Université de Montréal

Une généralisation des preuves en théorie de l'information du cas discret au cas continu

par

Simon Hennessey-Patry

Département de physique
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Physique

30 avril 2023

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Une généralisation des preuves en théorie de l'information du cas discret au cas continu

présenté par

Simon Hennessey-Patry

a été évalué par un jury composé des personnes suivantes :

Richard MacKenzie

(président-rapporteur)

Frédéric Dupont-Dupuis

(directeur de recherche)

Louis Salvail

(membre du jury)

Résumé

L'objectif principal de ce mémoire est de généraliser du cas discret au cas continu plusieurs quantités, inégalités et preuves qui surviennent en théorie de l'information.

Dans plusieurs cas, à la place de transposer la preuve ou les quantités d'intérêts au continu, le cas discret est étendu à l'extrême en prenant un très grand nombre de probabilités discrètes. Nous espérons que ce mémoire puisse servir de ressource pour faciliter la transition du discret au continu et que les différentes quantités trouvées puissent servir de fondation à toute autre preuve concernant les variables continues en théorie de l'information.

Les premières sections présenteront un survol des fondements de la théorie de l'information, une introduction aux probabilités ainsi que des fondements mathématiques requis pour la compréhension du reste du document. Les sections subséquentes introduiront les analogues continus à la théorie de l'information classique, en plus de différentes inégalités et preuves en rapport avec ces quantités.

Mots-clefs : Rényi, différentiel, divergence, entropie, théorie de l'information, continu

Abstract

This document's main goal is to generalize multiple quantities, inequalities, and proofs that arise in information theory. Many of these proofs use discrete variables. We seek here to generalize these proofs to the continuous case.

In many instances, instead of transposing the proofs to the continuous case, the discrete case is taken to the extreme by taking very large pools of discrete possibilities. We hope that this thesis can serve as a tool to ease the transition from the discrete case to the continuous case and that the various quantities and bounds found herein will help in establishing a framework to prove statements regarding continuous variables in information theory.

The first few sections will present a review of elementary information theory, as well as a primer on probabilities and fundamental mathematical concepts required for the rest of the document. The later sections will introduce the continuous counterparts of classical information theory, as well as various inequalities and proofs with respect to these new quantities.

Keywords: Rényi, differential, divergence, entropy, information theory, continuous.

Table des matières

Liste des figures

Remerciements

J'aimerais remercier Frédéric Dupuis pour son soutien continu dans la construction de ce document.

Je voudrais aussi remercier Alexander Fribergh pour nos discussions sur différents concepts mathématiques.

Introduction

Notes historiques

La théorie de l'information a fait ses premiers pas au début du 20^{e} siècle avec les travaux pionniers de Harry Nyquist et Ralph Vinton Lyon Hartley, tous les deux chercheurs aux Bell Laboratories [**1**]. Ce sera Nyquist qui publiera le premier article en lien avec la théorie de l'information : un ouvrage portant sur le taux de transmission des données, adéquatement nommé « Certain Factors Affecting Telegraph Speed » [**2**]. Il dérivera dans son travail une formule pour ces taux de transmission dans des canaux non bruités avec une bande passante finie. L'article de Hartley, intitulé « Transmission of Information », établit les premiers fondements d'une théorie mathématique de l'information [**3**].

Ce sera Claude Shannon qui, une vingtaine d'années plus tard, portera la théorie de l'information à l'attention du monde entier avec la publication de son article fondateur : « A Mathematical Theory of Communication » [**4**]. C'est dans cet article que Shannon introduira un modèle des communications à l'aide de processus statistiques et probabilistes. L'article introduit des quantités qui sont, à ce jour, utilisées dans de multiples domaines autre que les communications, tels la théorie des probabilités, les statistiques, l'économie, les sciences informatiques et la physique. Les concepts de surprenance, d'entropie de variables aléatoires, de bits, et d'information mutuelle et conditionelle y sont introduits. Le théorème de codage de source ainsi que la loi de Shannon-Hartley y font aussi leur apparence.

L'idée fondamentale derrière la théorie de Shannon est que la valeur en information d'un message dépend largement du degré de surprenance de ce message. En d'autres mots, si un évènement très probable survient, le message contient très peu d'informations. À l'inverse, si un évènement extrêmement improbable survient, le message contient beaucoup plus d'informations. Par exemple : le message qui vous indique que la chaîne de numéros 14-33-02-05-37-48 ne gagnera pas le prochain tirage à la loterie 6/49 n'est pas très surprenant. Les chances sont extrêment basses que cette chaîne soit tirée en premier lieu, *ergo* le message est carrément vide d'informations. À l'inverse, si un message vous est communiqué disant

que cette même chaîne de numéros sera gagnante au prochain tirage, le message est riche en informations! La raison étant (selon la théorie de Shannon) que le message nous communique le résultat d'un évènement qui est extrêmement improbable. Mathématiquement, la théorie de l'information codifie ce concept de surprenance de la manière suivante. Si E est un évènement au sens des probabilités et $p(E)$ est la probabilité que cet évènement arrive, alors la surprenance de E est définie de la manière suivante :

$$S(E) = \log_2 \left(\frac{1}{p(E)} \right) \tag{0.0.1}$$

Si un évènement est certain, $p(E) = 1$, et $S(E) = 0$ (l'évènement n'est aucunement surprenant), en revanche, si $p(E)$ est très petit, alors la surprenance devient très grande.

C'est à partir de cette définition que Shannon développe son idée de l'entropie d'une variable aléatoire. Selon lui, l'entropie devrait mesurer la valeur attendue (la moyenne) d'informations gagnées par la connaissance du résultat d'une expérience aléatoire. En d'autres mots, si X est une variable aléatoire discrète avec valeurs possibles x et $\mathbb{E}[X]$ dénote l'espérance de X, alors l'entropie de X est donnée par :

$$H(X) = \mathbb{E}[S(X)] = \mathbb{E}\left[\log_2 \left(\frac{1}{p(X)} \right) \right] = -\sum_x p(x) \log_2 p(x) \tag{0.0.2}$$

Il n'est pas toujours évident d'interpréter cette quantité. Shannon a montré dans son article que l'entropie représente une limite mathématique de la qualité de compression sans perte de données sur un canal sans bruit. Plus intuitivement, l'entropie nous donne la réponse à la question : Quelle est la longueur moyenne de la plus courte description de la variable aléatoire en bits? Une autre manière intuitive de comprendre l'entropie est la suivante : le nombre minimum (en moyenne) de questions binaires (questions oui/non) requises pour déterminer X se trouve entre $H(X)$ et $H(X) + 1$ [**5**].

Cette définition de l'entropie est sous-jacente à presque toutes les autres quantités en théorie de l'information. La grande majorité des développements subséquents dans le domaine utilise cette définition.

Entropie de Rényi

À la suite de la publication de l'article de Shannon, la théorie de l'information a connu des développements rapides dans plusieurs cadres et domaines, tels que l'exploration de transmission de messages avec canaux bruités, de compression de signaux, de communication avec ressources limitées, de codes correcteurs, ainsi que plusieurs travaux théoriques à propos

de la structure mathématique de l'information. L'un des développements subséquents les plus importants fut mis de l'avant par Alfréd Rényi en 1961 avec la publication de son article : « On Measures of Information and Entropy » [**6**].

Rényi avait comme objectif de généraliser plusieurs des quantités entropiques existantes à l'époque, telles que l'entropie de Shannon, l'entropie de Hartley, la min-entropie, et l'entropie de collision (souvent appelée elle aussi entropie de Rényi). Il s'agissait de trouver une définition de l'entropie utilisant un paramètre quelconque de manière à retrouver les définitions des différentes quantités pour des valeurs spécifiques du paramètre. C'est avec cette motivation en tête que Rényi proposa la définition d'entropie de Rényi, une quantité dépendant d'un paramètre α, donnée par :

$$H_\alpha(X) = \frac{1}{1-\alpha} \log_2 \left(\sum_{i=1}^{n} p_i^\alpha \right) \tag{0.0.3}$$

Où les valeurs possibles de X sont les p_i, et la sommation se fait sur toutes les valeurs possibles de X (dans le cas ci-dessus, il y a n valeurs possibles pour X).

Les différentes quantités énoncées ci-dessus peuvent être retrouvées en posant une valeur spécifique de α. L'entropie de collision se trouve en posant $\alpha = 2$, et la min-entropie en prenant la limite de α à l'infini. On remarque notamment que l'entropie classique de Shannon se trouve en prenant la limite $\alpha \to 1$. Il s'ensuit aussi qu'avec l'introduction de cette entropie de Rényi, il y a la possibilité de trouver d'autres analogues de Rényi aux quantités classiques (entropie conditionelle, divergence de Kullback-Leibler, information mutuelle, etc.). La divergence de Rényi, par exemple, est d'une importance significative, et a été extensivement explorée sous différents angles. [**7**]

Il y a, depuis l'introduction de cette quantité, un intérêt marqué à l'étude des propriétés de l'entropie de Rényi et des quantités liées à celle-ci. En effet, en connaissant bien l'entropie de Rényi, nous connaîtrons bien les cas spécifiques de l'entropie de Rényi qui nous sont d'intérêt. Plusieurs recherches modernes utilisent différents cas et limites du paramètre α pour bien caractériser le système qu'elles considèrent.

Applications en mécanique quantique

L'une des applications les plus importantes de cette nouvelle entropie de Rényi fut trouvée dans le domaine de la mécanique quantique. La mécanique quantique, ayant aussi pris son envol dans la première moitié du 20$^{\text{e}}$ siècle, s'est rapidement développée en parallèle à la théorie de l'information. Les applications de cette nouvelle théorie, à l'époque, étaient

multiples et il n'a pas fallu attendre longtemps avant que le domaine des communications passe à l'ère quantique. En faisant le passage de variables aléatoires à des matrices densités, de bits à des qubits, et de canaux classiques à des canaux quantiques (par exemple, une fibre optique), plusieurs des concepts de la théorie de l'information trouvent un analogue quantique qui peut être utilisé pour construire une théorie de l'information quantique. Il s'avère que l'entropie de Rényi mentionnée précédemment peut être utilisée, lorsque modifiée adéquatement à des fins de calculs en mécanique quantique, comme mesure d'intrication, l'un des phénomènes quantiques ayant le plus de potentiel pour les communications [**8**].

La théorie de l'information quantique est un domaine prometteur avec plusieurs avenues intrigantes et fascinantes, notamment la cryptographie quantique. À l'aide de protocoles tels que la téléportation quantique et la distribution de clefs de chiffrement avec ces protocoles, la porte est ouverte à des protocoles de communication théoriquement sécuritaires contre des attaquants possédant eux-mêmes des capacités computationelles quantiques [**9**]. L'étude de l'information quantique est aussi nécessaire face aux développements de protocoles et algorithmes compromettant la sécurité de protocoles de communication classique, l'exemple principal étant la sécurité du chiffrement RSA, communément utilisée de nos jours et compromise par le développement d'algorithmes quantiques tels que l'algorithme de Shor [**10**].

Il s'avère cependant que la grande majorité des travaux effectués en théorie de l'information (tant classique que quantique) utilise des variables aléatoires discrètes dans leurs preuves et développements. Classiquement, le bit reste restreint à 0 et 1, et le qubit, malgré la possibilité de superposition qui utilise des nombres complexes, est restreint à des systèmes pouvant être modélisés par deux états quantiques parfaitement discernables par une mesure (e.g. des systèmes de spin $\frac{1}{2}$, des polarisations de photons qui sont orthogonales, ou bien deux niveaux d'excitation d'un atome) [**11**].

Motivation de ce travail

Avec le développement des technologies modernes et des communications utilisant pleinement la mécanique quantique, il devient nécessaire d'adapter les travaux faits à ce jour. Avec les avancées faites en photonique et en optique quantique, le nouveau standard de communication de fine pointe n'est plus le bit, mais le paquet d'ondes. Ces paquets d'ondes sont plus souvent qu'autrement définis à l'aide d'une gaussienne, et donc à l'aide de variables continues.

Plusieurs travaux concernant les preuves de sécurité de ces nouvelles technologies utilisent toujours des variables aléatoires discrètes et prennent un très grand nombre de résultats possibles sur ces variables aléatoires pour se rapprocher du continu, mais ne font pas le passage strict au continu. La raison principale derrière ceci est qu'une grande partie des inégalités utilisées contiennent des termes dépendants de la cardinalité de l'espace des possibilités des variables aléatoires. Ceci vient effectivement rendre inutile toute inégalité concernant des variables aléatoires continues, celles-ci ayant une cardinalité d'évènements possibles infinie.

La motivation principale de cet ouvrage est alors de concevoir un outil qui peut servir de base à la construction de preuves en cryptographie et en théorie de l'information utilisant des variables aléatoires continues. Malgré qu'il soit difficile de rattraper et d'adapter des années de progrès, nous tentons ici de poser des bases solides sur lesquelles s'appuyer pour que de futurs développements dans la même direction puissent être faits rigoureusement.

Survol de la littérature

Certains articles ont exploré la transition de certains théorèmes de la théorie de l'information au cas quantique. Un exemple important est l'article de Tomamichel, Colbeck, et Renner [**12**], qui explore la version quantique de la propriété d'équipartition asymptotique.

La possibilité de communication utilisant des variables continues quantiques a été proposée il y a plusieurs années, et des articles comme celui de Weedbrook [**11**] explorent les avenues d'une telle approche. Des travaux plus récents [**13**] explorent plus en profondeur la sécurité de distribution de clefs quantiques avec des variables quantiques continues à l'aide de modulations discrètes. La version quantique de différents processus, tels la compression de données et l'extraction d'aléas, a aussi été explorée par Tomamichel et Hayashi [**14**]

D'autres approches, telles celle utilisée par Leverrier [**15**], utilisent des méthodes basées sur la réduction de Finetti pour montrer la sécurité de protocoles de distribution de clefs utilisant des états cohérents contre des attaques utilisant des états gaussiens.

Lignes directrices du travail

La première section (chapitre 1) du document couvrira les fondements de la théorie de la mesure ainsi que les prérequis à la construction d'une théorie de la probabilité systématique qui sera utilisée pour le reste du document. Les définitions d'espaces mesurables, de mesures, ainsi que les définitions connexes d'espaces de probabilités y seront

présentées rigoureusement.Nous introduirons les éléments de probabilité nécessaires à la compréhension et à la construction des sections subséquentes. En plus de définir les notions de base telles que les évènements et les notions d'indépendance et de conditionnalité, nous énoncerons quelques-uns des résultats importants de la théorie des probabilités qui serviront à la construction des preuves, ainsi que certaines quantités d'intérêt en probabilités.

La seconde (chapitre 2) section survolera les définitions et théorèmes fondamentaux de la théorie de l'information. Nous y définirons toutes les quantités significatives liées à l'entropie que nous avons évoquée plus haut. Nous énoncerons aussi quelques théorèmes d'importance significative en lien avec ces quantités. Cette section bouclera la présentation des notions préliminaires à la construction des quantités continues que nous désirons définir.

