Introduction à la Théorie de l'information

Raphaël Lachièze-Rey*

raphael.lachieze-rey@parisdescartes.fr

30 décembre 2023

Résumé

Le modèle de base de la théorie de l'information, dans lequel des messages sont émis par une source aléatoire, est présenté ici. Nous définissons alors les différentes versions de l'entropie de Shannon, qui permettent de quantifier le débit moyen d'information associé à une source aléatoire. La question du codage optimal, avec ou sans perte, de l'information émise par la source est alors traitée. On donne aussi une version discrète du Théorème de Sanov.

Table des matières

*Basé sur des notes initiales de Florent Benaych-Georges et Manon Defosseux

Introduction

Théorie des communications : moyen de transmettre une information depuis une source jusqu'à un utilisateur (cf figure 1) :

- Source = voix, signal électromagnétique, séquences symboles binaires,...
- Codeur = ens des opérations effectuées sur la sortie de la source avant transmission (modulation, compression,... but = combattre le bruit)
- Canal = ligne téléphonique, liaison radio, disque compact,...
- Bruit = pertubateur du canal : perturbations électriques, rayures,...
- Décodeur = restituer l'information de la source

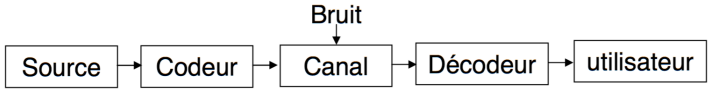


FIGURE 1 – Transmission d'information : schéma classique

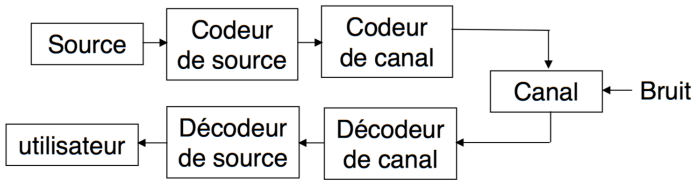


FIGURE 2 – Séparation modèles de sources et modèles de canaux

Simplification : séparation modèles de sources et modèles de canaux (cf figure 2).

Théorie de l'information (Shannon, 1916–2001, premier article sur le sujet en 1948) : C'est la partie de la théorie des communications qui, à l'aide de modèles mathématiques (proba, stats, codes,...), permet la mesure quantitave de l'information, son codage et le dimensionnement des cannaux de communication.

Résumé du cours : Le modèle de base de la théorie de l'information, dans lequel des messages sont émis par une source aléatoire, est présenté ici. Nous définissons alors les différentes versions de l'entropie de Shannon, qui permettent de quantifier le débit moyen d'information associé à une source aléatoire. La question du codage optimal, avec ou sans perte, de l'information émise par la source est traitée. Nous verrons que la solution fonctionnelle et optimale consiste grosso modo à coder un message X ayant n valeurs possibles $x_1, \ldots, x_n$, de probabilités respectives $p_1, \ldots, p_n$ en attribuant à chaque valeur x_i un nombre de bits d'information (binaire) $\approx -\ln_2 p_i$. Le nombre moyen de bits d'information nécessaire pour coder X est alors

$$\sum_{i=1}^{n} -p_i \ln_2 p_i,$$

une quantité (appelée *entropie de* X) jouant un rôle-clé dans ce cours.

1 Mesure quantitative de l'information

1.1 Introduction

Qu'est-ce exactement qu'un *message* ?

Symbol	Frequency	Huffman Code
[space]	67962112	111
e	37907119	010
t	28691274	1101
a	24373121	1011
o	23215532	1001
i	21820970	1000
n	21402466	0111
s	19059775	0011
h	18058207	0010
r	17897352	0001
l	11730498	10101
d	10805580	01101
c	8982417	00001
u	8022379	00000
f	7486889	110011
m	7391366	110010
w	6505294	110001
y	5910495	101001
p	5719422	101000
g	5143059	011001
b	4762938	011000
v	2835696	1100000
k	1720909	11000011
x	562732	110000100
j	474021	1100001011
q	297237	11000010101
z	93172	11000010100

Symbol	H4
[space]	0
e	32
t	31
a	23
o	22
i	21
n	13
s	12
h	11
r	10
l	333
d	332
c	331
u	330
f	303
m	302
w	301
y	203
p	202
g	201
b	200
v	3003
k	3002
x	3000
j	30012
q	30011
z	30010

FIGURE 3 – Nombre d'occurrences des lettres dans le corpus et codes de Huffman binaire et 4-aire

- *Message = suite finie de symboles appartenant à un ensemble fini, prédéterminé : l'alphabet.*
  Exemples d'alphabet :
  - ⋆ Lettres : a b c d e f...
  - ⋆ Alphabet binaire : 0 1

  Exemples de messages :

  "rendez-vous le 17 avril"

  ou

  "01101001010101100010100011101001011101"

- Envoi de messages par la source.
- Pour le destinataire, la source et le canal ont un comportement aléatoire, décrit en termes probabilistes.
- La communication n'a d'intérêt que si le contenu du message est inconnu a priori. *Plus un message est imprévu, improbable, plus il est informatif.*
- Qualitativement, fournir une information = lever une partie de l'incertitude sur l'issue d'une expérience aléatoire.

Exemple 1.1. L'information "il fera plus que 10°" demain est binaire, dans le sens où il n'y avait que 2 possibilités sur cette incertitude (Plus ou moins).

L'information "il fera 17°" est beaucoup plus précise, car il y avait plus de possibilités, elle apporte plus d'information.

Pour quantifier cela, imaginons que ces messages soit transmis par un code binaire, c'est-à-dire fait de "0" et "1". Les messages du premier type seront codés par des messages de longueur 1 ([0] ou [1]). Pour coder les messages du second type, on les remplace simplement par leur écriture binaire.

On observe que si on veut coder par exemple toutes les 32 températures possibles entre 0(= [00000] en binaire) et 31(= [11111]), il faut prévoir des codes de longueur $5 = \ln_2(32)$, et plus généralement si je veux coder toutes les températures entre 0 et $N-1$, il faut des codes de longueur au moins $\ln_2(N)$.

On peut remarquer sur cet exemple que si on suppose pour simplifier que toutes les températures sont équiprobables (de probabilité $p_i := \frac{1}{N}$), alors l'entropie définie plus bas nous donne la même valeur

$$\sum_i -p_i \ln_2(p_i) = \sum_{i=0}^{N-1} \frac{-1}{N} \ln_2(1/N) = \ln_2(N),$$

le comportement est donc bien similaire au nombre de bits nécessaire pour coder l'information.

En affinant le modèle et en attribuant des probabilités plus faibles à certaines températures, on fera baisser l'entropie. L'explication heuristique est que si certaines températures sont plus probables que d'autres, alors l'incertitude contenue dans le message est plus faible. Concrètement, on s'attend à avoir une température entre 10 et 25 degrés, ce qui est déjà "moins aléatoire" que n'importe quelle température entre 0 et 31...

Comment mesurer la " quantité d'information " d'un message ?

- Message = événements aléatoires produits par la source (exemple d'événement = émission d'une suite de symboles discrets choisis dans un ensemble fini de symboles (alphabet))
- La quantité d'information du message est proportionnelle à son degré d'improbabilité : un événement certain ou connu à l'avance du destinataire n'est pas trés informatif...
- Plus formellement : c'est le nombre de questions à réponse oui/non que cette connaissance nous évite d'avoir à poser. On voit de cette manière apparaître une quantité importante. En effet, pour déterminer avec certitude un nombre entre 0 et $2^n - 1$ par une série de questions fermées, il faudra procéder par dichotomie
  - ⋆ plus grand que 0? Oui.
  - ⋆ Plus grand que 16 ? Non.
  - ⋆ Plus grand que 8 ?
  - ⋆ Etc...,

  il faudra n questions (chacune correspondant à un 0 ou un 1 dans l'écriture en base 2 du nombre sur n bits), c'est-à-dire $\ln_2(2^n)$.

Bibliographie sommaire :

Les références apparaissant en premier dans cette liste sont celles qui correspondent le plus au cours.

- Marc Lelarge *Introduction à la Théorie de l'Information et au Codage*. Page web de l'auteur. (2014).
- Olivier Rioul *Théorie de l'information et du codage*, Lavoisier, 2007.
- Thomas M. Cover, Joy A. Thomas *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing) (2006).
- Raymond W. Yeung *A First Course in Information Theory*, Springer (2012).
- Marc Mézard, Andrea Montanari *Information, Physics, and Computation*, Oxford, 2009.
- Yann Ollivier *Aspects de l'entropie en mathématiques*. Page web de l'auteur. (2002).

1.2 Modèle probabiliste

Dans le modèle probabiliste, les messages possibles sont des variables aléatoires, et les ensembles de messages possibles sont donc des évènements.

Exemple 1.2. Soit $\Omega = \{-31, ..., 32\}$ l'ensemble des températures vraisemblables pour demain, et X une variable uniforme sur Ω. Soit A l'évènement $\{X > 10\}$.

Par ce qui précède, l'information propre d'un message, ou plus généralement, d'un ensemble A de messages envisagés, notée $\mathsf{I}(A)$, est d'autant plus petite que ce message est attendu, elle doit être une fonction décroissante de sa probabilité :

$$\mathsf{I}(A) = f(\mathbb{P}(A)),$$

avec f **décroissante**.

Par exemple l'information des messages $A = \{X > 10\}, B = \{X = 17\}$ sont

$$\mathsf{I}(A) = f(22/64) > \mathsf{I}(B) = f(1/64).$$

Que doit vérifier de plus une bonne fonction "d'information" ? Un autre axiome est que si les événements A et B sont statistiquement indépendants alors l'information totale qu'ils peuvent fournir ensemble est la somme des informations propres :

$$\mathsf{I}(A \cap B) = \mathsf{I}(A) + \mathsf{I}(B),$$

ce qui signifie que f doit vérifier

$$f(\mathbb{P}(A) \times \mathbb{P}(B))) = f(\mathbb{P}(A)) + f(\mathbb{P}(B)).$$

Enfin, bien entendu, on doit avoir

$$\mathsf{I}(\Omega) = 0,$$

càd

$$f(1) = 0.$$

Théorème 1.3. *L'ensemble des fonctions $f : [1, +\infty) \to \mathbb{R}_+$ continues telles que $f(xy) = f(x) + f(y)$ et $f(1) = 0$ sont les fonctions proportionnelles à* ln.

Démonstration. (preuve quand f est supposée dérivable) Toute fonction proportionnelle à $\ln_2$ vérifie $f(xy) = f(x) + f(y)$, et réciproquement, si $f : [1, +\infty) \to \mathbb{R}_+$ est dérivable telles que $f(xy) = f(x) + f(y)$, alors pour $g(x) := f(e^x)$, $x \geq 0$, on a $g(x+y) = g(x)+g(y)$ et $g(0) = 0$, ce qui implique que g' est constante, ce qui permet de conclure. □

On est donc amené à choisir, pour f, le $\ln_2$ dans une base choisie :

- $f = \ln_2$ unité : bit ou Shannon (Sh) (**choix fait dans ce cours**),
- $f = \ln_e$ unité : nat,
- $f = \ln_{10}$ unité : dit ou Hartley.

Définition 1.4. Soient A, B des événements (d'un même espace de probabilités). On définit :

- L'*information propre de* A :
$$\mathsf{I}(A) := -\ln_2(\mathbb{P}(A)).$$
- L'*information conjointe de* A et B :
$$\mathsf{I}(A, B) := \mathsf{I}(A \cap B).$$
- L'*information conditionnelle de* A *sachant* B :
$$\mathsf{I}(A|B) := -\ln_2(\mathbb{P}(A|B)).$$
- L'*information mutuelle de* A *et* B :
$$\mathsf{I}(A; B) := \ln_2(\mathbb{P}(A \cap B)/(\mathbb{P}(A)\mathbb{P}(B))) = \mathsf{I}(B; A).$$

Remarque 1.5. Par la formule de Bayes $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$, on a
$$\mathsf{I}(A, B) = \mathsf{I}(A) + \mathsf{I}(B|A) = \mathsf{I}(B) + \mathsf{I}(A|B).$$
On en déduit que
$$\mathsf{I}(A; B) = \mathsf{I}(A) - \mathsf{I}(A|B) = \mathsf{I}(B) - \mathsf{I}(B|A).$$
Autrement dit, le signe de l'info mutuelle est une mesure de la corrélation entre les événements ($\mathsf{I}(A; B) > 0$ si A et B sont positivement corrélés, $\mathsf{I}(A; B) < 0$ si A et B sont négativement corrélés et $\mathsf{I}(A; B) = 0$ si A et B sont indépendants).

Exemple 1.6. Considérons une source dont l'alphabet de sortie a 16 éléments $a_0, \dots, a_{15}$ équiprobables. Alors pour tout i, $\mathsf{I}(\{a_i\}) = 4\text{bits}$. Attention, cela est vrai car on a équiprobabilité ! Sinon, cela dépend de la probabilité d'occurence de a_i ! C'est tout l'objet de ce cours de comprendre ce qu'il en est du cas non équiprobable . Un exemple simple est celui où a_0, a_1 ont une proba $0.5 - 14 \cdot 10^{-100}$ d'advenir, alors que les autres lettres ont une proba 10^{-100} d'advenir. On comprend bien que dans ce cas, grosso-modo, 1 bit d'information suffit.

1.3 Mesures quantitatives moyennes de l'information : entropie

Supposons que la source est une variable aléatoire X qui prend les valeurs $x_1, \dots, x_n$, avec probas respectives $p_1, \dots, p_n$. Alors la quantité d'info moyenne émise par la source est

$$\mathsf{H}(X) := \sum_i p_i \, \mathsf{I}(X = x_i) = -\sum p_i \ln_2(p_i),$$

avec la convention $0 \ln_2(0) = 0$. On note aussi $\mathsf{H}(X) = \mathsf{H}(\mathcal{L}(X))$, où $\mathcal{L}(X)$ désigne la loi de X.

De façon plus générale, on introduit les notions suivantes.

Définition 1.7. Soient X qui prend les valeurs $x_1, \dots, x_n$, avec probas respectives $p_1, \dots, p_n$ et Y qui prend les valeurs $y_1, \dots, y_m$, avec probas respectives $q_1, \dots, q_m$. Posons

$$r_{ij} := \mathbb{P}(X = x_i, Y = y_j).$$

On définit

- L'*entropie de* X :

$$\mathsf{H}(X) := \sum_i p_i \, \mathsf{I}(X = x_i) = -\sum_i p_i \ln_2 p_i.$$

- L'*entropie conjointe de* X *et* Y :

$$\mathsf{H}(X, Y) := -\sum_{i,j} r_{ij} \ln_2 r_{ij},$$

qui n'est rien d'autre que l'entropie de la v.a. $Z := (X, Y)$:

$$\mathsf{H}(X, Y) = \mathsf{H}(Z).$$

- L'entropie conditionnelle de X sachant un évènement A :

$$\mathsf{H}(X \mid A) := -\sum_i p_{i|A} \ln_2(p_{i|A})$$

où $p_{i|A} := \mathbb{P}(X = i \mid A)$

- L'*entropie conditionnelle de X sachant Y* :

$$\mathsf{H}(X|Y) \;\; := \;\; \sum_j q_j \mathsf{H}(X \mid Y = j),$$

- L'*information mutuelle moyenne de X et Y* :

$$\mathsf{I}(X;Y) := \mathsf{H}(X) - \mathsf{H}(X \mid Y) = \mathsf{I}(Y;X)$$

(avec la convention $0 \ln_2 \frac{0}{0} = 0$), la seconde égalité doit être prouvée.

La symétrie de l'information mutuelle vient de la représentation

$$\mathsf{I}(X;Y) = \sum_{i,j} r_{ij} \ln_2 \frac{r_{ij}}{p_i q_j}$$

Exercice 1.1. Prouver cette représentation.

Remarque 1.8. Bien entendu, $\mathsf{H}(X)$ ne dépend que des probas $p_1, \ldots, p_n$ de la loi de X, on peut donc écrire $\mathsf{H}(p_1, \ldots, p_n)$ ou $\mathsf{H}(p)$ pour $p = (p_1, \ldots, p_n)$ ou pour p la loi de X. De même, $\mathsf{H}(X, Y)$, $\mathsf{H}(X|Y)$ et $\mathsf{I}(X;Y)$ ne dépendent que des nombres r_{ij}.

