1}^M 2^{-i\rho}$ are computed numerically for each M .

Chapter 2. Duality and Divergences

Lemma 5. *Let $\alpha \in \mathbb{R}^+ \setminus \{0, 1\}$ then one has that:*

$$D_\alpha(v\|\mu) = \sup_{g \in B(\mathcal{X})} \frac{\alpha}{\alpha-1} \log \int e^{(\alpha-1)g} d\nu - \log \left(\int e^{g^\alpha} d\mu \right). \quad (2.94)$$

We will gloss over some technical details in order to highlight the connection between Equation (2.92), Equation (2.93) and Equation (2.94). One can find several technical proofs of Lemma 5 in the literature (Atar et al., 2013; Anantharam, 2017; Birrell et al., 2021).

Proof. Let $g \in B(\mathcal{X})$ be an arbitrary function and let $\alpha > 1$. Setting $f = e^{(\alpha-1)g}$ in Equation (2.92) one recovers the following for every $\nu \ll \mu$

$$\int e^{(\alpha-1)g} d\nu \leq (H_\alpha(v\|\mu))^{\frac{1}{\alpha}} \left(\int e^{(\alpha-1)g\beta} d\mu \right)^{\frac{1}{\beta}}. \quad (2.95)$$

Applying the natural logarithm on both sides of Equation (2.95) and dividing by $\frac{1}{\alpha-1}$ gives us

$$\frac{1}{\alpha-1} \log \int e^{(\alpha-1)g} d\nu \leq \frac{1}{\alpha(\alpha-1)} \log H_\alpha(v\|\mu) + \frac{1}{\beta(\alpha-1)} \log \left(\int e^{(\alpha-1)g\beta} d\mu \right). \quad (2.96)$$

Now, using the fact that $\beta(\alpha-1) = \alpha$ and that $D_\alpha(v\|\mu) = \frac{1}{\alpha-1} \log H_\alpha(v\|\mu)$ one retrieves

$$\frac{1}{\alpha-1} \log \int e^{(\alpha-1)g} d\nu \leq \frac{1}{\alpha} D_\alpha(v\|\mu) + \frac{1}{\alpha} \log \left(\int e^{g^\alpha} d\mu \right), \quad (2.97)$$

or, re-arranging, one obtains that that for every $g \in B(\mathcal{X})$:

$$\frac{1}{\alpha} D_\alpha(v\|\mu) \geq \frac{1}{\alpha-1} \log \int e^{(\alpha-1)g} d\nu - \frac{1}{\alpha} \log \left(\int e^{g^\alpha} d\mu \right). \quad (2.98)$$

For the case of $\alpha < 1$ the proof follows from Equation (2.93) and the same steps as for $\alpha > 1$ with the main difference that dividing on both sides by $\frac{1}{\alpha-1}$ leads to an inversion of the inequality with respect to Equation (2.93) (which allows us to maintain the same variational representation for all values of α). Moreover, if $\nu \ll \mu$ and $\frac{d\nu}{d\mu} \in L^\alpha(\mu)$ then setting $g = \log \left(\frac{d\nu}{d\mu} \right)$ gives us equality in Equation (2.94). \square

Remark 18. *In (Atar et al., 2013; Anantharam, 2017) the variational representation is actually on the space of probability measures, i.e., given a function $g \in B(\mathcal{X})$ then, if $\alpha \in \mathbb{R} \setminus \{0, 1\}$*

$$\frac{1}{\alpha-1} \log \int e^{(\alpha-1)g} d\nu = \inf_{\mu \in \mathcal{M}_1(\mathcal{X})} \frac{1}{\alpha} \log \int e^{\alpha g} d\mu + \frac{1}{\alpha} D_\alpha(v\|\mu), \quad (2.99)$$

as well as

$$\frac{1}{\alpha} \log \int e^{\alpha g} d\mu = \sup_{\nu \in \mathcal{M}_1(\mathcal{X})} \frac{1}{\alpha-1} \log \int e^{(\alpha-1)g} d\nu - \frac{1}{\alpha} D_\alpha(v\|\mu). \quad (2.100)$$

These two expressions are equivalent (Anantharam, 2017, Section 4) and can also be proven using the same approach undertaken for Lemma 5. Indeed, for Equation (2.99), again using

Hölder's inequality:

$$\frac{1}{\alpha-1} \log \int e^{(\alpha-1)g} d\nu = \frac{1}{\alpha-1} \log \int e^{(\alpha-1)g} \frac{d\nu}{d\mu} d\mu \quad (2.101)$$

$$\leq \frac{1}{\beta(\alpha-1)} \log \int e^{(\alpha-1)\beta g} d\mu + \frac{1}{\alpha(\alpha-1)} \log \int \left(\frac{d\nu}{d\mu} \right)^\alpha d\mu \quad (2.102)$$

$$= \frac{1}{\alpha} \log \int e^{\alpha g} d\mu + \frac{1}{\alpha} D_\alpha(\nu \parallel \mu), \quad (2.103)$$

where the last steps follows as, given $\alpha, \beta = \frac{\alpha}{\alpha-1}$. Equality follows as one generally has a one-to-one mapping between the function achieving the supremum (or infimum) in the variational representation and the measures at play. In this case, one has that for a given g , $d\mu = e^{-g} d\nu$. Normalising μ so that it is a probability measure and plugging it in Equation (2.103) achieves the equality.

Remark 19. Glossing over technical details, writing the Rényi's entropy $H_\alpha(\mathcal{P})$ as $-D_\alpha(\mathcal{P} \parallel \xi)$, one has that, for every α and function f :

$$H_\alpha(\mathcal{P}) = -D_\alpha(\mathcal{P} \parallel \xi) \leq -\beta \log \int e^{(\alpha-1)f} d\mathcal{P} + \log \int e^{\alpha f} d\xi. \quad (2.104)$$

Assuming that ξ is the counting measure and setting $f = \frac{1}{\alpha-1} \log(g)$, then one has that for every positive, bounded and measurable function g :

$$H_\alpha(\mathcal{P}) = -D_\alpha(\mathcal{P} \parallel \xi) \leq -\beta \log \int g d\mathcal{P} + \log \int g^{\frac{\alpha}{\alpha-1}} d\xi \quad (2.105)$$

$$= -\beta \log \mathcal{P}(g) + \log \left(\sum_x g(x)^\beta \right) \quad (2.106)$$

$$= \log \|g\|_{\ell^\beta}^\beta - \beta \log \mathcal{P}(g), \quad (2.107)$$

thus recovering Lemma 4.

Remark 20. Interpreting H_α as a norm one can actually retrieve Equation (2.92) from the simple fact that the dual (in the sense of Definition 4) of the $L_\alpha(\mu)$ -norm $\|\cdot\|_{L^\alpha(\mu)}$ is simply the $L_\beta(\mu)$ -norm $\|\cdot\|_{L^\beta(\mu)}$, with $\frac{1}{\alpha} + \frac{1}{\beta} = 1$.

2.3 A Variational Representation for φ -Divergences

So far we have explored some applications of the variational representations for a variety of objects (Shannon's entropy, Rényi's entropy). It is possible to provide such a representation even for the general class of φ -Divergences. The resulting expression will be important for a variety of settings considered throughout this document. Moreover, analysing the proof will provide insights on how the dual of a φ -Divergence depends on the dual of φ itself. This is why it will be re-derived even though it has appeared in the literature before (Nguyen et al., 2008; Broniatowski and Keziou, 2010). In order to characterise such a variational representation, some clarifications on the considered spaces are in order. In particular, let $F(\mathcal{X})$ be an arbitrary

Chapter 2. Duality and Divergences

family of real-valued functions defined on \mathcal{X} . Denote $\langle F(\mathcal{X}) \cup B(\mathcal{X}) \rangle$ the linear span of $F(\mathcal{X}) \cup B(\mathcal{X})$. Then one can define the sets:

$$\mathcal{M}_1^F(\mathcal{X}) = \left\{ \nu \in \mathcal{M}_1(\mathcal{X}) : \int |f| d\nu < \infty \text{ for } f \in F(\mathcal{X}) \right\},$$

and

$$\mathcal{M}^F(\mathcal{X}) = \left\{ \nu \in \mathcal{M}(\mathcal{X}) : \int |f| d|\nu| < \infty \text{ for } f \in F(\mathcal{X}) \right\}.$$

Here $|\nu|$ denotes the total variation of the finite signed measure ν . If $F = B(\mathcal{X})$ then $\mathcal{M}_1^F(\mathcal{X}) = \mathcal{M}_1(\mathcal{X})$ and $\mathcal{M}^F(\mathcal{X}) = \mathcal{M}(\mathcal{X})$. Denote with τ_F the weakest topology on $\mathcal{M}^F(\mathcal{X})$ such that all mappings $\nu \rightarrow \nu(f)$ are continuous when $f \in \langle F(\mathcal{X}) \cup B(\mathcal{X}) \rangle$ and with τ_M the weakest topology on $\langle F(\mathcal{X}) \cup B(\mathcal{X}) \rangle$ such that all mappings $f \rightarrow \nu(f)$ are continuous when $\nu \in \mathcal{M}_F(\mathcal{X})$. One can show the following result

Proposition 2 ((Broniatowski and Keziou, 2010, Proposition 2.1)). *The space $\mathcal{M}_F(\mathcal{X})$ equipped with the τ_F -topology and the space $\langle F(\mathcal{X}) \cup B(\mathcal{X}) \rangle$ equipped with the τ_M are locally convex topological vector spaces and are the topological dual of the other.*

$D_\varphi(\cdot \| \mu) = \psi(\cdot)$ is thus a convex and lower semi-continuous mapping with respect to τ_F (Broniatowski and Keziou, 2010, Proposition 2.2) and it is possible to characterise its Legendre-Fenchel dual as follows

$$\psi^*(f) = (D_\varphi(\cdot \| \mu))^*(f) = \sup_{\nu} \int f d\nu - \int \varphi \left(\frac{d\nu}{d\mu} \right) d\mu = \int \varphi^*(f) d\mu. \quad (2.108)$$

In order to prove this result we need an intermediate result on convex functions:

Lemma 6 ((Broniatowski and Keziou, 2010)). *Let φ be a differentiable function over A , then*

$$\varphi^*(\varphi'(x)) = x\varphi'(x) - \varphi(x), \quad \forall x \in A, \quad (2.109)$$

moreover, if φ is strictly convex then

$$\varphi^*(x) = x(\varphi')^{-1}(x) - \varphi\left((\varphi')^{-1}(x)\right). \quad (2.110)$$

We can now state the variational representation for φ -Divergences bridging us between the two spaces $\mathcal{M}^F(\mathcal{X})$ and $\langle F(\mathcal{X}) \cup B(\mathcal{X}) \rangle$. For the additional technical condition required on φ (i.e, guaranteeing the uniqueness of the dual optimal solution the reader is referred to (Broniatowski and Keziou, 2010)).

Theorem 10 ((Broniatowski and Keziou, 2010, Proposition 4.2 and Theorem 4.3)). *Let φ be a strictly convex functional and let $\mu \in \mathcal{M}(\mathcal{X})$. One has that for every $\nu \in \mathcal{M}^F(\mathcal{X})$:*

$$D_\varphi(\nu \| \mu) = \sup_{f \in \langle F(\mathcal{X}) \cup B(\mathcal{X}) \rangle} \nu(f) - \mu(\varphi^*(f)), \quad (2.111)$$

where φ^* denotes the Legendre-Fenchel dual of φ . Moreover, one has that for a given $f \in$

$\langle F(\mathcal{X}) \cup B(\mathcal{X}) \rangle$:

$$\mu(\varphi^*(f)) = \sup_{\nu \in \mathcal{M}^F(\mathcal{X})} \nu(f) - D_\varphi(\nu \| \mu). \quad (2.112)$$

Proof. Glossing over technical details, starting from Young's Product Inequality we have that for every measure $\nu \ll \mu$ and function f

$$f \cdot \frac{d\nu}{d\mu} \leq \varphi^*(f) + \varphi\left(\frac{d\nu}{d\mu}\right). \quad (2.113)$$

Integrating on both sides with respect to μ and taking the supremum with respect to ν leads to

$$\sup_\nu \int f d\nu - \int \varphi\left(\frac{d\nu}{d\mu}\right) d\mu \leq \int \varphi^*(f) d\mu \quad (2.114)$$

Moreover, it is possible to see that, under suitable assumptions over φ , given a function f , the measure $\hat{\nu}$ achieving the supremum is actually equal to $d\hat{\nu} = (\varphi')^{-1}(f) d\mu$. Indeed, for every f

$$\int f d\hat{\nu} - D_\varphi(\hat{\nu} \| \mu) = \int f d\hat{\nu} - \int \varphi\left(\frac{d\hat{\nu}}{d\mu}\right) d\mu \quad (2.115)$$

$$= \int f(\varphi')^{-1}(f) d\mu - \int \varphi((\varphi')^{-1}(f)) d\mu \quad (2.116)$$

$$= \int (f(\varphi')^{-1}(f) - \varphi((\varphi')^{-1}(f))) d\mu \quad (2.117)$$

$$= \int \varphi^*(f) d\mu, \quad (2.118)$$

where Equation (2.118) follows from Equation (2.110). Similarly, one can show that, for a given measure ν , the function that achieves the supermum is $\hat{f} = \varphi'\left(\frac{d\nu}{d\mu}\right)$. \square

Several applications of this result will be explored in the following chapters.

Remark 21. *Such a variational representation does not allow us to recover the Donsker-Varadhan characterisation of the Kullback-Leibler divergence. Indeed, it is well-known that $D_\varphi(\nu \| \mu) = D(\nu \| \mu)$ if $\varphi(x) = x \log(x)$. However, instatiating Equation (2.111) to $\varphi(x) = x \log(x)$ leads to the following (also well-known) variational representation for the Kullback-Leibler Divergence:*

$$D(\nu \| \mu) = \sup_f \nu(f) - \mu(\exp(f) + 1). \quad (2.119)$$

It is also known that for a given function f :

$$D(\nu \| \mu) \geq \nu(f) - \log \mu(\exp(f)) \geq \nu(f) - \mu(\exp(f) + 1), \quad (2.120)$$

where the first inequality follows from Equation (2.3) and the second from the fact that $x - 1 \geq \log(x)$ for all $x > 0$. Thus, trying to achieve a variational representation for the Kullback-Leibler Divergence starting from Equation (2.111) leads to a (point-wise) worse expression when compared to Donsker-Varadhan's representation (when one is working with probability

Chapter 2. Duality and Divergences

measures, which for the purposes of Information Theory/Learning Theory/Concentration of Measure, is often the case). This is not specific to the Kullback-Leibler Divergence (Ruderman et al., 2012), indeed, let us fix φ and denote with $\psi_\mu(v)$ our φ -Divergence $D_\varphi(v\|\mu)$, one has that, for a given f the following two variational representations can be defined:

$$\psi_\mu^{R^*}(f) = \sup_{v \in \mathcal{P}(\mathcal{X})} v(f) - \psi_\mu(v) \quad (2.121)$$

$$\psi_\mu^*(f) = \sup_{v \in \mathcal{M}(\mathcal{X})} v(f) - \psi_\mu(v). \quad (2.122)$$

Clearly, since $\mathcal{P}(\mathcal{X}) \subset \mathcal{M}(\mathcal{X})$ one has that for a given f , $\psi_\mu^{R^*}(f) \leq \psi_\mu^*(f)$. Thus, for a given $v \in \mathcal{P}(\mathcal{X})$ and a given f ,

$$D_\varphi(v\|\mu) \geq v(f) - \psi_\mu^{R^*}(f) \geq v(f) - \psi_\mu^*(f). \quad (2.123)$$

3 Independence Vs Dependence

In this chapter¹ we will consider applications of the results presented so far. The purpose is to relate, through divergences, expected values of the same functions but with two different probability measures. We will start with probabilities of events (*i.e.*, expectations of indicator functions).

Let us consider a measurable space (Ω, \mathcal{F}) with the two measures μ and ν defined on it. Let $E \in \mathcal{F}$, we will consider bounds of the following form:

$$\nu(E) \leq \vartheta(\mu, E) \cdot \varpi \left(\frac{d\nu}{d\mu} \right), \quad (3.1)$$

for some functionals ϑ, ϖ . E represents some “undesirable” event (e.g., large [generalisation, cf. Section 4.3] error), whose measure under μ is known and whose measure under ν we wish to bound. The function $\frac{d\nu}{d\mu}$ denotes the Radon-Nikodym derivative of ν with respect to μ (assuming it exists). The quantity $\varpi \left(\frac{d\nu}{d\mu} \right)$ is often going to be a function of some divergence between ν and μ (e.g., Kullback-Leibler, Rényi’s α -Divergence, etc.). Of particular interest in concrete applications is the case where $\Omega = \mathcal{X} \times \mathcal{Y}$, $\nu = \mathcal{P}_{XY}$ (a joint measure), and $\mu = \mathcal{P}_X \mathcal{P}_Y$ (the corresponding product of the marginals). This allows us to bound the likelihood of $E \subseteq \mathcal{X} \times \mathcal{Y}$ when two random variables X and Y are dependent as a function of the likelihood of E when X and Y are independent (a setting that is typically much easier to analyse). The reason why we are trying to bound **ratios** of probabilities as opposed to, say, differences, will become much more clear when considering applications of the results we provide in this chapter. The main idea is to generalise concentration of measures results to settings where (some) random variables are not independent. For instance, if $E = \left\{ \frac{1}{n} \sum_i f(X_i) - \mathcal{P}_X(f(X)) \geq \eta \right\}$ it is possible to control the probability of such an event using a variety of inequalities and under a variety of assumptions (e.g., Azuma-Hoeffding’s Inequality (Hoeffding, 1963; Azuma, 1967), McDiarmid’s inequality (McDiarmid, 1989), etc.). One common assumption is that

¹The content of this chapter until Section 3.1.4 has appeared in the IEEE Transactions on Information Theory 2021, Volume: 67, Issue: 8 (Esposito et al., 2021a). Part of Section Section 3.1.5 has been presented at the International Symposium on Information Theory 2021 (Esposito et al., 2021b).

the function f is independent of the sequence X^n . While reasonable, said assumption limits the applicability of these inequalities in a variety of settings: hypothesis testing when the hypothesis is chosen looking at the data, analysis of the generalisation error of a learning algorithm, etc. In these settings, the idea is that f itself is a random variable and it is the outcome of some algorithm executed over X^n *i.e.*, $f = Y = g(X^n)$ (*e.g.*, data visualisation procedures to formulate the hypothesis, supervised learning algorithms, etc.). Hence, we typically know how to upper-bound $\mathcal{P}_X \mathcal{P}_Y(E)$ and we would like to use this bound to derive results on $\mathcal{P}_{XY}(E)$. Having a result in the shape of Equation (3.1) for opportune ϑ and ω would allow us to generalise these inequalities (McDiarmid's, Hoeffding's, etc.) to settings where the function f is **not** independent of the random variables. In particular, we need

1. $\vartheta(\mathcal{P}_X \mathcal{P}_Y, E) \leq \mathcal{P}_X \mathcal{P}_Y(E)$;
2. $\omega(1) \leq 1$.

Then one could achieve, for cases when $\mathcal{P}_{XY} = \mathcal{P}_X \mathcal{P}_Y$ the following expression which would indeed characterise a generalisation of these inequalities:

$$\mathcal{P}_{XY}(E) \leq \vartheta(\mathcal{P}_X \mathcal{P}_Y, E) \cdot \omega \left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right) = \vartheta(\mathcal{P}_X \mathcal{P}_Y, E) \cdot \omega(1) \leq \mathcal{P}_X \mathcal{P}_Y(E). \quad (3.2)$$

3.1 Ratio of Probabilities

In this section, we present our main results. First, we prove three general bounds on the probability of an event E under a joint distribution \mathcal{P}_{XY} with respect to its probability under the product of the marginals, using Luxemburg, Amemiya norms and φ -mutual Information. We subsequently derive several interesting corollaries that employ common information measures such as Rényi's α -Divergences, Sibson's α -Mutual Information and Maximal Leakage (*cf.* Section 3.1.2), Hellinger divergences and Hellinger Squared Distance (*cf.* Section 3.1.3).

3.1.1 General Results

Our first main bound employs the Luxemburg (*cf.* Equation (1.17)) and the Amemiya norm (*cf.* Equation (1.19)).

Theorem 11. *Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathcal{P}_X \mathcal{P}_Y)$ be two probability spaces and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X \mathcal{P}_Y$. Given $E \in \mathcal{F}$ and two Orlicz functions ζ, ψ :*

$$\mathcal{P}_{XY}(E) \leq \left\| \left\| \mathbb{1}_{\{X \in E_Y\}} \right\|_{L_\zeta^L(\mathcal{P}_X)} \right\|_{L_\psi^L(\mathcal{P}_Y)} \left\| \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right\|_{L_{\zeta_Y^A}^A(\mathcal{P}_X)} \right\|_{L_{\psi_Y^A}^A(\mathcal{P}_Y)} \quad (3.3)$$

$$= \left\| \left\| \frac{1}{\zeta^{-1} \left(\frac{1}{\mathcal{P}_X(E_Y)} \right)} \right\|_{L_\psi^L(\mathcal{P}_Y)} \right\| \left\| \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right\|_{L_{\zeta_Y^A}^A(\mathcal{P}_X)} \right\|_{L_{\psi_Y^A}^A(\mathcal{P}_Y)}, \quad (3.4)$$

where $\mathbb{1}_{\{y\}}$ is the indicator function, and for each $y \in \mathcal{Y}$, $E_y := \{x : (x, y) \in E\}$ (i.e., the “fiber” of E with respect to y), and ζ_Y^* and ψ_Y^* , ζ^{-1} are, respectively, the Young complementary functions of ζ and ψ and the generalised inverse of ζ .

Proof.

$$\mathcal{P}_{XY}(E) = \mathcal{P}_X \mathcal{P}_Y \left(\mathbb{1}_E \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right) \quad (3.5)$$

$$= \mathcal{P}_Y \left(\mathcal{P}_X \left(\mathbb{1}_{\{X \in E_y\}} \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right) \right) \quad (3.6)$$

$$\stackrel{(a)}{\leq} \mathcal{P}_Y \left(\left\| \mathbb{1}_{\{X \in E_y\}} \right\|_{L_\zeta^L(\mathcal{P}_X)} \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right\|_{L_{\zeta_Y^*}^A(\mathcal{P}_X)} \right) \quad (3.7)$$

$$\stackrel{(b)}{\leq} \left\| \left\| \mathbb{1}_{\{X \in E_y\}} \right\|_{L_\zeta^L(\mathcal{P}_X)} \right\|_{L_{\psi_Y^*}^L(\mathcal{P}_Y)} \left\| \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right\|_{L_{\zeta_Y^*}^A(\mathcal{P}_X)} \right\|_{L_{\psi_Y^*}^A(\mathcal{P}_Y)}, \quad (3.8)$$

where (a) and (b) follow from Lemma 2, i.e., generalised Hölder’s inequality. Moreover, one has the following for every $\sigma > 0$ and every $y \in \mathcal{Y}$

$$1 \geq \int \zeta \left(\frac{\mathbb{1}_{E_y}}{\sigma} \right) d\mathcal{P}_X = \zeta \left(\frac{1}{\sigma} \right) \mathcal{P}_X(E_y) \quad (3.9)$$

which, together with the non-decreasability of ζ , implies that

$$\sigma \geq \frac{1}{\zeta^{-1} \left(\frac{1}{\mathcal{P}_X(E_y)} \right)}. \quad (3.10)$$

Consequently, one has that

$$\left\| \mathbb{1}_{\{X \in E_y\}} \right\|_{L_\zeta^L(\mathcal{P}_X)} = \inf \left\{ \sigma > 0 : \int \zeta \left(\frac{\mathbb{1}_{E_y}}{\sigma} \right) \leq 1 \right\} = \frac{1}{\zeta^{-1} \left(\frac{1}{\mathcal{P}_X(E_y)} \right)}, \quad (3.11)$$

where $\zeta^{-1} \left(\frac{1}{\mathcal{P}_X(E_y)} \right) = \inf \left\{ \sigma \geq 0 : \zeta(\sigma) > \frac{1}{\mathcal{P}_X(E_y)} \right\}$ is the generalised inverse of ζ evaluated at $\frac{1}{\mathcal{P}_X(E_y)}$. \square

Remark 22. The expression stated in Equation (3.3) is symmetric with respect to the norms, meaning that one can invert the roles of the Radon-Nikodym and of the indicator function(s) and obtain a bound of the following form:

$$\mathcal{P}_{XY}(E) \leq \left\| \left\| \mathbb{1}_{\{X \in E_y\}} \right\|_{L_\zeta^A(\mathcal{P}_X)} \right\|_{L_{\psi_Y^*}^A(\mathcal{P}_Y)} \left\| \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right\|_{L_{\zeta_Y^*}^L(\mathcal{P}_X)} \right\|_{L_{\psi_Y^*}^L(\mathcal{P}_Y)}. \quad (3.12)$$

Moreover, similarly to Equation (3.4) one can compute the Amemiya norm of $\mathbb{1}_{X \in E_y}$ for every

Chapter 3. Independence Vs Dependence

$y \in \mathcal{Y}$. Indeed, given $y \in \mathcal{Y}$ and $E_y \subseteq \mathcal{X}$

$$\left\| \mathbb{1}_{\{X \in E_y\}} \right\|_{L^A_\zeta(\mathcal{P}_X)} = \inf_{t>0} \frac{1 + \int \zeta(t \mathbb{1}_{\{X \in E_y\}}) d\mathcal{P}_X}{t} \quad (3.13)$$

$$= \inf_{t>0} \frac{1 + \zeta(t) \mathcal{P}_X(E_y)}{t} \quad (3.14)$$

$$= \mathcal{P}_X(E_y) \inf_{t>0} \frac{\frac{1}{\mathcal{P}_X(E_y)} + \zeta(t)}{t} \quad (3.15)$$

$$= \mathcal{P}_X(E_y) (\zeta_Y^*)^{-1} \left(\frac{1}{\mathcal{P}_X(E_y)} \right). \quad (3.16)$$

Before investigating special cases of the above theorem (yielding explicit bounds in terms of known information measures), leveraging the reasoning in Remark 22, we prove a second result that is in the desired form of Equation (3.1) and employs the Luxemburg norm of the Radon-Nikodym Derivative.

Theorem 12. *Let (Ω, \mathcal{F}) be a measurable space and consider two measures μ and ν on this space, such that $\nu \ll \mu$. Given $E \in \mathcal{F}$ and an Orlicz function ψ :*

$$\nu(E) \leq \mu(E) \cdot (\psi_Y^*)^{-1} \left(\frac{1}{\mu(E)} \right) \left\| \frac{d\nu}{d\mu} \right\|_{L^L_{\psi_Y^*}(\mu)}, \quad (3.17)$$

where for $t \geq 0$, $\psi^{-1}(t) := \inf\{s \geq 0 : \psi(s) > t\}$

Proof. Let ψ_Y^* denote the Young's complementary function of ψ (cf. Equation (1.15)).

$$\nu(E) = \mu \left(\mathbb{1}_E \frac{d\nu}{d\mu} \right) \quad (3.18)$$

$$\stackrel{(a)}{\leq} \|\mathbb{1}_E\|_{L^A_\psi(\mu)} \cdot \left\| \frac{d\nu}{d\mu} \right\|_{L^L_{\psi_Y^*}(\mu)} \quad (3.19)$$

$$\stackrel{(b)}{=} \mu(E) \cdot (\psi_Y^*)^{-1} \left(\frac{1}{\mu(E)} \right) \left\| \frac{d\nu}{d\mu} \right\|_{L^L_{\psi_Y^*}(\mu)} \quad (3.20)$$

where (a) follows from the generalised Hölder's inequality, and (b) follows from the considerations in Remark 22. \square

One can then immediately derive the following Corollary of Theorem 12, setting $\Omega = \mathcal{X} \times \mathcal{Y}$, $\nu = \mathcal{P}_{XY}$ and $\mu = \mathcal{P}_X \mathcal{P}_Y$, i.e., setting μ to be a joint measure and ν the corresponding product of the marginals defined over suitable measurable spaces:

Corollary 1. *Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathcal{P}_X \mathcal{P}_Y)$ be two probability spaces and assume that*

$\mathcal{P}_{XY} \ll \mathcal{P}_X \mathcal{P}_Y$. Given $E \in \mathcal{F}$ and an Orlicz function ψ :

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X \mathcal{P}_Y(E) \cdot (\psi_Y^*)^{-1} \left(\frac{1}{\mathcal{P}_X \mathcal{P}_Y(E)} \right) \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right\|_{L_{\psi_Y^*}^L(\mathcal{P}_X \mathcal{P}_Y)}, \quad (3.21)$$

where for $t \geq 0$, $\psi^{-1}(t) := \inf\{s \geq 0 : \psi(s) > t\}$

Remark 23. With respect to Equation (3.1) we have that $\vartheta(t) = t(\psi_Y^*)^{-1}(1/t)$ and $\varpi(t) = \|t\|_{L_{\psi_Y^*}^L(\nu)}$.

Remark 24. Similarly to Theorem 11, Theorem 12 and Corollary 1 are symmetric in the sense that one can exchange the roles of ψ and ψ_Y^* . For instance, in Corollary 1, if one considers the Luxemburg norm $\|\cdot\|_{L_{\psi}^L(\mathcal{P}_X \mathcal{P}_Y)}$ and as its dual norm the Amemiya norm $\|\cdot\|_{L_{\psi_Y^*}^A(\mathcal{P}_X \mathcal{P}_Y)}$, then one achieves the following bound:

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X \mathcal{P}_Y(E) \cdot (\psi_Y^{**})^{-1} \left(\frac{1}{\mathcal{P}_X \mathcal{P}_Y(E)} \right) \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right\|_{L_{\psi}^L(\mathcal{P}_X \mathcal{P}_Y)}. \quad (3.22)$$

Moreover, whenever $\psi_Y^{**} = \psi$, one retrieves a bound involving only the Young function ψ and its generalised inverse (similarly to Equation (3.21), where one only considers ψ_Y^* and its generalised inverse):

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X \mathcal{P}_Y(E) \cdot \psi^{-1} \left(\frac{1}{\mathcal{P}_X \mathcal{P}_Y(E)} \right) \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right\|_{L_{\psi}^L(\mathcal{P}_X \mathcal{P}_Y)}. \quad (3.23)$$

Theorem 13. Let $\varphi : [0, +\infty) \rightarrow \mathbb{R}$ be a convex function such that $\varphi(1) = 0$, and assume φ is non-decreasing on $[0, +\infty)$. Suppose also that φ admits a generalized inverse, defined as $\varphi^{-1}(y) = \inf\{t \geq 0 : \varphi(t) > y\}$. Consider the measurable space $(\mathcal{Z}, \mathcal{F})$, and two measures μ and ν defined on the space. Given an event $E \in \mathcal{F}$, we have that if $\nu \ll \mu$

$$\nu(E) \leq \mu(E) \cdot \varphi^{-1} \left(\frac{D_{\varphi}(\nu \parallel \mu) + (1 - \mu(E))\varphi^*(0)}{\mu(E)} \right), \quad (3.24)$$

where φ^* is the Legendre-Fenchel dual of φ . Moreover, if $\varphi^*(0) \leq 0$, the bound simplifies to

$$\nu(E) \leq \mu(E) \cdot \varphi^{-1} \left(\frac{D_{\varphi}(\nu \parallel \mu)}{\mu(E)} \right). \quad (3.25)$$

Proof. Let $\lambda > 0$. Given a measurable event E one can set $f = \lambda \mathbb{1}_E$ in Equation (2.111) which leads us to the following:

$$D_{\varphi}(\nu \parallel \mu) \geq \lambda \int \mathbb{1}_E d\nu - \int \varphi^*(\lambda \mathbb{1}_E) d\mu \quad (3.26)$$

$$= \lambda \nu(E) - \left(\int_E \varphi^*(\lambda) d\mu + \int_{E^c} \varphi^*(0) d\mu \right) \quad (3.27)$$

$$= \lambda \nu(E) - \mu(E) \varphi^*(\lambda) - \mu(E^c) \varphi^*(0), \quad (3.28)$$

Chapter 3. Independence Vs Dependence

meaning that for every $\lambda > 0$

$$v(E) \leq \frac{D_\varphi(v\|\mu) + \mu(E)\varphi^*(\lambda) + \mu(E^c)\varphi^*(0)}{\lambda} \quad (3.29)$$

$$= \mu(E) \cdot \frac{\frac{D_\varphi(v\|\mu) + \mu(E^c)\varphi^*(0)}{\mu(E)} + \varphi^*(\lambda)}{\lambda}. \quad (3.30)$$

Denoting with $c = \frac{D_\varphi(v\|\mu) + \mu(E^c)\varphi^*(0)}{\mu(E)}$ and with $\vartheta(\lambda) = \varphi^*(\lambda)$

and choosing $\lambda = \vartheta'^{-1}(\vartheta^{\star-1}(c))$ gives us

$$v(E) \leq \mu(E) \cdot \varphi^{-1}\left(\frac{D_\varphi(v\|\mu) + \mu(E^c)\varphi^*(0)}{\mu(E)}\right). \quad (3.31)$$

Indeed, let us denote $\vartheta^{\star-1}(c) = t$, then

$$\frac{c + \vartheta(\vartheta'^{-1}(t))}{\vartheta'^{-1}(t)} = \frac{c + t\vartheta'^{-1}(t) - \vartheta^*(t)}{\vartheta'^{-1}(t)} \quad (3.32)$$

$$= t + \frac{c - \vartheta^*(t)}{\vartheta'^{-1}(t)} \quad (3.33)$$

$$= t = \vartheta^{\star-1}(c) = \varphi^{\star\star-1}(c) = \varphi^{-1}(c), \quad (3.34)$$

where Equation (3.32) follows from Equation (2.110) and Equation (3.34) follows from the fact that $\vartheta^*(\vartheta^{\star-1}(c)) = c$ and from the convexity of φ . In case the function φ is not invertible then, considering the generalised inverse $\vartheta^{\star-1}(c) = \inf\{t : \vartheta^*(t) > c\}$ one has that $\vartheta^*(\vartheta^{\star-1}(c)) > c$ and the argument still follow with an inequality sign in Equation (3.34). \square

Then, like before, one can set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\nu = \mathcal{P}_{XY}$ and $\mu = \mathcal{P}_X\mathcal{P}_Y$ and prove the following:

Theorem 14. *Let $\varphi : [0, +\infty) \rightarrow \mathbb{R}$ be a convex function such that $\varphi(1) = 0$, and assume φ is non-decreasing on $[0, +\infty)$. Suppose also that φ admits a generalized inverse, defined as $\varphi^{-1}(y) = \inf\{t \geq 0 : \varphi(t) > y\}$. Consider the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F})$, and the two measures \mathcal{P}_{XY} and $\mathcal{P}_X\mathcal{P}_Y$ defined on the space. Given an event $E \in \mathcal{F}$, we have that if $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$*

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E) \cdot \varphi^{-1}\left(\frac{I_\varphi(X, Y) + (1 - \mathcal{P}_X\mathcal{P}_Y(E))\varphi^*(0)}{\mathcal{P}_X\mathcal{P}_Y(E)}\right), \quad (3.35)$$

where φ^* is the Legendre-Fenchel dual of φ . Moreover, if $\varphi^*(0) \leq 0$, the bound simplifies to

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X\mathcal{P}_Y(E) \cdot \varphi^{-1}\left(\frac{I_\varphi(X, Y)}{\mathcal{P}_X\mathcal{P}_Y(E)}\right). \quad (3.36)$$

Remark 25. *The proof that has been presented for Theorem 13 is not the shortest nor the most simple one. It does however explicitly highlight how Equation (3.24) follows from Legendre-Fenchel duality. Moreover, it sets the stage for a similar proof in a subsequent section. A simpler*

proof can be found in Section 3.B.

3.1.2 Rényi's Information Measures

While Theorem 11, Theorem 12 and Corollary 1 are quite general, computing the Luxemburg or the Amemiya norm can be difficult for most functions. Moreover, our purpose is to retrieve, on the right-hand side of $\mathcal{P}_{XY}(E)$ some function of $\mathcal{P}_X\mathcal{P}_Y(E)$ and an information measure. With this drive, we will now compute some specific instances of these results for certain choices of ζ and ψ or ϕ , that allow us to retrieve well-known objects in information theory. The first result we derive is a specific instance of Theorem 11 but it will still be quite general. In particular, it will depend on four parameters $\alpha, \alpha', \beta, \beta'$. Different choices of these parameters give rise to bounds involving different Rényi information measures.

Theorem 15. *Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY}), (\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathcal{P}_X\mathcal{P}_Y)$ be two probability spaces, and assume that $\mathcal{P}_{XY} \ll \mathcal{P}_X\mathcal{P}_Y$. Given $E \in \mathcal{F}$ and $y \in \mathcal{Y}$, let $E_y = \{x : (x, y) \in E\}$, i.e. the "fibers" of E with respect to y . Then,*

$$\mathcal{P}_{XY}(E) \leq \left\| \mathbb{1}_{E_y} \right\|_{L^\beta(\mathcal{P}_X)} \left\| \cdot \right\|_{L^{\beta'}(\mathcal{P}_X)} \cdot \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{L^\alpha(\mathcal{P}_X)} \left\| \cdot \right\|_{L^{\alpha'}(\mathcal{P}_Y)} \quad (3.37)$$

$$= \left\| \mathcal{P}_X(E_y)^{\frac{1}{\beta}} \right\|_{L^{\beta'}(\mathcal{P}_X)} \cdot \left\| \frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X\mathcal{P}_Y} \right\|_{L^\alpha(\mathcal{P}_X)} \left\| \cdot \right\|_{L^{\alpha'}(\mathcal{P}_Y)} \quad (3.38)$$

where $\beta, \alpha, \beta', \alpha'$ are such that $1 = \frac{1}{\alpha} + \frac{1}{\beta} = \frac{1}{\alpha'} + \frac{1}{\beta'}$.

Remark 26. *A proof of this result follows from Theorem 11 choosing $\zeta(t) = \frac{|t|^\beta}{\beta}$ and $\psi(t) = \frac{|t|^{\beta'}}{\beta'}$ with $\beta, \beta' \geq 1$ or by choosing $\zeta(t) = |t|^\beta$ and $\psi(t) = |t|^{\beta'}$. With the second choice the link between Equation (3.4) and Equation (3.38) is quite clear as*

$$\frac{1}{\zeta^{-1}\left(\frac{1}{\mathcal{P}_X(E_y)}\right)} = \mathcal{P}_X(E_y)^{\frac{1}{\beta}}. \quad (3.39)$$

A direct proof can be derived using the classical Hölder's inequality twice (similarly to the proof of Theorem 11): once for \mathcal{P}_X and once for \mathcal{P}_Y .

Remark 27. *It is clear from the proof that one can similarly bound $\mathcal{P}_{XY}(g)$ (instead of $\mathcal{P}_{XY}(\mathbb{1}_E)$) for any positive function $g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ that is $\mathcal{P}_X\mathcal{P}_Y$ -integrable via Fubini's Theorem. The shape of the bound, however, becomes more complex as one in general does not have that $g^\beta = g$ for every $\beta \geq 1$. A more thorough discussion on Theorem 15 and expansions of this can be found in Section 3.1.5.*

Choosing $\alpha' = \alpha$ and thus $\beta' = \beta$ in Theorem 15, one recovers the following:

Corollary 2. *Given $E \in \mathcal{F}$ and $\alpha > 1$, we have that:*

$$\mathcal{P}_{XY}(E) \leq (\mathcal{P}_X\mathcal{P}_Y(E))^{\frac{\alpha-1}{\alpha}} \exp\left(\frac{\alpha-1}{\alpha} D_\alpha(\mathcal{P}_{XY} \parallel \mathcal{P}_X\mathcal{P}_Y)\right).$$

Chapter 3. Independence Vs Dependence

This result can, in fact, be proven in several alternative ways:

- via the data processing inequality for D_α ;
- via Corollary 1 with $\psi(t) = \frac{t^\alpha}{\alpha}$;
- via Theorem 14 with $f_\alpha(t) = (t^\alpha - 1)/(\alpha - 1)$ (which gives a bound in terms of Hellinger divergences that are in one-to-one mapping with Rényi's α -divergences (Sason and Verdú, 2016, Equation (80));
- setting $\nu = \mathcal{P}_{XY}$, $\mu = \mathcal{P}_X \mathcal{P}_Y$ and $g = \log(\mathbb{1}_E)$ (essentially, setting $g = 0$ over E and $-\infty$ outside of E) in Lemma 5.

Remark 28. *With respect to Equation (3.1) we again have that $\nu = \mathcal{P}_{XY}$ and $\mu = \mathcal{P}_X \mathcal{P}_Y$ but in this case $\vartheta(t) = t^{1/\beta}$ and $\varpi(\nu/\mu)$ does involve a divergence. In particular, $\varpi(t) = (\mathcal{P}_X \mathcal{P}_Y(t^\alpha))^{\frac{1}{\alpha}}$.*

Notice that Corollary 2 is essentially ignoring the fact that $\mathcal{P}_X \mathcal{P}_Y$ is a product measure, which is fundamental in order to achieve Theorem 11 and Theorem 15. In the case of Corollary 2, this property is not required and indeed one can show a “more general” version that applies to any pair of measures ν and μ defined on (Ω, \mathcal{F}) and such that $\nu \ll \mu$ (e.g., as a Corollary of Theorem 12 with $\psi_Y^*(x) = x^\alpha$).

