Entropies et critères entropiques

Jean-François Bercher

Chapitre 1

Entropies et critères entropiques

par J.-F. Bercher

LIGM, UMR 8049, ESIEE/Université Paris-Est.

1.1. Référence

Chapitre rédigé par J.-F. Bercher, LIGM, UMR 8049, ESIEE/Université Paris-Est
Extrait de l'ouvrage « Méthodes d'inversion appliquées au traitement du signal et de l'image», ISBN 2746245485, J.-F. Giovannelli et J. Idier éditeurs, publié chez Hermès-Lavoisier, décembre 2013.

1.2. Introduction

Ce chapitre est centré sur les notions d'entropies et de lois à maximum d'entropie qui seront caractérisées selon plusieurs angles. Au delà des liens avec les applications en ingénierie puis en physique, on montrera qu'il est possible de bâtir des fonctionnelles régularisantes fondées sur l'emploi d'une technique à maximum d'entropie, qui peuvent alors éventuellement être utilisées comme potentiels *ad hoc* dans des problèmes d'inversion de données.

Le chapitre débute par un tour d'horizon des principales propriétés des mesures d'information, et par l'introduction de différentes notions et définitions. En particulier, on définit la divergence de Rényi, on présente la notion de distribution escorte, et on commente le principe du maximum d'entropie qui sera utilisé par la suite. On présente ensuite un problème classique d'ingénierie, le problème du codage de source,

et on montre l'intérêt d'utiliser des mesures de longueur différentes de la mesure standard, et en particulier une mesure exponentielle, qui conduit à un théorème de codage de source dont la borne minimale est une entropie de Rényi. On montre également que les codes optimaux peuvent être calculés aisément grâce aux distributions escortes. En section 1.5, on introduit et on étudie un modèle simple de transition d'état. Ce modèle conduit à une distribution d'équilibre définie comme une distribution escorte généralisée, et conduit en sous-produit, à nouveau à une entropie de Rényi. On étudie le flux d'information de Fisher le long de la courbe définie par la distribution escorte généralisée, et on obtient des connections avec la divergence de Jeffreys. Finalement, on obtient différents arguments qui, dans ce cadre, conduisent à une méthode d'inférence fondée sur la minimisation de l'entropie de Rényi sous une contrainte de moyenne généralisée, *i.e.* prise vis-à-vis de la distribution escorte. À partir de la section 1.6.3, on s'intéresse alors à la minimisation de la divergence de Rényi sous une contrainte de moyenne généralisée. On donne et on caractérise la densité optimale qui résout ce problème, et la valeur de la divergence optimale correspondante. On définit, et on caractérise les principales propriétés des entropies qui peuvent y être associées. Enfin, on montre comment calculer pratiquement ces entropies et comment on peut envisager de les utiliser pour la résolution de problèmes linéaires.

1.3. Quelques entropies en théorie de l'information

Le concept d'information joue un rôle majeur dans nombre de champs scientifiques et techniques et dans leurs applications. Qui plus est, la théorie de l'information à la Shannon rejoint les théories de la physique, les unes et les autres se fertilisant mutuellement ; ces interactions ont été exploitées par Jaynes [JAY 57a, JAY 57b] dès 1957, sont discutées par exemple par Brillouin [BRI 62] et plus récemment dans le fascinant travail [MER 10]. Nous donnerons par la suite un modèle simple de transition de phase qui conduit à une entropie de Rényi.

Une question fondamentale en théorie de l'information est bien entendu la mesure, ou la définition, de l'information. Plusieurs approches sont possibles. La première est pragmatique et accepte comme mesure valable d'une information les mesures qui apparaissent d'elles même dans la résolution d'un problème pratique. La seconde est axiomatique, où l'on débute avec un certain nombre de propriétés ou de postulats raisonnables, et où il s'agit ensuite d'une dérivation mathématique des fonctions possédant ces propriétés. C'est le point de vue adopté à l'origine par Shannon, dans son article fondamental [SHA 48a, SHA 48b], et qui a conduit à nombre de développements ultérieurs, parmi lesquels on citera [ACZ 75] et [ACZ 84] (ou l'auteur met en garde contre l'excès de généralisations : *«I wish to urge here caution with regard to generalizations in general, and in particular with regard to those introduced through characterizations. (...) There is a large number of "entropies" and other "information measures" and their "characterizations", mostly formal generalizations of (1),*

(19), (16), (24), (17), (23) etc. popping up almost daily in the literature. It may be reassuring to know that most are and will in all probability be completely useless. »

De même, Rényi lui même [R´ 65][CSI 06] soulignait que ne devraient être considérées comme mesures d'information que les quantités qui peuvent effectivement être utilisées dans des problèmes concrets, rejoignant en cela l'approche pragmatique (*As a matter of fact, if certain quantities are deduced from some natural postulates (from "first principles") these certainly need for their final justification the control whether they can be effectively used in solving concrete problems*).

1.3.1. *Principales propriétés et définitions*

Nous rappellerons toutefois ici les principales propriétés utilisées pour les caractérisations de mesures d'information. Si P, Q, R désignent des distributions de probabilité discrètes portant sur n événements, et où p_k est la probabilité associée au k$^{\text{e}}$ événement, $k = 1..n$, alors, en notant $H(P) = H(p_1, p_2, ..., p_n)$ la mesure de l'information associée aux événements de distribution P, les principales propriétés sont les suivantes :

1) symétrie : $H(p_1, p_2, ..., p_n)$ ne dépend pas de l'ordre des événements,

2) $H(p, 1-p)$ est une fonction continue de p,

3) $H(\frac{1}{2}, \frac{1}{2}) = 1$,

4) récursivité (branchement) : $H_{n+1}(p_1q_1, p_1q_2, p_2, \ldots, p_n) = H_n(p_1, p_2, \ldots, p_n) + p_1 H_2(q_1, q_2)$,

5) expansible : $H_{n+1}(p_1, p_2, \ldots, p_n, 0) = H_n(p_1, p_2, \ldots, p_n)$,

6) sous-additivité : $H(PQ) \leq H(P) + H(Q)$ (et additivité dans le cas indépendant : $H(PQ) = H(P) + H(Q)$),

7) sous-additivité conditionnelle : $H(PQ|R) \leq H(P|R) + H(Q|R)$,

8) récursivité généralisée : $H_{n+1}(p_1q_1, p_1q_2, p_2, \ldots, p_n) = H_n(p_1, p_2, \ldots, p_n) + m(p_1)H_2(q_1, q_2)$.

Conséquences simples

Les quatre premiers postulats sont les axiomes de Faddeev [FAD 56], qui suffisent à caractériser uniquement l'entropie de Shannon

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i. \tag{1.1}$$

Si on lève le postulat de récursivité mais qu'on ajoute une exigence d'additivité, alors la classe des solutions possibles est bien plus large, et inclut en particulier l'entropie de Rényi, dont on reparlera plus loin. Le remplacement de la récursivité 4 par un postulat

de récursivité générale, 8, avec $m(p_1p_2)$ multiplicatif $m(p_1p_2) = m(p_1)m(p_2)$ et en particulier $m(p) = p^q$ conduit à l'entropie d'ordre q

$$H_q(P) = \frac{1}{2^{1-q} - 1} \left(\sum_{i=1}^{n} p_i^q - 1 \right), \tag{1.2}$$

qui a été introduite par [HAV 67], indépendamment par Daróczy [DAR 70], puis redécouverte dans le champ de la physique statistique par C. Tsallis [TSA 88]. Pour $q \geq 1$, ces entropies sont sous-additives, mais ne sont pas additives. Dans le cas $q = 1$, par la règle de l'Hôpital, l'entropie d'ordre $q = 1$ n'est autre que l'entropie de Shannon. En physique statistique, une communauté importante s'est formée autour de l'étude de la thermodynamique nonextensive (non-additive en fait) [TSA 09] reposant sur l'emploi de l'entropie de Tsallis, des distributions à maximum d'entropie associées et de l'extension de la thermodynamique classique.

Dans l'axiomatique de Faddeev, Rényi [R´ 61] a proposé de remplacer le postulat de récursivité par la propriété d'additivité, et d'ajouter une propriété de moyenne de l'entropie, qui spécifie que l'entropie de l'union de deux distributions de probabilité incomplètes est égale à la moyenne pondérée des entropies des deux distributions. Lorsque la moyenne utilisée est une moyenne arithmétique, la seule solution est l'entropie de Shannon. Par contre en utilisant une moyenne exponentielle, l'entropie qui apparaît est une entropie de Rényi

$$H_q(P) = \frac{1}{1-q} \log \sum_{i=1}^{n} p_i^q. \tag{1.3}$$

Une autre manière d'appréhender l'entropie de Rényi est de noter que l'entropie de Shannon est la moyenne arithmétique, avec les poids p_i, des informations élémentaires $I_i = -\log p_i$ associées aux différents événements. En remplaçant la moyenne arithmétique par une moyenne de Kolmogorov-Nagumo, l'entropie devient $H_\psi(p_1, \ldots, p_n) = \psi^{-1}\left(\sum p_i \psi(-\log p_i)\right)$. Sous une condition supplémentaire d'additivité et sous $\lim_{p\to 0} H_\psi(p, 1-p) = 0$, alors cette entropie est soit l'entropie de Shannon, soit l'entropie de Rényi, avec $q \geq 0$. À nouveau, par la règle de l'Hôpital, on retrouve l'entropie de Shannon pour $q = 1$. Par ailleurs, pour $q = 0$, l'entropie de Rényi devient l'entropie de Hartley, le logarithme du nombre d'événements de probabilité non nulle.