Remarque 1.9. Le poids de chaque p_i dans $\mathsf{H}(p_1, \ldots, p_n)$ est donc $-p_i \ln_2(p_i)$: la pondération forte des événements rares avec $\mathsf{I}(A) = -\ln_2(\mathbb{P}(A))$ est donc tempérée par le fait de prendre l'espérance. La figure 4 présente la courbe de la fonction $p \in [0,1] \mapsto -p \ln_2(p)$.

Exemple 1.10. Soit $X \sim B(p)$. Alors

$$\mathsf{H}(X) = -p \ln_2 p - (1-p) \ln_2(1-p) = -\ln_2(p^p (1-p)^{1-p})$$

représentée dans la figure 5.

Théorème 1.11. *On a $\mathsf{H}(X|Y) = 0$ si et seulement si $X = g(Y)$ pour une fonction g.*

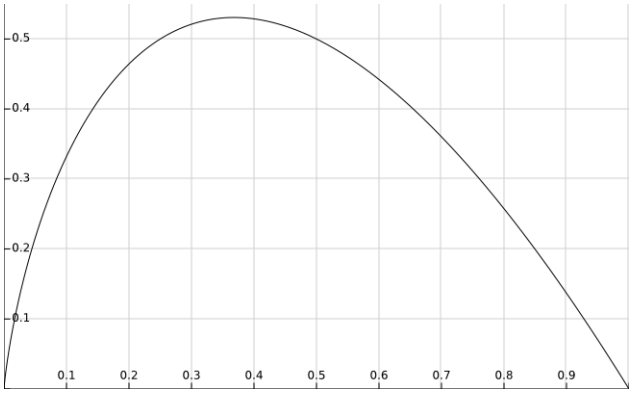


FIGURE 4 – Courbe de la fonction $p \in [0,1] \mapsto -p\ln_2(p)$ (la tangente à l'origine est verticale et le maximum est atteint en $p = \mathrm{e}^{-1}$).

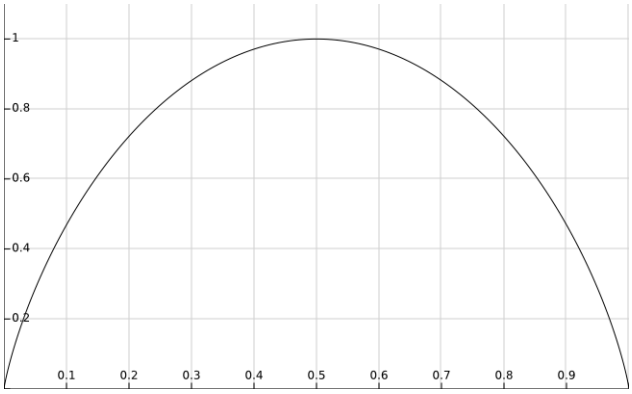


FIGURE 5 – Entropie $\mathsf{H}(X) = -\ln_2(p^p(1-p)^{1-p})$ d'une v.a. de Bernoulli en fonction de $p \in [0,1]$. Notons que cette courbe est (le double de) la symétrisée, par rapport à $1/2$, de la courbe de la figure 4.

Démonstration. Utilisons la formule

$$\mathsf{H}(X|Y) \;:=\; \sum_j q_j \mathsf{H}(\mathcal{L}(X|Y=y_j)),$$

où $\mathcal{L}(X|Y=y_j)$ désigne la loi de X sachant que $Y=y_j$. On a

$$\mathsf{H}(X|Y)=0 \iff \forall i, \mathsf{H}(\mathcal{L}(X|Y=y_j)) \iff \forall j, \mathcal{L}(X|Y=y_j) \text{ masse de Dirac,}$$

ce qui permet de conclure. □

Proposition 1.12. *Montrer que* $\mathsf{H}(\varphi(X)) \leqslant \mathsf{H}(X)$ *pour toute fonction* φ*, avec égalité ssi* φ *est injective.*

Exercice 1.2. Preuve.

Remarque 1.13. L'entropie propre de X se réécrit :

$$\mathsf{H}(X) = -\,\mathbb{E}\ln_2 p(X)$$

pour p la fonction définie par $p(x_i) = p_i$.

Remarque 1.14. Notons que, contrairement à l'espérance conditionnelle $\mathbb{E}[X|Y]$, $\mathsf{H}(X|Y)$ n'est pas aléatoire ! On pourrait introduire une notion aléatoire, similaire à $\mathbb{E}[X|Y]$, qui serait $\mathsf{H}_{\text{aleat}}(X|Y)$ définie par

$$\mathsf{H}_{\text{aleat}}(X|Y)(\omega) := \mathsf{H}(\mathcal{L}(X|Y = Y(\omega))).$$

Théorème 1.15 (Chain rule 1)**.** *On a toujours*

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

Démonstration. On a

$$\begin{aligned}
\mathsf{H}(X,Y) &= -\sum_{i,j} r_{ij}\ln_2 r_{ij} \\
&= -\sum_{i,j} r_{ij}\ln_2 q_j \mathbb{P}(X = x_i|Y = y_j) \\
&= -\sum_{i,j} r_{ij}\ln_2 q_j - \sum_{i,j} r_{ij}\ln_2 \mathbb{P}(X = x_i|Y = y_j) \\
&= -\sum_{j} q_j\ln_2 q_j - \sum_{i,j} r_{ij}\ln_2 \mathbb{P}(X = x_i|Y = y_j),
\end{aligned}$$

ce qui donne directement la formule. □

Par récurrence, on en déduit :

Corollaire 1.16 (Chain rule 2)**.** *On a toujours*

$$\mathsf{H}(X_1,\ldots,X_n) = \mathsf{H}(X_1 \mid X_2,\ldots,X_n) + \mathsf{H}(X_2 \mid X_3,\ldots,X_n) + \cdots + \mathsf{H}(X_{n-1} \mid X_n) + \mathsf{H}(X_n).$$

2 Concavité de l'entropie

Soit, pour $n \geq 1$,

$$\mathcal{P}_n := \{(p_1, \ldots, p_n) \in [0,1]^n \,;\, p_1 + \cdots + p_n = 1\}.$$

Proposition 2.1. *La fonction* H *est une fonction symétrique sur* $\mathcal{P}_n$ *(i.e. invariante par changement de l'ordre de ses arguments) et invariante par adjonction d'un* 0 *à la suite de ses arguments :*

$$\mathsf{H}(p_1, \ldots, p_n) = \mathsf{H}(p_1, \ldots, p_n, 0).$$

Théorème 2.2. *La fonction* H *est strictement concave sur le convexe* $\mathcal{P}_n$*, c'est-à-dire, pour tout* $p, q \in \mathcal{P}_n$*,* $\lambda \in (0,1)$*,*

$$\lambda \mathsf{H}(p) + (1-\lambda)\mathsf{H}(q) < \mathsf{H}(\lambda p + (1-\lambda)q),$$

avec égalité si et seulemnt si $p = q$*.*

Démonstration. Il suffit de montrer la stricte concavité de la fonction 1D

$$\varphi(\lambda) = \mathsf{H}(\lambda p + (1-\lambda)q) = -\sum_i \gamma(\lambda p_i + (1-\lambda)q_i)$$

avec $\gamma(x) = x \ln_2(x)$ strictement convexe, donc $\gamma'' > 0$ sur $(0,1)$. Donc

$$\varphi''(\lambda) = -\sum_i \frac{d^2}{d\lambda^2}\gamma(\lambda p_i + (1-\lambda)q_i) = -\sum_i (p_i - q_i)^2 \gamma''(\lambda p_i + (1-\lambda)q_i)$$

$\varphi'' < 0$ si il existe i tel que $p_i \neq q_i$, c'est-à-dire si $p \neq q$.

On peut aussi utiliser la hessienne : On a, pour tout i,

$$\frac{\partial}{\partial p_i}\mathsf{H}(p) = -(\ln_2 p_i + 1)/\ln_2 2$$

donc la matrice Hessienne de H en p est $\mathrm{diag}(-1/p_i, 1 \leq i \leq n)$, qui est bien définie négative [1] si p appartient à l'intérieur de $\mathcal{P}_n$. □

Si l'on fusionne deux états, par exemple $\{1\}$ et $\{2\}$, l'information, et donc l'entropie, devraient logiquement baisser :

1. et non pas positive

Corollaire 2.3. *L'entropie décroit strictement par fusion de plusieurs de ses arguments : si* $p_1, p_2 > 0$,

$$\mathsf{H}(p_1, \dots, p_n) > \mathsf{H}(p_1 + p_2, p_3, \dots, p_n).$$

Démonstration. On a, pour $p_1, p_2 > 0$,

$$(p_1, \dots, p_n) = \frac{p_2}{p_1 + p_2}(0, p_1 + p_2, p_3, \dots, p_n) + \frac{p_1}{p_1 + p_2}(p_1 + p_2, 0, p_3, \dots, p_n),$$

donc, par stricte concavité,

$$\mathsf{H}(p_1, \dots, p_n) > \frac{p_2}{p_1 + p_2}\mathsf{H}(0, p_1+p_2, p_3, \dots, p_n)+\frac{p_1}{p_1 + p_2}\mathsf{H}(p_1+p_2, 0, p_3, \dots, p_n).$$

Or, par la proposition 2.1,

$$\mathsf{H}(0, p_1 + p_2, p_3, \dots, p_n) = \mathsf{H}(p_1 + p_2, 0, p_3, \dots, p_n) = \mathsf{H}(p_1 + p_2, p_3, \dots, p_n),$$

ce qui permet de conclure. □

Exercice 2.1. Soit $u \in [0, 1]$. Montrer que pour tout $x \in [0, u]$,

$$-x \ln_2(x) - (u - x) \ln_2(u - x) \leq -u \ln_2(u/2).$$

Exercice 2.2. Prouver que pour tout $1 \leq i < j \leq n$,

$$\mathsf{H}(p_1, \dots, p_i \dots, p_j, \dots, p_n) \leq \mathsf{H}(p_1, \dots, \frac{p_i + p_j}{2}, \dots, \frac{p_i + p_j}{2}, \dots, p_n).$$

2.1 Propriétés de l'entropie conditionnelle

Utilisons ici les notations introduites à la définition 1.7 : soient X qui prend les valeurs $x_1, \dots, x_n$, avec probas respectives $p_1, \dots, p_n$ et Y qui prend les valeurs $y_1, \dots, y_m$, avec probas respectives $q_1, \dots, q_m$. Posons

$$r_{ij} := \mathbb{P}(X = x_i, Y = y_j).$$

Théorème 2.4. *On a*

$$\mathsf{H}(X, Y) \leq \mathsf{H}(X) + \mathsf{H}(Y),$$

avec égalité si et seulement si X, Y *indépendantes.*

Si l'on interprète l'entropie comme de l'incertitude, il est naturel que l'entropie apportée par deux variables est moindre que la somme de leurs entropies, sauf quand elles sont indépendantes.

Preuve du Théorème 2.4. Soient X qui prend les valeurs $x_1, \dots, x_n$, avec probas respectives $p_1, \dots, p_n$ et Y qui prend les valeurs $y_1, \dots, y_m$, avec probas respectives $q_1, \dots, q_m$. Posons

$$r_{ij} := \mathbb{P}(X = x_i, Y = y_j).$$

Par la formule des probabilités totales, on a

$$\mathsf{H}(X) = -\sum_{i,j} r_{ij} \ln_2 p_i \qquad \text{et} \qquad \mathsf{H}(Y) = -\sum_{i,j} r_{ij} \ln_2 q_j.$$

Donc

$$\mathsf{H}(X) + \mathsf{H}(Y) = -\sum_{i,j} r_{ij} \ln_2 p_i q_j = \underbrace{-\sum_{i,j} r_{i,j} \ln_2 \left(\frac{r_{i,j}}{p_i q_j} \right)}_{\geqslant \ln_2(r_{i,j} \sum_{i,j} \frac{p_i q_j}{r_{ij}}) = -\ln_2(1) = 0} - \sum_{i,j} r_{i,j} \ln_2(r_{i,j})$$

$$\geqslant -\sum_{i,j} r_{i,j} \ln_2 (r_{i,j}) = \mathsf{H}(X, Y),$$

avec égalité si et seulement si pour tout i, j, $r_{ij} = p_i q_j$, càd si et seulement si X, Y indépendantes. □

Corollaire 2.5. *a) On a $\mathsf{H}(X, Y) \geq \mathsf{H}(X)$, avec égalité si et seulement si $Y = g(X)$ pour une fonction g.*

b) On a $\mathsf{H}(X|Y) \leq \mathsf{H}(X)$, avec égalité si et seulement si X, Y indépendantes.

c) On a $\mathsf{I}(X; Y) \geq 0$, avec égalité si et seulement si X, Y indépendantes.

Le corollaire est illustré par la figure 6.

Démonstration: Traitons les cas d'égalité :

- Si $\mathsf{H}(X, Y) = \mathsf{H}(X)$, alors $\mathsf{H}(X|Y) = 0$, $\sum_j \mathsf{H}(X|Y = y_j) q_j = 0$, donc pour tout j, $\mathsf{H}(X|Y = y_j) = 0$, X est déterminée par $Y = y_j$, ce qui veut dire que $X = g(y_j)$ pour une certaine fonction g.
- Si $\mathsf{H}(X|Y) = \mathsf{H}(X)$, il faut utiliser le cas d'égalité dans le Théorème 2.4.

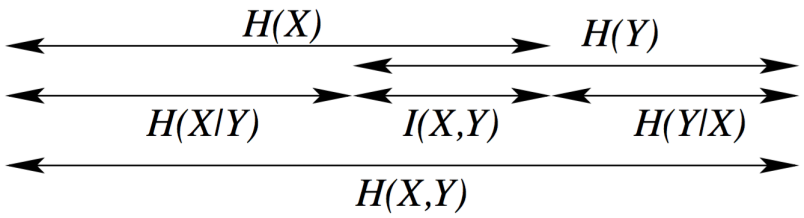


FIGURE 6 – Relations entre les différentes entropies.

□

Remarque 2.6. L'information mutuelle $\mathsf{I}(X;Y)$ entre X et Y correspond ainsi à la diminution d'incertitude sur Y causée par la connaissance de X, c'est-à-dire la quantité d'information sur Y contenue dans X. Elle est, chose non évidente a priori, symétrique en X et Y.

Exemple 2.7 (Alphabet)**.** • Si p est la loi uniforme sur un alphabet de 26 lettres, alors

$$\mathsf{H}(p) = \ln_2 26 \approx 4.7 \text{ bits.}$$

• Néanmoins, si, en considérant l'alphabet latin usuel, on tient compte des fréquences d'usage des lettres (qui dépendent des langues). En français, par exemple, dans l'ordre décroissant,

$$\begin{aligned} p_E =& 0,121 \\ p_A =& 0,0711 \\ p_I =& 0,0659 \\ etc.... & \end{aligned}$$

on obtient

$$\mathsf{H}(\text{Français}) = \sum_{\alpha\in\{A,B,C,\ldots\}} -p_\alpha \ln_2(p_\alpha) \approx 4.14 \text{ bits}$$

et de la même manière,

$$\mathsf{H}(\text{Anglais}) \approx 4.19 \text{ bits.}$$

Ces valeurs diminuent si on tient compte des suites de plusieurs lettres, appelées diagrammes. Par exemple pour les "bigrammes" en français (https:

//www.apprendre-en-ligne.net/crypto/stat/francais.html) :

$$\begin{aligned} p_{ES} &= 3,08\% \\ p_{LE} &= 2,25\% \\ p_{EN} &= 2,20\% \\ etc... & \\ \mathsf{H}(\mathcal{L}_2) &= -\sum_{\alpha,\beta\in\{A,B,C,...\}} -p_{\alpha\beta}\ln_2(p_{\alpha\beta}) \end{aligned}$$

Exercice 2.3. On note $\mathcal{L}_n$ la loi des textes de n lettres dans la langue

1. Montrer que la suite $\mathsf{H}(\mathcal{L}_n)$ est sous-additive, c'est-à-dire que $\mathsf{H}(\mathcal{L}_{n+m}) \leq \mathsf{H}(\mathcal{L}_n) + \mathsf{H}(\mathcal{L}_m)$.
2. En s"appuyant sur le lemme suivant, en déduire que cette limite existe

$$\mathsf{H}(\mathcal{L}) := \lim_{n\to\infty} \frac{\mathsf{H}(\mathcal{L}_n)}{n}.$$

Cette limite est une bonne définition de la langue $\mathcal{L}$.