Corollary 3. *Given $E \in \mathcal{F}$ and $\alpha > 1$, we have that:*

$$\nu(E) \leq (\mu(E))^{\frac{\alpha-1}{\alpha}} \exp\left(\frac{\alpha-1}{\alpha} D_\alpha(\nu\|\mu)\right).$$

Starting from Theorem 15 and considering the limit as $\alpha' \rightarrow 1$, which implies $\beta' \rightarrow +\infty$, we retrieve a bound in terms of Sibson mutual information:

Corollary 4. *Given $E \in \mathcal{F}$, we have that:*

$$\mathcal{P}_{XY}(E) \leq \left(\operatorname{ess\,sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_Y)\right)^{1/\beta} \cdot \mathcal{P}_Y\left(\mathcal{P}_X^{1/\alpha}\left(\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y}\right)^\alpha\right)\right) \quad (3.40)$$

$$= \left(\operatorname{ess\,sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_Y)\right)^{1/\beta} \cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X, Y)\right), \quad (3.41)$$

where $I_\alpha(X, Y)$ is the Sibson mutual information of order α , (Verdú, 2015). Moreover, α and β are such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$.

Remark 29. *An in-depth study of α -Mutual Information can be found in (Verdú, 2015), where a slightly different notation is used. For reference, we can restate Equation (3.40) in the notation of (Verdú, 2015) to obtain:*

$$\mathcal{P}_{XY}(E) \leq \left(\operatorname{ess\,sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_Y)\right)^{1/\beta} \cdot \mathbb{E}_{\mathcal{P}_Y}\left[\mathbb{E}_{\mathcal{P}_X}^{1/\alpha}\left[\left(\frac{d\mathcal{P}_{Y|X}}{d\mathcal{P}_Y}\right)^\alpha \middle| Y\right]\right]. \quad (3.42)$$

Given that α and β are Hölder's conjugates, the bound in Equation (3.41) can be rewritten as:

$$\mathcal{P}_{XY}(E) \leq \exp\left(\frac{1}{\beta} \left(I_\alpha(X, Y) + \log \operatorname{ess\,sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_Y) \right)\right).$$

A property of Sibson's α -Mutual Information is that it is non-decreasing with respect to α (Verdú, 2015). Considering the right hand side of Equation (3.43) we have that, for $\alpha_1 \leq \alpha_2$:

$$\frac{\alpha_1 - 1}{\alpha_1} I_{\alpha_1}(X, Y) \leq \frac{\alpha_2 - 1}{\alpha_2} I_{\alpha_2}(X, Y), \quad (3.43)$$

thus, choosing a smaller α yields a better dependence on $I_\alpha(X, Y)$ in the bound; but given that $\frac{1}{\beta_1} = \frac{\alpha_1 - 1}{\alpha_1} \leq \frac{\alpha_2 - 1}{\alpha_2} = \frac{1}{\beta_2}$ and $\log \operatorname{ess\,sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_Y) \leq 0$, the second term increases for smaller values of α . This leads to a trade-off between the two quantities. Interesting applications of this result find applications in a variety of fields as we will see in subsequent chapters. In such applications, X is typically a random vector of length n and $\mathcal{P}_X(E_Y)$ is exponentially decaying with the number of samples for every y , leading us to bounds of the following form:

$$\mathcal{P}_{XY}(E) \leq \exp\left(\frac{1}{\beta} \left(I_\alpha(X, Y) - 2n\eta^2 \right)\right), \quad (3.44)$$

for some fixed value of η assumed to be small. Thus, as long as, given n $I_\alpha(X, Y) < 2n\eta^2$, then we can guarantee exponential concentration even if X and Y depend on each other. One can also start from Corollary 4 and consider the limit of $\alpha \rightarrow +\infty$, which implies $\beta \rightarrow 1$. This leads to a bound in terms of Sibson's mutual information of order infinity, also known in the literature as Maximal Leakage (Issa et al., 2020):

Corollary 5. *Given $E \in \mathcal{F}$, we have that:*

$$\mathcal{P}_{XY}(E) \leq \left(\operatorname{ess\,sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_Y) \right) \exp(\mathcal{L}(X \rightarrow Y)). \quad (3.45)$$

Unfortunately, taking the limit of $\alpha \rightarrow 1$ in Corollary 2 or Corollary 4 in order to recover a bound involving the Kullback-Leibler Divergence or Mutual Information leads to a vacuous bound:

$$\mathcal{P}_{XY}(E) \leq 1. \quad (3.46)$$

3.1.3 Other divergences

As Hellinger divergences are of independent interest, and include important objects, like the χ^2 -divergence, we restate the bound explicitly in terms of Hellinger divergences. Recall that Hellinger divergences can be characterized by $\varphi_p(t) = (t^p - 1)/(p - 1)$ with $p \in (0, 1) \cup (1, +\infty)$ (Sason and Verdú, 2016, Equation (53)). Theorem 14 can, however, only be applied to $p \in (1, +\infty)$, as $\varphi_p(\cdot)$ is convex but non-increasing for $p \in (0, 1)$. Let $\mathcal{H}_p(X, Y)$ denote the Hellinger divergence of \mathcal{P}_{XY} from $\mathcal{P}_X \mathcal{P}_Y$ and with a slight abuse of notation let $\chi^2(X, Y)$

Chapter 3. Independence Vs Dependence

denote $\chi^2(\mathcal{P}_{XY} \parallel \mathcal{P}_X \mathcal{P}_Y)$. We can now state the following.

Corollary 6. *Let $E \subseteq \mathcal{X} \times \mathcal{Y}$ and let $p \in (1, +\infty)$ then $I_{\varphi_p}(X, Y) = \mathcal{H}_p(X, Y)$ and*

$$\mathcal{P}_{XY}(E) \leq \mathcal{P}_X \mathcal{P}_Y(E)^{\frac{p-1}{p}} \cdot ((p-1)\mathcal{H}_p(X, Y) + 1)^{1/p}. \quad (3.47)$$

In particular, for $p = 2$, we have

$$\mathcal{P}_{XY}(E) \leq \sqrt{(\chi^2(X, Y) + 1)\mathcal{P}_X \mathcal{P}_Y(E)} \quad (3.48)$$

$$\leq \sqrt{\exp(\mathcal{L}(X \rightarrow Y))\mathcal{P}_X \mathcal{P}_Y(E)}. \quad (3.49)$$

Equation (3.49) follows from the fact that $\chi^2(X, Y) \leq \exp(\mathcal{L}(X \rightarrow Y)) - 1$ (cf. (Issa and Gastpar, 2018)). In the family of Hellinger's Divergences, other than the well-known χ^2 divergence, one can also find the famous Hellinger Squared Distance. The main characteristic of this information measure is the fact that it is always guaranteed to be bounded. This is fundamentally different with respect to most of the information-measures presented so far, for which it is easy to find examples that render them equal to $+\infty$. In order to provide such a result with the tools presented in the previous section one needs the following extra assumption (which holds in most of the settings of interest to us): $\mathcal{P}_{XY}(E) \geq \mathcal{P}_X \mathcal{P}_Y(E)$. The reason for this is clear from the proof of the following result.

Corollary 7. *Let $E \in \mathcal{F}$ and assume that $\mathcal{P}_{XY}(E) \geq \mathcal{P}_X \mathcal{P}_Y(E)$, we have that:*

$$\mathcal{P}_{XY}(E) - \mathcal{P}_X \mathcal{P}_Y(E) \leq H^2(X; Y) + 2H(X; Y)\sqrt{\mathcal{P}_X \mathcal{P}_Y(E)}, \quad (3.50)$$

where $H^2(X; Y)$ denotes $H^2(\mathcal{P}_{XY} \parallel \mathcal{P}_X \mathcal{P}_Y)$.

Proof. Let $\varphi(t) = (\sqrt{t} - 1)^2$. We have that $I_{\varphi}(X, Y) = H^2(\mathcal{P}_{XY} \parallel \mathcal{P}_X \mathcal{P}_Y) = H^2(X; Y)$, i.e., the squared Hellinger Distance of the joint from the product of the marginals. Moreover, $\varphi(t)$ is strictly increasing and invertible when restricted to $[1, +\infty)$ and $\varphi^{-1}(t) = t + 1 + 2\sqrt{t}$. Equation (3.50) then follows from Theorem 14 as stated in Equation (3.36). In particular: starting from Equation (3.195) we have that the inverse of φ is applied to an inequality of the form $c \geq \varphi\left(\frac{\mathcal{P}_{XY}(E)}{\mathcal{P}_X \mathcal{P}_Y(E)}\right)$. Given that $\frac{\mathcal{P}_{XY}(E)}{\mathcal{P}_X \mathcal{P}_Y(E)} \geq 1$ by assumption and, using the invertibility of φ on $[+1, \infty)$, we recover Equation (3.50) after some algebraic manipulations. \square

Example 1. Independent case Suppose that X and Y are independent, one has that $H(X; Y) = H^2(X; Y) = 0$ and for every event E one has that Corollary 7 recovers

$$\mathcal{P}_X \mathcal{P}_Y(E) = \mathcal{P}_{XY}(E) \leq \mathcal{P}_X \mathcal{P}_Y(E). \quad (3.51)$$

Example 2. Strongly dependent case Suppose $X = Y \sim \mathcal{U}([n])$ and let $E = \{(x, y) \in [n] \times [n] \mid x = y\}$, then

$$1 = \mathcal{P}_{XY}(E) \leq 1 - \frac{1}{n^{3/2}} + 2\sqrt{\left(1 - \frac{1}{n^{3/2}}\right)\frac{1}{n}}. \quad (3.52)$$

Thus, the bound is asymptotically tight even in case there is a strong dependence.

3.1.4 Tightness

We illustrate the bound by first giving three examples where Inequality (3.45) is met with equality for varying scenarios of dependence: X is independent from Y , X and Y are equal, and X and Y are related but not equal.

Example 1*. *Independent case* Suppose that E is such that $\mathcal{P}_X(E_y) = c$ for all $y \in \mathcal{Y}$. In that case we have that, if X and Y are independent:

$$c = \mathcal{P}_Y(\mathcal{P}_X(E_y)) = \mathcal{P}_{XY}(E) \leq c. \quad (3.53)$$

Example 2*. *Strongly dependent case* Suppose $X = Y \sim \mathcal{U}([n])$ then we have that $\mathcal{L}(X \rightarrow Y) = \log n$ and if $E = \{(x, y) \in [n] \times [n] | x = y\}$ then,

$$1 = \mathcal{P}_{XY}(E) \leq \frac{1}{n} \cdot n = 1. \quad (3.54)$$

Example 3*. Suppose (X, Y) is a doubly-symmetric binary source with parameter p for some $p < 1/2$. Let $E = \{(x, y) : x = y\}$. Then,

$$1 - p = \mathcal{P}_{XY}(E) \leq \frac{1}{2}(2(1 - p)) = 1 - p. \quad (3.55)$$

The above examples show that when the worst-case behavior (i.e., $\max_y \mathcal{P}_X(E_y)$) matches with the average-case behavior (i.e., $\mathcal{P}_Y(\mathcal{P}_X(E_y)) = \mathcal{P}_X \mathcal{P}_Y(E)$), Corollary 5 represents a generalisation of the classical concentration of measure inequalities to adaptive settings. This is typically the case in learning scenarios of interest, where we generalise Hoeffding's and McDiarmid's inequalities to settings where the function of a sequence of random variables X^n (whose concentration around the mean one is trying to prove) can depend on the sequence X^n itself.

Moreover, the following proposition shows that the bound is tight in the following strong sense: if we want to bound the ratio $\mathcal{P}_{XY}(E) / (\text{ess sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_y))$ as a function of $\mathcal{P}_{Y|X}$ only (i.e., independently of \mathcal{P}_X and E), then $\exp(\mathcal{L}(X \rightarrow Y))$ is the best bound one could get:

Proposition 3. *Given finite alphabets \mathcal{X} and \mathcal{Y} , and a fixed conditional distribution $\mathcal{P}_{Y|X}$, then there exists \mathcal{P}_X and E such that (3.45) is met with equality. That is,*

$$\sup_{E \subseteq \mathcal{X} \times \mathcal{Y}} \sup_{\mathcal{P}_X} \log \frac{\mathcal{P}_{XY}(E)}{\text{ess sup}_{\mathcal{P}_Y} \mathcal{P}_X(E_y)} = \mathcal{L}(X \rightarrow Y). \quad (3.56)$$

Proof. Define a function $g: \mathcal{Y} \rightarrow \mathcal{X}$ such that $g(y) \in \arg \max_{x \in \mathcal{X}} \mathcal{P}_{Y|X}(y|x)$, and let $\mathcal{X}_g \subseteq \mathcal{X}$ be the image of g . Now, let \mathcal{P}_X be the uniform distribution over \mathcal{X}_g , and $E = \{(x, y) : x = g(y)\}$.

Then, for any $y \in \mathcal{Y}$,

$$E_y = \{g(y)\} \Rightarrow \mathcal{P}_X(E_y) = \frac{1}{|\mathcal{X}_g|}. \quad (3.57)$$

So we get

$$\mathcal{P}_{XY}(E) = \sum_{(x,y) \in E} \mathcal{P}_{XY}(x, y) \quad (3.58)$$

$$= \sum_{y \in \mathcal{Y}} \sum_{x \in E_y} \mathcal{P}_X(x) \mathcal{P}_{Y|X}(y|x) \quad (3.59)$$

$$= \sum_{y \in \mathcal{Y}} \mathcal{P}_X(g(y)) \mathcal{P}_{Y|X}(y|g(y)) \quad (3.60)$$

$$= \frac{1}{|\mathcal{X}_g|} \sum_{y \in \mathcal{Y}} \max_x \mathcal{P}_{Y|X}(y|x), \quad (3.61)$$

where the last equality follows from Equation (3.57) and the definition of g . \square

3.1.5 Sibson's α -Mutual Information and its functional inequalities

A more extensive discussion can be built around Theorem 15. This type of inequalities can be easily extended to more than two random variables and different ranges of parameters α, β and so on. These choices can lead us to different (or new) information measures whose operational meaning can be immediately deduced by the functional inequality one starts from. Or, inverting the order, one can define (starting from an approach similar to Equation (1.50)) a family of information measures and endow them with an operational meaning deriving from the corresponding functional inequality. The behaviour described in Theorem 15 contains all the possible nested norms for two random variables with parameters $\alpha, \alpha' > 1$. If one considers two random variables but $\alpha < 1$, the following result can be easily proven.

Theorem 16. *Let X, Y be two random variables whose joint measure is \mathcal{P}_{XY} and the corresponding marginals are given by \mathcal{P}_X and \mathcal{P}_Y . For every $g: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ one has that*

$$\mathcal{P}_{XY}(g) \geq \mathcal{P}_Y^{\frac{1}{\beta'}} \left(\mathcal{P}_X^{\frac{\beta'}{\beta}} \left(g^\beta \right) \right) \cdot \mathcal{P}_Y^{\frac{1}{\alpha'}} \left(\mathcal{P}_X^{\frac{\alpha'}{\alpha}} \left(\left(\frac{d\mathcal{P}_{XY}}{d\mathcal{P}_X \mathcal{P}_Y} \right)^\alpha \right) \right), \quad (3.62)$$

where $\frac{1}{\alpha} + \frac{1}{\beta} = 1 = \frac{1}{\alpha'} + \frac{1}{\beta'}$ and $\alpha, \alpha' < 1$.

Now, if $0 < \alpha < 1$ then one has that $\beta < 0$ while if $\alpha < 0$, then $0 < \beta < 1$. With $\alpha' \rightarrow 1^-$ (which implies $\beta' \rightarrow -\infty$) then one can provide a result involving:

- the regular $I_\alpha(X, Y)$, if $0 < \alpha < 1$;
- an unfamiliar object, if $\alpha < 0$. This object, to the best of our knowledge, has not appeared before in the literature and can be defined to be an extension of Sibson's I_α to negative values of α (cf. (Esposito et al., 2022, Definition 3)).

More formally,

Corollary 8. *Consider the same setting as in Theorem 16, that is: let X, Y be two random variables whose joint measure is \mathcal{P}_{XY} and the corresponding marginals are given by \mathcal{P}_X and \mathcal{P}_Y . For every $g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ and for every $\alpha < 1$ and $\beta = \frac{\alpha-1}{\alpha}$ one has that*

$$\mathcal{P}_{XY}(g) \geq \operatorname{ess\,inf}_{\mathcal{P}_Y} \left(\mathcal{P}_X^{\frac{1}{\beta}}(g^\beta) \right) \cdot \exp \left(\operatorname{sign}(\alpha) \cdot \frac{\alpha-1}{\alpha} I_\alpha(X, Y) \right). \quad (3.63)$$

Remark 30. *Corollary 8 holds for every non-negative function g . However, given $0 < \alpha < 1$, which in turn implies $\beta < 0$, an issue for null functions arises. Indeed, if there exists an x such that $\mathcal{P}_X(\{x\}) > 0$ and, for every y with $\mathcal{P}_Y(\{y\}) > 0$, one has $g(x, y) = 0$, then a trivial lower-bound of 0 on $\mathcal{P}_{XY}(g)$ is derived. Indeed for such x , $g(x, y)^\beta = 0^\beta = \infty$ for every y rendering $\operatorname{ess\,inf}_{\mathcal{P}_Y} \left(\mathcal{P}_X^{\frac{1}{\beta}}(g^\beta) \right) = 0$. This prevents us from setting $g = \mathbb{1}_E$ and, thus, from recovering a bound that involves probabilities if $0 < \alpha < 1$.*

Whenever $\alpha < 0$ (which implies $0 < \beta < 1$), one can consider functions that are null on spaces of non-zero measure and provide the following lower-bound. This bound connects the probability of any event E with respect to the joint measure \mathcal{P}_{XY} and the corresponding product of the marginals $\mathcal{P}_X \mathcal{P}_Y$:

Corollary 9. *Let X, Y be two random variables and consider the probability spaces $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathcal{P}_{XY})$ and $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathcal{P}_X \mathcal{P}_Y)$. Let $E \in \mathcal{F}$ and, given $y \in \mathcal{Y}$, denote with $E_y = \{x : (x, y) \in E\}$, then, for every $\alpha < 0$*

$$\mathcal{P}_{XY}(E) \geq \min_y \mathcal{P}_X(E_y)^{\frac{1}{\beta}} \cdot \exp \left(-\frac{\alpha-1}{\alpha} I_\alpha(X, Y) \right) \quad (3.64)$$

$$= \exp \left(\frac{1}{\beta} \left(\log(\min_y \mathcal{P}_X(E_y)) - I_\alpha(X, Y) \right) \right). \quad (3.65)$$

Taking the limit of $\alpha \rightarrow -\infty$ which implies $\beta \rightarrow 1$ one recovers the following:

$$\mathcal{P}_{XY}(E) \geq \min_y \mathcal{P}_X(E_y) \exp(-I_{-\infty}(X, Y)) \quad (3.66)$$

$$= \min_y \mathcal{P}_X(E_y) \exp(-\mathcal{L}^c(X \rightarrow Y)). \quad (3.67)$$

Remark 31. *The case when $\alpha \rightarrow -\infty$ is, in a way, symmetric to $\alpha \rightarrow \infty$ as, essentially:*

$$\exp(-I_{-\infty}(X, Y)) = \exp(-\mathcal{L}^c(X \rightarrow Y)) \leq \frac{\mathcal{P}_{XY}(E)}{\min_y \mathcal{P}_X(E_y)} \quad (3.68)$$

as opposed to

$$\exp(I_\infty(X, Y)) = \exp(\mathcal{L}(X \rightarrow Y)) \geq \frac{\mathcal{P}_{XY}(E)}{\max_y \mathcal{P}_X(E_y)}. \quad (3.69)$$

Similarly, one could obtain a result involving D_α with $\alpha < 0$ from Theorem 16, setting $\alpha = \alpha'$. This result is also, in a way, symmetric to Corollary 2. Indeed, if $\alpha \rightarrow \infty$, Corollary 2 boils down

Chapter 3. Independence Vs Dependence

to:

$$\frac{\mathcal{P}_{XY}(E)}{\mathcal{P}_X \mathcal{P}_Y(E)} \leq \exp(D_\infty(\mathcal{P}_{XY} \parallel \mathcal{P}_X \mathcal{P}_Y)) = \sup_F \frac{\mathcal{P}_{XY}(F)}{\mathcal{P}_X \mathcal{P}_Y(F)} \quad (3.70)$$

for any given event E , while an equivalent result obtained from from Theorem 16 when considering $\alpha \rightarrow -\infty$ would boil down to:

$$\frac{\mathcal{P}_{XY}(E)}{\mathcal{P}_X \mathcal{P}_Y(E)} \geq \exp(D_{-\infty}(\mathcal{P}_{XY} \parallel \mathcal{P}_X \mathcal{P}_Y)) = \inf_F \frac{\mathcal{P}_{XY}(F)}{\mathcal{P}_X \mathcal{P}_Y(F)} \quad (3.71)$$

for any event E .

Let us compare, through Table 3.1, Corollary 8, Corollary 9 and a straight-forward generalisation of Corollary 4 that we will now state for reference:

Corollary 10. Consider the same setting as in Theorem 16, that is: let X, Y be two random variables whose joint measure is \mathcal{P}_{XY} and the corresponding marginals are given by \mathcal{P}_X and \mathcal{P}_Y . For every $g: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ and for every $\alpha > 1$ and $\beta = \frac{\alpha-1}{\alpha}$ one has that

$$\mathcal{P}_{XY}(g) \leq \text{ess sup}_{\mathcal{P}_Y} \left(\mathcal{P}_X^{\frac{1}{\beta}}(g^\beta) \right) \cdot \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X, Y)\right). \quad (3.72)$$

Differently from I_α with $\alpha > 0$, I_α with $\alpha < 0$ is strongly connected to converse/negative results. For instance, it can find applications in estimation procedures that lower-bounds the Bayesian Risk. In these settings, one often has a non-negative loss function $\ell(\cdot, \cdot)$ that measures how far a parameter is from its estimation. The purpose is often to lower-bound the minimum expected value of this loss where the minimum is over all the possible estimators of the parameter. In such a framework Corollary 8 can be useful, as we will see in a following Chapter. However, Corollary 10 can also be immediately employed in such a setting with a simple trick (cf. Section 5.3).

Table 3.1: Behaviour of the bounds expressed in Corollary 8, Corollary 9, Corollary 10 and Corollary 4

Behaviour of the Bound $\mathcal{P}_{XY}(g) \stackrel{\leq}{\geq} h_\beta(g) \cdot \hat{h}(I_\alpha(X, Y))$			
Range of α	$\alpha < 0 \implies 0 < \beta < 1$	$0 < \alpha < 1 \implies \beta < 0$	$\alpha > 1 \implies \beta > 1$
Information-Measure $\hat{h}(I_\alpha)$	$\exp((1-\alpha)/\alpha \cdot I_\alpha(X, Y))$	$\exp((\alpha-1)/\alpha \cdot I_\alpha(X, Y))$	$\exp((\alpha-1)/\alpha \cdot I_\alpha(X, Y))$
Multiplicative Term $h_\beta(g)$	$\min_y \mathcal{P}_X^{\frac{1}{\beta}}(g(X, y)^\beta)$	$\min_y \mathcal{P}_X^{\frac{1}{\beta}}(g(X, y)^\beta)$	$\max_y \mathcal{P}_X^{\frac{1}{\beta}}(g(X, y)^\beta)$
Multiplicat. Term $h_\beta(\mathbb{1}_E)$	$\min_y (\mathcal{P}_X(E_y))^{\frac{1}{\beta}}$	cannot be provided	$\max_y (\mathcal{P}_X(E_y))^{\frac{1}{\beta}}$
Inequality	$\mathcal{P}_{XY}(g) \geq h_\beta(g) \cdot \hat{h}(I_\alpha(X, Y))$	$\mathcal{P}_{XY}(g) \geq h_\beta(g) \cdot \hat{h}(I_\alpha(X, Y))$	$\mathcal{P}_{XY}(g) \leq h_\beta(g) \cdot \hat{h}(I_\alpha(X, Y))$
References	Corollary 8 and Corollary 9	Corollary 8	Corollary 10 and Corollary 4

A general approach that can be undertaken is to start from results like Theorem 16, carefully choose the parameters and then, looking closely at the functional of the Radon-Nikodym derivative that appears, draw a connection to well-known information measures. In particular, when $\alpha' \rightarrow 1$ and $\alpha < 0$ in Theorem 16, one retrieves Corollary 8, where I_α can be defined as follows:

$$I_\alpha(X, Y) = -\max_{\mathcal{Q}_Y} D_\alpha(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{Q}_Y). \quad (3.73)$$

Once the link with known information measures is established, one can show a sequence of information-measure-like properties that such an object satisfies (cf. (Esposito et al., 2022, Theorem 3)).

A similar perspective can be undertaken with Theorem 15 but generalised to three random variables instead. In this case, the choice of parameters grows (as well as the number of orderings that one can use in defining the nested norms) and can lead to a variety of objects.

Theorem 17. *Let X, Y, Z be three random variables whose joint measure is \mathcal{P}_{XYZ} . Let \mathcal{P}_Z be the marginal with respect to Z of \mathcal{P}_{XYZ} and assume the existence of the conditional measures $\mathcal{P}_{X|Z}$ and $\mathcal{P}_{Y|Z}$. Assuming that $\mathcal{P}_{XYZ} \ll \mathcal{P}_{Y|Z} \mathcal{P}_{X|Z} \mathcal{P}_Z$, for every $g : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \rightarrow \mathbb{R}^+$ one has that:*

$$\mathcal{P}_{XYZ}(g) \leq \left\| \| \| g \|_{L^\beta(\mathcal{P}_{X|Z})} \|_{L^{\beta'}(\mathcal{P}_{Y|Z})} \|_{L^{\beta''}(\mathcal{P}_Z)} \cdot \left\| \left\| \frac{d\mathcal{P}_{XYZ}}{d\mathcal{P}_Z \mathcal{P}_{Y|Z} \mathcal{P}_{X|Z}} \right\|_{L^\alpha(\mathcal{P}_{X|Z})} \right\|_{L^{\alpha'}(\mathcal{P}_{Y|Z})} \right\|_{L^{\alpha''}(\mathcal{P}_Z)}, \quad (3.74)$$

where $\frac{1}{\alpha} + \frac{1}{\beta} = \frac{1}{\alpha'} + \frac{1}{\beta'} = \frac{1}{\alpha''} + \frac{1}{\beta''} = 1$ and $\alpha, \alpha', \alpha'' > 1$.

Setting $\alpha'' = \alpha$ and letting $\alpha \rightarrow 1$ (which, in turn, imply $\beta'' = \beta$ and $\beta \rightarrow \infty$) one then recovers on the right-hand side of Equation (3.74) a form of conditional Sibson's α -Mutual Information $I_\alpha^{Y|Z}(X, Y|Z)$ that has appeared in (Tomamichel and Hayashi, 2018, Section IV.C.2) and that corresponds to the following information measure

$$I_\alpha^{Y|Z}(X, Y|Z) = \min_{\mathcal{Q}_{Y|Z}} D_\alpha(\mathcal{P}_{XYZ} \| \mathcal{P}_Z \mathcal{Q}_{Y|Z} \mathcal{P}_{X|Z}). \quad (3.75)$$

The associated functional inequality is then the following:

Corollary 11. *Under the same setting as in Theorem 17 one has that:*

$$\mathcal{P}_{XYZ}(g) \leq \mathcal{P}_Z^{\frac{\alpha-1}{\alpha}} \left(\operatorname{ess\,sup}_{\mathcal{P}_{Y|Z}} \mathcal{P}_{X|Z}(g) \right) \cdot \exp \left(\frac{\alpha-1}{\alpha} I_\alpha^{Y|Z}(X, Y|Z) \right). \quad (3.76)$$

Choosing a different factorisation of the product measures in Theorem 17 one can obtain a result that is analogous to Corollary 11 (but that involves a different information measure):

Corollary 12. *Under the same setting as in Theorem 17 one has that:*

$$\mathcal{P}_{XYZ}(g) \leq \left(\operatorname{ess\,sup}_{\mathcal{P}_Z} (\mathcal{P}_{Y|Z} \mathcal{P}_{X|Z}(g)) \right)^{\frac{\alpha-1}{\alpha}} \cdot \exp \left(\frac{\alpha-1}{\alpha} I_\alpha^Z(X, Y|Z) \right). \quad (3.77)$$

Chapter 3. Independence Vs Dependence

This time the information measure considered corresponds to the following (cf. (Esposito et al., 2021b, Definition 4)):

$$I_\alpha^Z(X, Y|Z) = \min_{Q_Z} D_\alpha(\mathcal{P}_{XYZ} \| \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z} Q_Z). \quad (3.78)$$

Different factorisations can be chosen in Theorem 17 (cf. Equation (3.76), Equation (3.77)) which lead to different minimisations (respectively, Equation (3.75), Equation (3.78)) and thus to different information measures. These information measures will have similar properties (Esposito et al., 2021b).

The main idea, much like with Shannon’s conditional Mutual Information, is to estimate how far the joint probability measure is from one that formalises the Markov Chain $X - Z - Y$. In the case of Sibson’s α -Mutual Information, this is done using Rényi’s Divergences followed by a minimisation step (cf. Equation (3.75), Equation (3.78)). These objects will share similar properties while possessing some unique characteristics.

In particular, the one defined in Equation (3.75) is an asymmetric information measure and has the property that taking the limit of $\alpha \rightarrow \infty$ one retrieves conditional Maximal Leakage (Issa et al., 2020, Definition and Theorem 6):

$$I_\alpha^{Y|Z}(X, Y|Z) \xrightarrow{\alpha \rightarrow \infty} \mathcal{L}(X \rightarrow Y|Z). \quad (3.79)$$

The conditional I_α^Z defined in Equation (3.78), differently from Equation (3.75), is symmetric and can be connected to the strong data-processing coefficient of the Hellinger integral (cf. (Esposito et al., 2021b, Theorem 4)).

To conclude, let us reiterate the pattern analysed in this section:

1. each of these objects has a nested-norm structure that stems from D_α and the family of measures one is minimising over (e.g., minimising over Q_Z leads to an $L_1(\mathcal{P}_Z)$ -norm in the information measure and an $L_\infty(\mathcal{P}_Z)$ -norm in the corresponding dual-norm);
2. the nested-norm structure can be easily connected to functional inequalities similar to Theorem 17 via multiple application of Hölder’s inequality;
3. the inequalities allow us, in turn, to connect the information-measure to a variety of settings (lower-bounds on the Bayesian Risk in estimation procedures, cf. Chapter 5, Hypothesis Testing problems (Esposito et al., 2021b, Section IV), etc.).

One can, of course, consider different ranges of values for $\alpha, \alpha', \alpha''$ (e.g., $0 < \alpha < 1$ or $\alpha < 0$) and retrieve inequalities in the opposite direction with corresponding “new” information measures. Once again, duality (in this case, between the norms) is the key element to consider.

3.2 Difference of Expectations

In the previous part of the chapter we tackled the problem of bounding the ratio between the probability of an event under the joint and under (a functional of) the product of the marginals. This led to a sequence of bounds involving a variety of norms (Luxemburg, Amemiya) and of information measures (Sibson's Mutual Information, Rényi's Divergences, φ -Divergences). Another related and relevant question we will address, with similar tools, is the following: given two measures μ, ν and a function f can we provide bounds of the following form

$$|\mu(f) - \nu(f)| \leq \omega(\nu, \mu) ? \quad (3.80)$$

I.e., can one bound the difference of the expectations of the same function with respect to two different measures using, for instance, a function of a divergence $\psi_\mu(\nu) = D_\varphi(\nu \parallel \mu)$? This problem is not new and has appeared in a variety of forms and contexts in the literature: exploration bias (Russo and Zou, 2016; Jiao et al., 2017; Issa and Gastpar, 2018), learning theory (Xu and Raginsky, 2017c), etc. As we will see, this problem is also connected to Wasserstein Distances, Transportation-Cost Inequalities as well as with bounding the **expected** generalisation error of a learning algorithm. One can, in particular, provide the following result:

Theorem 18. *Let $f \in C_c(\mathcal{X})$ and ϕ be a strictly convex function such that its Legendre-Fenchel dual ϕ^* admits a generalised inverse $\phi^{\star-1}(t) = \inf\{s : \phi(s) \geq t\}$. Let $\psi^* : \mathcal{M}(\mathcal{X}) \rightarrow \mathbb{R}$ be the Legendre-Fenchel dual of $\psi : C_c(\mathcal{X}) \rightarrow \mathbb{R}$. If*

$$\psi(\lambda f) \leq \phi(\lambda) \text{ for every } \lambda > 0 \quad (3.81)$$

then, for every measure ν such that $\nu(f) < +\infty$ and $\phi^{\prime-1}(\phi^{\star-1}(\psi^(\nu))) > 0$,*

$$\nu(f) \leq \phi^{\star-1}(\psi^*(\nu)). \quad (3.82)$$

Proof. By definition of the Legendre-Fenchel transform we know that for a given function f and any given measure ν :

$$\psi^*(\nu) = \sup_{g \in C_c(\mathcal{X})} \langle g, \nu \rangle - \psi(g) \quad (3.83)$$

$$\geq \lambda \langle f, \nu \rangle - \psi(\lambda f) = \lambda \nu(f) - \psi(\lambda f). \quad (3.84)$$

Hence, we can say that, given $f \in C_c(\mathcal{X})$, $\nu \in \mathcal{M}(\mathcal{X})$ and $\lambda > 0$

$$\nu(f) \leq \frac{\psi(\lambda f) + \psi^*(\nu)}{\lambda} \leq \frac{\phi(\lambda) + \psi^*(\nu)}{\lambda}, \quad (3.85)$$

where in Equation (3.85) we used our assumption on ψ , *i.e.* Equation (3.81). Denoting with $c = \psi^*(\nu)$ and choosing $\lambda = \phi^{\prime-1}(\phi^{\star-1}(c))$ gives us that

$$\nu(f) \leq \phi^{\star-1}(c) = \phi^{\star-1}(\psi^*(\nu)). \quad (3.86)$$

Chapter 3. Independence Vs Dependence

Indeed, let us denote, for simplicity, $\phi^{\star-1}(c) = t$, then replacing λ with $\phi'^{-1}(\phi^{\star-1}(c))$ in Equation (3.85) one has

$$\frac{c + \phi(\phi'^{-1}(t))}{\phi'^{-1}(t)} = \frac{c + t\phi'^{-1}(t) - \phi^{\star}(t)}{\phi'^{-1}(t)} \quad (3.87)$$

$$= t + \frac{c - \phi^{\star}(t)}{\phi'^{-1}(t)} \quad (3.88)$$

$$= \phi^{\star-1}(c) + \frac{c - \phi^{\star}(\phi^{\star-1}(c))}{\phi'^{-1}(t)} \quad (3.89)$$

$$\leq \phi^{\star-1}(\psi^{\star}(v)), \quad (3.90)$$

where Equation (3.87) follows from Equation (2.110). \square

Remark 32. In most settings of interest the assumption that $\phi'^{-1}(\phi^{\star-1}(\psi^{\star}(v))) > 0$ is easily satisfied. If one considers $\psi^{\star}(v)$ to be a φ -Divergence, i.e., $\psi^{\star}(v) = D_{\varphi}(v\|\mu)$, with μ a probability measure fixed before-hand, then $\psi^{\star}(v) \geq 0$ for every $v \in \mathcal{P}(\mathcal{X})$. Typical functions ϕ will be of the form $\phi(x) = |x|^{\alpha}/\alpha$ with $\alpha > 1$. This implies that $\phi'^{-1}(\phi^{\star-1}(\psi^{\star}(v))) = (\beta\psi^{\star}(v))^{\frac{1}{\alpha}}$ with $\beta = \alpha/(\alpha - 1) > 0$, which is clearly positive for $\psi^{\star}(v) > 0$. In general it is possible to state a slightly less general result by adding some additional constraints on ϕ that allow us to compute the infimum in Equation (3.85) using (Boucheron et al., 2013, Lemma 2.4).

Remark 33. The pattern we highlighted in Chapter 2 appears once again in a slightly different form but with the same spirit. In order to provide these results, connecting the expected value of a function with a functional of measures, one has to:

- (a) use the Legendre-Fenchel transform and duality (in order to tie expected values and the functional of measures);
- (b) bound the Legendre-Fenchel dual of the divergence that we have chosen;

Thus, if we can provide a bound on ψ (cf. Equation (3.81)) then we can relate the expected value of a function f under the measure v and the functional $\psi^{\star}(v)$ (cf. Equation (3.82)). For reference, in the converse of Shannon's coding theorem (cf. Section 2.1.2) we had to bound the dual of entropy evaluated at the function l (length of the code, e.g., assuming Kraft's inequality). Through duality we then recovered a result connecting the expected value $\mathcal{P}(l)$ and the Shannon Entropy of \mathcal{P} .

Let us now proceed with examples with the purpose of better understanding both the assumptions and the implications of Theorem 18 and how it fits in the framework and the desiderata of this chapter.

Corollary 13. Let X be a random variable over the probability space $(\mathcal{X}, \mathcal{F}, \mu)$ and assume X to be a zero-mean σ^2 -sub-Gaussian random variable² with respect to μ . We have that for every

²Given a zero-mean random variable X we say that it is σ^2 -sub-Gaussian if the following holds true for every $\lambda \in \mathbb{R}$: $\log(\mu(e^{\lambda X})) \leq \frac{\lambda^2 \sigma^2}{2}$.

measure ν such that $\nu \ll \mu$ and such that $\nu(X) < +\infty$

$$\nu(X) \leq \sqrt{2\sigma^2 D(\nu\|\mu)}. \quad (3.91)$$

Proof. The proof follows from Theorem 18, selecting $\psi^*(\nu) = D(\nu\|\mu)$ with $\nu \in \mathcal{P}(\mathcal{X})$. The fact that X is 0 mean, σ^2 sub-Gaussian under μ means that for every λ

$$\log \int_{\mathcal{X}} e^{\lambda X} d\mu \leq \frac{\lambda^2 \sigma^2}{2}, \quad (3.92)$$

i.e., the assumption of sub-Gaussianity under μ of X is implicitly implying a bound on the Legendre-Fenchel dual of the Kullback-Leibler with a strictly convex function of λ , $\phi(\lambda) = \frac{\lambda^2 \sigma^2}{2}$. Computing the inverse of the convex-conjugate of ϕ we can thus bound $\nu(f)$ in terms of $D(\nu\|\mu)$. Indeed, to complete the argument, we observe that

$$\phi^*(\lambda^*) = \frac{\lambda^{*2}}{2\sigma^2} \implies \phi^{*-1}(t) = \sqrt{2\sigma^2 t},$$

which then establishes the claimed bound. □

Corollary 13 represents one of the simplest examples of applicability of Theorem 18. The choice of ψ_μ and ψ_μ^* allows us to immediately retrieve a well-known result on σ^2 -sub-Gaussian random variables. This result lies at the foundations of most of the bounds connecting Mutual Information and the expected generalisation error of learning algorithms like we will see in Section 4.2.

3.2.1 Generalising Transportation-cost Inequalities

Background

The origins of Optimal Transport date back to Monge's work in 1781. In his formulation, the problem represented the formalisation of a very intuitive and practical issue: redistribution (in the sense of transportation and reshaping) of some material (e.g., sand or soil) with the minimum effort/cost. The problem, as stated by Monge, does not always admit a solution (Ambrosio, 2003, Example 1.1.). Almost 200 years later, the problem was resurrected and rescued from oblivion by Kantorovich, who proposed a simple but powerful relaxation of the problem (Ambrosio, 2003, Section 2). Denote with $\mathcal{M}(\mathcal{X})$ the set of all Radon signed measures over \mathcal{X} and with $\mathcal{M}_1(\mathcal{X})$ the set of all probability measures over \mathcal{X} . Let $\mu, \nu \in \mathcal{M}_1(\mathcal{X})$, consider the set $\Pi(\mu, \nu)$ of all the joint probability measures $\pi \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$ with marginals equal to μ and ν , *i.e.*, such that $\pi(\cdot \times \mathcal{X}) = \mu(\cdot)$ and $\pi(\mathcal{X} \times \cdot) = \nu(\cdot)$. The problem advanced by Kantorovich was the following: given μ and ν and a Borel function $d: \mathcal{X} \times \mathcal{X} \rightarrow [0, +\infty]$, can we find a joint measure π that minimises

$$\pi(d(X, Y)) ?$$

Chapter 3. Independence Vs Dependence

The coupling π represents a **transport plan** between μ, ν . Under mild assumptions on d , optimal transport plans are guaranteed to exist for extremely general spaces \mathcal{X} (Villani, 2008, Theorem 4.1). Moreover, any transport map induces a transport plan and a transport plan can be seen as induced by a transport map under some additional conditions (Ambrosio, 2003, Proposition 2.1). If d itself is a metric over \mathcal{X} , then $\inf_{\pi} \pi(d(X, Y))$ represents a distance over the space of probability measures: the larger is the quantity the more “difficult” it is to transform μ into ν . In particular, given a metric d , let us denote with $\mathcal{M}_p(\mathcal{X})$ the set of probability measures μ on \mathcal{X} such that $\mu(d(X, x_0)^p)^{1/p} < +\infty$ for some $x_0 \in \mathcal{X}$.