1.3.2. *Entropies et divergences dans le cas continu*

Dans le cas continu, la définition utilisée pour l'entropie de Shannon associée à une densité $f(x)$ est

$$H[f] = -\int f(x) \log f(x) \mathrm{d}x. \tag{1.4}$$

Cependant, il faut bien noter que cette expression ne résulte du passage à la limite du cas discret qu'à une constante additive tendant vers l'infini près, voir par exemple [PAP 81]. On parle dès lors plutôt d'entropie différentielle. Cependant, Jaynes a lucidement noté, dès [JAY 63, p. 202], qu'il est nécessaire d'introduire une mesure $m(x)$ rendant compte de la « densité de points » dans un processus de passage à la limite ; cette mesure conférant de plus à l'information résultante une invariance par changement de coordonnées, ce qui n'est pas le cas de (1.4). L'entropie différentielle correspondante prend alors la forme

$$H[f] = -\int f(x) \log \frac{f(x)}{m(x)} \mathrm{d}x. \tag{1.5}$$

Cette forme est similaire à une divergence de Kullback-Leibler [KUL 59] (ou I-divergence dans la terminologie de Csiszár) entre deux distributions de probabilité de densités $f(x)$ et $g(x)$ par rapport à une mesure commune $\mu(x)$, et qui est définie par :

$$D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} \mathrm{d}\mu(x), \tag{1.6}$$

en supposant g absolument continue par rapport à f, et avec la convention $0 \log 0 = 0$. Lorsque g est uniforme, par rapport à μ, la divergence de Kullback devient, au signe près, une μ-entropie. Dans le cas où μ est la mesure de Lebesgue, on retrouve l'entropie différentielle de Shannon (1.4) ; dans le cas discret, si μ est la mesure de comptage, on retrouvera (1.1). On montre facilement, par une application de l'inégalité de Jensen, que la divergence de Kullback est définie non-négative, $D(f||g) \geq 0$ avec égalité ssi $f = g$. Elle peut ainsi être comprise comme une distance entre distributions, bien qu'elle ne soit pas symétrique et ne vérifie pas l'inégalité triangulaire.

De la même manière, on peut définir des versions continues des entropies de Rényi et Tsallis, pour un index entropique $q \neq 1$,

$$S_q[f] = \frac{1}{1-q} \left(\int f(x)^q \mathrm{d}\mu(x) - 1 \right), \tag{1.7}$$

est l'entropie de Tsallis et

$$H_q[f] = \frac{1}{1-q} \log \int f(x)^q \mathrm{d}\mu(x), \tag{1.8}$$

l'entropie de Rényi. Ces deux entropies se réduisent à l'entropie de Shannon pour $q = 1$. On peut également leur associer une divergence ; par exemple la divergence de Rényi

$$D_q(f||g) = \frac{1}{q-1} \log \int f(x)^q g(x)^{1-q} \mathrm{d}\mu(x), \tag{1.9}$$

qui est également définie non négative (par l'inégalité de Jensen), et se réduit à la divergence de Kullback lorsque $q \to 1$.

1.3.3. *Maximum d'entropie*

Le principe du maximum d'entropie est largement utilisé en physique, et peut reposer sur un grand nombre d'argumentations: dénombrements, axiomatiques, etc. Le principe a été particulièrement mis en exergue par Jaynes [JAY 57a] «*Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum entropy estimate. It is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information »,* et nous nous bornerons ici à en rappeler la pertinence en terme de statistiques, en suivant en cela Ellis [ELL 99] (théorème 2 par exemple).

Si f_N est la distribution empirique correspondant à un tirage de N variables aléatoires selon une loi de densité g par rapport à μ, alors la probabilité Q de trouver f_N dans un ensemble $\mathcal{B}$ est, *grossièrement* – cf Ellis [ELL 99] pour des formulations plus correctes, et pour N grand

$$Q\left(f_N \in \mathcal{B}\right) \approx \exp\left(-N \inf_{P\in\mathcal{B}} D(f||g)\right). \tag{1.10}$$

On en déduit donc, en itérant le raisonnement sur des sous ensembles de $\mathcal{B}$, que la distribution absolument prépondérante dans $\mathcal{B}$ est celle qui réalise le minimum de la distance de Kullback à g : on a concentration de toute la probabilité sur la distribution la plus proche de g. On en déduit donc *un principe de minimum de distance de Kullback*, ou de manière équivalente, si g est uniforme, *un principe de maximum d'entropie*. Parmi toutes les distributions d'un ensemble $\mathcal{B}$, il convient de sélectionner la densité qui minimise $D(f||g)$. Lorsque ce qui nous intéresse, comme en physique statistique, est la probabilité de trouver une moyenne empirique x_N, c'est-à-dire la moyenne sous f_N, dans un ensemble $\mathcal{C}$, alors on a un résultat de grandes déviation de niveau 1 qui indique que

$$Q\left(x_N \in \mathcal{C}\right) \approx \exp\left(-N \inf_{x\in\mathcal{C}} \mathcal{F}(\mathbf{x})\right), \tag{1.11}$$

où $\mathcal{F}(\mathbf{x})$ est la fonction de taux $\mathcal{F}(\mathbf{x}) = \inf_{P:\mathbf{x}=\mathrm{E}[\mathbf{X}]} D(P||\mu)$. Ce résultat suggère ainsi de sélectionner l'élément le plus probable, celui qui réalise le minimum de $\mathcal{F}(\mathbf{x})$ sur $\mathcal{C}$. Le passage d'une problématique sur les distributions à une problématique sur les moyennes est connu comme le principe de contraction.

1.3.4. *Distributions escortes*

Nous utiliserons également dans la suite de ce chapitre la notion de distribution escorte. Ces distributions escortes ont été introduites comme un outil dans le contexte des multifractales [CHH 89][BEC 93], avec d'intéressantes connections avec la

thermodynamique standard. Les distributions escortes se révèlent utiles en codage de source, où elles permettent d'obtenir des mots-codes optimaux dont la longueur moyenne est bornée par une entropie de Rényi [BER 09]. C'est ce que nous présenterons en 1.4.3. Nous retrouverons ensuite ces distributions escortes dans le cadre d'un problème de transition d'état, section 1.5.

Si $f(x)$ est une densité de probabilité, alors son escorte d'ordre $q \geq 0$ est donnée par

$$f_q(x) = \frac{f(x)^q}{\int f(x)^q \mathrm{d}\mu(x)}, \tag{1.12}$$

pourvu que la fonction génératrice informationnelle $M_q[f] = \int f(x)^q \mathrm{d}\mu(x)$ soit finie. On voit facilement que si $f_q(x)$ est l'escorte de $f(x)$, alors $f(x)$ est elle-même l'escorte d'ordre $1/q$ de $f_q(x)$. Quand q diminue, l'escorte se rapproche d'une distribution uniforme tandis que lorsque q augmente, les modes de la densité sont amplifiés. Ceci peut être précisé : on peut en effet montrer, dans le cas à support compact, que $D(f_q||U) > D(f||U)$ pour $q > 1$, ainsi que $D(f_q||U) < D(f||U)$ pour $q < 1$, ce qui signifie que f_q est plus éloignée de l'uniforme que f lorsque $q > 1$ et plus proche dans le cas contraire.

On peut également étendre la notion de distribution escorte pour prendre en compte deux densités $f(x)$ et $g(x)$ selon :

$$f_q(x) = \frac{f(x)^q g(x)^{1-q}}{\int f(x)^q g(x)^{1-q} \mathrm{d}\mu(x)}, \tag{1.13}$$

avec $M_q[f, g] = \int f(x)^q g(x)^{1-q} \mathrm{d}\mu(x) < \infty$. Cette distribution escorte généralisée est simplement une moyenne géométrique pondérée de $f(x)$ et $g(x)$. Bien entendu, si $g(x)$ est une mesure uniforme dont le support inclut celui de $f(x)$, alors l'escorte généralisée se réduit à l'escorte standard (1.12). Cette escorte généralisée apparaît dans l'analyse de l'efficacité de tests d'hypothèses [CHE 52] et permet de définir le meilleur exposant possible dans la probabilité d'erreur [COV 06, chapitre 11]. Quand q varie, l'escorte généralisée décrit une courbe qui connecte $f(x)$ et $g(x)$. Enfin, nous appellerons moments généralisés les moments pris par rapport à une distribution escorte : le moment généralisé d'ordre p associé à l'escorte standard d'ordre q sera

$$m_{p,q}[f] = \int |x|^p f_q(x) \mathrm{d}x = \frac{\int |x|^p f(x)^q \mathrm{d}\mu(x)}{\int f(x)^q \mathrm{d}\mu(x)}. \tag{1.14}$$

1.4. Codage de source avec des distributions escortes et des bornes de Rényi

Dans cette section, on illustre l'intérêt de l'entropie de Rényi et des distributions escortes dans le cadre du codage de source, l'un des problèmes fondamentaux de la théorie de l'information. Après un rappel très succinct du contexte du codage de

source, on décrit un théorème de codage de source reliant une nouvelle mesure de longueur moyenne et l'entropie de Rényi. On montre ensuite qu'il est possible de calculer pratiquement les codes optimaux en utilisant la notion de distribution escorte. Des détails sur ces éléments ainsi que d'autres résultats sont donnés dans [BER 09].

