Université de Rennes

Master de cryptographie

Probabilités et statistique pour la théorie de l'information

Notes de cours

Dimitri Petritis

UFR Mathématiques

Rennes
Janvier 2023

Dimitri Petritis
Institut de recherche mathématique de Rennes
Université de Rennes et CNRS (UMR6625)
Campus de Beaulieu
35042 Rennes Cedex
France

Mathematics subject classification (2010): 60-01 94-01

Table des matières

1 Aléa et information

Des nos jours, le codage, le stockage, la transmission, le traitement, l'extraction et le chiffrement de l'information se font de manière algorithmique sur des messages numériques que nous pouvons toujours considérer comme des suites de bits. On peut donc légitimement s'interroger qu'ont à faire les probabilités et la statistique — branches mathématiques étudiant le caractère aléatoire des phénomènes — dans cet univers algorithmique.

1.1 Rôle de l'aléa

De manière générale, il y a plusieurs phénomènes purement déterministes qui présentent une indétermination aléatoire. Par exemple le lancer d'une pièce est une expérience déterministe, l'aléa résulte de l'imprécision dans la définition de la condition initiale. Un système dynamique chaotique est un objet purement déterministe ; le comportement aléatoire est dû à l'imprécision avec laquelle on peut représenter des fonctions très irrégulières. La mesure d'une grandeur est un processus déterministe ; l'aléa dans l'estimation de la valeur et la probabilité que l'on attache à cette estimation (intervalle de confiance) est dû à l'imprécision inhérente à toute mesure physique.

Plus spécifiquement, il s'avère que les probabilités interviennent dans tous les étapes de la vie de l'information, tantôt de manière fondamentale pour définir la notion même d'information, tantôt comme modélisation d'une nuisance externe ou d'une richesse algorithmique, tantôt comme outil de chiffrement ; la statistique intervient tantôt de manière fondamentale pour définir la notion d'extraction d'information, tantôt comme outil d'estimation, tantôt comme outil de quantification de la confiance que nous attachons à nos prévisions.

En guise de motivation, quelques exemples de l'utilité de probabilités et de la statistique en théorie de l'information sont donnés ci-dessous :

— Quelle est la **quantité d'information** contenue dans une suite de bits ? Pour répondre à cette question, considérons le bit codant le résultat d'une expérience de pile (0) ou face (1) provenant du lancer d'une pièce honnête. L'incertitude avant que le résultat de l'expérience nous soit révélé est totale (100%). Après

que le résultat nous soit révélé, l'incertitude est nulle. L'information contenue dans le bit correspondant au résultat du lancer d'une pièce honnête est égale à la réduction de l'incertitude après révélation du résultat. Nous verrons au chapitre 7 que pour une pièce arbitraire, donnant face avec probabilité $p \in [0,1]$, l'information contenue dans le bit codant le résultat du lancer est la quantité

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p) \in [0,1], \text{ (avec } 0\log_2 0 = 0),$$

connue sous le nom d'**entropie** ou **quantité d'information** [63, 41]. Remarquer que $H(1/2) = 1$.

— Même si la suite de bits est construite de manière totalement déterministe (correspondant au cas $p = 0, 1$ dans l'exemple ci-dessus), pour être utilisée, elle doit être stockée dans un vecteur physique (laser, courant électrique, signal hertzien, molécule d'ADN) et propagée à travers un canal physique (fibre optique, câble, atmosphère, cellule). Le canal introduit *nécessairement* du bruit de sorte que la suite des bits reçus est toujours une suite aléatoire. Nous verrons dans le chapitre 9 que si le **bruit du canal** est un bruit symétrique, i.e.

$$\mathbb{P}(\text{bit transmis} = 1 | \text{bit émis} = 0) = \mathbb{P}(\text{bit transmis} = 0 | \text{bit émis} = 1) = p \in [0,1],$$

alors la **capacité du canal** est $C(p) = 1 - H(p)$ [64]. On remarque que si $p = 1/2$, le canal symétrique correspondant est totalement inutile car il ne peut transmettre aucune information !

— Une autre utilité des probabilités intervient dans les opérations de **chiffrement**. Supposons que nous voulions transmettre une suite de n bits. Si nous lui superposons une autre suite — la clé de chiffrement — de n bits aléatoires (par addition bit par bit modulo 2), la suite ainsi obtenue devient une suite purement aléatoire. Gilbert Vernam a déposé en 1919 le brevet US1310719 qui proposa une réalisation physique de ce chiffrement. Il a été démontré par Shannon durant la 2e guerre mondiale (et publié ultérieurement [64]) que le chiffrement de Vernam, pourvu que la suite de chiffrement soit utilisée une seule fois, est inviolable. La fonction entropie est de nouveau présente dans les estimations utilisées par Shannon pour montrer ce résultat (cf. chapitre 10).

— Les **codes correcteurs d'erreur** permettent de détecter et/ou corriger certaines erreurs de transmission. Cependant, l'utilisation d'un code correcteur dégrade nécessairement le taux de transmission d'un message car des ressources sont mobilisées pour transmettre les bits supplémentaires qui servent à la détection/correcion des erreurs. Nous verrons au chapitre 11 qu' il existe une frontière dans le plan probabilité d'erreur résiduelle / taux de transmission, appelée **frontière de Shannon**, au delà de laquelle il n'est pas possible de transmettre de message. La forme de cette frontière est encore déterminée par la fonction d'entropie [23, 24].

Cette liste d'exemples d'utilisation des probabilités en théorie de l'information est loin d'être exhaustive. Plusieurs autres aspects de la théorie (codage de la source, codage du canal, reconstitution du message, etc.) nécessitent l'utilisation de méthodes probabilistes.

Quant à l'utilité de la statistique, elle est vaste. Mentionons ici seulement les aspects inférentiels, c'est-à-dire comment à partir d'un nombre fini d'observations (ou de mesures) pouvons-nous inférer (déterminer, prédire) les valeurs des grandeurs observées et comment déterminer des intervalles de confiance pour nos prédictions.

1.2 Probabilités et intuition aléatoire

Lorsque nous lançons une pièce honnête, il est intuitivement clair que nous nous attendons à obtenir pile avec « probabilité » 50% et face avec la même probabilité et ceci malgré le fait que le lancer de la pièce est une opération purement déterministe, régie par les équations de la mécanique newtonienne. L'aléa ne traduit que la connaissance incomplète de la condition initiale et des caractéristiques mécaniques de l'usinage de la pièce.

Dans la phrase précédente, nous n'avons pas définie le terme « probabilité ». La signification intuitive que nous donnons est que si nous répétons l'expérience N fois (ou que nous faisons l'expérience en lançant simultanément N pièces identiques) et nous notons $N(F)$ le nombre de fois que nous obtenons « face », le rapport $\mathbb{P}_N(F) = \frac{N(F)}{N}$ tend vers $\mathbb{P}(F) = 1/2$ lorsque $N \to \infty$. Nous assignons alors le nombre $\mathbb{P}(F)$ à la « probabilité d'obtenir face ». Dans cette approche, la notion de probabilité est alors identifiée à la notion de la limite de la fréquence relative de réalisation de l'événement $F =$ « obtenir face » $= \{\text{face}\}$, F étant considéré comme partie d'un ensemble universel $\Omega = \{\text{pile,face}\}$ de tous les résultats possibles de l'expérience qui consiste à lancer une pièce.

La formulation axiomatique [42] de la **théorie des probabilités** introduit la notion d'espace probabilisé comme étant le couple [1] $(\Omega, \mathbb{P})$ où Ω est un ensemble universel de tous les résultats possibles d'une expérience et $\mathbb{P}$ une fonction qui à chaque partie $F \subseteq \Omega$ associe un nombre $\mathbb{P}(F) \in [0,1]$, sa probabilité. L'interprétation fréquentielle devient alors un théorème de la théorie des probabilités, connu sous le nom de **loi des grands nombres** (cf. chapitre 4).

1.3 Esquisse du problème d'estimation statistique

Il n'est pas aisé de donner une description des problématiques et des méthodes de la statistique mathématique sans une exposition préalable de la théorie des probabilités. C'est pourquoi dans ce chapitre introductif, nous nous limitons à présenter un exemple concret.

Supposons que nous disposions d'une pièce qui donne « face » avec une probabilité θ fixe mais inconnue. Nous voulons estimer cette valeur en effectuant une expérience avec cette pièce. Par exemple, on peut la lancer N fois (N est **nécessairement** fini pour que l'expérience soit physiquement réalisable), on note $X_n \in \{0,1\}$ le résultat de chaque lancer et on calcule

$$\theta_N = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{\{0\}}(X_n).$$

On verra que θ_N est un **estimateur asymptotique** de θ dans le sens que $\lim_{N\to\infty} \theta_N = \theta$ (la limite a lieu dans un certain sens qui sera précisé dans ce cours) et que pour un certain seuil de confiance c, on peut trouver un **intervalle de confiance** [2] $I_N = [\theta_N -$

1. Nous verrons dans le chapitre 2 que cette construction n'est pas possible si Ω n'est pas dénombrable ; la construction de la fonction $\mathbb{P}$ n'est en général possible que sur une famille spécifique $\mathcal{F} \subset \mathcal{P}(\Omega)$, appelée **tribu**. Le choix de $\mathcal{F} = \mathcal{P}(\Omega)$ n'est possible que dans le cas dénombrable ; dans le cas non-dénombrable, les tribus $\mathcal{F}$ *utiles* sont beaucoup plus petites que $\mathcal{P}(\Omega)$ (cf. théorème 2.1.13). Un espace probabilisé sera donc la donnée du triplet $(\Omega, \mathcal{F}, \mathbb{P})$ et non du couple $(\Omega, \mathbb{P})$.

2. Cet intervalle est aléatoire.

$a_N, \theta_N + a_N]$ tel que la probabilité pour que la vraie valeur (déterministe) θ appartienne dans I_N dépasse le seuil c, c'est-à-dire $\mathbb{P}(\theta \in I_N) \geq c$.

1.4 Exercices

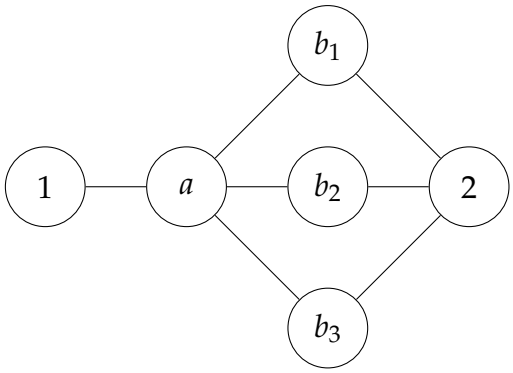
Le premier exercice est purement combinatoire ; nous introduisons les 4 types de dénombrement. Pour les exercices sur les probabilités élémentaires de ce chapitre, nous laissons de côté les questions de définition de la notion de probabilité en nous limitant à la définition empirique (fréquentielle) des probabilités exposée plus haut. Dans ce contexte, les probabilités de ces exercices se calculent comme des rapports de cardinaux.

Les quatre types de dénombrement

1. On rappelle les quatre types de dénombrement :
   — n tirages discernables à choisir parmi M possibilités avec remise ($n > 0, M > 0$) :
   $$\Omega_1 = \{\omega = (\omega_1, \ldots, \omega_n); \omega_i \in \{1, \ldots, M\}, i = 1, \ldots, n\}.$$
   $\mathsf{card}\Omega_1 = M^n$.
   — n tirages discernables à choisir parmi M possibilités sans remise ($0 < n \leq M$) :
   $$\Omega_2 = \{\omega = (\omega_1, \ldots, \omega_n); \omega_i \in \{1, \ldots, M\}, i = 1, \ldots, n; \omega_i \neq \omega_j, \text{pour } i \neq j\}.$$
   $\mathsf{card}\Omega_2 = \frac{M!}{(M-n)!}$.
   — n tirages indiscernables à choisir parmi M possibilités avec remise ($n > 0, M > 0$) :
   $$\Omega_3 = \{\omega = [\omega_1, \ldots, \omega_n]; \omega_i \in \{1, \ldots, M\}, i = 1, \ldots, n\}.$$
   $\mathsf{card}\Omega_3 = C^n_{n+M-1}$.
   — n tirages indiscernables à choisir parmi M possibilités sans remise ($0 < n \leq M$) :
   $$\Omega_4 = \{\omega = [\omega_1, \ldots, \omega_n]; \omega_i \in \{1, \ldots, M\}, i = 1, \ldots, n; \omega_i \neq \omega_j, \text{pour } i \neq j\}.$$
   $\mathsf{card}\Omega_4 = \frac{M!}{n!(M-n)!} = C^n_M$.
   Pour chacun de ces quatres types,
   (a) démontrer les formules des cardinaux,
   (b) donner un exemple concret d'application.