Par la suite, les prochaines sections (chapitres 3 à 9) présenteront la construction des quantités de Rényi continues. Nous passerons par l'entropie différentielle, la divergence, l'information mutuelle, ainsi que l'entropie conditionnelle. Nous montrerons aussi quelques propriétés d'intérêt sur ces quantités.

Les dernières sections (chapitres 10 et 11) présenteront les analogues continus de deux inégalités d'importance marquée, soit l'inégalité entre l'entropie et l'entropie conditionnelle, ainsi qu'une inégalité entre l'entropie différentielle de Shannon et l'entropie différentielle de Rényi.

Chapitre 1

Probabilités

La théorie fondamentale sous-tendant les probabilités est la théorie de la mesure. Nous commençons par établir quelques définitions sur les ensembles et les mesures qui nous aideront à définir des espaces de probabilité [**16**]. Nous commençons par la définition d'une tribu.

Définition 1.0.1. *(Tribu) Soit Ω un ensemble non vide et soit $\mathcal{A} \subset 2^\Omega$ (l'ensemble puissance de Ω) une classe de sous-ensembles de Ω. $\mathcal{A}$ est appelé une tribu et remplit les conditions suivantes :*

(i) $\Omega \in \mathcal{A}$.

(ii) $\mathcal{A}$ *est fermé sous complément* ($A^c = \Omega \setminus A \in \mathcal{A}$ *pour tout ensemble* $A \in \mathcal{A}$).

(iii) $\mathcal{A}$ *est fermé sous union fini ou infini dénombrable(pour un nombre fini ou infini dénombrable d'ensembles dans* $\mathcal{A}$, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ *pour tout choix d'ensembles* $A_1, A_2,... \in \mathcal{A}$).

Les tribus jouent un rôle crucial dans la construction d'espaces mesurables, et donc d'espaces de probabilités. Cependant, pour décrire pleinement les espaces mesurables, nous devons définir d'autres constructions mathématiques :

Définition 1.0.2. *(Fonction d'ensemble) Soit $\mathcal{A} \subset 2^\Omega$ et $\mu : \mathcal{A} \to [0,\infty]$ une fonction d'ensemble. On dit que μ est :*

(i) monotone *si $\mu(A) \leq \mu(B)$ pour n'importe quels ensembles $A, B \in \mathcal{A}$ avec $A \subset B$.*

(ii) σ-additif *si $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ pour tout choix dénombrables d'ensembles mutuellement disjoints $A_1, A_2,... \in \mathcal{A}$ avec $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.*

(iii) σ-subadditif *si pour tout choix dénombrable d'ensembles $A, A_1, A_2,... \in \mathcal{A}$ avec $A \subset \bigcup_{i=1}^{\infty} A_i$, on a $\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i)$.*

Avec ces définitions en place, nous sommes maintenant aptes à définir une mesure ainsi qu'une mesure de probabilité.

Définition 1.0.3. *(pré-mesure, mesure, mesure de probabilité) Soit $\mathcal{A}$ une classe d'ensembles telle que :*

(i) $\emptyset \in \mathcal{A}$.

(ii) Pour tout $A, B \in \mathcal{A}$, *l'ensemble différence* $B \setminus A$ *est une union finie d'ensembles mutuellement disjoints dans* $\mathcal{A}$.

(iii) $\mathcal{A}$ *est* $\cap$*-fermé (*$A \cap B \in \mathcal{A}$ *lorsque* A,$B \in \mathcal{A}$*).*

Soit $\mu : \mathcal{A} \to [0,\infty]$ *une fonction d'ensemble telle que* $\mu(\emptyset) = 0$. *On appelle* μ *:*

- contenu *si* μ *est additive.*
- pré-mesure *si* μ *est* σ*-additive.*
- mesure *si* μ *est une pré-mesure et* $\mathcal{A}$ *est une tribu.*
- mesure de probabilité *si* μ *est une mesure et* $\mu(\Omega) = 1$.

Les différents espaces qui émergent des différentes mesures et types d'ensembles considérés sont les suivants :

Définition 1.0.4. *(espace mesurable, espace de mesure, espace de probabilité)*

(i) Une paire $(\Omega, \mathcal{A})$ *composée d'un ensemble non vide* Ω *et d'une tribu* $\mathcal{A} \subset 2^\Omega$ *est appelée un* espace mesurable. *Les ensembles* $A \in \mathcal{A}$ *sont appelés* ensembles mesurables. *Si* Ω *est au plus infini dénombrable et si* $\mathcal{A} = 2^\Omega$, *alors l'espace de mesure* $(\Omega, 2^\Omega)$ *est appelé* discret.

(ii) Un triple $(\Omega, \mathcal{A}, \mu)$ *est appelé un* espace de mesure *si* $(\Omega, \mathcal{A})$ *est un espace mesurable et si* μ *est une mesure sur* $\mathcal{A}$.

(iii) Si en plus de cela, $\mu(\Omega) = 1$, *alors* $(\Omega, \mathcal{A}, \mu)$ *est appelé un* espace de probabilité. *Dans ce cas, les ensembles* $A \in \mathcal{A}$ *sont appelés des* évènements.

Enfin, il existe une classe spéciale de sous-ensembles que nous considérons ici et qui sont nécessaires pour notre traitement des espaces de probabilités.

Définition 1.0.5. *(Topologie) Soit* $\Omega \neq \emptyset$ *un ensemble arbitraire. Une classe de sous-ensembles* $\tau \subset 2^\Omega$ *est appelée une* topologie *sur* Ω *si elle satisfait les trois propriétés suivantes :*

(i) $\emptyset, \Omega \in \tau$.
(ii) $A \cap B \in \tau$ *pour tout* A,$B \in \tau$.
(iii) $(\bigcup_{A \in \mathcal{F}} A) \in \tau$ *pour tout* $\mathcal{F} \subset \tau$.

Le couple (Ω, τ) *est appelé un* espace topologique. *Les ensembles* $A \in \tau$ *sont appelés* ouverts, *et les ensembles* $A \subset \Omega$ *avec* $A^c \in \tau$ *sont appelés* fermés.

Cela nous permet de définir la catégorie suivante d'algèbres.

Définition 1.0.6. *(tribu de Borel) Soit* (Ω, τ) *un espace topologique. La tribu*

$$\mathcal{B}(\Omega) = \mathcal{B}(\Omega, \tau) = \sigma(\tau)$$

qui est généré par les ensembles ouverts est appelé la tribu de Borel *sur* Ω. *Les éléments* $A \in \mathcal{B}(\Omega, \tau)$ *sont appelés des* ensembles de Borel *ou* ensembles de Borel mesurables.

Il existe un théorème important qui est d'intérêt significatif pour nous et qui concerne les mesures et leurs compositions : le théorème de décomposition de Lebesgue. En gros et sans entrer trop dans les catégorisations des mesures, le théorème de Lebesgue affirme que certaines mesures (telles que les mesures sur les algèbres-σ de Borel) peuvent être décomposées comme la somme de différents types de mesures. Autrement dit, une mesure ν peut être décomposée comme suit :

$$\nu = \nu_{cont} + \nu_{sing} + \nu_{pp} \tag{1.0.1}$$

où ν_{cont} est la partie absolument continue, ν_{sing} est la partie singulière continue, et ν_{pp} est la mesure ponctuelle (ou discrète).

La théorie des mesures et des espaces mesurables est vaste. Les définitions données ci-dessus effleurent la surface d'une structure mathématique complexe et profonde. Il serait possible de continuer à parler des propriétés de ces espaces et de leurs applications dans d'autres théories. Nous commençons donc à présent à restreindre notre point de vue aux applications des espaces mesurables, aux probabilités et aux espaces de probabilités. Nous commençons par la définition d'une variable aléatoire ainsi que sa distribution. Nous considérons à partir de maintenant un espace de probabilités $(\Omega, \mathcal{A}, P)$.

Définition 1.0.7. *(Variabe aléatoire) Soit* $(\Omega', \mathcal{A}')$ *un espace mesurable et soit* $X : \Omega \to \Omega'$ *mesurable.*

(i) X *est appelée une* variable aléatoire *à valeurs dans* $(\Omega', \mathcal{A}')$. *Si* $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, *alors* X *est appelée une variable aléatoire réelle ou simplement une variable aléatoire.*

(ii) Pour $A' \in \mathcal{A}'$, on note $X \in A' := X^{-1}(A')$ et $P[X \in A'] := P[X^{-1}(A')]$. En particulier, on pose $\{X \geq 0\} := X^{-1}([0, \infty))$ et on définit $\{X \leq b\}$ de manière similaire et ainsi de suite.

Définition 1.0.8. *(Distributions) Soit X une variable aléatoire.*

(i) La mesure de probabilité $P_X := P \circ X^{-1}$ est appelée la distribution *de X.*

(ii) Pour une variable aléatoire réelle X, l'application $F_X : x \mapsto P[X \leq x]$ est appelée la distribution de X *(ou plus précisément, de P_X). Nous écrivons $X \sim \mu$ si $\mu = P_X$ et nous disons que X a pour distribution μ.*

(iii) Un groupe de variables aléatoires $(X_i)_{i \in I}$ est dit distribué identiquement *si $P_{X_i} = P_{X_j}$ pour tout $i,j \in I$. On écrit $X = Y$ si $P_X = P_Y$.*

Ces définitions posent les bases de la théorie des probabilités systématiques que nous utiliserons pour le reste de ce travail et concluent notre courte introduction à la théorie de la mesure. Elles sous-tendent les définitions plus classiques de la probabilité issues des axiomes de Kolmogorov que la plupart d'entre nous connaissons et avec lesquels nous travaillerons.

Définition 1.0.9. *(Axiomes de Kolmogorov) Considérons l'espace de probabilités (Ω, F, P). On appelle évènement tout ensemble de Ω appartenant à F. Si l'évènement est constitué d'un seul élément, on dit que l'évènement est un évènement élémentaire. L'ensemble vide est alors nommé évènement impossible, et l'ensemble Ω est nommé évènement certain. Les évènements respectent alors les axiomes suivants :*

1- La probabilité d'un évènement E est un nombre réel tel que :

$$0 \leq P(E) \leq 1, \forall E \in F. \tag{1.0.2}$$

2- La probabilité qu'au moins un évènement de l'espace d'échantillonnage se produise est égale à 1 (évènement certain), c'est-à-dire que :

$$P(\Omega) = 1. \tag{1.0.3}$$

3- Une séquence dénombrable d'ensembles mutuellement exclusifs satisfait l'additivité-σ :

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n). \tag{1.0.4}$$

En utilisant ces axiomes et définitions, nous pouvons définir à la fois l'indépendance des évènements ainsi que les probabilités conditionnelles.

Définition 1.0.10. *Les évènements A et B sont dits indépendants si :*

$$P(A \cap B) = P(A)P(B). \tag{1.0.5}$$

Définition 1.0.11. *La probabilité conditionnelle (probabilité de B sachant A) est donnée par :*

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \tag{1.0.6}$$

Une définition plus conviviale d'une variable aléatoire est la suivante :

Définition 1.0.12. *Une variable aléatoire X sur F est une fonction qui associe les évènements de l'espace d'échantillonnage à F ($X : \Omega \to F$). La probabilité que X prenne la valeur x est alors donnée par :*

$$P(X = x) = \sum_{\omega \in \Omega, X(\omega) = x} P(\omega). \tag{1.0.7}$$

Lorsque le contexte est clair, nous écrivons souvent $P(x)$ comme une abréviation de $P(X = x)$.

Une quantité utile qui apparaît souvent en probabilité est l'espérance, ou valeur d'attente d'une variable aléatoire.

Définition 1.0.13. *(Valeur d'attente) La* valeur d'attente (espérance) *d'une variable aléatoire X est la somme de tous les évènements possibles pour X pondérés par la probabilité de ces évènements. Nous écrivons l'espérance comme :*

$$\mathbb{E}[X] = \sum_i x_i P(x_i). \tag{1.0.8}$$

L'espérance a plusieurs propriétés utiles, notamment la linéarité, la non-négativité, la monotonicité et la loi de l'espérance d'une fonction d'une variable aléatoire (connue par le nom populaire anglais : *Law of the unconscious statistician*, ou pour faire court, LOTUS):

- **Linéarité**
  Pour deux variables aléatoires X et Y, et une constante a, nous avons:

$$\mathbb{E}[aX + Y] = a\,\mathbb{E}[X] + \mathbb{E}[Y]. \tag{1.0.9}$$

  PREUVE. En utilisant les propriétés de la sommation, nous avons:

$$\mathbb{E}[aX + Y] = \sum_i p_i(ax_i + y_i) \tag{1.0.10}$$

$$= a\sum_i p_i x_i + \sum_i p_i y_i \tag{1.0.11}$$

$$= a\,\mathbb{E}[X] + \mathbb{E}[Y]. \tag{1.0.12}$$

  □

- **Non-négativité**
  Pour $X \geq 0$ (c'est à dire, toutes les valeurs possibles de X sont plus grandes ou égales à 0), $\mathbb{E}[X] \geq 0$.

  PREUVE. Les probabilités sont définies comme des nombres positifs. Pour une variable aléatoire qui prend des valeurs dans $[0,+\infty)$, la somme des produits de nombres positifs doit être supérieure ou égale à 0. Nous pouvons alors écrire de manière triviale :

$$\mathbb{E}[X] = \sum_i x_i p_i \geq 0. \tag{1.0.13}$$

  □

- **Monotonicité**
  Si $X \leq Y$, alors $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

  PREUVE. Si $X \leq Y$, alors pour tout i, nous avons $x_i \leq y_i$. L'inégalité découle directement de cette condition :

$$\mathbb{E}[X] = \sum_i x_i p_i \leq \sum_i y_i p_i = \mathbb{E}[Y]. \tag{1.0.14}$$

  □

- **LOTUS**
  Si X a une fonction de probabilité $P(x)$, alors l'espérance d'une fonction de X, disons

$g(X)$, est donnée par:

$$\mathbb{E}[g(X)] = \sum_i g(x_i)P(x_i). \tag{1.0.15}$$

La preuve est particulièrement difficile.

PREUVE. Supposons que la fonction g est différentiable et qu'elle admet une inverse, g^{-1}, qui est monotone. Soit X une variable aléatoire avec une fonction de masse de probabilité $P_X(x)$ et soit $Y = g(X)$ une autre variable aléatoire. Nous pouvons écrire explicitement la valeur d'espérance de Y comme suit :

$$\begin{aligned}
\mathbb{E}[g(x)] = \mathbb{E}[Y] &= \sum_i y_i P(Y = y_i) && (1.0.16)\\
&= \sum_i y_i P(g(x) = y_i) && (1.0.17)\\
&= \sum_i y_i P(g^{-1}(g(x)) = g^{-1}(y_i)) && (1.0.18)\\
&= \sum_i y_i P(x = g^{-1}(y_i)) && (1.0.19)\\
&= \sum_i y_i \sum_{x_i = g^{-1}(y_i)} P_X(x_i) && (1.0.20)\\
&= \sum_i \sum_{x_i = g^{-1}(y_i)} y_i P_X(x_i) && (1.0.21)\\
&= \sum_i \sum_{x_i = g^{-1}(y_i)} g(x_i) P_X(x_i). && (1.0.22)
\end{aligned}$$

La double sommation est équivalente à la sommation de tous les x_i si g^{-1} est monotone (ce qui est le cas). Nous pouvons donc conclure que :

$$\mathbb{E}[g(X)] = \sum_i \sum_{x_i = g^{-1}(y_i)} g(x_i) P_X(x_i) = \sum_j g(x_j) P_X(x_j). \tag{1.0.23}$$

□

Une autre quantité qui découle directement de la valeur attendue et qui est souvent utile est la variance. Nous la définissons ici.