Lemme 2.8. *si (u_n) est une suite sous additive positive, alors $lim_n u_n/n$ existe et est égal à $\liminf u_n/n$.*

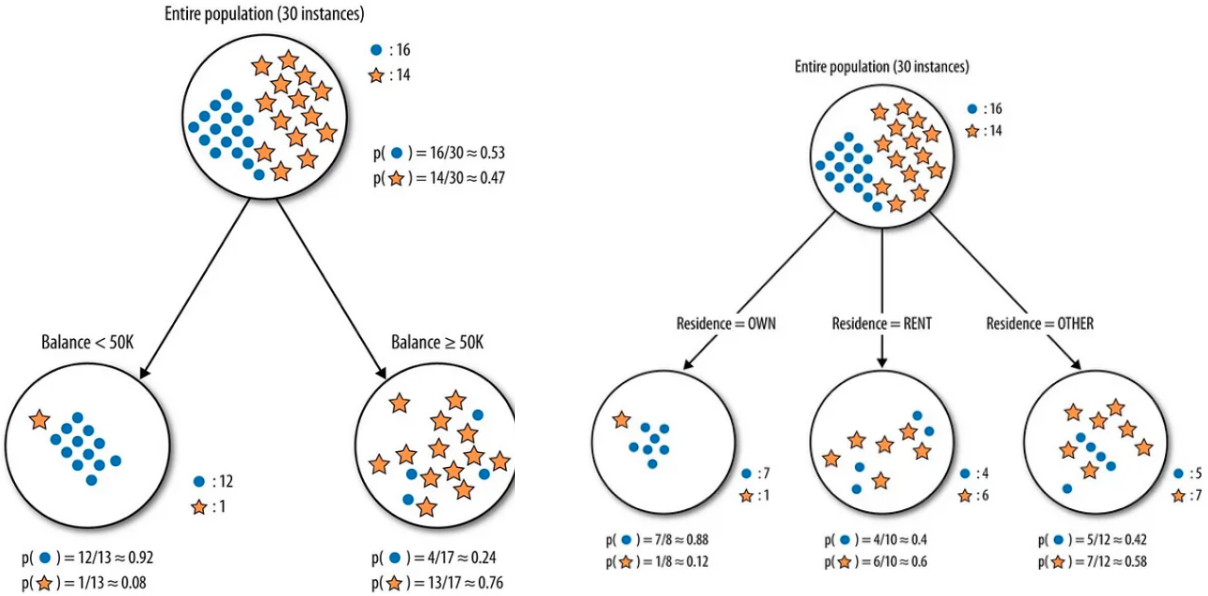
Démonstration. Soit $\alpha = \inf u_n/n$. Fixons $\epsilon > 0$. Il existe p tel que $\alpha < u_p/p < \alpha + \epsilon$. Soit $n \geq p$. On écrit $n = kp + r$, avec $0 \leq r \leq p-1$. On a par récurrence $u_n \leq ku_p + u_r$ par sous-additivité. Donc

$$\alpha \leq u_n/n \leq \frac{k}{kp+r}u_p + u_r/n \leq u_p/p + \frac{1}{n}\max\{u_1,\ldots,u_{p-1}\} \leq \alpha + 2\epsilon,$$

pour n assez grand, ce qui montre le résultat.

□

Exercice 2.4. On considère trente personnes ayant demandé un crédit immobilier, représentées par deux figures différentes.



Pour chaque individu, on sait s'ils ont remboursé leur crédit (étoile jaune) ou pas (rond bleu). On sait grâce à la 1re figure s'ils avaient plus de 50 K $ sur leur compte avant de souscrire le crédit ("Balance>50K") ou pas ("Balance<50K"). Sur la seconde figure, on a une information sur leur résidence actuelle : propriétaire ("OWN"), locataire ("RENT") ou autre ("OTHER"). On tire au hasard uniformément un individu dans cette population, que l'on appelle Z.

1. Que vaut $\mathsf{H}(Z)$?

   Les réponses aux questions suivantes sont, dans le désordre : $0,62$; $0,13$; $0,37$; $0,99$ (à vous de bien les associer et donner des formules littérales exactes).

2. Quelle est l'entropie de la variable $X = 1_{\text{Z rembourse son crédit}}$?
3. Quelle est l'entropie conditionnelle de la variable X par rapport à la variable aléatoire $Y = 1_{\text{Balance>50K}}$?
4. L'information mutuelle entre les variables X et Y, aussi appelé "gain d'information", représente l'information supplémentaire qu'apporte Y quand on connaît X (ou le contraire). Donnez cette valeur.
5. La deuxième figure donne les mêmes informations pour la variable "Residence". Donner l'expression du gain d'information de cette variable par rapport à X.

6. Laquelle de ces deux méthodes vous paraît la plus pertinente pour représenter les données ?

2.2 Distance de Kullback-Leibler discrète

Définition 2.9. L'*entropie relative* ou *distance de Kullback-Leibler* entre deux distributions p et q est définie par :

$$\mathsf{D}(p\|q) := \mathbb{E}_p \ln_2 \frac{p}{q} = \sum_u p(u) \ln_2 \frac{p(u)}{q(u)} = \mathbb{E}_q \frac{p}{q} \ln_2 \frac{p}{q}.$$

Exercice 2.5. Montrez les différentes égalités de la définition de la distance de KL.

Exercice 2.6. Soit $p(a) = .5$, $p(b) = .25 = p(c)$ et q la loi uniforme sur $\{a, b, c\}$. Calculer $\mathsf{H}(p)$, $\mathsf{H}(q)$, $\mathsf{D}(p\|q)$ et $\mathsf{D}(q\|p)$. Vérifier que $\mathsf{D}(p\|q) \neq \mathsf{D}(q\|p)$.

Exercice 2.7. Soient $p, q \in (0, 1)$ et μ, ν lois de Bernoulli de paramètres p, q. Calculer $\mathsf{D}(\mu\|\nu)$. Si $p, q = .5, .25$, a-t-on $\mathsf{D}(\mu\|\nu) = \mathsf{D}(\nu\|\mu)$?

Théorème 2.10. *Soit $p = (p_1, \ldots, p_n) \in \mathcal{P}_n$. Alors la fonction*

$$q \in \mathcal{P}_n \mapsto \mathsf{D}(p\|q)$$

est positive et atteint son minimum 0 en p uniquement.

Par strict concavité du $\ln_2$ on a

$$-\mathsf{D}(p\|q) = \sum p_i \ln_2 \frac{q_i}{p_i} \leq \ln_2 \sum_i p_i \frac{q_i}{p_i} = \ln_2(1) = 0,$$

avec égalité si et seulement si $p = q$, où elle s'annule.

Exercice Soit X une variable à valeurs dans $\{1, \ldots, m\}$, et Y à valeurs dans $\{m + 1, \ldots, n\}$. Soit Z une variable de Bernoulli de paramètre p indépendante de X et Y. Finalement, soit

$$T = \begin{cases} X \text{ si } Z = 1 \\ Y \text{ si } Z = 0. \end{cases}$$

On pose $p_k = \mathbb{P}(X = k), q_k = \mathbb{P}(Y = k)$, $p_X = (p_1, \ldots, p_m, 0, \ldots, 0), p_Y = (0, \ldots, 0, q_{m+1}, \ldots, q_n)$..

1. Donner $\mathbb{P}(T = k)$ pour $1 \leqslant k \leqslant n$, en distinguant les cas $k \leqslant m$ et $k > m$.
2. Montrez que $\mathsf{H}(T) \geqslant p\mathsf{H}(X) + (1-p)\mathsf{H}(Y)$. Quand a-t-on égalité ?
3. Montrer que $\mathsf{H}(T) \leqslant \mathsf{H}(X) + \mathsf{H}(Y) + 1$.
4. Calculer $h(p) := \mathsf{H}(T)$ en fonction de $p, \mathsf{H}(X), \mathsf{H}(Y)$.

La distribution uniforme joue un rôle central.

Théorème 2.11. *On note pour $n \in \mathbb{N}^*$,*

$$p_n = (1/n, \dots, 1/n).$$

On a

- $\mathsf{H}(p) = 0$ *si et seulement si p a toutes ses coordonnées nulles sauf une,*
- $\mathsf{H}(p) = \mathsf{H}(p_n) - \mathsf{D}(p \| p_n)$
- $\mathsf{H}(p) = \ln_2 n$ *si et seulement si $p = p_n$.*

Donc pour tout $p \in \mathcal{P}_n$,

$$0 \leq \mathsf{H}(p) \leq \mathsf{H}(p_n) = \ln_2 n.$$

Le Théorème est illustré par la figure 7.

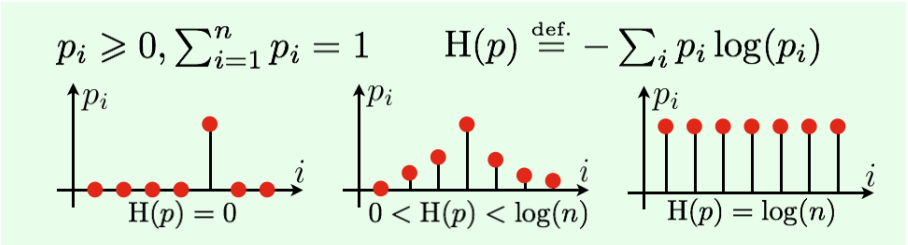


FIGURE 7 – L'entropie peut se comprendre comme une mesure de la *dispersion* (non métrique, les distances $|j - i|$ n'étant pas prises en compte) d'une loi de probabilité.

Preuve du Théorème 2.11. Pour le cas où $\mathsf{H}(p) = 0$, comme $p \log(p) > 0$ pour $p \in]0, 1[$, cela signifie que $p_i \log(p_i) = 0$ pour tous les p_i, ce qui veut dire $p_i = 0$ ou $p_i = 1$.

Vérifions ensuite

$$\mathsf{H}(p_n) = \sum_i \frac{1}{n} \log_2(n) = \log_2(n)$$

$$\mathsf{D}(p\|p_n) = \sum_i p_i \log_2\left(\frac{p_i}{1/n}\right) = \sum_i p_i \log_2(p_i) + (\sum_i p_i)\log_2(n) = \mathsf{H}(p_n) - \mathsf{H}(p)$$

Il en découle les résultats énoncés.

□

Exercice 2.8. Soit (X, Y) un vecteur aléatoire, et X', Y' deux variables aléatoires indépendantes telles que $X' \sim X, Y' \sim Y$. Montrez que

$$\mathsf{D}((X,Y)\|(X',Y')) = \mathsf{I}(X,Y).$$

Démonstration. On a tout simplement

$$\mathsf{I}(X,Y) = \sum_{i,j} r_{i,j} \ln_2\left(\frac{r_{ij}}{p_i q_j}\right)$$

qui correspond à la définition de la distance de KL. □

2.3 Variables discrètes à support infini

On peut définir similairement l'entropie et la distance de KL pour des variables discrètes à support quelconque X, Y. Soit $\mathcal{S}$ dénombrable où X prend ses valeurs, avec $p_x := \mathbb{P}(X = x)$ pour $x \in \mathcal{S}$ et $\sum_{x\in\mathcal{S}} p_x = 1$. On définit

$$\begin{aligned}\mathsf{H}(X) &= -\sum_{x\in\mathcal{S}} p_x \ln_2(p_x)\\ \mathsf{H}(Y|X) &= \sum_{x\in\mathcal{S}} \mathsf{H}(Y \mid X = x)p_x\\ \mathsf{I}(X;Y) &= \mathsf{H}(X) - \mathsf{H}(X|Y) = \mathsf{H}(Y) - \mathsf{H}(Y|X)\end{aligned}$$

Si Y prend des valeurs à l'extérieur de $\mathcal{S}$, $\mathsf{D}(Y\|X) = \infty$, et sinon on pose $q_x := \mathbb{P}(Y = x)$, et

$$\mathsf{D}(Y\|X) := \sum_{x\in\mathcal{S}} q_x \ln_2(q_x/p_x).$$

On a les même propriétés, prouvées comme dans le cadre fini essentiellement avec la concavité du ln, par exemple

- $\mathsf{D}(Y\|X) \geqslant 0$, avec égalité ssi X, Y ont la même loi.

Il n'est par contre pas évident de définir une distribution d'entropie maximale, on verra à la fin de la prochaine section comment traiter cette question.

TD 1 - Propriétés algébriques

Exercice 2.1. Que vaut $\mathsf{I}(X;X)$?

Exercice 2.2. Soient X, Y des variables aléatoires. telles que X est de loi uniforme sur l'ensemble $\{1,2,3,4\}^2$ et Y de loi uniforme sur $\{1,\dots,8\}$ telles que $\mathsf{H}(X|Y) = 4$.

a) Donner $\mathsf{H}(X)$, $\mathsf{H}(Y)$, $\mathsf{H}(X,Y)$, $\mathsf{I}(X;Y)$ et $\mathsf{H}(Y|X)$.

b) A-t-on X et Y indépendantes ?

Exercice 2.3. Soit (X,Y) dont la loi jointe est donnée figure 8. Calculer la

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

FIGURE 8 – Loi de (X,Y)

loi de X, celle de Y, ainsi que $\mathsf{H}(X)$, $\mathsf{H}(Y)$, $\mathsf{H}(X,Y)$, $\mathsf{H}(X|Y)$, $\mathsf{H}(Y|X)$ et $\mathsf{I}(X;Y)$.

Exercice 2.4. Considérons une source $\{a_0, a_1, a_2\}$ qui produit un évènement par unité de temps (par exemple, chaque μsec) selon la distribution de probabilité $p_0 = 0.8$, $p_1 = 0.02$, $p_2 = 0.18$. Supposons que notre flot de données doit passer par un canal qui n'accepte que 0.87 bits par unité de temps en moyenne. Aura-t-on congestion du canal ?

Exercice 2.5. Jean tire avec remise et une à une, des cartes dans un paquet de 13 cartes contenant toutes les cartes d'une couleur jusqu'à ce qu'il obtienne un as. Soit X le numéro du premier tirage où on obtient un as. Jacques ne connait pas la valeur de X et souhaite la déterminer en posant à Jean une suite de questions binaires (réponses oui/non).

1. Quelle est la loi de X ?
2. Trouver n tel que la question "est ce que $X > n$?" soit optimale, c'est-à-dire apporte en moyenne l'information la plus grande. Vous justifierez soigneusement votre réponse.
3. Supposons que Jean ait répondu oui à la question "est ce que $X > n$?", où n est l'entier déterminé à la question 1. Quelle deuxième question doit poser Jacques pour maximiser l'information que donnera la réponse de Jean ?
4. Supposons que Jean ait répondu non à la question "est ce que $X > n$?", où n est l'entier déterminé à la question 2. Quelle deuxième question doit alors poser Jacques pour maximiser l'information que donnera la réponse de Jean ?

Exercice 2.6. Soit X une v.a. prenant ses valeurs dans $\mathcal{X} = \{x_1, \dots, x_n\}$ et soit

$$p : x_i \in \mathcal{X} \mapsto p(x_i) := \mathbb{P}(X = x_i).$$

Prouver que pour tout $u \in (0, 1)$,

$$\mathbb{P}(p(X) \leq u) \ln_2(1/u) \leq \mathsf{H}(X).$$

Exercice 2.7. Soient p, q, u les lois sur l'ensemble $\{1, \dots, 5\}$ définies par

$$p_1 = 1,\ p_2 = \cdots = p_5 = 0, \qquad q_1 = 0.1,\ q_2 = 0.3, q_3 = q_4 = 0.25,\ q_5 = 0.1$$

et $u_1 = \cdots = u_5 = 0.2$. Sans faire le moindre calcul, classer les 3 nombres $\mathsf{H}(p), \mathsf{H}(q)$ et $\mathsf{H}(u)$.