Événements et ensembles

2. Sous quelles conditions les événements A et B satisfont l'égalité $A = A \cup B$? L'égalité $A = A \cap B$?
3. Soit $(A_i)_{i \in I}$ une collection d'ensembles. Déterminer $(\cup_{i \in I} A_i)^c$ et $(\cap_{i \in I} A_i)^c$.
4. La figure suivante décrit un circuit électrique entre les points 1 et 2 comportant les fusibles a, b_1, b_2 et b_3 qui peuvent tomber en panne.



Noter A (resp. B_i) les événements : « le fusible a (resp. le fusible b_i) est en panne », pour $i = 1, 2, 3$ et C l'événement « le courant passe de 1 à 2 ». Déterminer C^c et C en termes des événements A, B_1, B_2 et B_3.

5. Une cible est constituée de 10 disques concentriques de rayons $r_k, k = 1, \ldots, 10$. L'événement A_k signifie « le projectile a touché la cible à l'intérieur du disque de rayon r_k », pour $k = 1, \ldots, 10$. Quelle est la signification des événements B et C définis par
$$B = \cup_{k=1}^{5} A_k \text{ et } C = \cap_{k=1}^{10} A_k ?$$
6. 20 chevaux sont au départ d'une course. Trouver le nombre de tiercés, de quartés, de quintés dans l'ordre et dans le désordre.

Probabilités élémentaires

7. Un ouvrage de 4 volumes est placé dans un ordre aléatoire sur le rayonnage d'une bibliothèque. Quelle est la probabilité que les 4 volumes soient placés dans l'ordre (ascendant ou descendant) ? *Rép. : 1/12.*
8. Toutes les faces d'un cube en bois sont peintes. Ensuite le cube est découpé (selon des plans parallèles à ses faces) en 1000 cubes identiques. Un petit cube est choisi au hasard ; quelle est la probabilité qu'il comporte exactement 2 faces peintes ? *Rép. : 0,096.*
9. Dix livres sont placés dans un ordre aléatoire sur un rayonnage. Quelle est la probabilité que trois livres spécifiques se retrouvent côte-à-côte ? *Rép. : 1/15.*
10. Un sac contient 5 bâtonnets de longueurs 1, 3, 5, 7 et 9. On en extrait 3. Quelle est la probabilité que l'on puisse construire un triangle ayant ces bâtonnets comme côtés ? *Rappel : La longueur d'un côté d'un triangle ne peut pas excéder la somme des longueurs des autres côtés.*
11. Supposons qu'un entier entre 1 et 1000 est choisi au hasard. Quelle est la probabilité que les 2 derniers digits décimaux de son cube soient 1 ? *Suggestion : Il n'est pas nécessaire de consulter une table des cubes des tous les entiers entre 1 et 1000.* *Rép. : 0,01.*
12. Quelle est la probabilité qu'au moins deux étudiants suivant un cours auquel N étudiants sont inscrits ($N \geq 2$) aient leur anniversaire le même jour ? On néglige les années bissextiles et on suppose que tous les jours de l'année sont équiprobables en tant que jour de naissance.
13. Une tombola comporte M tickets dont n (avec $M \geq 2n$) sont gagnants. Une personne achète n tickets. Quelle est la probabilité pour qu'elle gagne au moins un lot ?

14. On gagne (au premier rang) une loterie si l'on coche 6 bons numéros parmi les 49 que comporte une grille. Quelle est la probabilité de gagner cette loterie au premier rang ?

Première partie

Théorie des probabilités ; statistique mathématique

2

Théorie élémentaire des probabilités

Nous présentons la théorie telle qu'elle a été formulée par Kolmogorov [42] et nous nous inspirons librement des exposés pédagogiques [29, 61, 65, 69], en particulier pour le choix de certains exemples. Les livres [2, 8, 52] peuvent utilement être consultés mais sont mathématiquement plus exigeants.

La formulation de Kolmogorov est une construction mathématique abstraite. Cependant, déjà dans l'œuvre originale [42] transparaît une forte intuition expérimentale ; pour comprendre les notions de base de la théorie, il est très utile de nous servir des modèles venant d'autres domaines de mathématiques ou de la physique.

La présentation naturelle de la théorie requiert la théorie de la mesure. Dans la suite nous essaierons de donner une présentation **élémentaire** i.e. qui ne fait pas appel à la théorie de la mesure.

2.1 Espace de probabilité

2.1.1 Espace des épreuves, espace des événements

Épreuves

Lors d'une expérience aléatoire, on note Ω l'ensemble de tous les résultats possibles et imaginables. Les éléments de Ω sont notés ω et appelés **épreuves**. L'**espace des épreuves** Ω peut être dénombrable (fini ou infini) ou non dénombrable et s'appelle aussi **univers**.

Événements

Les parties de Ω — plus précisément *certaines* parties de Ω — sont appelés **événements** ; Ω lui-même est toujours un événement (l'événement certain) ainsi que $\emptyset$ (l'événement impossible). Intuitivement, un événement est un ensemble d'épreuves qui vérifient une propriété. Si la famille $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ est une famille d'événements, il est naturel de demander qu'elle constitue une **algèbre**.

Définition 2.1.1. Une famille $\mathcal{A}$ de parties de Ω est une **algèbre** si

1. $\Omega \in \mathcal{A}$,
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$ (stabilité par complémentation) et
3. $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cup A_2 \in \mathcal{A}$ (stabilité aux réunions finies).

Remarque 2.1.2. De la définition 2.1.1 d'une algèbre $\mathcal{A}$ sur Ω découlent immédiatement les faits suivants :

1. $\emptyset \in \mathcal{A}$ et
2. $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cap A_2 \in \mathcal{A}$ (stabilité aux intersections finies).

Exemple 2.1.3 (Pile ou face)**.** L'espace des épreuves pour le lancer d'une pièce est $\Omega = \{\text{pile}, \text{face}\} \equiv \{0,1\}$. La famille $\mathcal{A} = \mathcal{P}(\Omega)$ est une algèbre qui correspond à tous les événements possibles.

Exemple 2.1.4 (Pile ou face répété n fois)**.** L'espace des épreuves pour cette expérience est $\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in \{0,1\}, i = 1, \ldots, n\} \equiv \{0,1\}^n$. On remarque que $\text{card}\Omega = 2^n$. La famille $\mathcal{A} = \mathcal{P}(\Omega)$ est une algèbre qui correspond à tous les événements possibles ; elle a $\text{card}\mathcal{A} = 2^{2^n}$.

Exemple 2.1.5 (Pile ou face répété indéfiniment)**.** Parfois on s'intéresse à des suites infinies d'expériences (par exemple une suite de bits indéfiniment longue). L'espace des épreuves pour cette expérience est $\Omega = \{\omega = (\omega_1, \omega_2, \ldots) : \omega_i \in \{0,1\}, i \in \mathbb{N}\} \equiv \{0,1\}^{\mathbb{N}}$. On remarque que $\text{card}\Omega = 2^{\aleph_0} = \mathfrak{c}$ (donc $\Omega \cong \mathbb{R}$).

Remarque 2.1.6. Le dernier exemple établit que si nous voulons considérer des limites de suites — considérer la suite donc dans son intégralité — nous sommes obligés de sortir du cadre discret.

Lorsque l'espace Ω est dénombrable infini (ou *a fortiori* non dénombrable), il est naturel de demander la stabilité de la famille des événements aux réunions (intersections) dénombrables. Ceci nous amène à la

Définition 2.1.7. Une famille $\mathcal{F}$ de parties de Ω est dite une **tribu** (ou σ-algèbre) si elle est une algèbre et vérifie la propriété de stabilité aux intersections dénombrables :

$$[\forall n \in \mathbb{N}, A_n \in \mathcal{F}] \Rightarrow [\cup_{n\in\mathbb{N}} A_n \in \mathcal{F}].$$

Exemple 2.1.8. Soit Ω un ensemble arbitraire. Alors $\mathcal{F}_1 = \{\emptyset, \Omega\}$ et $\mathcal{F}_2 = \mathcal{P}(\Omega)$ sont des tribus. $\mathcal{F}_1$ est appelée tribu grossière, $\mathcal{F}_2$ est appelée tribu exhaustive. Lorsque $\text{card}\Omega = N$, alors $\text{card}\mathcal{P}(\Omega) = 2^N$. Lorsque $\text{card}\Omega = \aleph_0$ alors $\text{card}\mathcal{P}(\Omega) = \mathfrak{c}$. Lorsque $\text{card}\Omega = \mathfrak{c}$ alors $\text{card}\mathcal{P}(\Omega) = 2^{\mathfrak{c}}$ ce qui montre que la tribu exhaustive dans le cas d'un espace des épreuves avec la puissance du continu est un objet énorme (dit aussi non-énumerable).

Définition 2.1.9. Soit Ω un espace des épreuves et $\mathcal{F}$ une tribu sur Ω. Le couple $(\Omega, \mathcal{F})$ est appelé **espace des événements** (ou espace mesurable dans la terminologie de la théorie de la mesure).

2.1.2 Probabilisation

Définition 2.1.10. Soit $(\Omega, \mathcal{F})$ un espace des événements. Supposons qu'il existe une application $\mathbb{P} : \mathcal{F} \to [0,1]$ vérifiant

1. $\mathbb{P}(\Omega) = 1$ (normalisation de la masse totale) et
2. $[\forall n \in \mathbb{N}, A_n \in \mathcal{F} \text{ et } A_m \cap A_n = \emptyset, m \neq n] \Rightarrow [\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)]$ (additivité dénombrable disjointe dite aussi σ-additivité disjointe).

Alors $\mathbb{P}$ est appelée (mesure de) **probabilité** sur $(\Omega, \mathcal{F})$ et le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ est alors appelé **espace probabilisé**.

Proposition 2.1.11. *Supposons que Ω est dénombrable (fini ou infini) et qu'il existe une application $\rho : \Omega \to [0,1]$ telle que $\sum_{\omega\in\Omega} \rho(\omega) = 1$. Alors ρ définit une mesure de probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{F})$, où $\mathcal{F} = \mathcal{P}(\Omega)$, par*

$$\mathcal{F} \ni A \mapsto \mathbb{P}(A) = \sum_{\omega\in A} \rho(\omega) \in [0,1].$$

Reciproquement, si $\mathbb{P}$ est une probabilité sur $(\Omega, \mathcal{P}(\Omega))$ et Ω est dénombrable, alors il existe $\rho : \Omega \to [0,1]$ telle que $\sum_{\omega\in\Omega} \rho(\omega) = 1$ permettant de définir $\mathbb{P}$ comme ci-dessus. (En fait $\rho(\omega) = \mathbb{P}(\{\omega\})$ pour tout $\omega \in \Omega$).

Démonstration. Cf. exercice 20. □

L'application ρ de la proposition précédente est appelée **vecteur de probabilité** ou **densité discrète** de $\mathbb{P}$. Cette proposition établit en outre une bijection entre les mesures de probabilités sur les espaces dénombrables et les vecteurs de probabilité.

Exemple 2.1.12. Probabilité sur des ensembles discrets.

1. Soient $\Omega = \{1,2,3,4,5,6\}$ et $\rho(\omega) = 1/6 = 1/\mathsf{card}(\Omega)$ pour tout $\omega \in \Omega$. Ce vecteur de probabilité définit une probabilité uniforme sur $(\Omega, \mathcal{P}(\Omega))$ correspondant à un dé honnête. Noter qu'une probabilité uniforme $\mathbb{P}$ sur un ensemble fini Ω s'exprime alors nécessairement comme un rapport de cardinaux : $\mathbb{P}(A) = \frac{\mathsf{card}(A)}{\mathsf{card}(\Omega)}$, pour tout $A \in \mathcal{P}(\Omega)$.
2. Soient $\Omega = \{1,2,3,4,5,6\}$ et $p(1) = p(6) = 1/4$ et $p(2) = p(3) = p(4) = p(5) = 1/8$ définit une probabilité non-uniforme sur $(\Omega, \mathcal{P}(\Omega))$ correspondant à un dé lesté qui donne 6 (ou 1) deux fois plus fréquemment que les autres faces.
3. Soient $\Omega = \mathbb{N}$ et $\rho(\omega) = 1/2^{\omega+1}$. Ce vecteur de probabilité définit une probabilité non-uniforme sur $(\Omega, \mathcal{P}(\Omega))$. Noter qu'il n'est pas possible de définir une probabilité uniforme sur un ensemble dénombrable infini.