Definition 20 ((Villani, 2008, Def. 6.1)). *Let (\mathcal{X}, d) be a Polish space and $p \in [1, +\infty)$. Let $\mu, \nu \in \mathcal{M}_p(\mathcal{X})$, the p -Wasserstein distance between μ and ν is defined as*

$$W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} (\pi(d(X, Y)^p))^{1/p}.$$

Wasserstein distances satisfy interesting properties (Raginsky and Sason, 2014, Lemma 3.4.1).

Lemma 7 ((Villani, 2008)). *Let (\mathcal{X}, d) be a Polish space we have that:*

- if $p \geq 1$ W_p is a metric on $\mathcal{M}_p(\mathcal{X})$;
- if $1 \leq p \leq q$ then $W_p(\mu, \nu) \leq W_q(\mu, \nu) \forall \mu, \nu \in \mathcal{M}_q(\mathcal{X})$;
- W_p metrizes weak convergence in $\mathcal{M}_p(\mathcal{X})$.

Moreover, when connected to Kullback Leibler-divergences, in what are known in the literature as “Transportation-cost Inequalities”, they have interesting implications in the concentration of measure phenomenon.

Definition 21 (Transportation-Cost Inequality). *Let (\mathcal{X}, d) be a Polish space and μ a probability measure on \mathcal{X} , we say that μ satisfies an L^p -transportation-cost inequality with constant c (or $T_p(c)$ in short) if for every $\nu \ll \mu$*

$$W_p(\mu, \nu) \leq \sqrt{2cD(\nu \parallel \mu)}. \quad (3.93)$$

When $p = 1$, for instance, these inequalities are equivalent to concentration in the following sense:

Theorem 19 ((Bobkov and Götze, 1999, Thm 3.1)). *Let $\mu \in \mathcal{M}_1(\mathcal{X})$ be a Borel probability measure. There exists a c such that for every $\lambda \in \mathbb{R}$*

$$\log \mu(\exp(\lambda f)) \leq \left(\frac{c\lambda^2}{2} \right) \quad (3.94)$$

for every 1-Lipschitz function f , if and only if μ satisfies a $T_1(c)$ inequality, i.e., for every $\nu \ll \mu$

$$W_1(\mu, \nu) \leq \sqrt{2cD(\nu \parallel \mu)}. \quad (3.95)$$

Example 6. *Let \mathcal{X} be a discrete space and $d(x, y) = \mathbb{1}_{x \neq y}$. We have that $W_1(\mu, \nu) = TV(\mu, \nu)$, i.e., the Total Variation distance between μ, ν (Raginsky and Sason, 2014, Prop. 3.4.1). In this case*

the transportation cost inequality is well-known under the name of Pinsker's inequality and it holds for every $\mu \in \mathcal{M}_1(\mathcal{X})$ with $c = 1/4$. In general, these inequalities are highly non-trivial and hold for specific distributions (e.g., Gaussian, etc.).

Transportation-Cost Inequalities via Duality

A recurring theme in the field that lies at the intersection between Information Theory and Optimal Transport is showing that Transportation-Cost Inequalities are equivalent to some form of concentration of measure. One such example is Theorem 19. An in-depth review on the topic with an information-theoretic perspective can be found in (Raginsky and Sason, 2014, Section 3.4). This type of results are generally made of two parts:

- the “if part”: if μ satisfies concentration for every function in some family then the $T_p(c)$ holds;
- the “only if” part: if μ satisfies a $T_p(c)$ inequality then one has concentration for every function in some family.

So far, Theorem 18 essentially establishes the “if part” for general functionals of measures. Let us now establish the “only if” part and then discuss how it relates to the classical framework.

Theorem 20. *Let $\psi : \mathcal{M}(\mathcal{X}) \rightarrow \mathbb{R}$ be a functional and let $f \in \mathcal{M}(\mathcal{X})^*$. If there exist a Young function ϕ and its complementary function ϕ_Y^* which admits a generalised inverse ϕ_Y^{*-1} such that for every $\nu \in \mathcal{F}$ with $\nu(f) < +\infty$ one has*

$$\nu(f) \leq \phi_Y^{*-1}(\psi(\nu)), \quad (3.96)$$

then

$$\psi^*(\lambda f) \leq \phi(\lambda) \text{ for every } \lambda > 0, \quad (3.97)$$

where $\psi^* : \mathcal{M}(\mathcal{X})^* \rightarrow \mathbb{R}$ is the Legendre-Fenchel dual of ψ .

Proof. Given that ϕ is a Young function, one has that by (Boucheron et al., 2013, Lemma 2.4)

$$\nu(f) \leq \phi_Y^{*-1}(\psi(\nu)) = \inf_{\lambda > 0} \frac{\phi(\lambda) + \psi(\nu)}{\lambda}. \quad (3.98)$$

This means that for every $\lambda > 0$, every ν and every ν -integrable function f ,

$$\phi(\lambda) \geq \lambda \nu(f) - \psi(\nu). \quad (3.99)$$

Taking the supremum with respect to ν one then recovers the statement: $\phi(\lambda) \geq \psi^*(\lambda f)$. \square

In order to see how Theorem 18 and Theorem 20 represent, respectively, the “if” and “only if” part connecting $T_p(c)$ -like inequalities to concentration let us recover Theorem 19. In particular, select

Chapter 3. Independence Vs Dependence

- $\psi^*(\nu) = D(\nu\|\mu)$ for a given μ (consequently, one has that $\psi(\lambda f) = \log \mu(\exp(\lambda f))$, cf. (Dembo and Zeitouni, 2009, Lemma 6.2.13));
- $\phi(\lambda) = \frac{c\lambda^2}{2}$ (which implies $\phi^{*-1}(\kappa) = \sqrt{2c\kappa}$);

Theorem 18 is what allows us to reach Equation (3.95) starting from Equation (3.94). Keeping the same ϕ but inverting the roles of ψ and ψ^* in Theorem 20 is what allows us to reach Equation (3.94) starting from Equation (3.95). Some extra technical steps are necessary in order to bring in Wasserstein Distances (which will be considered in the proof just below). Given the generality of the results we can, as an example, consider a setting similar to Theorem 19 but involving a different divergence than the Kullback-Leibler:

Theorem 21. *Let $\mu \in \mathcal{M}_1(\mathcal{X})$ and $\beta > 1$. There exists a c such that for every λ and every 1-Lipschitz function f*

$$\mu\left(|\lambda f|^\beta\right) \leq (c\lambda)^\beta, \quad (3.100)$$

if and only if, for every $\nu \ll \mu$

$$W_1(\mu, \nu) \leq (\alpha c^\alpha H_\alpha(\nu\|\mu))^\frac{1}{\alpha}, \quad (3.101)$$

where $\alpha = \frac{\beta}{\beta-1}$. Setting $\beta = 2$ we recover the following:

$$W_1(\mu, \nu) \leq \sqrt{c^2 2(\chi^2(\nu\|\mu) + 1)}. \quad (3.102)$$

Proof. Let μ be a probability measure and let $\psi^*(\nu) = H_\alpha(\nu\|\mu)$ with $\alpha > 1$, one has that $\psi(f) = \frac{\mu(|f|^\beta)}{\beta}$ (cf. Equation (2.111)). Moreover, let $\phi(\lambda) = \frac{|c\lambda|^\beta}{\beta}$, one has that $\phi_Y^*(\lambda^*) = \frac{|\lambda^*/c|^\alpha}{\alpha}$ with $\alpha = \frac{\beta}{\beta-1}$. Consequently, for positive κ , $\phi_Y^{*-1}(\kappa) = (\alpha c^\alpha \kappa)^\frac{1}{\alpha}$.

Let f be a function such that $\|f\|_{Lip} \leq 1$ and $\mu(f) = 0$. Assume that for every $\nu \ll \mu$, $W_1(\mu, \nu) \leq (\alpha c^\alpha H_\alpha(\nu\|\mu))^\frac{1}{\alpha}$. By the Kantorovich-Rubenstein dual representation of W_1 one has that

$$W_1(\mu, \nu) = \sup_{f:\|f\|_{Lip} \leq 1} |\mu(f) - \nu(f)|. \quad (3.103)$$

Hence, for every function f such that $\|f\|_{Lip} \leq 1$, $\mu(f) = 0$ one can rewrite Equation (3.102) as follows:

$$\nu(f) \leq (\alpha c^\alpha H_\alpha(\nu\|\mu))^\frac{1}{\alpha} = \phi_Y^{*-1}(H_\alpha(\nu\|\mu)). \quad (3.104)$$

Consequently, by Theorem 20 one has that for every such f

$$\psi^*(\lambda f) = \mu\left(\frac{|\lambda f|^\beta}{\beta}\right) \leq \frac{(c\lambda)^\beta}{\beta} = \phi(\lambda). \quad (3.105)$$

Similarly, assuming Equation (3.105) leads to Equation (3.104) for every such f via Theorem 18. Repeating the same argument with $-f$ one reaches the following statement:

$$|\mu(f) - \nu(f)| \leq (\alpha c^\alpha H_\alpha(\nu\|\mu))^\frac{1}{\alpha}. \quad (3.106)$$

The assumption that $\mu(f) = 0$, can be dropped replacing f with $f - \mu(f)$. Taking then the supremum over all the 1-Lipschitz functions f in Equation (3.106) and using again the Kantorovich-Rubenstein duality formula for Wasserstein Distances one reaches that for every $\nu \ll \mu$,

$$W_1(\mu, \nu) \leq (\alpha c^\alpha H_\alpha(\nu \parallel \mu))^{\frac{1}{\alpha}}.$$

□

Drawing inspiration from Theorem 21 one can consider almost any φ -Divergence. Some restrictions on the possible choices of φ arise naturally in order to have access to both variational representations (cf. (Broniatowski and Keziou, 2010, Theorems 3.3, 3.4)). *I.e.*, to be able to have both Equation (2.111) and Equation (2.112) and guarantee the existence of a unique optimiser (μ -a.e.).

Remark 34. *In the classical $T_1(c)$ -setting one assumes that $\log(\mu(\exp(\lambda f))) \leq (\lambda c)^2/2$ for every f that is 1-Lipschitz and consequently recovers Equation (3.95). Considering Theorem 21 instead, in Equation (3.100) we only asking for the β -th moment of every f to be bounded. Consider $\alpha = 2$, even though it is well known that $\chi^2(\nu \parallel \mu) \geq D(\nu \parallel \mu)$ (leading thus, to a worse bound on W_1), in order to obtain Equation (3.102) we only need to bound the second moment of f with respect to μ and such a bound can exist for functions with unbounded log-moment generating function, which are excluded from a classical $T_1(c)$ setting (more technical details regarding the setting will be presented in Remark 42).*

Theorem 21 highlights the following approach: starting from the variational representation of Wasserstein distances W_p one understands the restriction on the family of functions that needs to be considered (cf. Equation (3.103)). Let us denote such a family with \mathcal{F}_p . The next step is then to fix a measure μ and a functional ψ_μ^* (in Theorem 21, the choice of ψ_μ^* has fallen on the Hellinger integral). The upper-bound that one can provide on $\psi(\lambda f)$ for $f \in \mathcal{F}_p$, characterised by $\phi(\lambda)$ (cf. Equation (3.81), Equation (3.94) and Equation (3.100)), will determine the shape of the $T_p(c)$ -like inequality (through $\phi^{\star^{-1}}$, cf. Equation (3.82), Equation (3.95) and Equation (3.101)), here denoted ϕ_p -inequalities for convenience. Vice versa, bounding a Wasserstein distance W_p through a divergence ψ_μ^* via a ϕ_p -inequality (cf. Equation (3.95), Equation (3.96) and Equation (3.101)) implies a bound on ψ_μ (cf. Equation (3.94), Equation (3.97) and Equation (3.100)) which, in turn, can imply concentration according to ψ_μ , ϕ and \mathcal{F}_p . *E.g.*, if, like in Theorem 19, ψ_μ is the log-moment generating function, the connection to concentration is obvious and is characterised by ϕ for the functions in \mathcal{F}_p .

In particular, with respect to classical $T_p(c)$ inequalities, one has two extra degrees of freedom: ψ_μ^* and ϕ . Changing ψ_μ^* changes the divergence and takes us away from Kullback-Leibler but it also requires us to bound a different functional (cf. Equation (3.100)).

In Theorem 21, both ψ_μ^* and ϕ were different with respect to the usual transportation-cost inequalities. This additional freedom is relevant: trying to show a classical $T_p(c)$ -like inequality for some measure μ , fixing the shape of the inequality to be approximately $W_p(\mu, \nu) \leq \sqrt{kcD(\nu \parallel \mu)}$ (which means, essentially, fixing ϕ) can be impossible (as one might not have that $\psi_\mu(\lambda f) \leq \phi(\lambda)$ for all λ). This strongly depends on the concentration properties of μ or the

Chapter 3. Independence Vs Dependence

family of functions that we have to consider \mathcal{F}_p . Asking for the dual of the Kullback-Leibler divergence to be bounded for every λ and $f \in \mathcal{F}_p$ implies asking for all the moments of f to be bounded as well. Theorem 21, for instance, relaxes this assumption by changing ψ_μ^* and only requires for the β -th moment of (λf) to be bounded.

It is also possible, using such a framework, to maintain the same ψ_μ^* (e.g., the Kullback-Leibler Divergence) but change ϕ and thus ask a behaviour of the log-moment generating function which is different from Gaussian-like, like we will see in the next section.

3.2.2 Beyond Sub-Gaussianity

Consider, for instance, a 0-mean random variable such that, for a given $\theta > 0$ there exists a $K > 0$ such that for all $\lambda > 0$

$$\mu(\exp(\lambda X)) \leq \exp\left((\lambda K)^{\frac{1}{\theta}}\right). \quad (3.107)$$

Remark 35. If $\theta = \frac{1}{2}$ and $K = \frac{\sigma}{\sqrt{2}}$, then one is using the usual characterisation of a σ^2 -sub-Gaussian random variable. One could, in principle, interpolate between lighter and heavier tailed distributions through the parameter θ . Setting $\theta = 1$ one would be tempted to say that Equation (3.107) recovers sub-Exponential random variables. However, notice that such a bound on the moment generating function, for a sub-exponential random variable would only hold for a restricted range of parameters λ (cf. (Vershynin, 2018, Proposition 2.7.1.d)) while, in Equation (3.107), we are asking for the bound to hold for every $\lambda > 0$.

Similarly, one could consider random variables such that

$$\mu\left(\exp(\lambda X)^{\frac{1}{\theta}}\right) \leq \exp\left((\lambda K)^{\frac{1}{\theta}}\right). \quad (3.108)$$

for all the $\lambda < \frac{1}{K}$ and recover the characterisation of sub-Weibull random variables (Vladimirova et al., 2020, Theorem 2.1), correctly interpolating between sub-Exponentials and sub-Gaussians. However, such a characterisation would not be perfectly suitable for a ϕ_p -inequality (or, a correct generalisation of the usual $T_p(c)$ -inequalities beyond σ^2 -sub-Gaussianity) as we will see in a later remark. It does, however, lend itself to a different type of inequalities that involve the Kullback-Leibler Divergences (cf. Section 3.2.4).

Corollary 14. *Let X be a positive random variable over the probability space $(\mathcal{X}, \mathcal{F}, \mu)$. Assume that $\mu(X) = 0$ and that it satisfies Equation (3.107) for every $\lambda > 0$. Let $\phi(\lambda) = (\lambda K)^{\frac{1}{\theta}}$, one has that for every measure ν such that $\nu \ll \mu$*

$$\nu(X) \leq \left(\frac{D(\nu \parallel \mu) K^{\frac{1}{1-\theta}}}{\theta^{\frac{\theta}{1-\theta}} - \theta^{\frac{1}{1-\theta}}} \right)^{1-\theta}. \quad (3.109)$$

Proof. Let us set $\psi_{\mu}^{\star}(\nu) = D(\nu \parallel \mu)$ and $\phi(\lambda) = (\lambda K)^{\frac{1}{\theta}}$. Moreover, one has that

$$\phi^{\star}(\lambda^{\star}) = \left(\frac{\lambda^{\star}}{K} \right)^{\frac{1}{1-\theta}} \left(\theta^{\frac{\theta}{1-\theta}} - \theta^{\frac{1}{1-\theta}} \right) \quad (3.110)$$

and that

$$\phi^{\star-1}(t) = \left(\frac{t K^{\frac{1}{1-\theta}}}{\theta^{\frac{\theta}{1-\theta}} - \theta^{\frac{1}{1-\theta}}} \right)^{1-\theta}. \quad (3.111)$$

The proof then follows from Theorem 18. □

Remark 36. *It is easy to see that setting $\theta = \frac{1}{2}$ and $K = \frac{\sigma}{\sqrt{2}}$ in Equation (3.109) one does indeed recover Equation (3.91). One could provide a version of Corollary 14 that applies to sub-Weibull random variables as characterised by Equation (3.108), however the following additional constraint would be needed $\lambda = \phi^{\star-1}(D(\nu \parallel \mu)) < 1/K$. This allows us to select the λ that leads to the desired expression in Equation (3.86). The corresponding bound would, however, have $\nu\left(X^{\frac{1}{\theta}}\right)$ on the left-hand side and could not lead to a generalisation of the usual Transportation-Cost inequalities. A big advantage of the characterisation of sub-Gaussianity like the one in Equation (3.107) with $\theta = 2$ comes from the lack of the power 2 on the left hand-side of 3.107. This seems not to be possible considering sub-Weibull random variables with a generic parameter θ (Vladimirova et al., 2020, Theorem 2.1).*

3.2.3 Transportation-cost inequalities with Rényi's α -Divergences

Continuing along the lines of transportation-cost inequalities, we will now try to connect (with a slightly less general approach) Rényi's α -Divergences with $0 < \alpha < 1$ to concentration of measure. This section will leverage the variational representation of the Rényi's divergences that stems from the Hellinger Integral, H_{α} (cf. Lemma 5). This representation, to the best of our knowledge, is not necessarily linked to the Legendre-Fenchel dual of D_{α} , which prevents us from directly using the framework described in the previous section (cf. Theorem 20). We can, nonetheless, use similar tools in an ad-hoc fashion and tighten results like Theorem 19. More precisely, given that $D_{\alpha}(\nu \parallel \mu) \leq D(\nu \parallel \mu)$ for $0 < \alpha < 1$, the question that we would like to pose is the following:

Chapter 3. Independence Vs Dependence

“Assuming a bound on the Wasserstein Distance through D_α with $0 < \alpha < 1$ (rather than the Kullback-Leibler Divergence), can one obtain stronger concentration of measure properties?”

More formally, fixed a probability measure μ , if we assume that for every $\nu \ll \mu$

$$W_1(\nu, \mu) \leq \phi^{\star^{-1}}(D_\alpha(\nu \parallel \mu)) \quad (3.112)$$

can we bound $\log \mu(\exp(\lambda f))$ using Equation (2.94) in a way that is different from Equation (3.94)?

Before we start, let us quickly introduce a new object: φ -Entropies.

Definition 22. Let X be a non-negative real-valued random variable with measure ξ defined over a proper probability space and let $\varphi : \mathbb{R}^+ \rightarrow \mathbb{R}$ be a convex function, the φ -entropy of X is defined as follows

$$Ent_\varphi[X] = \xi(\varphi(X)) - \varphi(\xi(X)). \quad (3.113)$$

Remark 37. Such an object is clearly always positive by Jensen's inequality, and is also connected to concentration and the so-called Strong Data Processing Inequalities in Information Theory (Raginsky, 2016).

Similarly to before, from the Kantorovich-Rubenstein's representation of W_1 , assuming Equation (3.112) one can deduce that if ϕ is an Orlicz function, for every f that is 1-Lipschitz then:

$$\nu(f) - \mu(f) \leq \phi_Y^{\star^{-1}}(D_\alpha(\nu \parallel \mu)) = \inf_{\lambda > 0} \frac{\phi(\lambda) + D_\alpha(\nu \parallel \mu)}{\lambda}. \quad (3.114)$$

Consequently, for every $f \in \mathcal{F}_1$ and for every $\lambda > 0$

$$\nu(f) - \mu(f) \leq \frac{\phi(\lambda) + D_\alpha(\nu \parallel \mu)}{\lambda}. \quad (3.115)$$

Thus, for a given f and (assuming without loss of generality that $\mu(f) = 0$) one has that, for every $\nu \ll \mu$ and $\lambda > 0$

$$D_\alpha(\nu \parallel \mu) \geq \lambda \nu(f) - \phi(\lambda). \quad (3.116)$$

Let us now consider the measure $\hat{\nu}$ defined as follows

$$d\hat{\nu} = \frac{g d\mu}{\int g d\mu}, \quad (3.117)$$

where g is a μ -measurable and integrable function. Clearly, $\hat{\nu}$ is a probability measure and it is absolutely continuous with respect to μ by definition. One has that

$$D_\alpha(\hat{\nu} \parallel \mu) = \frac{1}{\alpha-1} \log \int \left(\frac{d\hat{\nu}}{d\mu} \right)^\alpha d\mu \quad (3.118)$$

$$= \frac{1}{\alpha-1} \log \int \left(\frac{g}{\int g d\mu} \right)^\alpha d\mu \quad (3.119)$$

$$= \frac{1}{\alpha-1} \log \int g^\alpha d\mu - \frac{\alpha}{\alpha-1} \log \int g d\mu. \quad (3.120)$$

Choosing $g = \exp(\lambda f)$ in Equation (3.117) and setting $\nu = \hat{\nu}$ in Equation (3.116) one has the following:

$$\frac{1}{\alpha-1} \log \int \exp(\lambda f)^\alpha d\mu - \frac{\alpha}{\alpha-1} \log \int \exp(\lambda f) d\mu \geq \lambda \nu(f) - \phi(\lambda). \quad (3.121)$$

Our purpose is to provide a bound on $\log(\mu(\exp(\lambda f)))$ thus, re-arranging one obtains

$$\frac{\alpha}{\alpha-1} \log \int \exp(\lambda f) d\mu \leq \frac{1}{\alpha-1} \log \int \exp(\lambda f)^\alpha d\mu - \lambda \nu(f) + \phi(\lambda). \quad (3.122)$$

Let us now manipulate the right-hand side of Equation (3.122) in order to get a more meaningful expression. In particular, one has that:

$$\frac{1}{\alpha-1} \log \int \exp(\lambda f)^\alpha d\mu = \frac{1}{\alpha-1} \log \int \exp(\lambda f)^{\alpha-1} \exp(\lambda f) d\mu \quad (3.123)$$

$$= \frac{1}{\alpha-1} \log \int \exp(\lambda f)^{\alpha-1} \left(\int \exp(\lambda f) d\mu \right) d\nu \quad (3.124)$$

$$= \frac{1}{\alpha-1} \log \int \exp(\lambda f)^{\alpha-1} d\nu + \frac{1}{\alpha-1} \log \int \exp(\lambda f) d\mu. \quad (3.125)$$

Substituting back in Equation (3.122) one obtains the following

$$\log \int \exp(\lambda f) d\mu \leq \frac{1}{\alpha-1} \log \int \exp(\lambda f)^{\alpha-1} d\nu - \nu(\lambda f) + \phi(\lambda). \quad (3.126)$$

It is easy to see that the right-hand side of Equation (3.126) contains a φ -Entropy, *i.e.*,

$$\frac{1}{\alpha-1} \log \int \exp(\lambda f)^{\alpha-1} d\nu - \nu(\lambda f) = - \left(\nu(\lambda f) - \frac{1}{\alpha-1} \log \nu(\exp(\lambda f)^{\alpha-1}) \right) \quad (3.127)$$

$$= -\text{Ent}_\varphi[\exp(\lambda f)^{(\alpha-1)}], \quad (3.128)$$

with $\varphi(x) = \frac{1}{\alpha-1} \log(x)$, a convex function for $0 < \alpha < 1$. Going back to Equation (3.126), one has that for every 1-Lipschitz function f and every $\lambda > 0$:

$$\log \int \exp(\lambda f) d\mu \leq -\text{Ent}_\varphi[\exp(\lambda f)^{(\alpha-1)}] + \phi(\lambda) \leq \phi(\lambda). \quad (3.129)$$

Hence, one can provide a tighter-bound on the cumulant-generating function of (λf) with

Chapter 3. Independence Vs Dependence

respect to a generalised version of Equation (3.94) (e.g., Theorem 20 with $\psi(v) = D(v\|\mu)$). Setting $\phi(\lambda) = \left(\frac{c\lambda^2}{2}\right)$ and assuming that for every $v \ll \mu$

$$W_1(v, \mu) \leq \sqrt{2cD_\alpha(v\|\mu)} \quad (3.130)$$

(something stronger than Equation (3.95), as $D_\alpha(v\|\mu) < D(v\|\mu)$ for $\alpha < 1$) it is possible to recover the following:

$$\log(\mu(\exp(\lambda f))) \leq -\text{Ent}_\phi[\exp(\lambda f)^{(\alpha-1)}] + \phi(\lambda) \quad (3.131)$$

i.e., an upper-bound on $\log(\mu(\exp(\lambda f)))$ that is potentially tighter than Equation (3.94), as $\text{Ent}_\phi[\exp(\lambda f)^{(\alpha-1)}] \geq 0$.

3.2.4 Orlicz spaces, tails of random-variables and the Kullback-Leibler Divergence

Stepping away from the Transportation-cost inequality framework (but maintaining the same tools), an equivalent characterisation of the sub-Weibull (and sub-Gaussian) random variables through Luxemburg norms can lead to immediate connections with the Kullback-Leibler Divergence. In particular, consider sub-Gaussian random variables. Given a random variable X such that $\mu(X) = 0$, the usual requirement would be that, if there exists a \tilde{K} such that for every $\lambda \in \mathbb{R}$

$$\log \mu(\exp(\lambda X)) \leq \tilde{K}^2 \lambda^2, \quad (3.132)$$

then X is sub-Gaussian. However, an alternative (and equivalent) characterisation would be the following (cf. (Vershynin, 2018, Proposition 2.5.2.iv)): if there exists a K such that

$$\mu\left(\exp\left(\frac{X^2}{K^2}\right)\right) \leq 2, \quad (3.133)$$

then X is said to be sub-Gaussian. One can then characterise the smallest constant K such that Equation (3.133) holds true:

$$\inf\left\{K > 0 : \mu\left(\exp\left(\frac{X^2}{K^2}\right)\right) \leq 2\right\}. \quad (3.134)$$

It is easy to see how Equation (3.134) represents the Luxemburg norm (cf. Equation (1.17)) of X with $\psi(x) = e^{x^2} - 1$.

Remark 38. In the literature, $\|X\|_{L_\psi^L(\mu)}$ goes under the name of sub-Gaussian norm (Vershynin, 2018, Definition 2.5.6).

Continuing the parallel, one has that $\psi(x) = e^{x^2} - 1$ is indeed an Orlicz function and the space of random variables that satisfy Equation (3.133) under μ for some K is precisely the Orlicz space $L_\psi(\mu)$ (cf. Equation (1.14)). Given a measure μ denote with $\mathcal{M}(\mathcal{X}, \mu)$ the space of measures on \mathcal{X} that are absolutely continuous with respect to μ and with $\mathcal{M}_1(\mathcal{X}, \mu)$ the space of probability measures on \mathcal{X} that are absolutely continuous with respect to μ . We are now

ready to prove our result.

Theorem 22. Consider the space $(\mathcal{X}, \mathcal{F}, \mu)$ and let $\psi : \mathcal{X} \rightarrow \mathbb{R}^+$ be such that $\psi(x) = e^{x^2} - 1$. Assume that $f \in (L_\psi(\mu) \cap B(\mathcal{X}))$ then, for every $\nu \in \mathcal{M}_1(\mathcal{X}, \mu)$ one has that

$$\nu(f^2) \leq \|f\|_{L_\psi^L(\mu)}^2 \cdot (D(\nu\|\mu) + \log 2) \quad (3.135)$$

Proof. Given that $f \in L_\psi(\mu)$ one has that $\|f\|_{L_\psi^L(\mu)} < \infty$ and, consequently

$$\mu \left(\exp \left(\frac{f^2}{\|f\|_{L_\psi^L(\mu)}^2} \right) \right) \leq 2. \quad (3.136)$$

For every $\nu \in \mathcal{M}_1(\mu)$ and $f \in (L_\psi(\mu) \cap B(\mathcal{X}))$ one has by Equation (2.3):

$$D(\nu\|\mu) = \sup_g \nu(g) - \log \mu(\exp(g)) \quad (3.137)$$

$$\geq \nu \left(\frac{f^2}{\|f\|_{\psi, \mu}^2} \right) - \log 2. \quad (3.138)$$

Meaning that

$$\nu(f^2) \leq \|f\|_{\psi, \mu}^2 \cdot (D(\nu\|\mu) + \log 2). \quad (3.139)$$

□

Similarly, one can characterise a connection between sub-Weibull random variables and the Kullback-Leibler Divergence. Indeed, consider $\psi_\theta(x) = \exp(|x|^{\frac{1}{\theta}}) - 1$, it is easy to see that ψ_θ is an Orlicz function for $0 < \theta \leq 1$. Similarly to before one can consider the space $L_{\psi_\theta}(\mu)$ and easily see that this is the space of θ -sub-Weibull random variables with finite Luxemburg norm $\|\cdot\|_{L_{\psi_\theta}^L(\mu)}$ (cf. (Vladimirova et al., 2020, Theorem 2.1.4)). One can then show the following:

Theorem 23. Consider the space $(\mathcal{X}, \mathcal{F}, \mu)$, let $0 < \theta \leq 1$ and let $\psi_\theta : \mathcal{X} \rightarrow \mathbb{R}^+$ be such that $\psi_\theta(x) = \exp(|x|^{\frac{1}{\theta}}) - 1$. Assume that $f \in (L_{\psi_\theta}(\mu) \cap B(\mathcal{X}))$ then, for every $\nu \in \mathcal{M}_1(\mathcal{X}, \mu)$ one has that

$$\nu(f^{\frac{1}{\theta}}) \leq \|f\|_{\psi_\theta, \mu}^{\frac{1}{\theta}} \cdot (D(\nu\|\mu) + \log 2) \quad (3.140)$$

Proof. The proof follows from the one of Theorem 22 with g selected to be $|f|^{\frac{1}{\theta}} / \|f\|_{L_{\psi_\theta}^L(\mu)}^{\frac{1}{\theta}}$. □

3.3 Application: Hypothesis Testing

An almost immediate application of the results presented in the previous sections comes from the realm of Statistics and Hypothesis Testing. Assume the setting of binary Hypothesis Testing. Assume we are given n iid random variables X_1, \dots, X_n . The idea is to infer whether $X_i \sim \mu$ or $X_i \sim \nu$ looking at the samples X_1, \dots, X_n . We will call the hypothesis that $X_i \sim \mu$ the null

Chapter 3. Independence Vs Dependence

hypothesis H_0 and the hypothesis that $X_i \sim \nu$ the alternative hypothesis H_1 . A classical setting is the following: upon observing the n samples X_1, \dots, X_n one can construct the empirical distribution $\xi_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x=X_i\}}$ and then use it to pick H_0 over H_1 or vice versa:

- if ξ_n is closer to μ than ν with respect to the Kullback-Leibler Divergence, then H_0 is chosen;
- if ξ_n is closer to ν than μ with respect to the Kullback-Leibler Divergence, then H_1 is chosen.

Given a constant $c \geq 0$, the test function is then the following $Y_c(X_1^n) = \mathbb{1}_{\{D(\xi_n \parallel \mu) - D(\xi_n \parallel \nu) > c\}}$. The chosen hypothesis is thus $H_{Y_c(X_1^n)}$.

Indeed,

- if $D(\xi_n \parallel \mu) - D(\xi_n \parallel \nu) > c$ then ξ_n is closer to ν than to μ and we pick the hypothesis 1 which is the outcome of $Y_c(X_1^n)$;
- if $D(\xi_n \parallel \mu) - D(\xi_n \parallel \nu) \leq c$ then ξ_n is closer to μ than to ν and we pick the hypothesis 0 which is the outcome of $Y_c(X_1^n)$.

Consider a discrete setting and denote with m and n the densities of μ and ν with respect to the counting measure, then one can show that

$$D(\xi_n \parallel \mu) - D(\xi_n \parallel \nu) = \frac{1}{n} \log \left(\frac{\prod_{i=1}^n m(X_i)}{\prod_{i=1}^n n(X_i)} \right). \quad (3.141)$$

In this case, $Y_c(X_1^n)$ is simply the likelihood ratio test, which is known to be optimal (Cover and Thomas, 2006, Theorem 11.7.1). The objects that one would like to control, in this setting, are typically the type I error probability and the type II error probability:

1. the type I error probability (denoted with ζ) occurs whenever we pick H_1 but the samples are distributed according to μ ;
2. the type II error probability (denoted with δ) occurs whenever we pick H_0 but the samples are distributed according to ν .

Hence, we can denote, in the usual notation:

$$\zeta = \mu^{\otimes n}(Y_c) = \mu^{\otimes n}(D(\xi_n \parallel \mu) - D(\xi_n \parallel \nu) > c) \quad (3.142)$$

$$\delta = \nu^{\otimes n}(1 - Y_c) = 1 - \nu^{\otimes n}(Y_c) = 1 - \nu^{\otimes n}(D(\xi_n \parallel \mu) - D(\xi_n \parallel \nu) > c). \quad (3.143)$$

Well-known negative results state that controlling both types of errors in an arbitrary fashion is not possible. In particular, if one requires ζ to decay exponentially with the number of samples n then it is possible to bound δ away from 0. For instance, using the Total Variation distance (and its variational representation) together with Pinsker's inequality, it is possible to show that:

$$\zeta + \delta \geq 1 - \sqrt{\frac{2}{n} D(\nu \parallel \mu)}. \quad (3.144)$$

The immediate application of our results comes from the fact that Y_c is clearly bounded. Being bounded, it is possible to show a bound on its log-moment generating function through Hoeffding's Lemma. In particular one has that:

$$\log(\mu(e^{\lambda(Y_c - \mu(Y_c))})) \leq \frac{\lambda^2}{8} = \phi(\lambda) \quad (3.145)$$

and using this in Theorem 18 with $\psi^*(v) = D(v\|\mu)$ (or simply Corollary 13) one recovers that:

$$(\mu^{\otimes n}(Y_c) - \nu^{\otimes n}(Y_c)) \leq \sqrt{\frac{n}{2} D(v\|\mu)}, \quad (3.146)$$

which can be re-written as follows:

$$\zeta + \delta = \mu^{\otimes}(Y_c) + 1 - \nu^{\otimes}(Y_c) = 1 - (\nu^{\otimes}(Y_c) - \mu^{\otimes}(Y_c)) \geq 1 - \sqrt{\frac{n}{2} D(v\|\mu)}. \quad (3.147)$$

Remark 39. *It is possible to derive Equation (3.147) from Theorem 18 as Pinsker's Inequality is essentially just a $T_1(1/4)$ -Inequality in the Transportation-Cost Inequality framework (Raginsky and Sason, 2014, Page 112). Setting $\psi_\mu^*(v) = D(v\|\mu)$ and assuming Equation (3.145) sets Theorem 18 in the classical Transportation-Cost Inequality framework, i.e., Theorem 19.*

This setting lends itself to tighter bounds as well. Let $\psi(x) = e^{x^2} - 1$ be the ‘‘sub-Gaussian’’ Orlicz function. One can show that given a function $f \in L_\psi(\mu)$ there exists a constant $C > 0$ such that $C\|f\|_{L_\psi^L(\mu)}$ is the tightest sub-Gaussian constant of f with respect to μ in the sense that (Vershynin, 2018, Equation (2.16))

$$\mu(\exp(\lambda f)) \leq \exp\left(C\lambda^2 \|f\|_{L_\psi^L(\mu)}^2\right), \forall \lambda \in \mathbb{R}. \quad (3.148)$$

Thus, one has that:

$$\mu(\exp(\lambda Y_c)) \leq \exp\left(C\lambda^2 \|Y_c\|_{L_\psi^L(\mu)}^2\right), \forall \lambda \in \mathbb{R}. \quad (3.149)$$

Consequently, using Theorem 18, one can provide the following bound:

$$|\mu^{\otimes n}(Y_c) - \nu^{\otimes n}(Y_c)| \leq \sqrt{4Cn \|Y_c\|_{L_\psi^L(\mu)}^2 D(v\|\mu)}, \quad (3.150)$$

or, alternatively:

$$\zeta + \delta \geq 1 - \sqrt{4Cn \|Y_c\|_{L_\psi^L(\mu)}^2 D(v\|\mu)}. \quad (3.151)$$

Similarly, one could pick $\psi_\beta(x) = \frac{|x|^\beta}{\beta}$ with $\beta > 1$. This is also an Orlicz function and one can consider the corresponding Orlicz space $L_{\psi_\beta}(\mu)$ endowed with the Luxemburg norm $\|\cdot\|_{L_{\psi_\beta}^L(\mu)}$.

Chapter 3. Independence Vs Dependence

Under this setting, if $f \in L_{\psi_\beta}(\mu)$ is such that $\|f\|_{L_{\psi_\beta}^L(\mu)} < \infty$ then one has that, if $\lambda > 0$ then

$$\frac{\mu(|\lambda f|^\beta)}{\beta} \leq (\lambda \|f\|_{L_{\psi_\beta}^L(\mu)})^\beta = \phi(\lambda). \quad (3.152)$$

The function $\phi(\lambda)$ is also a Young function if one considers the restriction to \mathbb{R}^+ . Setting $\sigma = \|f\|_{L_{\psi_\beta}^L(\mu)}$ and $c(\beta, \sigma) = (\beta\sigma^\beta)^{-\frac{1}{\beta-1}}$, one has that

$$\phi^*(\lambda^*) = \frac{1}{\alpha} (\lambda^*)^\alpha c(\beta, \sigma) \quad (3.153)$$

with $\alpha = \frac{\beta}{\beta-1}$ and, consequently,

$$\phi^{*-1}(k) = \left(\frac{k\alpha}{c(\beta, \sigma)} \right)^{\frac{1}{\alpha}}. \quad (3.154)$$

We are thus in the setting of Theorem 18 and one can use it to show that:

$$|\mu^{\otimes n}(Y_c) - \nu^{\otimes n}(Y_c)| \leq \left(\alpha \frac{H_\alpha(\nu^{\otimes n} \|\mu^{\otimes n}\|)}{c(\beta, \sigma)} \right)^{\frac{1}{\alpha}} \quad (3.155)$$

$$= (\alpha H_\alpha(\nu^{\otimes n} \|\mu^{\otimes n}\|)^{\frac{1}{\alpha}} \sigma \beta^{\frac{1}{\beta}}), \quad (3.156)$$

which can then be translated in the following lower-bound:

$$\zeta + \delta \geq 1 - (\alpha H_\alpha(\nu^{\otimes n} \|\mu^{\otimes n}\|)^{\frac{1}{\alpha}} \sigma \beta^{\frac{1}{\beta}}) \quad (3.157)$$

$$= 1 - (H_\alpha(\nu^{\otimes n} \|\mu^{\otimes n}\|)^{\frac{1}{\alpha}} \frac{\sigma \beta}{(1-\beta)^{\frac{1}{\beta}}}) \quad (3.158)$$

$$= 1 - \exp\left(\frac{(\alpha-1)}{\alpha} D_\alpha(\nu^{\otimes n} \|\mu^{\otimes n}\|)\right) \frac{\sigma \beta}{(1-\beta)^{\frac{1}{\beta}}} \quad (3.159)$$

$$= 1 - \exp\left(\frac{n(\alpha-1)}{\alpha} D_\alpha(\nu \|\mu)\right) \frac{\sigma \beta}{(1-\beta)^{\frac{1}{\beta}}} \quad (3.160)$$

$$= 1 - \exp\left(\frac{n(\alpha-1)}{\alpha} D_\alpha(\nu \|\mu) + \log\left(\frac{\sigma \beta}{(1-\beta)^{\frac{1}{\beta}}}\right)\right) \quad (3.161)$$

$$= 1 - \exp\left(\frac{n(\alpha-1)}{\alpha} D_\alpha(\nu \|\mu) + \log\left(\sigma \alpha (1-\beta)^{\frac{1}{\alpha}}\right)\right) \quad (3.162)$$

$$= 1 - \exp\left(\frac{n(\alpha-1)}{\alpha} D_\alpha(\nu \|\mu) + \log\left(\|Y_c\|_{L_{\psi_\beta}^L(\mu)} \alpha (1-\beta)^{\frac{1}{\alpha}}\right)\right). \quad (3.163)$$

3.3.1 Bounding the ratio of errors in Hypothesis Testing

So far we focused on bounding $\mu(f) - \nu(f)$ but using the tools described in this chapter, one could also bound $\mu(f)/\nu(f)$ in a hypothesis testing framework. Considering the same setting

as before but using, for instance, Equation (2.92) one has that, for $\alpha > 1$:

$$\nu(Y_c) \leq H_\alpha(\nu\|\mu)^{\frac{1}{\alpha}} \mu(Y_c^\beta)^{\frac{1}{\beta}}, \quad (3.164)$$

given that $Y_c \in \{0, 1\}$ we have that

$$\delta^c = 1 - \delta \leq H_\alpha(\nu\|\mu)^{\frac{1}{\alpha}} \zeta^{\frac{1}{\beta}} \quad (3.165)$$

which, after re-arranging becomes

$$\frac{\zeta^{\frac{1}{\beta}}}{\delta^c} \geq H_\alpha(\nu\|\mu)^{-\frac{1}{\alpha}}. \quad (3.166)$$

Why is this also a negative result? Because if we assume that ζ (the Type I error probability) is exponentially decaying with the number of samples n with rate R , *i.e.*,

$$\mu^{\otimes n}(Y_c) \leq \exp(-nR), \quad (3.167)$$

then Equation (3.164) tells us that

$$\nu^{\otimes n}(Y_c) \leq \exp\left(-n \frac{\alpha-1}{\alpha} (R - D_\alpha(\nu\|\mu))\right). \quad (3.168)$$

and since the Type II error probability δ is equal to $1 - \nu^{\otimes n}(Y_c)$ one has that

$$1 - \delta \leq \exp\left(-n \frac{\alpha-1}{\alpha} (R - D_\alpha(\nu\|\mu))\right). \quad (3.169)$$

Re-arranging Equation (3.169), one obtains the following:

$$\delta \geq 1 - \exp\left(-n \frac{\alpha-1}{\alpha} (R - D_\alpha(\nu\|\mu))\right). \quad (3.170)$$

What we can extrapolate from this is that: if, for a given α

$$R \geq D_\alpha(\nu\|\mu), \quad (3.171)$$

then δ will approach 1 exponentially fast. Once again, there is a trade-off between the quantities here. One would be tempted to pick α as small as possible so that the condition in Equation (3.171) is more easily satisfied. That would, however, also affect the rate of convergence to 1 as the smaller α is, the smaller the multiplicative factor $\frac{\alpha-1}{\alpha}$ becomes and, consequently, the slower the convergence to 0 of the right-hand side of Equation (3.169). The approach presented here can be extended to Sibson's α -Mutual Information (both conditionals and not) and can be applied to hypothesis testing settings. As we have seen in Section 3.1.5, to each of these information-measures one can associate a Hölder's-like inequality which will be pivotal in introducing the information-measure in the hypothesis testing framework (like in Equation (3.164)). This idea is explored further in (Esposito et al., 2021b).