Définition 1.0.14. *La variance d'une variable aléatoire X est écrite Var(X) et est donnée par :*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{1.0.24}$$

Plusieurs de ces définitions se transposent au cas où Ω, F et P sont définis sur des

espaces continus. La fonction de probabilité $P(x)$ se traduit par une fonction de densité de probabilité $p(x)$, et est donnée par la définition suivante :

Définition 1.0.15. *Une variable aléatoire continue X a une fonction de densité de probabilité $p(x)$ si, pour n'importe quel intervalle réel $[a,b]$ avec $a \leq b$, la probabilité que X prenne une valeur dans $[a,b]$ est donnée par :*

$$P(a \leq X \leq b) = \int_a^b p(x)\, d\mathrm{x}. \tag{1.0.25}$$

Nous écrirons souvent $p(x)$ plutôt que $p_X(x)$ pour plus de commodité si le contexte est clair.

Il existe de nombreuses inégalités d'intérêt en théorie de la probabilité. L'une des plus importantes utilisées dans ce document est l'inégalité de Jensen.

Définition 1.0.16. *Supposons que $p(x)$ est une fonction de densité de probabilité. Si g est une fonction à valeurs réelles et φ est convexe sur le domaine de g, alors :*

$$\varphi\left(\int_{-\infty}^{\infty} g(x)p(x)\, d\mathrm{x}\right) \leq \int_{-\infty}^{\infty} \varphi(g(x))p(x)\, d\mathrm{x}. \tag{1.0.26}$$

Une formulation courante est

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]. \tag{1.0.27}$$

Les inégalités sont inversées si φ est concave.

Ceci conclut notre brève introduction sur les probabilités ainsi que sur la théorie de la mesure.

Nous notons que dans les sections qui suivent, nous considérons des distributions de probabilités discrètes, ainsi que des distributions continues. Il s'ensuit que lorsque nous considérons des distributions discrètes, nous faisons référence à des distributions de probabilités prises sur un espace avec une mesure discrète. En d'autres mots, la mesure sera un ensemble de masse ponctuelle sur la ligne réelle. Lorsque nous considérons des distributions continues et des densités de probabilités, nous faisons référence à la mesure absolument continue sur l'ensemble réel considéré.

Chapitre 2

Théorie de l'information

Les fondements de la théorie de l'information ont été posés par Claude Shannon en 1948. Beaucoup des définitions qui sont utilisées à ce jour proviennent de son article phare intitulé « A Mathematical Theory of Communication ». Nous présentons ici certaines des définitions qui sont d'une importance fondamentale [**5**], en commençant par l'entropie d'une variable aléatoire discrète.

Le logarithme dominant en théorie de l'information est en base 2. Sans spécifier la base, nous supposons que $\log_2(x) = \log(x)$.

Définition 2.0.1. *(Entropie) Soit X une variable aléatoire discrète avec un alphabet (ensemble des résultats possibles) $\mathcal{X}$ et une fonction de masse de probabilité $p_X(x) = P\{X = x\}, x \in \mathcal{X}$. L'*entropie*, $H(X)$, de X est donnée par :*

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log(p_X(x)) = \mathbb{E}_p \left[\log \left(\frac{1}{p_X(x)} \right) \right]. \tag{2.0.1}$$

L'entropie peut être interprétée comme la longueur moyenne de la plus courte description de la variable aléatoire. Il peut également être compris (et prouvé) que le nombre minimal attendu de questions binaires requises pour déterminer X est compris entre $H(X)$ et $H(X)+1$.

Une fois que l'entropie est définie, nous pouvons introduire quelques quantités connexes.

Définition 2.0.2. *(Entropie jointe)*

*Soit (X,Y) un couple de variables aléatoires discrètes distribuées selon $p(x,y)$. L'*entropie conjointe *est définie comme suit :*

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log(p(x,y)). \tag{2.0.2}$$

Définition 2.0.3. *(Entropie conditionnelle) Soit $(X,Y) \sim p(x,y)$. L'*entropie conditionnelle *$H(Y|X)$ est définie comme suit :*

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \tag{2.0.3}$$
$$= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\log(p(y|x)) \tag{2.0.4}$$
$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(y|x)). \tag{2.0.5}$$

Il peut être démontré que les deux quantités précédentes sont liées par une expression communément appelée la règle de la chaîne.

Théorème 1. (Règle de la chaîne)

$$H(X,Y) = H(X) + H(Y|X). \tag{2.0.6}$$

PREUVE. Nous développons et appliquons la théorie de probabilité de base et l'algèbre.

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x,y)) \tag{2.0.7}$$
$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)p(x|y)) \tag{2.0.8}$$
$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x)) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x|y)) \tag{2.0.9}$$
$$= -\sum_{x\in\mathcal{X}} p(x)\log(p(x)) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x|y)) \tag{2.0.10}$$
$$= H(X) + H(Y|X). \tag{2.0.11}$$

□

Souvent, nous traitons de multiples distributions et espaces d'évènements. Il est alors utile de savoir à quel point les différentes distributions sont proches ou combien d'informations une distribution contient sur une autre. Les deux objets mathématiques les plus importants pour répondre à ces questions sont l'entropie relative (distance de Kullback-Leibler) et l'information mutuelle.

Définition 2.0.4. *(Entropie relative, distance de Kullback-Leibler, divergence) Soient $p(x)$ et $q(x)$ deux distributions sur $\mathcal{X}$. L'*entropie relative *entre $p(x)$ et $q(x)$ est définie comme suit :*

$$D(p \parallel q) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right). \tag{2.0.12}$$

Par convention, nous définissons $0\log\left(\frac{0}{q}\right) = 0\log\left(\frac{0}{0}\right) = 0$ et $p\log\left(\frac{p}{0}\right) = \infty$.

Il peut être démontré que la divergence est 0 si et seulement si $p = q$ et que $D(p \parallel q) \geq 0$. Nous notons également que, en général, $D(p \parallel q) \neq D(q \parallel p)$.

Définition 2.0.5. *(Information mutuelle) Considérons deux variables aléatoires discrètes $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ avec une fonction de masse de probabilité conjointe $p(x,y)$ et des fonctions de masse de probabilité marginales $p_X(x)$ et $p_Y(y)$. L'*information mutuelle *est définie comme suit:*

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \left(\frac{p(x,y)}{p_X(x)p_Y(y)} \right) \tag{2.0.13}$$

$$= D(p(x,y) \parallel p_X(x)p_Y(y)) \tag{2.0.14}$$

$$= H(Y) - H(Y|X). \tag{2.0.15}$$

L'information mutuelle est une mesure de la quantité d'information qu'une variable aléatoire contient à propos d'une autre. Cela est particulièrement évident lorsque nous lisons la dernière égalité de notre définition. Tout comme pour la divergence, nous pouvons montrer que $I(X;Y) \geq 0$ (comme le montre la deuxième ligne, l'information mutuelle peut être exprimée comme une divergence, et la divergence est non négative). Cependant, contrairement à la divergence, l'information mutuelle est symétrique dans ses arguments, de sorte que $I(X;Y) = I(Y;X)$.

Tout comme pour l'entropie, on peut définir la *mutualité conditionnelle* comme la réduction d'incertitude de X due à notre connaissance de Y lorsque nous connaissons Z.

Définition 2.0.6. L'information mutuelle conditionnelle *des variables aléatoires X et Y donnée Z est définie par*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z). \tag{2.0.16}$$

Comme pour l'entropie, cette quantité satisfait une règle de la chaîne.

Théorème 2. (Règle en chaîne pour l'information mutuelle) Soient $X_1, X_2,...,X_n$ et Y des variables aléatoires, alors

$$I(X_1, X_2,...,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_{i-1},X_{i-2},...,X_1). \tag{2.0.17}$$

PREUVE.

$$I(X_1, X_2,...,X_n; Y) \tag{2.0.18}$$

$$= H(X_1, X_2,...,X_n) - H(X_1,X_2,...,X_n|Y) \tag{2.0.19}$$

$$= \sum_{i=1}^{n} H(X_i; Y|X_{i-1},...,X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1},...,X_1,Y) \tag{2.0.20}$$

$$= \sum_{i=1}^{n} I(X_i; Y|X_1, X_2,...,X_{i-1}). \tag{2.0.21}$$

□

Il y a des théorèmes qui sont centraux en théorie de l'information. Certains des plus importants sont les suivants.

Théorème 3. (Réduction de l'entropie par conditionnement)

$$H(X|Y) \leq H(X), \tag{2.0.22}$$

avec égalité si et seulement si X et Y sont indépendantes.

PREUVE. Directement :

$$I(X; Y) = H(X) - H(X|Y) \geq 0 \Rightarrow H(X) \geq H(X|Y). \tag{2.0.23}$$

□

Théorème 4. (L'inégalité du traitement des données)

Soient X,Y et Z des variables aléatoires telles que X et Y forment une chaîne de Markov conditionnelle sur Z : $X \to Y \to Z$. Alors

$$I(X; Y) \geq I(X; Z). \tag{2.0.24}$$

En particulier, si $Z = g(Y)$, alors nous avons $I(X; Y) \geq I(X; g(Y))$.

PREUVE. Nous appliquons la règle de la chaîne pour développer la mutualité de l'information de deux manières différentes :

$$I(X; Y,Z) = I(X; Z) + I(X; Y|Z) \tag{2.0.25}$$

$$= I(X; Y) + I(X; Z|Y). \tag{2.0.26}$$

En rappelant que X et Z sont conditionnellement indépendants étant donné Y, on doit avoir $I(X; Z|Y) = 0$. Puisque l'information mutuelle est non-négative, on a :

$$I(X; Y) \geq I(X; Z). \tag{2.0.27}$$

Avec égalité si et seulement si $X \to Y \to Z$ (X, Y et Z forment une chaîne de Markov). □

Cette inégalité énonce essentiellement qu'aucun traitement physique local d'un signal ne peut augmenter l'information contenue dans le signal.

Il est intéressant de transposer autant que possible ces définitions au cas où les variables aléatoires considérées sont continues. Dans de nombreux cas, le passage de discret à continu se fait sans problème en substituant les sommes par des intégrales. Dans d'autres cas, la transition nécessite un ajustement de nos définitions. Nous présentons ici certaines des principales définitions liées aux variables aléatoires continues.

Définition 2.0.7. *(Entropie différentielle) Soit X une variable aléatoire continue avec une fonction de densité de probabilité $p(x)$ et un support S. L'*entropie différentielle *est définie comme*

$$h(X) = -\int_S p(x)\,\log(p(x))\,\mathrm{d}x. \tag{2.0.28}$$

Définition 2.0.8. *(Entropie conditionnelle différentielle) Si X et Y ont une fonction de densité de probabilité jointe $f(x,y)$. L'*entropie différentielle conditionnelle *est alors définie comme*

$$h(X|Y) = -\iint f(x,y)\log(f(x|y))\,\mathrm{d}x\mathrm{d}y. \tag{2.0.29}$$

Définition 2.0.9. *La* divergence différentielle *entre deux densités f et g, dont le support de f est contenu dans celui de g, est donnée par*

$$D(f \parallel g) = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)\,\mathrm{d}x. \tag{2.0.30}$$

Définition 2.0.10. *L'* information mutuelle différentielle *entre deux variables aléatoires avec une densité jointe $f(x,y)$ est définie comme*

$$I(X;Y) = \iint f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right)\mathrm{d}x\mathrm{d}y. \tag{2.0.31}$$

Ces quantités et théorèmes posent les fondements de la théorie moderne de l'information, et pratiquement toutes les avancées en découlent. Un développement important de la théorie a été proposé par Alfréd Rényi, qui cherchait à généraliser l'entropie de Shannon. Dans son article de 1961, Rényi a proposé l'expression suivante pour atteindre cet objectif :

Définition 2.0.11. *(Entropie de Rényi) Soit $\alpha \geq 0, \alpha \neq 1$, et X une variable aléatoire discrète avec une fonction de masse de probabilité $p(x)$. L'*entropie de Rényi *d'ordre α est définie comme suit :*

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n} p_i^\alpha(x)\right). \tag{2.0.32}$$

En particulier, la limite $\alpha \to 1$ nous redonne l'entropie de Shannon.

Comme pour le passage du cas discret au cas continu, nous sommes intéressés à trouver des quantités qui peuvent être généralisées en tant que quantités « Rényi ». Ces quantités, tout en généralisant nos concepts, récupéreraient les valeurs de Shannon habituelles lorsque la limite $\alpha \to 1$ est appliquée. La plupart des recherches modernes utilisent l'entropie de Rényi et des valeurs associées dans leurs preuves, car les différents ordres d'entropies de Rényi donnent différents types d'informations sur le système considéré.

Ces quantités de Rényi ont été largement développées dans le cadre discret. Nous cherchons à partir de maintenant à trouver, définir et catégoriser les analogues continus de Rényi aux cas discrets classiques.

Chapitre 3

Entropie de Rényi différentielle ($H_\alpha \to h_\alpha$)

Notre premier objectif est de définir correctement l'analogue continu de l'entropie de Rényi.
Théorème 5. (Entropie de Rényi différentielle) Soit X une variable aléatoire continue avec une fonction de densité de probabilité $f(x)$. Si le domaine de X est divisé en intervalles de largeur Δ, nous pouvons définir une variable aléatoire quantifiée de X comme X^Δ, telle que :

$$X^\Delta = x_i \ , \ i\Delta \leq X < (i+1)\Delta. \tag{3.0.1}$$

L'entropie différentielle de Rényi de X est alors donnée par la limite suivante :

$$h_\alpha(X) = \lim_{\Delta \to 0} \left[H_\alpha(X^\Delta) - \frac{1}{1-\alpha} \log\left(\Delta^{\alpha-1}\right) \right]. \tag{3.0.2}$$

et celle-ci est donnée par :

$$h_\alpha(X) = \frac{1}{1-\alpha} \log\left(\int f^\alpha(x)\, \mathrm{d}x \right). \tag{3.0.3}$$

Preuve. Soit X une variable aléatoire avec une densité $f(x)$ et soit le domaine de X divisé en intervalles de longueur Δ. Le théorème des accroissements finis nous garantit qu'il existe une valeur x_i dans chaque intervalle tel que :

$$\int_{i\Delta}^{(i+1)\Delta} f(x)\, \mathrm{d}x = f(x_i)\Delta. \tag{3.0.4}$$

Nous pouvons introduire une variable quantifiée. Soit X^Δ telle que :

$$X^\Delta = x_i \ , \ i\Delta \leq X < (i+1)\Delta. \tag{3.0.5}$$

La probabilité que $X^\Delta = x_i$ est donnée par le théorème des accroissements finis :

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)\, \mathrm{d}x = f(x_i)\Delta. \tag{3.0.6}$$

L'entropie de Rényi pour la variable quantifiée est alors donnée par :

$$\begin{aligned} H_\alpha(X^\Delta) &= \frac{1}{1-\alpha}\log\left(\sum_i p^\alpha(x_i)\right) && (3.0.7)\\ &= \frac{1}{1-\alpha}\log\left(\sum_i (f(x_i)\Delta)^\alpha\right) && (3.0.8)\\ &= \frac{1}{1-\alpha}\log\left(\Delta^{\alpha-1}\sum_i f^\alpha(x_i)\Delta\right) && (3.0.9)\\ &= \frac{1}{1-\alpha}\log\left(\sum_i f^\alpha(x_i)\Delta\right) + \frac{1}{1-\alpha}\log\left(\Delta^{\alpha-1}\right). && (3.0.10) \end{aligned}$$

Si nous prenons la limite $\Delta \to 0$ et si la fonction $f^\alpha(x)$ est intégrable au sens de Riemann, le terme à l'intérieur du premier logarithme se rapproche de l'intégrale de $f^\alpha(x)$. Nous combinons cela avec les lignes précédentes pour conclure que :

$$H_\alpha(X^\Delta) - \frac{1}{1-\alpha}\log\left(\Delta^{\alpha-1}\right) \xrightarrow[\Delta\to 0]{} h_\alpha(X) \quad (3.0.11)$$

et

$$h_\alpha(X) = \frac{1}{1-\alpha}\log\left(\int f^\alpha(x)\,\mathrm{d}x\right). \quad (3.0.12)$$

□

La preuve et le théorème se généralisent facilement au cas multivarié, comme nous le montrons ci-dessous.