Exercice 2.8. Soit (X, Y) dont la loi jointe est donnée figure 9.
Comparer $\mathsf{H}(X)$, $\mathsf{H}(X|Y = 1)$, $\mathsf{H}(X|Y = 2)$ et $\mathsf{H}(X|Y)$.

Exercice 2.9. On pose, pour $q \in [0, 1]$, $\mathsf{h}(q)$ l'entropie de la loi de Bernoulli de paramètre q. Donner la formule de $\mathsf{h}(q)$. Soit (X, Y) dont la loi jointe est donnée figure 10. Donner, en exprimant les résultats à partir de la fonction h, $\mathsf{H}(X)$, $\mathsf{H}(Y)$, $\mathsf{H}(X|Y)$, $\mathsf{I}(X;Y)$ et $\mathsf{H}(Y|X)$.

backslashboxYX	1	2
1	0	3/4
2	1/8	1/8

FIGURE 9 – Loi de (X, Y)

backslashboxYX	0	1
0	1/4	1/2
1	1/8	1/8

FIGURE 10 – Loi de (X, Y)

Exercice 2.10. 1. A-t-on toujours $\mathsf{H}(X|Y) = \mathsf{H}(Y|X)$?
2. Donner une CNS pour que $\mathsf{H}(X|Y) = \mathsf{H}(Y|X)$.

Exercice 2.11. On lance un dé à 6 faces équilibré.

1. Quelle est l'information mutuelle moyenne entre la face en haut H et la face en bas B ?
2. Quelle est l'information mutuelle moyenne entre la face en haut H et celle, notée C, qui est de côté et tournée vers le joueur ?

Exercice 2.12. Soit X une variable à valeurs dans $\{1, \dots, m\}$, et Y à valeurs dans $\{m+1, \dots, n\}$. Soit Z une variable de Bernoulli de paramètre p indépendante de X et Y. Finalement, soit

$$T = \begin{cases} X \text{ si } Z = 1 \\ Y \text{ si } Z = 0. \end{cases}$$

On pose $p_k = \mathbb{P}(X = k), q_k = \mathbb{P}(Y = k)$, $p_X = (p_1, \dots, p_m, 0, \dots, 0), p_Y = (0, \dots, 0, q_{m+1}, \dots, q_n)$..

1. Donner $\mathbb{P}(T = k)$ pour $1 \leqslant k \leqslant n$, en distinguant les cas $k \leqslant m$ et $k > m$.
2. Montrez que $\mathsf{H}(T) \geqslant p\mathsf{H}(X) + (1-p)\mathsf{H}(Y)$. Quand a-t-on égalité ?
3. Montrer que $\mathsf{H}(T) \leqslant \mathsf{H}(X) + \mathsf{H}(Y) + 1$.
4. Calculer $h(p) := \mathsf{H}(T)$ en fonction de $p, \mathsf{H}(X), \mathsf{H}(Y)$.
5. Trouver p^* où h atteint un extremum.

6. Est-ce un minimum ou un maximum ? Est-il strict ?
7. * Déduisez-en que

$$2^{\mathsf{H}(T)} \leqslant 2^{\mathsf{H}(X)} + 2^{\mathsf{H}(Y)}$$

Exercice 2.13. Soit $X_1, \ldots, X_n$ une suite de n v.a. à valeurs dans $\{0, 1\}$. Soit $R = (R_1, R_2, \ldots)$ la suite des longueurs de valeurs identiques dans $X_1, \ldots, X_n$ (par exemple, si $(X_1, \ldots, X_n) = 0001100100$, alors $R = (3, 2, 2, 1, 2)$). Comparer les nombres

$$\mathsf{H}(X_1, \ldots, X_n), \quad \mathsf{H}(R), \quad \mathsf{H}(X_1, \ldots, X_n, R).$$

Exercice 2.14. a) Soient $X_1, \ldots, X_n$ vaiid de loi $\frac{\delta_0+\delta_1}{2}$. Calculer $\mathsf{H}(X)$ de 2 façons différentes pour $X = (X_1, \ldots, X_n)$.

b) Soient $Y_1, \ldots, Y_n$ vaiid de loi $(1-p)\delta_0 + p\delta_1$. Calculer $\mathsf{H}(Y)$ pour $Y = (Y_1, \ldots, Y_n)$.

c) Prouver que $\mathsf{H}(X) \geq \mathsf{H}(Y)$. Commenter.

TD 2 - Partiel 22

Master 1ère année, 2022-2023, Théorie de l'information

Partiel - Novembre 2022

Durée : 2. Documents, téléphone et appareils électroniques interdits.

Exercice 2.1. Soit X une variable à valeurs dans $\{1, \dots, m\}$, et Y à valeurs dans $\{m+1, \dots, n\}$. Soit Z une variable de Bernoulli de paramètre p indépendante de X et Y. Finalement, soit

$$T = \begin{cases} X \text{ si } Z = 1 \\ Y \text{ si } Z = 0. \end{cases}$$

On pose $p_k = \mathbb{P}(X = k), q_k = \mathbb{P}(Y = k)$, $p_X = (p_1, \dots, p_m, 0, \dots, 0), p_Y = (0, \dots, 0, q_{m+1}, \dots, q_n)$..

1. Donner $\mathbb{P}(T = k)$ pour $1 \leqslant k \leqslant n$, en distinguant les cas $k \leqslant m$ et $k > m$.
2. Montrez que $\mathsf{H}(T) \geqslant p\mathsf{H}(X) + (1-p)\mathsf{H}(Y)$. Quand a-t-on égalité ?
3. Montrer que $\mathsf{H}(T) \leqslant \mathsf{H}(X) + \mathsf{H}(Y) + 1$.
4. Calculer $h(p) := \mathsf{H}(T)$ en fonction de $p, \mathsf{H}(X), \mathsf{H}(Y)$.
5. Trouver p^* où h atteint un extremum.
6. Est-ce un minimum ou un maximum ? Est-il strict ?
7. * Déduisez-en que

$$2^{\mathsf{H}(T)} \leqslant 2^{\mathsf{H}(X)} + 2^{\mathsf{H}(Y)}$$

Exercice 2.2. On va étudier dans cette exercice l'entropie de variables aléatoires à valeurs dans $\mathbb{N}$. Pour X une telle variable, on définit l'entropie de X par

$$\mathsf{H}(X) = \sum_{k=0}^{\infty} (-p_k \ln(p_k)) \in \mathbb{R}_+ \cup \{+\infty\}, k \in \mathbb{N},$$

où $p_k = \mathbb{P}(X = k)$, avec la convention $0 \ln(0) = 0$.

1. Pourquoi $\mathsf{H}(X)$ est-elle toujours bien définie ?
2. Montrer qu'il existe $C > 0$ finie telle que $p_k = \frac{C}{(k+1)\ln(k+2)^2}$ définisse bien une loi de probabilité. Montrer que $\mathsf{H}(X) = \infty$, où $p_k = \mathbb{P}(X = k)$.
   Rappel sur les séries de Bertrand :

   $$\sum_{k=2}^{\infty} \frac{1}{k \ln(k)^{\alpha}} < \infty$$

   ssi $\alpha > 1$.
3. Soit X une variable géométrique de paramètre $p \in]0, 1]$, on note $q = 1 - p$.
   (a) Donner $p_k = \mathbb{P}(X = k)$ et $\mathbb{E}(X)$ (inutile de justifier).
   (b) Donner $h(p) := \mathsf{H}(X)$ en fonction de p.
   (c) Que vaut $\lim_{p\to 0} h(p)$? $\lim_{p\to 1} h(p)$? Cela vous semble-t-il cohérent avec le concept d'entropie ?
4. Soit Y une autre variable aléatoire à valeurs dans $\mathbb{N}$, $q_k := \mathbb{P}(Y = k)$. On définit la distance de Kullback Leibler par

   $$\mathsf{D}(Y\|X) = \sum_k \ln_2 \left(\frac{q_k}{p_k}\right) q_k.$$

   Montrer que $\mathsf{D}(Y\|X) \geqslant 0$, et que l'on a égalité ssi Y et X ont la même loi.
5. Soit Y à valeurs dans $\mathbb{N}^*$. Soit $X \sim \mathcal{G}(1/\mathbb{E}(Y))$. Montrer que $\mathsf{D}(Y\|X) = \mathsf{H}(X) - \mathsf{H}(Y)$.
6. Soit $m > 0$. Quelles sont les variables X telles que $\mathbb{E}(X) = m$ qui maximisent l'entropie ?

3 Entropie continue, et TCL entropique

3.1 Entropie de variables aléatoires à densité

Nous allons dire un mot de l'entropie de distribution à densité. En effet, nous voulons énoncer des résultats de convergence d'entropie de somme de variables aléatoires comme elles apparaissent dans le TCL, vers l'entropie d'une Gaussienne. Il n'y a pas vraiment de sens à considérer la convergence de $\mathsf{H}(\frac{1}{\sqrt{n}}S_n)$ dans un cadre discret, puisque dans ce cadre, $\mathsf{H}(\frac{1}{\sqrt{n}}S_n) = \mathsf{H}(S_n)$. Toutes les densités que l'on considère sont continues, sauf éventuellement en un nombre fini de points.

Définition 3.1. Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$, et ayant une densité f_X. On définit l'entropie H de X par

$$\mathsf{H}(X) = -\int \ln(f_X(x)) f_X(x) dx = -\mathbb{E}(\ln(f_X(X)) \in \bar{\mathbb{R}}.$$

$\mathsf{H}(X)$ ne dépendant que de la distribution de X, on note aussi $\mathsf{H}(X) = \mathsf{H}(f_X)$. Il faut faire la supposition que

$$\int f(x) |\ln(f(x))| dx < \infty$$

pour que cette quantité soit bien définie, cette hypothèse est très faible et sera implicite dans de nombreux résultats.

Exercice 3.1. Soit X une variable de Cauchy. Montrer que $\exp(X)$ n'admet pas d'entropie.

Exercice 3.2.

1. Calculer l'entropie d'une Gaussienne de moyenne m et de variance σ^2.
2. Soit (X, Y) un vecteur Gaussien de dimension 2 de matrice de covariance Σ. Calculer $\mathsf{H}(X, Y)$ dans le cas où X et Y sont indépendantes. On rappelle la densité de (X, Y) :

$$f(x, y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp(-\frac{1}{2}(x, y)\Sigma^{-1}\begin{pmatrix} x \\ y \end{pmatrix})$$

Remarque 3.2. $\mathsf{H}(X) \to -\infty$ pour $\sigma \to 0$. La possibilité d'entropies négatives est une différence majeure avec le cas discret et montre qu'on ne peut pas faire les mêmes interepretations.

Exercice 3.3. Calculer l'entropie d'une loi uniforme sur $[a, b]$.

Exercice 3.4. Calculer l'entropie d'une loi v.a. X de loi Gamma de paramètre $(m, 1)$, en fonction de $\psi(m) = \frac{1}{\Gamma(m)} \int_0^\infty x^{m-1} e^{-x} \ln(x)\, dx$. On rappelle que la densité de X est $f_X(x) = \frac{1}{\Gamma(m)} x^{m-1} e^{-x} 1_{x>0}$.

Proposition 3.3. *Soit X une variable aléatoire sur $\mathbb{R}^d$, à densité f_X et $\alpha \in \mathbb{R}^d$. Alors*

$$\mathsf{H}(X) = \mathsf{H}(X + \alpha).$$

Exercice 3.5. Soit $\alpha \neq \pm 1$, et X une Gaussienne de moyenne m et de variance σ^2. Montrer que $\mathsf{H}(X) \neq \mathsf{H}(\alpha X)$.

Exercice 3.6. Soit $\alpha \neq 1$, et X une variable aléatoire réelle à densité admettant une entropie finie. Exprimer $\mathsf{H}(\alpha X)$ en fonction de $\mathsf{H}(X)$. Est-il possible d'avoir $\mathsf{H}(\varphi(X)) > \mathsf{H}(X)$ pour une fonction mesurable φ ?

Définition 3.4. Soit X et Y deux variables aléatoires à valeurs dans $\mathbb{R}^m$ et $\mathbb{R}^n$, admettant pour densité respective f_X et f_Y, et pour densité jointe $f_{X,Y}$. On définit l'entropie de X conditionnellemnt à Y par

$$\mathsf{H}(X|Y) = -\int_{\mathbb{R}^n} f_Y(y) \int_{\mathbb{R}^m} f_{X|Y=y}(x) \ln f_{X|Y=y}(x)\, dx\, dy,$$

où $f_{X|Y=y}(x) = f_{X,Y}(x, y)/f_Y(y)$. On note aussi $f_{X|Y=y}(x) = f_{X|Y}(x|y) = f(x|y)$.

Proposition 3.5. *Soit X et Y deux variables aléatoires à densité. Alors*

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

Démonstration. basée sur

$$\ln(f(x, y)) = \ln(f(x)) + \ln(f(y|x))$$

ça donne

$$\mathsf{H}(X, Y) = -\int \ln(f(x)) f(x, y) dx dy - \int \ln(f(y|x)) f(y|x) f(x) dx dy = -\mathsf{H}(X) - \mathsf{H}(Y|X)$$

□

Exercice 3.7. Soit A une matrice inversible et $u \in \mathbb{R}^m$, X une variable a densité sur $\mathbb{R}^m$ admettant une entropie. Que vaut $\mathsf{H}(AX+u)$? Retrouver le résultat précédent ?

Exercice 3.8. Montrer qu'il existe des variables continues dont l'entropie n'est pas approchable par l'entropie de variables discrètes : il existe X à densité telle que $\mathsf{H}(X_n)$ ne converge pas vers X quelle que soit la suite $(X_n)_n$ de VA discètes.

3.2 Distance de Kullback continue

Définition 3.6. Soit X et Y deux variables aléatoires à valeurs dans $\mathbb{R}^n$ et $\mathbb{R}^m$, de densités respectives f_X, f_Y. On définit (quand c'est possible) la Distance de Kulback de X à Y par

$$\mathsf{D}(X\|Y) = \mathbb{E}\ln(\frac{f_X(X)}{f_Y(X)}) = \int \ln(f_X(x)/f_Y(x)) f_X(x).$$

Cette quantité ne dépend que des distributions f_X et f_Y. On note aussi $\mathsf{D}(X||Y) = \mathsf{D}(f_X||f_Y)$.

Remarque 3.7. On a les conventions $0\ln(0/0) = 0$ et $1\ln(1/0) = \infty$, donc s'il existe A non-négligeable tel que $f_Y > 0$ sur A et $f_X = 0$ sur A, alors $\mathsf{D}(X\|Y) = +\infty$.

Autrement dit, $\mathsf{D}(X\|Y) = \infty$ si la loi de Y n'est pas absolument continue par rapport à la loi de X.

Théorème 3.8. *Soit $\sigma^2 > 0$ et g la densité gaussienne centrée de variance σ^2. Soit f, h des autres densités centrées de variance σ^2. Alors*

- $\mathsf{D}(f\|g) = \mathsf{H}(g) - \mathsf{H}(f)$.
- $\mathsf{D}(h\|f) \geqslant 0$.
- $\mathsf{D}(h\|f) = 0$ *ssi* $f = h$ *p.p.*

La gaussienne maximise donc l'entropie des variables centrées de variance σ^2.

Exercice 3.9.

1. Montrer que
$$\mathsf{D}(f||g) = \mathsf{H}(g) - \mathsf{H}(f).$$
2. Conclure grâce au

Théorème 3.9 (inégalité de Jensen)**.** *Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$ et ϕ une application convexe définie sur $\mathbb{R}^d$. On suppose de X et $\phi(X)$ admettent un moment d'ordre 1. On a alors*

$$\phi(\mathbb{E}(X)) \leq \mathbb{E}(\phi(X)).$$

Si ϕ est strictement convexe, alors l'inégalité est une égalité si et seuelement si X est presque surement une constante.

Exercice 3.10. Calculer la distance de Kullback entre deux gaussiennes $\mathcal{N}(m_1, \sigma_1^2)$ et $\mathcal{N}(m_2, \sigma_2^2)$.