1.4.1. *Codage de source*

En codage de source, on considère un ensemble de symboles $\mathcal{X} = \{x_1, x_2, \dots x_N\}$, produits par un source avec des probabilités respectives p_i où $\sum_{i=1}^{N} p_i = 1$. Le rôle du codage de source est d'associer à chaque symbole x_i un mot-code c_i, le longueur l_i, exprimé avec un alphabet de D éléments. Il est bien connu que si les longueurs vérifient l'inégalité de Kraft-Mac Millan

$$\sum_{i=1}^{N} D^{-l_i} \leq 1, \tag{1.15}$$

alors il existe un code uniquement décodable avec ces longueurs élémentaires. De plus, tout code uniquement décodable satisfait l'inégalité de Kraft-Mac Millan (1.15). Le théorème du codage de source de Shannon indique que la longueur moyenne $\bar{L}$ des mots-code est bornée inférieurement par l'entropie de la source, $H_1(p)$, et que le meilleur code uniquement décodable satisfait

$$H_1(p) \leq \bar{L} = \sum_i p_i l_i < H_1(p) + 1, \tag{1.16}$$

où le logarithme utilisé dans l'entropie de Shannon est pris en base D, et noté $\log_D$. Ce résultat indique que l'entropie de Shannon $H_1(p)$ est une limite fondamentale à la longueur moyenne minimale pour tout code construit pour la source. Les longueurs des mots-codes optimaux sont données par

$$l_i = -\log_D p_i. \tag{1.17}$$

La caractéristique de ces codes optimaux est qu'ils assignent les mots les plus courts aux symboles les plus probables et les mots les plus longs aux symboles les plus rares.

1.4.2. *Codage de source avec la mesure de Campbell*

Il est bien connu que l'algorithme de Huffman fournit un code à préfixe qui minimise la longueur moyenne et approche les longueurs limites optimales $l_i = -\log_D p_i$. Cependant d'autres formes de mesure de longueur ont également été considérées ; parmi lesquelles la première, et fondamentale, contribution est celle de Campbell [CAM 65]. On a vu, par la relation (1.17), que les probabilités les plus faibles conduisent aux mots-codes les plus longs. Cependant, le coût d'utilisation d'un code n'est

pas forcément une fonction linéaire de sa longueur, et il est possible que que l'ajout d'une lettre sur un mot long soit bien plus coûteux que le coût de l'ajout de la même lettre sur un mot court. Ceci conduisit Campbell à proposer une nouvelle mesure de longueur moyenne, en introduisant une pénalisation exponentielle des longueurs des mots-codes. Cette longueur, la longueur de Campbell, est une moyenne généralisée de Kolmogorov-Nagumo associée à un fonction exponentielle :

$$C_\beta = \frac{1}{\beta} \log_D \sum_{i=1}^{N} p_i D^{\beta l_i}, \tag{1.18}$$

avec $\beta > 0$. Le résultat remarquable de Campbell est que, de la même manière que l'entropie de Shannon borne inférieurement la longueur moyenne des mots-codes, l'entropie de Rényi d'ordre q, avec $q = 1/(\beta + 1)$, est la borne inférieure de la longueur moyenne de Campbell (1.18):

$$C_\beta \geq H_q(p). \tag{1.19}$$

Une démonstration simple du résultat est donnée dans [BER 09]. Il est facile de voir que l'égalité est obtenue pour

$$l_i = -\log_D P_i = -\log_D \left(\frac{p_i^q}{\sum_{j=1}^{N} p_j^q} \right). \tag{1.20}$$

Clairement, les longueurs l_i obtenues de cette manière peuvent être rendues plus faibles que les longueurs optimales de Shannon, en choisissant un paramètre q assez faible, qui a alors tendance à uniformiser la distribution, rehaussant alors de fait les probabilités les plus faibles. Ainsi, la procédure pénalise effectivement les mots codes les plus longs et fournit des mots-codes de longueur différentes de celle de Shannon, avec éventuellement des mots-codes plus courts associés aux faibles probabilités.

1.4.3. *Codage de source avec une moyenne escorte*

Pour la mesure habituelle de longueur moyenne $\bar{L} = \sum_i p_i l_i$, on a une combinaison linéaire des longueurs élémentaires, pondérées par les probabilités p_i. De manière à accroître l'impact des longueurs les plus importantes associées aux probabilités faibles, la longueur de Campbell utilise une exponentielle des longueurs élémentaires. Une autre idée est de modifier les poids dans la combinaison linéaire, de manière à augmenter l'importance des termes avec des probabilités faibles. Une manière simple de réaliser cela est de déformer, uniformiser la distribution de probabilité initiale, et d'utiliser les poids donnés par cette nouvelle distribution plutôt que les p_i. Naturellement, ceci conduit à utiliser une moyenne prise sous la distribution escorte :

$$M_q = \sum_{i=1}^{N} \frac{p_i^q}{\sum_j p_j^q} l_i = \sum_{i=1}^{N} P_i l_i. \tag{1.21}$$

Pour la source imaginaire qui aurait une distribution P, la moyenne statistique standard est M_q, et le théorème de source classique de Shannon s'applique immédiatement :

$$M_q \geq H_1(P), \tag{1.22}$$

avec égalité si

$$l_i = -\log_D P_i \tag{1.23}$$

soit exactement les longueurs (1.20) obtenues pour la mesure de Campbell. La simple relation $l_i = -\log_D P_i$ obtenue pour la minimisation de M_q sous la contrainte fournie par l'inégalité de Kraft-Mac Millan a une application immédiate mais importante. En effet, il suffit de fournir la distribution escorte P plutôt que la distribution initial p à un algorithme de codage standard, par exemple un algorithme de Huffman, pour obtenir un code optimisé pour la longueur de Campbell C_β, ou, de manière équivalente, pour la mesure de longueur M_q. Un exemple simple est donné Table 1.1 dans le cas $D = 2$: nous avons utilisé un algorithme de Huffman standard, avec la distribution initiale, puis ses escortes d'ordre $q = 0.7$ et $q = 0.4$.

p_i	$q = 1$	$q = 0.7$	$q = 0.4$
0.48	0	0	00
0.3	10	10	01
0.1	110	1100	100
0.05	1110	1101	101
0.05	11110	1110	110
0.01	111110	11110	1110
0.01	111111	11111	1111

Tableau 1.1 – Exemple de codes obtenus dans le cas binaire, pour différentes valeurs de q.

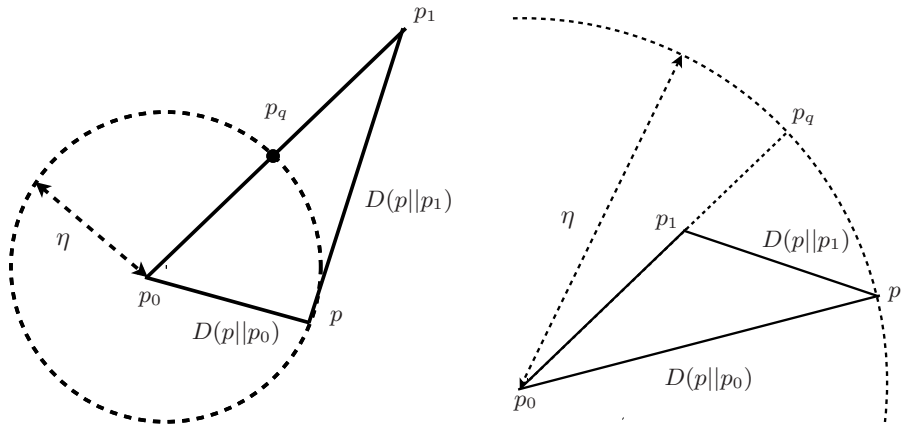
Il est important de noter que des algorithmes spécifiques ont été développés pour la longueur moyenne de Campbell. La connexion ci-dessus fournit une alternative aisée et immédiate. Un autre point important est que ces codes ont des applications pratiques : ils sont optimaux pour la minimisation de la probabilité de débordement de buffers [HUM 81] ou, avec $q > 1$, pour maximiser la probabilité de réception d'un message en un seul envoi de taille limitée.