Théorème 6. Soit $\mathbf{X} = (X_1, X_2,...,X_n)$ une variable aléatoire multivariée avec une densité $f(\mathbf{x})$. En divisant l'espace en volumes de taille Δ^n, on peut définir une variable aléatoire quantifiée de $\mathbf{X}$ comme $\mathbf{X}^\Delta$ telle que :

$$\mathbf{X}^\Delta = \mathbf{x}_i,\ \forall j,\ 0 \le j \le n,\ i\Delta \le X_j < (i+1)\Delta. \quad (3.0.13)$$

L'entropie différentielle de Rényi de $\mathbf{X}$ est alors donnée par la limite suivante :

$$H_\alpha(\mathbf{X}^\Delta) - \frac{1}{1-\alpha}\log\left((\Delta^{\alpha-1})^n\right) \xrightarrow[\Delta\to 0]{} h_\alpha(\mathbf{X}). \quad (3.0.14)$$

et l'entropie différentielle de Rényi est écrite :

$$h_\alpha(\mathbf{X}) = \frac{1}{1-\alpha}\log\left(\int f^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x}\right). \quad (3.0.15)$$

Preuve. Soit le domaine de $f(\mathbf{x})$ divisé en volumes de taille Δ^n (nous divisons effectivement chaque dimension en intervalles de largeur Δ). Le théorème des accroissements finis multivarié nous assure que, pour chaque volume, il existe un vecteur de valeurs $\mathbf{x}_0$ tel que :

$$\int_{\Delta^n} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = f(\mathbf{x}_0)\Delta^n. \tag{3.0.16}$$

La variable aléatoire discrétisée $\mathbf{X}^\Delta$ est définie par :

$$\mathbf{X}^\Delta = \mathbf{x}_i,\ \forall j,\ 0 \leq j \leq n,\ i\Delta \leq X_j < (i+1)\Delta. \tag{3.0.17}$$

La probabilité que $\mathbf{X}^\Delta = \mathbf{x}_i$ est alors donnée par :

$$p_i = \int_{\Delta^n} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = f(\mathbf{x}_i)\Delta^n. \tag{3.0.18}$$

L'entropie de Rényi est alors donnée par :

$$H_\alpha(\mathbf{X}^\Delta) = \frac{1}{1-\alpha}\log\left(\sum_i p^\alpha(\mathbf{x}_i)\right) \tag{3.0.19}$$

$$= \frac{1}{1-\alpha}\log\left(\sum_i (f(\mathbf{x}_i)\Delta^n)^\alpha\right) \tag{3.0.20}$$

$$= \frac{1}{1-\alpha}\log\left((\Delta^n)^{\alpha-1}\sum_i (f^\alpha(\mathbf{x}_i)\Delta^n)\right) \tag{3.0.21}$$

$$= \frac{1}{1-\alpha}\log\left(\sum_i f^\alpha(\mathbf{x}_i)\Delta^n\right) + \frac{1}{1-\alpha}\log\left((\Delta^n)^{\alpha-1}\right). \tag{3.0.22}$$

En prenant la limite $\Delta \to 0$, on obtient notre résultat :

$$H_\alpha(\mathbf{X}^\Delta) - \frac{1}{1-\alpha}\,\log\left((\Delta^{\alpha-1})^n\right) \xrightarrow[\Delta\to 0]{} h_\alpha(\mathbf{X}) \tag{3.0.23}$$

et

$$h_\alpha(\mathbf{X}) = \frac{1}{1-\alpha}\,\log\left(\int f^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x}\right). \tag{3.0.24}$$

□

Chapitre 4

Divergence $D_\alpha(P^\Delta \parallel Q^\Delta)$

La divergence de Rényi discrète est donnée par :

$$D_\alpha(P \parallel Q) = \frac{1}{\alpha - 1} \log \left(\sum_i \frac{p_i^\alpha}{q_i^{\alpha-1}} \right). \tag{4.0.1}$$

Nous pouvons généraliser au cas continu par la même méthode utilisée pour trouver l'entropie différentielle de Rényi. Le résultat est donné par le théorème suivant.

Théorème 7. (Divergence différentielle de Rényi) Considérons deux fonctions de densité de probabilité $f(x)$ et $g(x)$, le support de $f(x)$ étant un sous-ensemble du support de $g(x)$. La divergence différentielle de Rényi entre les deux distributions est donnée par :

$$D_\alpha(P \parallel Q) = \frac{1}{\alpha - 1} \log \left(\int \frac{f^\alpha(x)}{g^{\alpha-1}(x)} \, \mathrm{d}x \right). \tag{4.0.2}$$

Preuve. En utilisant essentiellement la même procédure que celle que nous avons utilisée dans la section 1, nous pouvons considérer les distributions continues P et Q avec une densité $f(x)$ et $g(x)$, respectivement. En appliquant le théorème des accroissements finis et en divisant le domaine en intervalles de longueur Δ, nous obtenons la divergence de Rényi quantifiée :

$$D_\alpha(P^\Delta \parallel Q^\Delta) = \frac{1}{\alpha - 1} \log \left(\sum_i \frac{f^\alpha(x_i)\Delta^\alpha}{g^{\alpha-1}(y_i)\Delta^{\alpha-1}} \right). \tag{4.0.3}$$

La plupart des Δ se simplifient et nous laissent avec :

$$D_\alpha(P^\Delta \parallel Q^\Delta) = \frac{1}{\alpha - 1} \log \left(\sum_i \frac{f^\alpha(x_i)\Delta}{g^{\alpha-1}(y_i)} \right). \tag{4.0.4}$$

En prenant la limite $\Delta \to 0$, nous obtenons le résultat souhaité (à condition que les fonctions f et g soient Riemann-intégrables):

$$\lim_{\Delta \to 0} D_\alpha(P^\Delta \parallel Q^\Delta) = D_\alpha(P \parallel Q) = \frac{1}{\alpha - 1} \log \left(\int \frac{f^\alpha(x)}{g^{\alpha-1}(x)} \, \mathrm{d}x \right). \tag{4.0.5}$$

□

Nous notons que, bien que le théorème des accroissements finis extrait des points différents dans chaque intervalle (x_i et y_i), l'intégrabilité de Riemann de f et g rend le choix de points dans l'intervalle sans importance lorsque nous prenons la limite; les deux fonctions convergent vers leur intégrale. En d'autres mots, il suffit de définir une fonction $F(x) = \frac{f^\alpha(x)}{g^{\alpha-1}(x)}$ et d'appliquer le théorème des accroissements finis à cette fonction $F(x)$ pour retrouver le même résultat.

Nous notons également que si nous prenons la limite $\alpha \to 1$ en utilisant la règle de L'Hôpital, nous retrouvons la divergence de Kullback-Leibler pour des distributions continues :

$$\lim_{\alpha \to 1} D_\alpha(p \parallel q) = D_{KL}(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) \, \mathrm{d}x. \tag{4.0.6}$$

Nous remarquons également que, pour les distributions multivariées, nous pouvons appliquer le même raisonnement utilisé pour l'entropie différentielle de Rényi. En divisant chaque dimension en intervalles et en appliquant le théorème des accroissements finis, puis en prenant la limite $\Delta \to 0$, nous obtenons :

$$D_\alpha(p(x_1,x_2,...,x_n) \parallel q(x_1,x_2,...,x_n)) = \frac{1}{\alpha - 1} \log \left(\int \frac{p^\alpha(\mathbf{x})}{q^{\alpha-1}(\mathbf{x})} \, \mathrm{d}\mathbf{x} \right). \tag{4.0.7}$$

En particulier, pour les distributions jointes à deux variables, on a :

$$D_\alpha(p_{XY} \parallel q_{XY}) = \frac{1}{\alpha - 1} \log \left(\iint \frac{(p(x,y))^\alpha}{(q(x,y))^{\alpha-1}} \, \mathrm{d}x\mathrm{d}y \right). \tag{4.0.8}$$

Chapitre 5

Information mutuelle I_α

Nous pouvons trouver l'information mutuelle différentielle de Rényi en généralisant la définition de l'information mutuelle. L'information mutuelle est donnée par

$$I(X,Y) = D(p(X,Y) \parallel p(X)p(Y)). \tag{5.0.1}$$

Nous observons que cette quantité est équivalente à

$$I(X,Y) = \min_{q(Y)} D(p(X,Y) \parallel p(X)q(Y)). \tag{5.0.2}$$

En étendant cette définition, on peut s'attendre à ce que

$$I_\alpha(X,Y) = \min_{q(Y)} D_\alpha(p(X,Y) \parallel p(X)q(Y)). \tag{5.0.3}$$

Et de la définition de la divergence que nous avons trouvée dans la section 2, ceci est

$$I_\alpha(X,Y) = \min_{q(Y)} \frac{1}{\alpha - 1} \log \left(\iint \frac{p^\alpha(x,y)}{p^{\alpha-1}(x)q^{\alpha-1}(y)} \, \mathrm{d}x\mathrm{d}y \right). \tag{5.0.4}$$

Pour soutenir cette définition, nous remarquons que le cas limite $\alpha \to 1$ converge vers la définition de l'information mutuelle pour des variables continues via L'Hôpital :

$$\lim_{\alpha \to 1} I_\alpha(X,Y) = I(X,Y) = \min_{q(Y)} \iint p(x,y) \log \left(\frac{p(x,y)}{p(x)q(y)} \right) \mathrm{d}x\mathrm{d}y. \tag{5.0.5}$$

Théorème 8. (Information mutuelle différentielle de Rényi) Étant donné deux variables aléatoires X et Y avec une fonction de densité de probabilité jointe $p_{XY}(x,y)$ et les marginales $p_X(x)$ et $q_Y(y)$, l'information mutuelle différentielle de Rényi est donnée par :

$$I_\alpha(X,Y) = \min_{q(Y)} D_\alpha(p(X,Y) \parallel p(X)q(Y)) = \min_{q(Y)} \frac{1}{\alpha - 1} \log \left(\iint \frac{p^\alpha(x,y)}{p^{\alpha-1}(x)q^{\alpha-1}(y)} \, \mathrm{d}x\mathrm{d}y \right). \tag{5.0.6}$$

Chapitre 6

Non-négativité de la divergence et de l'information mutuelle

Théorème 9. (Divergence est positive) Pour n'importe quelle paire de fonctions de densité de probabilité, nous avons

$$D_\alpha(p \parallel q) \geq 0. \tag{6.0.1}$$

Preuve. Nous rappelons que :

$$D_\alpha(p \parallel q) = \frac{1}{\alpha - 1} \log \left(\int \frac{p^\alpha(x)}{q^{\alpha-1}(x)} \, \mathrm{d}x \right) = \frac{1}{\alpha - 1} \log \left(\int p(x) \left(\frac{q(x)}{p(x)} \right)^{1-\alpha} \, \mathrm{d}x \right). \tag{6.0.2}$$

Pour démontrer l'inégalité, nous appliquons l'inégalité de Jensen à l'intégrande à l'intérieur du logarithme. L'inégalité de Jensen pour les distributions de probabilités indique que pour une fonction de densité de probabilité $f(x)$ et pour une fonction mesurable réelle $g(x)$ avec $\varphi(x)$ une fonction convexe sur l'image de $g(x)$:

$$\varphi \left(\int_{-\infty}^{\infty} g(x) f(x) \, \mathrm{d}x \right) \leq \int_{-\infty}^{\infty} \varphi(g(x)) f(x) \, \mathrm{d}x. \tag{6.0.3}$$

Si φ est concave sur l'image de $g(x)$, l'inégalité est inversée :

Dans notre cas, posons $f(x) = p(x)$, $g(x) = \frac{q(x)}{p(x)}$, et $\varphi(x) = x^{1-\alpha}$. Nous notons que $g(x)$ a une image de valeurs de $[0, +\infty)$, puisque $p(x)$ et $q(x)$ sont toutes deux des fonctions de densité de probabilité. Nous pouvons facilement vérifier si $\varphi(x)$ est convexe ou concave en vérifiant si sa deuxième dérivée est positive ou négative. Un calcul rapide montre que :

$$\varphi''(x) = \frac{\alpha(\alpha - 1)}{x^{\alpha+1}} < 0 \quad \text{pour} \quad 0 < \alpha < 1, \quad x > 0, \quad \text{(Concave)} \tag{6.0.4}$$

$$\varphi''(x) = \frac{\alpha(\alpha-1)}{x^{\alpha+1}} > 0 \quad \text{pour} \quad 1 < \alpha < +\infty, \quad x > 0, \quad \text{(Convexe)} \tag{6.0.5}$$

Nous pouvons donc appliquer l'inégalité de Jensen à notre intégrande. Dans le cas où nous avons $0 < \alpha < 1$, nous obtenons :

$$\int p(x) \left(\frac{q(x)}{p(x)}\right)^{1-\alpha} \mathrm{d}x \leq \left(\int p(x) \left(\frac{q(x)}{p(x)}\right) \mathrm{d}x\right)^{1-\alpha} = \left(\int q(x) \mathrm{d}x\right)^{1-\alpha} = 1. \tag{6.0.6}$$

Nous pouvons alors facilement observer que la divergence doit être non négative, puisque l'intérieur du logarithme est ≤ 1 et $0 < \alpha < 1$:

$$D_\alpha(p \parallel q) = \underbrace{\frac{1}{\alpha - 1}}_{<0} \underbrace{\log\left(\int p(x) \left(\frac{q(x)}{p(x)}\right)^{1-\alpha} \mathrm{d}x\right)}_{\leq 0} \geq 0. \tag{6.0.7}$$

En utilisant un raisonnement similaire, nous pouvons montrer que pour $1 < \alpha < +\infty$, l'intérieur du logarithme sera supérieur ou égal à 1, et donc :

$$D_\alpha(p \parallel q) = \underbrace{\frac{1}{\alpha - 1}}_{>0} \underbrace{\log\left(\int p(x) \left(\frac{q(x)}{p(x)}\right)^{1-\alpha} \mathrm{d}x\right)}_{\geq 0} \geq 0. \tag{6.0.8}$$

Nous pouvons conclure que :

$$D_\alpha(p \parallel q) \geq 0, \forall p,q, \text{with } \alpha > 0. \tag{6.0.9}$$

□

Nous pouvons rapidement montrer à partir de ce résultat que les informations mutuelles doivent également être non négatives.