Appendix

3.A Hölder's inequality and information measures

Hölder's inequality represents the foundations of the variational representations for the information measures that are related to $L^\alpha(\cdot)$ -norms. Indeed, it is what allowed us to provide a variational representation for Rényi's Entropy (cf. Section 2.2.2), for the Hellinger Integral and, in a way, for the Sibson's α -Mutual Information (cf. Section 3.1.2). In particular, the starting point is the pairing. Given a function f and a measure μ the key step was the following re-writing:

$$\nu(g) = \mu \left(g \frac{d\nu}{d\mu} \right). \quad (3.172)$$

If ν is a generic measure that belongs to $\mathcal{M}(\mathcal{X}, \mu)$ (space of measures on \mathcal{X} that are absolutely continuous with respect to μ) then this leads to a functional inequality involving the Hellinger Integral (cf. Definition 19) or the Hellinger Divergence of order α (Liese and Vajda, 1987, Definition 2.10). *I.e.*, if $\alpha > 1$ then one has that for every positive-valued function f :

$$\left((\alpha - 1) \mathcal{H}_\alpha(\nu \parallel \mu) + 1 \right)^{\frac{1}{\alpha}} = H_\alpha^\alpha(\nu \parallel \mu) = \left\| \frac{d\nu}{d\mu} \right\|_{L^\alpha(\mu)} \geq \frac{\nu(g)}{\|g\|_{L^\beta(\mu)}}, \quad (3.173)$$

where $\beta = \frac{\alpha}{\alpha-1}$ is the Hölder's conjugate of α . Equality is met if $f = \left(\frac{d\nu}{d\mu} \right)^{\alpha-1}$ μ -a.e. If $\alpha < 1$ then Equation (3.173) holds with a reversed inequality sign.

Otherwise, if ν is absolutely continuous with the counting (Lebesgue) measure and μ is selected to be the counting (Lebesgue) measure, then Hölder's inequality leads us to the variational representation of Rényi's Entropy (cf. Equation (2.52)), *i.e.*:

$$H_\alpha(\nu) = \inf_g \log \|g\|_{L^\beta(\mu)}^\beta - \beta \log(\nu(g)). \quad (3.174)$$

If we instantiate Equation (3.172) to $g = \mathbb{1}_E$, *i.e.*

$$\nu(E) = \nu \left(\mathbb{1}_E \frac{d\nu}{d\mu} \right) \quad (3.175)$$

we can connect probabilities of the same event under different (but absolutely continuous with respect to each other) measures. This led to the results presented in Section 3.1.2 and Sec-

tion 3.1.5. Moreover, to conclude this discussion, we will once again underline the connection between Hölder's inequality and Legendre-Fenchel duality, recovering Equation (3.173) (*i.e.*, Corollary 6) from Theorem 14 for two general measures (rather than a joint and product of marginals). For illustrative purposes we will present such a proof in detail, showing how the various results in Section 3.1 can all be derived from the same proof technique. Given a function g and setting $f = \lambda g$ in Equation (2.111) with $\varphi(x) = \frac{|x|^\alpha}{\alpha}$ one has that for every $\lambda > 0$:

$$v(|g|) \leq \frac{\frac{1}{\alpha} H_\alpha(v\|\mu) + \frac{\mu(|g|^\beta)}{\beta} \lambda^\beta}{\lambda} \quad (3.176)$$

$$= \mu(|g|^\beta) \frac{\frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)\alpha} + \frac{\lambda^\beta}{\beta}}{\lambda} \quad (3.177)$$

where $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Continuing the parallel with the proof of Theorem 14, we have that $\vartheta(\lambda) = \varphi^*(\lambda) = \frac{\lambda^\beta}{\beta}$ and that $c = \frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)\beta}$. Thus, we need to select $\lambda = \vartheta'^{-1}\left(\vartheta^{*-1}\left(\frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)\alpha}\right)\right)$. In particular, $\vartheta^*(y) = \phi(y) = \frac{|y|^\alpha}{\alpha}$ and thus $\vartheta^{*-1}(x) = (\alpha x)^{\frac{1}{\alpha}}$ for x positive. Moreover, one has that $\vartheta'(\lambda) = \lambda^{\beta-1}$ and, consequently, $\vartheta'^{-1}(z) = z^{\frac{1}{\beta-1}}$. Putting everything together we have that

$$\lambda = \vartheta'^{-1}\left(\vartheta^{*-1}\left(\frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)\alpha}\right)\right) \quad (3.178)$$

$$= \frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)^{\frac{1}{\alpha(\beta-1)}}}. \quad (3.179)$$

Substituting λ in Equation (3.177) and using the fact that $\frac{1}{\alpha} + \frac{1}{\beta} = 1 \implies \alpha = \frac{\beta}{\beta-1}$ we retrieve

$$v(|g|) = \mu\left(|g| \frac{dv}{d\mu}\right) \leq \mu(|g|^\beta) \frac{\frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)\alpha} + \frac{1}{\beta} \frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)^{\frac{\beta}{\alpha(\beta-1)}}}}{\frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)^{\frac{1}{\alpha(\beta-1)}}}} \quad (3.180)$$

$$= \mu(|g|^\beta) \frac{\frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)\alpha} + \frac{H_\alpha(v\|\mu)}{\beta\mu(|g|^\beta)}}{\frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)^{\frac{1}{\alpha(\beta-1)}}}} \quad (3.181)$$

$$= \mu(|g|^\beta) \frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)} \left(\frac{\frac{1}{\beta} + \frac{1}{\alpha}}{\frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)^{\frac{1}{\alpha(\beta-1)}}}} \right) \quad (3.182)$$

$$= H_\alpha(v\|\mu) \frac{\mu(|g|^\beta)^{\frac{1}{\alpha(\beta-1)}}}{H_\alpha(v\|\mu)} \quad (3.183)$$

$$= \mu(|g|^\beta)^{\frac{1}{\beta}} H_\alpha(v\|\mu)^{1-\frac{1}{\beta}} \quad (3.184)$$

$$= \mu(|g|^\beta)^{\frac{1}{\beta}} H_\alpha(v\|\mu)^{\frac{1}{\alpha}} \quad (3.185)$$

$$= \| |g| \|_{L^\beta(\mu)} \left\| \frac{dv}{d\mu} \right\|_{L^\alpha(\mu)}. \quad (3.186)$$

This is equivalent to setting $\varphi(x) = \frac{|x|^\alpha}{\alpha}$ in Theorem 13, *i.e.*,

$$v(|g|) = \mu \left(|g| \frac{dv}{d\mu} \right) \leq \mu(|g|^\beta) \varphi^{\star-1} \left(\frac{H_\alpha(v\|\mu)}{\alpha \mu(|g|^\beta)} \right) \quad (3.187)$$

$$= \mu(|g|^\beta) \left(\frac{H_\alpha(v\|\mu)}{\mu(|g|^\beta)} \right)^{\frac{1}{\alpha}} \quad (3.188)$$

$$= \mu(|g|^\beta)^{\frac{1}{\beta}} H_\alpha(v\|\mu)^{\frac{1}{\alpha}} \quad (3.189)$$

$$= \|g\|_{L^\beta(\mu)} \left\| \frac{dv}{d\mu} \right\|_{L^\alpha(\mu)}, \quad (3.190)$$

which represents Equation (3.173): yet another incarnation of Roger-Hölder's inequality. Considering product measures and applying the inequality twice leads us to Theorem 15.

3.B Alternative proof of Theorem 13

In this section we will present a simpler proof of Theorem 13, hereby re-stated for reference **Theorem**. *Let $\varphi : [0, +\infty) \rightarrow \mathbb{R}$ be a convex function such that $\varphi(1) = 0$, and assume φ is non-decreasing on $[0, +\infty)$. Suppose also that φ admits a generalized inverse, defined as $\varphi^{-1}(y) = \inf\{t \geq 0 : \varphi(t) > y\}$. Consider the measurable space $(\mathcal{Z}, \mathcal{F})$, and two measures μ and ν defined on the space. Given an event $E \in \mathcal{F}$, we have that if $\nu \ll \mu$*

$$\nu(E) \leq \mu(E) \cdot \varphi^{-1} \left(\frac{D_\varphi(\nu\|\mu) + (1 - \mu(E))\varphi^*(0)}{\mu(E)} \right), \quad (3.191)$$

where φ^* is the Legendre-Fenchel dual of φ . Moreover, if $\varphi^*(0) \leq 0$, the bound simplifies to

$$\nu(E) \leq \mu(E) \cdot \varphi^{-1} \left(\frac{D_\varphi(\nu\|\mu)}{\mu(E)} \right). \quad (3.192)$$

Proof. Let us denote with $n = \nu(E)$, $m = \mu(E)$, $n^c = 1 - \nu(E)$, $m^c = 1 - \mu(E)$. For every $y \geq 0$ we have

$$D_\varphi(\nu\|\mu) \stackrel{(a)}{\geq} D_\varphi(\text{Ber}(n)\|\text{Ber}(m)) = m\varphi\left(\frac{n}{m}\right) + m^c\varphi\left(\frac{n^c}{m^c}\right) \quad (3.193)$$

$$\stackrel{(b)}{\geq} m\varphi\left(\frac{n}{m}\right) + m^c\left(\frac{n^c}{m^c}y - \varphi^*(y)\right), \quad (3.194)$$

where (a) follows from the Data-Processing Inequality for φ -divergences and (b) follows from Young's inequality. Choosing $y = 0$ in Equation (3.194) and re-arranging the terms we retrieve:

$$\frac{D_\varphi(\nu\|\mu) + n^c\varphi^*(0)}{m} \geq \varphi\left(\frac{n}{m}\right) \iff m\varphi^{-1}\left(\frac{D_\varphi(\nu\|\mu) + n^c\varphi^*(0)}{m}\right) \geq n. \quad (3.195)$$

□

The Applications **Part II**

4 Learning Theory

The primary purpose of this chapter¹ is to apply the results presented so far to a Learning Theory framework. More precisely, we will bound the generalisation error of a learning algorithm, and it can be done in two ways:

- bounding the **expected** generalisation error;
- bounding the **probability** of having a large generalisation error.

When bounding the expected value of the generalisation error, we will use the results presented in Section 3.2. When bounding the probability of having a large generalisation error, we will leverage the results presented in Section 3.1.

4.1 Background

This section will provide some basic background knowledge on learning algorithms and concepts like the generalisation error. We are mainly interested in supervised learning, where the algorithm learns a *classifier* by looking at points in a proper space and the corresponding labels.

More formally, suppose we have an instance space \mathcal{Z} and a hypothesis space \mathcal{H} . The hypothesis space is a set of functions that, given a data point $s \in \mathcal{Z}$, outputs the corresponding label $y \in \mathcal{Y}$. Suppose we are given a training data set $\mathcal{Z}^n \ni S = \{z_1, \dots, z_n\}$ made of n points sampled in an i.i.d. fashion from some distribution \mathcal{P} . Given some $n \in \mathbb{N}$, a learning algorithm is a (possibly stochastic) mapping $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{H}$ that given as an input a finite sequence of points $S \in \mathcal{Z}^n$ outputs some classifier $h = \mathcal{A}(S) \in \mathcal{H}$. In the most straightforward setting, we can think of \mathcal{Z} as a product between the space of data points and the space of labels, *i.e.*, $\mathcal{Z} = \mathcal{D} \times \mathcal{C}$ and suppose that \mathcal{A} is fed with n pairs data-label $(d, c) \in \mathcal{Z}$. In this work, we will view \mathcal{A} as a family of conditional distributions $\mathcal{P}_{H|S}$ and provide a stochastic analysis of its generalisation capabilities using the information measures presented so far. The goal is to

¹The content of this chapter (excluding Section 4.2) has appeared in the IEEE Transactions on Information Theory 2021, Volume: 67, Issue: 8 (Esposito et al., 2021a).

generate a hypothesis $h : \mathcal{D} \rightarrow \mathcal{C}$ that has good performance on both the training set and newly sampled points from \mathcal{X} . To ensure said property, the concept of generalisation error was introduced.

Definition 23. Let \mathcal{P} be some distribution over \mathcal{Z} . Let $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}$ be a loss function. The error (or risk) of a prediction rule h with respect to \mathcal{P} is defined as

$$L_{\mathcal{P}}(h) = \mathcal{P}(\ell(h, Z)), \quad (4.1)$$

while, given a sample $S = (z_1, \dots, z_n)$, the empirical error of h with respect to S is defined as

$$L_S(h) = \frac{1}{n} \sum_{i=1}^n \ell(h, z_i). \quad (4.2)$$

Moreover, given a learning algorithm $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{H}$, its generalisation error with respect to S is defined as:

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) = L_{\mathcal{P}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S)). \quad (4.3)$$

The definition just stated considers general loss functions. An important instance for the case of supervised learning is the 0-1 loss. Suppose again that $\mathcal{Z} = \mathcal{D} \times \mathcal{C}$ and that \mathcal{H} is a family of hypotheses, *i.e.*, functions $h : \mathcal{D} \rightarrow \mathcal{C}$. Given a couple $(d, c) \in \mathcal{Z}$ and a hypothesis $h \in \mathcal{H}$, the 0 – 1 loss is defined as follows:

$$\ell(h, (d, c)) = \mathbb{1}_{h(d) \neq c}. \quad (4.4)$$

The corresponding errors become:

$$L_{\mathcal{P}}(h) = \mathcal{P}(\mathbb{1}_{h(d) \neq c}) = \mathcal{P}(h(d) \neq c). \quad (4.5)$$

and

$$L_S(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{h(d_i) \neq c_i}. \quad (4.6)$$

4.2 The expected generalisation error

Given a learning algorithm \mathcal{A} , one has that the expected generalisation error can be written as (Xu and Raginsky, 2017c, Equation (9)):

$$\mathcal{P}_{SH}(\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S)) = \mathcal{P}_{SH}(\ell(H, S)) - \mathcal{P}_S \mathcal{P}_H(\ell(H, S)), \quad (4.7)$$

i.e., it represents the difference of the expectations of the same function with respect to two different probability measures: the joint \mathcal{P}_{SH} and the corresponding product of the marginals $\mathcal{P}_S \mathcal{P}_H$. Hence, general bounds can be provided using Theorem 18 or the well-known Corollary 13 (cf. (Russo and Zou, 2016; Xu and Raginsky, 2017c)). Given a convex functional φ , we will consider the following family of functionals over measures $\psi_{\mu}^{\star}(\cdot) = D_{\varphi}(\cdot \| \mu)$, with μ a measure fixed beforehand. Moreover, ψ_{μ}^{\star} denotes the Legendre-Fenchel dual of ψ_{μ} . We will

assume, as above, that given a function $f \in C_c(\mathcal{X})$, $\psi_\mu(\lambda f) \leq \phi(\lambda)$ for some (strictly) convex function ϕ and every $\lambda > 0$.

In order to match the framework that is typically utilised in the Learning Theory literature, we will work with $\psi_{\mathcal{P}_S \mathcal{P}_H}^*(\cdot) = D_\varphi(\cdot \| \mathcal{P}_S \mathcal{P}_H)$ and, following the structure of Theorem 18, one has to assume something about $\psi_{\mathcal{P}_S \mathcal{P}_H}$ in order to provide a bound that involves $\psi_{\mathcal{P}_S \mathcal{P}_H}^*$. However, our assumptions (again, matching the literature) will involve $\psi_{\mathcal{P}_S}$ or $\psi_{\mathcal{P}_Z}$. The reason why we can do this is the following. Given the choice of $\psi_\mu^* = D_\varphi(\cdot \| \mu)$, we typically know the shape that ψ_μ will have. If ψ_μ^* is the Kullback-Leibler Divergence, then $\psi_\mu(f) = \log \mu(\exp(f))$ (cf. Equation (2.3)). If ψ_μ^* is a φ -Divergence, then $\psi_\mu(f) = \mu(\varphi^*(f))$ (cf. Equation (2.11)). This naturally implies that if we consider product measures, one has the following characterisation for ψ :

$$\psi_{\mu \times \xi}(f) = \xi(\mu(\varphi^*(f))) = \xi(\psi_\mu(f))$$

if ψ_μ^* is a φ -Divergences and

$$\exp(\psi_{\mu \times \xi}(f)) = \xi(\mu(\exp(f))) = \xi(\exp(\psi_\mu(f)))$$

if ψ_μ^* is the Kullback-Leibler Divergence.

Consequently, since we will consider the product measure $\mathcal{P}_S \mathcal{P}_H$ we have that, given the structure of ψ , an upper-bound on $\psi_{\mathcal{P}_S}$ for every $h \in \mathcal{H}$ naturally implies an upper-bound on $\psi_{\mathcal{P}_S \mathcal{P}_H}$. Another important consideration is that, an assumption of the form

$$\psi_{\mathcal{P}_Z}(\lambda(\ell(h, Z) - \mathcal{P}_Z(\ell(h, Z)))) \leq \phi(\lambda)$$

for every h in many cases implies an assumption of the form

$$\psi_{\mathcal{P}_S} \left(\frac{\lambda}{n} \left(\sum_i (\ell(h, Z_i) - \mathcal{P}_S(\ell(h, Z_i))) \right) \right) \leq \frac{\phi(\lambda)}{n},$$

with S a sequence of n iid samples distributed according to $\mathcal{P}_Z^{\otimes n}$. This property is something that we will informally define as the “ n -sum property” of $\psi_{\mathcal{P}_Z}$, $\psi_{\mathcal{P}_S}$ and ϕ . Under this framework, we can state the following result.

Corollary 15. *Let $\psi_{\mathcal{P}_S \mathcal{P}_H}^*(\cdot) = D_\varphi(\cdot \| \mathcal{P}_S \mathcal{P}_H)$ for some convex functional φ . Assume that*

$$\psi_{\mathcal{P}_S \mathcal{P}_H}(\lambda(L_S(H) - \mathcal{P}_H(L_{\mathcal{P}}(H)))) \leq \frac{\phi(\lambda)}{n}$$

for every $h \in \mathcal{H}$ and $n, \lambda > 0$. One has that

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) \leq \frac{\phi^*{}^{-1}(n D_\varphi(\mathcal{P}_{SH} \| \mathcal{P}_S \mathcal{P}_H))}{n}. \quad (4.8)$$

Proof. Let us denote with $\phi_n(\lambda) = \frac{\phi(\lambda)}{n}$ one has that $\phi_n^*(\lambda^*) = \frac{1}{n} \phi^*(\lambda n)$ and consequently

$\phi_n^{\star^{-1}}(t) = \frac{1}{n}\phi^{\star^{-1}}(nt)$. The statement then follows from Theorem 18. \square

From this we can now easily recover all the bounds on the expected generalisation-error that involve the Kullback-Leibler Divergence and that are present in the literature, for instance:

Corollary 16. *Let $\varphi(x) = x \log x$ in Corollary 15, hence $D_\varphi(\cdot \| \mathcal{P}_S \mathcal{P}_H)$ is the Kullback-Leibler Divergence. Assume that the loss function $\ell(h, Z_i)$ is σ^2 -sub-Gaussian under \mathcal{P}_Z for every h . We have that*

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) \leq \sqrt{\frac{2\sigma^2 I(S; \mathcal{A}(S))}{n}}. \quad (4.9)$$

Proof. Since $\ell(h, Z_i)$ is σ^2 -sub-Gaussian under \mathcal{P}_Z , then

$$\psi_{\mathcal{P}_Z}(\lambda(\ell(h, Z_i) - \mathcal{P}_Z(\ell(h, Z)))) \leq \frac{\lambda^2 \sigma^2}{2}$$

for every i , where $\psi_{\mathcal{P}_Z}(f) = \log \mathcal{P}_Z(\exp(f))$. Moreover, one also has that

$$\psi_{\mathcal{P}_S} \left(\frac{1}{n} \sum_i \lambda(\ell(h, Z_i) - \mathcal{P}_Z(\ell(h, Z))) \right) \leq \frac{\phi(\lambda)}{n} = \phi(\lambda).$$

Hence, $\psi_{\mathcal{P}_Z}, \psi_{\mathcal{P}_S}$ and ϕ satisfy the “ n -sum property”. The argument then follows directly from Corollary 15 along with the fact that $\frac{1}{n}\phi^{\star^{-1}}(nt) = \sqrt{\frac{2\sigma^2 t}{n}}$. \square

Remark 40. *Both Corollary 13 and Corollary 16 have appeared in the literature in a variety of forms (Russo and Zou, 2016, Prop. 1), (Xu and Raginsky, 2017c, Lemma 1).*

Remark 41. *In a learning theory framework this essentially means that, for a given divergence D_φ , if we can control $\psi_{\mathcal{P}_Z}(\lambda \ell(h, Z))$ for every h (or we can control $\psi_{\mathcal{P}_H \mathcal{P}_Z}(\lambda \ell(H, Z))$, on average with respect to \mathcal{P}_H) using a convex function ϕ , such that $\phi^{\star^{-1}}(x)$ is sub-linear in x , then we can aim at a generalisation error bound that decays with n . The effective decay, though, will also depend on the behaviour of $D_\varphi(\mathcal{P}_{S_H} \| \mathcal{P}_S \mathcal{P}_H)$ as a function of n .*

To conclude the section and emphasise the generality of this approach, let us choose a different divergence.

Corollary 17. *Let $\varphi(x) = \frac{x^2}{2}$ in Corollary 15, hence*

$$\psi_{\mathcal{P}_S \mathcal{P}_H}^{\star}(\cdot) = D_\varphi(\cdot \| \mathcal{P}_S \mathcal{P}_H) = \frac{\chi^2(\cdot \| \mathcal{P}_S \mathcal{P}_H) + 1}{2}.$$

Assume that, given the loss function $\ell(h, Z_i)$, there exists a constant $K > 0$ such that for every h

$$\mathcal{P}_Z((\lambda(\ell(h, Z_i) - \mathcal{P}_Z(\ell(h, Z))))^2) \leq K\lambda^2,$$

then one has that

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) \leq \sqrt{\frac{2K(\chi^2(\mathcal{P}_{S_H} \| \mathcal{P}_S \mathcal{P}_H) + 1)}{n}}. \quad (4.10)$$

Proof. Given that $\psi_{\mathcal{P}_S \mathcal{P}_H}^*(v) = \frac{1}{2}(\chi^2(v \|\mathcal{P}_S \mathcal{P}_H\| + 1))$ one has that $\psi_{\mathcal{P}_S \mathcal{P}_H}(f) = \frac{1}{2} \mathcal{P}_S \mathcal{P}_H(f^2)$ (cf. Equation (2.111)).

Since the Z_i are assumed to be iid random variables the variables $\ell(h, Z_i)$ are also iid and by assumption one has that

$$\mathcal{P}_S \left(\left(\lambda \frac{1}{n} \sum_{i=1}^n (\ell(h, Z_i) - \mathcal{P}_Z(\ell(h, Z))) \right)^2 \right) \leq \frac{\lambda^2}{n} K. \quad (4.11)$$

Hence $\psi_{\mathcal{P}_Z}, \psi_{\mathcal{P}_S}$ and ϕ satisfy the “ n -sum property”. The argument then follows from Corollary 15 and by noticing that $\phi^*(\lambda^*) = \frac{\lambda^{*2}}{2K}$ which in turn implies that $\phi^{*-1}(t) = \sqrt{2Kt}$. \square

Remark 42. Assuming the σ^2 -sub-Gaussianity of $\ell(h, Z_i) - \mathcal{P}_Z(\ell(h, Z))$ naturally implies a bound on $\mathcal{P}_Z((\lambda(\ell(h, Z_i) - \mathcal{P}_Z(\ell(h, Z))))^\kappa)$ for every $\kappa \geq 1$. For instance, σ^2 -sub-Gaussianity implies that in Corollary 17, $\mathcal{P}_Z((\lambda(\ell(h, Z_i) - \mathcal{P}_Z(\ell(h, Z))))^2) \leq K\lambda^2$ holds with $K = 2\sigma^2 e^{2/e}$. However, assuming that ℓ has bounded variance (for every h or on expectation with respect to \mathcal{P}_H) is much less restrictive than assuming it is sub-Gaussian. As an example of this, suppose that the loss ℓ is σ^2 -sub-Gaussian. One has that ℓ^2 is sub-Exponential with parameters $(4\sqrt{2}\sigma^2, 4\sigma^2)$ (Vershynin, 2018, Lemma 2.7). This means that ℓ^2 has a finite log-moment generating function only for $\lambda < 1/(4\sigma^2)$ (Vershynin, 2018, Proposition 2.7.1.v). Considering the framework given by Theorem 18 and Corollary 16, in order to solve the infimum in Equation (3.85) (which, in turn, provides Equation (4.9)), one needs to select $\lambda^* = \sqrt{\frac{2I(S;H)}{\sigma^2}}$. Consequently, if $I(S;H) \geq \frac{1}{25\sigma^2}$ the bound is not valid as, for that choice of λ , the log-moment generating function of ℓ^2 with respect to $\mathcal{P}_S \mathcal{P}_H$ is actually unbounded (and, thus, cannot be bounded by $\phi(\lambda)$). However, a sub-Exponential random variable is such that all of its moments are bounded, hence, an approach as the one suggested by Corollary 17 can be successful.

4.2.1 Recovering other known results

Other known results can be recovered, and they all hinge on creatively re-writing the generalisation error and use similar tools on said re-writing.

Individual Samples Mutual-Information Bound

One can re-write the generalisation error as follows

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) = \frac{1}{n} \sum_{i=1}^n (\mathcal{P}_{HZ}(\ell(H, Z_i)) - \mathcal{P}_H \mathcal{P}_Z(\ell(H, Z))) \quad (4.12)$$

$$= \frac{1}{n} \sum_{i=1}^n (\mathcal{P}_{HZ}(\ell(H, Z_i)) - \mathcal{P}_H \mathcal{P}_Z(\ell(H, Z))). \quad (4.13)$$

In Equation (4.13), the summands are once again difference of expectations of the same function under different distribution. In order to provide a bound on each term inside the sum, one can thus use Corollary 13, which requires us to provide (or assume) an upper-bound

on the dual of the Kullback-Leibler Divergence evaluated at $\lambda(\ell(H, Z_i) - \mathcal{P}_H \mathcal{P}_Z(\ell(H, Z)))$. If one assumes, like before, that $\ell(h, Z)$ is σ^2 -sub-Gaussian under \mathcal{P}_Z for every h (which means setting $\phi(\lambda) = \frac{\lambda^2 \sigma^2}{2}$ in Corollary 13) one recovers (Bu et al., 2019, Proposition 1)

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, \mathcal{S}) = \frac{1}{n} \sum_{i=1}^n (\mathcal{P}_{HZ}(\ell(H, Z_i)) - \mathcal{P}_H \mathcal{P}_Z(\ell(H, Z))) \quad (4.14)$$

$$\leq \frac{1}{n} \sum_{i=1}^n \left(\sqrt{2\sigma^2 I(Z_i; H)} \right). \quad (4.15)$$

The main result therein (cf. (Bu et al., 2019, Theorem 2)) after the re-writing in Equation (4.13), becomes then a corollary of Theorem 18 with $\varphi(x) = x \log(x)$, $\nu = \mathcal{P}_{ZH}$ and $\mu = \mathcal{P}_Z \mathcal{P}_H$.

Conditional Mutual-Information Bound

Another interesting re-writing comes from Steinke and Zakyntinou. The main idea is that that one can consider a super-sample $\tilde{\mathcal{S}}$ of length $2n$ and take random subsets of length n in the following way: let N denote a random sequence drawn uniformly from $\{0, 1\}^n$, one can denote with $\tilde{\mathcal{S}}_N$ a sub-sequence of samples of length n selected at random from $\tilde{\mathcal{S}}$. More precisely, given N and $\tilde{\mathcal{S}} \in \mathcal{Z}^{n \times 2}$, then $(\tilde{\mathcal{S}}_N)_i = \tilde{\mathcal{S}}_{i, N_i+1}$, *i.e.* we see the super-sample of length $2n$ as a matrix with two rows of length n (regular samples of size n) and N formalises the fact that we choose component-wise from the first or the second sample uniformly at random. This allows us to rewrite the generalisation error as follows

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, \tilde{\mathcal{S}}) = \mathcal{P}_{\tilde{\mathcal{S}}}(\mathcal{P}_{N H | \tilde{\mathcal{S}}}(L_{\tilde{\mathcal{S}}_N}(H) - L_{\tilde{\mathcal{S}}_N}(H))). \quad (4.16)$$

One can then use Theorem 18 for every given choice of $\tilde{\mathcal{S}}$ on $\mathcal{P}_{N H | \tilde{\mathcal{S}}}(L_{\tilde{\mathcal{S}}_N}(H) - L_{\tilde{\mathcal{S}}_N}(H))$ by proposing different upper-bounds on the dual of $D(\mathcal{P}_{H N | \tilde{\mathcal{S}}=\tilde{\mathcal{S}}} \| \mathcal{P}_{N | \tilde{\mathcal{S}}=\tilde{\mathcal{S}}} \mathcal{P}_{H | \tilde{\mathcal{S}}=\tilde{\mathcal{S}}})$, (*i.e.*, upper-bounding $\log \mathcal{P}_{N | \tilde{\mathcal{S}}=\tilde{\mathcal{S}}} \mathcal{P}_{H | \tilde{\mathcal{S}}=\tilde{\mathcal{S}}}(\exp(\lambda L_{\tilde{\mathcal{S}}_N}(H) - L_{\tilde{\mathcal{S}}_N}(H)))$) and recover (Steinke and Zakyntinou, 2020, Theorem 5.1, Corollary 5.2, 5.3). In particular, assuming like in (Steinke and Zakyntinou, 2020, Theorem 5.1) that there exists $\Delta : \mathcal{Z}^2 \rightarrow \mathbb{R}$ such that $|\ell(h, z_1) - \ell(h, z_2)| \leq \Delta(z_1, z_2)$ for every $h \in \mathcal{H}$ and $z_1, z_2 \in \mathcal{Z}$ one can then show that for every $\tilde{\mathcal{S}}$

$$\log \mathcal{P}_{N | \tilde{\mathcal{S}}=\tilde{\mathcal{S}}} \mathcal{P}_{H | \tilde{\mathcal{S}}=\tilde{\mathcal{S}}}(\exp(\lambda L_{\tilde{\mathcal{S}}_N}(H) - L_{\tilde{\mathcal{S}}_N}(H))) \leq \frac{\lambda^2}{2n} \sup_h \frac{1}{n} \sum_{i=1}^n \Delta^2(\tilde{\mathcal{S}}_{i,1}, \tilde{\mathcal{S}}_{i,2}) \quad (4.17)$$

$$= \frac{\lambda^2}{2n} c(\tilde{\mathcal{S}}) = \phi_{\tilde{\mathcal{S}}}(\lambda). \quad (4.18)$$

Applying Theorem 18 then one retrieves that

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, \tilde{S}) = \mathcal{P}_{\tilde{S}}(\mathcal{P}_{NH|\tilde{S}}(L_{\tilde{S}_N}(H) - L_{\tilde{S}_N}(H))) \quad (4.19)$$

$$\leq \mathcal{P}_{\tilde{S}}(\phi_Y^{\star^{-1}}(D(\mathcal{P}_{HN|\tilde{S}}\|\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}}))) \quad (4.20)$$

$$= \mathcal{P}_{\tilde{S}}\left(\inf_{\lambda>0} \frac{D(\mathcal{P}_{HN|\tilde{S}}\|\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}}) + \phi_{\tilde{S}}(\lambda)}{\lambda}\right) \quad (4.21)$$

$$\leq \inf_{\lambda>0} \frac{I(N; H|\tilde{S}) + \mathcal{P}_{\tilde{S}}(\phi_{\tilde{S}}(\lambda))}{\lambda} \quad (4.22)$$

$$= \inf_{\lambda>0} \frac{I(N; H|\tilde{S}) + \phi_{\mathcal{P}_{\tilde{S}}(\tilde{S})}(\lambda)}{\lambda} \quad (4.23)$$

$$= \phi_Y^{\star^{-1}}(I(N; H|\tilde{S})) \quad (4.24)$$

$$= \sqrt{\frac{2\mathcal{P}_{\tilde{S}}(c(\tilde{S}))}{n} I(N; H|\tilde{S})} \quad (4.25)$$

thus recovering (Steinke and Zakyntinou, 2020, Theorem 5.1). Different bounds on $\psi_{\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}}}$ lead to different bounds on the expected generalisation error (cf. (Steinke and Zakyntinou, 2020, Corollary 5.2,5.3, Theorem 5.4)).

4.2.2 Generalising Individual Samples and Conditional Mutual-Information

Clearly the possible combinations are numerous, however the pattern remains the same: a bound on the Legendre-Fenchel dual ψ of the targeted divergence ψ^{\star} implies a bound on a difference of expectations (that involve said divergence ψ^{\star}). Following the approach undertaken in (Bu et al., 2019) and described just above, but with the general spirit that characterises this work, one can show the following:

Corollary 18. *Consider a supervised learning setting and consider the sequence of iid samples $Z^n = S$ with $Z_i \sim \mathcal{P}_Z$ and thus $S \sim \mathcal{P}_Z^{\otimes n} = \mathcal{P}_S$. Let $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{H}$ be a learning algorithm mapping sequences of samples to the space of classifiers \mathcal{H} . Denote with \mathcal{P}_{SH} the joint measure induced by \mathcal{P}_S and the Markov Kernel induced by \mathcal{A} on \mathcal{H} . Assume that, given the loss function $\ell(h, Z_i)$, there exists a constant $K > 0$ such that for every h*

$$\mathcal{P}_Z((\lambda(\ell(h, Z_i) - \mathcal{P}_Z(\ell(h, Z))))^2) \leq K\lambda^2,$$

then one has that

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, S) \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2K(\chi^2(\mathcal{P}_{Z_iH}\|\mathcal{P}_{Z_i}\mathcal{P}_H) + 1)}. \quad (4.26)$$

Similarly, one can follow the approach undertake in (Steinke and Zakyntinou, 2020) and prove the following result:

Corollary 19. *Consider a supervised learning setting and consider the sequence of iid samples $Z^n = S$ with $Z_i \sim \mathcal{P}_Z$ and thus $S \sim \mathcal{P}_Z^{\otimes n} = \mathcal{P}_S$. Let $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{H}$ be a learning algorithm mapping*

sequences of samples to the space of classifiers \mathcal{H} . Denote with $\mathcal{P}_{\mathcal{S}H}$ the joint measure induced by $\mathcal{P}_{\mathcal{S}}$ and the Markov Kernel induced by \mathcal{A} on \mathcal{H} . Assume that there exists $\Delta : \mathcal{Z}^2 \rightarrow \mathbb{R}$ such that for every $h \in \mathcal{H}$ and every $z_1, z_2 \in \mathcal{Z}$ one has that $|\ell(h, z_1) - \ell(h, z_2)| \leq \Delta(z_1, z_2)$, then one has that

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, \tilde{\mathcal{S}}) \leq \sqrt{4\mathcal{P}_{\tilde{\mathcal{S}}}(\chi^2(\mathcal{P}_{HN|\tilde{\mathcal{S}}}\|\mathcal{P}_{N|\tilde{\mathcal{S}}}\mathcal{P}_{H|\tilde{\mathcal{S}}}))\mathcal{P}_{Z_1}\mathcal{P}_{Z_2}(\Delta^2(Z_1, Z_2))\left(\frac{1}{n} + 1\right)}, \quad (4.27)$$

where N is a random variable uniformly distributed over $\{0, 1\}^n$ and $\tilde{\mathcal{S}} \in \mathcal{Z}^{n \times 2}$ is a super-sample of size $2n$ (a matrix with two rows and n columns) and such that $\tilde{\mathcal{S}}_N$ represents a regular sample of size n defined as follows $(\tilde{\mathcal{S}}_N)_i = (\tilde{\mathcal{S}})_{(i, N_i+1)}$.