1.5. Un modèle simple de transition

Dans la section précédente, nous avons vu apparaître, et apprécié l'intérêt, de l'entropie de Rényi et des distributions escortes pour un problème de codage de source. Dans cette section, nous montrerons que ces deux quantités interviennent également

dans le cadre d'un modèle d'équilibre, ou de transition, entre deux états. Il a en effet été noté que la thermodynamique étendue, associée aux entropies de Tsallis et Rényi, semble particulièrement pertinente dans la cas de déviations à l'équilibre classique de Boltzmann-Gibbs. Ceci suggère alors d'amender la formulation classique de l'approche classique du maximum d'entropie (ou du minimum de divergence) et d'imaginer un équilibre caractérisé par deux (et non plus une seule) distributions : plutôt que de sélectionner la distribution la plus proche d'une distribution de référence sous une contrainte de moyenne, on recherchera une distribution $p_q(x)$ intermédiaire, en un sens à préciser, entre deux références $p_0(x)$ et $p_1(x)$. Cette construction, ainsi que certaines de ses conséquences, sont aussi décrits dans [BER 12].

1.5.1. *Le modèle*



(a) Cas $\eta < D(p_1||p_0)$ (b) Cas $\eta > D(p_1||p_0)$

Figure 1.1 – Équilibre entre les états p_0 and p_1 : la distribution d'équilibre est recherchée dans l'ensemble des distributions situées à une divergence fixée à p_0, $D(p||p_0) = \eta$, et à une distance de Kullback minimale de p_1. La distribution d'équilibre résultante p_q, la distribution escorte généralisée est « alignée » avec p_0 et p_1 et à l'intersection de l'ensemble $D(p||p_0) = \eta$.

On considère deux états de densités de probabilité $p_0(x)$ et $p_1(x)$ en un point x de l'espace des phases, et on recherche un état intermédiaire selon le scénario suivant. Le système d'état initial p_0, soumis à une force généralisée, est déplacé, et maintenu, à une distance $\eta = D(p||p_0)$ de p_0. D'autre part, le système est attiré vers un état final p_1. Par suite, le nouvel état intermédiaire p_q est choisi comme celui qui minimise sa divergence à l'attracteur p_1 tout en étant maintenu à une distance η de p_0. Comme illustré sur la Figure 1.1, la densité de probabilité intermédiaire est «alignée» avec p_0

et p_1 et à l'intersection avec l'ensemble $D(p||p_0) = \eta$, un cercle de rayon η centré sur p_0. Plus précisément, en prenant les densités par rapport à la mesure de Lebesgue, le problème peut être formulé comme suit :

$$\left\{ \begin{array}{c} \min_p \ D(p||p_1) \\ \text{sous } D(p||p_0) = \eta \\ \text{et } \int p(x)\mathrm{d}x = 1. \end{array} \right. \tag{1.24}$$

La solution est donnée par la proposition suivante.

PROPOSITION 1.1.– *Si q est un réel positif tel que $D(p_q||p_0) = \eta$ et si $M_q(p_1, p_0) = \int p_1(x)^q p_0(x)^{1-q}\mathrm{d}x < \infty$, alors, la solution du problème (1.24) est donnée par*

$$p_q(x) = \frac{p_1(x)^q p_0(x)^{1-q}}{\int p_1(x)^q p_0(x)^{1-q}\mathrm{d}x}. \tag{1.25}$$

REMARQUE.– Lorsque p_0 est uniforme, avec un support compact, on retrouve la distribution escorte standard (1.12). Si le support n'est pas compact et la distribution uniforme impropre, il est possible de modifier simplement la formulation en prenant pour contrainte une entropie fixée $H(p) = -\eta$, et l'on obtient alors la distribution escorte (1.12).

Évaluons la divergence $D(p||p_q)$. Pour toute les densités p telles que la contrainte $D(p||p_0) = \eta$ soit satisfaite, on a

$$\begin{aligned} D(p||p_q) = & \int p(x) \log \frac{p(x)}{p_q(x)}\mathrm{d}x = \int p(x) \log \frac{p(x)^q p(x)^{1-q}}{p_1(x)^q p_0(x)^{1-q}}\mathrm{d}x + \log M_q(p_1, p_0) \\ = & \ q \int p(x) \log \frac{p(x)}{p_1(x)}\mathrm{d}x + (1-q) \int p(x) \log \frac{p(x)}{p_0(x)}\mathrm{d}x + \log M_q(p_1, p_0) \\ = & \ q\, D(p||p_1) + (1-q)\eta + \log M_q(p_1, p_0) \end{aligned} \tag{1.26}$$

En prenant $p = p_q$, la dernière égalité devient

$$D(p_q||p_q) = q\, D(p_q||p_1) + (1-q)\eta + \log M_q(p_1, p_0). \tag{1.27}$$

Finalement en soustrayant (1.26) et (1.27), on obtient

$$D(p||p_q) - D(p_q||p_q) = q\ \left(D(p||p_1) - D(p_q||p_1)\right). \tag{1.28}$$

Puisque $q \geq 0$ et $D(p||p_q) \geq 0$ avec égalité ssi $p = p_q$, on obtient finalement $D(p||p_1) \geq D(p_q||p_1)$ ce qui prouve la Proposition.

Lorsque η varie, la fonction $q(\eta)$ est monotone croissante, avec toujours $D(p_q||p_0) = \eta$. Pour $\eta = 0$ on a $q = 0$ et pour $\eta = D(p_1||p_0)$ on a $q = 1$. Ainsi, lorsque q varie, p_q

définit une courbe qui connecte p_0 ($q = 0$) à p_1 ($q = 1$), et au delà pour $q > 1$, voir la Figure 1.1.

REMARQUE.– Il est intéressant de noter aussi que des résultats ont montré que le travail dissipé durant une transition peut être exprimé comme une divergence de Kullback-Leibler [PAR 09]. Dans ce contexte, avec un Halmitonien pair suivant l'impulsion, la contrainte $D(p||p_k) = \eta$, $k = 0$ ou 1, peut être interprétée comme une borne sur le travail moyen dissipé durant la transition de p à p_k.

1.5.2. *La divergence de Rényi comme conséquence*

Finalement, il est intéressant de constater que la divergence de Rényi apparaît comme sous-produit de notre construction. En effet, par une conséquence directe de (1.27) et de la définition de la divergence de Rényi (1.9), le minimum de l'information de Kullback peut-être exprimé comme

$$D(p_q||p_1) = \left(1 - \frac{1}{q}\right)(\eta - D_q(p_1||p_0))\,. \tag{1.29}$$

En prenant une mesure uniforme pour p_0, on fait apparaître l'entropie de Rényi

$$D(p_q||p_1) = \left(1 - \frac{1}{q}\right)(\eta + H_q\,[p_1])\,. \tag{1.30}$$

La divergence de Kullback-Leibler n'est pas symétrique. Dès l'origine, Kullback et Leibler ont introduit une version symétrique, retrouvant en cela la divergence de Jeffreys. Dans notre cas, cette divergence de Jeffreys est une simple fonction affine de la divergence de Rényi :

$$J(p_1, p_q) = D(p_1||p_q) + D(p_q||p_1) = \frac{(q-1)^2}{q}\,(D_q(p_1||p_0) - \eta)\,. \tag{1.31}$$

Cette égalité est une conséquence simple de la relation (1.26), avec $p = p_1$, et de la relation (1.27). On peut noter, comme conséquence importante, que la minimisation de la divergence de Jeffreys entre p_1 et p_q sous certaines contraintes, est donc équivalent à la minimisation de la divergence de Rényi avec les mêmes contraintes.

1.5.3. *Information de Fisher sur le paramètre* q

La distribution escorte généralisée p_q définit une courbe indexée par q connectant les distributions p_0 et p_1 pour $q = 0$ et $q = 1$. Il est intéressant d'évaluer l'information attachée au paramètre q de la distribution généralisée. Cette information de Fisher est donnée par

$$I(q) = \int \frac{1}{p_q(x)} \left(\frac{dp_q(x)}{dq} \right)^2 \mathrm{d}x = \int \frac{dp_q(x)}{dq} \log \frac{p_1(x)}{p_0(x)} \, \mathrm{d}x. \tag{1.32}$$

où le terme de droite s'obtient en utilisant la relation

$$\frac{\mathrm{d}p_q(x)}{\mathrm{d}q} = p_q(x) \left(\log \frac{p_1(x)}{p_0(x)} - \frac{\mathrm{d} \log M_q}{\mathrm{d}q} \right), \tag{1.33}$$

et le fait que $\int \frac{\mathrm{d}p_q(x)}{\mathrm{d}q} \mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}q} \int p_q(x) \mathrm{d}x = 0$ par la règle de Leibniz. On peut montrer également que cette information de Fisher est égale à la variance, par rapport à la loi p_q, du rapport de vraisemblance.