Théorème 10. Pour tout couple de variables aléatoires, on a

$$I_\alpha(X,Y) \geq 0. \tag{6.0.10}$$

Preuve. Il s'ensuit que l'information mutuelle, étant un cas particulier de divergence, doit également être non négative. Autrement dit,

$$D_\alpha(p \parallel q) \geq 0 \implies I_\alpha(X,Y) \geq 0. \tag{6.0.11}$$

Où, selon le théorème 8, p serait défini comme étant $p(X,Y)$ et q serait défini comme étant $p_X \cdot p_Y$, avec p_Y minimisé sur toutes les distributions possibles pour Y. □

Tout comme avec la version classique de l'information mutuelle, on note que l'information mutuelle est en moyenne non négative. Il existe certains scénarios où la connaissance d'une variable aléatoire augmente notre incertitude sur une autre.

Chapitre 7

Entropie de Rényi d'une distribution normale

Il est utile de dériver l'entropie de Rényi de certaines fonctions courantes. Nous considérons ici l'entropie de Rényi différentielle d'une distribution normale. Soit X une variable aléatoire distribuée selon une distribution gaussienne $\phi(x)$ avec une déviation standard σ et une moyenne de zéro ($\mu = 0$) :

$$X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right). \tag{7.0.1}$$

En utilisant notre définition de l'entropie de Rényi différentielle, nous obtenons :

$$\begin{aligned} h_\alpha(\phi) &= \frac{1}{1-\alpha} \log\left(\int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma^2)^{\frac{\alpha}{2}}} \exp\left(\frac{-\alpha x^2}{2\sigma^2}\right) \mathrm{d}x\right) &(7.0.2)\\ &= \frac{1}{\ln(2)(1-\alpha)} \ln\left(\frac{1}{(2\pi\sigma^2)^{\frac{\alpha}{2}}} \int_{-\infty}^{+\infty} \exp\left(\frac{-\alpha x^2}{2\sigma^2}\right) \mathrm{d}x\right). &(7.0.3)\end{aligned}$$

L'intégrale dans le logarithme est une intégrale gaussienne standard avec la substitution $b = \frac{\alpha}{2\sigma^2}$:

$$\int_{-\infty}^{+\infty} \exp\left(\frac{-\alpha x^2}{2\sigma^2}\right) \mathrm{d}x = \int_{-\infty}^{+\infty} \exp\left(-bx^2\right) \mathrm{d}x = \sqrt{\frac{\pi}{b}} = \sqrt{\frac{2\pi\sigma^2}{\alpha}}. \tag{7.0.4}$$

En substituant ce résultat dans notre équation pour l'entropie de Rényi, en regroupant les différents facteurs et en effectuant un peu d'algèbre, nous obtenons :

$$
\begin{aligned}
h_\alpha(\phi) &= \frac{1}{\ln(2)(1-\alpha)} \ln\left(\frac{1}{(2\pi\sigma^2)^{\frac{\alpha}{2}}}\sqrt{\frac{2\pi\sigma^2}{\alpha}}\right) && (7.0.5)\\
&= \frac{1}{\ln(2)(1-\alpha)} \ln\left(\left(\sqrt{\frac{2\pi\sigma^2}{\alpha^{\frac{1}{1-\alpha}}}}\right)^{1-\alpha}\right) && (7.0.6)\\
&= \frac{1}{\ln(2)(1-\alpha)}(1-\alpha)\ln\left(\sqrt{\frac{2\pi\sigma^2}{\alpha^{\frac{1}{1-\alpha}}}}\right) && (7.0.7)\\
&= \frac{1}{2ln(2)} \ln\left(\frac{2\pi\sigma^2}{\alpha^{\frac{1}{1-\alpha}}}\right) && (7.0.8)\\
&= \frac{1}{2} \log\left(\frac{2\pi\sigma^2}{\alpha^{\frac{1}{1-\alpha}}}\right). && (7.0.9)
\end{aligned}
$$

Chapitre 8

Entropie conditionnelle $h_\alpha(X|Y)$

Il existe plusieurs définitions de l'entropie conditionnelle de Rényi dans la littérature. Nous adoptons la définition suivante, notre principale motivation étant que le cas limite $\alpha \to 0$ nous redonne l'entropie conditionnelle différentielle.

Définition 8.0.1. *Considérons les deux variables aléatoires X et Y, avec des densités $p(x)$ et $q(x)$, respectivement. L'entropie conditionnelle différentielle $h_\alpha(X|Y)$ est alors donnée par:*

$$h_\alpha(X|Y) = -\log\left[\left(\int_y p(y)\left(\int_x p_{x|y}^\alpha(x|y)\,\mathrm{d}x\right)^{\frac{1}{\alpha}}\,\mathrm{d}y\right)^{\frac{\alpha}{\alpha-1}}\right]. \tag{8.0.1}$$

Pour vérifier le cas limite $\alpha \to 1$, nous commençons par réécrire l'équation comme suit :

$$h_\alpha(X|Y) = \frac{-1}{1-\frac{1}{\alpha}}\log\left[\int_y p(y)\left(\int_x p_{x|y}^\alpha(x|y)\right)^{\frac{1}{\alpha}}\,\mathrm{d}x\,\mathrm{d}y\right]. \tag{8.0.2}$$

La limite se prend directement :

$$\lim_{\alpha\to 1} h_\alpha(X|Y) = \frac{0}{0}. \tag{8.0.3}$$

Nous pouvons appliquer la règle de L'Hôpital. En dérivant le numérateur et le dénominateur par rapport à α :

$$\lim_{\alpha\to 1} h_\alpha(X|Y) = \lim_{\alpha\to 1}\frac{-1}{\left(\frac{1}{\alpha^2}\right)}\frac{1}{\ln(2)\int_y p(y)\left(\int_x p_{x|y}^\alpha(x|y)\right)^{\frac{1}{\alpha}}}\frac{\partial}{\partial\alpha}\left(\int_y p(y)\left(\int_x p_{x|y}^\alpha(x|y)\right)^{\frac{1}{\alpha}}\right) \tag{8.0.4}$$

$$= \frac{-1}{\ln(2)}\lim_{\alpha\to 1}\frac{\partial}{\partial\alpha}\left(\int_y p(y)\left(\int_x p_{x|y}^\alpha(x|y)\right)^{\frac{1}{\alpha}}\right) \tag{8.0.5}$$

$$= \frac{-1}{\ln(2)}\lim_{\alpha\to 1}\int_y p(y)\frac{\partial}{\partial\alpha}\exp\left[\frac{1}{\alpha}\ln\left(\int_x p_{x|y}^\alpha(x|y)\right)\right]. \tag{8.0.6}$$

Il ne reste qu'à prendre la dérivée du terme exponentiel. En nous épargnant l'algèbre, on obtient

$$\lim_{\alpha\to 1} h_\alpha(X|Y) = \frac{-1}{\ln(2)} \lim_{\alpha\to 1} \left(\int_y p(y) \left(\int_x p_{x|y}^\alpha(x|y) \right)^{\frac{1}{\alpha}} \right) \left[\frac{-\ln\left(\int_x p_{x|y}^\alpha(x|y) \right)}{\alpha^2} + \frac{\int_x p_{x|y}^\alpha \ln\left(p_{x|y}(x|y) \right)}{\alpha \int_x p_{x|y}^\alpha(x|y)} \right]. \tag{8.0.7}$$

En prenant la limite, le premier terme entre les parenthèses tend vers 0 après intégration de la distribution marginale à l'intérieur du logarithme, et le dénominateur du deuxième terme tend également vers 1 après intégration. Nous pouvons également intégrer à 1 la distribution marginale à l'extérieur des parenthèses. Il reste

$$\lim_{\alpha\to 1} h_\alpha(X|Y) = \frac{-1}{\ln(2)} \int_y p(y) \int_x p_{x|y}(x|y) \ln\left(p_{x|y}(x|y) \right). \tag{8.0.8}$$

Réintroduisant le logarithme en base 2 et la distribution jointe à partir du produit de $p(y)$ et de la distribution marginale, nous obtenons notre résultat final

$$\lim_{\alpha\to 1} h_\alpha(X|Y) = -\int_y \int_x p(x,y) \log\left(p_{x|y}(x|y) \right) = h(X|Y). \tag{8.0.9}$$

Nous notons que d'autres définitions peuvent être utilisées tout en préservant le fait que la limite $\alpha \to 1$ conduit à l'entropie conditionnelle de Rényi habituelle. Cependant, bon nombre de ces définitions alternatives ne satisfont pas des propriétés souhaitables, telles que la monotonie et la règle de chaîne faible [**17**].

Chapitre 9

Propriétés de D_α (Axiomes de Rényi généralisés)

Rényi a proposé à l'origine un ensemble d'axiomes pour l'entropie de Rényi [**6**]. La plupart d'entre eux sont facilement généralisables à notre cas.

9.0.1. Continuité

$D_\alpha(p \parallel q)$ est une fonction continue, à condition que les fonctions p et q soient continues. Cela découle automatiquement de la continuité de la composition de fonctions continues et de l'intégrabilité de telles compositions.

9.0.2. Multiplication par des constantes

À travers quelques manipulations algébriques élémentaires, nous pouvons montrer que :

$$D_\alpha(ap \parallel bq) = D_\alpha(p \parallel q) + \frac{\alpha}{\alpha - 1}\log(a) - \log(b). \tag{9.0.1}$$

9.0.3. Additivité

Si p_1 et q_1 sont tirés selon l'alphabet $\mathcal{X}$ et p_2 et q_2 selon $\mathcal{Y}$, alors :

$$D_\alpha(p_1 p_2 \parallel q_1 q_2) = D_\alpha(p_1 \parallel q_1) + D_\alpha(p_2 \parallel q_2). \tag{9.0.2}$$

Ceci n'est qu'une conséquence directe des propriétés des logarithmes et de la séparabilité des intégrales multiples.

9.0.4. Moyenne générale

Soient p_1 et q_1 tirés selon $\mathcal{X}$ et p_2 et q_2 tirés selon $\mathcal{Y}$. Nous définissons également l'opération $p \oplus q$ comme une union disjointe telle que :

$$p(x) \oplus q(x) = (p;q) = \begin{cases} p(x) & \text{si } x \in \mathcal{X} \\ q(x) & \text{si } x \in \mathcal{Y}. \end{cases}$$

Nous pouvons alors définir la propriété de la moyenne générale. Soit $\lambda \in (0,1)$ et $g_\alpha(x) = 2^{(\alpha-1)x}$. Si p_1 et q_1 sont pondérés avec un poids λ et p_2 et q_2 avec un poids $1-\lambda$, alors :

$$D_\alpha(\lambda p_1 \oplus (1-\lambda)p_2 \parallel \lambda q_1 \oplus (1-\lambda)q_2) = g^{-1}\left[\lambda g(D_\alpha(p_1 \parallel q_1)) + (1-\lambda)g(D_\alpha(p_2 \parallel q_2))\right]. \tag{9.0.3}$$

9.0.5. Inégalité du traitement de l'information

Le théorème habituel d'inégalité de traitement de l'information énonce que, pour une chaîne de Markov $X \to Y \to Z$, aucun traitement de Y ne peut augmenter l'information que Y contient sur X. Cela est souvent formulé comme :

$$I(X;Z) \leq I(X;Y). \tag{9.0.4}$$

Un cas spécifique qui est pertinent pour nous est donné par l'expression suivante :

$$D_\alpha(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq D_\alpha(\lambda p_1 \oplus (1-\lambda)p_2 \parallel \lambda q_1 \oplus (1-\lambda)q_2). \tag{9.0.5}$$

Chapitre 10

Inégalité de l'entropie conditionnelle $\left(h_\alpha(X|Y) \leq h_\alpha(X)\right)$

Nous démontrons cette inégalité en utilisant la divergence de Rényi ainsi que ses propriétés.

Considérons quatre fonctions de densité de probabilité p_1, p_2, q_1, q_2, ainsi que $\lambda \in (0,1)$ et $g(x)$ une fonction strictement monotone croissante et continue. Nous prenons également $\alpha > 1$. Nous pouvons alors constater que

$$g(D_\alpha(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2)) \tag{10.0.1}$$
$$\leq g(D_\alpha(\lambda p_1 \oplus (1-\lambda)p_2 \parallel \lambda q_1 \oplus (1-\lambda)q_2)) \tag{10.0.2}$$
$$= \lambda g(D_\alpha(\lambda p_1 \parallel \lambda q_1)) + (1-\lambda) g(D_\alpha((1-\lambda)p_2 \parallel (1-\lambda)q_2)). \tag{10.0.3}$$

Ici, nous avons appliqué l'inégalité de traitement de l'information de la première à la deuxième ligne et la propriété de la moyenne générale de la deuxième à la troisième. La dernière étape pour montrer notre inégalité est réalisée grâce à de l'algèbre élémentaire :

$$\lambda g(D_\alpha(\lambda p_1 \parallel \lambda q_1)) + (1-\lambda) g(D_\alpha((1-\lambda)p_2 \parallel (1-\lambda)q_2)) \tag{10.0.4}$$
$$\leq \lambda g\left[D_\alpha(\lambda p_1 \parallel \lambda q_1) + \frac{1}{\alpha-1}\log\left(\frac{1}{\lambda}\right)\right] + (1-\lambda) g\left[D_\alpha((1-\lambda)p_2 \parallel (1-\lambda)q_2) + \frac{1}{\alpha-1}\log\left(\frac{1}{\lambda}\right)\right]. \tag{10.0.5}$$

La quantité ajoutée à l'argument de la fonction g sera toujours positive pour $\alpha > 1$ et $\lambda \in (0,1)$. Nous pouvons ainsi établir cette inégalité en nous rappelant que g est une fonction strictement monotone croissante. En développant les divergences dans l'équation précédente et en simplifiant l'expression résultante, nous obtenons :

$$\lambda g\left[D_\alpha(\lambda p_1 \parallel \lambda q_1) + \frac{1}{\alpha-1}\log\left(\frac{1}{\lambda}\right)\right] + (1-\lambda)g\left[D_\alpha((1-\lambda)p_2 \parallel (1-\lambda)q_2) + \frac{1}{\alpha-1}\log\left(\frac{1}{\lambda}\right)\right] \tag{10.0.6}$$

$$= \lambda g\left[D_\alpha(p_1 \parallel p_2)\right] + (1-\lambda)g\left[D_\alpha(p_2 \parallel q_2)\right]. \tag{10.0.7}$$

Ce qui nous permet finalement de déclarer que

$$g(D_\alpha(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2)) \leq \lambda g\left[D_\alpha(p_1 \parallel p_2)\right] + (1-\lambda)g\left[D_\alpha(p_2 \parallel q_2)\right]. \tag{10.0.8}$$

Cette inégalité énonce que la quantité $gD_\alpha(\cdot \parallel \cdot)$ est conjointement convexe.