L'exercice 21 montre que l'ensemble fini $\Omega = \{0,1\}^n$ peut être probabilisé par un vecteur de probabilité vérifiant en outre la propriété d'invariance à l'application T_k « renversement du k^{e} bit » définie par

$$\Omega \ni \omega = (\omega_1, \ldots, \omega_n) \mapsto T_k(\omega) = ((\omega_1, \ldots, \omega_{k-1}, 1-\omega_k, \omega_{k+1}, \ldots, \omega_n), k = 1, \ldots, n.$$

Cette construction ne peut pas s'étendre au cas $\Omega = \{0,1\}^{\mathbb{N}}$ comme le montre le

Théorème 2.1.13 (Vitali 1905)**.** *Soit $\Omega = \{0,1\}^{\mathbb{N}}$ l'ensemble des suites infinies de bits. Il n'existe pas d'application $\mathbb{P} : \mathcal{P}(\Omega) \to [0,1]$ vérifiant sumultanément*

— *(normalisation) :* $\mathbb{P}(\Omega) = 1$,

— *(additivité dénombrable disjointe) : si* (A_n) *est une suite d'événements mutuellement disjoints,* $\mathbb{P}(\sqcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$,
— *(invariance aux renversements des bits) : si* $A \in \mathcal{P}(\Omega)$ *alors* $\mathbb{P}(T_k A) = \mathbb{P}(A)$ *pour tout* k.

La démonstration fait appel à l'axiome de choix [1], rappelé en 2.1.14 ; elle peut être omise en première lecture.

Axiome 2.1.14 (de choix). *Pour toute collection* $\mathcal{C}$ *d'ensembles non-vides, il existe une* ***fonction de choix*** *définie sur* $\mathcal{C}$ *qui à chaque ensemble* $A \in \mathcal{C}$ *associe un élément de cet ensemble. Plus précisément :*

$$\forall \mathcal{C}[\emptyset \notin \mathcal{C} \Rightarrow \exists f : \mathcal{C} \to \cup\mathcal{C} \text{ t.q. } \forall A \in \mathcal{C}, (f(A) \in A].$$

Démonstration du théorème 2.1.13. On introduit une relation d'équivalence sur Ω en décrétant que deux suites $\omega, \omega' \in \Omega$ sont équivalentes si elles sont identiques à partir d'une certaine position, i.e.

$$[\omega \sim \omega'] \Leftrightarrow [\exists n_0 \in \mathbb{N} : \omega_n = \omega'_n, \forall n \geq n_0].$$

Par l'axiome de choix, il existe une partie $A \subseteq \Omega$ qui contient exactement un élément de chaque classe d'équivalence de $\Omega/\sim$.

On note $\mathcal{S} = \{S \subset \mathbb{N} : \text{card} S < \infty\} \subseteq \sqcup_{\ell\in\mathbb{N}} \mathcal{S}_\ell$, où $\mathcal{S}_\ell = \{S \subset \mathbb{N} : \max S = \ell\}$. De toute évidence, $\mathcal{S}$ est dénombrable. Si $S \in \mathcal{S}$, on définit $T_S = \circ_{k\in S} T_k$ (l'ordre de composition des fonctions est sans importance car les applications T_k et $T_{k'}$ commutent pour $k \neq k'$). On observe que

— La famille $(T_S A)_{S\in\mathcal{S}}$ est composée de parties deux-à-deux disjoints. En effet, supposons que $T_S A \cap T_{S'} A \neq \emptyset$ pour $S, S' \in \mathcal{S}$. Il existe alors $\omega, \omega' \in A$ tels que $T_S\omega = T_{s'}\omega'$. Nous aurons alors,
$$\omega \sim T_S\omega = T_{s'}\omega' \sim \omega'.$$
Par le choix de A, ceci signifie que $\omega = \omega'$ et par conséquent $S = S'$.
— Par ailleurs $\Omega = \sqcup_{S\in\mathcal{S}} T_S A$, car, en effet, pour tout $\omega \in \Omega$, il existe $\omega' \in A$, tel que $\omega \sim \omega'$. Par conséquent, il existe un $S \in \mathcal{S}$ tel que $\omega = T_S\omega' \sim \omega' \in T_S A$.
— En utilisant successivement les propriétés de normalisation, d'additivité dénombrable disjointe et d'invariance aux renversements des bits, nous arrivons à la conclusion absurde :

$$1 = \mathbb{P}(\Omega) = \sum_{S\in\mathcal{S}} \mathbb{P}(T_S A) = \sum_{S\in\mathcal{S}} \mathbb{P}(A) \in \{0, \infty\}.$$

□

Remarque 2.1.15. Dans le théorème précédent, $\text{card}\Omega = \mathfrak{c}$; ce que nous apprend ce théorème est qu'il n'est pas possible d'avoir une mesure de probabilité, invariante aux renversements de bits, sur la tribu exhaustive $\mathcal{P}(\Omega)$ lorsque l'ensemble Ω a le cardinal du continu. L'introduction de la notion abstraite de tribu nous permet de définir une probabilité pour certaines tribus *plus petites* que la tribu exhaustive. L'ensemble A, dans la démonstration précédente, est un ensemble de Vitali ; il fournit un exemple concret d'ensemble non mesurable (voir remarque 2.1.17 plus loin).

Dans l'exercice 18 nous introduisons la notion de **tribu engendrée** par une famille arbitraire (non-vide) $\mathcal{A}$ de parties de $\mathcal{P}(\Omega)$ comme étant la plus petite tribu qui contient $\mathcal{A}$ — notée $\sigma(\mathcal{A})$.

Définition 2.1.16 (Tribu borélienne)**.** Soient $\Omega = \mathbb{R}^d$, pour $d \geq 1$ entier, et

$$\mathcal{R}_d = \{\prod_{i=1}^{d} [a_i, b_i] : a_i < b_i, a_i, b_i \in \mathbb{Q}\},$$

1. L'axiome de choix a été introduit par Zermelo en 1904 pour formaliser la théorie axiomatique des ensembles.

le système de **pavés rectangulaires** avec des arêtes parallèles aux axes et sommets avec des coordonnées rationnelles. La tribu $\sigma(\mathcal{R}_d)$ engendrée par cette famille est appelée **tribu borélienne** sur $\mathbb{R}^d$ et notée $\mathcal{B}_d := \mathcal{B}(\mathbb{R}^d)$. (Lorsque $d = 1$, nous abrégeons la notation en $\mathcal{B}$). Les événements de $\mathcal{B}_d$ sont appelés boréliens.

Remarque 2.1.17. La tribu $\mathcal{B}_d$ est beaucoup plus grosse que l'on pouvait supposer à première vue. En particulier elle contient les classes
- des ouverts ;
- des fermés ;
- G_δ (intersections dénombrables d'ouverts) ;
- F_σ (réunions dénombrables de fermés) ;
- $G_{\delta\sigma}$ (réunions dénombrables d'éléments de G_δ) ;
- $F_{\sigma\delta}$ (réunions dénombrables d'éléments de F_σ)
- ...

Pour les besoins de ce cours il est suffisant de connaître que toutes les parties de $\mathbb{R}^d$ que nous rencontrons en pratique sont boréliennes. Montrer l'existence de parties non-boréliennes n'est pas tâche aisée, il faut recourir à l'axiome du choix. On peut en donner une en s'inspirant de la démonstration du théorème 2.1.13 ; on peut en trouver une dans [8, pp. 24–28] par exemple.

Définition 2.1.18 (Tribu produit)**.** Soient $(\Omega_n, \mathcal{F}_n)_{n\in\mathbb{N}}$ une suite d'espaces mesurables, $\Omega = \times_{n\in\mathbb{N}}\Omega_n$ et $\mathcal{R} = \times_{n\in\mathbb{N}}\mathcal{F}_n$. La tribu $\mathcal{F} = \sigma(\mathcal{R})$ sur Ω, engendrée par le produit cartésien $\mathcal{R}$ de tribus est appelée **tribu produit** et notée $\mathcal{F} = \otimes_{n\in\mathbb{N}}\mathcal{F}_n$.

Théorème 2.1.19 (Propriétés de $\mathbb{P}$)**.** *Toute mesure de probabilité $\mathbb{P}$ sur un espace d'événements $(\Omega, \mathcal{F})$ vérifie les propriétés suivantes pour des événements arbitraires $A, B, A_1, A_2, \ldots \in \mathcal{F}$:*

1. $\mathbb{P}(\emptyset) = 0$.
2. *Additivité finie.* $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$. *En particulier* $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.
3. *Monotonie. Si* $A \subseteq B$ *alors* $\mathbb{P}(A) \leq \mathbb{P}(B)$.
4. *σ-sous-additivité.* $\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$.
5. *σ-continuité monotone. Si soit $A_n \uparrow A$ soit $A_n \downarrow A$ (i.e. la suite (A_n) est soit croissante avec réunion A soit décroissante avec intersection A), alors* $\mathbb{P}(A_n) \xrightarrow[n\to\infty]{} \mathbb{P}(A)$.

Démonstration. 1. Puisque l'ensemble vide est disjoint de tout événement, donc de lui-même, la suite constante $\emptyset, \emptyset, \emptyset, \ldots$ est composée d'événements mutuellement disjoints. La propriété de additivité dénombrable disjointe de $\mathbb{P}$ (propriété 2, déf. 2.1.10) devient

$$\mathbb{P}(\emptyset) = \mathbb{P}(\emptyset \cup \emptyset \cup \ldots) = \sum_{n\in\mathbb{N}} \mathbb{P}(\emptyset).$$

Cette égalité n'est possible que si $\mathbb{P}(\emptyset) = 0$.

2. Supposons pour commencer que A et B sont disjoints. Or la propriété 2 de la définition 2.1.10 requiert une suite infinie d'événements disjoints. On complète donc par des ensembles vides et on a

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \cup B \cup \emptyset \cup \emptyset \cup \ldots) = \mathbb{P}(A) + \mathbb{P}(B) + 0 + 0 + \ldots.$$

Dans le cas général, nous décomposons

$$\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + 2\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B).$$

Le cas particulier s'obtient en posant $B = A^c$ et en utilisant la condition de normalisation (propriété 1, déf. 2.1.10).

3. Si $A \subseteq B$, alors de la propriété d'additivité finie établie en 2, on obtient $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$ car les probabilités sont toujours positives.
4. Toute réunion $\cup_{n\in\mathbb{N}} A_n$ peut s'écrire comme réunion d'événements mutuellement disjoints : On aura alors

$$\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \mathbb{P}(\sqcup_{n\in\mathbb{N}}(A_n \setminus \cup_{m<n} A_m) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n \setminus \cup_{m<n} A_m) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

5. Supposons que $A_n \uparrow A$; alors nécessairement $A \in \mathcal{F}$ (pourquoi ?). Sans perte de généralité, on peut supposer que $A_0 = \emptyset$. On a alors

$$\mathbb{P}(A) = \mathbb{P}(\sqcup_{n\geq 1}(A_n \setminus A_{n-1})) = \sum_{n\geq 1} \mathbb{P}(A_n \setminus A_{n-1})$$
$$= \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{P}(A_n \setminus A_{n-1}) = \lim_{N\to\infty} \mathbb{P}(A_N).$$

Le cas de suite décroissante est laissé en exercice.

□

Pour montrer que deux probabilités $\mathbb{P}$ et $\mathbb{P}'$ sur $(\Omega, \mathcal{F})$ coïncident, il faut montrer qu'elles coïncident sur tous les éléments de la tribu $\mathcal{F}$. Or, nous avons vu que les tribus sont des objets compliqués, difficilement maniables. Le théorème suivant nous donne un moyen très efficace de montrer plusieurs propriétés de probabilités sans avoir recours explicitement à la tribu sous-jacente.

Théorème 2.1.20 (Théorème d'unicité). *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et supposons que $\mathcal{F} = \sigma(\mathcal{G})$ où $\mathcal{G} \subseteq \mathcal{P}(\Omega)$ est une famille génératrice stable par intersections finies. Alors $\mathbb{P}$ est uniquement déterminée par sa restriction $\mathbb{P}\restriction_{\mathcal{G}}$.*

La démonstration de ce théorème peut être omise dans le cadre de ce cours ; elle peut être trouvée dans [29, Théorème 1.12], par exemple.

Nous sommes maintenant en mesure de construire une probabilité sur un espace non-dénombrable.