Finalement, il est possible d'identifier l'intégrale de l'information de Fisher le long de la courbe, l'"énergie» de la courbe, à la divergence de Jeffreys. Plus précisément, on a

PROPOSITION 1.2.– *L'intégrale de l'information de Fisher, de $q = r$ à $q = s$ est proportionnelle à la divergence de Jeffreys entre p_r et p_s :*

$$(s - r) \int_r^s I(q) \mathrm{d}q = J(p_s, p_r) = D(p_s || p_r) + D(p_r || p_s). \tag{1.34}$$

Avec $r = 0$ et $s = 1$, on obtient donc que

$$\int_0^1 I(q) \mathrm{d}q = J(p_1, p_0) = D(p_1 || p_0) + D(p_0 || p_1). \tag{1.35}$$

Pour démontrer (1.34), il suffit d'intégrer (1.32) :

$$\begin{aligned} \int_r^s I(q) \mathrm{d}q &= \int_r^s \int \frac{\mathrm{d}p_q(x)}{\mathrm{d}q} \log \frac{p_1(x)}{p_0(x)} \, \mathrm{d}x \, \mathrm{d}q. \\ &= \int (p_s(x) - p_r(x)) \log \frac{p_1(x)}{p_0(x)} \, \mathrm{d}x. \end{aligned}$$

En tenant compte du fait que $\log p_s/p_r = (s - r) \log p_1/p_0$, on obtient alors (1.34).

Finalement, si θ_i, $i = 1..M$ sont un ensemble de variables intensives dépendant de q, alors $\frac{d\log p}{dq} = \sum_{i=1}^{M} \frac{\partial \log p}{\partial \theta_i} \frac{d\theta_i}{dq}$ et l'information de Fisher de q peut s'exprimer en fonction de la matrice d'information de Fisher sur θ. Dans ces conditions, et pour la distribution escorte généralisée, on obtient que la «divergence thermodynamique» sur la transition n'est autre que la divergence de Jeffreys :

$$\mathcal{J} = \int_0^1 I(q)\mathrm{d}q = \sum_{i=1}^{M}\sum_{j=1}^{M} \int_0^1 \frac{\mathrm{d}\theta_i}{\mathrm{d}q} \left[I(\theta)\right]_{i,j} \frac{\mathrm{d}\theta}{\mathrm{d}q}\,\mathrm{d}q = D(p_1||p_0) + D(p_0||p_1). \tag{1.36}$$

1.5.4. *Inférence d'une distribution sous contrainte de moment généralisé*

Supposons maintenant que la distribution p_1 soit imparfaitement connue, mais qu'une information complémentaire soit disponible sous la forme d'une valeur moyenne, prise sous la distribution p_q. Cette moyenne est la moyenne généralisée (1.14), qui est utilisée en physique statistique nonextensive ; elle a ici l'interprétation claire d'une moyenne prise sous la distribution d'équilibre p_q. Le problème qui se pose maintenant est alors la détermination de la distribution la plus générale compatible avec cette contrainte.

On peut conserver l'idée de minimiser la divergence à p_1, comme dans le problème (1.24) qui nous a mené à la distribution d'équilibre en escorte généralisée. Comme la divergence de Kullback est dirigée, on conservera la direction en minimisant $D(p_q||p_1)$ pour $q < 1$ et $D(p_1||p_q)$ pour $q > 1$. Dans les deux cas, la divergence s'exprime comme une fonction affine de la divergence de Rényi $D_q(p_1||p_0)$, cf (1.29), et ces minimisations sont finalement équivalentes à la minimisation de la divergence de Rényi sous la contrainte de moyenne généralisée.

Dans le même ordre d'idée, on pourrait s'intéresser à la minimisation de la divergence symétrique de Jeffreys entre p_q et p_1. Or, nous avons noté en (1.31) que celle-ci s'exprime également comme une fonction affine de la divergence de Rényi $D_q(p_1||p_0)$. Sa minimisation est donc équivalente à la minimisation de la divergence de Rényi sous contrainte de moyenne généralisée.

Enfin, la divergence de Jeffreys $J(p_1, p_q)$ est proportionnelle à la divergence thermodynamique, l'intégrale de l'information de Fisher, comme indiqué en (1.34), pour $q > 1$ comme pour $q < 1$. Dès lors, la minimisation de la divergence thermodynamique entre p_q et p_1 est également équivalente à la minimisation de la divergence de Rényi.

Ces différents arguments nous conduisent donc très légitimement à rechercher la distribution p_1 comme la distribution minimisant la divergence de Rényi d'index q, sous la contrainte de moyenne généralisée.

1.6. Minimisation de la divergence de Rényi et entropies associées

Dans les paragraphes précédents nous avons décrit un cadre permettant de faire apparaître de manière naturelle l'information de Rényi, les distributions escortes et les moments généralisés. De plus, nous en avons déduit une méthode d'inférence de distribution : la minimisation de l'information de Rényi, avec une information disponible sous forme de moments généralisés. Dans cette section, nous donnerons d'abord l'expression de la densité qui minimise la divergence de Rényi, puis nous décrirons certaines propriétés des fonctions de partition associées. Enfin, nous montrerons comment on peut en déduire de nouvelles fonctionnelles entropiques dont on donnera quelques exemples. Une partie de ces résultats, mais aussi certaines extensions pourront être consultés dans [BER 08].

1.6.1. *Minimisation de la divergence de Rényi sous contrainte de moment généralisé*

Nous considérerons d'abord un moment généralisé d'ordre quelconque (1.14), dont l'expression est rappelée ci-dessous

$$m_{p,q}[f] = \int |x|^p f_q(x) \mathrm{d}\mu(x) = \frac{\int |x|^p f(x)^q g(x)^{1-q} \mathrm{d}\mu(x)}{\int f(x)^q g(x)^{1-q} \mathrm{d}\mu(x)}. \tag{1.37}$$

On considère alors le problème

$$\mathcal{F}_q(m) = \left\{ \begin{array}{cc} & \min_f D_q(f||g) \\ \text{sous} & m = m_{p,q}[f] \\ \text{et} & \int f(x) \mathrm{d}\mu(x) = 1. \end{array} \right. \tag{1.38}$$

Le minimum obtenu est bien sûr une fonction de la contrainte $m = m_{p,q}[f]$, que l'on notera $\mathcal{F}_q(m)$. Il s'agit d'une version contractée de la divergence de Rényi, qui définit une « entropie » dans l'espace des moyennes possibles m. Dans [BER 11], nous avons considéré un problème plus général, dans lequel les index du moment généralisé et de la divergence de Rényi ne sont pas identiques. Quoi qu'il en soit, on a ici le résultat suivant.

PROPOSITION 1.3.– *La densité G_γ qui réalise le minimum dans le problème (1.38) est donnée par*

$$G_\gamma(x) = \frac{1}{Z_\nu(\gamma, \bar{x}_p)} \left(1 - (1-q)\gamma \left(|x|^p - \bar{x}_p\right)\right)_+^\nu g(x) \tag{1.39}$$

ou de façon équivalente par

$$G_{\bar{\gamma}}(x) = \frac{1}{Z_\nu(\bar{\gamma})} \left(1 - (1-q)\bar{\gamma}|x|^p\right)_+^\nu g(x) \tag{1.40}$$

avec $\nu = 1/(1-q)$, $\bar{x}_p$ un paramètre de translation éventuel, γ et $\bar{\gamma}$ des paramètres d'échelle choisis tels que la contrainte de moment généralisé soit satisfaite, et enfin où $(x)_+ = \max(0, x)$. Les quantités $Z_\nu(\gamma, \bar{x}_p)$ et $Z_\nu(\bar{\gamma})$ sont les fonctions de partition qui permettent de normaliser la densité. Pour $q = 1$, la densité $G_\gamma(x)$ devient une densité exponentielle

$$G_\gamma(x) = \frac{1}{Z_\nu(\gamma)} \exp\left(-\gamma\left(|x|^p - \bar{x}_p\right)\right) g(x) \tag{1.41}$$

par rapport à $g(x)$.

Dans le cas $p = 2$, on retrouve ainsi une densité gaussienne. La densité G_γ est parfois appelée « gaussienne généralisée ». On notera encore que γ et $\bar{\gamma}$ sont reliés par la relation

$$\bar{\gamma} = \frac{\gamma}{1 + \frac{\gamma}{\nu}\bar{x}_p}. \tag{1.42}$$

On propose ici la démonstration dans le cas de l'expression (1.40). La démarche est tout-à-fait similaire dans la cas de la densité (1.39).