Pour poursuivre notre argument, nous citons ici un théorème provenant des notes de Michael Wolf [**18**].

Théorème 11. Considérons une fonctionnelle $F : \mathcal{D} \subset \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_n \to \mathbb{R}$ qui est conjointement convexe. Alors F est monotone par rapport à toutes les applications T qui préservent la somme des probabilités, au sens que pour tout $A \in \mathcal{D}$:

$$F(T(A_1),...,T(A_n)) \leq F(A_1,...,A_n) \tag{10.0.9}$$

L'approche habituelle pour montrer l'inégalité d'entropie conditionnelle à partir de ce point serait d'introduire une permutation π dans l'alphabet $\mathcal{Y}$, prendre une valeur d'attente uniforme de la divergence sur toutes les permutations possibles, et appliquer le théorème 11 pour obtenir un chemin clair vers le résultat souhaité. En effet, si p_{XY} et q_{XY} sont des fonctions de masse de probabilités (et donc tirées d'alphabets discrets $\mathcal{X}$ et $\mathcal{Y}$), nous pouvons introduire la fonction de permutation bijective :

$$\pi : \mathcal{Y} \to \mathcal{Y} \tag{10.0.10}$$

$$y_i \mapsto y_j. \tag{10.0.11}$$

Nous choisissons les indices i et j de manière à ce que π soit une bijection. Nous notons que l'application de cette permutation aux deux arguments d'une divergence ne change pas sa valeur car la permutation est essentiellement un changement d'étiquetage des éléments de l'ensemble, et la divergence mesure une distance entre deux distributions sur ces éléments. Nous pouvons alors écrire :

$$g\left[D_\alpha(p_{XY} \parallel q_{XY})\right] = g\left[D_\alpha(\pi(p_{XY}) \parallel \pi(q_{XY}))\right]. \tag{10.0.12}$$

Nous pouvons prendre la valeur d'attente uniforme sur toutes les permutations possibles sans perturber notre égalité.

$$g\left[D_\alpha(\pi(p_{XY}) \parallel \pi(q_{XY}))\right] = E_\pi\left[g\left[D_\alpha(\pi(p_{XY}) \parallel \pi(q_{XY}))\right]\right]. \tag{10.0.13}$$

Nous pouvons appliquer le théorème 11 pour établir que :

$$E_\pi\left[g\left[D_\alpha(\pi(p_{XY}) \parallel \pi(q_{XY}))\right]\right] \geq g\left[D_\alpha(E_\pi[\pi(p_{XY})] \parallel E_\pi[\pi(q_{XY})])\right]. \tag{10.0.14}$$

La valeur d'attente réduira effectivement la distribution sur $\mathcal{Y}$ à une distribution uniforme et détruira toute corrélation qui aurait pu exister entre les distributions. Nous pouvons maintenant appliquer la propriété d'additivité pour obtenir :

$$g\left[D_\alpha(E_\pi[\pi(p_{XY})] \parallel E_\pi[\pi(q_{XY})])\right] = g\left[D_\alpha(p_X \cdot u_Y \parallel q_X \cdot u_Y)\right] = g\left[D_\alpha(p_X \parallel q_X)\right]. \tag{10.0.15}$$

En se souvenant que la fonction g est croissante, on peut déduire que les divergences doivent respecter ce qui suit

$$D_\alpha(p_{XY} \parallel q_{XY}) \geq D_\alpha(p_X \parallel q_X). \tag{10.0.16}$$

Il s'agit en fait d'une expression de l'inégalité de traitement des données sous forme de divergence. Nous pouvons obtenir l'inégalité d'entropie habituelle avec quelques manipulations et en appliquant le résultat nouvellement trouvé :

$$\begin{aligned} h_\alpha(X|Y) &= \log(|\mathcal{X}|) - D_\alpha(p_{XY} \parallel u_X \cdot p_Y) && (10.0.17)\\ &\leq \log(|\mathcal{X}|) - D_\alpha(p_X \parallel u_X) && (10.0.18)\\ &. = h_\alpha(X) && (10.0.19) \end{aligned}$$

Ce qui est le résultat que nous cherchions à obtenir.

Le même argument ne fonctionne pas aussi facilement lorsqu'on considère des ensembles denses (tels que les ensembles de nombres réels) comme alphabets pour nos distributions. Bien que l'idée d'une permutation d'un ensemble dense ait un sens lorsque nous considérons simplement une bijection de l'ensemble sur lui-même, il n'y a aucune garantie que l'expression résultante de cette permutation sera intégrable (que ce soit l'intégrale de Riemann ou de Lebesgue). Autrement dit, toutes les permutations ne donnent pas lieu à une intégrale de Lebesgue sensée. Cela pose de nombreux problèmes lorsqu'il s'agit de choisir quelles permutations conserver pour la valeur d'attente. Nous pourrions considérer la valeur d'attente uniquement sur les bijections mesurables pour éviter ce dilemme, mais nous pouvons montrer que même pour certaines bijections mesurables, l'intégrale serait toujours mal définie.

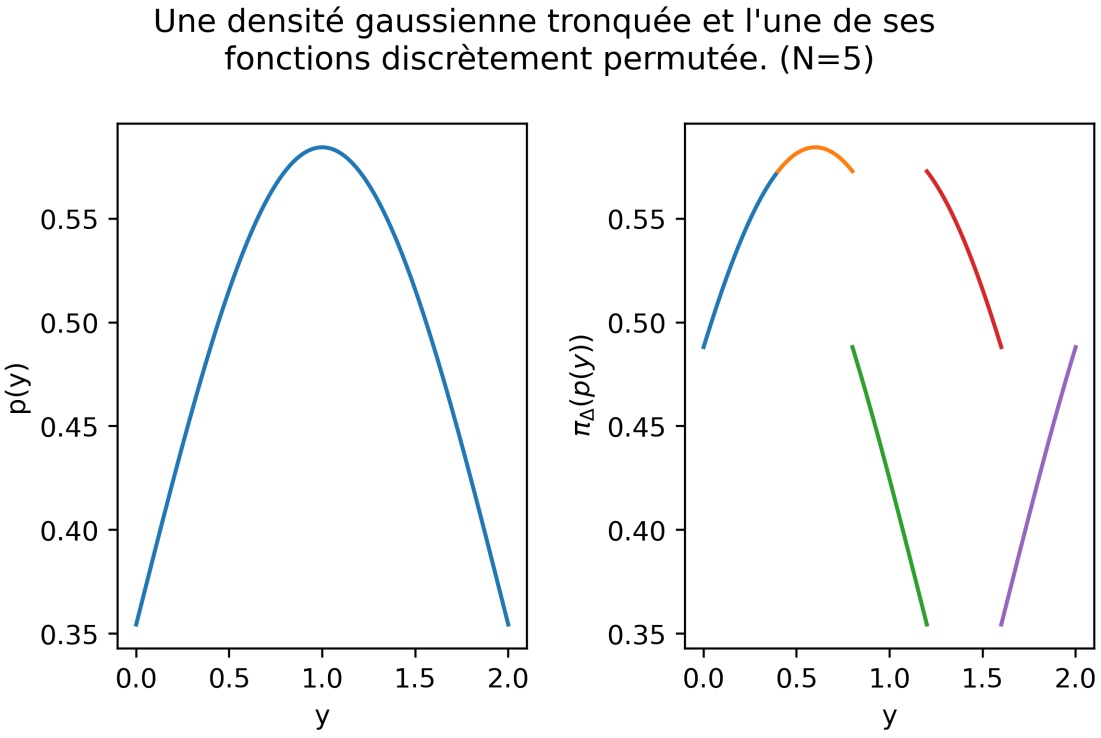

Fig. 10.1. La fonction est effectivement découpée en morceaux et redistribuée sans recouvrement sur le domaine considéré. Elle reste donc une fonction, et son intégrale est égale à l'intégrale de la fonction non permutée.

Nous devons considérer une approche plus analytique du problème. Nous essayons ici d'établir la fonction de permutation π comme la limite de permutations plus discrètes. Considérons une fonction de densité de probabilité intégrable q_Y avec un alphabet $\mathcal{Y}$. Nous divisons le domaine de q_Y en intervalles de longueur égale étiquetés par L_i. La longueur de chaque intervalle est donnée par $||L_i|| = \Delta$ (il en découle logiquement que pour N intervalles et pour un domaine borné, $\Delta = (b-a)/N$, où b et a sont les bornes du domaine considéré avec $b > a$). Les ensembles L_i peuvent être explicitement écrits comme $L_i = [y_{i-1}, y_i)$, où les points y_{i-1} et y_i sont le début et la fin de l'intervalle L_i. À partir de ces définitions, nous pouvons considérer une permutation discrète, disons π^Δ, de la fonction q_Y comme suit :

$$\pi^\Delta : \mathcal{Y} \to \mathcal{Y} \tag{10.0.20}$$

$$q_Y(L_i) \mapsto q_Y(L_j). \tag{10.0.21}$$

Avec i et j choisis de telle sorte que π^Δ soit une bijection. Cette fonction mélange effectivement les intervalles tout en conservant la propriété d'intégrabilité de chaque intervalle individuel. La figure 10.1 montre le résultat de l'application de la fonction π^Δ à une

distribution.

Si q_Y était comparé à une autre distribution p_Y par la divergence de Rényi, le résultat serait le même que π^Δ soit appliqué ou non aux deux distributions (nous ne faisons que re-étiqueter p_Y et q_Y). Autrement dit, nous avons toujours que :

$$D_\alpha(p_Y \parallel q_Y) = D_\alpha(\pi^\Delta(p_Y) \parallel \pi^\Delta(q_Y)). \tag{10.0.22}$$

Nous pouvons alors toujours prendre la valeur d'attente sur toutes les permutations possibles, sans changer notre égalité.

$$\mathbb{E}_{\pi^\Delta}[D_\alpha(\pi^\Delta(p_Y) \parallel \pi^\Delta(q_Y))]. \tag{10.0.23}$$

Nous satisfaisons toujours les hypothèses du théorème 11. Nous pouvons l'appliquer pour obtenir :

$$\mathbb{E}_{\pi^\Delta}[D_\alpha(\pi^\Delta(p_Y) \parallel \pi^\Delta(q_Y))] \geq D_\alpha(\mathbb{E}_{\pi^\Delta}\, \pi^\Delta(p_Y) \parallel \mathbb{E}_{\pi^\Delta}\, \pi^\Delta(q_Y)) \tag{10.0.24}$$

$$= D_\alpha(\tilde{u}_{p_Y} \parallel \tilde{u}_{q_Y}). \tag{10.0.25}$$

Au lieu d'obtenir une distribution strictement uniforme, la discrétisation de π en π^Δ introduit une erreur lors de la prise de la valeur d'espérance sur toutes les permutations. Les distributions résultantes diffèrent d'une distribution uniforme dans une certaine mesure en fonction de l'épaisseur des intervalles. Nous appelons ces distributions $\tilde{u}_{p_Y}$ et $\tilde{u}_{q_Y}$.

Les fonctions $\tilde{u}_{p_Y}$ et $\tilde{u}_{q_Y}$ sont assez intéressantes en elles-mêmes. En raison de la valeur d'attente sur toutes les permutations, ce sont des fonctions de densité de probabilité périodiques de fréquence Δ. Il est donc suffisant d'étudier le comportement de ces fonctions dans un seul intervalle pour comprendre leur ensemble. Il est également plus facile de considérer un seul intervalle, car nous avons maintenant simplement à moyenner tous les segments d'intervalles pour trouver $\tilde{u}_{p_Y}$ dans cet intervalle. Nous pouvons ensuite appliquer la périodicité pour dessiner la totalité de la fonction, ce qui est beaucoup plus rapide que de trouver toutes les permutations possibles et de les moyenner. Pour illustrer cette méthode, un exemple d'une telle fonction est donné dans la figure 10.2. Avec un nombre suffisamment grand d'intervalles, nous pouvons obtenir une distribution très proche d'une distribution uniforme. Le compromis est une fréquence croissante des oscillations autour de la distribution uniforme.

Pour autant que nous sachions, la convergence de π^Δ vers π lorsque $\Delta \to 0$ dépendra de la fonction de densité de probabilité. Par exemple, il est possible de trouver des exemples où

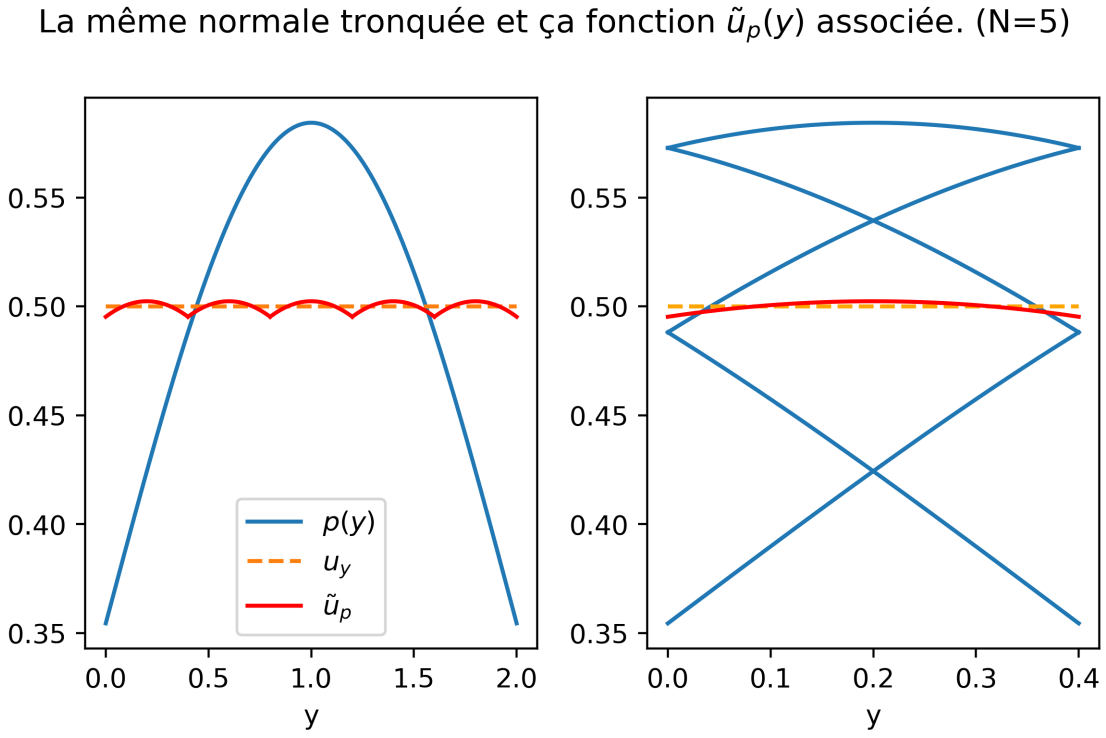

Fig. 10.2. Le graphique de gauche montre la fonction de densité de probabilité, la distribution uniforme sur le domaine de cette fonction, et la fonction $\tilde{u}_p$ qui résulte de la valeur d'espérance sur toutes les permutations π^{Δ}. Le graphique de droite représente une seule période de la fonction $\tilde{u}_p$ avec toutes les courbes que nous moyennons pour la trouver. Les courbes bleues sur le graphique de droite sont les courbes dans chaque intervalle. Puisque l'intégrale de ces courbes bleues donne 1 et que le segment rouge est moyenné avec un facteur $1/N$, l'aire sous la courbe rouge est $1/N$. L'aire sous $\tilde{u}_p$ doit alors être 1.