Proof. Similarly to the discussion above, we will leverage the following re-writing of the generalisation error:

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, \tilde{\mathcal{S}}) = \mathcal{P}_{\tilde{\mathcal{S}}}(\mathcal{P}_{NH|\tilde{\mathcal{S}}}(L_{\tilde{\mathcal{S}}_N}(H) - L_{\tilde{\mathcal{S}}_{N'}}(H))). \quad (4.28)$$

As before, in order to prove a bound on the expected generalisation error through (in this case) the χ^2 -divergence, it is necessary to provide a bound on the Legendre-Fenchel of the χ^2 -divergence evaluated at $(\lambda L_{\tilde{\mathcal{S}}_N}(H) - L_{\tilde{\mathcal{S}}_{N'}}(H))$. Leveraging the assumptions and the setting one can prove the following:

$$\mathcal{P}_{H|\tilde{\mathcal{S}}=\tilde{\mathcal{S}}}\mathcal{P}_{N|\tilde{\mathcal{S}}=\tilde{\mathcal{S}}}\left(\left(\lambda L_{\tilde{\mathcal{S}}_N}(H) - L_{\tilde{\mathcal{S}}_{N'}}(H)\right)^2\right) = \mathcal{P}_{H|\tilde{\mathcal{S}}=\tilde{\mathcal{S}}}\mathcal{P}_{N|\tilde{\mathcal{S}}=\tilde{\mathcal{S}}}\left(\left(\lambda \frac{1}{n} \sum_{i=1}^n \ell(H, (\tilde{\mathcal{S}}_N)_i) - \ell(H, (\tilde{\mathcal{S}}_{N'})_i)\right)^2\right) \quad (4.29)$$

$$= \mathcal{P}_{H|\tilde{\mathcal{S}}=\tilde{\mathcal{S}}}\mathcal{P}_{N|\tilde{\mathcal{S}}=\tilde{\mathcal{S}}}\left(\left(\lambda \frac{1}{n} \sum_{i=1}^n (1 - 2N_i) \ell(H, (\tilde{\mathcal{S}}_{i,1}) - \ell(H, (\tilde{\mathcal{S}}_{i,2}))\right)^2\right) \quad (4.30)$$

$$\leq \max_h \mathcal{P}_{N|\tilde{\mathcal{S}}=\tilde{\mathcal{S}}}\left(\left(\lambda \frac{1}{n} \sum_{i=1}^n (1 - 2N_i) \ell(h, (\tilde{\mathcal{S}}_{i,1}) - \ell(h, (\tilde{\mathcal{S}}_{i,2}))\right)^2\right) \quad (4.31)$$

$$\leq \max_h \left(\lambda \frac{1}{n} \sum_{i=1}^n \ell(h, (\tilde{\mathcal{S}}_{i,1}) - \ell(h, (\tilde{\mathcal{S}}_{i,2}))\right)^2 \quad (4.32)$$

$$= \frac{\lambda^2}{n^2} \max_h \left(\sum_{i=1}^n \ell(h, (\tilde{\mathcal{S}}_{i,1}) - \ell(h, (\tilde{\mathcal{S}}_{i,2}))\right)^2 = \phi_{\tilde{\mathcal{S}}}(\lambda). \quad (4.33)$$

Using then Theorem 18 with $\psi_{\mathcal{P}_{H|\tilde{\mathcal{S}}}\mathcal{P}_{N|\tilde{\mathcal{S}}}}(\nu) = \chi^2(\nu\|\mathcal{P}_{H|\tilde{\mathcal{S}}}\mathcal{P}_{N|\tilde{\mathcal{S}}})$ one can then provide the following

bound:

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, \tilde{S}) = \mathcal{P}_{\tilde{S}}(\mathcal{P}_{HN|\tilde{S}}(L_{\tilde{S}_N}(H)) - L_{\tilde{S}_N}(H)) \quad (4.34)$$

$$\leq \mathcal{P}_{\tilde{S}}(\phi_Y^{\star -1}(\chi^2(\mathcal{P}_{HN|\tilde{S}}\|\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}}))) \quad (4.35)$$

$$= \mathcal{P}_{\tilde{S}}\left(\inf_{\lambda>0} \frac{\chi^2(\mathcal{P}_{HN|\tilde{S}}\|\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}}) + \phi_{\tilde{S}}(\lambda)}{\lambda}\right) \quad (4.36)$$

$$\leq \inf_{\lambda>0} \frac{\mathcal{P}_{\tilde{S}}(\chi^2(\mathcal{P}_{HN|\tilde{S}}\|\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}})) + \phi_{\mathcal{P}_{\tilde{S}}}(\lambda)}{\lambda} \quad (4.37)$$

$$= \phi_Y^{\star -1}(\mathcal{P}_{\tilde{S}}(\chi^2(\mathcal{P}_{HN|\tilde{S}}\|\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}}))) \quad (4.38)$$

$$= \sqrt{\frac{4\mathcal{P}_{\tilde{S}}(\chi^2(\mathcal{P}_{HN|\tilde{S}}\|\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}}))\mathcal{P}_{\tilde{S}}(c(\tilde{S}))}{n^2}}, \quad (4.39)$$

where $c(\tilde{S}) = \max_h (\sum_{i=1}^n \ell(h, (\tilde{s}_{i,1})) - \ell(h, (\tilde{s}_{i,2})))^2$. To conclude, let us further upper-bound $\mathcal{P}_{\tilde{S}}(c(\tilde{S}))$ as follows:

$$\mathcal{P}_{\tilde{S}}(c(\tilde{S})) = \mathcal{P}_{\tilde{S}}\left(\max_h \left(\sum_{i=1}^n \ell(h, (\tilde{s}_{i,1})) - \ell(h, (\tilde{s}_{i,2}))\right)^2\right) \quad (4.40)$$

$$\leq \mathcal{P}_{\tilde{S}}\left(\sum_{i=1}^n \Delta^2(\tilde{s}_{i,1}, \tilde{s}_{i,2}) + \sum_{i \neq j} \Delta(\tilde{s}_{i,1}, \tilde{s}_{i,2})\Delta(\tilde{s}_{j,1}, \tilde{s}_{j,2})\right) \quad (4.41)$$

$$= \left(\sum_{i=1}^n \mathcal{P}_{Z_1}\mathcal{P}_{Z_2}(\Delta^2(Z_1, Z_2)) + \sum_{i \neq j} \mathcal{P}_{Z_1}\mathcal{P}_{Z_2}(\Delta^2(Z_1, Z_2))\right) \quad (4.42)$$

$$\leq n\mathcal{P}_{Z_1}\mathcal{P}_{Z_2}(\Delta^2(Z_1, Z_2)) + n^2\mathcal{P}_{Z_1}\mathcal{P}_{Z_2}(\Delta^2(Z_1, Z_2)). \quad (4.43)$$

This leads us to the following upper-bound on the expected generalisation error:

$$\text{gen-err}_{\mathcal{P}}(\mathcal{A}, \tilde{S}) \leq \sqrt{\frac{4\mathcal{P}_{\tilde{S}}(\chi^2(\mathcal{P}_{HN|\tilde{S}}\|\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}}))\mathcal{P}_{\tilde{S}}(c(\tilde{S}))}{n^2}} \quad (4.44)$$

$$\leq \sqrt{4\mathcal{P}_{\tilde{S}}(\chi^2(\mathcal{P}_{HN|\tilde{S}}\|\mathcal{P}_{N|\tilde{S}}\mathcal{P}_{H|\tilde{S}}))\mathcal{P}_{Z_1}\mathcal{P}_{Z_2}(\Delta^2(Z_1, Z_2))\left(\frac{1}{n} + 1\right)}. \quad (4.45)$$

□

4.3 The probability of having a large generalisation error

4.3.1 Sibson's α -Mutual Information

Consider now the learning setup as defined in Section 4.1. The following result can be used to give a concentration bound on the generalisation error defined in Equation (4.3):

Corollary 20. *Let $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{H}$ be a learning algorithm that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution \mathcal{P} over \mathcal{Z} .*

Let $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}$ be a loss function such that $\ell(h, Z)$ is σ^2 -sub-Gaussian random variable for every $h \in \mathcal{H}$. Given $\eta \in (0, 1)$, let $E = \{(S, h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Fix $\alpha \geq 1$. Then,

$$\mathcal{P}_{SH}(E) \leq 2 \exp\left(\frac{\alpha-1}{\alpha} \left(I_{\alpha}(S, \mathcal{A}(S)) - n \frac{\eta^2}{2\sigma^2}\right)\right). \quad (4.46)$$

Consequently, in order to ensure a confidence of $\delta \in (0, 1)$, i.e. $\mathcal{P}_{SH}(E) \leq \delta$, it is sufficient to have n samples where

$$n \geq \frac{2\sigma^2}{\eta^2} \left(I_{\alpha}(S, \mathcal{A}(S)) + \log 2 + \frac{\alpha}{\alpha-1} \log\left(\frac{1}{\delta}\right)\right). \quad (4.47)$$

Proof. Fix $\eta \in (0, 1)$. Let us denote with E_h the fiber of E over h for some $h \in \mathcal{H}$, i.e. $E_h = \{S : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. By assumption we have that $\ell(h, Z)$ is σ^2 -sub-Gaussian for every h . We can thus use Hoeffding's inequality and retrieve that for every $h \in \mathcal{H}$:

$$\mathcal{P}_S(E_h) \leq 2 \cdot \exp\left(-n \frac{\eta^2}{2\sigma^2}\right). \quad (4.48)$$

Then it follows from Corollary 4 and Inequality (4.48) that:

$$\begin{aligned} \mathcal{P}_{SH}(E) &\leq \exp\left(\frac{\alpha-1}{\alpha} I_{\alpha}(S, \mathcal{A}(S))\right) \cdot \left(2 \exp\left(-n \frac{\eta^2}{2\sigma^2}\right)\right)^{\frac{1}{\alpha}} \\ &= 2 \exp\left(\frac{\alpha-1}{\alpha} \left(I_{\alpha}(S, \mathcal{A}(S)) - n \frac{\eta^2}{2\sigma^2}\right)\right). \end{aligned} \quad (4.49)$$

□

Remark 43. Corollary 20 applies to the special case in which $\mathcal{Z} = \mathcal{D} \times \mathcal{C}$ and ℓ is the 0-1 loss function as defined in (4.4). Indeed, one can show that ℓ is σ^2 -sub-Gaussian for $\sigma = \frac{1}{2}$. Moreover, in this case, we only need to assume that the samples S are independent, as the use of Hoeffding's inequality in the proof can be replaced by McDiarmid's inequality (for functions with bounded differences) (McDiarmid, 1989).

Smaller α means that $I_{\alpha}(S, \mathcal{A}(S))$ will be smaller, but it will imply a worse dependence on $\log(1/\delta)$ in the sample complexity. It is worth noticing that, for fixed α , the sample complexity dependence on $\log(1/\delta)$ is optimal (up to constants). In particular, consider the setup of PAC learning with finite \mathcal{H} with VC dimension d , e.g. assume that $\mathcal{D} = [d]$ and $\mathcal{H} = \{0, 1\}^{\mathcal{D}}$, we have that the VC-dimension of \mathcal{H} is d (Bassily et al., 2018; Shalev-Shwartz and Ben-David., 2014). By (Shalev-Shwartz and Ben-David., 2014, Theorem 6.8), we know that the number of necessary samples for learning, in the “realizable case”, satisfies $n \geq c \frac{d + \log(1/\delta)}{\eta}$, for some constant c . Assume also that \mathcal{A} is the ERM algorithm, in which case the generalisation error and the true error are the same and $I_{\alpha}(S, \mathcal{A}(S)) \leq \log(|\mathcal{H}|) = d$. From Equation (4.47) for the 0-1 loss we have that $n \geq c \frac{d + \gamma \log(1/\delta)}{\eta^2}$, where $\gamma = \frac{\alpha}{\alpha-1}$. Hence, for a given α , the dependence on δ is optimal. A similar reasoning could be applied to the agnostic case in order to tackle the optimality with respect to η as well.

4.3.2 Maximal Leakage

An interesting particular case of Corollary 4 appears when $\alpha \rightarrow \infty$. In this scenario, on the right-hand side of Equation (3.41) one obtains Maximal Leakage (cf. Corollary 5). Maximal Leakage enjoys a variety of properties that are of particular interest to us and we will soon analyse them.

Remark 44. *Note that, considering the other extreme i.e., $\alpha \rightarrow 1$, one retrieves a trivial bound. Indeed, letting $\alpha \rightarrow 1$ in the results presented in Section 3.1.2 leads to a trivial upper-bound of 1 on $\mathcal{P}_{XY}(E)$. Hence, this approach does not provide bounds that exploit either the Kullback-Leibler Divergence or the Mutual Information. Nonetheless, we will provide a comparison with a similar result obtained for Mutual Information (that can be obtained through the Legendre-Fenchel transform of the Kullback-Leibler Divergence, cf. Equation (2.10) and has been provided in (Bassily et al., 2018, Theorem 8)) in Section 4.4.1.*

This result is helpful for the following reasons:

- Maximal Leakage is more amenable to analysis due to its semi-closed form (e.g., it is possible to easily compute the Maximal Leakage of noise-addition mechanisms);
- The absence of the power $\frac{1}{\gamma}$ in Equation (3.45) as compared to the right-hand side of Equation (3.41) allows us to provide a generalisation of the classical concentration of measure results in adaptive settings;
- A conditional version of Maximal Leakage will enable us to provide adaptive composition results (cf. Section 4.5).

Generalisation Error Bounds

We will now explore how Corollary 5 can be applied in providing bounds on the generalisation error of learning algorithms.

Corollary 21. *Let $A : \mathcal{Z}^n \rightarrow \mathcal{H}$ be a learning algorithm that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution \mathcal{P} over \mathcal{Z} . Let $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}$ be a loss function such that $\ell(h, Z)$ is σ^2 -sub-Gaussian random variable for every $h \in \mathcal{H}$. Given $\eta \in (0, 1)$, let $E = \{(S, h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$. Then,*

$$\mathcal{P}_{SH}(E) \leq 2 \cdot \exp\left(\mathcal{L}(S \rightarrow \mathcal{A}(S)) - n \frac{\eta^2}{2\sigma^2}\right). \quad (4.50)$$

Consequently, in order to ensure a confidence of $\delta \in (0, 1)$, i.e. $\mathbb{P}(E) \leq \delta$, it is sufficient to have n samples where

$$n \geq \frac{2\sigma^2}{\eta^2} \left(\mathcal{L}(S \rightarrow \mathcal{A}(S)) + \log\left(\frac{2}{\delta}\right) \right). \quad (4.51)$$

The proof follows from Corollary 5 and the same technique used to prove Corollary 20.

Remark 45. *Similarly to Corollary 20, this bound applies to the case in which ℓ is the 0-1 loss function with $\sigma = \frac{1}{2}$. Moreover, as discussed following Corollary 20, the dependence on $\log(1/\delta)$*

is optimal. Indeed, let us consider the very same example as in the previous section. Let $\mathcal{D} = [d]$ and $\mathcal{H} = \{0, 1\}^{\mathcal{D}}$, we have that the VC-dimension of \mathcal{H} is d . Choosing again \mathcal{A} to be the ERM algorithm we have that $\mathcal{L}(S \rightarrow \mathcal{A}(S)) = d$. Looking at Equation (4.51) with $\sigma = 1/2$, we can see how the lack of the $\alpha/(\alpha - 1)$ term (that one can find in Equation (4.47) instead) allows us to make a clear analogy with the VC-dimension bound stated in (Shalev-Shwartz and Ben-David., 2014, Theorem 6.8). More precisely, from Equation (4.51) we have that $n \geq \frac{d + \log(2/\delta)}{2\eta^2}$ while (Shalev-Shwartz and Ben-David., 2014, Theorem 6.8.3) (realizable case) tells us that $n \geq c \frac{d + \log(1/\delta)}{\eta}$ for some constant c . A similar reasoning could be applied to the agnostic case in order to tackle the optimality with respect to η as well.

Whenever \mathcal{A} is independent from the samples S , we have that $\exp(\mathcal{L}(S \rightarrow \mathcal{A}(S))) = 1$ and we immediately fall back on the non-adaptive scenario:

$$\mathcal{P}_{SH}(E) \leq 2 \cdot \exp\left(-n \frac{\eta^2}{2\sigma^2}\right) = \mathcal{P}_S \mathcal{P}_H(E), \quad (4.52)$$

i.e., we recover (and, thus, generalise) Hoeffding's inequality.

4.3.3 Analysing Schemes via Maximal Leakage

A simple way of keeping the Maximal Leakage of an algorithm $\mathcal{A}(X)$ bounded (and thus ensure generalisation) is to add noise (e.g., $\hat{Y} = \mathcal{A}(X) + N$ with \mathcal{A} a real-valued function). The proofs for this section can be found in Section 4.B.

Lemma 8 (Laplacian Noise). *Let $h : \mathcal{X}^n \rightarrow \mathbb{R}$ be a function such that $h(x) \in [a, c]$, $a < c \forall x \in \mathcal{X}^n$. The mechanism $\mathcal{M}(x) = h(x) + N$ where $N \sim \text{Lap}(b)$ is such that:*

$$\mathcal{L}(X \rightarrow \mathcal{M}(X)) = \log\left(1 + \frac{(c - a)}{b}\right). \quad (4.53)$$

Similar results can be obtained by analysing different types of noise.

Lemma 9 (Gaussian Noise). *Let $h : \mathcal{X}^n \rightarrow \mathbb{R}$ be a function such that $\forall x \in \mathcal{X}^n$ $h(x) \in [a, c]$, $a < c$. The mechanism $\mathcal{M}(x) = h(x) + N$ where $N \sim \mathcal{N}(0, \sigma^2)$ is such that:*

$$\mathcal{L}(X \rightarrow \mathcal{M}(X)) = \log\left(1 + \frac{(c - a)}{\sqrt{2\pi\sigma^2}}\right). \quad (4.54)$$

Lemma 10 (Exponential Noise). *Let $h : \mathcal{X}^n \rightarrow \mathbb{R}$ be a function such that $\forall x \in \mathcal{X}^n$ $h(x) \in [a, c]$, $c > 0$. The mechanism $\mathcal{M}(x) = h(x) + N$ where $N \sim \text{Exp}(\lambda)$ (i.e., $\mathbb{E}[N] = (1/\lambda) = b$) is such that:*

$$\mathcal{L}(X \rightarrow \mathcal{M}(X)) = \log\left(1 + \frac{(c - a)}{b}\right). \quad (4.55)$$

The addition of carefully calibrated noise to control Maximal Leakage can be used in practice to obtain generalisation guarantees of learning algorithms. As an exact analogy to (Xu and

4.3 The probability of having a large generalisation error

Raginsky, 2017c, Corollary 4) we can state the following result, involving a noisy version of the Empirical Risk Minimization (ERM) algorithm.

Corollary 22. *Let us consider the following algorithm:*

$$\mathcal{A}(S) = \operatorname{argmin}_{h \in \mathcal{H}} (L_S(h) + N_h), \quad (4.56)$$

where N_h is exponential noise drawn independently from the input, added to the empirical risk of each hypothesis on a given data-set S . Suppose \mathcal{H} is countable (i.e., finite or countably infinite), and denote with N_i the noise added to the hypothesis h_i with mean b_i . Then, for every $\eta \in (0, 1)$:

$$\mathcal{P}_{SH}(\text{gen-err}(\mathcal{A}) \geq \eta) \leq 2 \exp \left(\sum_{i=1}^{|\mathcal{H}|} \log \left(1 + \frac{1}{b_i} \right) - 2n\eta^2 \right). \quad (4.57)$$

Choosing $b_i = i^{1.1} / n^{1/3}$, we retrieve:

$$\mathcal{P}_{SH}(\text{gen-err}(\mathcal{A}) \geq \eta) \leq 2 \exp \left(-n(2\eta^2 - 11/n^{2/3}) \right). \quad (4.58)$$

This example shows how easily the Maximal Leakage bound can be used, in contrast with the Mutual Information one. Indeed, following the proof of (Xu and Raginsky, 2017c, Corollary 4), the Mutual Information of the same mechanism analysed here is hard to compute directly and the quantity $I(S; H)$ is, in the end, upper-bounded using Maximal Leakage:

$$I(S; H) \leq \sum_{i=1}^{|\mathcal{H}|} \log \left(1 + \frac{L_\mu(h_i)}{b_i} \right) \quad (4.59)$$

$$\leq \sum_{i=1}^{|\mathcal{H}|} \log \left(1 + \frac{1}{b_i} \right) = \mathcal{L}(S \rightarrow H). \quad (4.60)$$

Remark 46. *The noisy version of the ERM algorithm does not provide the same guarantees (with respect to the classical ERM) in terms of training error. Having a small generalisation error means that the training and testing errors are close, but they could be large. In this case, by adding noise, we reduce the information measure (by Data-Processing Inequality) and, as a consequence of our bounds, the probability of having a large generalisation error is also reduced. The addition of noise, however, can increase the training error of the new algorithm. In particular, we will no longer choose the empirical risk minimiser ($h^* = \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$) but some hypothesis h that minimizes $L_S(h) + N_h$ and whose error can be more or less close to $L_S(h^*)$ (depending on the noise). Hence, while the generalisation error may be smaller, training and testing errors could both get larger.*

4.3.4 Other Divergences

This section presents new bounds in terms of Rényi's α -Divergences and φ -Divergences, particularly Hellinger divergences. Applying Corollary 6 to a learning setting, one gets:

Corollary 23. *Let $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{H}$ be a learning algorithm that, given a sequence S of n points,*

returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution \mathcal{P} over \mathcal{Z} . Let $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}$ be a loss function such that $\ell(h, Z)$ is a σ^2 -sub-Gaussian random variable, for some σ and for every $h \in \mathcal{H}$. Given $\eta \in (0, 1)$, let $E = \{(S, h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$ and $p \in (1, +\infty)$. Then,

$$\mathcal{P}_{SH}(E) \leq 2^{\frac{p-1}{p}} \exp\left(\frac{\log((p-1)H_p(S, \mathcal{A}(S)) + 1)}{p} - \frac{n\eta^2}{2p\sigma^2}\right).$$

In particular,

$$\mathcal{P}_{SH}(E) \leq \sqrt{2} \exp\left(\frac{1}{2} \left(\log(\chi^2(S, \mathcal{A}(S)) + 1) - \frac{n\eta^2}{2\sigma^2}\right)\right), \quad (4.61)$$

and in order to ensure a confidence of $\delta \in (0, 1)$, i.e. $\mathcal{P}_{SH}(E) \leq \delta$, it is sufficient to have n samples where

$$n \geq \frac{2\sigma^2}{\eta^2} \left(\log(\chi^2(S, \mathcal{A}(S)) + 1) + 2\log\left(\frac{\sqrt{2}}{\delta}\right)\right).$$

An implication of Equation (4.61), say for the 0-1 loss function, is the following: if, for a given n ,

$$\chi^2(S, \mathcal{A}(S)) < \exp(2n\eta^2) - 1$$

then we can guarantee an exponential decay in the probability of having a large generalisation error. Doing the same with Equation (3.49), one gets:

$$\mathcal{P}_{SH}(E) \leq \sqrt{2} \exp\left(\frac{1}{2} (\mathcal{L}(S \rightarrow \mathcal{A}(S)) - 2n\eta^2)\right). \quad (4.62)$$

Given the relationship between these two measures, one has that every time

$$\exp(\mathcal{L}(X \rightarrow Y)) \leq 2n\eta^2 \text{ then } \chi^2(X, Y) \leq \exp(2n\eta^2) - 1$$

and thus, generalisation with Maximal Leakage implies generalisation with χ^2 . An advantage of using $\chi^2(X, Y)$ is that it can be significantly smaller than $\mathcal{L}(X \rightarrow Y)$. Indeed:

Example 7. Let $X \sim \text{Ber}(1/2)$ and let $Y = \text{BSC}(p)$, with $p < 1/2$. Thus, $\mathcal{P}_{Y|X=x}(x) = 1 - p$. In this case $\chi^2(X, Y) = (1 - 2p)^2$ while $\exp(\mathcal{L}(X \rightarrow Y)) - 1 = (1 - 2p)$. It is easy to see that, since $(1 - 2p) < 1$ then $(1 - 2p)^2$ can be much smaller than $(1 - 2p)$.

Nonetheless, using Maximal Leakage can be advantageous. In fact, the Maximal Leakage $\mathcal{L}(X \rightarrow Y)$ depends on \mathcal{P}_X only through the support. This allows us to provide bounds that depend only loosely on the distribution over the training samples. The $\chi^2(X, Y)$ instead, cannot be exactly computed unless one has access to \mathcal{P}_X . The distribution over the data \mathcal{P}_{X^n} can be very complicated and is typically defined on large-dimensional spaces (e.g., images, audio recordings, etc.). In general, one only has access to (and control over) the conditional distributions $\mathcal{P}_{Y|X^n}$ induced by the chosen learning algorithm. This can render bounds

4.3 The probability of having a large generalisation error

like Equation (4.61) difficult to use in practice, although “tighter” in theory.

Another important characteristic of Maximal Leakage is that, due to the “chain rule” it satisfies, it composes adaptively (cf. Section 4.5). This property is not known to hold, in general, for either φ - or Sibson’s α -Mutual Information.

To conclude this section, let us state the generalisation error and sample complexity bounds provided by Hellinger distance.

Corollary 24. *Let $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{H}$ be a learning algorithm that, given a sequence S of n points, returns a hypothesis $h \in \mathcal{H}$. Suppose S is sampled i.i.d according to some distribution \mathcal{P} over \mathcal{Z} . Let $\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}$ be a loss function such that $\ell(h, Z)$ is a σ^2 -sub-Gaussian random variable, for some σ and for every $h \in \mathcal{H}$. Given $\eta \in (0, 1)$, let $E = \{(S, h) : |L_{\mathcal{P}}(h) - L_S(h)| > \eta\}$.*

$$\mathcal{P}_{SH}(E) \leq 2 \exp\left(-n \frac{\eta^2}{2\sigma^2}\right) + H^2(S; \mathcal{A}(S)) + 2^{3/2} H(S; \mathcal{A}(S)) \exp\left(-n \frac{\eta^2}{4\sigma^2}\right) \quad (4.63)$$

$$\leq 2 \exp\left(-n \frac{\eta^2}{2\sigma^2}\right) + H^2(S; \mathcal{A}(S)) + 2^{3/2} H(S; \mathcal{A}(S)). \quad (4.64)$$

and in order to ensure a confidence of $\delta \in (0, 1)$, i.e. $\mathcal{P}_{SH}(E) \leq \delta$, it is sufficient to have n samples where

$$n \geq \frac{\log\left(\frac{1}{\delta - H^2(S; \mathcal{A}(S)) + 2^{3/2} H(S; \mathcal{A}(S))}\right)}{4\eta^2}. \quad (4.65)$$

The main advantage of the Hellinger squared distance (similarly to the Total Variation distance) is that it is always guaranteed to be bounded by 1. Regardless, the same reasoning that compared Maximal Leakage to χ^2 applies: computing $H(X; Y)$ requires access to the marginal distributions $\mathcal{P}_X, \mathcal{P}_Y$ and can be very complicated. Indeed, even for simple additive noise channels, no closed-form expression is known for $H(X^n; Y)$ (or even for $TV(X^n; Y)$). In the context of learning instead, even for simple gradient descent mechanisms (cf. (Wang et al., 2019)), computing such measures can be very hard.

4.3.5 From Probability to Expected Value

Given the bounds on the probability of having a large generalisation error proposed so far, one might ask how these reflect upon the expected value of the generalisation error. To give a meaningful result some assumptions on the probability of our event E are required. In particular, we will assume this probability to be exponentially decreasing with the number of samples n (as it often happens in the literature). The following result is inspired by (Shalev-Shwartz and Ben-David., 2014, p. 419) with a slightly different proof.

Lemma 11. *Let X be a random variable and let $\hat{x} \in \mathbb{R}$. Suppose that there exist $a \geq 0$ and $b \geq e$ such that for every $\eta > 0$ $\mathcal{P}_X(|X - \hat{x}| \geq \eta) \leq 2b \exp(-\eta^2 / a^2)$ then*

$$\mathcal{P}_X(|X - \hat{x}|) \leq a \min\left\{3\sqrt{\log b}, 2\sqrt{\log 2b}\right\}.$$

Proof. Since $|X - \hat{x}|$ is a positive random variable one has that:

$$\mathcal{P}_X(|X - \hat{x}|) = \int_0^{+\infty} \mathcal{P}_X(|X - \hat{x}| \geq \eta) d\eta. \quad (4.66)$$

One, thus, has the following:

$$\mathcal{P}_X(|X - \hat{x}|) = \int_0^{+\infty} \mathcal{P}_X(|X - \hat{x}| \geq \eta) d\eta \quad (4.67)$$

$$\leq \int_0^{+\infty} \min(1, 2b \exp(-\eta^2/a^2)) d\eta \quad (4.68)$$

$$= \int_0^{\sqrt{a^2 \log 2b}} d\eta + \int_{\sqrt{a^2 \log 2b}}^{+\infty} 2b \exp(-\eta^2/a^2) d\eta \quad (4.69)$$

$$\leq \sqrt{a^2 \log 2b} + \frac{a^2}{\sqrt{a^2 \log 2b}} \int_{\sqrt{a^2 \log 2b}}^{+\infty} \frac{2b\eta}{a^2} \exp(-\eta^2/a^2) d\eta \quad (4.70)$$

$$= a \left(\sqrt{\log 2b} + \frac{1}{\sqrt{\log 2b}} \right) \leq a \min \{ 3\sqrt{\log b}, 2\sqrt{\log 2b} \}. \quad (4.71)$$

□

This section will focus only on Sibson's α -Mutual Information. It is possible to extend these results also to φ -Mutual Information. However, in order to do so, one needs explicitly choose φ (or a family of convex functions that satisfy Theorem 14). For instance, using the same techniques and starting from Corollary 23, one can state a bound on the expected generalisation error for p -Hellinger divergences. It is, however, unclear how to derive a result involving all increasing and convex functions φ .

Theorem 24. *Let $\mathcal{A} : \mathcal{Z}^n \rightarrow \mathcal{H}$ be a learning algorithm and let $I_\alpha(S, \mathcal{A}(S))$ (i.e., Sibson's Mutual Information of order α) be the dependence measure chosen. Suppose that the loss function $\ell : \mathcal{Z} \times \mathcal{H} \rightarrow \mathbb{R}$ is such that $\forall h, \mathcal{P}_{S \sim \mathcal{D}^n}(|L_S(h) - \mathcal{P}_S(L(h))| > \eta) \leq 2 \exp\left(-\frac{\eta^2}{2\sigma^2 n}\right)$ for some $\sigma > 0$ (e.g. $\ell(h, Z) - \mathbb{E}[\ell(h, Z)]$ is σ^2 -sub-Gaussian for each h), then:*

$$\mathcal{P}_{SH}(|L_S(H) - \mathcal{P}_S(L(H))|) \leq \sqrt{\frac{8\sigma^2(\log(2) + I_\alpha(S, \mathcal{A}(S)))}{n}}.$$

Proof. The proof is a simple application of Lemma 11 and Corollary 20 with $a = \sqrt{2\gamma\sigma^2}/\sqrt{n}$ and $b = 2^{\frac{1}{\gamma}-1} \exp\left(\frac{1}{\gamma} I_\alpha(S, \mathcal{A}(S))\right)$. □

Remark 47. *Notice that, even though we provide a concrete example (cf. Theorem 24) that uses σ^2 -sub-Gaussianity the assumption is not strictly necessary. Lemma 11 only requires that $\mathcal{P}_X(|X - \hat{x}|)$ decays exponentially fast. This can be true also for other classes of random variables, like sub-Weibull random variables with an opportune choice of parameters (Vladimirova et al., 2020).*

4.3 The probability of having a large generalisation error

An important result, obtained through a different route, is the bound on the expected generalisation error via Mutual Information (cf. (Xu and Raginsky, 2017c, Theorem 1)). We restate it here for ease of reference. Under the assumption that $\ell(h, Z)$ is σ^2 -sub Gaussian for each h :

$$|\mathcal{P}_{SH}(L_S(H) - \mathcal{P}_S \mathcal{P}_H(L(H)))| \leq \sqrt{\frac{2\sigma^2}{n} I(S; \mathcal{A}(S))}. \quad (4.72)$$

In the spirit of comparison, let us also state a similar bound using Theorem 24 but with $\alpha \rightarrow \infty$ and using $3a\sqrt{\log b}$ as a bound on the expected value. Setting $a = \sqrt{\frac{2\sigma^2}{n}}$, $b = \exp(\mathcal{L}(S \rightarrow \mathcal{A}(S)))$ one retrieves the following:

$$|\mathcal{P}_{SH}(L_S(H) - \mathcal{P}_S(L(H)))| \leq \mathcal{P}_{SH}(|L_S(H) - \mathcal{P}_S(L(H))|) \quad (4.73)$$

$$\leq 3\sqrt{\frac{2\sigma^2}{n} \mathcal{L}(S \rightarrow \mathcal{A}(S))}. \quad (4.74)$$

We have seen before (cf., Section 4.4.1) that Sibson's α -Mutual-Information brings an exponential improvement in the dependence over δ (cf. Section 4.3). However, the same information-measure does not seem to bring any improvement in controlling the expected generalisation error when the technique described in this section is used.

4.3.6 Some considerations

While the relationship among I_α for various α 's is clear, a detailed and complete understanding of the relationship among all the φ -Divergences and Rényi's α -Divergences is still lacking. Many works are trying to address the issue (e.g., (Sason and Verdú, 2016; Harremoës and Vajda, 2010)). Restricting ourselves to χ^2 -like divergences, a summary of our current understanding is the following:

- $I_\alpha(X, Y) = \min_{\mathcal{Q}_Y} D_\alpha(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{Q}_Y) \leq D_\alpha(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y)$;
- $I_{\alpha_1}(X, Y) \leq I_{\alpha_2}(X, Y)$ if $\alpha_1 \leq \alpha_2$;
- $\mathcal{L}(X \rightarrow Y) \geq \log(\chi^2(X, Y) + 1) \geq I_2(X, Y)$.

A naïve consideration would be the following: in this family of divergences the tightest sample-complexity bound would be given by I_2 (where both $1/\alpha = 1/\beta = 1/2$). Taking $\alpha \in (1, 2)$ will consistently improve the dependence of the bound with respect to the information measure term while rendering the bound closer and closer to 1 (and, thus, worsening the dependence with respect to δ in the sample complexity, as β will tend to $+\infty$, cf. Equation (4.47)). The best trade-off between these two quantities remains an open problem.

Considering Corollary 24 instead, while the dependence with respect to δ and η seems to be the right one, the role played by the information measure is not as clear. Finding the optimal function φ in Theorem 14, for a given behaviour of $\mathcal{P}_X \mathcal{P}_Y(E)$, could also shed some light on whether or not one should consider functions φ outside of the χ^2 -like family (polynomials). A Taylor expansion argument shows that most φ -divergences are, in the end, χ^2 -like, but

while those measures blow up in some deterministic settings, others like Total Variation and Hellinger Distance do not. This different behaviour could be key in obtaining the tightest bound in the learning theory framework as well.

4.4 Comparison with other bounds

Table 4.4.1: Comparison between bounds.

	Robust (cf. Lemma 16)	Adaptive (cf. Lemma 17)	Bound	Sample Complexity
γ -Stability (Bousquet and Elisseeff (2002))	No	No	exp. decay in n	$f(\gamma, \eta) \times \log\left(\frac{2}{\delta}\right)$
ϵ -DP (Dwork et al. (2015b))	Yes	Yes	$\frac{1}{4} \exp\left(\frac{-n\eta^2}{12}\right)$, $\epsilon \leq \eta/2$	$\frac{12 \cdot \log(1/4\delta)}{\eta^2}$
Mutual Information(Bassily et al. (2018))	Yes	Yes	$\frac{I(S;Y)+1}{2n\eta^2-1}$	$\frac{I(S;Y)}{\eta^2} \cdot \frac{1}{\delta}$
Maximal Leakage	Yes	Yes	$2 \cdot \exp(\mathcal{L}(S \rightarrow Y) - 2n\eta^2)$	$\frac{\mathcal{L}(S \rightarrow Y) + \log\left(\frac{2}{\delta}\right)}{2\eta^2}$
Sibson's α -Mutual Information	Yes	Unknown	$\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(S, Y) - 2n\eta^2\right)$	$\frac{I_\alpha(S, Y) + \gamma \log\left(\frac{1}{\delta}\right)}{2\eta^2}$
χ^2 -Divergence	Yes	Unknown	$2 \exp(-n\eta^2) \sqrt{2\chi^2(S, Y) + 1}$	$\frac{\log(\chi^2(S, Y) + 1) + 2 \log\left(\frac{\sqrt{2}}{\delta}\right)}{\eta^2}$
VC-Dim. d (Shalev-Shwartz and Ben-David. (2014))			$2 \cdot \exp(\log(K) - 2n\eta^2)$	$\frac{d + \log\left(\frac{2}{\delta}\right)}{2\eta^2}$

This section compares the proposed new bounds to the existing ones from the literature. A summary is provided in Table 4.4.1, where the bound involving the Hellinger squared distance (cf. Corollary 24) has been omitted since it has a very different shape (both in terms of the high-probability bound and, as a byproduct, in terms of the sample complexity bound). At a glance, from Table 4.4.1, it is easy to see that most of the information measures bounds, with the sole exception of the Mutual Information, can have an exponential decay with the number of samples n . This is indeed the desired behaviour with respect to n . The reason is that the event E we consider is the event that the empirical average of some function (empirical risk) evaluated on a sequence of iid random variables (e.g., S , the training sample) diverges from its actual average (risk) more than some constant η . If this function is independent from the samples S , the decay one typically finds in the literature is exponential with the number of samples n (e.g., McDiarmid's and Hoeffding's inequality). More detailed comparisons will appear in the following subsections.

4.4.1 Maximal Leakage and Mutual Information

A result that connects the probability of having a large generalisation error with Mutual Information, under the same assumptions as Corollary 20, is the following (cf. (Bassily et al.,

2018, Theorem 8):

$$\mathcal{P}_{SH}(E) \leq \frac{I(S; \mathcal{A}(S)) + \log 2}{2\eta^2 - \log 2}. \quad (4.75)$$

Let us compare this result with Corollary 21 in terms of sample complexity. From Corollary 21, it follows that using a sample size of

$$n \geq \left(\frac{\mathcal{L}(S \rightarrow \mathcal{A}(S)) + \log(2/\delta)}{2\eta^2} \right), \quad (4.76)$$

yields a learner for \mathcal{H} with accuracy η and confidence δ . Using the same reasoning with Equation (4.75), one gets:

$$n \geq \left(\frac{I(S; \mathcal{A}(S)) + \log 2 + \delta \log 2}{2\eta^2 \delta} \right). \quad (4.77)$$

In the regime where Maximal Leakage and Mutual Information behave similarly, the reduction in the sample complexity is exponential in δ . However, in general, $\mathcal{L}(X \rightarrow Y) \geq I(X; Y)$ and the gap between the two information measures can be arbitrarily large. This means that, in some settings, the two information measures could behave quite differently and, in those cases, a comparison between Equation (4.76) and Equation (4.77) is not straightforward. The same reasoning can be applied to the sample complexity of I_α for a given $\alpha \in (1, +\infty]$: the exponential improvement in δ remains, although with a worse constant that multiplies the $\log(1/\delta)$ term. Another source of comparison can be found from the Examples presented in Section 3.1.4. Indeed, re-stating them for Mutual Information one obtains the following:

Example 1* . Independent case

In the same setting as in Example 1, that is X independent from Y and E is such that $\mathcal{P}_X(E_y) = c$ for all $y \in \mathcal{Y}$, from Equation (4.75) one obtains:*

$$c = \mathcal{P}_{XY}(E) \leq \frac{1}{-\log(\max_y \mathcal{P}_X(E_y))} = \frac{1}{\log(1/c)}, \quad (4.78)$$

which is weaker than the bound $\mathcal{P}_{XY}(E) \leq c$ that can be obtained from Equation (3.45).

Example 2* . Strongly Dependent Case

In the same setting as in Example 2, that is $X = Y \sim \mathcal{U}([n])$ and $E = \{(x, y) \in [n] \times [n] | x = y\}$, from Equation (4.75) one obtains:*

$$1 = \mathcal{P}_{XY}(E) \leq 1 + \frac{1}{\log n}. \quad (4.79)$$

However, from Equation (3.45) one obtains:

$$1 = \mathcal{P}_{XY}(E) \leq \frac{1}{n} \cdot n = 1. \quad (4.80)$$

Thus, while the bound obtained via Mutual Information is asymptotically tight, the bound obtained via Maximal Leakage is met with equality.

4.4.2 Maximal Leakage and Differential Privacy

This section will compare our results with the generalisation guarantees provided by differential privacy (DP). The definition of ϵ -differential privacy (ϵ -DP for short) is the following:

Definition 24. Let $\mathcal{A} : \mathcal{X}^n \rightarrow \mathcal{Y}$ be a randomised algorithm. \mathcal{A} is ϵ -DP if for every $S \subseteq \mathcal{Y}$ and every $x, \hat{x} \in \mathcal{X}^n$ that differ only in one position:

$$\mathbb{P}(\mathcal{A}(x) \in S) \leq e^\epsilon \mathbb{P}(\mathcal{A}(\hat{x}) \in S). \quad (4.81)$$

A relationship with Maximal Leakage can be established:

Lemma 12. Let $\mathcal{A} : \mathcal{X}^n \rightarrow \mathcal{Y}$ be an ϵ -DP algorithm, then

$$\mathcal{L}(X \rightarrow \mathcal{A}(X)) \leq \epsilon \cdot n. \quad (4.82)$$

Proof. Let $Y = \mathcal{A}(X)$ and assume, for simplicity, that Y is a discrete random variable (the proof for continuous Y follows from very similar arguments). Fix some $\hat{x} \in \mathcal{X}^n$, $\forall x \in \mathcal{X}^n$ we have that x and \hat{x} differ in at most n positions and, iteratively applying the definition of Differential Privacy, we have that $\mathbb{P}(Y = y | X = x) \leq e^{\epsilon \cdot n} \mathbb{P}(Y = y | X = \hat{x})$. Thus:

$$\mathcal{L}(X \rightarrow Y) = \log \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}^n} \mathbb{P}(Y = y | X = x) \quad (4.83)$$

$$\leq \log \sum_{y \in \mathcal{Y}} e^{\epsilon \cdot n} \mathbb{P}(Y = y | X = \hat{x}) \quad (4.84)$$

$$= n \cdot \epsilon \quad (4.85)$$

□

This suggests an immediate application of Corollary 21 to a learning setting. Indeed, suppose the learning algorithm \mathcal{A} is also ϵ -Differentially Private, then:

$$\exp(\mathcal{L}(X \rightarrow Y) - 2n\eta^2) \leq \exp(\epsilon n - 2n\eta^2) \quad (4.86)$$

$$= \exp(-n(2\eta^2 - \epsilon)). \quad (4.87)$$

In order for the bound to be decreasing with n , we need $2\eta^2 - \epsilon > 0$ leading us to $\epsilon < 2\eta^2$, where η represents the accuracy of the learning algorithm and ϵ the privacy parameter. Thus, for a given $\eta > 0$, as long as the privacy parameter is smaller than $2\eta^2$, we have guaranteed generalisation capabilities for \mathcal{A} with an exponentially decreasing bound. For $\epsilon \leq \eta/2$, it is shown in (Dwork et al., 2015a, Theorem 9) that

$$\mathcal{P}_{SH}(E) \leq 1/4 \exp(-n\eta^2/12).$$

It is easy to check that, for large enough n , Equation (4.87) is tighter if $\epsilon \leq \eta^2 \cdot 23/12$.

It is also possible to see that enforcing differential privacy on some algorithm \mathcal{A} induces

generalisation guarantees similar to those stated in Corollary 5: let \mathcal{A} be an ϵ -DP algorithm such that $\epsilon \leq \sqrt{\frac{\log(1/\gamma)}{2n}}$, and let $\max_y \mathcal{P}_X(E_y) \leq \gamma$ then one has that (cf. (Dwork et al., 2015a, Theorem 11)):

$$\mathcal{P}_{XY}(E) \leq 3\sqrt{\gamma}. \quad (4.88)$$

The results that information-measures provide are qualitatively different. We do not require the imposition of some (possibly strong) privacy criteria on the algorithm to be able to analyse its performances. Instead, we propose a way of estimating how the probabilities we are interested in change (when considering settings where the function depends on the samples), measuring the dependence through information measures.