Comme dans [BER 11], posons $A(\bar{\gamma}) = 1/Z(\bar{\gamma})$. On a immédiatement

$$\begin{aligned}\int f^q G_{\bar{\gamma}}^{1-q} \mathrm{d}\mu(x) &= A(\bar{\gamma})^{1-q} M_q[f,g] \times \int \left(1 - (1-q)\bar{\gamma}|x|^p\right)_+ \frac{f^q g^{1-q}}{M_q[f,g]} \mathrm{d}\mu(x) \\ &\geq A(\bar{\gamma})^{1-q}(1 - (1-q)\bar{\gamma}\, m_{p,q}[f])\, M_q[f,g], \end{aligned} \tag{1.43}$$

avec $M_q[f,g] = \int f^a g^{1-q} \mathrm{d}\mu(x)$, où $m_{p,q}[f]$ désigne le moment généralisé, et où l'inégalité résulte du fait que le support de $\left(1 - (1-q)\bar{\gamma}|x|^p\right)_+$peut être inclut dans celui de $f^q g^{1-q}$. L'inégalité devient une égalité dans le cas $q \geq 1$. À partir de (1.43) on a directement, avec $f = G_{\bar{\gamma}}$:

$$M_1[G_{\bar{\gamma}}] = 1 = A(\bar{\gamma})^{1-q}\left(1 - (1-q)\bar{\gamma}\, m_{q,p}[G_{\bar{\gamma}}]\right) M_q[G_{\bar{\gamma}}, g]. \tag{1.44}$$

Ainsi, pour toutes les distributions f de moment généralisé $m_{p,q}[f] = m$ et pour $\bar{\gamma}$ tel que $G_{\bar{\gamma}}$ ait le même moment $m_{p,q}[G_{\bar{\gamma}}] = m$, alors la combinaison de (1.43) et (1.44) entraîne

$$\int f^q G_{\bar{\gamma}}^{1-q} \mathrm{d}\mu \geq \frac{M_q[f,g]}{M_q[G_{\bar{\gamma}}, g]},$$

avec égalité si $q \geq 1$. Finalement la divergence de Rényi d'ordre q peut ainsi être exprimée comme

$$D_q(f||G_{\bar{\gamma}}) = \log\left(\int f^q G_{\bar{\gamma}}^{1-q} \mathrm{d}\mu(x)\right)^{\frac{1}{q-1}} \tag{1.45}$$

$$\leq \log\left(\frac{M_q[f,g]}{M_q[G_{\bar{\gamma}}, g]}\right)^{\frac{1}{q-1}} = D_q(f||g) - D_q(G_{\bar{\gamma}}||g). \tag{1.46}$$

Par la non négativité de la divergence, on obtient donc que

$$D_q(f||g) \geq D_q(G_{\bar{\gamma}}||g) \tag{1.47}$$

pour toutes les distributions f de moment généralisé $m_{p,q}[f] = m_{p,q}[G_{\bar{\gamma}}] = m$, et avec égalité ssi $f = G_{\bar{\gamma}}$.

1.6.2. *Quelques propriétés des fonctions de partition*

Dans ce paragraphe, on donne quelques propriétés essentielles des fonctions de partition $Z_\nu(\gamma, \bar{x}_p)$ associées à la densité optimale G_γ, voir [BER 08]. Ces propriétés seront essentielles pour la caractérisation des fonctionnelles entropiques $\mathcal{F}_q(x)$. On désigne par E_ν la moyenne statistique prise vis-à-vis de la loi optimum de densité (1.39), avec $\nu = 1/(1-q)$. Il est aussi important de réaliser, dès à présent, que la densité escorte d'ordre q de (1.39) n'est autre que cette même densité G_γ mais avec un exposant $\nu - 1$, de sorte que l'on ait

$$m_{p,q}[G_{\bar{\gamma}}] = E_{\nu-1}\left[X\right]. \tag{1.48}$$

Les fonctions de partition successives sont liées par

$$Z_\nu(\gamma, \overline{x}_p) = E_{\nu-1}\left[1 - \frac{\gamma}{\nu}\left(|x|^p - \overline{x}_p\right)\right] Z_{\nu-1}(\gamma, \overline{x}). \tag{1.49}$$

Comme conséquence directe, on voit que $Z_\nu(\gamma, \overline{x}_p) = Z_{\nu-1}(\gamma, \overline{x}_p)$ ssi $\overline{x}_p = E_{\nu-1}\left[|X|^p\right]$.

En utilisant la règle de Leibniz, on peut obtenir que la dérivée par rapport à γ est donnée par

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} Z_\nu(\gamma, \overline{x}_p) = \left(-E_{\nu-1}\left[|X|^p - \overline{x}_p\right] + \gamma \frac{\mathrm{d}\overline{x}_p}{\mathrm{d}\gamma}\right) Z_{\nu-1}(\gamma, \overline{x}_p), \tag{1.50}$$

sous réserve que $\overline{x}_p$ soit bien différentiable par rapport à γ. De même

$$\frac{\mathrm{d}}{\mathrm{d}\overline{x}_p} Z_\nu(\gamma, \overline{x}_p) = \left(-\frac{\mathrm{d}\gamma}{\mathrm{d}\bar{x}_p} E_{\nu-1}\left[|X|^p - \overline{x}_p\right] + \gamma\right) Z_{\nu-1}(\gamma, \overline{x}_p) \tag{1.51}$$

Ainsi, si $\overline{x}_p = E_{\nu-1}\left[|X|^p\right]$, alors en tenant compte de l'égalité des fonctions de partition de rang ν et $\nu - 1$, on a

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} \log Z_\nu(\gamma, \overline{x}_p) = \gamma \frac{\mathrm{d}\overline{x}_p}{\mathrm{d}\gamma}, \tag{1.52}$$

ou encore

$$\frac{\mathrm{d}}{\mathrm{d}\overline{x}_p} \log Z_\nu(\gamma, \overline{x}_p) = \gamma. \tag{1.53}$$

Par ailleurs, lorsque $\overline{x}_p$ est un paramètre indépendant de γ, disons $\overline{x}_p = m$, alors

$$\frac{\mathrm{d}^2 Z_\nu(\gamma, m)}{\mathrm{d}\gamma^2} = \frac{(\nu-1)}{\nu} E_{\nu-2}\left[(X-m)^2\right] Z_{\nu-2}(\gamma, m), \tag{1.54}$$

et de même

$$\frac{\mathrm{d}^2 Z_\nu(\gamma, m)}{\mathrm{d}m^2} = \frac{(\nu-1)}{\nu} \gamma^2 E_{\nu-2}\left[(X-m)^2\right] Z_{\nu-2}(\gamma, m), \tag{1.55}$$

ce qui, compte tenu du fait que $(\nu-1)/\nu = q > 0$, du fait que les fonctions de partition sont strictement positives, montre que si $\overline{x}_p = m$ et γ sont indépendants, alors la fonction de partition $Z_\nu(\gamma, m)$ est convexe en ses deux variables.

Enfin, on peut exprimer la solution du problème (1.38), c'est-à-dire $\mathcal{F}_q(m)$, à la fonction de partition. Par calcul direct, on a en effet

$$D_q(G_\gamma || g) = \frac{1}{q-1} \log Z_{q\nu}(\gamma, \bar{x}_p) - \frac{q}{q-1} \log Z_\nu(\gamma, \bar{x}_p), \tag{1.56}$$

ce qui se réduit simplement à

$$\mathcal{F}_q(m) = D_q(G_\gamma || g) = -\log Z_\nu(\gamma, m) = -\log Z_{\nu-1}(\gamma, m), \tag{1.57}$$

pour la valeur de γ telle que la contrainte soit satisfaite, soit $m_{p,q}[G_{\tilde{\gamma}}] = E_{\nu-1}[X] = \bar{x}_p = m$.

1.6.3. *Fonctionnelles entropiques issues de la divergence de Rényi*

Ainsi, la solution du problème de minimisation de la divergence de Rényi d'ordre q, vue comme une fonction de la contrainte, définit une «fonctionnelle entropique ». Différentes fonctionnelles seront associées aux différentes spécifications de la densité de référence $g(x)$, ainsi qu'aux différentes valeurs de l'index q. Nous allons voir que les fonctions en question présentent des propriétés intéressantes. On dispose ainsi potentiellement d'un ensemble de fonctions qui peuvent être utilisées, éventuellement, comme fonctions objectif, ou termes de régularisation.

La principale caractérisation de $\mathcal{F}_q(m)$ est la suivante.

PROPOSITION 1.4.– *L'entropie $\mathcal{F}_q(m)$, définie par (1.38), est non négative, avec un minimum unique en m_g, la moyenne de g, et $\mathcal{F}_q(m_g) = 0$. L'entropie est une fonction pseudo-convexe pour $q \in [0, 1)$ et strictement convexe pour $q \geq 1$.*

La divergence de Rényi $D_q(f||g)$ est toujours non négative, et nulle uniquement pour $f = g$. Puisque les fonctionnelles $\mathcal{F}_q(m)$ sont définies comme le minimum de