Proposition 2.1.21. *Soit $\Omega = \mathbb{R}^d$, muni de sa tribu borélienne. Supposons qu'il existe $\rho : \Omega \to \mathbb{R}_+$ vérifiant les propriétés :*

1. *pour tout $c > 0$, on a $\{\rho \leq c\} := \{\omega \in \Omega : \rho(\omega) \leq c\} \in \mathcal{B}_d$,*
2. *$\int_\Omega \rho(\omega) d\omega_1 \cdots d\omega_d = 1$.*

Alors, il existe une unique probabilité $\mathbb{P}$ sur $(\mathbb{R}^d, \mathcal{B}_d)$, définie par

$$\mathbb{P}(B) = \int_B \rho(\omega)\lambda_d(d\omega) = \int_\Omega \mathbb{1}_B(\omega)\rho(\omega)\lambda_d(d\omega),$$

où $\lambda_d(d\omega) = d\omega_1 \ldots d\omega_d$ désigne la mesure de Lebesgue en dimension d.

Démonstration. La seule chose à montrer est la propriété d'additivité dénombrable disjointe. Or si $(B_n)_n$ est une suite de boréliens deux-à-deux disjoints, on a $\mathbb{1}_{\cup_{n\in\mathbb{N}} B_n} = \sum_{n\in\mathbb{N}} \mathbb{1}_{B_n}$. La propriété découle du théorème de convergence monotone. □

Remarque 2.1.22. La proposition 2.1.21 est l'analogue continu de la proposition 2.1.11, établi dans le cas discret. La fonction ρ est l'analogue de la notion de vecteur de probabilité et porte le nom de **densité de probabilité** (par rapport à la mesure de Lebesgue). Noter cependant que contrairement au cas dénombrable — où les probabilités sont en bijection avec les densités discrètes —, Il n'est pas vrai que toutes les probabilités sur $(\mathbb{R}^d, \mathcal{B}_d)$ s'expriment sous cette forme. En général, toute probabilité sur $(\mathbb{R}^d, \mathcal{B}_d)$ se décompose en trois parties : une partie discrète, une partie qui s'écrit sous la forme de la proposition 2.1.21 (appelée absolument continue par rapport à Lebesgue) et une partie qui n'est ni discrète, ni absolument continue et s'appelle singulière continue (cf. [2, exemple III.2.5, p. 47]). Cependant, nous n'aurons pas à considérer des probabilités avec partie singulière continue dans ce cours. Pour utiliser une terminologie unifiée dans les cas discret et continu, nous pouvons appeler le vecteur de probabilité ρ de la proposition 2.1.11 aussi densité de probabilité (par rapport à la mesure de dénombrement) ou densité discrète.

Exemple 2.1.23. Probabilité sur des ensembles non-dénombrables.

1. Soient $\Omega = [0,1]$ et $\rho(\omega) = 1$ pour tout $\omega \in \Omega$. La fonction densité ρ définit une probabilité sur $([0,1], \mathcal{B}([0,1]))$, qui est appelée probabilité uniforme sur $[0,1]$.
2. Soient $\Omega = \mathbb{R}$ et $\rho(\omega) = \frac{\exp(-\omega^2/2\sigma^2)}{\sqrt{2\pi\sigma}}$. Alors ρ est une densité, i.e. $\int_{\mathbb{R}} \rho(\omega)\lambda(d\omega) = 1$ (pouvez-vous le montrer ?) qui définit une probabilité non-uniforme sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ appelée **probabilité gaussienne centrée de variance** σ pour des raisons qui seront explicitées plus loin dans ce texte. Il s'agit d'une des lois les plus importantes en théorie des probabilités, nous donnons donc ci-dessous quelques détails sur elle.

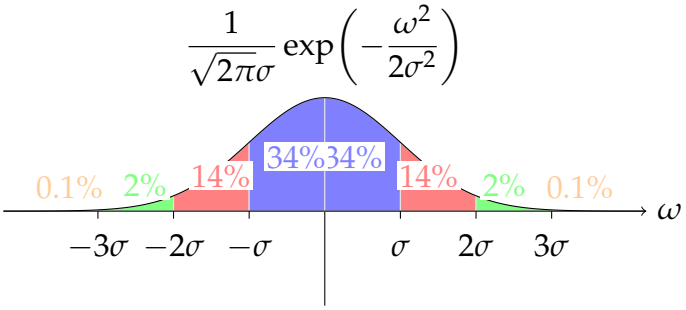


FIGURE 2.1 – La masse de la loi gaussienne s'étend sur tout l'axe $\mathbb{R}$. Cependant, elle reste essentiellement concentrée sur l'intervalle $[-3\sigma, 3\sigma]$.

Définition 2.1.24. Sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, un événement $N \in \mathcal{F}$ avec $\mathbb{P}(N) = 0$ est appelé **négligeable**; un événement $S \in \mathcal{F}$ avec $\mathbb{P}(S) = 1$ est appelé **presque sûr**.

Il n'est pas vrai en général qu'un événement négligeable est vide; ni qu'un ensemble presque sûr coïncide avec l'univers tout entier comme le contre-exemple suivant nous enseigne.

Contre-exemple 2.1.25. Soient $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \lambda)$, $N = [0,1] \cap \mathbb{Q}$ et $S = [0,1] \setminus \mathbb{Q}$. Alors $\mathbb{P}(N) = 0$ et $\mathbb{P}(S) = 1 - \mathbb{P}(N) = 1$. Pourtant, N n'est pas vide (il est même dense dans $[0,1]$) et $S = [0,1] \setminus \mathbb{Q}$ (il manque à S une infinité de points pour qu'il devienne égal à Ω).

2.2 Variables aléatoires

Supposons que nous lançons une pièce 3 fois de suite. L'espace des épreuves sera alors l'espace de configurations possibles de 3 bits : $\Omega = \{\omega = (\omega_1, \omega_2, \omega_3) : \omega_i \in \{0,1\}\}$; il sera muni de la tribu exhaustive $\mathcal{F} = \mathcal{P}(\Omega)$. Mais supposons que l'information que nous intéresse est le nombre de fois que la pièce tombe sur la face « 1 ». L'information codée dans ω est surabondante. Ce qui nous intéresse est une information plus sommaire, à savoir la valeur que prend l'application $X : \Omega \to \mathbb{X} := \{0,1,2,3\}$ définie par $X(\omega) = \sum_{i=1}^{3} \mathbb{1}_{\{1\}}(\omega_i)$. Si la pièce est honnête, l'espace des événements $(\Omega, \mathcal{F})$ est probabilisé par la densité discrète uniforme $\rho(\omega) = 1/2^3$ pour tout ω. Cette probabilité induit une probabilité $\mathbb{P}_X$ sur $(\mathbb{X}, \mathcal{P}(\mathbb{X}))$ définie par la densité non uniforme $\rho_X(0) = \rho_X(3) = 1/8$ et $\rho_X(1) = \rho_X(2) = 3/8$. Cet exemple fournit l'archétype d'une variable aléatoire, notion formalisée dans la

Définition 2.2.1. Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, $(\mathbb{X}, \mathcal{X})$ un espace d'événements ($\mathcal{F}$ et $\mathcal{X}$ sont des tribus) et $X : \Omega \to \mathbb{X}$ telle que pour tout $A \in \mathcal{X}$, on a $X^{-1}(A) \in \mathcal{F}$. Alors X est appelée **variable aléatoire** sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $(\mathbb{X}, \mathcal{X})$. Elle induit une probabilité $\mathbb{P}_X$ sur $(\mathbb{X}, \mathcal{X})$ par

$$\mathcal{X} \ni A \mapsto \mathbb{P}_X(A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}) =: \mathbb{P}(X^{-1}(A)).$$

La probabilité $\mathbb{P}_X$ sur $(\mathbb{X}, \mathcal{X})$ est appelée **loi de la variable aléatoire** X.

Remarque 2.2.2. La condition $X^{-1}(A) \in \mathcal{F}$ pour tout $A \in \mathcal{X}$ est essentielle dans les cas où les tribus ne sont pas exhaustives. Considérer par exemple $\Omega = \{0,1\}^3$, $\mathbb{X} = \{0,1,2,3\}$ et $X(\omega) = \sum_{i=1}^{3} \omega_i$. Si l'on munit $\mathbb{X}$ avec la tribu exhaustive $\mathcal{X} = \mathcal{P}(\mathbb{X})$ tandis que Ω est muni de la tribu $\mathcal{F} = \sigma(F) = \{\emptyset, \Omega, F, F^c\}$ avec $F := \{011, 101, 111\}$, alors X *n'est pas une variable aléatoire* car par exemple $X^{-1}(\{2\}) = \{011, 101, 110\} \notin \mathcal{F}$. Par contre, la même application *est une variable aléatoire* si $\mathcal{F}$ est la tribu exhaustive.

Remarque 2.2.3. La donnée importante pour une variable aléatoire est sa loi $\mathbb{P}_X$ (une probabilité sur $(\mathbb{X}, \mathcal{X})$). L'espace abstrait $(\Omega, \mathcal{F}, \mathbb{P})$ n'est qu'une construction auxiliaire. Le choix du triplet $(\Omega, \mathcal{F}, \mathbb{P})$ reflète le moyen où le système décrit par la variable aléatoire X est construite et il n'est pas unique comme le montre l'exemple 2.2.4.

Exemple 2.2.4. (Trois manières radicalement différentes de jouer au pile ou face). Jouer au pile ou face équivaut à construire une probabilité $\mathbb{P}_X = \frac{1}{2}\varepsilon_0 + \frac{1}{2}\varepsilon_1$ sur $\mathbb{X} = \{0,1\}$, muni de sa tribu exhaustive $\mathcal{X} = \mathcal{P}(\mathbb{X})$.

Monsieur Tout-le-monde joue au pile ou face. On lance une pièce sur une table approximativement considérée infiniment étendue dans le plan horizontale et infiniment plastique. L'état instantané de la pièce est un élément de $\Omega = (\mathbb{R}_+ \times \mathbb{R}^2) \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{S}^2$. Cet espace code la position du centre de masse $\mathbf{R}$, la vitesse du centre de masse $\mathbf{V}$, le moment angulaire $\mathbf{M}$ et l'orientation $\mathbf{N}$ de la normale extérieure sur la face « face ». On choisit pour $\mathcal{F} = \mathcal{B}(\Omega)$ et $\mathbb{P}$ désigne

une probabilité à support compact dans Ω, décrivant l'incertitude de la condition initiale du système mécanique. Une fois la pièce lancée, son mouvement est régi par les équations de Newton. La pièce touche la table dans un temps aléatoire $T(\omega) = \inf\{t > 0 : R_3(t) = 0; \mathbf{V}(t) = 0, \mathbf{M}(t) = 0\}$ et elle s'arrête immédiatement à cause de la plasticité de la table. On définit

$$X(\omega) = \begin{cases} 0 & \text{si} \quad \mathbf{N}(T(\omega)) \cdot \mathbf{e}_3 = -1 \\ 1 & \text{si} \quad \mathbf{N}(T(\omega)) \cdot \mathbf{e}_3 = 1, \end{cases}$$

où $\mathbf{e}_3$ est le vecteur unité vertical de $\mathbb{R}^3$. La cause de l'aléa est la stratification très fine (cf. figure 2.2) de l'espace Ω. Ce système mécanique set étudié en détail dans [39, 19].

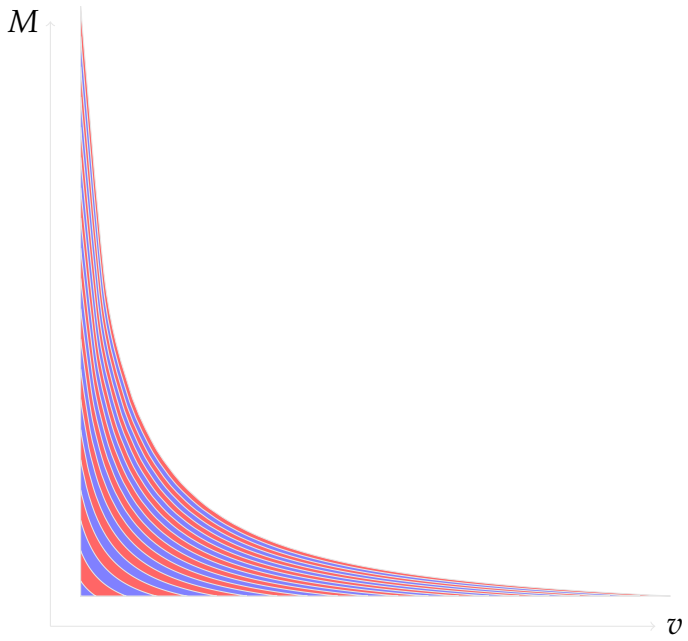


FIGURE 2.2 – L'espace de phases (v, M) (sous quelques conditions simplificatrices), où v désigne la vitesse verticale et M le moment angulaire, est stratifié en régions rouges et bleues selon les résultats possibles (pile ou face). Seulement les premières strates sont présentées ; cependant, la stratification couvre tout le quadrant.