Focusing on Maximal Leakage, for instance, given an ϵ -DP algorithm, the bound obtained via Equation (4.82) can be tighter than Equation (4.88) for certain values of ϵ . Indeed, let:

$$\epsilon < \frac{\log(3/\sqrt{\gamma})}{n} \leq \sqrt{\frac{\log(1/\gamma)}{2n}}, \quad (4.89)$$

using Equation (4.88) one retrieves *fixed* bound of $3\sqrt{\gamma}$ while, with Corollary 5 and Lemma 12, one obtains the following:

$$\exp(\mathcal{L}(X \rightarrow Y)) \cdot \gamma < \exp(\log(3/\sqrt{\gamma})) \cdot \gamma = 3\sqrt{\gamma}. \quad (4.90)$$

Hence, whenever the privacy parameter is lower than $\frac{\log(3/\sqrt{\gamma})}{n}$ Corollary 5 provides a better bound. Notice that Lemma 12 can be provide a loose upper-bound on Maximal Leakage. Indeed, using Lemma 8, it is possible to see that for classical mechanisms that imply ϵ -Differential Privacy, Maximal Leakage can be much lower than $\epsilon \cdot n$. For instance:

Corollary 25. *Let $h : \mathcal{X}^n \rightarrow \mathbb{R}$ be a function of sensitivity $\frac{1}{n}$ and let $N \sim \text{Lap}(\frac{1}{n\epsilon})$ then the mechanism $\mathcal{M}(x) = h(x) + N$ is ϵ -DP. Without loss of generality we have that $|h(x)| \leq 1$ and thus:*

$$\mathcal{L}(X \rightarrow \mathcal{M}(X)) = \log(1 + \epsilon \cdot n) < \epsilon \cdot n. \quad (4.91)$$

Moreover, the family of algorithms with bounded Maximal Leakage is not restricted to the differentially private ones. It is easy to see, for instance, that whenever there is a deterministic mapping, and ϵ -DP is enforced on it, $\epsilon = +\infty$. Trying to relax it to (ϵ, δ) -Differential Privacy does not help either, as one would need $\delta \geq 1$, rendering it useless in practice. If the algorithm has a bounded range, however, the Maximal Leakage from its input to its output is always guaranteed to be bounded, since $\mathcal{L}(X \rightarrow Y) \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$. These simple observations allow us to also immediately retrieve another result (Dwork et al., 2015b, Theorem 9):

$$\mathcal{P}_{XY}(E) \leq |\mathcal{Y}| \cdot \gamma,$$

where γ is such that $\mathcal{P}_X(E_y) \leq \gamma$ for every y . Indeed, given a random variable Y with bounded

support, $\mathcal{L}(X \rightarrow Y) \leq \log |\mathcal{Y}|$ and from Corollary 5 one has that:

$$\mathcal{P}_{XY}(E) \leq \max_y \mathcal{P}_X(E_y) \exp(\mathcal{L}(X \rightarrow Y)) \leq \gamma \cdot |\mathcal{Y}|. \quad (4.92)$$

This shows that Corollary 5 is more general than Theorems 6 and 9 of (Dwork et al., 2015b). To conclude the comparison let us now state Corollary 21 for functions with sensitivity² c :

$$\mathcal{P}_{XY}(E) \leq 2 \cdot \exp\left(\mathcal{L}(X \rightarrow Y) - \frac{2\eta^2}{c^2 n}\right). \quad (4.93)$$

By contrast, (Dwork et al., 2015b, Corollary 7) states that whenever an algorithm $\mathcal{A}: \mathcal{X}^n \rightarrow \mathcal{Y}$ outputs a function $H: \mathcal{X}^n \rightarrow \mathbb{R}$ of sensitivity c (i.e., $H = \mathcal{A}(X^n)$) and is $\eta/(cn)$ -Differentially Private then, denoting with S a random variable distributed over \mathcal{X}^n one has that:

$$\mathcal{P}_{S, \mathcal{A}(S)}(H(S) - \mathcal{P}_S(H) \geq \eta) \leq 3 \exp\left(-\frac{\eta^2}{c^2 n}\right). \quad (4.94)$$

It is easy to see that Equation (4.93) provides a tighter bound whenever the accuracy $\eta > nc$. Indeed:

$$\mathcal{P}_{S, H}(H(S) - \mathcal{P}_S(H) \geq \eta) \leq 2 \exp\left(\mathcal{L}(S \rightarrow H) - \frac{2\eta^2}{c^2 n}\right) \quad (4.95)$$

$$\leq 2 \exp\left(\frac{\eta}{cn} n - \frac{2\eta^2}{c^2 n}\right) \quad (4.96)$$

$$= 2 \exp\left(\frac{nc\eta - 2\eta^2}{c^2 n}\right) \stackrel{?}{<} 3 \exp\left(-\frac{\eta^2}{c^2 n}\right). \quad (4.97)$$

We are thus asking: “when is it that the quantity on the right-hand side of Equation (4.95) is smaller than the quantity on right-hand side of Equation (4.94)?”. Ignoring the constants in front (respectively 2 and 3) and solving the inequality in (4.97), one finds the condition $\eta > nc$.

4.4.3 Sibson’s Mutual Information, Maximal Leakage and Max Information

Another tool used in the line of work started by Dwork et al. is the concept of max-information. The definition is the following:

Definition 25 ((Dwork et al., 2015b, Definition 10)). *Let X, Y be two random variables jointly distributed according to \mathcal{P}_{XY} and with marginals $\mathcal{P}_X, \mathcal{P}_Y$. The max-information between X and Y , is defined as follows:*

$$I_\infty^M(X, Y) = \log \sup_{(x, y) \in \mathcal{X} \times \mathcal{Y}} \frac{\mathcal{P}_{XY}(\{(x, y)\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{y\})}, \quad (4.98)$$

²The definition of sensitivity selected by Dwork et al. is the following (cf. (Dwork et al., 2015b, Page 6)):

Definition. *Let $f: \mathcal{X}^n \rightarrow \mathbb{R}$. The function f has sensitivity c if, for all $i \in \{1, \dots, n\}$, $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ and $\hat{x} \in \mathcal{X}^n$, then*

$$|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, \hat{x}, \dots, x_n)| < c.$$

while, the γ -approximate max-information is defined as:

$$I_{\infty}^{M,\gamma}(X, Y) = \log \sup_{\mathcal{O} \subseteq \mathcal{X} \times \mathcal{Y}, \mathcal{P}_{XY}(\mathcal{O}) > \gamma} \frac{\mathcal{P}_{XY}(\mathcal{O}) - \gamma}{\mathcal{P}_X \mathcal{P}_Y(\mathcal{O})}. \quad (4.99)$$

Remark 48. Notice that the notation from (Dwork et al., 2015b) was slightly changed to avoid confusion. $I_{\infty}^M(X, Y)$ does **not** correspond to Sibson's Mutual Information of order infinity, I_{∞} , but it actually corresponds to Rényi's D_{∞} , i.e., $I_{\infty}^M(X, Y) = D_{\infty}(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y)$. The quantity $I_{\infty}^{M,\gamma}(X, Y)$ represents instead a new object related to Rényi's D_{∞} and to Rényi's Differential Privacy (Mironov, 2017).

One of the main reasons that led to the definition of approximate max-information is related to the generalisation guarantees it provides, here recalled for convenience.

Lemma 13 ((Dwork et al., 2015b, Theorem 13)). Let X be a random dataset in \mathcal{X}^n and let $\mathcal{A}: \mathcal{X}^n \rightarrow \mathcal{Y}$ be such that for some $\gamma \geq 0$, $I_{\infty}^{M,\gamma}(X, \mathcal{A}(X)) = k$. Let $Y = \mathcal{A}(X)$ then, for any event $E \subseteq \mathcal{X}^n \times \mathcal{Y}$:

$$\mathcal{P}_{XY}(E) \leq e^k \mathcal{P}_X \mathcal{P}_Y(E) + \gamma. \quad (4.100)$$

The result looks quite similar to Corollary 21, but the two measures, Max-Information and Maximal Leakage, although related, can be quite different. In this section we will analyse the connections and differences between the two measures underlining the corresponding implications.

Lemma 14. $I_{\infty}^M(X, Y) \geq D_{\alpha}(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y) \geq I_{\alpha}(X, Y)$ for every $\alpha \in [1, +\infty]$.

Proof. We have that $I_{\infty}^M(X, Y) = D_{\infty}(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y) \geq \mathcal{L}(X \rightarrow Y) \geq I_{\alpha}(X, Y)$ for any $\alpha \in [1, +\infty]$. \square

With respect to γ -approximate max-information instead, one can state the following:

Lemma 15. Let $\mathcal{A}: \mathcal{X}^n \rightarrow \mathcal{Y}$ be a randomised algorithm. Let X be a random variable distributed over \mathcal{X}^n and let $Y = \mathcal{A}(X)$. Suppose X, Y are discrete random variables and denote with \mathcal{P}_{XY} the joint distribution and with $\mathcal{P}_X, \mathcal{P}_Y$ the corresponding marginals. For any $\gamma \in (0, 1)$ and $\alpha \in (1, +\infty]$

$$I_{\infty}^{M,\gamma}(X, \mathcal{A}(X)) \leq \frac{\alpha - 1}{\alpha} I_{\alpha}(X, \mathcal{A}(X)) + \log\left(\frac{1}{\gamma}\right). \quad (4.101)$$

Proof. Fix any $\gamma > 0$. Using (Dwork et al., 2015b, Lemma 18) we have that

$$\text{if } \mathcal{P}_{XY} \left(\left\{ (x, y) \in \mathcal{X} \times \mathcal{Y} \mid \frac{\mathcal{P}_{XY}(\{x, y\})}{\mathcal{P}_X(\{x\}) \mathcal{P}_Y(\{y\})} \geq e^k \right\} \right) \leq \gamma, \text{ then } I_{\infty}^{M,\gamma}(X, Y) \leq k.$$

Denote with $Y = \mathcal{A}(X)$, and with $F = \left\{ (x, y) \in \mathcal{X} \times \mathcal{Y} \mid \frac{\mathcal{P}_{XY}(\{x, y\})}{\mathcal{P}_X(\{x\})\mathcal{P}_Y(\{y\})} \geq \frac{\exp(\frac{\alpha-1}{\alpha} I_\alpha(X, Y))}{\gamma} \right\}$, then

$$\mathcal{P}_{XY}(F) \leq \frac{\mathcal{P}_{XY}\left(\frac{\mathcal{P}_{XY}(\{X, Y\})}{\mathcal{P}_X(\{X\})\mathcal{P}_Y(\{Y\})}\right) \cdot \gamma}{\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X, Y)\right)} \leq \frac{\mathcal{P}_Y\left(\left(\mathcal{P}_X\left(\left(\frac{\mathcal{P}_{Y|X}(\{Y\})}{\mathcal{P}_Y(\{Y\})}\right)^\alpha \mid Y\right)\right)^{1/\alpha}\right) \cdot \gamma}{\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(X, Y)\right)} = \gamma. \quad (4.102)$$

Hence, $I_\infty^{M, \gamma}(X, \mathcal{A}(X)) \leq \log\left(\frac{\exp(\frac{\alpha-1}{\alpha} I_\alpha(X, Y))}{\gamma}\right) = \frac{\alpha-1}{\alpha} I_\alpha(X, Y) + \log\left(\frac{1}{\gamma}\right)$.

Taking the limit of $\alpha \rightarrow \infty$ one also gets that $I_\infty^{M, \gamma}(X, \mathcal{A}(X)) \leq \mathcal{L}(X \rightarrow Y) + \log\left(\frac{1}{\gamma}\right)$. \square

The role of γ in Equation (4.101) can lead to undesirable behaviours of γ -approximate Max-Information. The following example shows how γ -approximate max-Information can be unbounded while, in the discrete case, the Maximal Leakage between two random variables is always guaranteed to be bounded.

Example 8. Let us fix a $\gamma \in (0, 1)$. Suppose $X \sim \text{Ber}(2\gamma)$. We have that $\mathcal{L}(X \rightarrow X) = \log|\mathcal{X}| = \log 2$. For the γ -Approximate Max-information we have: $I_\infty^{M, \gamma}(X, X) \geq \log((2\gamma - \gamma)/\gamma^2) = \log(1/\gamma)$. It can thus be arbitrarily large.

Another interesting characteristic of max-information is that, differently from differential privacy, it can be bounded even if one has deterministic algorithms: this observation is implied by the connection with what in the literature is known as “description length” of an algorithm, and synthesised in the following result (Dwork et al., 2015b):

Let $\mathcal{A} : \mathcal{X}^n \rightarrow \mathcal{Y}$ be a randomised algorithm, for every $\gamma > 0$,

$$I_\infty^{M, \gamma}(\mathcal{A}, n) \leq \log\left(\frac{|\mathcal{Y}|}{\gamma}\right). \quad (4.103)$$

In contrast, with Sibson’s I_α for every $\alpha \in [1, +\infty)$ one has that

$$I_\alpha(X, \mathcal{A}(X)) \leq \mathcal{L}(X \rightarrow \mathcal{A}(X)) \leq \log(|\mathcal{Y}|). \quad (4.104)$$

Since γ is typically small in applications, the corresponding multiplicative factors in the bounds are $(|\mathcal{Y}|/\gamma)$ and $|\mathcal{Y}|$, and the difference between the two quantities can be substantial. It is also worth noticing that Equation (4.103) can be seen as a consequence of Lemma 15 and Equation (4.104). The difference between the two measures is not uniquely restricted to deterministic mechanisms. The following is a simple example of a randomised mapping where Maximal Leakage is smaller than γ -approximate Max-Information, for small γ .

Example 9. Consider $X \sim \text{Ber}(1/2)$ and a random variable Y with support $\mathcal{Y} = \{0, 1, e\}$. Consider also the following randomised mapping: $\mathbb{P}(Y = e|X = x) = \xi$ and $\mathbb{P}(Y = x|X = x) = 1 - \xi$. That is, Y can be interpreted as passing X through a binary erasure channel with erasure probability ξ . In this case, the Maximal Leakage is $\mathcal{L}(X \rightarrow Y) = \log(2 - \xi)$ (Issa et al., 2020); while, for γ -Approximate Max-Information one finds (after a series of computations) that:

$I_{\infty}^{M,\gamma}(X, Y) = \log(2 \cdot \max\{(1 - \xi - \gamma)/(1 - \xi), (1 - \gamma)/(1 + \xi)\})$; It is easy to see how for a fixed ξ and for γ going to 0, Approximate Max-Information approaches $\log 2$ while Maximal Leakage is strictly smaller.

4.5 Adaptive Data Analysis

Other than providing a generalisation of the classical bounds to adaptive settings, Maximal Leakage can also be employed in adaptive data analysis. The model of adaptive composition we will be considering is identical to the setting in (Dwork et al., 2015a,b; Rogers et al., 2016) and defined as follows:

Definition 26 (Adaptive Composition). *Let \mathcal{X} be a set. Let S be a random variable over \mathcal{X}^n . Let $(\mathcal{A}_1, \dots, \mathcal{A}_k)$ be a sequence of algorithms such that $\forall i : 1 \leq i \leq k$*

$$\mathcal{A}_i : \mathcal{X}^n \times \mathcal{Y}_1 \times \dots \times \mathcal{Y}_{i-1} \rightarrow \mathcal{Y}_i.$$

Denote with $Y_1 = \mathcal{A}_1(S), Y_2 = \mathcal{A}_2(S, Y_1), \dots, Y_k = \mathcal{A}_k(S, Y_1, \dots, Y_{k-1})$. The adaptive composition of $(\mathcal{A}_1, \dots, \mathcal{A}_k)$ is an algorithm that takes as an input S and sequentially executes the algorithms $(\mathcal{A}_1, \dots, \mathcal{A}_k)$ as described by the sequence $(Y_i, 1 \leq i \leq k)$.

This level of generality allows us to formalise the behavior of a data analyst who, after viewing the previous outcomes of the analyses performed, decides what to do next. A potential analyst would typically execute a sequence of algorithms that are known to have a certain property (e.g., generalise well) when used without adaptivity. The question we would like to address is the following: is this property also maintained by the adaptive composition of the sequence? The answer is not trivial as, for every i , the outcome of \mathcal{A}_i depends both on S and on the previous outputs (that depend on the data themselves). However, when this property is guaranteed by some measure that composes adaptively itself (like differential privacy or, as we will show soon, Maximal Leakage) then it can be preserved.

Indeed, being robust to post-processing (like every other information measure, due to the Data-Processing Inequality), Maximal Leakage allows us to retain the generalisation guarantees it provides, regardless of how one may manipulate the outcome of the algorithm:

Lemma 16 (Robustness to post-processing). *Let \mathcal{X} be the sample space and let X be distributed over \mathcal{X} . Let \mathcal{Y} and \mathcal{Y}' be output spaces, and consider $\mathcal{A} : \mathcal{X} \rightarrow \mathcal{Y}$ and $\mathcal{B} : \mathcal{Y} \rightarrow \mathcal{Y}'$. Then,*

$$\mathcal{L}(X \rightarrow \mathcal{B}(\mathcal{A}(X))) \leq \mathcal{L}(X \rightarrow \mathcal{A}(X)).$$

The proof is a direct application of the data processing inequality for Maximal Leakage. The useful implication of this result is as follows: in terms of Maximal Leakage, any generalisation guarantees provided by \mathcal{A} cannot be invalidated by further processing the output of \mathcal{A} . Regarding adaptive composition of two algorithms, we retrieve the following:

Lemma 17 (Adaptive Composition of Maximal Leakage). *Let $\mathcal{A} : \mathcal{X} \rightarrow \mathcal{Y}$ be an algorithm such*

that $\mathcal{L}(X \rightarrow \mathcal{A}(X)) \leq k_1$. Let $\mathcal{B}: \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}$ be an algorithm such that:

$$\text{for all } y \in \mathcal{Y}, \mathcal{L}(X \rightarrow \mathcal{B}(X, y)) \leq k_2,$$

then

$$\mathcal{L}(X \rightarrow (\mathcal{A}(X), \mathcal{B}(X, \mathcal{A}(X)))) \leq k_1 + k_2.$$

The proof of this lemma relies crucially on the fact that Maximal Leakage depends on the marginal \mathcal{P}_X only through its support and can be found in Section 4.A, along with the other proofs for this section. In order to generalise the result to the adaptive composition of n algorithms, we need to lift the property stated in Equation (1.55) to more than two random variables.

Lemma 18. *Let $k \geq 1$ and X, A_1, \dots, A_k be random variables.*

$$\mathcal{L}(X \rightarrow (A_1, \dots, A_k)) \leq \mathcal{L}(X \rightarrow A_1) + \mathcal{L}(X \rightarrow A_2 | A_1) + \dots + \mathcal{L}(X \rightarrow A_k | (A_1, \dots, A_{k-1})).$$

The proof can be found in Section 4.A. An immediate application of Lemma 18 leads us to the following result.

Lemma 19. *Consider a sequence of $k \geq 1$ algorithms: $(\mathcal{A}_1, \dots, \mathcal{A}_k)$ where for each $1 \leq i \leq k$, $\mathcal{A}_i: \mathcal{X} \times \mathcal{Y}_1 \times \dots \times \mathcal{Y}_{i-1} \rightarrow \mathcal{Y}_i$. Suppose that for all $1 \leq i \leq k$ and for all $(y_1, \dots, y_{i-1}) \in \mathcal{Y}_1 \times \dots \times \mathcal{Y}_{i-1}$,*

$$\mathcal{L}(X \rightarrow \mathcal{A}_i(X, y_1, \dots, y_{i-1})) \leq j_i.$$

Then, denoting with A_1, \dots, A_k the (random) outputs of the algorithms $(\mathcal{A}_1, \dots, \mathcal{A}_k)$ when executed as described in Definition 26:

$$\mathcal{L}(X \rightarrow (A_1, \dots, A_k)) = \mathcal{L}(X \rightarrow A^k) \leq \sum_{i=1}^k j_i. \quad (4.105)$$

The conclusion to be drawn is straightforward: given a collection of algorithms that have bounded Maximal Leakage (and thus good generalisations capabilities) even if the outcome of one of them is used to inform a subsequent analysis (hence, creating multiple dependencies on the data) the generalisation guarantees of the composition (as defined in Definition 26) can still be maintained.

Another interesting application of Corollary 5 in adaptive settings may be the following (same setting as in (Rogers et al., 2016)): consider the problem of bounding the probability of making a false discovery, when the statistics to apply is selected with some data dependent algorithm \mathcal{T} . In this context, the classical guarantees that allow to upper-bound this probability by the significance value no longer hold. Measuring the information leaked from the data through \mathcal{T} with the Maximal Leakage we retrieve the following:

Corollary 26. *Let $\mathcal{A}: \mathcal{X}^n \rightarrow \mathcal{T}$ be a data dependent algorithm for selecting a test statistic $t \in \mathcal{T}$.*

Let X be a random dataset over \mathcal{X}^n . Suppose that $\sigma \in [0, 1]$ is the significance level chosen to control the false discovery probability. Denote with E the event that \mathcal{A} selects a statistic such that the null hypothesis is true but its p -value is at most σ . Then,

$$\mathbb{P}(E) \leq \exp(\mathcal{L}(X \rightarrow \mathcal{A}(X))) \cdot \sigma.$$

If the analyst wishes to achieve a bound of δ on the probability of making a false discovery in adaptive settings, the significance level σ to be used should be no higher than $\frac{\delta}{\exp(\mathcal{L}(X \rightarrow \mathcal{A}(X)))}$. Once again, if \mathcal{A} is independent from X , we recover the known bound of σ .

Appendix

4.A Properties of Maximal Leakage

In this appendix we will provide proofs for the properties of Maximal Leakage. Let us start with the Adaptive Composition of the measure and let us recall the statement for reference:

Lemma. *Let $\mathcal{A} : \mathcal{X} \rightarrow \mathcal{Y}$ be an algorithm such that $\mathcal{L}(X \rightarrow \mathcal{A}(X)) \leq k_1$. Let $\mathcal{B} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}$ be an algorithm such that for all $y \in \mathcal{Y}$, $\mathcal{L}(X \rightarrow \mathcal{B}(X, y)) \leq k_2$. Then $\mathcal{L}(X \rightarrow (\mathcal{A}(X), \mathcal{B}(X, \mathcal{A}(X)))) \leq k_1 + k_2$.*

The proof of this lemma relies crucially on the fact that maximal leakage depends on the marginal \mathcal{P}_X only through its support.

Proof. Let us denote with R_X the support of a random variable X . If we consider the second constraint in our assumption and denoting with $Z_y = \mathcal{B}(X, y)$, we get:

$$\forall y \in \mathcal{Y} \mathcal{L}(X \rightarrow Z_y) \leq k_2 \iff \forall y \in \mathcal{Y} \sum_{z_y \in R_{Z_y}} \max_{x \in R_X} \mathbb{P}(z_y | x) \leq \exp(k_2) \quad (4.106)$$

$$\iff \quad (4.107)$$

$$\forall y \in \mathcal{Y} \sum_{z_y \in R_{Z_y}} \max_{x \in R_X} \mathbb{P}(z | x, y) \leq \exp(k_2). \quad (4.108)$$

The last step holds, since every y generates a family of conditional distributions $\mathbb{P}(z_y | x)$ through \mathcal{B} and this probability is just $\mathbb{P}(z | x, y)$, with $z = \mathcal{B}(x, y)$. Using this observation in the conditional leakage of (1.55):

$$\mathcal{L}(X \rightarrow Z | Y) = \log \max_{y \in R_Y} \sum_{z \in R_{Z|Y=y}} \max_{x \in R_{X|Y=y}} \mathbb{P}(z | x, y) \quad (4.109)$$

$$\leq \log \max_{y \in R_Y} \sum_{z \in R_{Z|Y=y}} \max_{x \in R_X} \mathbb{P}(z | x, y) \quad (4.110)$$

$$\leq \log \max_{y \in R_Y} \exp(k_2) \quad (4.111)$$

$$= k_2, \quad (4.112)$$

leading us to the desired bound. □

Let us now show the generalization of this property to k random variables. The statement reads:

Lemma. *Let $k \geq 1$ and X, A_1, \dots, A_k be random variables.*

$$\mathcal{L}(X \rightarrow (A_1, \dots, A_k)) \leq \mathcal{L}(X \rightarrow A_1) + \mathcal{L}(X \rightarrow A_2 | A_1) + \dots + \mathcal{L}(X \rightarrow A_k | (A_1, \dots, A_{k-1})). \quad (4.113)$$

Proof.

$$\mathcal{L}(X \rightarrow (A_1, \dots, A_k)) = \mathcal{L}(X \rightarrow A^k) = \mathcal{L}(X \rightarrow (A^{k-1}, A_k)) \quad (4.114)$$

$$\leq \mathcal{L}(X \rightarrow A^{k-1}) + \mathcal{L}(X \rightarrow A_k | A^{k-1}), \quad (4.115)$$

then the result follows from recursively applying the same argument to $\mathcal{L}(X \rightarrow A^{k-1})$. \square

4.B Examples

4.B.1 Proof of Lemma 8

We will now compute the value of the Maximal Leakage for an additive noise mechanism, where the noise is a Laplace random variable. Recall the statement of Lemma 8 is:

Lemma. *Let $h : \mathcal{X}^n \rightarrow \mathbb{R}$ be a function such that $h(x) \in [a, c]$, $a < c \forall x \in \mathcal{X}^n$. The mechanism $\mathcal{M}(x) = h(x) + N$ where $N \sim \text{Lap}(b)$ is such that:*

$$\mathcal{L}(X \rightarrow \mathcal{M}(X)) = \log\left(1 + \frac{(c-a)}{2b}\right) \quad (4.116)$$

Proof. Let $Y = g(X) + N$, starting from Equation (1.53),

$$\exp(\mathcal{L}(X \rightarrow Y)) = \int_{\mathbb{R}} \sup_{x: f_X(x) > 0} f_{Y|X}(y|x) dy \quad (4.117)$$

$$= \int_{\mathbb{R}} \sup_{x: f_X(x) > 0} f_N(y - h(x)) dy \quad (4.118)$$

$$= \frac{1}{2b} \left(\int_{-\infty}^{+\infty} \sup_{x: \mathcal{P}_X(x) > 0} \exp\left(\frac{-|y - h(x)|}{b}\right) dy \right) \quad (4.119)$$

$$= \frac{1}{2b} \left(\int_{-\infty}^a \exp\left(\frac{-|y - a|}{b}\right) dy + \int_a^c dy \right) + \frac{1}{2b} \left(\int_c^{+\infty} \exp\left(\frac{-|y - c|}{b}\right) dy \right) \quad (4.120)$$

$$= \frac{1}{2b} \left(\int_{-\infty}^0 \exp\left(\frac{-|z|}{b}\right) dz + (c - a) \right) + \frac{1}{2b} \left(\int_0^{+\infty} \exp\left(\frac{-|w|}{b}\right) dw \right) \quad (4.121)$$

$$= \frac{1}{2b} \left((c - a) + 2 \int_0^{+\infty} \exp\left(\frac{-w}{b}\right) dw \right) \quad (4.122)$$

$$= \frac{1}{2b} ((c - a) + 2b) = \left(1 + \frac{(c - a)}{2b}\right). \quad (4.123)$$

\square

The proofs of the other additive noise mechanisms (Gaussian and Exponential) follow an approach similar to the one just showed.

Proof of Corollary 22

Suppose the hypothesis space is countable and let $k := |\mathcal{H}|$ (could be infinite). Suppose also that $\mathcal{P}_{N_i}(N_i) = b_i$ (cf. Xu and Raginsky (2017c)), with N_i being the noise added to the i -th hypothesis. Since the choice of the hypothesis depends only on the noisy empirical errors, the following is a Markov Chain $S - (L_S(h_i))_{i \in [k]} - (L_S(h_i) + N_i)_{i \in [k]} - H$. Then by the data-processing inequality for Maximal Leakage:

$$\mathcal{L}(S \rightarrow H) \leq \mathcal{L}\left((L_S(h_i))_{i \in [k]} \rightarrow (L_S(h_i) + N_i)_{i \in [k]}\right). \quad (4.124)$$

Also, denoting with $X_i = L_S(h_i)$ and with $Y_i = X_i + N_i$:

$$\exp(\mathcal{L}((X_1, \dots, X_k) \rightarrow (Y_1, \dots, Y_k))) = \int \cdots \int_{-\infty}^{+\infty} \max_{x^n} f(y^n | x^n) dy^n \quad (4.125)$$

$$= \int \cdots \int_{-\infty}^{+\infty} \max_{x^n} \left(\prod_{i=1}^k f_{N_i}(y_i - x_i) \right) dy^n \quad (4.126)$$

$$= \int \cdots \int_{-\infty}^{+\infty} \max_{x^n} \left(\prod_{i=1}^k \frac{1}{b_i} e^{-(y_i - x_i)/b_i} \right) dy^n \quad (4.127)$$

$$= \prod_{i=1}^k \int_{-\infty}^{+\infty} \max_{x_i} \left(\frac{1}{b_i} e^{-(y_i - x_i)/b_i} \right) dy \quad (4.128)$$

$$= \prod_{i=1}^k \left(1 + \frac{1}{b_i} \right). \quad (4.129)$$

Equation (4.129), along with Corollary 21, implies that:

$$\mathcal{P}_{SH}(\text{gen-err}(\mathcal{A}) \geq \eta) \leq 2 \exp(\mathcal{L}(S \rightarrow H) - 2n\eta^2) \quad (4.130)$$

$$= 2 \exp\left(\sum_{i=1}^k \log\left(1 + \frac{1}{b_i}\right) - 2n\eta^2\right). \quad (4.131)$$

Now, suppose that $b_i = i^{1.1}/n^{1/3}$,

$$\mathcal{L}(S \rightarrow H) \leq \sum_{i=1}^k \log(1 + n^{1/3}/i^{1.1}) \quad (4.132)$$

$$\leq n^{1/3} \sum_{i=1}^{+\infty} \frac{1}{i^{1.1}} \quad (4.133)$$

$$\leq (n^{1/3}) \cdot 11. \quad (4.134)$$

We have that

$$\mathcal{P}_{SH}(\text{gen-err} \geq \eta) \leq 2 \exp(-n(2\eta^2 - 11/n^{2/3})). \quad (4.135)$$

5 Bayesian Risk in Estimation Procedures

5.1 Introduction

In this chapter¹ we consider a different type of application for the results presented in Chapter 3. We consider the problem of parameter estimation in a Bayesian setting. More precisely, we propose an approach to *lower-bounding* the Bayesian risk leveraging a variety of information measures. Through a sequence of tools, we can shift the focus from the Bayesian risk itself to the computation of two objects:

1. an information measure (*e.g.*, Sibson's α -Mutual Information, φ -Mutual Information, etc.);
2. a functional of the probability of some event under independence (*e.g.*, a small-ball probability (Xu and Raginsky, 2017a));

We look at the problem through an information-theoretic lens, similarly to (Xu and Raginsky, 2017a). We thus treat the parameter to be estimated as a message sent through a channel. This allows us to include frameworks where, in a distributed fashion, m processors observe noisy samples of this parameter. The processors will then send a version of their observations to a central node. The central node will then proceed to estimate the parameter.

An advantage of using this type of bounds is that one can render the functional in Item 2 (*e.g.*, the small-ball probability) independent of the specific estimator. Similarly, the information measure can also be rendered independent of the estimator via Data Processing Inequalities. Therefore, these lower bounds can be applied to any estimation framework that matches this one, regardless of the specific choice of the estimator.

It is important to notice that, although the problem can be interpreted as a transmission problem, a fundamental difference is that the size of the quantised messages may not grow with the number of samples. This might render the reconstruction of the samples impossible but the estimation of the parameter may remain feasible (Xu and Raginsky, 2017a). Our main focus will not be on asymptotic results but rather on finite number of samples lower-bounds.

¹Part of the content of this chapter has been presented at the International Symposium on Information Theory 2021 (Esposito and Gastpar, 2021)

5.2 Problem Setting

Let \mathcal{W} denote the parameter space and assume that we have access to a prior distribution over \mathcal{W} , \mathcal{P}_W . Suppose that we observe $W \sim \mathcal{P}_W$ through the family of distributions $\mathcal{P} = \{\mathcal{P}_{X|W=w} : w \in \mathcal{W}\}$. Given a function $\phi : \mathcal{X} \rightarrow \hat{\mathcal{W}}$ one can then estimate W from $X \sim \mathcal{P}_{X|W}$ via $\phi(X) = \hat{W}$. Let us denote with $\ell : \mathcal{W} \times \hat{\mathcal{W}} \rightarrow \mathbb{R}^+$ a loss function, the Bayesian risk is defined as:

$$R = \inf_{\phi} \mathcal{P}_{W\hat{W}}(\ell(W, \phi(X))) = \inf_{\phi} \mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})). \quad (5.1)$$

Our purpose is to lower-bound R using the tools described in the previous chapters. To this end, among other tools, we will be using a simple Markov's inequality approach: *i.e.*, for every estimator ϕ and $\rho \geq 0$, one can do the following

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \geq \rho) \geq \rho (\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \geq \rho)). \quad (5.2)$$

This allows us to shift the focus from the expected value of ℓ to the probability that ℓ is larger than some constant ρ . We aim at providing a lower bound on the risk R that is *as independent as possible* of the specific choice of ϕ . With that drive, one can use the results presented in Chapter 3 to manipulate the probability of the event under the joint measure $\mathcal{P}_{W\hat{W}}$. In particular, a family of suitable results is the one involving Sibson's α -Mutual Information, as it naturally allows us to "substitute" $\mathcal{P}_{W\hat{W}}(E)$ with a function of $\max_{\hat{w}} \mathcal{P}_W(E_{\hat{w}})$ (that will thus depend on ϕ only through the support of \hat{W}) and the Sibson's Mutual Information $I_{\alpha}(W, \hat{W})$. More precisely, let us denote $E = \{\ell(W, \hat{W}) \leq \rho\}$ and with $L_W(\hat{W}, \rho) = \mathcal{P}_W \mathcal{P}_{\hat{W}}(\ell(W, \hat{W}) \leq \rho)$, our desideratum will be a lower-bound of the following form:

$$R \geq \varpi \left(\frac{d\mathcal{P}_{W\hat{W}}}{d\mathcal{P}_W \mathcal{P}_{\hat{W}}} \right) \varphi(\rho) \vartheta(\mathcal{P}_W \mathcal{P}_{\hat{W}}, E), \quad (5.3)$$

with the purpose of then rendering the right-hand side of Equation (5.3) as independent as possible of the estimator ϕ . A functional ϑ of particular interest to us, is the one that leads to (a function of) the so-called small-ball probability

$$L_W(\rho) = \sup_{\hat{w} \in \hat{\mathcal{W}}} L_W(\hat{w}, \rho) = \sup_{\hat{w} \in \hat{\mathcal{W}}} \mathcal{P}_W(\ell(W, \hat{w}) \leq \rho). \quad (5.4)$$

As in Chapter 3, the choice of ϑ (and of φ) will depend on the choice of ϖ and vice versa. In the following sections, we will explore different choices of ϖ and ϑ that lead to interesting results in the field.

5.3 The lower-bounds

The very first result one can provide stems from an immediate application of Corollary 4 in conjunction with Equation (5.2) (and that makes the $\frac{1}{\beta}$ -th power of the small-ball probability appear):

Theorem 25. Consider the Bayesian framework described in Section 5.2. The following must hold for every $\alpha > 1$ and $\rho > 0$:

$$R \geq \rho \left(1 - \exp \left(\frac{\alpha - 1}{\alpha} (I_\alpha(W, X) + \log(L_W(\rho))) \right) \right). \quad (5.5)$$

Proof. We have that

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \leq \rho) \leq \left(\sup_{\hat{w} \in \hat{\mathcal{W}}} \mathcal{P}_W(\ell(W, \hat{w}) \leq \rho) \right)^{\frac{\alpha-1}{\alpha}} \exp \left(\frac{\alpha-1}{\alpha} I_\alpha(W, \hat{W}) \right) \quad (5.6)$$

$$= \exp \left(\frac{\alpha-1}{\alpha} (I_\alpha(W, \hat{W}) + \log(L_W(\rho))) \right) \quad (5.7)$$

$$\leq \exp \left(\frac{\alpha-1}{\alpha} (I_\alpha(W, X) + \log(L_W(\rho))) \right). \quad (5.8)$$

Equation (5.6) follows from Corollary 4, Equation (5.8) follows from the Data Processing Inequality for I_α and the Markov Chain $W - X - Y - \hat{W}$. Moreover, starting from Equation (5.2) and using Markov's inequality one has that

$$R \geq \rho \cdot \mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \geq \rho) \quad (5.9)$$

$$= \rho \cdot (1 - \mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \leq \rho)). \quad (5.10)$$

The statement follows from lower bounding Equation (5.10) using Equation (5.8). \square

Two remarks are in order:

- It is important to notice that the behaviour of Equation (5.5) is fundamentally different from (Xu and Raginsky, 2017a, Theorem 1). In (Xu and Raginsky, 2017a, Theorem 1) the dependence is linear with respect to the Mutual Information and logarithmic in $L_W(\rho)$ while in Theorem 25 there is an exponential dependence on I_α and linear in $L_W(\rho)$.
- Theorem 25 introduces a new parameter $\alpha > 1$ to optimise over. The presence of α leads to a trade-off between the two quantities for a given ρ , $I_\alpha(W, X)$ and $L_W(\rho)$: $\frac{\alpha-1}{\alpha} I_\alpha(W, X)$ will increase with α whereas $L_W(\rho)^{\frac{\alpha-1}{\alpha}}$ will decrease with α .

Taking the limit of $\alpha \rightarrow \infty$ in Theorem 25 allows us to get rid of the parameter α while maintaining the same type of behaviour. This also brings in Maximal Leakage:

Corollary 27. Consider the Bayesian framework described in Section 5.2.

$$R \geq \sup_{\rho > 0} \rho \left(1 - \exp(\mathcal{L}(W \rightarrow X) + \log(L_W(\rho))) \right). \quad (5.11)$$

An interesting characteristic of Corollary 27 is that $\mathcal{L}(W \rightarrow X)$ depends on W only through the support. This allows us to provide, essentially for free, an even more general lower-bound on the risk. Indeed, ignoring $L_W(\rho)$ for a moment, for a fixed family of $\mathcal{P}_{X|W}$, $\mathcal{L}(W \rightarrow X)$ has the

same value regardless of \mathcal{P}_W (as long as the support of W remains the same). We can also walk the same path undertaken in Chapter 3 and derive a variety of lower-bounds involving a variety of information-measures.

Theorem 26. *Consider the Bayesian framework described in Section 5.2. Let $\varphi : [0, +\infty) \rightarrow \mathbb{R}$ be an increasing convex function and suppose that the generalized inverse, defined as $\varphi^{-1}(y) = \inf\{t \geq 0 : \varphi(t) > y\}$, exists. Then the following must hold for every $\rho > 0$ and every estimator \hat{W} :*

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \rho \left(1 - L_W(\hat{W}, \rho) \cdot \varphi^{-1} \left(\frac{I_\varphi(W, \hat{W}) + (1 - L_W(\hat{W}, \rho))\varphi^*(0)}{L_W(\hat{W}, \rho)} \right) \right). \quad (5.12)$$

Moreover, if $\varphi^*(0) \leq 0$, the bound simplifies to

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \rho \left(1 - L_W(\hat{W}, \rho) \cdot \varphi^{-1} \left(\frac{I_\varphi(W, \hat{W})}{L_W(\hat{W}, \rho)} \right) \right). \quad (5.13)$$

Proof. One has that for every function φ with the desired properties, given $E = \{\ell(W, \hat{W}) \leq \rho\}$, Theorem 14 tells us that:

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \leq \rho) \leq L_W(\hat{W}, \rho) \cdot \varphi^{-1} \left(\frac{I_\varphi(W, \hat{W}) + (1 - L_W(\hat{W}, \rho))\varphi^*(0)}{L_W(\hat{W}, \rho)} \right). \quad (5.14)$$

In particular, when $\varphi^*(0) \leq 0$ the bound reduces to

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \leq \rho) \leq L_W(\hat{W}, \rho) \cdot \varphi^{-1} \left(\frac{I_\varphi(W, \hat{W})}{L_W(\hat{W}, \rho)} \right). \quad (5.15)$$

Rewriting $\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \geq \rho)$ as $1 - \mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \leq \rho)$ and combining this with Equation (5.2) and Equation (5.14) concludes the proof. \square

Remark 49. *An approach closely connected to the one just proposed with φ -Divergences is in (Chen et al., 2016). The authors therein focused on the notion of φ -informativity (cf. (Csiszár, 1972b)) and leveraged the data processing inequality similarly to Equation (3.36). In particular, φ -informativities are more general than the φ -Mutual Informations considered in this work (cf. Definition 11) and they can potentially lead to tighter results. More precisely, φ -informativities (similarly to Sibson's α -Mutual Information) are defined as follows:*

$$\hat{I}_\varphi(X, Y) = \inf_{\mathcal{Q}_Y} D_\varphi(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{Q}_Y). \quad (5.16)$$

Clearly $\hat{I}_\varphi(X, Y) \leq I_\varphi(X, Y)$. Moreover, the authors of (Chen et al., 2016) in order to achieve lower-bounds on the Risk for losses between 0 and 1, use the Data-Processing Inequality of \hat{I}_φ along with directly inverting the resulting binary divergence (thus obtaining a tighter bound that, however, does not always admit a closed-form expression). In this document, we were looking for bounds that admitted a closed-form expression and that represented a generalisation of Hölder's inequality for general convex functions. Indeed, following the steps that lead to Theorem 14

via Data-Processing Inequality (cf. Equation (3.194)), in order to achieve such form the binary φ -Divergence is lower-bounded further. This naturally leads to a worse bound connecting $\mathcal{P}_{XY}(E)$, $\mathcal{P}_X\mathcal{P}_Y(E)$ and $I_\varphi(X, Y)$ (with respect to (Chen et al., 2016, Theorem 3.2)) that possesses a different shape and is easier to analyse. The technique used to provide lower-bounds on the Bayesian risk for general non-negative losses (cf. (Chen et al., 2016, Section 4)) is, however, different. It is unclear whether the results provided in this work are equivalent (or weaker) with respect to those obtained in (Chen et al., 2016).