la divergence $D_q(f||g)$, elles sont toujours non négatives. À partir de (1.53), on a $\frac{d}{d\overline{x}} \log Z_\nu(\gamma, \overline{x}) = \gamma$. Ainsi, les fonctionnelles $\mathcal{F}_q(x)$ ne présentent qu'un seul point singulier en $\gamma = 0$. Pour cette valeur de γ, on a $G_{\gamma=0} = g$, et $D_q(g||g) = 0$. Dans ces conditions, $\mathcal{F}_q(x)$ possède un minimum unique pour $x = m_g$, la moyenne de g, et $\mathcal{F}_q(m_g) = 0$. On obtient donc que $\mathcal{F}_q(x)$ est unimodale et ne présente pas de points d'inflexion avec une tangente horizontale ; ceci suffit pour affirmer que $\mathcal{F}_q(x)$ est pseudo convexe, au sens de Mangasarian [MAN 87]. Examinons maintenant la convexité pour $q \geq 1$. Si f_q est la distribution escorte généralisée donnée par (1.13), alors on a l'égalité $D_q(f||g) = D_{\frac{1}{q}}(f_q||g)$. Par suite, rechercher la distribution f qui réalise le minimum de $D_q(f||g)$ avec une contrainte de moyenne généralisée, *i.e.* prise vis-à-vis de f_q, est équivalente à rechercher la distribution f_q qui minimise $D_{\frac{1}{q}}(f_q||g)$, sous une contrainte de moment standard. Dans ces conditions, soient p_1 et p_2 les deux densités qui minimisent $D_{\frac{1}{q}}(f_q||g)$ sous les contraintes $x_1 = E_{f_q}[X]$ et $x_2 = E_{f_q}[X]$. Alors, $\mathcal{F}_q(x_1) = D_{\frac{1}{q}}(p||g)$, et $\mathcal{F}_q(x_2) = D_{\frac{1}{q}}(p_2||g)$. De la même manière, soit $\mathcal{F}_q(\mu x_1 + (1-\mu)x_2) = D_{\frac{1}{q}}(\hat{P}||Q)$, où $\hat{P}$ est la distribution escorte optimale de moyenne $\mu x_1 + (1-\mu)x_2$. Les distributions $\hat{P}$ et $\mu p_1 + (1-\mu)p_2$ ont alors même moyenne. Ainsi, lorsque $D_{\frac{1}{q}}(f_q||g)$ est une fonction strictement convexe de f_q, c'est-à-dire pour $q \geq 1$, on a $D_{\frac{1}{q}}(\hat{P}||g) < \mu D_{\frac{1}{q}}(p_1||g) + (1-\mu) D_{\frac{1}{q}}(p_2||g)$, soit $\mathcal{F}_q(\mu x_1 + (1-\mu)x_2) < \mu \mathcal{F}_q(x_1) + (1-\mu)\mathcal{F}_q(x_2)$ et l'entropie $\mathcal{F}_q(x)$ est une fonction strictement convexe.

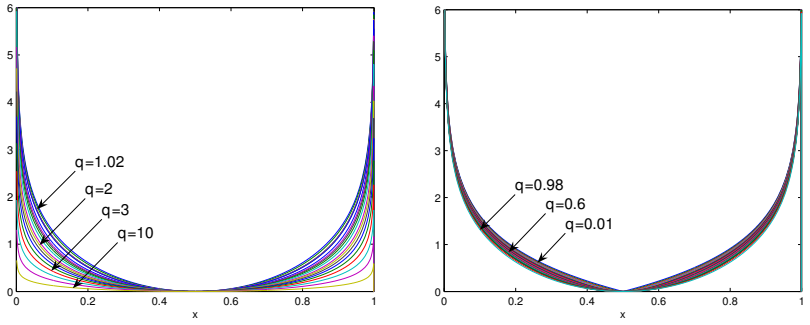
Même munis de cette caractérisation intéressante, il reste un problème pratique d'importance : comment déterminer analytiquement, ou numériquement les entropies $\mathcal{F}_q(x)$ pour une densité de référence g et un index entropique q donnés. Le problème revient à déterminer le paramètre γ tel que la moyenne généralisée de la densité optimale (1.39) ait une valeur spécifiée m. Une manière simple de procéder est de revenir au fait que si $\bar{x}_p$ est un paramètre fixe m, indépendant de γ, alors la relation de dérivation (1.50) se réduit à

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} Z_\nu(\gamma, m) = \left(m - E_{\nu-1}\left[|X|^p\right]\right) Z_{\nu-1}(\gamma, m). \tag{1.58}$$

Dès lors, on voit qu'il suffit de rechercher les extrema de la fonction de partition $Z_\nu(\gamma, m)$ pour obtenir un γ tel que $m = E_{\nu-1}\left[|X|^p\right]$. Comme nous avons vu que $Z_\nu(\gamma, m)$, à m fixé, est strictement convexe, alors cet extremum est unique et est un minimum. Enfin, la valeur de l'entropie est simplement donnée par (1.57) : $\mathcal{F}_q(m) = -\log Z_\nu(\gamma, m)$.

La recherche de l'expression de $\mathcal{F}_q(m)$ requiert donc de calculer la fonction de partition puis de résoudre $\frac{\mathrm{d}}{\mathrm{d}\gamma} Z_\nu(\gamma, m) = 0$ par rapport à γ. Hors quelques cas particuliers, cette résolution ne semble pas possible analytiquement, et l'entropie $\mathcal{F}_q(m)$ est donnée implicitement. Dans le cas particulier où g est une mesure de Bernoulli,

il est possible d'obtenir une expression analytique pour $\mathcal{F}_q(m)$, ceci pour tout $q > 0$. Pour d'autres densités de référence g, il est possible d'obtenir des expressions analytiques quand $q \to 1$. Ces points sont détaillés dans [BER 08], où par ailleurs différentes densités de référence g sont étudiées, et les entropies correspondantes évaluées numériquement, selon le schéma indiqué précédemment. À titre d'exemple, on donne figures 1.2a et 1.2b les résultats numériques obtenus dans le cas $p = 1$ et d'une densité uniforme sur l'intervalle $[0, 1]$. Pour $q \geq 1$, on obtient bien une famille de fonctions convexes sur $(0, 1)$, minimum pour la moyenne de g, soit 0.5, comme nous l'avions indiqué plus haut. Pour $q < 1$, on obtient une famille de fonctions non négatives, unimodales, et également minimales en $x = m_g = 0.5$.



(a) Cas $q > 1$. (b) Cas $q \in (0, 1)$.

Figure 1.2 – Entropie $\mathcal{F}_q(x)$ pour une référence uniforme, respectivement pour $q \geq 1$ et $q \in (0, 1)$.

1.6.4. *Critères entropiques*

Au vu des figures précédentes, il est apparent que la minimisation de $\mathcal{F}_q(x)$ sous certaines contraintes fournit automatiquement une solution dans l'intervalle $(0, 1)$. De plus, le paramètre q peut être utilisé pour ajuster la courbure de la fonction ou la pénalisation sur les bords du domaine. Il est ainsi intéressant d'utiliser ces entropies en vue de la résolution de problèmes inverses. Plus précisément, on peut utiliser un critère entropique tel que $\mathcal{F}_q(x)$ comme fonction objectif. On se restreindra dans toute cette section au cas $p = 1$. Si on considère un problème inverse linéaire $\mathbf{y} = \mathbf{A}\mathbf{x}$, avec x_k les composantes de $\mathbf{x}$, alors ceci peut se formuler comme

$$\begin{cases} \min_{\mathbf{x}} \sum_k \mathcal{F}_q(x_k) \\ \text{sous } \mathbf{y} = \mathbf{A}\mathbf{x}. \end{cases} \tag{1.59}$$

Ceci correspond alors à sélectionner parmi les solutions possibles la solution dont les composantes sont d'entropie minimale. Remarquons que l'on a supposé ici, implicitement, que le critère était séparable en ses composantes. En réalité, si on définit $\mathcal{F}_q(\mathbf{x})$ comme le minimum de la divergence de Rényi sous une contrainte de moyenne généralisée, alors, même en supposant les composantes indépendantes, on obtient une densité sur $\mathbf{x}$ analogue à (1.39), qui n'est pas séparable sur ses composantes. Afin d'obtenir un critère séparable, qui est à la fois plus conforme à l'intuition et d'une utilisation plus aisée, on amende la formulation en recherchant la densité produit qui réalise le minimum de la divergence de Rényi sous contrainte de moyenne généralisée, ce qui conduit effectivement au critère séparable. Ainsi, le problème précédent (1.59) peut aussi se lire

$$\left\{ \begin{array}{l} \min_{\mathbf{x}} \left\{ \begin{array}{c} \min_{\mathbf{f}} D_q(\mathbf{f}||\mathbf{g}) \\ \text{sous } \mathbf{f} = \prod_k f_k \\ \text{et } \mathbf{x} = E_{\mathbf{f}_q}[X] \end{array} \right. \\ \text{sous } \mathbf{y} = \mathbf{A}\mathbf{x}, \end{array} \right. \tag{1.60}$$

où $E_{\mathbf{f}_q}[X]$ désigne la moyenne généralisée, *i.e.* prise par rapport à la distribution escorte d'ordre q. Il s'agit donc bien ici d'un problème de «maximum d'entropie » qui consiste à sélectionner une solution $\mathbf{x}$, vue comme moyenne généralisée d'une loi à minimum de divergence de Rényi à une densité de référence g. Les entropies $\sum_k \mathcal{F}_q(x_k)$ étant pseudo-convexes, on sait que la minimisation sous contraintes linéaires conduit à un minimum unique, *c.f.* par exemple [CAM 08, théorème 4.4.1]. Examinons maintenant comment on peut obtenir une solution de (1.59), même dans le cas où les entropies n'ont pas d'expression explicite. La solution correspond à un point stationnaire du lagrangien $L(\boldsymbol{\lambda}, \mathbf{x})$ associé au problème (1.59), et il s'agit donc de résoudre

$$\min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}, \mathbf{x}) = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} \sum_k \mathcal{F}_q(x_k) + \boldsymbol{\lambda}^T (\mathbf{y} - \mathbf{A}\mathbf{x}) \tag{1.61}$$

$$= \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} \sum_k \mathcal{F}_q(x_k) - c_k x_k + \boldsymbol{\lambda}^T \mathbf{y} \tag{1.62}$$

avec $c_k = \left[\boldsymbol{\lambda}^T \mathbf{A}\right]_k$. En utilisant le fait que $\mathcal{F}_q(x_k) = -\log Z_\nu(\gamma_*, x_k) = -\inf_\gamma \log Z_\nu(\gamma, x_k)$, comme nous l'avons vu en section 1.6.3, on a ainsi

$$\min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}, \mathbf{x}) = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} \sum_k -\log Z_\nu(\gamma_*, x_k) - c_k x_k + \boldsymbol{\lambda}^T \mathbf{y} \tag{1.63}$$

$$= \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} - \sum_k \max_{x_k} \left(\log Z_\nu(\gamma_*, x_k) + c_k x_k\right). \tag{1.64}$$