3.3 Convergence de variables aléatoires

Si l'on a des variables (continues ou discrètes) $X_n, n \geqslant 1$ et une variable X telles que $\mathsf{D}(X_n \| X) \to 0$, est-ce que cela implique que $X_n \to X$ en loi ?

Remarque 3.10. Si Y et X sont de même loi avec $Y \neq X$ p.s., $\mathsf{D}(X\|Y) = 0$. Cela nous rappelle que la distance entre lois ne peut impliquer de la convergence presque sûre.

Définition 3.11. On définit pour des VA (discrètes ou à densité),

$$d_{TV}(X, Y) = \sup_A |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|.$$

Exercice 3.11. Montrer que pour des variables discrètes de distributions $(p_i)_i$ et $(q_i)_i$,

$$d_{TV}(X, Y) = \frac{1}{2} \sum_i |p_i - q_i|$$

et pour des variables continues de densités f et g,

$$d_{TV}(f, g) = \frac{1}{2} \int |f(x) - g(x)| dx.$$

Théorème 3.12 (Inégalité de Pinsker)**.** *On a pour des variables continues X, Y*

$$\sqrt{\mathsf{D}(X\|Y)} \geqslant \sqrt{\ln(2)} d_{TV}(X, Y).$$

et pour des variables discrètes X, Y

$$\sqrt{\mathsf{D}(X\|Y)} \geqslant d_{TV}(X,Y).$$

On a donc les implications

$$\mathsf{D}(X_n\|X) \to 0 \ \Rightarrow X_n \xrightarrow[n\to\infty]{d_{VT}} X \Rightarrow X_n \xrightarrow[n\to\infty]{\mathcal{L}} X.$$

Démonstration. Montrons d'abord l'inégalité.

Exercice 3.12. 1. Prouver l'inégalité dans le cas de variables de Bernoulli. On pourra introduire la fonction

$$f(x) = p\ln x + (1-p)\ln(1-x).$$

2. Montrer que pour tout A, pour des variables continues X, Y

$$\mathsf{D}(X\|Y) \geqslant \ln(2)\mathsf{D}(\mathbf{1}_{\{X\in A\}}\|\mathbf{1}_{\{Y\in A\}})$$

et pour des variables discrètes

$$\mathsf{D}(X\|Y) \geqslant \mathsf{D}(\mathbf{1}_{\{X\in A\}}\|\mathbf{1}_{\{Y\in A\}})$$

On pourra décomposer $f = f_A p_A + f_{A^c} p_{A^c}$ où $f_A = \mathbf{1}_{\{A\}} f(\int_A f)^{-1}$.

3. Conclure

Si $\mathsf{D}(X_n\|X) \to 0$, alors pour tout A,

$$\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A).$$

En particulier avec $A = [t, \infty[$ pour $t \in \mathbb{R}$, on en déduit la convergence de la fonction de répartition de X_n, $F_{X_n}(t) = \mathbb{P}(X_n \geqslant t)$ vers celle de X, et donc la convergence en loi. □

3.4 Maximisation de l'entropie continue

Exercice 3.13. Soit $K = [0,1]^d$ et f une densité sur K. Montrez que l'entropie de f est plus petite que celle de la loi uniforme sur K.

Exercice 3.14. Quelle variable aléatoire X sur $\mathbb{N}$ maximise l'entropie

$$\mathsf{H}(X) = -\sum_{k=0}^{\infty} \mathbb{P}(X=k)\ln(\mathbb{P}(X=k))?$$

sous la contrainte $\mathbb{E}(X) = 1$.

Principe de maximisation d'entropie ou *Pourquoi modélise-t-on toujours tout avec une loi gaussienne ?* [2] Le principe général de maximum d'entropie, énoncé par exemple par Jaynes dans les années 1950, consiste à dire que la loi de probabilité à choisir parmi un ensemble de lois possibles lors d'une modélisation doit être celle qui maximise l'entropie. Les contraintes proviennent alors de l'information dont on dispose.

Théorème 3.13. *Soit $\eta : A \subset \mathbb{R}^d \to \mathbb{R}$ une fonction mesurable telle que*

$$Z_\eta := \int_A \exp(-\eta(x))dx < \infty$$

et soit alors la densité correspondante

$$f_\eta(x) := \frac{1}{Z_\eta} \exp(-\eta(x)), x \in A.$$

Soit f une densité telle que $\mathsf{H}(f)$ soit bien définie et telle que $f(x)\eta(x)$ soit intégrable sur A. Alors pour tout $a \in \mathbb{R}$, sous la contrainte

$$\int_A f(x)\eta(x)dx = \int f_\eta(x)\eta(x)dx =: a,$$

on a

$$\mathsf{H}(f) \leqslant \mathsf{H}(f_\eta) = (a + \ln(Z_\eta)),$$

il n'y a égalité que si $f = f_\eta$ presque partout.

Démonstration.

Exercice 3.15. Montrez $\mathsf{H}(f_\eta) - \mathsf{H}(f) = \mathsf{D}(f\|f_\eta)$ et prouver le théorème. □

Exemple 3.14 (Les variables positives de moyenne fixée de plus grande entropie sont les exponentielles)**.** Soit $\eta(x) = \lambda x \mathbf{1}_{\{x>0\}}$, $\lambda \in \mathbb{R}_+$. La densité $f_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ qui maximise strictement l'entropie $\mathsf{H}(f)$ sous la contrainte

$$\int_0^\infty x f(x)dx = \lambda < \infty$$

est la loi exponentielle $\mathcal{E}(\lambda)$ de paramètre λ, de densité

$$f_\lambda(x) = \lambda \exp(-\lambda x),$$

avec

$$\mathsf{H}(f_\lambda) = \lambda + \ln(Z_\lambda) = \lambda - \ln(\lambda).$$

Dans $\mathbb{R}^d$ on pose $\eta(x) = \lambda\|x\|_1$ et on parle de lois de Laplace (ou double-exponentielle).

Exemple 3.15 (Les variables centrées réduites de plus grande entropie sont les gaussiennes)**.** Soit $\sigma > 0$,. La densité f_σ centrée qui maximise strictement l'entropie $\mathsf{H}(f)$ sous la contrainte

$$\int x^2 dx = \sigma^2$$

est la loi gaussienne $\mathcal{N}(0, \sigma^2)$ de densité

$$f_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

On retrouve

$$\mathsf{H}(f_\sigma) = \frac{1}{2} - \ln(\sqrt{2\pi\sigma^2}).$$

Exercice 3.16. Dans quelles classes des variables de loi de type Gamma maximisent-elles l'entropie ? Retrouver le cas de l'exponentielle.

Exercice 3.17. Sous quelles contraintes les vecteurs gaussiens maximisent-ils l'entropie ?

D'après le principe de Jaynes, sous contrainte de support, on considérera une loi uniforme, sous contrainte de moyenne, on choisira une loi exponentielle, sous contrainte de variance, on considèrera une loi gaussienne, etc...

Il y a une interprétation du théorème central limite via l'entropie. Pour des variables $X_1, \ldots, X_n$ iid centrées on note

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

On sait que

Théorème 3.16 (Théorème Central Limite)**.** *Si les variables iid centrée X_i sont L^2, alors U_n converge en loi vers $\mathcal{N}(0, 1)$.*

Il existe une preuve de ce résultat grâce à l'entropie, en montrant notamment que l'entropie est croissante :

Théorème 3.17. *Pour tout* $n, \mathsf{H}(S_n) \leqslant \mathsf{H}(S_{n+1})$*, et*

$$\lim_n \uparrow \mathsf{H}(S_n) = \mathsf{H}(\mathcal{N}(0, \sigma^2)).$$

La preuve de ce résultat est difficile, on peut en trouver une version relativement courte (et un historique) dans *Monotonicity of entropy and Fisher information : a quick proof via maximal correlation*, de T. Courtade, *Communications in Information and Systems Volume 16, Number 2, 111-115, 2016.*

On en déduit cependant le TCL :

$$\mathsf{D}(H_n \| \mathcal{N}(0, \sigma^2)) \to 0,$$

Exercice 3.18 (Exam 2022)**.** Soit $\mu_1 \in \mathbb{R}^d, \mu_2 \in \mathbb{R}^d$, et Σ_1, Σ_2 deux matrices symétriques définies positives d'ordre d. Soit X_1 le vecteur gaussien de loi $\mathcal{N}(\mu_1, \Sigma_1)$, et X_2 le vecteur gaussien de loi $\mathcal{N}(\mu_2, \Sigma_2)$. Le but de l'exercice est de montrer que la distance de Kullback-Leibler est

$$\mathsf{D}(X_1 \| X_2) = \frac{1}{2} \ln |\det \Sigma_1 \Sigma_2^{-1}| - \frac{d}{2} + \frac{1}{2} \mathrm{Trace}(\Sigma_1 \Sigma_2^{-1}) + \frac{1}{2} \langle \mu_2 - \mu_1, \Sigma_2^{-1}(\mu_2 - \mu_1) \rangle.$$

1. Commençons par la dimension $d = 1$.
   (a) Donner f_1 et f_2 en dimension 1.
   (b) Montrer qu'en dimension 1, pour $i = 1, 2$,
   $$\int f_1 \ln(f_i) = -\frac{1}{2} \ln(2\pi) - \ln(\sigma_i) - \frac{1}{2\sigma_i^2}(\sigma_1^2 + (\mu_1 - \mu_2)^2),$$
   où $\sigma_i^2 = \mathrm{Var}(X_i)$
   (c) En déduire la formule demandée.
2. Retour en dimension quelconque. Rappeler la densité f_1 de X_1 en dimension quelconque. Donnez l'expression de $\ln(f_1(x))$ pour $x \in \mathbb{R}^d$.
3. On suppose tout d'abord $\mu_1 = \mu_2 = 0$.

(a) Montrez que pour toute matrice A,

$$\mathbb{E}(\langle X_1, AX_1\rangle) = \text{Trace}(\Sigma_1 A).$$

(b) Déduisez-en $\mathsf{D}(X_1 \| X_2)$. Qu'obtient-on si $\Sigma_1 = \Sigma_2$?

4. Retour au cas général (μ_1 et μ_2 sont quelconques). Montrez que

$$\mathbb{E}(\langle X_1 - \mu_1, A(X_1 - \mu_1)\rangle) = \text{Trace}(\Sigma_1 A).$$

5. Déduisez-en la formule dans le cas général.

4 Codage sans perte

Le but de ce chapitre est le suivant : étant donné une source finie $\mathcal{X}$, et un alphabet $\mathcal{D}$, on souhaite coder des éléments de $\mathcal{X}$ par des suites d'éléments de $\mathcal{D}$, de manière à minimiser la longueur des mots une fois codés. Par exemple, $\mathcal{X} = \{a, b, c, \ldots, z\}$, $\mathcal{D} = \{0, 1\}$, chaque code sera alors un nombre binaire, i.e. un élément de $\mathcal{W} = \{0, 1, 00, 01, 10, 11, 000, 001, ...\}$. Un exemple élémentaire de code

$$C : x \in \mathcal{X} \mapsto \mathcal{W}; a \mapsto 0, b \mapsto 1, c \mapsto 00, ...$$

constitue une possibilité. On note $\mathcal{D}^* = \mathcal{D}$ l'ensemble de toutes les suites d'éléments de $\mathcal{D}$.

Afin de minimiser les longueurs des codes des messages composés sur $\mathcal{X}$, il faut se munir d'une fréquence d'apparition des différents éléments, i.e. une distribution $p = (p(u); u \in \mathcal{X})$ dont la somme vaut 1. On peut dès lors adapter le code C et attribuer les codes les plus courts aux éléments les plus fréquents afin de diminuer la longueur moyenne.

Dans ce cours, $\mathcal{D}$ sera uniquement $\{0, 1\}$, comme précédemment, mais on peut imaginer des codes ternaires, i.e. constitués de $0, 1, 2$, ou p-aire pour n'importe quel p. Les résultats énoncés pour les codes binaires se généralisent sans difficultés aux codes p-aires, mais sont un peu plus fastidieux en termes de notation.

Définition 4.1. Soient $\mathcal{X}$ et $\mathcal{D}$ deux ensembles finis. Un *code de la source* $\mathcal{X}$ *dans* $\mathcal{D}^*$ est une fonction $C : \mathcal{X} \to \mathcal{D}^*$. Si $\mathcal{X}$ est muni d'une loi de probabilité p, la *longueur moyenne* de C est alors la moyenne

$$L_p(C) = \sum_{u \in \mathcal{X}} l(C(u))p(u)$$

.

Remarque 4.2. Notons qu'une telle fonction admet toujours une extension, par concaténation, à $\mathcal{X}^*$.

Exemple 4.3. Si $\mathcal{X} = \{a, b, c, d\}$, muni de $p = (.5, , .25, .125, .125)$, $D = 2$ et $C(a) = 0$, $C(b) = 10$, $C(c) = 110$ et $C(d) = 111$, alors $\mathsf{H}(p) = 1.75$ bits $= L_p(C)$.

Définition 4.4. Un code C est dit :

- *non ambigu* si C est injectif sur $\mathcal{X}$,
- *uniquement décodable* si son extension à $\mathcal{X}^*$ est injective,
- *instantané* ou *préfixe* si aucun mot $C(u)$, $u \in \mathcal{X}$ n'est préfixe d'un mot $C(v)$, $v \in \mathcal{X}\backslash\{u\}$.

Exercice 4.1. Parmi les codes suivants sur $\{0, 1\}$,

u	$C_1(u)$	$C_2(u)$	$C_3(u)$
a	0	10	0
b	010	00	10
c	01	11	110
d	10	110	111

dire lesquels sont non ambigus, uniquement décodables et instantanés. Quelles sont leurs longueurs moyenne si les lettres sont équiprobables ?

Les représenter sous forme d'arbre binaire s'ils sont préfixe.

Remarque 4.5. Un code instantané est uniquement décodable. De plus un code instantané peut être décodé sans référence aux mots code future puisque la fin d'un mot code est reconnaissable immédiatement.

Remarque 4.6. On a instantané $\implies$ uniquement décodable $\implies$ non ambigu.

Remarque 4.7. Un code instantané peut-être codé par un arbre, ou chaque code s'apparente à un chemin fini descendant partant de la racine.

Remarque 4.8. Il y a une équivalence entre les codes binaires uniquement décodables et les jeux de questions-réponse. Considérons un jeu où un mot w est tiré au hasard avec une probabilité p sur $\mathcal{X}^*$. Le jeu consiste à poser un

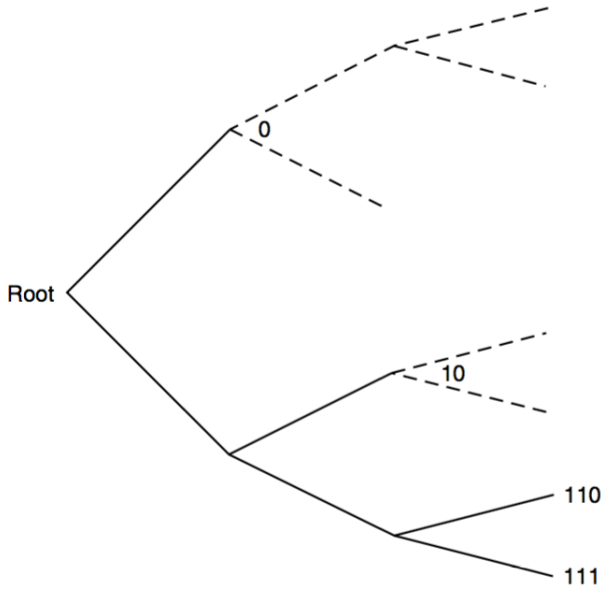


FIGURE 11 – Arbre associé à $\mathcal{X} = \{a, b, c, d\}$ avec $C(a) = 0, C(b) = 10, C(c) = 110 C(d) = 111$.

minimum de questions binaires (la réponse est "oui" ou "non") pour deviner w. On suppose donc qu'on a établi un "arbre" de questions à explorer en fonction des réponses qui permettent de déterminer le mot. Associons à un mot w la suite de réponses "oui,oui,non,oui,..." qui permet de le déterminer codée par sa suite binaire 1101..., ce qui fournit bien un code binaire via $C(w) = 1101....$ Le fait que le code soit instantané et uniquement décodable est impliqué par la propriété élémentaire que la même suite de réponses mène au même mot. Réciproquement, la suite de questions à poser est "Est-ce que le premier bit de $C(w)$est 1, est-ce que le second bit est 1, ..."