Although Theorem 26 represents a quite general result, in order to apply it to the Bayesian Risk setting (and provide an estimator-independent lower-bound) one has to carefully select φ . In particular, one has to render the right-hand side of Equation (5.12) (or Equation (5.13)) independent of $\hat{W} = \phi(X)$. In order to do that, the following two quantities need to be rendered independent of \hat{W} :

1. The information-measure (e.g., through the data-processing inequality $I_\varphi(W, \hat{W}) \leq I_\varphi(W, X)$);
2. The quantity $L_W(\hat{W}, \rho)$ (which can be easily upper-bounded in the following way: $L_W(\hat{W}, \rho) \leq \sup_{\hat{w}} L_W(\hat{w}, \rho) = L_W(\rho)$).

For simplicity, consider Equation (5.13) and introduce the following object

$$G_\varphi(I_\varphi, L_W) = L_W(\hat{W}, \rho) \cdot \varphi^{-1} \left(\frac{I_\varphi(W, \hat{W})}{L_W(\hat{W}, \rho)} \right). \quad (5.17)$$

To use the two inequalities just stated in Item 1) and Item 2), one needs that for a given choice of φ , $G_\varphi(I_\varphi, L_W)$ is increasing in I_φ for a given value of L_W and vice versa. This lets us further lower-bound Equation (5.13) and render the quantity independent of the specific choice of ϕ . Hence, starting from Equation (5.1) one can provide a lower bound on the risk R that is independent of ϕ .

Let us now look at some specific choices of φ such that G_φ satisfies the desired properties and, thus, for which a bound on the Bayesian risk can be retrieved.

Corollary 28. *Consider the Bayesian framework described in Section 5.2. The following must hold for every $p > 1$ and $\rho > 0$:*

$$R \geq \rho \left(1 - L_W(\rho)^{\frac{p-1}{p}} \cdot ((p-1)\mathcal{H}_p(W, X) + 1)^{\frac{1}{p}} \right). \quad (5.18)$$

Proof. The statement follows from Theorem 26 with $\varphi(x) = \frac{x^{p-1}}{p-1}$ (or, from Corollary 6). Hence, for every estimator $\hat{W} = \phi(X^n)$,

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W}) \geq \rho) \leq L_W(\hat{W}, \rho) \cdot \varphi^{-1} \left(\frac{I_\varphi(W, \hat{W}) + (1 - L_W(\hat{W}, \rho))\varphi^*(0)}{L_W(\hat{W}, \rho)} \right) \quad (5.19)$$

$$\leq L_W(\rho)^{\frac{p-1}{p}} \left((p-1)\mathcal{H}_p(W, X) + 1 \right)^{\frac{1}{p}}, \quad (5.20)$$

where Equation (5.20) follows from the data-processing inequality for φ -divergences. One thus retrieves that for every estimator \hat{W}

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \rho \left(1 - L_W(\rho)^{\frac{p-1}{p}} \left((p-1)\mathcal{H}_p(W, X) + 1 \right)^{\frac{1}{p}} \right). \quad (5.21)$$

Since the right-hand side of Equation (5.21) is independent of $\hat{W} = \phi(X)$ one can use it to lower-bound the risk R . \square

Restricting the choice of φ to this family of polynomials one can thus state the following lower-bound on the risk:

$$R \geq \sup_{\rho > 0} \sup_{p > 1} \rho \left(1 - L_W(\rho)^{\frac{p-1}{p}} \cdot \left((p-1)\mathcal{H}_p(W, \hat{W}) + 1 \right)^{\frac{1}{p}} \right). \quad (5.22)$$

Remark 50. Using the one-to-one mapping connecting Hellinger divergences and Rényi's α -Divergence, the bound above can be re-written as follows:

$$R \geq \sup_{\rho > 0} \sup_{\alpha > 1} \rho \left(1 - L_W(\rho)^{\frac{\alpha-1}{\alpha}} \cdot \exp \left(\frac{\alpha-1}{\alpha} D_\alpha(\mathcal{P}_{W\hat{W}} \| \mathcal{P}_W \mathcal{P}_{\hat{W}}) \right) \right). \quad (5.23)$$

Clearly, a number of results can be derived from Theorem 26. Each of them with potentially interesting applications in a specific Bayesian Estimation setting. In this chapter we will mostly focus on Sibson's α -Mutual Information and Hellinger p -Divergences (or, essentially, Rényi's α -Divergences). In the spirit of leveraging the generality of Theorem 26, we also provide a bound involving a novel information measure (strongly inspired by the so-called E_γ -Divergence (cf. (Sason and Verdú, 2016, Equation (66)), (Polyanskiy et al., 2010, Page 2314)), also known in the literature as the Hockey-stick Divergence, whose application in this framework has been explored in (Asoodeh et al., 2021)). The information measure is the following:

Definition 27. Let (Ω, \mathcal{F}) be a measurable space and let μ and ν be two probability measures defined on the space. Denote with $\varphi(x) = \max\{0, \delta x - \gamma\}$ with $\delta > 0$ and $\gamma \geq \delta$. The function $\varphi(x)$ is convex, increasing and is such that $\varphi(1) = 0$. Assume that $\nu \ll \mu$, then one can define the following

$$E_{\gamma, \delta}(\nu \| \mu) = D_\varphi(\nu \| \mu). \quad (5.24)$$

Moreover, whenever $\Omega = \mathcal{P}_{XY}$, $\nu = \mathcal{P}_{XY}$ and $\mu = \mathcal{P}_X \mathcal{P}_Y$ we denote (with a slight abuse of notation) $E_{\gamma, \delta}(\mathcal{P}_{XY} \| \mathcal{P}_{XY})$ with $E_{\gamma, \delta}(X, Y)$. If $\delta = 1$ then one recovers the usual E_γ -divergence.

Leveraging it, one can provide the following result in this framework:

Corollary 29. Consider the Bayesian framework described in Section 5.2. The following must hold for every $\delta > 0$, $\gamma \geq \delta$, and $\rho > 0$:

$$R \geq \rho \left(1 - \frac{E_{\gamma, \delta}(W, \hat{W}) - \gamma L_W(\rho)}{\delta} \right). \quad (5.25)$$

Proof. Let $\varphi(x) = \max\{0, \delta x - \gamma\}$ in Theorem 26, along with the fact that $\varphi^*(0) \leq 0$ one has that for every estimator $\hat{W} = \phi(X^n)$,

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \rho \left(1 - \frac{E_{\gamma, \delta}(W, \hat{W}) + \gamma L_W(\hat{W}, \rho)}{\delta} \right) \quad (5.26)$$

$$\geq \rho \left(1 - \frac{E_{\gamma, \delta}(W, X) + \gamma L_W(\rho)}{\delta} \right). \quad (5.27)$$

Since Equation (5.27) is independent of $\hat{W} = \phi(X)$ one can use it to lower-bound the risk R . \square

One can thus retrieve the following lower-bound on the risk:

$$R \geq \sup_{\rho > 0} \sup_{\delta > 0, \gamma \geq \delta} \rho \left(1 - \frac{E_{\gamma, \delta}(W, \hat{W}) + \gamma L_W(\rho)}{\beta} \right). \quad (5.28)$$

Remark 51. *Setting $\delta = 1$ in Equation (5.25) one recovers (Asoodeh et al., 2021, Remark 1). In fact, by introducing an additional degree of freedom through the δ parameter in Equation (5.28), the resulting lower-bound can only be tighter than (Asoodeh et al., 2021, Remark 1).*

Using these results one can provide meaningful lower-bounds on the Risk in a variety of settings of interest, as we will see in Section 5.4.

5.4 Examples

In this section we apply the results presented in the previous section to three estimation settings. The first two are classical settings: estimation of the mean of a Bernoulli random variable and estimation of the mean of a Gaussian random variable with different prior distributions over the mean. The loss function is going to be in both cases the L^1 -distance. The third example shows how one can provide a meaningful lower-bound for the ‘‘Hide-and-seek’’ problem presented in (Shamir, 2014).

5.4.1 Bernoulli Bias

Example 1. *Suppose that $W \sim \mathcal{U}([0, 1])$ and that for each $i \in [n]$, $X_i | W = w \sim \text{Ber}(w)$. Also, assume that $\ell(w, \hat{w}) = |w - \hat{w}|$.*

It is easy to see that using the sample mean estimator *i.e.*, $\hat{W} = \frac{1}{n} \sum_{i=1}^n X_i$, one has that

$$R \leq \frac{1}{\sqrt{6n}}. \quad (5.29)$$

Let us now lower-bound the risk in this setting. Without making any further assumptions on

Chapter 5. Bayesian Risk in Estimation Procedures

W , we can only trivially upper-bound $L_W(\rho)$ with 2ρ . Moreover one has that

$$\mathcal{P}_{X^n|W=w}(x^n) = w^k(1-w)^{(n-k)}$$

where k is the hamming weight of x^n . As per assumption, $\mathcal{P}_W(w) = 1$ if $0 \leq w \leq 1$ and, consequently, one has that

$$\mathcal{P}_{W|X^n=x^n}(w) = (n+1) \binom{n}{k} (1-w)^{n-k} w^k.$$

One can now proceed to analyse the Maximal Leakage in this setting.

$$\mathcal{L}(W \rightarrow X^n) = \log \sum_{x^n} \max_w \mathcal{P}_{X^n|W=w}(x^n) \quad (5.30)$$

$$= \log \sum_{k=0}^n \binom{n}{k} \max_w w^k (1-w)^{n-k} \quad (5.31)$$

$$= \log \sum_{k=0}^n \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k} \quad (5.32)$$

$$\leq \log \left(2 + \sum_{k=1}^{n-1} \sqrt{\frac{n}{2\pi k(n-k)}} \right) \quad (5.33)$$

$$\leq \log \left(2 + \sqrt{\frac{\pi n}{2}} \right). \quad (5.34)$$

Substituting Equation (5.34) in Equation (5.11) provides us with the following lower-bound on the risk:

$$R \geq \sup_{\rho>0} \rho (1 - \exp(-\mathcal{L}(W \rightarrow X^n))) L_W(\rho) \quad (5.35)$$

$$\geq \sup_{\rho>0} \rho \left(1 - \left(2 + \sqrt{\frac{\pi n}{2}} \right) 2\rho \right). \quad (5.36)$$

The quantity in Equation (5.36) is a convex function of ρ and thus we can maximise it. In particular, the maximiser is $\hat{\rho} = \frac{1}{4(2 + \sqrt{\frac{\pi n}{2}})}$ and plugging it in Equation (5.36) one gets the following:

$$R \geq \frac{1}{8 \left(2 + \sqrt{\frac{\pi n}{2}} \right)}, \quad (5.37)$$

which, for n large enough (*i.e.*, $n \geq 127/\pi \approx 41$), can be further lower-bounded as follows

$$R \geq \frac{1}{5\sqrt{2\pi n}}.$$

Surprisingly, Maximal Leakage already offers a lower-bound that matches the upper-bound up to a constant (cf. Equation (5.29)) without any extra machinery. Equation (5.37) provides a larger lower-bound than the one provided using Mutual Information (cf. (Xu and Raginsky,

2017a, Corollary 2)) for $n \geq 1$. Moreover, the proof in (Xu and Raginsky, 2017a) needs a more complicated setting involving a conditioning with respect to an independent copy of X^n and can only provide an *asymptotic* lower bound on the risk of $1/(16\sqrt{2\pi n})$ (that thus, only holds for n large enough).

On the contrary, given the closed-form expression, Maximal Leakage can be quite easy to compute or upper-bound. Moreover, the information measure depends on \mathcal{P}_W only through the support. This means that if one has access to an upper-bound on $L_W(\rho)$ that does not employ any knowledge of \mathcal{P}_W except for the support (*e.g.*, if W were to be discrete, an upper-bound of 1 over the probability mass function could suffice) the resulting lower-bound on the risk (in this example), would apply to any W whose support is the interval $[0, 1]$.

One can also provide a more general lower-bound involving I_α . Indeed, one has that, in this setting:

$$\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W, X^n)\right) = \sum_{k=0}^n \binom{n}{k} \left(\frac{\Gamma(k\alpha+1)\Gamma((n-k)\alpha+1)}{\Gamma(n\alpha+2)}\right)^{\frac{1}{\alpha}}. \quad (5.38)$$

Plugging Equation (5.38) in Equation (5.5) one obtains the following lower-bound on the risk:

$$R \geq \sup_{\rho>0} \sup_{\alpha>1} \rho \left(1 - (2\rho)^{\frac{\alpha-1}{\alpha}} \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W, X^n)\right)\right). \quad (5.39)$$

The lower-bound in Equation (5.39) can clearly only improve the one provided in Equation (5.36), as $\mathcal{L}(W \rightarrow X^n) = I_\infty(W, X^n)$. However, differently from Equation (5.36), it does not admit a close-form expression and needs to be computed numerically in order to assess how far it is from the upper-bound. Similarly, one could try to employ a lower-bound that includes Hellinger- p Divergences. The lower-bound on the risk induced by Corollary 28 is given by

$$R \geq \sup_{\rho>0} \sup_{p>1} \rho \left(1 - (2\rho)^{\frac{p-1}{p}} \cdot (\mathcal{H}_p(W, X^n))^{\frac{1}{p}}\right). \quad (5.40)$$

The expressions represented in Equation (5.40) and Equation (5.39) are extremely similar. Moreover, for a given $\alpha > 1$, one has that:

$$\exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W, X^n)\right) = \exp\left(\frac{\alpha-1}{\alpha} \inf_{\mathcal{Q}_{X^n}} D_\alpha(\mathcal{P}_{WX^n} \parallel \mathcal{P}_W \mathcal{Q}_{X^n})\right) \quad (5.41)$$

$$\leq \exp\left(\frac{\alpha-1}{\alpha} D_\alpha(\mathcal{P}_{WX^n} \parallel \mathcal{P}_W \mathcal{P}_{X^n})\right) \quad (5.42)$$

$$= (\mathcal{H}_\alpha(W, X^n))^{\frac{1}{\alpha}}. \quad (5.43)$$

Thus, one can easily argue that Equation (5.39) will always provide a larger lower-bound than Equation (5.40) and indeed this is confirmed in the simulations. It is also true that, for some values of p , one can actually provide nice closed-form expressions for the lower-bound

provided by Equation (5.40). Indeed, in general, one has that:

$$((p-1)\mathcal{H}_p(W, X^n) + 1) = \left\| \frac{d\mathcal{P}_{WX^n}}{d\mathcal{P}_W\mathcal{P}_{X^n}} \right\|_{L_p(\mathcal{P}_W\mathcal{P}_{X^n})}^p \quad (5.44)$$

$$= \sum_{x^n} \int_0^1 \left(\frac{\mathcal{P}_{W|X^n=x^n}(w)}{\mathcal{P}_W(w)} \right)^p dw \quad (5.45)$$

$$= (n+1)^{p-1} \sum_{k=0}^n \int_0^1 \left(\binom{n}{k} w^k (1-w)^{(n-k)} \right)^p dw \quad (5.46)$$

$$= (n+1)^{p-1} \sum_{k=0}^n \binom{n}{k}^p \frac{\Gamma(kp+1)\Gamma((n-k)p+1)}{\Gamma(np+2)}, \quad (5.47)$$

where Equation (5.47) uses the identity relating the Beta function with the Gamma function *i.e.*,

$$\text{Beta}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \quad (5.48)$$

Then, with $p = 2$ one recovers:

$$\mathcal{H}_2(W, X^n) + 1 = \chi^2(W, X^n) = \frac{n+1}{2n+1} \cdot \frac{4^n}{\binom{2n}{n}} \leq \frac{16\sqrt{\pi n}}{21}. \quad (5.49)$$

Hence, specialising Equation (5.40) to $p = 2$ leads us to:

$$R \geq \sup_{\rho>0} \rho \left(1 - \sqrt{2\rho(\chi^2(W, X^n) + 1)} \right). \quad (5.50)$$

Solving then the maximization over ρ and using Equation (5.49) one can conclude that:

$$R \geq \frac{2}{27} \cdot \frac{1}{\chi^2(W, X^n) + 1} \geq \frac{7}{72\sqrt{\pi n}}. \quad (5.51)$$

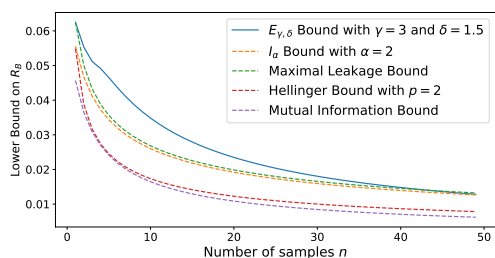
Notice that Equation (5.51) also matches the upper-bound up to a constant and, similarly to Maximal Leakage, tightens the result in (Xu and Raginsky, 2017a, Corollary 2) while not requiring that $n \rightarrow \infty$.

Remark 52. *Stirling's approximation yields $(\chi^2(W, X^n) + 1) \sim \frac{\sqrt{\pi n}}{2}$ when n is large. This implies that, for n large, one can show that $R \gtrsim \frac{4}{27\sqrt{\pi n}}$, thus leading to a slight improvement over Equation (5.51).*

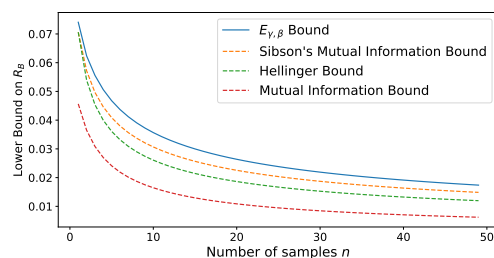
To conclude, one can apply the same steps with the $E_{\gamma, \delta}$ -Divergence. The lower-bound on the risk in this example can thus be expressed as

$$R \geq \sup_{\rho>0} \rho \left(1 - \frac{(E_{\gamma, \delta}(W, X^n) + 2\rho\gamma)}{\delta} \right) \quad (5.52)$$

$$= \frac{(\delta - E_{\gamma, \delta}(W, X^n))^2}{8\gamma\delta}. \quad (5.53)$$



(a) The picture shows the behaviour of Equation (5.36), Equation (5.39) with $\alpha = 2$, Equation (5.51), Equation (5.53) with $\gamma = 3$ and $\delta = 1.5$ and (Xu and Raginsky, 2017a, Equation (19)) as a function of n . The values of $E_{3,1.5}(W, X^n)$ for each n are computed numerically. Solid lines mean that the corresponding lower-bound is the largest.



(b) Comparison between the largest lower-bounds one can retrieve for different information measures in Example 1: that is between Equation (5.39), Equation (5.40), Equation (5.52) and (Xu and Raginsky, 2017a, Equation (19)). The quantities are analytically maximized over ρ and numerically optimized over, respectively, $\alpha > 1$, $p > 1$, $\delta > 0$, and $\gamma \geq \delta$. Solid lines mean that the corresponding lower-bound is the largest.

The lower-bound in Equation (5.53) can be empirically seen to be the best among the ones presented so far (thus beating Hellinger, I_α and, consequently, Maximal Leakage and Mutual Information). A direct comparison between the bounds provided here and those already present in the literature can be seen in Figure 5.4.1a and Figure 5.4.1b. The lower-bounds are computed as a function of the number of samples n , which we consider to be in the range $\{1, \dots, 50\}$. The figure shows that all the divergences we considered in this work provide a larger (and thus, better) lower-bound on the Bayesian risk when compared with results that stem from using Shannon's Mutual Information (cf. (Xu and Raginsky, 2017a, Corollary 2)). In particular, the lower-bound involving the $E_{\gamma,\delta}$ -Mutual Information represents the largest among the ones we consider. Given the lack of a closed-form expression for $E_{\gamma,\delta}$ in this example, the quantity in Equation (5.53) was computed numerically. Moreover, in order to verify whether the behaviour (and ordering) of the lower-bounds in Figure 5.4.1a was determined by the specific choices of the parameters p, γ, δ and α , in Figure 5.4.1b the lower-bounds on the risk have also been numerically optimised over the respective parameters $p, \gamma, \delta, \alpha$. As Figure 5.4.1b shows, the lower-bound provided by $E_{\gamma,\delta}$ remains the best. Notice that the lower-bound involving Mutual Information has no parameter to optimise over (other than ρ). Maximal Leakage does not provide the best bound, but it possesses the interesting characteristic of depending on \mathcal{P}_W only through the support, thus leading to potential applicability in a variety of settings in which \mathcal{P}_W is not accessible. Differently, Mutual Information, the Hellinger Divergence and the $E_{\gamma,\delta}$ -Divergence all require to know \mathcal{P}_W . The lower-bounds on the Risk in this Example can thus be summarised as follows:

Corollary 30. *Consider the setting of Example 1 one has that*

$$R \geq \max \left\{ \max_{\delta > 0, \gamma \geq \delta} \left\{ \frac{(\beta - E_{\gamma,\delta}(W, X^n))^2}{8\gamma\delta} \right\}, \max_{\alpha > 1} \left\{ \left(\frac{(2\alpha - 1)}{2\alpha} \exp \left(\frac{\alpha - 1}{\alpha} I_\alpha(W, X^n) \right) \right)^{\frac{-\alpha}{\alpha-1}} \frac{(\alpha - 1)}{(2\alpha - 1)} \right\} \right\}. \quad (5.54)$$

5.4.2 Gaussian prior with Gaussian noise (and absolute error)

Another classical and interesting setting is given by the following example:

Example 2. Assume that $W \sim N(0, \sigma_W^2)$ and that for $i \in [n]$, $X_i = W + Z_i$ where $Z_i \sim N(0, \sigma^2)$. Assume also that the loss is s.t. $\ell(w, \hat{w}) = |w - \hat{w}|$.

Using the sample mean estimator one has that:

$$R \leq \sqrt{\frac{\sigma_W^2}{1 + n\sigma_W^2/\sigma^2}}. \quad (5.55)$$

Moreover, given that $\ell(w, \hat{w}) = |w - \hat{w}|$ it is also possible to show that:

$$L_W(\rho) \leq \rho \sqrt{\frac{2}{\sigma_W^2 \pi}}. \quad (5.56)$$

In this setting, $\mathcal{L}(W \rightarrow X^n)$ is infinite. However, $I_\alpha(W, X^n)$ is finite for every $\alpha < +\infty$. One can thus provide a lower-bound on the Risk, resorting to I_α via Equation (5.5). Given that the empirical mean is a sufficient statistic for W in this case, one has that:

$$I_\alpha(W, X^n) = I_\alpha\left(W, \frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{2} \log\left(1 + \alpha n \frac{\sigma_W^2}{\sigma^2}\right). \quad (5.57)$$

These considerations imply that :

$$R \geq \sup_{\alpha > 1} \max_{\rho > 0} \rho \left(1 - \exp\left(-\frac{\alpha-1}{\alpha} I_\alpha(W, X^n)\right)\right) \left(\rho \sqrt{\frac{2}{\sigma_W^2 \pi}}\right)^{\frac{\alpha-1}{\alpha}} \quad (5.58)$$

$$= \sup_{\alpha > 1} \max_{\rho > 0} \rho \left(1 - \left(\rho \sqrt{\left(1 + \alpha n \frac{\sigma_W^2}{\sigma^2}\right) \frac{2}{\sigma_W^2 \pi}}\right)^{\frac{\alpha-1}{\alpha}}\right) \quad (5.59)$$

$$= \sup_{\alpha > 1} \frac{1}{(\beta+1)} \left(\frac{\beta}{\beta+1}\right)^\beta \left(\sqrt{\left(1 + \alpha n \frac{\sigma_W^2}{\sigma^2}\right) \frac{2}{\sigma_W^2 \pi}}\right)^{-\frac{1}{\beta}}, \quad (5.60)$$

remembering that $\beta = \frac{\alpha}{\alpha-1}$. Stepping away from Sibson's α -Mutual Information one can look at Hellinger p -Divergences and $E_{\gamma, \delta}$ once again. In particular, one has that for $p > 1$:

$$\mathcal{H}_p(W, X) = \left(\frac{\left(1 + \frac{\sigma_W^2}{\sigma^2}\right)^p}{1 + (2-p)p \frac{\sigma_W^2}{\sigma^2}}\right)^{\frac{1}{2}}. \quad (5.61)$$

Thus, the family of bounds provided by Corollary 28 can be expressed as follows

$$R \geq \sup_{p>1} \max_{\rho>0} \rho \left(1 - \left(\frac{2\rho}{\sqrt{2\pi\sigma_W^2}} \right)^{\frac{p-1}{p}} \mathcal{H}_p^{\frac{1}{p}}(W, X^n) \right) \quad (5.62)$$

$$= \sup_{p>1} \max_{\rho>0} \rho \left(1 - \left(\frac{2\rho}{\sqrt{2\pi\sigma_W^2}} \right)^{\frac{p-1}{p}} \left(\frac{\left(1 + \frac{\sigma_W^2}{\sigma^2}\right)^p}{1 + (2-p)p \frac{\sigma_W^2}{\sigma^2}} \right)^{\frac{1}{2p}} \right) \quad (5.63)$$

$$= \sup_{p>1} \frac{1}{q+1} \left(\frac{q}{(q+1)} \right)^q \left(\left(\frac{2}{\sqrt{2\pi\sigma_W^2}} \right)^{\frac{p-1}{p}} \left(\frac{\left(1 + \frac{\sigma_W^2}{\sigma^2}\right)^p}{1 + (2-p)p \frac{\sigma_W^2}{\sigma^2}} \right)^{\frac{1}{2p}} \right)^{-\frac{1}{q}}, \quad (5.64)$$

where q represents the Hölder's conjugate with respect to p , i.e., $q = \frac{p}{p-1}$.

In particular, setting $p = 3/2$ one obtains:

$$\mathcal{H}_{3/2}(W, X) = \sqrt{\frac{\left(1 + \frac{\sigma_W^2}{\sigma^2}\right)^{\frac{3}{2}}}{1 + \frac{3\sigma_W^2}{4\sigma^2}}}, \quad (5.65)$$

leading us to a lower bound on the Bayesian risk given by:

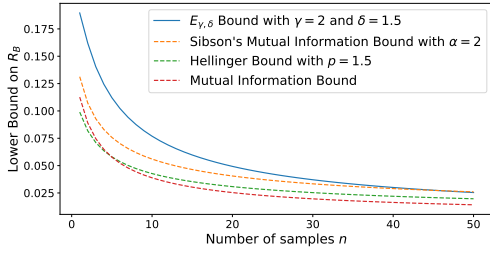
$$R \geq \frac{81\sqrt{2\pi}}{2048} \sqrt{\frac{\sigma_W^2}{1 + n \frac{\sigma_W^2}{\sigma^2}}}. \quad (5.66)$$

Similarly to the previous example, one has that Equation (5.66) matches the upper-bound up to a constant factor, and provides a strengthening of the bounds obtained in (Xu and Raginsky, 2017a, Corollary 1). Repeating the analysis with the $E_{\gamma,\delta}$ -Divergence, one obtains the following:

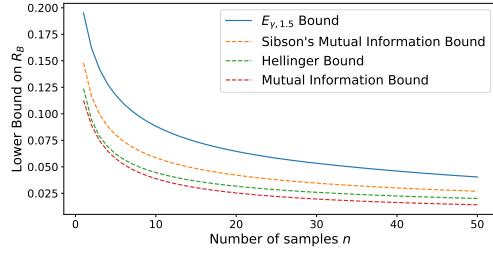
$$R \geq \sup_{\rho>0} \rho \left(1 - \frac{\left(E_{\gamma,\delta}(W, X^n) + \frac{2\rho\gamma}{\sqrt{2\sigma_W^2\pi}} \right)}{\delta} \right) \quad (5.67)$$

$$= \frac{\sqrt{2\sigma_W^2\pi} (\delta - E_{\gamma,\delta}(W, X^n))^2}{8\gamma\delta}. \quad (5.68)$$

Like in Example 1, one can numerically evaluate Equation (5.68) and compare it with Equation (5.60), Equation (5.66) and (Xu and Raginsky, 2017a, Equation (16)). Figure 5.4.1a and Figure 5.4.1b show the resulting lower-bounds as a function of the number of samples n . One can observe similar behaviors when comparing with the results from previous example: the



(a) Setting: Example 2 with $\sigma_W^2 = 1$ and $\sigma^2 = 2$. The picture shows the behaviour of Equation (5.60) with $\alpha = 2$, Equation (5.68) with $\gamma = 2$ and $\delta = 1.5$ and (Xu and Raginsky, 2017a, Equation (16)) as a function of n . The values of $E_{2,1.5}(W, X^n)$ for each n are computed numerically. Solid lines mean that the corresponding lower-bound is the largest.



(b) Comparison between the largest lower-bounds one can retrieve for different information measures in Example 2: that is between, Equation (5.60), Equation (5.64), Equation (5.68) with $\delta = 1.5$, and (Xu and Raginsky, 2017a, Equation (16)). The quantities are numerically optimised over, respectively, $\gamma \geq 1$, $p > 1$ and $\alpha > 1$. The numerical optimisation over the parameter δ is not carried out for computational reasons. Solid lines mean that the corresponding lower-bound is the largest.

bounds retrieved through the \mathcal{H}_p - and $E_{\gamma, \delta}$ -Divergences are both able to improve on the lower-bound relying on Shannon's Mutual Information. Once again, $E_{\gamma, \delta}$, (cf. Equation (5.68)) provides the largest lower-bound, while Sibson's α -Mutual Information is still able to provide a stronger result than Equation (5.64). Similarly to before, one can also numerically optimise the bounds with respect to the corresponding parameters $\alpha > 1$, $p > 1$, $\delta > 0$ and $\gamma \geq \delta$ and the resulting comparison is depicted in Figure 5.4.1b.

5.4.3 Hide-and-seek problem

To conclude, let us consider next a d -dimensional distributed estimation problem, known as the "Hide and seek" problem. It has been first presented in Shamir (2014) and also studied in Xu and Raginsky (2017a).

Example 3. Consider a family of distributions $\mathcal{P} = \{\mathcal{P}_w : w = 1, \dots, d\}$ on $\{0, 1\}^d$. Under \mathcal{P}_w , the w -th coordinate of the random vector $X \in \{0, 1\}^d$ has bias $\frac{1}{2} + \rho$ while the other coordinates of X are independently drawn from $\text{Ber}(1/2)$. For $i = 1, \dots, m$, the i -th processor observes n samples X_i^n drawn independently from \mathcal{P}_W , and sends a b -bits message $Y_i = \varphi(X_i^n, Y_i^{i-1})$ to the estimator. The estimator computes $\hat{W} = \psi(Y^m)$ from the received messages. The risk in this example is defined as:

$$R_M = \inf_{\varphi^m, \psi} \max_{w \in [d]} \mathbb{P}[W \neq \hat{W}]. \quad (5.69)$$

A lower-bound for R_M derived by Shamir is as follows:

$$R_M \geq 1 - \left(\frac{3}{d} + 5 \sqrt{\min \left\{ \frac{10\rho nmb}{d}, mn\rho^2 \right\}} \right) \quad (5.70)$$

and only holds for $0 \leq \rho \leq 1/(4n)$. Additionally, in (Xu and Raginsky, 2017a) a quite different

lower-bound has been proposed:

$$R_M \geq 1 - \frac{1}{\log d} \min \left\{ \left[1 - \left(\frac{1-2\rho}{1+2\rho} \right)^n \right] mb + 1, \min(4mn\rho^2, \log d) + 1 \right\}, \quad (5.71)$$

and it holds for $0 \leq \rho \leq 1/2$. Let us now use a naïve approach with Maximal Leakage. We have that $W - X^{n \times m} - Y^m - \hat{W}$ forms a Markov Chain. Thus,

$$\mathcal{L}(W \rightarrow \hat{W}) \leq \min(\mathcal{L}(W \rightarrow X^{n \times m}), \mathcal{L}(W \rightarrow Y^m)).$$

We also have that $\mathcal{L}(W \rightarrow Y^m) \leq mb$ and that:

$$\mathcal{L}(W \rightarrow X^{n \times m}) \leq nm \mathcal{L}(W \rightarrow X) \quad (5.72)$$

$$= nm \log \sum_x \max_w \mathcal{P}_{X|W=w}(x) \quad (5.73)$$

$$\leq nm \log \sum_x \frac{1}{2} \left(\frac{1}{2} + \rho \right) \quad (5.74)$$

$$= nm \log(2^d(2^{-d} + 2^{-d+1}\rho)) \quad (5.75)$$

$$= nm \log(1 + 2\rho), \quad (5.76)$$

Hence:

$$\mathcal{L}(W \rightarrow \hat{W}) \leq \min(nm \log(1 + 2\rho), \log d, mb). \quad (5.77)$$

Using Equation (5.77) in Corollary 27 we get the following:

$$\mathbb{P}(\{\hat{W} \neq W\}) \geq 1 - \frac{\exp(\min\{mb, \log d, nm \log(1 + 2\rho)\})}{d}. \quad (5.78)$$

Notice that Equation (5.78) is such that the right-hand side is always greater or equal than 0. Indeed, assuming d to be fixed and letting n and m grow, we have that the minimum is achieved by $\log d$ and in that case we have $\mathbb{P}(\{\hat{W} \neq W\}) \geq 0$. Here, the difference in behaviour of Corollary 27 with respect to (Xu and Raginsky, 2017a, Theorem 1) is pivotal. Let us now compare the results on a common setting. The setting chosen in (Xu and Raginsky, 2017a), where $d = 512, b = 3d, m = 10$ and $\rho = 1/(4n)$ does not represent a choice of parameters where Equation (5.78) is interesting. Indeed, for large enough n , $nm \log(1 + 2\rho) = nm \log(1 + 1/2n) \approx m/2$ and, as a consequence, the expression will converge to a constant determined by the minimum between $mb, \log d, m/2$. Furthermore, both Equation (5.70) and Equation (5.71) have a term that depends on $mn\rho^2$ which, for $\rho = 1/(4n)$, will decay with n . Thus, choosing $\rho \sim n^{-p}$ with $p > 1$ represents an interesting setting for the bound in Equation (5.78), as the plots in Figure 5.4.1a and Figure 5.4.1b show.

Thanks to the different behaviour of Equation (5.78) (reaching 1 exponentially fast) we can see a much sharper jump towards 1 with respect to Equation (5.71), which instead plateaus below 1, and with respect to Equation (5.70) that reaches 1 more slowly. The growth towards 1 of Equation (5.78) becomes even sharper with larger p and converges towards a specific

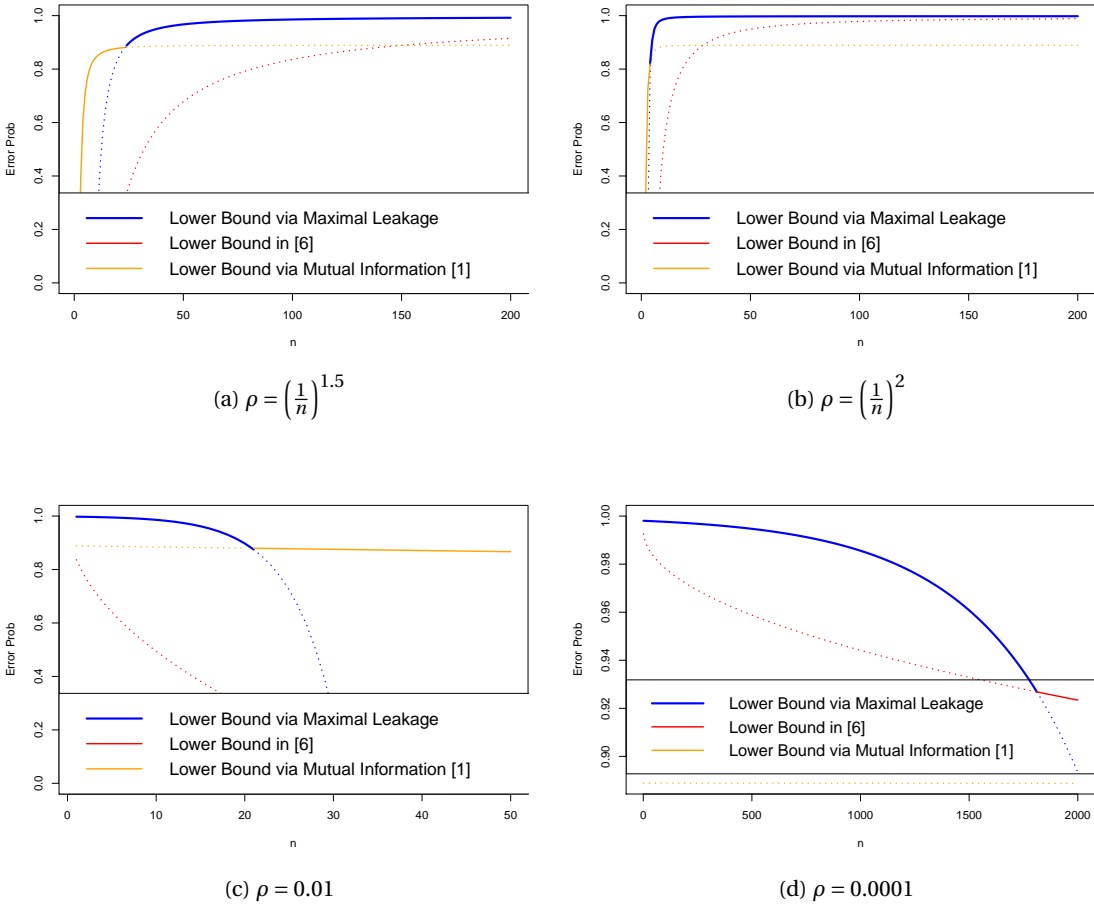


Figure 5.4.1: Behaviour of Equation (5.78) for various values of ρ . Solid lines mean that the corresponding lower-bound is the largest.

behaviour at $p \approx 2$. Increasing p any further does not alter the behaviour of the bound meaningfully. As for the behaviour of the bound for fixed ρ , if $\rho = 0.01$. then Equation (5.78) provides a larger lower-bound only for $n < 25$. If the parameter is brought down to $\rho = 0.0001$ then Equation (5.78) is larger than Equation (5.71) for all n but only larger than Equation (5.70) for $n < 1850$. Regardless of the considerations related to the specific settings, it is interesting how a very simple application of Corollary 27 can provide a tighter lower-bound. Moreover, in the proof of Equation (5.71) in (Xu and Raginsky, 2017a), in order to compute $I(W; X)$ an assumption on the distribution of W was necessary and the choice fell on W uniform on $[d]$. With Maximal Leakage, $\mathcal{L}(W \rightarrow X)$ does not depend on the specific distribution over W , rendering the bound potentially more general. Other divergences could be explored in this setting as well. However, one in general does not have a chain rule for any other φ -Divergence (or Sibson's α -Mutual Information with $\alpha < +\infty$) which is a fundamental step in the proof for Maximal Leakage (cf. Equation (5.72)). Moreover, some assumption (or maximisation over) \mathcal{P}_W would be necessary. In general, some additional machinery would be required in order to employ them in this setting. Given that this is outside of the scope of this thesis, these approaches will not be explored in this document.

5.5 Other approaches and generalisations

5.5.1 Inverting the roles

The Sibson's α -mutual information is an asymmetric quantity. A natural question is: can one provide a result similar to Theorem 25 involving $I_\alpha(X, W)$ instead? Indeed, by inverting the roles of W and \hat{W} , such a bound can be given but it will involve the small ball probability for \hat{W} i.e.,

$$L_{\hat{W}}(\rho) = \sup_w \mathcal{P}_{\hat{W}}(d(w, \hat{W}) \geq \rho). \quad (5.79)$$

This quantity hinges on the marginal distribution of \hat{W} , which, in turn, depends on the estimator used. In terms of $L_{\hat{W}}(\rho)$, one can give the following general bound:

Lemma 20. *Consider the Bayesian framework described in Section 5.2. The following holds for every $\alpha > 1$ and $\rho > 0$:*

$$R \geq \rho \left(1 - \exp \left(\frac{\alpha - 1}{\alpha} (I_\alpha(X, W) + \log(L_{\hat{W}}(\rho))) \right) \right). \quad (5.80)$$

Moreover, taking the limit of $\alpha \rightarrow \infty$ one has:

$$R \geq \rho \left(1 - \exp \left(\mathcal{L}(X \rightarrow W) + \log(L_{\hat{W}}(\rho)) \right) \right). \quad (5.81)$$

To apply this lemma in concrete cases, one needs to compute or upper bound the small ball probability $L_{\hat{W}}(\rho)$. Leveraging basic properties of the estimator, one can sometimes

bound it. For example, if the estimator is a linear function of the noisy observations one can leverage results related to Lévy’s concentration functions of sums of independent random variables. *E.g.*, if Y_1, \dots, Y_m are uncorrelated and have log-concave distributions, then for every $\rho \geq 0$ (Bobkov and Chistyakov, 2015, Theorem 1.1),

$$L_{\sum_{i=1}^m Y_i}(\rho) \leq \frac{2\rho}{\sqrt{\text{Var}(\sum_{i=1}^m Y_i) + \rho^2/3}} = \frac{2\rho}{\sqrt{m\text{Var}(Y_1) + \rho^2/3}}. \quad (5.82)$$

More general statements can be made, assuming $\phi(Y^m) = \sum_{i=1}^m a_i Y_i$ under different constraints over a_i (Nguyen and Vu, 2013). To appreciate the promise of this approach, let us also discuss the behaviors of $I_\alpha(W, X)$ and $I_\alpha(X, W)$. More specifically, let us consider again the “Hide and Seek” problem. Assuming, as in (Xu and Raginsky, 2017a, Example 12), that \mathcal{P}_W is uniform over $[d]$, one has that

$$\mathcal{L}(X^{n \times m} \rightarrow W) = \log \frac{d(1/2 + \rho)}{(d-1)(1/2 - \rho) + (1/2 + \rho)} = \log \kappa(d, \rho) < \log d. \quad (5.83)$$

In case ρ and d are constant and the estimator ϕ is a linear combination of the observations, using Equation (5.82) in Lemma 20 one gets:

$$R \geq \rho \left(1 - \frac{\kappa(d, \rho) 2\rho}{\sqrt{m\text{Var}(Y_i) + \rho^2/3}} \right). \quad (5.84)$$

This lower bound approaches ρ as m grows, rather than providing the trivial lower bound of 0, as it happens in Equation (5.78).