Le simulateur (sur ordinateur) joue au pile ou face. Soient $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$ et $\mathbb{P}$ la mesure de Lebesgue sur $[0,1]$. Le résultat d'une pièce honnête est modélisé par la variable aléatoire

$$X(\omega) = \begin{cases} 0 & \text{si} \quad \omega < 1/2 \\ 1 & \text{si} \quad \omega > 1/2. \end{cases}$$

Le mathématicien joue au pile ou face. Soient $\Omega = \{0,1\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ et $\mathbb{P} = \frac{1}{2}\varepsilon_0 + \frac{1}{2}\varepsilon_1$. Le résultat d'une pièce honnête est modélisé par la variable aléatoire $X = \mathrm{id}$. En outre, $\mathbb{P}_X = \mathbb{P}$. Une telle réalisation est appelée **minimale**.

La nature probabiliste d'une variable aléatoire X est totalement codée en sa loi $\mathbb{P}_X$. On va s'intéresser à de variables aléatoires à valeurs dans $\mathbb{X} \subseteq \mathbb{R}$; dans ce cas, la loi $\mathbb{P}_X$ est à son tour totalement spécifiée par une fonction réelle.

Définition 2.2.5. Soit X une variable aléatoire réelle (i.e. à valeurs dans $\mathbb{X} \subseteq \mathbb{R}$) définie sur l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On appelle **fonction de répartition** de X (ou de sa loi $\mathbb{P}_X$) l'application

$$\mathbb{R} \ni x \mapsto F_X(x) := \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}_X(]-\infty, x]).$$

Remarque 2.2.6. Si $\mathrm{card}\Omega = m$, alors on peut choisir $\mathbb{X} = X(\Omega)$ avec $\mathrm{card}\mathbb{X} = n \leq m$, i.e. on peut écrire $\mathbb{X} = \{x_1, \ldots, x_n\}$ pour des réels $x_i, i = 1, \ldots, n$ avec $x_i \leq x_j$ si $i < j$. Alors, on pourra écrire

$$F_X(x) = \sum_{i:x_i \leq x} \mathbb{P}_X(\{x_i\})$$

où $\mathbb{P}_X(\{x_i\}) = \rho_X(x_i) = \Delta F_X(x_i) := F_X(x_i) - F_X(x_i-)$.

Exemple 2.2.7. Soit $\mathbb{X} = \{0, 1, 2\} \subset \mathbb{R}$ et $\mathbf{p} = (0.2, 0.5, 0.3)$ un vecteur de probabilité sur $\mathbb{X}$. La fonction de répartition est cette loi est représentée sur la figure 2.3.

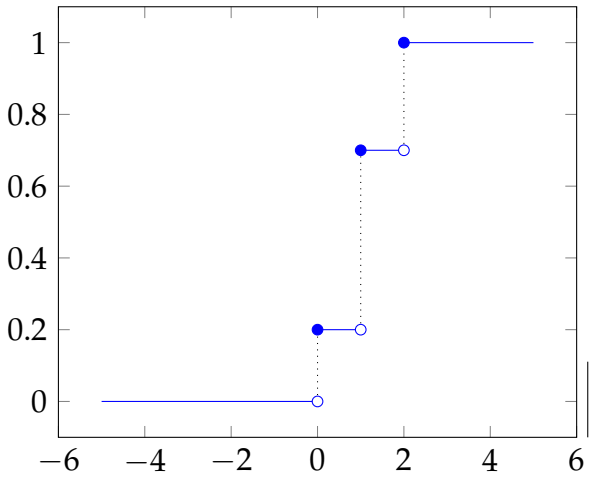


FIGURE 2.3 – Fonction de répartition de la loi sur $\mathbb{X} = \{0, 1, 2\}$ définie par le vecteur de probabilité $\mathbf{p} := (0.2, 0.5, 0.3)$.

Proposition 2.2.8. *La fonction de répartition F_X d'une variable aléatoire réelle vérifie*

1. *F_X est croissante.*
2. *$\lim_{x\to-\infty} F_X(x) = 0$ et $\lim_{x\to+\infty} F_X(x) = 1$.*
3. *F_X est continue à droite, i.e. $F_X(x+) = F_X(x)$ pour tout $x \in \mathbb{R}$.*

Démonstration. Cf. exercice 23. □

2.3 Exercices

Algèbres, tribus

15. Soit $\mathcal{A}$ une algèbre sur Ω (fini) et soient A, B et C trois événements quelconques de $\mathcal{A}$. Exprimer les événements suivants. Parmi A, B, C
    (a) A seul se produit,
    (b) A et B se produisent et C ne se produit pas,

(c) les trois événements se produisent,
(d) l'un au moins des événements se produit,
(e) au moins deux des événements se produisent,
(f) un seul événement se produit,
(g) aucun des événements ne se produit.

16. Soit $(\mathcal{A}_i)_{i\in I}$ une famille d'algèbres sur Ω. Montrer que $\cap_{i\in I}\mathcal{A}_i$ est une algèbre sur Ω. En déduire que si $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ est une famille arbitraire de parties de Ω, il existe une algèbre minimale qui contient $\mathcal{C}$; on la note $\alpha(\mathcal{C})$ et on l'appelle **algèbre engendrée** par $\mathcal{C}$. La famille $\mathcal{C}$ est alors appelée **famille génératrice** de l'algèbre.
17. Soit $\mathcal{A}$ une algèbre d'événements sur un espace fini Ω. Montrer qu'il existe une unique partition $\mathcal{D}$ de Ω génératrice de $\mathcal{A}$, i.e. $\mathcal{A} = \alpha(\mathcal{D})$.
18. Soit $(\mathcal{F}_i)_{i\in I}$ une famille de tribus sur Ω. Montrer que $\cap_{i\in I}\mathcal{F}_i$ est une tribu [2] sur Ω. En déduire que si $\mathcal{G} \subseteq \mathcal{P}(\Omega)$ est une famille arbitraire (non-vide) de parties de Ω, il existe une tribu minimale qui contient $\mathcal{G}$; on la note $\sigma(\mathcal{G})$ et on l'appelle **tribu engendrée** par $\mathcal{G}$. La famille $\mathcal{G}$ est appelée **famille génératrice** de la tribu.

Propriétés des probabilités discrètes

19. Le but de cet exercice est de donner une interprétation géométrique de l'ensemble des probabilités $\mathcal{M}_1(\Omega, \mathcal{F})$ sur un univers fini Ω et $\mathcal{F}$ une tribu sur Ω.
    (a) Rappeler la définition d'un ensemble convexe.
    (b) Montrer que $\mathcal{M}_1(\Omega, \mathcal{F})$ est convexe.
    (c) Rappeler la définition d'un point extrémal d'un ensemble convexe et montrer que $\mathbb{P} \in \mathcal{M}_1(\Omega, \mathcal{F})$ est extrémal si, et seulement si, pour tout $F \in \mathcal{F}$, on a $\mathbb{P}(F) \in \{0, 1\}$. On note par $\partial_e\mathcal{M}_1(\Omega, \mathcal{F})$ l'ensemble des probabilités extrémales.
    (d) Conclure que $E = \{\varepsilon_\omega : \omega \in \Omega\}$ est un ensemble composé de probabilités qui sont toutes extrémales.
    (e) Donner un exemple où $E \neq \partial_e\mathcal{M}_1(\Omega, \mathcal{F})$.
    (f) Montrer que si $\mathcal{F}$ est dénombrablement engendrée et contient les singletons, alors $E = \partial_e\mathcal{M}_1(\Omega, \mathcal{F})$.
    (g) Donner la structure géométrique de $\mathcal{M}_1(\Omega, \mathcal{F})$ dans le cas où $|\Omega| < \infty$ et $\mathcal{F} = \mathcal{P}(\Omega)$. On peut utiliser sans démonstration les deux théorèmes suivants :
    Théorème A : *(Carathéodory) Tout point dans l'enveloppe convexe* $\mathsf{co}(A)$ *d'un ensemble non-vide* $A \subset \mathbb{R}^d$ *peut être exprimé comme combinaison convexe d'au plus* $d + 1$ *points de* A.
    Théorème B : *(Krein-Milman) Si* X *est un espace vectoriel de dimension finie, toute partie convexe compacte non-vide de* X *est la collection des combinaisons convexes de ses points extrémaux.*
    (h) Lorsque $\mathsf{card}\Omega = 3$, dessiner la partie de $\mathbb{R}^2$ qui est isomorphe à $\mathcal{M}_1(\Omega, \mathcal{P}(\Omega))$ et y placer le point qui correspond au vecteur de probabilité $\mathbf{p} = (\frac{4}{19}, \frac{12}{19}, \frac{3}{19})$.

2. *Attention :* La réunion $\cup_{i\in I}\mathcal{F}_i$ n'est pas, en général, une tribu. Même dans le cas où $(\mathcal{F}_i)_{i\in\mathbb{N}}$ est une filtration dénombrable (i.e. une suite des tribus strictement emboîtées $\mathcal{F}_i \subset \mathcal{F}_{i+1}$, pour $i \in \mathbb{N}$), la réunion $\cup_{i\in\mathbb{N}}\mathcal{F}_i$ n'est **jamais** une tribu ; cf. [12]. La réunion engendre cependant une tribu, notée $\vee_{i\in\mathbb{N}}\mathcal{F}_i$, par le procédé décrit dans cet exercice, i.e. $\vee_{i\in\mathbb{N}}\mathcal{F}_i = \sigma(\cup_{i\in I}\mathcal{F}_i)$.

20. Supposons que Ω est dénombrable (fini ou infini) et qu'il existe une application $\rho : \Omega \to [0,1]$ telle que $\sum_{\omega\in\Omega} \rho(\omega) = 1$. Montrer que ρ définit une mesure de probabilité sur $(\Omega, \mathcal{F})$ définie par

$$\mathcal{F} \ni A \mapsto \mathbb{P}(A) = \sum_{\omega\in A} \rho(\omega) \in [0,1].$$

Suggestion : la seule chose à démontrer est la propriété d'additivité dénombrable disjointe.

21. Soit $\Omega = \{0,1\}^n$ l'espace des épreuves de n lancers d'une pièce, muni de sa tribu exhaustive $\mathcal{F} = \mathcal{P}(\Omega)$. L'espace des événements est probabilisé à l'aide du vecteur de probabilité uniforme ρ, chargeant chaque $\omega \in \Omega$ par $\rho(\omega) = \frac{1}{2^n}$. Pour chaque $k = 1, \ldots, n$, on définit l'application

$$\Omega \ni \omega = (\omega_1, \ldots, \omega_k, \ldots, \omega_n) \mapsto T_k\omega := (\omega_1, \ldots, 1-\omega_k, \ldots, \omega_n) \in \Omega.$$

L'application T_k est appelée « k^{th} bit-flip » (renversement du k^{e} bit). Montrer que pour tout $A \in \mathcal{F}$, on a l'invariance $\mathbb{P}(T_k A) = \mathbb{P}(A)$.

Variables aléatoires, fonctions de répartition

22. On lance une pièce lestée (donnant face (1) avec probabilité $p \in [0,1]$) n fois.
    (a) Décrire très précisément l'espace $(\Omega, \mathcal{F}, \mathbb{P})$ qui modélise cette expérience en déterminant la densité discrète ρ de $\mathbb{P}$.
    (b) On s'intéresse au nombre de fois que la pièce tombe sur face (1). Décrire très précisément la variable aléatoire X qui décrit cette grandeur et déterminer sa loi $\mathbb{P}_X$ et la densité discrète ρ_X correspondante.
    (c) Vérifier que la fonction de répartition vérifie $F_X(x) = 1$ pour $x \geq n$.
23. Montrer les propriétés de la fonction de répartition F_X d'une variable aléatoire réelle X (à savoir qu'elle est croissante, continue à droite et vérifie $\lim_{x\to-\infty} F_X(x) = 0$ et $\lim_{x\to+\infty} F_X(x) = 1$).
24. Montrer que si F est une fonction de répartition, elle a *au plus* une infinité dénombrable de discontinuités. *Suggestion : étant donné que F est continue à droite, un point x est une discontinuité s'il est une discontinuité gauche, i.e. si $DF(x) = F(x) - F(x-) > 0$. Écrire alors $\{x \in \mathbb{R} : DF(x) > 0\}$ comme une réunion dénombrable de parties de $\mathbb{R}$ et ... réfléchir un peu sur le cardinal de chacun de ces ensembles.*
25. Soient $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \lambda)$ et $X : \Omega \to \mathbb{R}$ une application continue. Suppposons que l'espace image $\mathbb{R}$ est muni de sa tribu borélienne. Montrer que X est une variable aléatoire.
26. Soit $F : \mathbb{R} \to [0,1]$ une fonction croissante, continue à droite, vérifiant $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$. Montrer qu'il existe un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et une variable aléatoire réelle X sur cet espace ayant F comme fonction de répartition.
27. Soit X une variable aléatoire réelle dont la loi admet une densité ρ_X (par rapport à la mesure de Lebesgue) sur $\mathbb{R}$. Montrer que sa fonction de répartition F_X s'exprime comme l'intégrale

$$F_X(x) = \int_{]-\infty,x]} \rho_X(t)\lambda(dt).$$

Noter que si F_X a une densité comme ci-dessus, alors F_X est continue et différentiable avec $F'_X = \rho_X$. (La démonstration de ces affirmations n'est pas demandée. Tandis que la démonstration de la continuité est facile, celle de la différentiabilité est longue et non-triviale.