Dans la suite de ce cours, on se concentrera sur les codes binaires, i.e. on pose $\mathcal{D} = \{0, 1\}$. La plupart des résultats, et le code de Huffman, s'étendent à D quelconque, voir par exemple Thomas M. Cover, Joy A. Thomas *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing) (2006).

Exercice 4.2. Soit $\mathcal{D} = \{0, 1\}$. Pour chaque $m \geq 2$, donner le nombre $\ell \geq 1$

minimal tel qu'il existe $\mathcal{W} = \{w_1, \ldots, w_m\}$ un ensemble de m mots de $\mathcal{D}^*$ de même longueur ℓ (ce code est forcément préfixe, et donc UD).

Exercice 4.3. On considère l'ensemble source $\{1, \ldots, 8\}$ muni de la probabilité p et des deux codes C_1 et C_2 présentés à la figure 12.

1. Que doit-on vérifier pour pouvoir affirmer que ces deux codes sont instantanés. Le sont-ils ?
2. Comparer les longueurs moyennes des deux codes. Commenter.

x	$p(x)$	$w_1(x)$	$w_2(x)$
1	1/2	000	0
2	1/4	001	10
3	1/8	010	110
4	1/16	011	1110
5	1/32	100	11110
6	1/64	101	111110
7	1/128	110	1111110
8	1/128	111	11111110

FIGURE 12 – Ensemble source, probabilités et codes

4.1 Inégalité de Kraft - McMillan

Le théorème suivant peut se démontrer grâce à la notion d'arbre associé à un ensemble de mots (voir figure 11), mais on en donne une démonstration plus élémentaire.

Théorème 4.9 (Inégalité de Kraft - McMillan)**.** • *Soit* $\mathcal{W} = \{w_1, \ldots, w_m\}$ *un ensemble de* m *mots de* $\{0,1\}^*$ *uniquement décodable, de longueurs respectives* $l_1, \ldots, l_m$*. Alors on a*

$$\sum_i \frac{1}{2^{l_i}} \leq 1.$$

• *Réciproquement, pour toute suite* $l_1, \ldots, l_m$ *telle que*

$$\sum_i \frac{1}{2^{l_i}} \leq 1,$$

il existe un ensemble $\mathcal{W} = \{w_1, \ldots, w_m\}$ *de* m *mots de* $\{0,1\}^*$ *dont aucun n'est préfixe d'un autre, de longueurs respectives* $l_1, \ldots, l_m$*. En particulier, le code est UD.*

Faisons quelques exercices avant la preuve.

Exercice 4.4. a) Parmi les codes suivants sur $\{0,1\}$,

u	$C_1(u)$	$C_2(u)$	$C_3(u)$
a	1	11	01
b	110	00	00
c	01	10	100
d	10	100	111

dire lesquels sont non ambigus et lesquels sont instantanés (en expliquant pourquoi).

b) Pour lesquels de ces codes l'inégalité de McMillan est-elle vérifiée par les mots images de a, b, c, d ?

c) Existe-t-il un code instantané pour lequel les longueurs des images de a, b,c, d soient les mêmes que pour C_2 ? Si oui, le donner.

d) Décoder 01100100010011101 pour C_3.

e) On souhaite étendre le code C_3 en ajoutant une valeur $C_3(e)$ pour une nouvelle lettre e.

(i) Donner la longueur minimale $\ell_{\min}$ à donner à $C_3(e)$ pour que l'inégalité de Kraft reste valable.

(ii) Existe-t-il un mot w sur $\{0,1\}$ de longueur $\ell_{\min}$ tel que en posant $C_3(e) = w$, le code C_3 reste instantané ?

(iii) Étendre C_3 en ajoutant une valeur $C_3(e)$ (de longueur aussi petite que possible) de façon à maintenir le code instantané.

On va faire la preuve de l'inégalité sous forme d'exercice dans le cas où le code est préfixe.

Exercice 4.5. On note $n = \max_i l_i$ la longueur maximale. On note T_n l'ensemble de tous les mots / nombres binaires de longueur n (i.e. l'ensemble de tous les nombres/mots binaires de longueur n). Pour chaque i, on appelle D_{w_i} le sous-ensemble de T_n contitué de tous les mots commençant par w_i.

1. Montrer que les D_{w_i} sont disjoints : $\mathcal{W}_i \cap \mathcal{W}_j = \emptyset$ pour $i \neq j$.
2. Donner le cardinal de chaque D_{w_i}.

3. En déduire l'inégalité.
4. Réciproquement, étant donné une suite de longueurs $l_i, 1 \leqslant i \leqslant m$ qui vérifient cette inégalité, donner un code préfixe ayant ces longueurs. (Indice : par récurrence)

Exercice 4.6. Pour chaque $m = 2, 3, 4, 5, 6$, trouver $\mathcal{W} = \{w_1, \ldots, w_m\}$ un ensemble de m mots de $\mathcal{D}^*$ tels que aucun n'est préfixe d'un autre. On essaiera de le choisir optimal, càd tels que

$$\sum_{i=1}^{m} \frac{1}{2^{l(w_i)}}$$

(qui doit être ≤ 1) dépasse strictement 1 dès que la longueur d'un des mots est réduite.

Preuve de l'inégalité de McMillan pour un code UD général (pas forcément préfixe).

Démonstration. Soit $L := \max_u l(C(u))$. Soit $k \geq 1$. La preuve est basée sur la considération de toutes les phrases de k mots, i.e. de la forme $u_1 \ldots u_k$, où les $u_i \in \mathcal{W}$.

On a

$$\begin{aligned}
\left(\sum_{u\in\mathcal{X}} 2^{-l(C(u))}\right)^k &= \sum_{u_1\in\mathcal{X}} \cdots \sum_{u_k\in\mathcal{X}} 2^{-l(C(u_1))-\cdots-l(C(u_k))} \\
&= \sum_{u_1,\ldots,u_k\in\mathcal{X}} 2^{-l(C(u_1\cdots u_k))} \\
&= \sum_{m=1}^{kL} \underbrace{\left|\{(u_1,\ldots,u_k)\in\mathcal{X}^k\,;\, l(C(u_1\cdots u_k))=m\}\right|}_{=:A_m} 2^{-m}
\end{aligned}$$

le nombre de termes est borné par kL car il est impossible de faire un code de longueur $> kL$ avec k mots vu que chaque code a L lettres ou moins.

Il y a au plus 2^m codes binaires de longueur m, et chaque code correspond a au plus une phrase car le code est UD, donc le nombre de phrases (de longueur k) dont le code est de longueur m est au plus 2^m :

$$A_m \leq 2^m$$

, donc

$$\left(\sum_{u\in\mathcal{X}} 2^{-l(C(u))}\right)^k \leq kL,$$

càd

$$\sum_{u\in\mathcal{X}} 2^{-l(C(u))} \leq (kL)^{1/k},$$

donc, comme $(kL)^{1/k} \underset{k\to\infty}{\longrightarrow} 1$, on a bien

$$\sum_{u\in\mathcal{X}} 2^{-l(C(u))} \leq 1.$$

□

4.2 Le code de Huffman

Quelle longueur moyenne peut-on espérer atteindre ?

Exemple 4.10. Supposons que $\mathcal{X} = \{a, b, c, d, e\}$, muni de $p = (.3, .25, .25, .1, .1)$. Voici comment fonctionne l'algo :

- $\mathcal{X}_0 = \{a, b, c, d, e\}$, muni de $p_0 = (.3, .25, .25, .1, .1)$
- $\mathcal{X}_1 = \{a, b, c, de\}$, muni de $p_1 = (.3, .25, .25, .2)$
- $\mathcal{X}_2 = \{cde, a, b\}$, muni de $p_2 = (.45, 3, .25)$
- $\mathcal{X}_3 = \{ab, cde\}$, muni de $p_3 = (.55, .45)$
- $\mathcal{X}_4 = \{abcde\}$, muni de $p_4 = (1)$

On dessine ensuite l'arbre suivant (cf figure 13) pour écrire le code suivant (cf figure 14).

Exercice 4.7. On considère la source $\mathcal{X} = \{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta, \eta\}$ avec probabilités respectives $5\%, 10\%, 25\%, 15\%, 5\%, 35\%, 5\%$.

1. Construire un codage binaire en utilisant l'algorithme de Huffman.
2. Calculer $\sum_i 2^{-l_i}$.
3. Calculer la longueur moyenne du code et la comparer à l'entropie de la source.

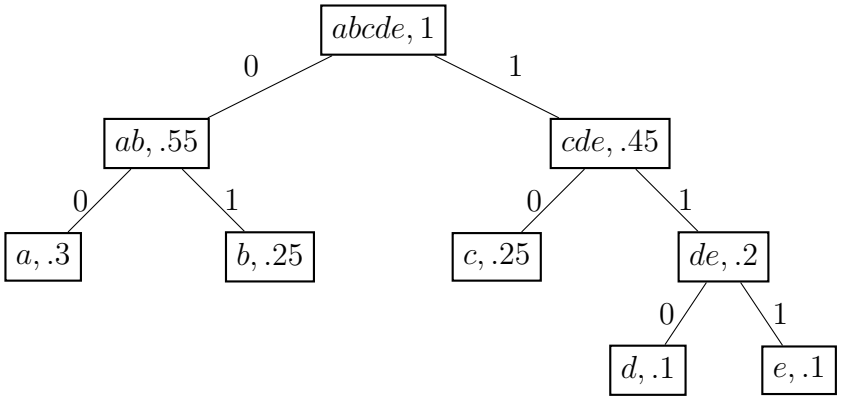


FIGURE 13 – Arbre de Huffman associé à $\mathcal{X} = \{a, b, c, d, e\}$, muni de $p = (.3, .25, .25, .1, .1)$

u	$C(u)$
a	00
b	01
c	10
d	110
e	111

FIGURE 14 – Code de Huffman obtenu pour $\mathcal{X} = \{a, b, c, d, e\}$, muni de $p = (.3, .25, .25, .1, .1)$

4. Encoder le mot $\beta\alpha\gamma\alpha\gamma\varepsilon$.
5. A une étape de l'algorithme, on peut choisir une autre manière de fusionner les lettres, ce qui donne un autre code final. Donner ce code.
6. Refaire les questions précédentes avec le nouveau code.

Nous allons voir qu'une procédure de codage apparaît naturellement comme optimale.

Algorithme de codage de Huffman (version *avec les mains*) : posons $\mathcal{X} = \{u_1, \ldots, u_n\}$ $(n = |\mathcal{X}|)$, avec $p(u_1) \geq \cdots \geq p(u_n)$.

• Si $n = 1$, on pose $C \equiv \emptyset$ et si $n = 2$, on pose $C(u_1) = 0$ et $C(u_2) = 1$.

• Sinon, on fusionne les u_i deux à deux en fusionnant toujours les deux qui ont la plus faible proba, et en associant à leur fusion la somme des probas des fusionnés, de façon à arriver à $n = 2$. On dessine l'arbre des fusions et pour obtenir le code binaire de chaque u_i, on remonte l'arbre à partir de la

racine jusqu'aux feuilles en rajoutant à chaque fois au code un 0 ou un 1 selon la branche suivie.

Algorithme de codage de Huffman (formalisation) :

• Soit un ensemble $\mathcal{X} = \{u_1, \dots, u_m\}$ avec $m \geqslant 3$ muni d'une loi de probabilités $p = (p_1, \dots, p_m)$. Quitte à changer de numérotation, on suppose que les p_i sont décroissants

$$p_1 \leqslant \cdots \leqslant p_m$$

et on définit $(\mathcal{X}', p') := \mathcal{T}(\mathcal{X}, p)$ via

$$\begin{aligned} \mathcal{X}' &= \{u_1, \dots, u_{k-2}, u'_{m-1} := \{u_{m-1}, u_m\}\} \\ p' &= (p_1, \dots, p_{m-2}, p'_{m-1} = p_{m-1} + p_m). \end{aligned}$$

• On construit la suite d'ensembles

$$(\mathcal{X}_n, p^n) = (\mathcal{X}, p), (\mathcal{X}_{n-1}, p^{n-1}) = \mathcal{T}(\mathcal{X}_n, p^n), \dots, (\mathcal{X}_2, p^2) = \mathcal{T}(\mathcal{X}_3, p^3).$$

On remarque que l'ordre des $p_i^k, 1 \leqslant i \leqslant k$ peut changer à chaque itération afin d'être toujours en ordre décroissant.

• On construit ensuite le code de manière récursive :

• En notant α, β les éléments de $\mathcal{X}_2$ avec $p_\alpha \leqslant p_\beta$, on définit le code sur $\mathcal{X}_2$:

$$C_2(\alpha) = 0, C_2(\beta) = 1$$

• Pour $k \geqslant 3$, on construit le code sur $\mathcal{X}_k$ à partir de $\mathcal{X}_{k-1} = \{\alpha_1, \dots, \alpha_{k-1}\}$ en remarquant que l'un des α_i est de la forme $\alpha_i = \{\beta, \beta'\} \subset \mathcal{X}_k$ car $(\mathcal{X}_{k-1}, p^{k-1}) = \mathcal{T}(\mathcal{X}_k, p^k)$, et on constuit le code par

$$\begin{aligned} C_k(\beta') &= C_{k-1}(\alpha_i)0 \\ C_k(\beta) &= C_{k-1}(\alpha_i)1 \\ C_k(\alpha_j) &= C_{k-1}(\alpha_j), j \neq i. \end{aligned}$$

Parfois, on affecte par convention le bit 0 à l'élèment le moins probable, i.e. $p(\beta) \geqslant p(\beta')$ avec les notations ci-dessus.

Rappelons que par construction, β, β' sont les élèments dont les probabilités p_k, p_{k-1} sont les plus faibles de p^k. Ainsi chaque code C_k de l'algorithme possède la propriété fondamentale :

$$\begin{array}{r} \text{Les deux éléments ayant les codes les plus longs} \\ \text{ne diffèrent que par le dernier bit.} \end{array} \qquad \text{(P)}$$

Il se trouve que ces deux éléments sont parmi les moins probables (il se peut qu'il y en ait plus que 2).

Remarque 4.11. Il n'y a pas qu'une seule solution car il se peut que $p^k_{k-1} = p^k_{k-2}$ au stade k, ce qui veut dire qu'on peut fusionner de deux manières différentes : p_k avec p_{k-1} ou p_k avec p_{k-2}. Dans ce cas, une bonne pratique est de fusionner les mots les plus courts, afin de ne pas avoir trop de disparités dans les longueurs finales.

4.3 Codes optimaux

On considère toujours un ensemble $\mathcal{X} = \{x_1, \ldots, x_m\}$ muni d'une probabilité $p = (p_i; i = 1, \ldots, m)$.

On note $\mathcal{C}_m$ l'ensemble des codes UD sur un ensemble à m éléments. On note $L_p(C)$, ou juste $L(C)$ pour simplifier, la longueur moyenne avec les fréquences p,

$$L_p(C) = \sum_{u \in \mathcal{X}} p(u) l(C(u)).$$

Si un code minimise L_p, il est dit p-optimal, ou juste optimal.

Proposition 4.12. *1. Si*

$$\sum_i 2^{-l_i} < 1,$$

on peut trouver $C' \in \mathcal{C}_m$ code UD dont les longueurs l'_i vérifient $l'_i \leqslant l_i; i = 1, \ldots, m$ et qui vérifie

$$\sum_i 2^{-l_i} = 1$$

2. En déduire que tout code p-optimal satisfait l'égalité

$$\sum_{i=1}^{m} 2^{-l_i} = 1.$$

Exercice 4.8. Faire la preuve.