Or, par la relation (1.53), il vient

$$\frac{\mathrm{d}}{\mathrm{d}x_k} \left(\log Z_\nu(\gamma_*, x_k) + c_k x_k\right) = \gamma_* + c_k \tag{1.65}$$

ce qui entraîne $\gamma_* = -c_k$, et x_k est la moyenne généralisée associée. Il s'agit donc finalement de résoudre

$$\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} - \sum_k \left(\log Z_\nu(-c_k, x_k) + c_k x_k \right), \tag{1.66}$$

où, pour tout c_k, la moyenne généralisée correspondante x_k peut être calculée comme solution unique du problème

$$x_k = \arg \min_x \left(\log Z_\nu(-c_k, x) + c_k x \right). \tag{1.67}$$

En suivant cette démarche, il est ainsi possible de résoudre le problème (1.59). Ceci conduit alors à une solution, unique, « à maximum d'entropie de Rényi » au problème inverse linéaire $\mathbf{y} = \mathbf{A}\mathbf{x}$, problème où l'on peut inclure différentes contraintes, notamment de support, via la densité de référence g, et où la forme du critère peut être ajustée par l'intermédiaire de l'index entropique q.

Dans le cas où les données $\mathbf{y}$ seraient imparfaites, il est possible de minimiser le critère entropique sous une contrainte fournie par une statistique (par exemple du χ^2) sur le résidu, plutôt qu'avec une contrainte exacte. Il est également possible d'utiliser le critère entropique avec un terme d'attache aux données.

Dans le cas $q = 1$, la divergence de Rényi se réduit à la divergence de Kullback, les moments généralisés aux moments habituels, et la densité optimale à une densité exponentielle (1.41) par rapport à g. Dans ces conditions, la log-fonction de partition s'écrit $\log Z_\infty(-c_k, x_k) = -c_k x_k + \log \int \exp\left(c_k x_k\right) g(x_k) \mathrm{d}\mu(x_k)$, le problème (1.66) devient

$$\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} - \sum_k \log \int \exp\left(c_k x_k\right) g(x_k) \mathrm{d}\mu(x_k), \tag{1.68}$$

et la solution optimale est donnée par la dérivée de la log-fonction de partition par rapport aux c_k. Cette dernière approche a été développée dans un travail dirigé par Guy Demoment [LEB 99].

1.7. Bibliographie

[ACZ 75] ACZÉL J., DAROCZY Z., *On measures of information and their characterizations*, Academic Press, 1975.

[ACZ 84] ACZÉL J., "Measuring information beyond communication theory—Why some generalized information measures may be useful, others not", *Aequationes Mathematicae*, vol. 27, n° 1, p. 1–19, mars 1984.

[BEC 93] BECK C., SCHLOEGL F., *Thermodynamics of Chaotic Systems*, Cambridge University Press, 1993.

[BER 08] BERCHER J.-F., "On some entropy functionals derived from Rényi information divergence", *Information Sciences*, vol. 178, n° 12, p. 2489–2506, juin 2008.

[BER 09] BERCHER J.-F., "Source coding with escort distributions and Rényi entropy bounds", *Physics Letters A*, vol. 373, n° 36, p. 3235–3238, août 2009.

[BER 11] BERCHER J.-F., "Escort entropies and divergences and related canonical distribution", *Physics Letters A*, vol. 375, n° 33, p. 2969–2973, août 2011.

[BER 12] BERCHER J.-F., "A simple probabilistic construction yielding generalized entropies and divergences, escort distributions and -Gaussians", *Physica A: Statistical Mechanics and its Applications*, vol. 391, n° 19, p. 4460–4469, octobre 2012.

[BRI 62] BRILLOUIN L., *Science and Information Theory*, Academic Pr, 2 édition, juin 1962.

[CAM 65] CAMPBELL L. L., "A coding theorem and Rényi's entropy", *Information and Control*, vol. 8, n° 4, Page423–429, 1965.

[CAM 08] CAMBINI A., MARTEIN L., *Generalized convexity and optimization*, Springer, novembre 2008.

[CHE 52] CHERNOFF H., "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations", *The Annals of Mathematical Statistics*, vol. 23, n° 4, p. 493–507, 1952.

[CHH 89] CHHABRA A., JENSEN R. V., "Direct determination of the $f(\alpha)$ singularity spectrum", *Physical Review Letters*, vol. 62, n° 12, Page1327, mars 1989.

[COV 06] COVER T. M., THOMAS J. A., *Elements of Information Theory 2nd Edition*, Wiley-Interscience, 2 édition, juillet 2006.

[CSI 06] CSISZÁR I., "Stochastics: Information Theory", HORVÁTH J., Ed., *A Panorama of Hungarian Mathematics in the Twentieth Century I*, vol. 14, p. 523–535, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[DAR 70] DARÓCZY Z., "Generalized information functions", *Information and Control*, vol. 16, n° 1, p. 36–51, mars 1970.

[ELL 99] ELLIS R. S., "The theory of large deviations: from Boltzmann's 1877 calculation to equilibrium macrostates in 2D turbulence", *Physica D*, vol. 133, n° 1-4, p. 106–136, septembre 1999.

[FAD 56] FADDEEV D., "On the concept of entropy of a finite probabilistic scheme", *Uspekhi Mat. Nauk*, vol. 11, n° 1(67), p. 227–231, 1956, (in russian).

[HAV 67] HAVRDA J., CHARVÁT F. S., "Quantification method of classification processes. Concept of structural a-entropy", *Kybernetika*, vol. 3, p. 30–35, 1967.

[HUM 81] HUMBLET P., "Generalization of Huffman coding to minimize the probability of buffer overflow", *IEEE Transactions on Information Theory*, vol. 27, n° 2, p. 230–232, 1981.

[JAY 57a] JAYNES E. T., "Information Theory and Statistical Mechanics", *Physical Review*, vol. 106, n° 4, p. 620–630, mai 1957.

[JAY 57b] JAYNES E. T., "Information Theory and Statistical Mechanics. II", *Physical Review*, vol. 108, n° 2, p. 171–190, octobre 1957.

[JAY 63] JAYNES E. T., "Information theory and statistical mechanics.", *1962 Brandeis Summer Institute in Theoretical Physics*, vol. 3, p. 182–218, K. W. Ford, W. A. Benjamin Inc., New York, 1963, reprinted in: E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics, éd. R. D. Rosencrantz, Synthèse Library, Vol. 138, Reidel, 1983.

[KUL 59] KULLBACK S., *Information Theory and Statistics*, John Wiley & Sons, New York, 1959, Republished by Dover Publications, 1997.

[LEB 99] LE BESNERAIS G., BERCHER J.-F., DEMOMENT G., "A new look at entropy for solving linear inverse problems", *IEEE Transactions on Information Theory*, vol. 45, n° 5, p. 1565–1578, 1999.

[MAN 87] MANGASARIAN O. L., *Nonlinear Programming*, Society for Industrial Mathematics, janvier 1987.

[MER 10] MERHAV N., "Statistical Physics and Information Theory", *Foundations and Trends in Communications and Information Theory*, vol. 6, n° 1-2, p. 1–212, 2010.

[PAP 81] PAPOULIS A., "Maximum entropy and spectral estimation: A review", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, n° 6, p. 1176–1186, décembre 1981.

[PAR 09] PARRONDO J. M. R., BROECK C. V. D., KAWAI R., "Entropy production and the arrow of time", *New Journal of Physics*, vol. 11, n° 7, Page073008, juillet 2009.

[R´ 61] RÉNYI A., "On measures of entropy and information", *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, Berkeley, Calif., Univ. California Press, p. 547–561, 1961.

[R´ 65] RÉNYI A., "On the Foundations of Information Theory", *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, vol. 33, n° 1, p. 1–14, 1965.

[SHA 48a] SHANNON C., "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, n° 3, p. 379–423, 1948.

[SHA 48b] SHANNON C., "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, n° 4, Page623–656, 1948.

[TSA 88] TSALLIS C., "Possible generalization of Boltzmann-Gibbs statistics", *Journal of Statistical Physics*, vol. 52, n° 1, p. 479–487, juillet 1988.

[TSA 09] TSALLIS C., *Introduction to Nonextensive Statistical Mechanics*, Springer, avril 2009.