$D_\alpha(\tilde{u}_{p_Y} \parallel \tilde{u}_{q_Y})$ est plus proche de 0 pour un Δ plus grand que pour certains plus petits. Cependant, si nous choisissons des intervalles suffisamment petits, nous pouvons nous attendre à ce que $D_\alpha(\tilde{u}_{p_Y} \parallel \tilde{u}_{q_Y})$ soit aussi proche de 0 que nous le souhaitons. Mathématiquement, cela signifie que, pour tout choix de ϵ, il existe une largeur d'intervalles Δ telle que:

$$\epsilon > D_\alpha(\tilde{u}_{p_Y} \parallel \tilde{u}_{q_Y}) \geq 0. \tag{10.0.26}$$

Cela étant dit, nous pouvons viser à réduire la divergence entre $\tilde{u}_{p_Y}$ et $\tilde{u}_{q_Y}$ en dessous d'un certain seuil pour un grand nombre d'intervalles, avec l'intention de rendre cette divergence égale à 0 en prenant la limite d'un très grand nombre d'intervalles. Pour ce faire, nous introduisons le théorème suivant :

Théorème 12. Soit p_Y une fonction de densité de probabilité définie sur $[a,b], a < b$. Soit u_Y la distribution uniforme sur $[a,b]$. Si nous divisons le domaine en N intervalles de longueur égale Δ, nous pouvons définir le segment de fonction dans le n-ème intervalle comme étant p_{Yn}. Il en résulte qu'il existe un nombre d'intervalles N_ϵ et $\epsilon > 0$ tels que :

$$\left(\frac{1}{N_\epsilon}\sum_{n=1}^{N_\epsilon}\max(p_{Yn})\right) - u_Y < \epsilon. \tag{10.0.27}$$

PREUVE. Nous procédons par contradiction. Supposons qu'il n'existe pas un tel epsilon. Cela signifie qu'il doit y avoir une certaine valeur pour laquelle la différence entre la moyenne des maximums de chaque intervalle et la distribution uniforme est bornée au-dessus de zéro.

Cependant, si nous prenons la limite $\Delta \to 0$ et nous nous rappelons que $N\Delta = (b-a)$, nous retrouvons une distribution uniforme via une somme de Riemann supérieure :

$$\frac{1}{N}\sum_{n=1}^{N}\max(p_{Yn}) = \frac{\Delta}{(b-a)}\sum_{n=1}^{N}\max(p_{Yn}) \xrightarrow[\Delta\to 0]{} \frac{1}{(b-a)}\int p_Y \,\mathrm{d}y = u_Y. \tag{10.0.28}$$

Ainsi, dans le cas limite $\Delta \to 0$, la différence entre les deux termes est égale à 0. Notre hypothèse de départ doit donc être fausse, et il doit exister une valeur pour $N_\epsilon, \Delta_\epsilon$, avec $N_\epsilon\Delta_\epsilon = (b-a)$, telle que pour tout $\epsilon > 0$, nous avons :

$$\left(\frac{1}{N_\epsilon}\sum_{n=1}^{N_\epsilon}\max(p_{Yn})\right) - u_Y < \epsilon. \tag{10.0.29}$$

□

Nous notons que nous pouvons montrer un théorème similaire où nous prenons, mutatis mutandis, la moyenne des minimums dans les intervalles.

Nous pouvons maintenant procéder à la borne de la divergence entre $\tilde{u}p_Y$ et $\tilde{u}q_Y$. Nous rappelons que

$$D_\alpha(\tilde{u}_{p_Y} \parallel \tilde{u}_{q_Y}) = \frac{1}{\alpha-1}\log\left(\int_a^b \frac{(\tilde{u}_{p_Y})^\alpha}{(\tilde{u}_{q_Y})^{\alpha-1}}\,\mathrm{d}y\right). \tag{10.0.30}$$

Nous pouvons commencer notre recherche de borne en utilisant les opérations suivantes :

$$D_\alpha(\tilde{u}_{p_Y} \parallel \tilde{u}_{q_Y}) = \frac{1}{\alpha-1}\log\left(\int_a^b \frac{(\tilde{u}_{p_Y})^\alpha}{(\tilde{u}_{q_Y})^{\alpha-1}}\,\mathrm{d}y\right) \leq \log\left(\frac{(\max(\tilde{u}_{p_Y}))^\alpha}{(\min(\tilde{u}_{q_Y}))^{\alpha-1}}\int_a^b \mathrm{d}y\right) \tag{10.0.31}$$

$$= \log\left((b-a)\frac{(\max(\tilde{u}_{p_Y}))^\alpha}{(\min(\tilde{u}_{q_Y}))^{\alpha-1}}\right) \tag{10.0.32}$$

$$\leq \log\left((b-a)\frac{\left(\frac{1}{N}\sum_{n=1}^{N}\max(p_{Yn})\right)^\alpha}{\left(\frac{1}{N}\sum_{n=1}^{N}\min(q_{Yn})\right)^{\alpha-1}}\right). \tag{10.0.33}$$

La dernière inégalité nécessite l'observation que pour une fonction $\tilde{u}_{p_Y}$, le maximum de cette fonction est nécessairement inférieur ou égal à la moyenne de tous les maximums des segments de courbe sur lesquels nous effectuons la moyenne pour obtenir $\tilde{u}_{p_Y}$. De même, le minimum de $\tilde{u}_{p_Y}$ est supérieur ou égal à la moyenne de tous les minimums des segments de courbe.

À ce stade, nous pouvons appliquer le théorème 12 au numérateur et au dénominateur. Il suffit de choisir un N suffisamment grand pour que :

$$\log\left((b-a)\frac{\left(\frac{1}{N}\sum_{n=1}^{N}\max(p_{Yn})\right)^{\alpha}}{\left(\frac{1}{N}\sum_{n=1}^{N}\min(q_{Yn})\right)^{\alpha-1}}\right) \leq \log\left((b-a)\frac{(\epsilon_1+u_Y)^{\alpha}}{(-\epsilon_1+u_Y)^{\alpha-1}}\right). \tag{10.0.34}$$

Nous pouvons nous assurer que N est également suffisamment grand pour que le rapport dans le logarithme soit inférieur à un autre ϵ. Autrement dit :

$$\left|\frac{(\epsilon_1+u_Y)^{\alpha}}{(-\epsilon_1+u_Y)^{\alpha-1}} - u_Y\right| < \epsilon_2. \tag{10.0.35}$$

À partir de cela, nous pouvons dire que

$$\log\left((b-a)\frac{(\epsilon_1+u_Y)^{\alpha}}{(-\epsilon_1+u_Y)^{\alpha-1}}\right) \leq \log\left((b-a)(\epsilon_2+u_Y)\right) \tag{10.0.36}$$

$$= \log(\epsilon_2(b-a)+1) \tag{10.0.37}$$

$$\leq \epsilon. \tag{10.0.38}$$

Et donc, pour des intervalles suffisamment petits :

$$D_{\alpha}(\tilde{u}_{p_Y} \parallel \tilde{u}_{q_Y}) \leq \epsilon. \tag{10.0.39}$$

À l'étape suivante, nous devons prendre trois limites : les limites lorsque $a \to -\infty$, $b \to +\infty$ et $\epsilon_2 \to 0$. La dernière limite est finalement équivalente à prendre $\Delta \to 0$. En assurant que ϵ tend vers 0 plus rapidement que $b-a$ ne peut tendre vers l'infini, nous obtenons notre résultat :

$$\lim_{a\to-\infty}\lim_{b\to+\infty}\lim_{\epsilon_2\to 0}\log(\epsilon_2(b-a)+1) = \log(1) = 0. \tag{10.0.40}$$

En appliquant tout cet argument aux fonctions de densité de probabilité jointes telles que celles considérées pour le cas discret, nous constatons que :

$$g\left[D_\alpha(\pi^\Delta(p_{XY}) \parallel \pi^\Delta(q_{XY}))\right] = \mathbb{E}_{\pi^\Delta}\left[g\left[D_\alpha(\pi^\Delta(p_{XY}) \parallel \pi^\Delta(q_{XY}))\right]\right] \tag{10.0.41}$$

$$\geq g\left[D_\alpha(p_X \parallel q_X) + \epsilon)\right] \tag{10.0.42}$$

$$\geq g\left[D_\alpha(p_X \parallel q_X)\right]. \tag{10.0.43}$$

Et ainsi, pour des intervalles suffisamment petits :

$$D_\alpha(p_{XY} \parallel q_{XY}) \geq D_\alpha(p_X \parallel q_X). \tag{10.0.44}$$

Bien que nous ayons montré l'inégalité d'entropie conditionnelle, il existe d'autres résultats qui sont intéressants en ce qui concerne $g\left[D_\alpha(\cdot \parallel \cdot)\right]$. Nous pouvons, par un développement similaire à celui utilisé pour montrer la convexité jointe, montrer que $g\left[D_\alpha(\cdot \parallel \cdot)\right]$ est conjointement concave pour $\alpha \in (0,1)$. Pour un tel choix de α, nous avons :

$$g(D_\alpha(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2)) \geq \lambda g\left[D_\alpha(p_1 \parallel p_2)\right] + (1-\lambda) g\left[D_\alpha(p_2 \parallel q_2)\right]. \tag{10.0.45}$$

Nous pouvons aller un peu plus loin avec la dernière équation en écrivant les termes de la fonction g et en rappelant que $(\alpha - 1)$ est négatif :

$$2^{(\alpha-1)D_\alpha(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2)} \geq \lambda 2^{(\alpha-1)D_\alpha(p_1\|q_1)} + (1-\lambda)2^{(\alpha-1)D_\alpha(p_2\|q_2)} \tag{10.0.46}$$

$$\Rightarrow D_\alpha(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \frac{1}{\alpha - 1}\log_2(\lambda 2^{(\alpha-1)D_\alpha(p_1\|q_1)} + (1-\lambda)2^{(\alpha-1)D_\alpha(p_2\|q_2)}). \tag{10.0.47}$$

Nous pouvons ensuite appliquer l'inégalité de Jensen en définissant une fonction $\varphi_\alpha(x)$ comme suit :

$$\varphi_\alpha(x) = \frac{1}{\alpha - 1}\log_2(x). \tag{10.0.48}$$

Cette fonction est convexe pour tout x et $\alpha \in (0,1)$. En appliquant l'inégalité de Jensen, on obtient :

$$\frac{1}{\alpha - 1}\log_2(\lambda 2^{(\alpha-1)D_\alpha(p_1\|q_1)} + (1-\lambda)2^{(\alpha-1)D_\alpha(p_2\|q_2)}) \tag{10.0.49}$$

$$\leq \frac{\lambda}{\alpha - 1}\log_2(2^{(\alpha-1)D_\alpha(p_1\|q_1)}) + \frac{(1-\lambda)}{\alpha - 1}\log_2(2^{(\alpha-1)D_\alpha(p_2\|q_2)}) \tag{10.0.50}$$

$$= \lambda D_\alpha(p_1 \parallel q_1) + (1-\lambda)D_\alpha(p_2 \tag{10.0.51}$$

Ce qui nous permet finalement d'écrire :

$$D_\alpha(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \lambda D_\alpha(p_1 \parallel q_1) + (1-\lambda)D_\alpha(p_2 \parallel q_2). \quad (10.0.52)$$

Nous pouvons donc conclure que $D_\alpha(\cdot \parallel \cdot)$ est jointement convexe pour $0 < \alpha < 1$.

Cependant, la convexité jointe de D_α est perdue pour $\alpha > 1$. Certaines propriétés sont récupérables pour $\alpha > 1$, comme la quasi-convexité et la convexité dans un argument de D_α [**7**].

Chapitre 11

Inégalité entre entropie de Rényi et de Shannon ($h(X) < h_\alpha(X) + [...]$)

Il est d'un intérêt significatif d'essayer de borner l'entropie différentielle avec son entropie de Rényi. Une part significative des preuves de sécurité de distribution quantique de clefs de chiffrements repose sur des inégalités entre l'entropie de Rényi et l'entropie de Shannon. Plus précisément, ces inégalités bornent l'entropie de Shannon par des fonctions qui dépendent de l'entropie de Rényi et, plus particulièrement, de la cardinalité de l'espace des possibilités de la variable aléatoire considérée. Dans le cas où l'espace des possibilités est discret et fini, la cardinalité et l'entropie de Rényi offrent une borne raisonnable et utilisable pour des preuves de sécurité.

Avec les avancées en informatique, notamment en informatique quantique, le domaine des communications prend lui aussi son envol dans l'ère quantique. Pour pleinement extraire les bénéfices de la théorie quantique, le bit classique laisse place à des paquets d'ondes et à des superpositions d'états. Avec cette transition, plusieurs espaces de probabilités passent au continu. La taille des ensembles des valeurs possibles pour une variable devient infinie et la grande majorité des inégalités qui nous étaient utiles classiquement perdent leur valeur.

C'est de là que vient la demande pour de nouvelles bornes, soit des inégalités entre l'entropie différentielle de Rényi et de Shannon qui ne dépendent pas de la cardinalité de l'espace de la variable considérée. Préférablement, ces bornes dépenderaient des propriétés intrinsèques de la distribution de la variable aléatoire considérée, tels que les différents moments de la distribution en question.

Nous explorons différentes approches qui ont été entreprises dans le passé et essayons de les transposer au cadre continu. Une telle approche est donnée dans le livre de Tomamichel

(*Quantum Information Processing with Finite Resources*) [**19**].