Proposition 4.13. *Tout code UD vérifie*

$$L_p(C) \geqslant \mathsf{H}(p)$$

et de plus

$$\min_{C \in \mathcal{C}_m} L_p(C) < \mathsf{H}(p) + 1.$$

On a égalité $L(C) = \mathsf{H}(p)$ uniquement pour un code p-optimal si les probabilités sont de la forme

$$p_i = 2^{-l_i}.$$

Démonstration. On note $l_i, i = 1, \ldots, m$ les longueurs $C(u_i), u_i \in \mathcal{X}$. La preuve est basée sur les probabilités

$$\tilde{p}_i = \frac{2^{-l_i}}{s}$$

où $s = \sum_i 2^{-l_i}$. L'idée est que dans ce cas, $\ln_2(\tilde{p}_i s) = l_i$, et donc

$$L_{\tilde{p}}(C) = \sum_i \tilde{p}_i l_i = \sum_i \tilde{p}_i \ln_2(\tilde{p}_i) + \ln_2(s) = \mathsf{H}(p) + \ln_2(s).$$

Si le code est optimal, on a vu que $s = 1$, et donc on a bien $L_{\tilde{p}}(C) = \mathsf{H}(\tilde{p})$.

Dans le cas général, comme le code est UD on a $\sum_i 2^{-l_i} \leqslant 1$, et la différence entre $\mathsf{H}(p)$ et $L(C)$ est dictée par la différence entre les p_i et les $\tilde{p}_i$:

$$\begin{aligned} L(C) - \mathsf{H}(p) &= \sum_i p_i l_i - \mathsf{H}(p) \\ &= -\sum_i p_i \ln_2(2^{-l_i}) - \mathsf{H}(p) \\ &= -\sum_i p_i \ln_2(\tilde{p}_i) - \sum_i p_i \ln_2(s) - \mathsf{H}(p) \\ &= \mathsf{D}(p\|\tilde{p}) - \underbrace{\ln_2(s)}_{\leqslant 0 \text{ car } s \leqslant 1} \geqslant 0. \end{aligned}$$

On a en effet égalité ssi $p = \tilde{p}$ et $s = 1$.

Pour la borne supérieure, on pose

$$l'_i = [-\ln_2(p_i)] + 1 \geqslant -\ln_2(p_i),$$

on a alors l'inégalité de Kraft qui est vérifiée,

$$\sum_i 2^{-l'_i} \leqslant \sum_i 2^{-\ln_2(p_i)} = 1,$$

donc il existe un code préfixe avec ces longueurs. On a

$$\begin{aligned} -\ln_2(p_i) \leqslant l_i &\leqslant -\ln_2(p_i) + 1 \\ -p_i \ln_2(p_i) \leqslant p_i l_i &\leqslant -p_i \ln_2(p_i) + p_i \\ \mathsf{H}(p) \leqslant L(p) &\leqslant \mathsf{H}(p) + 1 \end{aligned}$$

□

Théorème 4.14. *Un code de Huffman C est p-optimal.*

Faisons la preuve avec ces 2 exercices.

Exercice 4.9. Soit un code optimal C. On suppose toujours $p_1 \leqslant \cdots \leqslant p_m$ quitte à réordonner les éléments.

1. Montrer que $l_i \leqslant l_j$ pour $p_i > p_j$.
2. Montrez que $l_m = l_{m-1}$.
3. Montrez qu'on peut construire un code préfixe de mêmes longueurs qui vérifie (P).

Exercice 4.10. Soit

$$\begin{aligned} (\mathcal{X}, p) =& (\mathcal{X}_{k+1}, p^{k+1}) \text{ muni du code } C = C_{k+1}, \\ (\mathcal{X}', p') =& \mathcal{T}(\mathcal{X}, p) = (\mathcal{X}_k, p^k) \text{ muni du code } C' = C_k \end{aligned}$$

deux codes successifs dans l'algorithme de Huffman ($2 \leqslant k < n$). On note

$$\begin{aligned} p^{k+1} =& (p_1, \ldots, p_{k+1}) \\ p^k =& (p_1, \ldots, p_{k-1}, p'_k = p_k + p_{k+1}) \text{ (pas forcément dans l'ordre décroissant).} \end{aligned}$$

1. Montrez que
$$L(C) = L(C') + p'_k. \tag{1}$$
2. Montrez que $\mathsf{H}(p') - \mathsf{H}(p) = -p'_k \mathsf{H}(\mathcal{B}(p_k/p'_k))$.
3. Montrez que si C' est optimal sur $(\mathcal{X}', p')$ parmi tous les codes préfixes, C l'est également sur $(\mathcal{X}, p)$.
4. Déduisez-en que le code de Huffman est un code préfixe ayant la meilleure longueur moyenne.

Corollaire 4.15. *Les codes de Huffman sont optimaux parmi tous les codes UD.*

Démonstration: Soit C un code UD optimal. D'après l'inégalité de McMillan, ses longueurs $l_1, \ldots, l_m$ vérifient l'inégalité de Kraft. Donc on peut construire un code préfixe C' qui a ces longueurs-là. Comme d'après l'exo précédent tout code de Huffman C_H est optimal parmi les codes préfixe, $L(C_H) \leqslant L(C)$. □

4.4 Exercices

Exercice 4.11. Un individu A choisit en pensée un nombre entier X uniformément entre 1 et 32 (inclus) et un autre individu, B, va essayer de déterminer ce nombre en posant à A des questions auxquelles il ne répond que par oui ou par non.

1. Donner une façon certaine de déterminer X en 5 questions.
   Indication : $32 = 2^5$.
2. Est-il possible, parfois, avec de la chance par exemple, de deviner X en moins de 5 questions ?
3. Existe-t-il une façon de déterminer X en moins de 5 questions en moyenne ?

Exercice 4.12. Des mots comme "stop" ou "feu" sont petits, non parce que leur utilisation est fréquente mais peut-être parce qu'on souhaite minimiser le temps nécessaire pour les dire. Considèrons une variable aléatoire $X : \Omega \to \{x_1, \ldots, x_n\}$. Soit l_i le nombre de bits nécessaires pour coder x_i, c_i le coût par lettres du mot x_i, et $p_i = \mathbb{P}(X = x_i)$. Le coût moyen du code est alors
$$C = \sum_{i=1}^{n} p_i c_i l_i.$$

1. Donner un moyen de trouver un code qui minimise C en vous inspirant de la procédure pour construire un code de Huffman.
2. Déterminer un code optimal dans le cas suivant.

x	$\mathbb{P}(X = x)$	$c(x)$
1	0.5	0.5
2	0.25	0.125
3	0.125	0.125
4	0.0625	0.125
5	0.0625	0.125

Exercice 4.13. On dispose de 6 bouteilles de vin. On sait qu'une bouteille est empoisonnée. On dispose d'un produit qui réagit lorsqu'il est mis en présence de produit empoisonné.

1. Trouver une procédure qui permet de déterminer la mauvaise bouteille avec uniquement 3 doses de réactif. *indice : Penser à mélanger plusieurs vins.*
2. Avec la méthode précédente, étant donné n bouteilles dont 1 empoisonnée, combien faut-il de doses pour trouver la mauvaise bouteille ?
3. En notant X la variable aléaoire donnant le numéro de la bouteille mauvaise. On donne :

$$\mathbb{P}(X = 1) = 7/26, \mathbb{P}(X = 2) = 5/26, \mathbb{P}(X = 3) = 4/26,$$

$$\mathbb{P}(X = 4) = 4/26, \mathbb{P}(X = 5) = 3/26, \mathbb{P}(X = 6) = 3/26.$$

   Trouver un code de Huffman associé à la distibution de X.
4. Quel est le nombre moyen de tests à réaliser avant de trouver la mauvaise bouteille avec la distribution précédente ?

Exercice 4.14. Combien de mots binaires de longueur l existe-t-il ? En déduire qu'on peut associer à tout ensemble de m mots binaire un arbre de hauteur $l = [\ln_2(m)] + 1$.

Exercice 4.15. Soit un ensemble source à coder 12 symboles. Votre code binaire uniquement décodable a déjà 4 mots de longueur 3, et 6 mots de longueur 4. Combien de mots de longueur 5 pouvez-vous ajouter ?

Exercice 4.16. Considérons un alphabet de 4 lettres. Existe-t-il un code uniquement décodable sur cet alphabet qui consiste en 2 mots de longueur 1, 4 mots de longueur 2, 10 mots de longueur 3 et 16 mots de longueur 4 ?

Exercice 4.17 (Code de Huffman et loi uniforme)**.** Soit $\mathcal{X}$ un alphabet de taille $n \geq 1$, muni d'une loi p.

a) Donner une expression précise de l'entier strictement positif k tel que

$$2^{k-1} < n \leq 2^k.$$

b) Montrer qu'il existe un code instantanné C de $\mathcal{X}$ tel que pour tout $x \in \mathcal{X}$, $\ell(C(x)) = k$.

c) Montrer que si p est la loi uniforme, on a bien

$$\mathsf{H}(p) \leq L(C) < \mathsf{H}(p) + 1.$$

d) Prouver que si p est la loi uniforme et n est une puissance de 2, il n'existe pas de code instantanné de longueur moyenne inférieure.

e) Si p est la loi uniforme et n n'est pas une puissance de 2, existe-t-il un code instantanné de longueur moyenne inférieure ?

4.5 Bilan

Soit $\mathcal{X} = \{x_1, \ldots, x_m\}$ un ensemble à m éléments, et $p = (p_1, \ldots, p_m)$ une distribution sur $\mathcal{X}$.

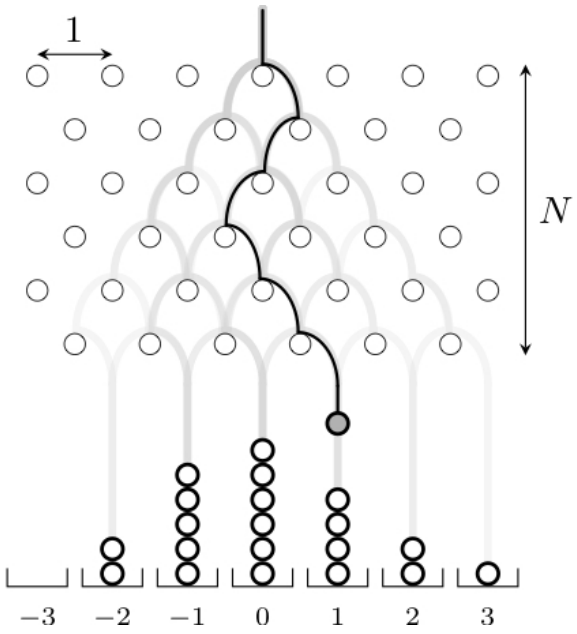
- Etant donné des longueurs $l_1, \ldots, l_m$ il existe un code $C : \mathcal{X} \to \{0,1\}^*$ préfixe tel que $\ell(C(x_i)) = l_i$.
- Si un code est optimal sur $\mathcal{X}$, $\sum_i 2^{-l(C(x_i))} = 1$ (la réciproque est fausse car l'optimalité dépend de p, alors que cette égalité ne dépend que des longueurs, et pas de p)
- Tout code de Huffman est $p-$optimal et préfixe, mais il peut exister d'autres codes UD $p-$optimaux (et en particuliers d'autres codes de Huffman)
- Tout code $p-$optimal vérifie

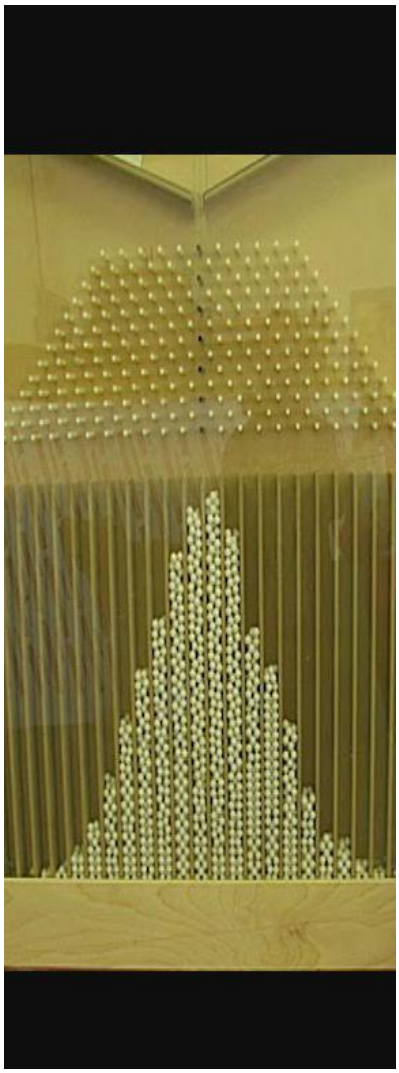
$$\mathsf{H}(p) \leqslant L_p(C) < \mathsf{H}(p) + 1.$$

5 Ensembles typiques et grandes déviations

Le but de cette section est d'étudier précisément la loi de $S_n = \sum_{i=1}^n X_i$ lorsque les X_i sont des variables discrètes i.i.d. dans un espace fini $\mathcal{X} = \{a_1, \dots, a_m\}$.

Introduisons cette section par l'exemple de la planche de Galton, qui permet de visualiser la loi de S_n lorsque les X_i sont des variables de Bernoulli :





La valeur de $\mathbb{P}(S_n = k)$ correspond à la hauteur de la colonne de billes dans la case k. On voit que la valeur la plus probable est $n/2$, ce qui corres-

pond à la LGN :

$$S_n \sim \frac{n}{2}.$$

Pourquoi ? Tout simplement parce qu'il y a plus de "trajectoires" qui donnent 0 qu'un autre nombre : pour $n = 4$:

$$\begin{aligned}
0 &= ++--=--++=-+-+=+-+-=+--+=-++- \\
2 &= +++-=++-+=+-++=-+++ \\
-2 &= (idem...) \\
4 &= ++++ \\
-4 &= ----
\end{aligned}$$

On observe même une décroissance de cette probabilité lorsqu'on s'éloigne de la moyenne, suivant le profil d'une loi gaussienne, ce qui correspond au TCL. On va en fait pouvoir étudier très précisément la valeur de $\mathbb{P}(S_n = k)$ en utilisant l'entropie.

Pour étudier $S_n = X_1 + \cdots + X_n$, la première observation est que l'ordre des X_i n'est pas important, autrement dit, si on a $\bar{x} = (x_1, \ldots, x_n)$ des valeurs possibles, on s'intéresse non pas à

$$\mathbb{P}(X = \bar{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

mais à

$$\mathbb{P}(\forall a \in \mathcal{X}, \#\{i : X_i = a\} = \#\{i : x_i = a\}). \qquad (*)$$

Ce qui est important c'est donc le nombre de fois où X prend une valeur $a \in \mathcal{X}$, et non pas le moment où elle la prend :

$$n_{\bar{x}}(a) = \#\{i : x_i = a\}$$

ou de manière équivalente la *distribution empirique* $p_{\bar{x}} = (p_{\bar{x}}(a))_{a \in \mathcal{X}}$ avec

$$p_{\bar{x}}(a) = \left(\frac{n_{\bar{x}}(a)}{n} \right)$$

on ajoute la loi empirique $(n_{\bar{x}}(+1), n_{\bar{x}}(-1))$:

$$\begin{aligned}
0 &= ++--=--++=-+-+=+-+-=+--+=-++- && (2+, 2-) \\
2 &= +++-=++-+=+-++=-+++ && (3+, 1-) \\
-2 &= (idem...) && (1+, 3-) \\
4 &= ++++ && (4+, 0-) \\
-4 &= ---- && (0+, 4-)
\end{aligned}$$

Cette quantité est représentée par la mesure empirique :

Définition 5.1. pour $\bar{x} \in \mathcal{X}^n$,

$$\mu_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} = \frac{1}{n}\sum_{a\in\mathcal{X}} n_{\bar{x}}(a)\delta_a = \sum_a p_{\bar{x}}(a).$$

L'avantage de la seconde somme est qu'elle a un nombre borné de termes lorsque $n \to \infty$. $\mu_{\bar{x}}$ est une mesure constituée d'un nombre fini d'atomes. Le $\frac{1}{n}$ sert à conserver une masse totale de 1.