3

Probabilité conditionnelle et indépendance

3.1 Conditionnement

Revenons pour l'instant à l'interprétation fréquentielle de la probabilité, présentée dans le chapitre 1, et considérons l'expérience suivante. On jette un dé N fois et on compte deux grandeurs : le nombre de fois N_B que l'événement $B =$ « la face supérieure montre 6 » se réalise et le nombre N_A de réalisations de l'événement $A =$ « la face supérieure porte un numéro pair ». Lorsque N est très grand, nous estimons la probabilité d'apparition de la face 6 par $\mathbb{P}(B) \simeq N_B/N \simeq 1/6$ et la probabilité d'apparition d'une face paire par $\mathbb{P}(A) \simeq N_A/N \simeq 1/2$. Mais supposons que quelqu'un nous informe que lors de cette expérience une face paire est apparue. En quoi cette information va modifier notre estimation de la probabilité d'obtenir 6 ? Un instant de réflexion, nous indique que la probabilité d'obtenir 6 **sachant que** le dé montre une face paire est $\mathbb{P}_A(B) \simeq N_B/N_A \simeq 1/3$; la connaissance que l'événement A s'est réalisé a transformé la probabilité de $\mathbb{P}$ en $\mathbb{P}_A$ qui doit logiquement vérifier les propriétés suivantes :

1. $\mathbb{P}_A(A) = 1$, i.e. l'événement A, sachant que A s'est réalisé, est maintenant certain,
2. il existe une constante $c_A > 0$ telle que pour tout événement $D \subseteq A$ on ait $\mathbb{P}_A(D) = c_A\mathbb{P}(D)$, i.e. les événements sont répondérés.

Proposition 3.1.1. *Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $A \in \mathcal{F}$ avec $\mathbb{P}(A) > 0$. Il existe une unique probabilité $\mathbb{P}_A$ sur $(\Omega, \mathcal{F})$ vérifiant les deux propriétés ci-dessus, définie par la formule*

$$\mathbb{P}_A(B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \forall B \in \mathcal{F}.$$

Démonstration. Supposons que $\mathbb{P}_A$ vérifie les deux propriétés ci-dessus. Pour $B \in \mathcal{F}$ arbitraire,

$$\mathbb{P}_A(B) = \mathbb{P}_A(B \cap A) + \mathbb{P}_A(B \setminus A) = \mathbb{P}_A(B \cap A) + 0 = c_A\mathbb{P}(B \cap A).$$

Pour $B = A$, nous avons $\mathbb{P}_A(A) = c_A\mathbb{P}(A)$, par conséquent $c_A = 1/\mathbb{P}(A)$ et nous obtenons ainsi la formule pour $\mathbb{P}_A$. Inversement, si $\mathbb{P}_A$ est donnée par cette formule, elle vérifie les deux propriétés 1 et 2 ci-dessus. □

Ces considérations nous amènent à la

Définition 3.1.2. Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $A, B \in \mathcal{F}$ avec $\mathbb{P}(A) > 0$. On appelle **probabilité conditionnelle** de l'événement B — sachant que l'événement A s'est réalisé — la quantité

$$\mathbb{P}_A(B) := \mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

Remarque 3.1.3. Lorsque $\mathbb{P}(A) = 0$, alors $\mathbb{P}(A \cap B) = 0$ aussi. Par conséquent la probabilité conditionnelle n'est pas bien définie dans ce cas ; plus précisément, on peut définir la probabilité conditionnelle comme prenant une valeur pré-déterminée arbitraire dans $[0, 1]$.

Théorème 3.1.4. *Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $(B_n)_{n\in I}$ une partition au plus dénombrable de Ω) avec $B_n \in \mathcal{F}$ pour tout $n \in I$. Alors, pour tout $A \in \mathcal{F}$ on a :*

$$\mathbb{P}(A) = \sum_{n\in I} \mathbb{P}(A|B_n)\mathbb{P}(B_n) \text{ (formule de probabilité totale)},$$

et, pour tout $k \in I$,

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_{n\in I} \mathbb{P}(A|B_n)\mathbb{P}(B_n)} \text{ (formule de Bayes)}.$$

Démonstration. Nous avons

$$\sum_{n\in I} \mathbb{P}(A|B_n)\mathbb{P}(B_n) = \sum_{n\in I} \mathbb{P}(A \cap B_n) = \mathbb{P}(A \cap (\cup_{n\in I} B_n)) = \mathbb{P}(A).$$

La formule de Bayes s'obtient immédiatement de la définition de la probabilité conditionnelle et de la formule précédente. □

Exercice 3.1.5. (Important ; résolvez-le avant de continuer). Une urne contient deux pièces de monnaie. Une honnête et une biaisée qui donne face avec probabilité $1/3$. On en extrait une, on la lance et on obtient face. Quelle est la probabilité qu'il s'agissait de la pièce honnête ?

Proposition 3.1.6. *Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $A_1, \ldots, A_n \in \mathcal{F}$. Alors*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1) \cdots \mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

Démonstration. Par simple substitution et simplification de termes dans le produit télescopique. □

Cette proposition nous fournit un moyen efficace de construction de modèles réalistes. Supposons que $\mathbf{X}$ est une variable aléatoire définie sur un espace de probabilité abstrait $(\Omega, \mathcal{F}, \mathbb{P})$, à valeurs dans $\mathbb{X} = \mathbb{A}^{N+1}$ où $\mathbb{A}$ est un ensemble dénombrable. Alors $\mathbf{X} = (X_0, \dots, X_N) \in \mathbb{A}^N$ peut-être considérée comme un mot aléatoire de N lettres construit sur l'alphabet $\mathbb{A}$. Tout joueur de scrabble connaît que les différentes lettres dans un mot français ne sont pas choisies purement au hasard ; la fréquence d'apparition d'une lettre en deuxième position dépend de la lettre déjà apparue en première position.

Si $\mathbf{x} = (x_0, \dots, x_N) \in \mathbb{X}$ est un mot de longueur $N+1$ construit sur l'alphabet $\mathbb{A}$ et $k : 1 \leq k \leq N$, on note $\mathbf{x}\restriction_k = (x_0, \dots, x_k)$ la restriction du mot à ses $k+1$ premières lettres.

Théorème 3.1.7. *Soient $N \in \mathbb{N}$ et $\mathbb{X}_0, \dots, \mathbb{X}_N$ une famille d'ensembles dénombrables (finis ou infinis). On note $\mathbb{X} = \times_{n=0}^N \mathbb{X}_n$ et $\mathbf{x} = (x_0, \dots, x_N)$ les éléments de $\mathbb{X}$. L'espace $\mathbb{X}$ est supposé muni de la tribu $\mathcal{X} = \otimes_{n=0}^N \mathcal{P}(\mathbb{X}_n)$. Supposons que ρ_0 est un vecteur de probabilité sur $\mathbb{A}_0$ et, pour tout $k = 1, \dots N$ et $\mathbf{x} \in \mathbb{X}$, $\rho_{k,\mathbf{x}\restriction_{k-1}}$ est une vecteur de probabilité sur $\mathbb{A}_k$. Alors, il existe un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, avec $(\Omega, \mathcal{F})$ suffisamment riche pour que la suite $\mathbf{X} = (X_0, \dots, X_n)$ de variables aléatoires soit définie de façon que, pour tout $\mathbf{x} = (x_0, \dots, x_N) \in \mathbb{X}$, $\mathbb{P}$ est l'unique probabilité sur $(\Omega, \mathcal{F})$ vérifiant*

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \rho_0(x_0)\rho_{1,x_0}(x_1) \cdots \rho_{N,\mathbf{x}\restriction_{N-1}}(x_N).$$

Dans le cas considéré dans le théorème précédant (où $N < \infty$) la tribu $\mathcal{X} = \otimes_{n=0}^N \mathcal{P}(\mathbb{X}_n)$ coïncide avec la tribu exhaustive $\mathcal{X} = \mathcal{P}(\mathbb{X})$, engendrée par les singletons $\mathcal{X} = \sigma(\{\mathbf{x}\}, \mathbf{x} \in \mathbb{X})$. La démonstration donc du théorème est élémentaire. Pour établir par exemple l'unicité de la construction, il suffit de vérifier que toutes les probabilités définies par le membre de droite de la formule de l'énoncé coïncident sur les singletons, sont donc identiques sur $\mathcal{X}$.

Nous serons cependant amenés à considérer des suites infinies de variables aléatoires, par exemple pour étudier les suites de bits générées par une source. Nous avons donc besoin d'un résultat analogue au théorème 3.1.7 mais pour $N \to \infty$. La difficulté essentielle provient du fait que, même pour une suite $(\mathbb{X}_n)_{n\in\mathbb{N}}$ d'alphabets finis, l'espace $\mathbb{X} = \times_{n\in\mathbb{N}} \mathbb{X}_n$ n'est plus dénombrable. Le théorème 3.1.8, démontré ci-dessous, fournit la généralisation du théorème 3.1.7 dans le cas $N = \infty$.

Théorème 3.1.8. *Soit $(\mathbb{X}_n)_{n\in\mathbb{N}}$ une suite d'alphabets dénombrables. On note $\mathbb{X} = \times_{n\in\mathbb{N}} \mathbb{X}_n$ le produit cartésien des ces alphabets. L'espace $\mathbb{X}$ est supposé muni de la tribu $\mathcal{X} = \otimes_{n\in\mathbb{N}} \mathcal{P}(\mathbb{X}_n)$. Supposons que ρ_0 est un vecteur de probabilité sur $\mathbb{X}_0$ et pour tout $k \geq 1$, et tout $\mathbf{x} = (x_0, x_1, \dots) \in \mathbb{X}$, $\rho_{k,\mathbf{x}\restriction_{k-1}}$ un vecteur de probabilité sur $\mathbb{X}_k$. Alors, il existe un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, avec $(\Omega, \mathcal{F})$ suffisamment riche pour que toute la suite $\mathbf{X} = (X_0, X_1, \dots) = (X_n)_{n\in\mathbb{N}}$ de variables aléatoires soit définie de façon que, pour tout $N \geq 1$ et tout $\mathbf{x} \in \mathbb{X}$, la probabilité $\mathbb{P}$ soit l'unique probabilité sur $(\Omega, \mathcal{F})$ vérifiant*

$$\mathbb{P}(\mathbf{X}\restriction_N = \mathbf{x}\restriction_N) = \rho_0(x_0)\rho_{1,x_0}(x_1) \cdots \rho_{N,\mathbf{x}\restriction_{N-1}}(x_N).$$

Remarque 3.1.9. La tribu $\mathcal{X} = \otimes_{n\in\mathbb{N}} \mathcal{P}(\mathbb{X}_n)$ n'est pas la tribu exhaustive. Elle est engendrée par la famille des « **rectangles** » $\mathcal{R} = \cup_{N\in\mathbb{N}} \mathcal{R}_N$, où

$$\mathcal{R}_N = \{\times_{k\in\mathbb{N}} B_k : B_k \in \mathcal{P}(\mathbb{X}_k) \text{ et } \sharp\{k \in \mathbb{N} : B_k \neq \mathbb{X}_k\} < \infty\}.$$

Elle est aussi engendrée par la famille des « **cylindres** » $\mathcal{C} = \cup_{N\in\mathbb{N}}\mathcal{C}_N$, où

$$\mathcal{C}_N = \{\{x_0\cdots x_N\}\times_{k>N}\mathbb{X}_k,\ x_0\cdots x_N\in(\times_{\ell=0}^N\mathbb{X}_\ell)\}.$$

On remarque que pour un $N\in\mathbb{N}$ arbitraire, un cylindre arbitraire de la famille $\mathcal{C}_N$ s'écrit comme

$$[x_0\cdots x_N] := \{x_0\cdots x_N\}\times\mathbb{X}_{N+1}\times\mathbb{X}_{N+2}\times\cdots.$$

Pour un $\mathbf{x}\in\mathbb{X}$ on note donc le cylindre $[x_0\cdots x_N] = [\mathbf{x}\restriction_N]$.