The assumptions required, along with the need of specifying a prior over W , clearly restrict the domain of applicability of Lemma 20 with respect to Theorem 25 and Corollary 27. However, this approach can provide results in settings where Theorem 25 and Corollary 27 become vacuous.

5.5.2 Conditioning

Following the approach undertaken in (Xu and Raginsky, 2017a), it is also possible to propose a conditional version of the theorems proposed above, in order to retrieve tighter bounds. For this to happen one needs a definition of conditional information measures. For φ -Divergences the choice would typically fall on objects of the following form

$$I_\varphi(X, Y|Z) = D_\varphi(\mathcal{P}_{XYZ} \| \mathcal{P}_Z \mathcal{P}_{X|Z} \mathcal{P}_{Y|Z}). \quad (5.85)$$

As for Sibson’s I_α , the matter becomes slightly more complicated. $I_\alpha(X, Y)$ itself is not simply defined as $D_\alpha(\mathcal{P}_{XY} \| \mathcal{P}_X \mathcal{P}_Y)$ but it requires an additional projection step. It has been shown in (Esposito et al., 2021b) that several definitions of conditional I_α can be proposed, depending on the operational meaning and corresponding probability bound one needs.

With the purpose of this chapter in mind, we will consider the conditional version of I_α defined

in Equation (3.75) and hereby re-stated for reference:

$$I_\alpha^{Y|Z}(X, Y|Z) = \min_{Q_{Y|Z}} D_\alpha(\mathcal{P}_{XYZ} \| \mathcal{P}_{X|Z} Q_{Y|Z} \mathcal{P}_Z). \quad (5.86)$$

The choice of this specific definition is necessary in order to be able to use Corollary 11, which is in turn necessary to provide a conditional version of Theorem 25 and Corollary 27 similar to (Xu and Raginsky, 2017a, Theorem 1 eq. (5)).

Leveraging said definition and the fact that:

$$I_\alpha^{Y|Z}(X, Y|Z) \xrightarrow{\alpha \rightarrow \infty} \mathcal{L}(X \rightarrow Y|Z)$$

one can thus give a conditional version of Theorem 25 and Corollary 27, introducing the notion of conditional small-ball probability. In particular, one can define the following quantity:

$$L_{W|U}(U, \rho) = \sup_{\hat{w} \in \hat{\mathcal{W}}} \mathcal{P}_{W|U}(\ell(W, \hat{w}) \leq \rho), \quad (5.87)$$

and show the following result:

Theorem 27. *Consider the Bayesian framework described in Section 5.2,*

$$R \geq \sup_{\mathcal{P}_{U|W, X}} \sup_{\rho > 0, \alpha \geq 1} \rho \left(1 - \exp \left(\frac{\alpha - 1}{\alpha} (I_\alpha(W, X|U) + \log(\mathcal{P}_U(L_{W|U}(U, \rho))) \right) \right), \quad (5.88)$$

Moreover, taking the limit of $\alpha \rightarrow \infty$ one has:

$$R \geq \sup_{\mathcal{P}_{U|W, X}} \sup_{\rho > 0} \rho \left(1 - \exp \left(\mathcal{L}(W \rightarrow X|U) + \log(\mathcal{P}_U(L_{W|U}(U, \rho))) \right) \right). \quad (5.89)$$

The main idea lying behind the conditional version for Mutual Information is that, choosing an appropriate U , it is possible to control the growth of $I(W; X|U)$ and obtain tighter bounds. In particular, let us assume we have n samples X^n . If the family $\mathcal{P} = \{\mathcal{P}_{X|W=w} : w \in \mathcal{W}\}$ is a subset of a finite-dimensional exponential family and W has a density supported on a compact subset of \mathbb{R}^d , choosing U to be a conditionally independent copy \hat{X}^n of X^n (given W) the mutual information $I(W; X^n | \hat{X}^n)$ will converge to a constant as n grows (rather than grow with n). This property seems to be specific to Shannon's Mutual Information. In the examples considered above, there does not appear to be a suitable U that tightens the bounds further for the divergences considered. Nonetheless we state the result as it may be of interest in other settings.

5.5.3 Leveraging SDPIs

An important step utilised both here and in (Xu and Raginsky, 2017a) to provide a lower-bound on the Risk is the application of the Data-Processing Inequality. The Data-Processing

Inequality allows us to go from² $I_\varphi(W, \hat{W})$ to $I_\varphi(W, X^n)$. In particular, one typically has that $I_\varphi(W, \hat{W}) \leq I_\varphi(W, X^n)$, because of the Markov Chain $W - X^n - \hat{W}$.

This inequality can be tightened considering the so-called “Strong Data-Processing Inequalities”. *I.e.*, given an information measure I_φ that satisfies the DPI and a Markov Chain $W - X - \hat{W}$ one can define and compute the following quantity if $\mathcal{P}_{\hat{W}|X}$ is fixed

$$\sup_{\mathcal{P}_{W|X}} \frac{I_\varphi(W, \hat{W})}{I_\varphi(W, X)} = \eta_\varphi(\mathcal{P}_X, \mathcal{P}_{\hat{W}|X}), \quad (5.90)$$

as well as

$$\sup_{\mathcal{P}_{WX}} \frac{I_\varphi(W, \hat{W})}{I_\varphi(W, X)} = \eta_\varphi(\mathcal{P}_{\hat{W}|X}), \quad (5.91)$$

if $\mathcal{P}_{\hat{W}|X}$ is fixed. Whenever the quantities in Equation (5.90) and Equation (5.91) are strictly smaller than 1, one obtains a tightening in the usual Data-Processing Inequalities:

$$I_\varphi(W, \hat{W}) \leq I_\varphi(W, X) \eta_\varphi(\mathcal{P}_X, \mathcal{P}_{\hat{W}|X}) \leq I_\varphi(W, X) \eta_\varphi(\mathcal{P}_{\hat{W}|X}). \quad (5.92)$$

In general, it is not easy to compute such quantities but a number of upper-bound can be found in the literature (Xu and Raginsky, 2017a, Section III). In particular, given any convex φ it is possible to show the following result:

Lemma 21 ((Cohen et al., 1993, Theorem 4.1)). *Let $\mathcal{P}_{X|W}$ represent a Markov Kernel, and let φ be a convex functional, one has that*

$$\eta_\varphi(\mathcal{P}_{X|W}) \leq \vartheta(\mathcal{P}_{X|W}), \quad (5.93)$$

where

$$\vartheta(\mathcal{P}_{X|W}) = \max_{w, w'} \|\mathcal{P}_{X|W=w} - \mathcal{P}_{X|W=w'}\|_{TV} \quad (5.94)$$

represents the Dobrushin contraction coefficient of $\mathcal{P}_{X|W}$.

These results apply to all φ -Divergences but not necessarily to Sibson’s α -Mutual Information. Indeed, in general, one cannot directly upper-bound $\eta_{I_\alpha}(\mathcal{P}_{\hat{W}|X})$ using the Dobrushin’s coefficient. One can, however, provide a different type of SDPI for this family of information measures that allows us to leverage these results once again. In particular, given the Markov

²In this section we will slightly abuse the notation and, for simplicity, consider Sibson’s I_α as an instance of I_φ since it satisfies a DPI even though it is not, technically, a φ -Divergence.

Chain $W - X - \hat{W}$ one can do the following:

$$I_\alpha(W, \hat{W}) = \frac{\alpha}{\alpha-1} \log \mathcal{P}_{\hat{W}} \left(\frac{\alpha-1}{\alpha} \exp(D_\alpha(\mathcal{P}_{W|\hat{W}} \parallel \mathcal{P}_W)) \right) \quad (5.95)$$

$$= \frac{\alpha}{\alpha-1} \log \mathcal{P}_{\hat{W}} \left(H_\alpha(\mathcal{P}_{W|\hat{W}} \parallel \mathcal{P}_W) \right) \quad (5.96)$$

$$\leq \frac{\alpha}{\alpha-1} \log \mathcal{P}_{\hat{W}} \left(\eta_{\varphi_\alpha}^{1/\alpha}(\mathcal{P}_{W|X}) H_\alpha(\mathcal{P}_{X|\hat{W}} \parallel \mathcal{P}_X) \right) \quad (5.97)$$

$$= \frac{1}{\alpha-1} \log(\eta_{\varphi_\alpha}(\mathcal{P}_{W|X}) + I_\alpha(X, \hat{W})). \quad (5.98)$$

Bringing in the SDPI constant of a classical φ -Divergence (*i.e.*, the Hellinger Integral H_α that stems from $\varphi_\alpha(x) = x^\alpha$) one can then tighten the DPI for I_α using the tools described in (Xu and Raginsky, 2017a, Section III). However, given the asymmetry of Sibson's Mutual Information, the sequence of steps just above only works if the second random variable in Equation (5.95) and Equation (5.98) remains the same *i.e.*, one considers $I_\alpha(\cdot, \hat{W})$. This is due to the fact that the representation depicted in Equation (5.96) requires us to rewrite both $I_\alpha(W, \hat{W})$ and $I_\alpha(X, \hat{W})$ as an expectation with respect to $\mathcal{P}_{\hat{W}}$ of an Hellinger integral. This is only possible if the second random variable is the same in both I_α 's. Using the same approach, for instance, one cannot tighten the inequality $I_\alpha(W, \hat{W}) \leq I_\alpha(W, X)$ (as the second random variable changes from \hat{W} to X).

One could, however, follow the strategy suggested in Section 5.5.1 and employ $I_\alpha(\hat{W}, W)$ to provide a lower-bound on the Risk. Such a choice would then allow to provide an upper-bound that involves $I_\alpha(X, W)$ and the Strong Data-Processing inequality $\eta_{\varphi_\alpha}(\mathcal{P}_{X|W})$.

5.5.4 Lower-Bounding the Risk Directly

An alternative route can be undertaken that does not use Markov's inequality as a first step and can possibly lead to tighter bounds. These results follow from the functional inequalities expressed in Section 3.1.5, in particular Theorem 16 and Corollary 8. Such a result would involve, for instance, Sibson's α Mutual Information with $\alpha < 1$:

Corollary 31. *Consider the Bayesian framework described in Section 5.2. The following must hold for every $\alpha < 1$ and $\rho > 0$:*

$$R \geq \text{ess inf}_{\mathcal{P}_{\hat{W}}} \left(\mathcal{P}_{\hat{W}}^{\frac{1}{\beta}} \left(\ell(W, \hat{W})^\beta \right) \right) \cdot \exp \left(\text{sign}(\alpha) \cdot \frac{\alpha-1}{\alpha} I_\alpha(W, X) \right). \quad (5.99)$$

Proof. Using Corollary 8 with $g = \ell$ one recovers the following:

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \text{ess inf}_{\mathcal{P}_{\hat{W}}} \left(\mathcal{P}_{\hat{W}}^{\frac{1}{\beta}} \left(\ell(W, \hat{W})^\beta \right) \right) \cdot \exp \left(\text{sign}(\alpha) \cdot \frac{\alpha-1}{\alpha} I_\alpha(W, \hat{W}) \right) \quad (5.100)$$

Now, if $0 < \alpha < 1$ then $\frac{1}{\beta} < 0$. By the Data-Processing Inequality for I_α with $0 < \alpha < 1$ (along

with the negativity of $\frac{1}{\beta}$) one has that

$$\exp\left(\text{sign}(\alpha) \cdot \frac{\alpha-1}{\alpha} I_\alpha(W, \hat{W})\right) = \exp\left(\frac{1}{\beta} I_\alpha(W, \hat{W})\right) \geq \exp\left(\frac{1}{\beta} I_\alpha(W, X)\right). \quad (5.101)$$

Otherwise, if $\alpha < 0$ then $0 < \frac{1}{\beta} < 1$. By the by the Data-Processing Inequality for I_α with $\alpha < 0$ (cf. (Esposito et al., 2022, Theorem 3.5)) one has that

$$\exp\left(\text{sign}(\alpha) \cdot \frac{\alpha-1}{\alpha} I_\alpha(W, \hat{W})\right) = \exp\left(-\frac{1}{\beta} I_\alpha(W, \hat{W})\right) \geq \exp\left(-\frac{1}{\beta} I_\alpha(W, X)\right). \quad (5.102)$$

The lower-bound on the Risk follows by noticing that the right-hand side Equation (5.100) can be rendered independent of \hat{W} for every $\alpha < 1$ (*i.e.*, it will only depend on the support of \hat{W} through the ess inf) via Equation (5.101) and Equation (5.102). \square

Corollary 31 is different from the results presented in the previous section. While in Section 5.3 the only dependence on ℓ was through the small-ball probability, in Corollary 31 one is required to have access to the expected value of the β -th moments of ℓ with respect to \mathcal{P}_X . Such an object may not be as easy to bound as the small-ball probability.

Remark 53. *If $W = \hat{W}$ then $\ell(W, \hat{W}) = 0$ and $I_\alpha(W, \hat{W}) = 0$. If $0 < \alpha < 1$, given that $\beta < 0$, one recovers the following lower-bound on the risk:*

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq 0,$$

which matches with our intuition.

A limiting behaviour of Corollary 31 is of interest as well. For instance, taking the limit of $\alpha \rightarrow -\infty$ one has that (whenever the so-called regular conditional probabilities $\mathcal{P}_{\hat{W}|W}$ exist, *e.g.*, in Radon Spaces):

$$\exp\left(-\frac{\alpha-1}{\alpha} I_\alpha(W, X^n)\right) \rightarrow \exp(-\mathcal{L}^c(W \rightarrow X^n)) \quad (5.103)$$

$$= \int_{\mathcal{Y}} \text{ess inf}_{\mathcal{P}_W} \left(\frac{d\mathcal{P}_{W\hat{W}}}{d\mathcal{P}_W \mathcal{P}_{\hat{W}}} \right) d\mathcal{P}_{\hat{W}} \quad (5.104)$$

$$= \int_{\mathcal{Y}} \text{ess inf}_{\mathcal{P}_W} \mathcal{P}_{\hat{W}|W}. \quad (5.105)$$

Moreover, one has that, whenever $\alpha \rightarrow -\infty$, then $\beta \rightarrow 1$, and the resulting lower-bound would be the following:

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \left(\text{ess inf}_{\mathcal{P}_{\hat{W}}} \mathcal{P}_W(\ell(W, \hat{w})) \right) \left(\int_{\mathcal{Y}} \text{ess inf}_{\mathcal{P}_W} \mathcal{P}_{\hat{W}|W} \right). \quad (5.106)$$

In a discrete setting this amounts to:

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \left(\min_{\hat{w}} \mathcal{P}_W(\ell(W, \hat{w})) \right) \left(\sum_{\hat{w}} \min_w p_{\hat{W}|W=w}(\hat{w}) \right), \quad (5.107)$$

where $p_{\hat{W}|W=w}$ denotes the probability mass function of \hat{W} given that $W = w$. This result can also be easily proved directly, indeed:

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) = \sum_{w, \hat{w}} \ell(w, \hat{w}) p_{W\hat{W}}(w, \hat{w}) \quad (5.108)$$

$$= \sum_{w, \hat{w}} \ell(w, \hat{w}) p_W(w) p_{\hat{W}|W=w}(\hat{w}) \quad (5.109)$$

$$\geq \sum_{\hat{w}} \min_w p_{\hat{W}|W=w}(\hat{w}) \sum_w \ell(w, \hat{w}) p_W(w) \quad (5.110)$$

$$\geq \left(\min_{\hat{w}} \sum_w \ell(w, \hat{w}) p_W(w) \right) \sum_{\hat{w}} \min_w p_{\hat{W}|W=w}(\hat{w}) \quad (5.111)$$

$$= \left(\min_{\hat{w}} \mathcal{P}_W(\ell(W, \hat{w})) \right) \left(\exp(-\mathcal{L}^c(W \rightarrow \hat{W})) \right). \quad (5.112)$$

The main difference with the results presented in Section 5.3 is that there is no small-ball probability involved and it is thus required to have access to an object of the form $\min_{\hat{w}} \|\ell(W, \hat{w})\|_{L^\beta(\mathcal{P}_W)}$, which might be harder to compute than $L_W(\rho)$. It is, however, possible to relate the quantity to a small-ball probability, restricting the value of α to $[-\infty, 0)$ and using a combination of Markov's inequality and Corollary 9. This approach leads us to the following lower-bound on the risk

Corollary 32. *Consider the Bayesian framework described in Section 5.2. The following must hold for every $\alpha < 0$ and $\rho > 0$:*

$$R \geq \rho \left(\min_{\hat{w}} \mathcal{P}_W(\ell(W, \hat{w}) \geq \rho)^{\frac{1}{\beta}} \cdot \exp\left(-\frac{\alpha-1}{\alpha} I_\alpha(W, X)\right) \right) \quad (5.113)$$

And, in particular,

$$\min_{\hat{w}} \mathcal{P}_W(\ell(W, \hat{w}) \geq \rho) = \min_{\hat{w}} (1 - \mathcal{P}_W(\ell(W, \hat{w}) \leq \rho)) \quad (5.114)$$

$$= 1 - \max_{\hat{w}} \mathcal{P}_W(\ell(W, \hat{w}) \leq \rho) \quad (5.115)$$

$$= 1 - L_W(\rho). \quad (5.116)$$

Hence, if we can upper-bound $L_W(\rho)$ like we assumed in the previous section, then we can lower-bound $\min_{\hat{w}} \mathcal{P}_W(\ell(W, \hat{w}) \geq \rho)$ and, consequently, we can also lower-bound the risk using the approach depicted in this section (cf. Corollary 32).

5.5.5 Concave conjugates

In Chapter 2 and 3 we have seen that convex conjugates are a great tool to provide upper-bounds on the expected values of functions. In this chapter, however, the purpose is to provide a lower-bound on the Risk. With this in mind, we will explore the application of concave conjugates. Let us start with the well-known Kullback-Leibler Divergence. It is possible to show that given a measure μ and $\nu \in \mathcal{M}_1(\mathcal{X}, \mu)$:

$$-D(\nu\|\mu) = \inf_{g \in B(\mathcal{X})} \{\nu(g) + \log \mu(\exp(-g))\} \quad (5.117)$$

i.e., $-\log \mu(\exp(-g))$ is the concave conjugate of $-D(\nu\|\mu) = -\psi_\mu(\nu)$. We can thus set $g = \lambda \ell$ with $\lambda > 0$. We obtain a parametrised bound of the following shape:

$$\nu(\ell) \geq \frac{-D(\nu\|\mu) - \log \mu(\exp(-\lambda \ell))}{\lambda} \quad (5.118)$$

and in order to get the tightest bound one would have to compute the following

$$\nu(\ell) \geq \sup_{\lambda > 0} \frac{-D(\nu\|\mu) - \log \mu(\exp(-\lambda \ell))}{\lambda}. \quad (5.119)$$

Given the ℓ is normally assumed to be non-negative in this setting, in order to have a non-trivial lower-bound one would need $-D(\nu\|\mu) - \log(\mu(\exp(-\lambda \ell))) > 0$ for some value of λ . Hence one would need

$$D(\nu\|\mu) < -\log(\mu(\exp(-\lambda \ell))). \quad (5.120)$$

Since $D(\nu\|\mu) > 0$, the right-hand side of Equation (5.120) should also be positive, meaning that one would need $0 < \mu(\exp(-\lambda \ell)) < 1$ for some value of λ . Because of this, undertaking an approach similar to upper-bounding the expected-value of a function (cf. Section 3.2) and thus, assuming for instance sub-gaussianity of the function (in order to control its log-moment generating function) would not be enough for the right-hand side of Equation (5.119) to be positive. Indeed, one would have that if ℓ is σ^2 -sub-Gaussian under μ then for every $\lambda \in \mathbb{R}$:

$$-\log(\mu(\exp(-\lambda \ell))) \geq -\frac{\lambda^2 \sigma^2}{2}.$$

which is always guaranteed to be negative.

One can, however, propose a general approach for general φ -Divergences. Indeed, since $-D_\varphi(\nu\|\mu)$ is a concave functional of ν , for a given μ , denoting with $\psi_\mu(\nu) = D_\varphi(\nu\|\mu)$ one has that

$$(-\psi_\varphi)^*(g) = -\psi_\varphi^*(-g) = -\mu(\varphi^*(-g)) = -\int \varphi^*(-g) d\mu. \quad (5.121)$$

Consequently, one has the following general lower-bound:

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \sup_{\lambda > 0} \frac{-I_\varphi(W, \hat{W}) - \mathcal{P}_W \mathcal{P}_{\hat{W}}(\varphi^*(-\lambda \ell))}{\lambda}. \quad (5.122)$$

Assume that one has one has $\mathcal{P}_W \mathcal{P}_{\hat{W}}(\varphi^*(-\lambda\ell)) \leq \phi(\lambda)$ for a strictly convex ϕ . Assume also that ϕ^* admits a generalised inverse $\phi^{*-1}(y) = \inf\{t : \phi^*(t) > y\}$. Choosing $\lambda = \phi'^{-1}(\phi^{*-1}(I_\varphi(W, \hat{W})))$ one has that

$$\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})) \geq \frac{-I_\varphi(W, \hat{W}) - \mathcal{P}_W \mathcal{P}_{\hat{W}}(\varphi^*(-\lambda\ell))}{\lambda} \quad (5.123)$$

$$\geq \frac{-I_\varphi(W, \hat{W}) - \phi(\lambda)}{\lambda} \quad (5.124)$$

$$= \frac{-I_\varphi(W, \hat{W}) - \phi(\phi'^{-1}(\phi^{*-1}(I_\varphi(W, \hat{W}))))}{\phi'^{-1}(\phi^{*-1}(I_\varphi(W, \hat{W})))} \quad (5.125)$$

$$= \frac{-I_\varphi(W, \hat{W}) + \phi^*(\phi^{*-1}(I_\varphi(W, \hat{W}))) - \phi^{*-1}(I_\varphi(W, \hat{W}))\phi'^{-1}(\phi^{*-1}(I_\varphi(W, \hat{W})))}{\phi'^{-1}(\phi^{*-1}(I_\varphi(W, \hat{W})))} \quad (5.126)$$

$$= \frac{-I_\varphi(W, \hat{W}) + \phi^*(\phi^{*-1}(I_\varphi(W, \hat{W})))}{\phi'^{-1}(\phi^{*-1}(I_\varphi(W, \hat{W})))} - \phi^{*-1}(I_\varphi(W, \hat{W})) \quad (5.127)$$

$$> -\phi^{*-1}(I_\varphi(W, \hat{W})). \quad (5.128)$$

This also means that $I_\varphi(W, \hat{W}) > \phi^*(-\mathcal{P}_{W\hat{W}}(\ell(W, \hat{W})))$.

With a similar reasoning but selecting $g = \lambda \mathbb{1}_{\{\ell(W, \hat{W}) \geq \rho\}}$ (rather than $\lambda\ell$ and thus, not upper-bounding $\mathcal{P}_W \mathcal{P}_{\hat{W}}(\varphi^*(g))$ but computing it exactly) and then choosing

$$c = \frac{I_\varphi(W, \hat{W}) + \mathcal{P}_W \mathcal{P}_{\hat{W}}(\ell(W, \hat{W}) \leq \rho) \varphi^*(0)}{\mathcal{P}_W \mathcal{P}_{\hat{W}}(\ell(W, \hat{W}) \geq \rho)}$$

and, consequently,

$$\lambda = \varphi^{*'-1}(\varphi^{-1}(c))$$

would lead to Equation (5.12) along with Markov's inequality.

Bibliography

- S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966. ISSN 00359246. URL <http://www.jstor.org/stable/2984279>.
- N. Alon and A. Orlitsky. A lower bound on the expected length of one-to-one codes. *IEEE Transactions on Information Theory*, 40(5):1670–1672, 1994. doi: 10.1109/18.333891.
- Luigi Ambrosio. *Lecture Notes on Optimal Transport Problems*, pages 1–52. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-540-39189-0. doi: 10.1007/978-3-540-39189-0_1. URL https://doi.org/10.1007/978-3-540-39189-0_1.
- Venkat Anantharam. A variational characterization of rényi divergences. *CoRR*, abs/1701.07796, 2017. URL <http://arxiv.org/abs/1701.07796>.
- E. Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Transactions on Information Theory*, 42(1):99–105, 1996. doi: 10.1109/18.481781.
- E. A. Arutjunjan. Bounds for the Exponent of the Probability of Error for a Semicontinuous Memoryless Channel. *Probl. Peredachi Inf.*, 4(4):37–48, 1968. ISSN 0555-2923.
- Amir R. Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 7245–7254, Red Hook, NY, USA, 2018. Curran Associates Inc.
- Shahab Asoodeh, Maryam Aliakbarpour, and Flavio P. Calmon. Local differential privacy is equivalent to contraction of an f -divergence. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 545–550, 2021. doi: 10.1109/ISIT45174.2021.9517999.
- Rami Atar, Kamaljit Chowdhary, and Paul Dupuis. Robust bounds on risk-sensitive functionals via renyi divergence, 2013.
- Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357 – 367, 1967. doi: 10.2748/tmj/1178243286. URL <https://doi.org/10.2748/tmj/1178243286>.

Bibliography

- R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff. Learners that use little information. volume 83 of *Proceedings of Machine Learning Research*, pages 25–55. PMLR, 07–09 Apr 2018.
- M. Ben-Bassat and J. Raviv. Renyi’s entropy and the probability of error. *IEEE Transactions on Information Theory*, 24(3):324–331, 1978. doi: 10.1109/TIT.1978.1055890.
- Sterling K. Berberian. *Lectures in functional analysis and operator theory*. Springer Verlag New York, 1974. ISBN 0387900802.
- Jeremiah Birrell and Luc Rey-Bellet. Uncertainty quantification for markov processes via variational principles and functional inequalities. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):539–572, jan 2020. doi: 10.1137/19m1237429. URL <https://doi.org/10.1137/2F19m1237429>.
- Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational representations and neural network estimation of rényi divergences, 2021.
- Sergey G. Bobkov and Gennadiy P. Chistyakov. On concentration functions of random variables. *Journal of Theoretical Probability volume*, 28, 2015.
- S.G Bobkov and F Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1 – 28, 1999. ISSN 0022-1236.
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2: 499–526, 3 2002. ISSN 1532-4435. doi: 10.1162/153244302760200704.
- R. I. Boş, S. M. Grad, and G. Wanka. Fenchel-Lagrange Duality Versus Geometric Duality in Convex Optimization. *Journal of Optimization Theory and Applications*, 129(1):33–54, April 2006. doi: 10.1007/s10957-006-9047-2. URL https://ideas.repec.org/a/spr/joptap/v129y2006i1d10.1007_s10957-006-9047-2.html.
- Michel Broniatowski and Amor Keziou. Minimization of divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43, 03 2010. doi: 10.1556/SScMath.43.2006.4.2.
- Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591, 2019. doi: 10.1109/ISIT.2019.8849590.
- Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130, 2020. doi: 10.1109/JSAIT.2020.2991139.

- L.L. Campbell. A coding theorem and rényi's entropy. *Information and Control*, 8(4):423–429, 1965. ISSN 0019-9958. doi: [https://doi.org/10.1016/S0019-9958\(65\)90332-3](https://doi.org/10.1016/S0019-9958(65)90332-3). URL <https://www.sciencedirect.com/science/article/pii/S0019995865903323>.
- Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On bayes risk lower bounds. *J. Mach. Learn. Res.*, 17(1):7687–7744, jan 2016. ISSN 1532-4435.
- H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Math. Stat.*, 23:493–509, 1952.
- Joel E. Cohen, Yoh Iwasa, Gh. Rautu, Mary Beth Ruskai, Eugene Seneta, and Gh. Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear Algebra and its Applications*, 179:211–235, 1993. ISSN 0024-3795. doi: [https://doi.org/10.1016/0024-3795\(93\)90331-H](https://doi.org/10.1016/0024-3795(93)90331-H). URL <https://www.sciencedirect.com/science/article/pii/002437959390331H>.
- Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.
- I. Csiszar. Generalized cutoff rates and renyi's information measures. *IEEE Transactions on Information Theory*, 41(1):26–34, 1995. doi: 10.1109/18.370121.
- I. Csiszar. Generalized cutoff rates and Rényi's information measures. *IEEE Transactions on Information Theory*, 41(1):26–34, Jan 1995. ISSN 0018-9448. doi: 10.1109/18.370121.
- Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutató Int. Közl.*, 8:85–108, 1963.
- Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967. URL <https://ci.nii.ac.jp/naid/10028997448/en/>.
- Imre Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2:191–213, 1972a.
- Imre Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2:191–213, 1972b.
- A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2009. ISBN 9783642033100. URL <https://books.google.ch/books?id=d3nnjwEACAAJ>.
- John C. Duchi and Martin J. Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *CoRR*, abs/1311.2669, 2013. URL <http://arxiv.org/abs/1311.2669>.
- N. Dunford and J.T. Schwartz. *Linear Operators, Part 1: General Theory*. Wiley Classics Library. Wiley, 1988. ISBN 9780471608486.

Bibliography

- J. Dunham. Optimal noiseless coding of random variables (corresp.). *IEEE Transactions on Information Theory*, 26(3):345–345, 1980. doi: 10.1109/TIT.1980.1056200.
- C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, pages 117–126, New York, NY, USA, 2015a. ACM. ISBN 978-1-4503-3536-2. doi: 10.1145/2746539.2746580.
- Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, Cambridge, MA, USA, 2015b. MIT Pressf.
- Amedeo Roberto Esposito and Michael Gastpar. Lower-bounds on the bayesian risk in estimation procedures via Sibson’s α -mutual information. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 748–753, 2021. doi: 10.1109/ISIT45174.2021.9517954.
- Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8):4986–5004, 2021a. doi: 10.1109/TIT.2021.3085190.
- Amedeo Roberto Esposito, Diyuan Wu, and Michael Gastpar. On conditional sibson’s α -mutual information. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1796–1801, 2021b. doi: 10.1109/ISIT45174.2021.9517944.
- Amedeo Roberto Esposito, Adrien Vandenbroucque, and Michael Gastpar. On sibson’s α -mutual information, 2022.
- W. Fenchel. On conjugate convex functions. *Canadian Journal of Mathematics*, 1(1):73–77, 1949. doi: 10.4153/CJM-1949-007-x.
- G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013. ISBN 9781118626399. URL <https://books.google.ch/books?id=wI4fAwAAQBAJ>.
- J.P. Gossez, E.J. LamiDozo, J. Mahwin, and L. Waelbroeck. *Nonlinear Operators and the Calculus of Variations*. Springer Berlin Heidelberg, 1976.
- Konstantinos Gourgoulis, Markos A. Katsoulakis, Luc Rey-Bellet, and Jie Wang. How biased is your model? concentration inequalities, information and model bias. *IEEE Transactions on Information Theory*, 66(5):3079–3097, 2020. doi: 10.1109/TIT.2020.2977067.
- Timothy Gowers, June Barrow-Green, and Imre Leader, editors. *The Princeton Companion to Mathematics*. Princeton University Press, 2010. ISBN 9781400830398. doi: doi:10.1515/9781400830398. URL <https://doi.org/10.1515/9781400830398>.
- Robert Graczyk and Igal Sason. On two-stage guessing. *Information*, 12(4), 2021. ISSN 2078-2489. doi: 10.3390/info12040159. URL <https://www.mdpi.com/2078-2489/12/4/159>.

- Peter Harremoës and Igor Vajda. Joint range of f-divergences. In *2010 IEEE International Symposium on Information Theory*, pages 1345–1349, 2010. doi: 10.1109/ISIT.2010.5513445.
- Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459. URL <http://www.jstor.org/stable/2282952>.
- O. Hölder. Ueber einen Mittelwertsatz. *Gött. Nachr.*, 1889:38–47, 1889.
- Henryk Hudzik and Lech Maligranda. Amemiya norm equals orlicz norm in general. *Indagationes Mathematicae*, 11(4):573 – 585, 2000. ISSN 0019-3577. doi: [https://doi.org/10.1016/S0019-3577\(00\)80026-9](https://doi.org/10.1016/S0019-3577(00)80026-9). URL <http://www.sciencedirect.com/science/article/pii/S0019357700800269>.
- I. Issa, A. B. Wagner, and S. Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2020. doi: 10.1109/TIT.2019.2962804.
- Ibrahim Issa and Michael Gastpar. Computable bounds on the exploration bias. In *2018 IEEE International Symposium on Information Theory, ISIT Vail, CO, USA, June 17-22, 2018*, pages 576–580, 2018. doi: 10.1109/ISIT.2018.8437470.
- Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1475–1479. IEEE, 2017.
- E. Kreyszig. *Introductory Functional Analysis with Applications*. Wiley classics library. Wiley India Pvt. Limited, 2007. ISBN 9788126511914. URL <https://books.google.ch/books?id=osXw-pRsptoC>.
- S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.
- A. Lapidoth and C. Pfister. Testing Against Independence and a Rényi Information Measure. In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2018.
- S. Leung-Yan-Cheong and T. Cover. Some equivalences between shannon entropy and kolmogorov complexity. *IEEE Transactions on Information Theory*, 24(3):331–338, 1978. doi: 10.1109/TIT.1978.1055891.
- F. Liese and I. Vajda. *Convex Statistical Distances*. Teubner, Leipzig, 1987.
- F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theor.*, 52(10):4394–4412, 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.881731. URL <http://dx.doi.org/10.1109/TIT.2006.881731>.
- Jingbo Liu. *Information Theory from A Functional Viewpoint*. PhD thesis, Princeton, NJ : Princeton University, 2018.

Bibliography

- Adrian Tovar Lopez and Varun S. Jog. Generalization error bounds using wasserstein distances. *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2018.
- David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., USA, 1st edition, 1997. ISBN 047155359X.
- Lech Maligranda. Why hölder’s inequality should be called rogers’ inequality. *Mathematical Inequalities & Applications*, 1, 01 1998. doi: 10.7153/mia-01-05.
- J.L. Massey. Guessing and entropy. In *Proceedings of 1994 IEEE International Symposium on Information Theory*, pages 204–, 1994. doi: 10.1109/ISIT.1994.394764.
- Colin McDiarmid. *On the method of bounded differences*, page 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.
- Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017. doi: 10.1109/CSF.2017.11.
- Z.D. Ren M.M. Rao. *Theory of Orlicz Spaces*. New York: M. Dekker, 1991.
- Tetsuzo Morimoto. Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, 1963. doi: 10.1143/JPSJ.18.328. URL <https://doi.org/10.1143/JPSJ.18.328>.
- Hoi H. Nguyen and Van H. Vu. *Small Ball Probability, Inverse Theorems, and Applications*, pages 409–463. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-39286-3. doi: 10.1007/978-3-642-39286-3_16.
- XuanLong Nguyen, Martin J Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1089–1096. Curran Associates, Inc., 2008.
- Tomohiro Nishiyama and Igal Sason. On relations between the relative entropy and χ^2 -divergence, generalizations and applications. *Entropy*, 22(5), 2020. ISSN 1099-4300. doi: 10.3390/e22050563. URL <https://www.mdpi.com/1099-4300/22/5/563>.
- Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550, 2018.
- David Pollard. *A User’s Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2001. doi: 10.1017/CBO9780511811555.002.

- Yury Polyanskiy and Sergio Verdú. Arimoto channel coding converse and rényi divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1327–1333, 2010. doi: 10.1109/ALLERTON.2010.5707067.
- Yury Polyanskiy, H. Vincent Poor, and Sergio Verdu. Channel coding rate in the finite block-length regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, 2010. doi: 10.1109/TIT.2010.2043769.
- M. Raginsky and I. Sason. *Concentration of Measure Inequalities in Information Theory, Communications, and Coding: Second Edition*. Now Foundations and Trends, 2014.
- Maxim Raginsky. Strong data processing inequalities and ϕ -sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016. doi: 10.1109/TIT.2016.2549542.
- Firas Rassoul-Agha and Timo Seppäläinen. *A course on large deviations with an introduction to Gibbs measures*. 05 2015. ISBN 978-0-8218-7578-0.
- R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.
- Borja Rodríguez-Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. Tighter expected generalization error bounds via wasserstein distance, 2021. URL <https://arxiv.org/abs/2101.09315>.
- L.J. Rogers. An extension of a certain theorem in inequalities. *Messenger of Math.*, 17:145–150, 1888.
- Ryan M. Rogers, Aaron Roth, Adam D. Smith, and Om Dipakbhai Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–494, 2016.
- Avraham Ruderman, Mark D. Reid, Darío García-García, and James Petterson. Tighter variational representations of f-divergences via restriction to probability measures. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, page 1155–1162, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.
- Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240. PMLR, 09–11 May 2016.
- A. Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10:193–215, 1959. doi: 10.1007/BF02063299.
- A. Rényi. On measures of entropy and information. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, 1:547–561, 1960.

Bibliography

- Igal Sason. On f -divergences: Integral representations, local behavior, and inequalities. *Entropy*, 20(5), 2018a. ISSN 1099-4300. doi: 10.3390/e20050383. URL <https://www.mdpi.com/1099-4300/20/5/383>.
- Igal Sason. Tight bounds on the rényi entropy via majorization with applications to guessing and compression. *Entropy*, 20(12), 2018b. ISSN 1099-4300. doi: 10.3390/e20120896. URL <https://www.mdpi.com/1099-4300/20/12/896>.
- Igal Sason and Sergio Verdú. f -divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. doi: 10.1109/TIT.2016.2603151.
- Igal Sason and Sergio Verdú. Arimoto–rényi conditional entropy and bayesian m -ary hypothesis testing. *IEEE Transactions on Information Theory*, 64(1):4–25, 2018. doi: 10.1109/TIT.2017.2757496.
- S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.
- Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 163–171. Curran Associates, Inc., 2014.
- C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.
- C. E. Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1):50–64, 1951. doi: 10.1002/j.1538-7305.1951.tb01366.x.
- R. Sibson. Information radius. *Z. Wahrscheinlichkeitstheorie verw Gebiete* 14, pages 149–160, 1969.
- Thomas Steinke and Lydia Zakyntinou. Reasoning About Generalization via Conditional Mutual Information. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452. PMLR, 09–12 Jul 2020. URL <https://proceedings.mlr.press/v125/steinke20a.html>.
- M. Tomamichel and M. Hayashi. Operational interpretation of rényi information measures via composite hypothesis testing against product and markov distributions. *IEEE Transactions on Information Theory*, 64(2):1064–1082, 2018. doi: 10.1109/TIT.2017.2776900.
- T. van Erven and P. Harremoës. Rényi divergence and kullback-keibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, July 2014.
- S.R.S. Varadhan. *Large Deviations and Applications*. 1984.

- Sergio Verdú. α -mutual information. In *2015 Information Theory and Applications Workshop, ITA 2015, San Diego, CA, USA, February 1-6, 2015*, pages 1–6, 2015.
- E. Verriest. An achievable bound for optimal noiseless coding of a random variable (corresp.). *IEEE Transactions on Information Theory*, 32(4):592–594, 1986. doi: 10.1109/TIT.1986.1057200.
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.
- O.G. Smolyanov V.I. Bogachev. *Topological Vector Spaces and Their Applications*. Springer Monographs in Mathematics. Springer, Cham, 2017.
- C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003. ISBN 9780821833124. URL https://books.google.ch/books?id=R_nWqjq89oEC.
- C. Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.
- Mariia Vladimirova, Sté phane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1), jan 2020. doi: 10.1002/sta4.318.
- Hao Wang, Mario Diaz, José Cândido S. Santos Filho, and Flavio P. Calmon. An information-theoretic view of generalization via wasserstein distance. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 577–581, 2019. doi: 10.1109/ISIT.2019.8849359.
- Yihong Wu and Sergio Verdú. Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Transactions on Information Theory*, 56(8):3721–3748, 2010. doi: 10.1109/TIT.2010.2050803.
- A. Xu and M. Raginsky. Information-theoretic lower bounds on bayes risk in decentralized estimation. *IEEE Transactions on Information Theory*, 63(3):1580–1600, 2017a.
- A. Xu and M. Raginsky. Information-theoretic lower bounds for distributed function computation. *IEEE Transactions on Information Theory*, 63(4):2314–2337, 2017b.
- A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, page 2521–2530, 2017c.
- Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- Marek Śmieja and Jacek Tabor. Rényi entropy dimension of the mixture of measures. pages 685–689, 08 2014. doi: 10.1109/SAI.2014.6918261.