(*) se réécrit

$$\mathbb{P}(p_X = p_{\bar{x}})$$

où $X = (X_1, \ldots, X_n)$. Le simplexe est l'ensemble des distributions possibles

$$\mathcal{S}^n = \{(\frac{n_1}{n}, \ldots, \frac{n_m}{n}) : n_i \in \mathbb{N} : \sum_{i=1}^{m} n_i = n\}..$$

On note

$$\mathcal{S} = \{(p_1, \ldots, p_m) \in [0,1]^n : \sum_{i=1}^{m} p_i = 1\}$$

Exercice 5.1. Montrer que $\cup_n \mathcal{S}^n \subsetneq \mathcal{S}$.

On confondra souvent les deux objets via la notation $\equiv$. La seconde observation importante est qu'on peut avoir $\bar{x} \neq \bar{y}$ mais $p_{\bar{x}} = p_{\bar{y}}$, par exemple

$$p_{(1,1,0)} = p_{(1,0,1)} = p_{(0,1,1)} \equiv (1/3, 2/3).$$

Exemple 5.2. Pour revenir à la planche de Galton, prenons $\mathcal{X} = \{0,1\}$ et la loi de X est $\mathcal{B}(p), p \in [0,1]$: étant donné $\bar{x} \in \mathcal{X}^n$, $p_{\bar{x}}$ se définit simplement par le nombre de fois où 0 et 1 sont atteints :

$$p_{(0,1,1,0,1)} = (2/5, 3/5).$$

Application pour une somme :

$$\mathbb{P}(\sum_i X_i = s) = \sum_{\bar{x}:\sum_i a n_{\bar{x}}(a) = s} \mathbb{P}(p_X = p_{\bar{x}})$$

5.1 Ensembles typiques et Loi des Grands Nombres

Le théorème suivant nous dit que lorsque l'on observe l'échantillon sous l'angle de sa distribution empirique p_X, on observe le plus souvent une distribution p_X de $\mathcal{S}^n$ pour lesquelles la distance $\mathsf{D}(p_X\|p)$ est faible où p est la loi des X_i, i.e.

$$\mathbb{P}(p_X = q) = 2^{-n\mathsf{D}(q\|p)}\theta_q, q \in \mathcal{S}^n$$

où θ_q n'est pas "très petit" ou "très grand".

Théorème 5.3. *Pour $q \in \mathcal{S}^n, X = (X_1, \ldots, X_n)$, des variables i.i.d. de distribution $p \in \mathcal{S}$,*

$$\frac{\mathbb{P}(p_X = q)}{2^{-n\mathsf{D}(q\|p)}} \in \left[\frac{1}{(n+1)^{|\mathcal{X}|}}, 1\right]$$

Formulation logarithmique

$$\ln_2(\mathbb{P}(p_{\bar{X}} = q)) = -n\mathsf{D}(q\|p) + O(\ln_2(n)).$$

Le terme exponentiel $2^{-n\mathsf{D}(q\|p)}$ est très faible si la distance est grande, cela permet de dire que $\mathsf{D}(\mu_X\|q)$ sera proche de 0 avec une grande probabilité, et on peut facilement borner la marge d'erreur de cette affirmation :

Exercice 5.2. Prenons comme exemple $\mathcal{X} = \{-1, 1\}$, $p = (1/2, 1/2)$. Soit X_i des Rademacher indépendantes et

$$S_n = \sum_{i=1}^{n} X_i.$$

On code p_X sous la forme $p_X = (N_+/n, N_-/n)$, ce qui veut dire que N_+ variables X_i prennent la valeur $+1$, et N_- variables la valeur -1. On a en particulier

$$S_n = N_+ - N_-.$$

Nécessairement, $N_+ + N_- = n$.

1. Donner une approximation de $\mathbb{P}(N_+ = n_+, N_- = n_-)$.

2. Soit $c > 0$ et a_n^+, a_n^- des suites telles que pour tout n, $|a_n^+/n - 1/2| > c$. Que vaut $\lim_n \mathbb{P}(N_+ = a_n^+, N_- = a_n^-)$? Déduisez-en $\lim_n \mathbb{P}(|S_n|/n > \varepsilon)$ pour $\varepsilon > 0$.
3. Soit a_n^+, a_n^- des suites telles que

$$\delta_n := \frac{a_n^+}{n} - \frac{1}{2} \to 0.$$

Donner une approximation de

$$\mathbb{P}(N_+ = a_n^+, N_- = a_n^-)$$

et montrez que ça tend vers 0 si $n^{-1/2} = o(\delta_n)$.

Proposition 5.4. *On a*

$$\mathbb{P}(\mathsf{D}(p_X \| p) \geqslant \frac{|\mathcal{X}| \ln_2(n)}{n}) \to 0.$$

et que $\mathsf{D}(p_X \| p) \to 0$ *p.s. De plus pour* $\varphi : \mathcal{X} \to \mathbb{R}$ *on remontre la LGN*

$$\frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) \to \mathbb{E}(\varphi(p)) := \sum_a \varphi(a) p(a).$$

Exercice 5.3 (Preuve)**.**

Exercice 5.4. Soit X_k des variables iid de loi $\mathcal{B}(p)$, Bernoulli de paramètre $p \in [0, 1]$, et $S_n = \sum_{k=1}^n X_k$.

1. Donner des borne inférieures et supérieures sur $\mathbb{P}(S_n = k)$ en fonction de la fonction $d(q) := \mathsf{D}(\mathcal{B}(q) \| \mathcal{B}(p))$.
2. Soit $C_p > c_p := \frac{1}{2\ln(2)}(\frac{1}{p} + \frac{1}{1-p})$. Montrez que pour $\varepsilon > 0$

$$c_p \varepsilon^2 \leqslant d(p + \varepsilon) \leqslant C_p \varepsilon^2.$$

l'inégalité de droite n'étant vraie que ε suffisament petit. On pourra écrire le DL à l'odre 2 de $\ln_2(1 + \varepsilon)$, ou utiliser l'inégalité de Pinsker.

3. Déduisez-en que pour $k \in \mathbb{N}$,

$$\mathbb{P}(S_n - np = k) \leqslant 2^{-c_p k^2/n}.$$

4. Montrez qu'$\exists c > 0$ tel que $\mathbb{P}(S_n - np > \sqrt{n}t) \geqslant c 2^{-C_p t^2}$.

5.2 Preuve du Théorème

On note

$$T_{\bar{x}} = \{\bar{y} : p_{\bar{x}} = p_{\bar{y}}\},$$

estimer son cardinal sera une question importante.

Exercice 5.5. On peut avoir une écriture explicite :

$$\#T_{\bar{x}} = \binom{n_{\bar{x}}(a_1)}{n}\binom{n_{\bar{x}}(a_1)}{n - n_{\bar{x}}(a_1)}\binom{n_{\cdot x}(a_2)}{n - n_{\bar{x}}(a_1) - n_{\bar{x}}(a_2)}\cdots\binom{n_{\bar{x}}(a_m)}{n_{\bar{x}}(a_m)} = \frac{n!}{n_{\bar{x}}(a_1)!n_{\bar{x}}(a_2)!\ldots n_{\bar{x}}(a_m)!}$$

Il y a beaucoup plus de possibilités pour X que pour p_X (pour $|\mathcal{X}|$ fini) :

$$\begin{aligned} &\#\mathcal{X}^n = |\mathcal{X}|^n && \text{(croissance exponentielle en } \exp(n\ln_2(\mathcal{X}))) \\ &\#\mathcal{S}^m \leqslant (n+1)^{|\mathcal{X}|} && \text{(croissance polynomiale).} \end{aligned}$$

Exercice 5.6. Montrer ces inégalités.

En conséquence, il y a des mesures empiriques $\mu_{\bar{x}}$ qui vont recevoir une très grande probabilité par rapport à d'autres. Le but de la première partie est de les identifier.

Exercice 5.7.

1. Soit $\bar{x} \in \mathcal{X}^n$. Montrez que si $p = p_{\bar{x}}$
$$\mathbb{P}(X = \bar{x}) = 2^{-n\mathsf{H}(p_{\bar{x}})}.$$
2. Montrer que pour $\bar{x} \in \mathcal{X}^n$ quelconque
$$\mathbb{P}(X = \bar{x}) = 2^{-n(\mathsf{H}(p_{\bar{x}}) + \mathsf{D}(p_{\bar{x}}\|p))}$$
3. * On appelle $T_{\bar{x}} \subset \mathcal{X}^n$ la classe des $\bar{y} \in \mathcal{X}^n$ donnant la même mesure empirique,
$$T_{\bar{x}} = \{\bar{y} \in \mathcal{X}^n : p_{\bar{x}} = p_{\bar{y}}\}.$$
Le but est de montrer que pour $\bar{x} \in \mathcal{X}^n$,
$$\frac{2^{n\mathsf{H}(p_{\bar{x}})}}{(n+1)^{|\mathcal{X}|}} \leqslant \#T_{\bar{x}} \leqslant 2^{n\mathsf{H}(p_{\bar{x}})}.$$

(a) Montrez que $\sum_{\bar{y}} 2^{-n\mathsf{H}(p_{\bar{x}})} \leqslant 1$, déduisez-en la borne supérieure.

(b) On note μ^n la mesure produit sur $\mathcal{X}^n$ définie par

$$\mu^n(\{\bar{x}\}) = \mu(\{x_1\})\mu(\{x_2\})...\mu(\{x_n\}) = \prod_{a\in\mathcal{X}} \mu(\{a\})^{n_{\bar{x}}(a)}.$$

Montrez que pour tout $\bar{y}$,

$$\mu^n_{\bar{x}}(T_{\bar{x}}) \geqslant \mu^n_{\bar{x}}(T_{\bar{y}}).$$

en utilisant l'inégalité $\frac{m!}{p!} \geqslant p^{m-n}$ pour tout $m, p \in \mathbb{N}^*$ (distinguer $m \geqslant p$ et $m < p$).

(c) Déduisez-en la borne inférieure.

4. Conclure.

On voit donc qu'au premier ordre, la quantité importante qui définit si une valeur ν va être atteinte est la distance $\mathsf{D}(\nu\|\mu)$.

Dans la section suivante, on étudie le convergence $p_X \to p$ plus quantitativement.

5.3 Théorème de Sanov

Un corollaire très pratique est la théorème de Sanov. On va considérer un sous-ensemble de la classe de toutes les mesures empiriques. Puisqu'on étudie souvent la moyenne $\frac{1}{n}\sum_i x_i$ si $\mathcal{X} \subset \mathbb{R}$, un exemple typique est quand la moyenne est dans un certain intervalle :

$$\mathcal{E}_{[a,b]} := \{p : \sum_a p(a) \in [a, b]\}.$$

On peut imaginer des exemples plus biscornus comme

$$\mathcal{E} = \{p : p(a) \neq p(b) \forall a \neq b \in \mathcal{X}\}.$$

Théorème 5.5 (Sanov)**.** *Soit $\mathcal{E}$ un ensemble de distributions sur $\mathcal{X}$. Alors*

$$-n \min_{q\in\mathcal{S}^n\cap\mathcal{E}} \mathsf{D}(q\|p) - |\mathcal{X}|\ln_2(n+1) \leqslant \ln_2 \mathbb{P}(p_X \in \mathcal{E}) \leqslant -n \inf_{q\in\mathcal{E}} \mathsf{D}(q\|p) + |\mathcal{X}|\ln_2(n+1).$$

Exercice 5.8. Soit $\mathcal{X} = \{0, 1\}$, et $p = \mathcal{B}(t)$ où $t \in [0, 1]$, $\mathcal{E} = \{q : \sum_{a\in\mathcal{X}} aq(a) > t + \alpha\}$ où $\alpha > 0$.

1. Trouver le minimum sur $\mathcal{E}$, et sur $\mathcal{E}_{\mathcal{X}^n} \cap \mathcal{E}$.
2. Donnez la limite de
$$\frac{1}{n} \ln_2 \mathbb{P}(\frac{1}{n} \sum_i X_i > p + \alpha)$$

Démonstration. La borne supérieure est assez facile grâce au Théorème 5.3 :
$$\begin{aligned}\mathbb{P}(p_X \in \mathcal{E}) &= \sum_{p_{\bar{x}} \in \mathcal{E}} \mathbb{P}(p_X = p_{\bar{x}}) \\ &\leqslant \sum_{p_{\bar{x}}} \underbrace{2^{-n\mathsf{D}(p_{\bar{x}}\|p)}}_{\leqslant 2^{-n \inf_q \mathsf{D}(q\|p)}} \\ &\leqslant \underbrace{\#|\{\mu_{\bar{x}} : \bar{x} \in \mathcal{X}^n\}|}_{\leqslant (n+1)^{|\mathcal{X}|}} 2^{-n\mathsf{D}(q^*\|p)}\end{aligned}$$

Pour la borne inférieure, on choisit $\bar{x}_n^* \in \mathcal{X}^n$ qui minimise la distance sur les vecteurs de taille n : pour tout $\bar{x} \in \mathcal{X}^n$,
$$\mathsf{D}(p_{\bar{x}_n^*}\|p) \leqslant \mathsf{D}(p_{\bar{x}}\|p),$$
et on a alors
$$\mathbb{P}(p_X \in \mathcal{E}) = \sum_{\substack{p_{\bar{x}} \in \mathcal{E} \\ \bar{x} \in \mathcal{X}^n}} \mathbb{P}(p_X = p_{\bar{x}}) \geqslant \sum_{\substack{p_{\bar{x}} \in \mathcal{E} \\ \bar{x} \in \mathcal{X}^n}} \frac{2^{-n\mathsf{D}(p_{\bar{x}}\|p)}}{(n+1)^{|\mathcal{X}|}} \geqslant \frac{2^{-n\mathsf{D}(p_{\bar{x}_n^*}\|p)}}{(n+1)^{|\mathcal{X}|}}$$
puisque $p_{\bar{x}_n^*}$ est un élément de la somme. On a finalement
$$-n\mathsf{D}(p_{\bar{x}_n^*}\|p) + O(\ln_2(n)) \leqslant \ln_2(\mathbb{P}(p_X \in \mathcal{E})) \leqslant -n \inf \mathsf{D}(p_{\bar{x}^*}\|p) + O(\ln_2(n))$$
□

La question qui demeure quand $n \to \infty$ est donc topologique : a-t-on
$$\mathsf{D}(p_{\bar{x}_n^*}\|\mu) \to \inf_{p \in \mathcal{E}} \mathsf{D}(\mu_{\bar{x}}\|\mu)?$$

Exercice 5.9. Montrez que si $\mathcal{E}$ est la fermeture de son intérieur pour la topologie de la convergence en loi, alors
$$\lim_n \inf_{\bar{x} \in \mathcal{X}^n} \mathsf{D}(p_{\bar{x}}\|p) = \mathsf{D}(p^*\|p).$$
On note souvent $p^* = p^*(\mathcal{E}, X, \mu)$ la distribution de $\mathcal{E}$ dont l'entropie est la plus proche de celle de p :
$$\mathsf{D}(p^*\|p) = \min_{q \in \mathcal{E}} \mathsf{D}(q\|p).$$

Références

[1] Thomas M. Cover, Joy A. Thomas *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing) (2006).

[2] Djalil Chafai, *Quelques mots sur l'entropie*, http://www.lsp.ups-tlse.fr/Chafai/

[3] , Amir Dembo, Ofer Zeitouni *Large Deviation Techniques and Applications*. Springer.

[4] Friedli, S., Velenik, Y. *Statistical Mechanics of Lattice Systems : A Concrete Mathematical Introduction* Cambridge : Cambridge University Press, 2017.

[5] Marc Lelarge *Introduction à la Théorie de l'Information et au Codage*. Page web de l'auteur. (2014)

[6] MacKenzie, I. S., Soukoreff, R. W., Helga, J. (2011). *1 thumb, 4 buttons, 20 words per minute : Design and evaluation of H4-Writer*. Proceedings of the ACM Symposium on User Interface Software and Technology - UIST 2011, 471–480. New York : ACM.

[7] Marc Mézard, Andrea Montanari *Information, Physics, and Computation*, Oxford, 2009.

[8] Yann Ollivier *Aspects de l'entropie en mathématiques*. Page web de l'auteur. (2002)

[9] Olivier Rioul *Théorie de l'information et du codage*, Lavoisier, 2007.

[10] Raymond W. Yeung *A First Course in Information Theory*, Springer (2012).