L'équation 4.102 du livre stipule que pour tout intervalle [a,b] contenant 1 et avec $\alpha \in [a,b]$, il existe une constante $K \in [0, +\infty]$ (assurée par le théorème de Taylor) telle que :

$$\left| \overline{D}_\alpha(\rho \parallel \sigma) - D_\alpha(\rho \parallel \sigma) - \frac{(\alpha-1)}{2\log(e)} V(\rho \parallel \sigma) \right| \leq K(\alpha-1)^2. \tag{11.0.1}$$

Avec V et $\overline{D}$ donnés par:

$$V(\rho \parallel \sigma) = \mathrm{Tr}\left(\rho(\log \rho - \log \sigma - D(\rho \parallel \sigma))^2\right). \tag{11.0.2}$$

$$\overline{D}_\alpha(\rho \parallel \sigma) = \begin{cases} \frac{1}{\alpha-1}\log\left(\frac{\mathrm{Tr}(\rho^\alpha \sigma^{1-\alpha})}{\mathrm{Tr}\rho}\right) & \text{si } (\alpha < 1 \wedge \rho \not\perp \sigma) \vee \rho \subseteq \sigma \\ +\infty & \text{sinon} \end{cases}$$

Nous pouvons facilement traduire cette expression de la mécanique quantique en termes classiques. En remplaçant les matrices de densité par les fonctions de densité de probabilités et les traces par des intégrales sur ces densités de probabilités, nous pouvons manipuler l'expression ci-dessus pour obtenir l'inégalité suivante :

$$-K(\alpha-1)^2 + D(p \parallel q) + \frac{(\alpha-1)}{2\log(e)}\left[\int_x p(\log(p) - \log(q) - D(p \parallel q))^2 \,\mathrm{d}x\right] \leq D_\alpha(p \parallel q). \tag{11.0.3}$$

En développant le terme au carré à l'intérieur de l'intégrale et en déplaçant les termes, on obtient ce qui suit :

$$D(p \parallel q) + \frac{\alpha-1}{2\log(e)}\left[\int_x pD^2(p \parallel q) + 2p\log(q)D(p \parallel q) - 2p\log(p)D(p \parallel q)\,\mathrm{d}x\right] \tag{11.0.4}$$

$$\leq D_\alpha(p \parallel q) + K(\alpha-1)^2 - \frac{\alpha-1}{2\log(e)}\left[\int_x p(\log^2(p) + \log^2(q) - 2\log(p)\log(q))\,\mathrm{d}x\right]. \tag{11.0.5}$$

L'intégrale à gauche de l'inégalité peut être simplifiée :

$$\frac{\alpha-1}{2\log(e)}\left[\int_x pD^2(p\parallel q)+2p\log(q)D(p\parallel q)-2p\log(p)D(p\parallel q)\,\mathrm{d}x\right] \tag{11.0.6}$$

$$=\frac{\alpha-1}{2\log(e)}\left[D(p\parallel q)\int_x pD(p\parallel q)+2p\log(q)-2p\log(p)\,\mathrm{d}x\right] \tag{11.0.7}$$

$$=\frac{\alpha-1}{2\log(e)}\left[D(p\parallel q)\int_x pD(p\parallel q)-2p\log\left(\frac{p}{q}\right)\,\mathrm{d}x\right] \tag{11.0.8}$$

$$=\frac{\alpha-1}{2\log(e)}\left[D(p\parallel q)\left(D(p\parallel q)\int_x p\,\mathrm{d}x-2D(p\parallel q)\right)\right] \tag{11.0.9}$$

$$=\frac{\alpha-1}{2\log(e)}\left[D^2(p\parallel q)-2D^2(p\parallel q)\right] \tag{11.0.10}$$

$$=\frac{\alpha-1}{2\log(e)}\left[-D^2(p\parallel q)\right]. \tag{11.0.11}$$

La partie gauche de notre inégalité peut ensuite être écrite comme suit :

$$\frac{1-\alpha}{2\log(e)}D^2(p\parallel q)+D(p\parallel q). \tag{11.0.12}$$

L'intégrale du côté droit peut être écrite comme suit :

$$-\frac{\alpha-1}{2\log(e)}\left[\int_x p(\log(p)-\log(q))^2\,\mathrm{d}x\right] \tag{11.0.13}$$

$$=-\frac{\alpha-1}{2\log(e)}\left[\int_x p\log^2\left(\frac{p}{q}\right)\,\mathrm{d}x\right] \tag{11.0.14}$$

$$=-\frac{\alpha-1}{2\log(e)}E_p\left[\log^2\left(\frac{p}{q}\right)\right]. \tag{11.0.15}$$

Ce qui rend le côté droit égal à :

$$D_\alpha(p\parallel q)+K(\alpha-1)^2+\frac{1-\alpha}{2\log(e)}E_p\left[\log^2\left(\frac{p}{q}\right)\right]. \tag{11.0.16}$$

L'inégalité est ainsi réduite à :

$$\frac{1-\alpha}{2\log(e)}D^2(p\parallel q)+D(p\parallel q)\leq D_\alpha(p\parallel q)+K(\alpha-1)^2+\frac{1-\alpha}{2\log(e)}E_p\left[\log^2\left(\frac{p}{q}\right)\right]. \tag{11.0.17}$$

À ce stade, il est intéressant de définir p et q de manière plus précise. Nous considérons d'abord le cas où les deux fonctions de densité de probabilités sont données par des distributions normales non centrales. Soit p une distribution normale avec une moyenne μ_1 et une variance σ_1^2, et q une distribution normale avec une moyenne μ_2 et une variance σ_2^2. La distribution du ratio des deux distributions est donnée par :

$$\frac{p(x)}{q(x)} = \frac{\sigma_2}{\sigma_1} \exp\left\{\frac{1}{2}\left[-\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]\right\}. \tag{11.0.18}$$

En appliquant $\log^2$ à notre expression :

$$\log^2\left(\frac{p(x)}{q(x)}\right) = \frac{1}{\ln^2(2)} \ln^2\left(\frac{p(x)}{q(x)}\right) \tag{11.0.19}$$

$$= \frac{1}{\ln^2(2)}\left[\ln\left(\left(\frac{\sigma_2}{\sigma_1}\right)\exp\left\{\frac{1}{2}\left[-\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]\right\}\right)\right]^2 \tag{11.0.20}$$

$$= \frac{1}{\ln^2(2)}\left[\ln\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left(\left(\frac{x-\mu_2}{\sigma_2}\right)^2 - \left(\frac{x-\mu_1}{\sigma_1}\right)^2\right)\right]^2 \tag{11.0.21}$$

$$= \frac{1}{\ln^2(2)}\left[\ln\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left(\frac{(x^2 - 2x\mu_2 + \mu_2^2)}{\sigma_2^2} - \frac{(x^2 - 2x\mu_1 + \mu_1^2)}{\sigma_1^2}\right)\right]^2 \tag{11.0.22}$$

$$= \frac{1}{\ln^2(2)}\left[\ln\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left(\frac{(x^2 - 2x\mu_2 + \mu_2^2)}{\sigma_2^2} - \frac{(x^2 - 2x\mu_1 + \mu_1^2)}{\sigma_1^2}\right)\right]^2 \tag{11.0.23}$$

$$= \frac{1}{\ln^2(2)}\left[\ln\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2\sigma_1^2\sigma_2^2}(x^2(\sigma_1^2 - \sigma_2^2) - 2x(\sigma_1^2\mu_2 - \sigma_2^2\mu_1) + (\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2))\right]^2. \tag{11.0.24}$$

Pour plus de lisibilité, nous définissons la variable Z comme suit :

$$Z(x) = x^2(\sigma_1^2 - \sigma_2^2) - 2x(\sigma_1^2\mu_2 - \sigma_2^2\mu_1) + (\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2). \tag{11.0.25}$$

Nous pouvons développer le carré et distribuer pour obtenir une expression plus concise :

$$\log^2\left(\frac{p(x)}{q(x)}\right) = \log^2\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\log\left(\frac{\sigma_2}{\sigma_1}\right)Z}{\ln(2)\sigma_1^2\sigma_2^2} + \frac{Z^2}{4ln^2(2)\sigma_1^4\sigma_2^4}. \tag{11.0.26}$$

Enfin, nous pouvons prendre la valeur d'espérance de la dernière équation pour obtenir le terme que nous avions initialement dans nos développements :

$$\mathbb{E}_p\left[\log^2\left(\frac{p(x)}{q(x)}\right)\right] = \mathbb{E}_p\left[\log^2\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\log\left(\frac{\sigma_2}{\sigma_1}\right)Z}{\ln(2)\sigma_1^2\sigma_2^2} + \frac{Z^2}{4ln^2(2)\sigma_1^4\sigma_2^4}\right] \tag{11.0.27}$$

$$= \log^2\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\log\left(\frac{\sigma_2}{\sigma_1}\right)}{\ln(2)\sigma_1^2\sigma_2^2}E_p[Z] + \frac{1}{4ln^2(2)\sigma_1^4\sigma_2^4}E_p[Z^2]. \tag{11.0.28}$$

Notre inégalité sur la divergence est donc :

$$\frac{1-\alpha}{2\log(e)}D^2(p \parallel q) + D(p \parallel q) \leq D_\alpha(p \parallel q) + K(\alpha-1)^2 + \frac{1-\alpha}{2\log(e)}\log^2\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\log\left(\frac{\sigma_2}{\sigma_1}\right)}{\ln(2)\sigma_1^2\sigma_2^2}E_p[Z] + \frac{1}{4ln^2(2)\sigma_1^4} \tag{11.0.29}$$

Les valeurs d'espérance de Z et Z^2 sont ultimement des fonctions des différents moments de $p(x)$. Nous avons, explicitement :

$$E_p[Z] = (\sigma_1^2-\sigma_2^2)E_p[x^2] - 2(\sigma_1^2\mu_2 - \sigma_2^2\mu_1)E_p[x] + (\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2) \tag{11.0.30}$$

$$E_p[Z^2] = (\sigma_1^2-\sigma_2^2)^2E_p[x^4] - 4(\sigma_1^2-\sigma_2^2)(\sigma_1^2\mu_2 - \sigma_2^2\mu_1)E_p[x^3] + 2(\sigma_1^2-\sigma_2^2)(\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2)E_p[x^2] \tag{11.0.31}$$

$$+ 4(\sigma_1^2\mu_2 - \sigma_2^2\mu_1)^2E_p[x^2] - 4(\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2)E_p[x] + (\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2)^2. \tag{11.0.32}$$

Si p et q sont des distributions centrées, alors $\mu_1 = \mu_2 = E_p[x] = 0$, $E_p[x^2] = Var_p(x)$ et les expressions sont significativement simplifiées :

$$E_p[Z] = (\sigma_1^2-\sigma_2^2)\mathrm{Var}_p(x) \tag{11.0.33}$$

$$E_p[Z^2] = (\sigma_1^2-\sigma_2^2)E_p[x^4]. \tag{11.0.34}$$

Nous pouvons également tenter de définir p et q comme des distributions uniformes. Soit p prenant des valeurs dans $[a,b]$ et q prenant des valeurs dans $[c,d]$, où $[a,b] \subseteq [c,d]$. La valeur d'espérance de $\log^2$ de la distribution de rapport est simplement :

$$E_p\left[\log^2\left(\frac{p(x)}{q(x)}\right)\right] = \log^2(d-c) - 2\log(d-c)\log(b-a) + \log^2(b-a) = \log^2\left(\frac{d-c}{b-a}\right). \tag{11.0.35}$$

Et l'inégalité est réduite à une forme beaucoup plus simple :

$$\frac{1-\alpha}{2\log(e)}D^2(p \parallel q) + D(p \parallel q) \leq D_\alpha(p \parallel q) + K(\alpha-1)^2 + \frac{1-\alpha}{2\log(e)}\log^2\left(\frac{d-c}{b-a}\right). \tag{11.0.36}$$

Ces inégalités nous donnent, ultimement, des bornes sur la divergence de Kullback-Leibler qui dépendent sur la divergence de Rényi (et donc sur α).

Chapitre 12

Conclusion

L'objectif initial de ce travail était de trouver et de bien définir plusieurs des quantités et inégalités en théorie de l'information dans le cas où nous travaillons avec Rényi et dans le cas continu. Le travail requis pour bien définir les quantités continues de Rényi est loin d'être trivial, mais comme nous avons pu le démontrer ici, il est possible de les trouver morceau par morceau. De la construction d'une théorie des probabilités sur les variables discrètes et continues à l'établissement de quantités continues de Rényi, il y a un grand nombre de problèmes à prendre en considération. Nous souhaitons cependant que les développements de ce mémoire offrent une base pour tout autre travail désirant utiliser la théorie de l'information continue dans ses développements.

Les développements donnés dans ce document sont une petite partie de ce qui pourait être un travail beaucoup plus extensif et profond. Il reste plusieurs inégalités et notions possibles à généraliser au cas de Rényi continu. Nous n'avons toujours pas touché aux analogues quantiques de ces quantités continues; il y a un grand nombre de possibilités de développements à partir de ceux donnés ici. Même dans le cas classique, il reste plusieurs inégalités qui pourraient être adaptées, optimisées, ou renforcées par un passage au continu.

Références bibliographiques

[1] G. Markowsky, "Information theory." `https://www.britannica.com/science/information-theory`, mars 2023.

[2] H. Nyquist, "Certain factors affecting telegraph speed," *Bell System Technical Journal*, 1924.

[3] R. Hartley, "Transmission of information," *Bell System Technical Journal*, 1928.

[4] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, 1948.

[5] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 2006.

[6] A. Rényi, "On measures of information and entropy," *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics, and Probability*, 1961.

[7] T. van Erven and P. Harremoes, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, pp. 3797–3820, jul 2014.

[8] R. Islam, R. Ma, P. M. Preiss, M. E. Tai, A. Lukin, M. Rispoli, and M. Greiner, "Measuring entanglement entropy in a quantum many-body system," *Nature*, vol. 528, pp. 77–83, December 2015.

[9] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," *Theoretical Computer Science*, vol. 560, pp. 7–11, dec 2014.

[10] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM Journal on Computing*, vol. 26, pp. 1484–1509, oct 1997.

[11] C. Weedbrook, S. Pirandola, R. García-Patrón, N. J. Cerf, T. C. Ralph, J. H. Shapiro, and S. Lloyd, "Gaussian quantum information," *Reviews of Modern Physics*, vol. 84, pp. 621–669, may 2012.

[12] M. Tomamichel, R. Colbeck, and R. Renner, "A fully quantum asymptotic equipartition property," *IEEE Transactions on Information Theory*, vol. 55, pp. 5840–5847, dec 2009.

[13] S. Ghorai, P. Grangier, E. Diamanti, and A. Leverrier, "Asymptotic security of continuous-variable quantum key distribution with a discrete modulation," *Phys. Rev. X*, vol. 9, p. 021059, Jun 2019.

[14] M. Tomamichel and M. Hayashi, "A hierarchy of information quantities for finite block length analysis of quantum tasks," *IEEE Transactions on Information Theory*, vol. 59, pp. 7693–7710, nov 2013.

[15] A. Leverrier, "Security of continuous-variable quantum key distribution via a gaussian de finetti reduction," *Phys. Rev. Lett.*, vol. 118, p. 200501, May 2017.

[16] A. Klenke, *Probability Theory: A Comprehensive Course*. Springer, 2007.

[17] S. Fehr and S. Berens, "On the conditional rényi entropy," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 6801–6810, 2014.

[18] M. M. Wolf, "Quantum channels and operations - guided tour." `https://mediatum.ub.tum.de/node?id=1701036`, 2012. Graue Literatur.

[19] M. Tomamichel, *Quantum Information Processing with Finite Resources*. Springer International Publishing, 2016.