Démonstration du théorème 3.1.8. Existence. Choisir $\Omega = I_\emptyset := [0,1[$ muni de la tribu borélienne $\mathcal{B}([0,1[)$ et probabilisé par la mesure de Lebesgue λ. On introduit une partition de $I_\emptyset$ en $\mathrm{card}\mathbb{X}_0$ intervalles semi-ouverts contigus $I_\emptyset = \sqcup_{x_0\in\mathbb{X}_0}I_{x_0}$ en imposant $\lambda(I_{x_0}) = \rho_0(x_0)$. Supposons que nous ayons construit une partition à l'ordre $k-1$, i.e. pour $x_0\cdots x_{k-1}\in\mathbb{X}_0\times\cdots\times\mathbb{X}_{k-1}$, nous disposons de la famille exhaustive d'intervalles disjoints contigus $I_{x_0\cdots x_{k-1}}$ dont les longueurs vérifient

$$\lambda(I_{x_0\cdots x_{k-1}}) = \rho_0(x_0)\cdots\rho_{k-2,\mathbf{x}\restriction_{k-2}}(x_{k-1}).$$

Nous obtenons une partition à l'ordre k, en écrivant chacun de ceux intervalles comme une partition plus fine

$$I_{x_0\cdots x_{k-1}} = \sqcup_{x_k\in\mathbb{X}_k}I_{x_0\cdots x_k},$$

avec $\lambda(I_{x_0\cdots x_k}) = \lambda(I_{x_0\cdots x_{k-1}})\rho_{k,\mathbf{x}\restriction_{k-1}}(x_k)$. Cette construction peut être répétée *ad infinitum*.

Pour chaque k, la famille $(I_{\mathbf{x}})_{\mathbf{x}\in\mathbb{X}_0\times\cdots\times\mathbb{X}_k}$ forme une partition de Ω. Il s'ensuit que pour chaque $\omega\in\Omega$, il existe un unique intervalle de la génération k qui le contient, i.e. il existe une suite $\mathbf{X}(\omega) = (X_0(\omega), X_1(\omega),\ldots)\in\mathbb{X}\simeq\Omega$ telle que $\omega\in I_{X_0(\omega)\cdots X_k(\omega)}$, pour tout $k\in\mathbb{N}$.

L'application $\mathbf{X}:\Omega\to\mathbb{X}$ est une variable aléatoire. En effet, considérer la famille $\mathcal{G} = \{\emptyset\}\cup\bigcup_{k\geq0}\mathcal{G}_k$, où

$$\mathcal{G}_k = \{X_0 = x_0,\ldots,X_k = x_k, x_i\in\mathbb{X}_i, i = 0,\ldots,k\} = \{\mathbf{X}\restriction_k = \mathbf{x}\restriction_k\}.$$

Cette famille constitue une algèbre qui engendre une tribu $\mathcal{F}$, sous-tribu de $\mathcal{B}([0,1[)$. Par ailleurs, pour tout cylindre $C = [x_0\cdots x_N]\in\mathcal{C}$, on aura $\mathbf{X}^{-1}(C) = \{\mathbf{X}\restriction_N = \mathbf{x}\restriction_N\} = I_{\mathbf{x}\restriction_N}\in\mathcal{F}\subseteq\mathcal{B}([0,1[)$. En outre, $\lambda\circ\mathbf{X}^{-1}$ définit bien une probabilité $\mathbb{P}$ sur $(\Omega,\mathcal{F})$ ayant les propriétés requises.

Unicité. Supposons qu'il existe deux mesures de probabilité $\mathbb{P}$ et $\mathbb{P}'$ sur $(\Omega,\mathcal{F})$ telles que

$$\mathbb{P}(\mathbf{X}\restriction_N = \mathbf{x}\restriction_N) = \rho_0(x_0)\rho_{1,x_0}(x_1)\cdots\rho_{N,\mathbf{x}\restriction_{N-1}}(x_N) = \mathbb{P}'(\mathbf{X}\restriction_N = \mathbf{x}\restriction_N).$$

On remarque que ces probabilités coïncident sur $\mathcal{G}$. Comme $\mathcal{G}$ contient l'ensemble vide, est stable par intersection finie et engendre $\mathcal{F}$, elles coïncideront sur $\mathcal{F}$. □

Définition 3.1.10. La probabilité $\mathbb{P}$, dont l'existence et l'unicité sont établies dans le théorème 3.1.8 est appelée **loi ou probabilité conjointe** de la suite aléatoire $(X_0, X_1,\ldots,)$. La probabilité sur $\mathbb{X}_0\times\cdots\times\mathbb{X}_N$ définie par sommation partielle sur toutes les valeurs $x_{N+1}, x_{N+2},\ldots$ non observées est appelée **probabilité marginale**.

3.2 Indépendance

La signification intuitive de l'indépendance entre deux événements A et B est que la probabilité de B n'est pas influencée par la révélation que A s'est réalisé (ou vice versa avec les rôles de A et B interchangés). Ceci nous incite à introduire la notion d'indépendance par la

Définition 3.2.1. Soit $(\Omega,\mathcal{F},\mathbb{P})$ un espace de probabilité. Deux événement $A,B\in\mathcal{F}$ sont **indépendants** par rapport à la probabilité $\mathbb{P}$ si

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

Remarque 3.2.2. — Un fait qui est souvent négligé est que l'indépendance entre deux événements n'est pas une propriété intrinsèque des événements en considération mais fait intervenir de manière cruciale la probabilité sous-jacente. (Cf. exercice 38). Deux événements A et B sur $(\Omega, \mathcal{F})$ peuvent être indépendants par rapport à une probabilité $\mathbb{P}$ sur $\mathcal{F}$ et dépendants par rapport à une probabilité $\mathbb{P}'$.

— L'indépendance définie en 3.2.1, appelée aussi indépendance stochastique, ne doit pas être confondue avec un autre type d'indépendance, l'indépendance causale, entre deux événements, signifiant qu'il n'y a pas de relation de cause à effet entre eux. (Cf. exercice 39).

Définition 3.2.3. Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $I \neq \emptyset$ une famille arbitraire d'indices.

1. Une famille $(A_i)_{i\in I}$ d'événements de $\mathcal{F}$ est dite **indépendante** par rapport à $\mathbb{P}$ si pour toute partie *finie* J, $\emptyset \neq J \subseteq I$, nous avons
$$\mathbb{P}\left(\cap_{j\in J} A_j\right) = \prod_{j\in J} \mathbb{P}(A_j).$$
2. Soient $(\mathbb{X}_i, \mathcal{X})_{i\in I}$ une famille d'espace d'événements et $(X_i)_{i\in I}$ une famille de variables aléatoires avec $X_i : \Omega \to \mathbb{X}_i$. La famille est **indépendante** si pour tout choix d'événements $B_i \in \mathcal{X}_i$, la famille d'événements $(\{X_i \in B_i\})_{i\in I}$ est indépendante, i.e. si pour toute partie finie J avec $\emptyset \neq J \subseteq I$, nous avons
$$\mathbb{P}\left(\cap_{j\in J}\{X_j \in B_j\}\right) = \prod_{j\in J} \mathbb{P}(\{X_j \in B_j\}).$$

(Le cas trivial $\mathrm{card} J = 1$ est inclus dans cette définition pour de raisons de simplicité).

Une famille d'événements (ou de variables aléatoires) peut être telle que si l'on considère deux éléments arbitraires de la famille, ils sont indépendants sans que la famille soit indépendante comme le montre le

Contre-exemple 3.2.4. On lance une pièce deux fois de suite. L'univers est $\Omega = \{0,1\}^2$; on probabilise sa tribu exhaustive avec la probabilité uniforme sur Ω. Considérer les événements

$$\begin{aligned} A &= \{\text{lors du premier lancer la pièce tombe sur face}\},\\ B &= \{\text{lors du second lancer la pièce tombe sur face}\},\\ C &= \{\text{lors des deux lancers la pièce tombe du même côté}\}. \end{aligned}$$

Alors
$$\mathbb{P}(A\cap B\cap C) = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C),$$
tandis que les événements pris deux-à-deux sont indépendants.

La définition 3.2.3 appliquée à la construction faite dans le théorème 3.1.8 implique que les vecteurs de probabilité ne dépendent pas de x. Une famille de variables aléatoires $(X_n)_{n=1,\dots,N}$, avec $X_n : \Omega \to \mathbb{B}_n$, est indépendante si sur chaque $(\mathbb{X}_n, \mathcal{P}(\mathbb{X}_n))$, il existe un vecteur de probabilité ρ_n (indépendant des x_i pour $i < n$) et
$$\mathbb{P}(X_1 = x_1, \dots, X_N = x_n) = \prod_{n=1}^{N} \mathbb{P}(X_n = x_n) = \prod_{n=1}^{N} \rho_n(x_n).$$

(Étant donné que les ρ_n ne dépendent pas dex $x_i, i < n$, nous pouvons calculer les marginales arbitraires). Ceci reste valable même pour des suites infinies $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires $X_n : \Omega \to \mathbb{X}_n$ indépendantes, sur des alphabets dénombrables $(\mathbb{X}_n)$. Le théorème 3.1.8 combiné avec la propriété d'indépendance implique qu'il existe une suite infinie de vecteurs de probabilité (ρ_n) et une unique probabilité $\mathbb{P}$ telles que pour toute partie finie non-vide $J \subseteq \mathbb{N}$, on ait

$$\mathbb{P}\left(\cap_{j\in J}\{X_j = x_j\}\right) = \prod_{j\in J} \rho_j(x_j).$$

3.3 Exercices

Probabilité conditionelle

Dans tout ce paragraphe, on travaille sur un espace probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. Tous les événements utilisés, $A, B, (A_k), \ldots$, appartiennent à $\mathcal{A}$.

28. Montrer que **les formules**

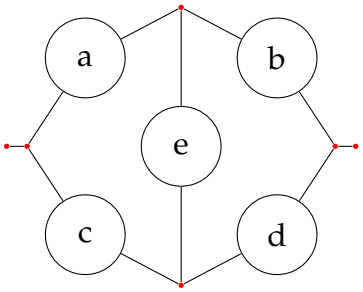
$$\mathbb{P}(B|A) + \mathbb{P}(B|A^c) = 1 \text{ et } \mathbb{P}(B|A) + \mathbb{P}(B^c|A^c) = 1$$

sont fausses.

29. Pour une famille d'événements $(A_k)_{k=1,\ldots,n}$ indépendante, noter $p_k = \mathbb{P}(A_k)$. Calculer la probabilité de l'événement « aucun des (A_k) ne se réalise ».
30. Soient A et B deux événements indépendants, avec $p = \mathbb{P}(A)$ et $q = \mathbb{P}(B)$. Calculer les probabilités pour que

    (a) exactement k,
    (b) au moins k,
    (c) au plus k,

    des événements A et B se réalisent pour $k = 0, 1, 2$.
31. Exprimer $\mathbb{P}(\cup_{k=1}^n A_k)$ lorsque les A_k sont indépendants.
32. Un lot de 100 objets manufacturés subit un contrôle par échantillonnage : un échantillon de 5 objets est examiné et s'il contient un article défectueux, tout le lot est rejeté. Quelle est la probabilité qu'un lot soit rejeté s'il contient 5% d'objets défectueux ?
33. Une personne a oublié le dernier digit du numéro de téléphone de son correspondant ; elle compose donc au hasard. Quelle est la probabilité p qu'elle soit obligée de composer moins de 3 fois (≤ 3) ? Que devient cette probabilité, si la personne se souvient que le dernier digit est impair ?
34. Une personne écrit n lettres différentes, chacune destinée a un destinataire spécifique. Elle les a scellées et posées bien rangées en pile sur son bureau afin d'écrire le lendemain les adresses sur les enveloppes. Un plaisantin est passé par là, il a renversée la pile par erreur et il a remis les enveloppes sur le bureau mais dans un ordre aléatoire par rapport à l'ordre correct. Ignorant ce fait, la personne qui a rédigé les lettres a inscrit le lendemain les adresses. Quelle est la probabilité qu'une lettre parvienne à la personne à laquelle elle était destinée ?
35. Le circuit électrique suivant comporte 5 interrupteurs, notés a à e ; chacun d'eux peut être fermé (laisse le courant passer) avec probabilité p ou ouvert (coupe le courant) avec probabilité $q = 1 - p$.



(a) Quelle est la probabilité pour que le courant passe de gauche à droite ?
(b) Sachant que le courant passe de gauche à droite, quelle est la probabilité pour que l'interrupteur e soit fermé ?

Variables aléatoires ; probabilités conjointes

36. On lance une pièce honnête et on désigne par $X_1 \in \{0,1\}$ le résultat obtenu. Si $X_1 = 1$, on lance une pièce honnête sinon on lance une pièce lestée qui donne 1 avec probabilité $2/3$; on note X_2 le résultat du deuxième lancer. Si on a obtenu 2 fois 1 lors des 2 premiers lancers, on lance un dé honnête, sinon on lance un dé lesté qui donne 6 avec probabilité $1/2$ et les autres faces équiprobables.
    (a) Calculer explicitement les vecteurs de probabilité ρ_1, ρ_{2,x_1} et $\rho_{3,(x_1,x_2)}$.
    (b) Déterminer la probabilité conjointe $\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ pour $x_1, x_2 \in \{0,1\}$ et $x_3 \in \{1,\dots,6\}$.
    (c) Calculer les probabilités marginales $\mathbb{P}(X_1 = x_1)$, $\mathbb{P}(X_1 = x_1, X_2 = x_2)$, $\mathbb{P}(X_2 = x_2)$, $\mathbb{P}(X_3 = x_3)$,
37. (Formule de Bayes).

    Une usine a produit 2 lots, notés a et b, d'objets. Tous les objets du lot a sont fonctionnels tandis que seulement 75% d'objets du lot b sont fonctionnels, les autres étant défectueux. On note $\mathbf{p} = (1/2, 1/2)$ le vecteur de probabilité uniforme sur l'ensemble de lots $\mathbb{L} = \{a, b\}$ et $\mathbb{X} = \{d, f\}$.
    (a) On modélise par la variable aléatoire L à valeurs dans $\mathbb{L}$, de loi *a priori* $\mathbf{p}$, le choix du lot et par une variable aléatoire X à valeurs dans $\mathbb{X}$ l'état de l'objet choisi. On choisit un lot au hasard (selon $\mathbf{p}$) et un objet au hasard (selon la loi uniforme) dans le lot ; l'objet se trouve être fonctionnel. Déterminer le vecteur de probabilité *a posteriori* $\mathbf{q}$ sur $\mathbb{L}$, obtenue après que cette expérience ait eu lieu (i.e. sachant que l'objet choisi est fonctionnel).
    (b) On remet l'objet dans le lot duquel il a été extrait et on choisit au hasard un deuxième objet **de ce même lot**. Quelle est la probabilité que cet objet soit défectueux ? (On note Y la variable aléatoire à valeurs dans $\mathbb{X}$ qui modélise l'état du second objet extrait).

Indépendance

38. Une urne contient n boules discernables noires et r boules discernables rouges et on considère l'espace $\Omega = \{1,\dots,n+r\}^2$ (qui peut décrire toutes les expériences d'extraction de deux boules avec ou sans remise). On note

$$A = \{\text{la première boule est rouge}\} = \{\omega \in \Omega : \omega_1 \in \{1,\dots,r\}\}$$

et

$$B = \{\text{la seconde boule est rouge}\} = \{\omega \in \Omega : \omega_2 \in \{1, \dots, r\}\}.$$

(a) On en extrait une boule, on la remet dans l'urne et on en extrait une seconde ; on note $\mathbb{P}$ la probabilité uniforme sur Ω. Les événements A et B sont-ils indépendants (par rapport à $\mathbb{P}$) ?

(b) On extrait une première boule et sans la remettre dans l'urne, on en extrait une seconde. On note $\mathbb{P}'$ la probabilité uniforme sur

$$\Omega' = \{\omega = (\omega_1, \omega_2), \omega_i \in \{1, \dots, n+r\}, \omega_1 \neq \omega_2\} \subset \Omega.$$

La probabilité $\mathbb{P}'$ peut être étendue en une probabilité, aussi notée $\mathbb{P}'$, sur Ω chargeant avec masse 0 les éléments de $\Omega \setminus \Omega'$. Les événements A et B sont-ils indépendants (par rapport à $\mathbb{P}'$) ?

39. On jette un dé deux fois de suite et on note $\Omega = \{1, \dots, 6\}^2$ l'espace des épreuves correspondant. On munit la tribu exhaustive de cet espace de la probabilité uniforme et on considère les événements

$$A = \{\text{la somme est } 7\} \text{ et } B = \{\text{le premier dé tombe sur } 6\}.$$

(a) Montrer que A et B sont indépendants (sous la probabilité uniforme).

(b) Soient $\mathbb{G} = (\mathbb{G}^0, \mathbb{G}^1)$ un graphe dirigé acyclique, $\mathbb{U} = \{U_v, v \in \mathbb{G}^0\}$ une famille de variables aléatoires indexée par les sommets du graphe et $\mathbb{F} = \{f_v, v \in \mathbb{G}^0\}$ une famille d'applications réelles à plusieurs variables. Une famille de variables aléatoires $(X_v)_{v \in \mathbb{G}^0}$ suit un **modèle causal** si pour tout sommet $v \in \mathbb{G}^0$, on peut exprimer $X_v = f_v(X_u, u \in \mathsf{pa}(v); U_v)$. Montrer que A est déterminé de B de manière causale.

4

Espérance, variance ; théorèmes des grands nombres

4.1 Espérance

Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $(\mathbb{X}, \mathcal{X})$ un espace d'événements avec $\mathbb{X} \subset \mathbb{R}$ et $\text{card}\mathbb{X} = n$. Alors, il existe des réels distincts $x_i, i = 1, \ldots, n$ tels que $\mathbb{X} = \{x_1, \ldots, x_n\}$. Une variable aléatoire $X : \Omega \to \mathbb{X}$ peut alors s'écrire à l'aide de la partition $A_i = \{X = x_i\}, i = 1, \ldots, n$ comme $X(\omega) = \sum_{i=1}^{n} x_i \mathbb{1}_{A_i}(\omega)$. La loi de X s'exprime en termes du vecteur de probabilité ρ vérifiant $p_i := \rho(x_i) = \mathbb{P}(X = x_i) = \mathbb{P}(A_i)$, pour $i = 1, \ldots, n$. Si nous observons N réalisations de la variable aléatoire X, nous nous attendons à ce que x_i soit obtenu approximativement Np_i fois. Si nous exprimons la moyenne pondérée des N réalisations de X, nous obtenons

$$\frac{1}{N}\left[Np_1x_1 + \ldots + Np_nx_n\right] = \sum_{i=1}^{n} p_i x_i.$$

L'expression du second membre ci-dessus est ce que nous appelons espérance de la variable aléatoire X (voir définition 4.1.1). Sa signification est le barycentre (pondéré par les probabilités $p_i = \mathbb{P}(X = x_i)$) de l'ensemble de valeurs distinctes (x_i) que peut prendre X.

4.1.1 Cas discret

Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $X : \Omega \to \mathbb{R}$ une variable aléatoire réelle. La variable aléatoire X est appelée **discrète** si $X(\Omega)$ est dénombrable (fini ou infini).

Définition 4.1.1. 1. Une variable aléatoire réelle discrète X sur $(\Omega, \mathcal{F}, \mathbb{P})$ est dite **intégrable** (plus précisément $\mathbb{P}$-intègrable) si $\sum_{x \in X(\Omega)} |x| \mathbb{P}(\{X = x\}) < \infty$. On note alors $X \in \mathcal{L}^1 := \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.

2. Si $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, on définit son **espérance** $\mathbb{E}X$ par

$$\mathbb{E}(X) = \mathbb{E}_{\mathbb{P}}(X) := \sum_{x \in X(\Omega)} x\mathbb{P}(X = x).$$

Remarque 4.1.2. — Si $A \in \mathcal{F}$, l'application $\mathbb{1}_A : \Omega \to \{0,1\}$ est une variable aléatoire. Son espérance est $\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$.

— Si une variable aléatoire X est discrète, nous pouvons toujours la décomposer comme

$$X(\omega) = \sum_{x \in X(\Omega)} x\mathbb{1}_{A_x}(\omega),$$

où $A_x = \{\omega \in \Omega : X(\omega) = x\}$. Nous aurons alors

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x \in X(\Omega)} x\mathbb{P}(A_x) \\
&= \sum_{x \in X(\Omega)} x\mathbb{P}_X(\{x\}) \\
&= \sum_{x \in X(\Omega)} x\rho_X(x) \\
&= \sum_{x \in X(\Omega)} x\Delta F_X(x), \quad \text{où } \Delta F_X(x) = F_X(x) - F_X(x-).
\end{aligned}$$

Proposition 4.1.3. *Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $X, Y, (X_n), (Y_n)$ de variables aléatoires réelles discrètes et intégrables sur $(\Omega, \mathcal{F}, \mathbb{P})$. Alors nous avons les propriétés suivantes :*

1. *Si $X \leq Y$ alors $\mathbb{E}(X) \leq \mathbb{E}(Y)$ (monotonie).*
2. *Pour tout $c \in \mathbb{R}$, $\mathbb{E}(cX) = c\mathbb{E}(X)$ et $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ (linéarité).*
3. *Si pour tout $n \in \mathbb{N}$ on $X_n \geq 0$ et $X = \sum_{n\in\mathbb{N}} X_n$, alors $\mathbb{E}(X) = \sum_{n\in\mathbb{N}} \mathbb{E}(X_n)$ (σ-additivité).*
4. *Si $Y_n \uparrow Y$, alors $\mathbb{E}(Y) = \lim_{n\to\infty} \mathbb{E}(Y_n)$ (convergence monotone).*
5. *Si X et Y sont indépendantes, alors $XY \in \mathcal{L}^1$ et $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ (mulitiplicativité en cas d'indépendance).*

Démonstration. 1. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé sur lequel les deux variables aléatoires $X : \Omega \to \mathbb{X} = X(\Omega)$ et $Y : \Omega \to \mathbb{Y} = Y(\Omega)$ sont simultanément définies. La condition $X \leq Y$ signifie que pour toute réalisation $\omega \in \Omega$, $X(\omega) \leq Y(\omega)$. On conclut que

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x\in\mathbb{X}} x\mathbb{P}(X^{-1}(\{x\})) = \sum_{x\in\mathbb{X}} x\mathbb{P}(X^{-1}(\{x\}), Y^{-1}(\mathbb{Y})) \\
&= \sum_{x\in\mathbb{X}, y\in\mathbb{Y}} x\mathbb{P}(X^{-1}(\{x\}), Y^{-1}(\{y\})) \leq \sum_{x\in\mathbb{X}, y\in\mathbb{Y}} y\mathbb{P}(X^{-1}(\{x\}), Y^{-1}(\{y\})) \\
&= \sum_{y\in\mathbb{Y}} y\mathbb{P}(X^{-1}(\mathbb{X}), Y^{-1}(\{y\})) = \sum_{y\in\mathbb{Y}} y\mathbb{P}(Y^{-1}(\{y\})) = \mathbb{E}(Y).
\end{aligned}$$

2. L'égalité $\mathbb{E}(cX) = c\mathbb{E}(X)$ découle immédiatement de la définition de l'espérance. Pour montrer la deuxième égalité, on remarque que si X et Y sont des variables aléatoires réelles discrètes définies sur le même espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, à valeurs respectivement sur $\mathbb{X} = X(\Omega)$ et $\mathbb{Y} = Y(\Omega)$, la variable « somme », $W = X + Y$ est une variable aléatoire discrète définie sur

$(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{W}$, image de $\mathbb{X} \times \mathbb{Y}$ par l'application $\mathbb{X} \times \mathbb{Y} \ni (x, y) \mapsto x + y = w \in \mathbb{W}$. La variable aléatoire est intégrable car

$$\begin{aligned}\mathbb{E}(|W|) &= \sum_{w\in\mathbb{W}} |w|\mathbb{P}(X+Y=w) = \sum_{x\in\mathbb{X},w\in\mathbb{W}} |w|\mathbb{P}(X=x, Y=w-x)\\ &\leq \sum_{x\in\mathbb{X},w\in\mathbb{W}} (|x|+|w-x|)\mathbb{P}(X=x, Y=w-x)\\ &= \sum_{x\in\mathbb{X},y\in\mathbb{Y}} (|x|+|y|)\mathbb{P}(X=x, Y=y) = \mathbb{E}(|X|)+\mathbb{E}(|Y|).\end{aligned}$$

L'égalité requise s'obtient en enlevant la valeur absolue dans la suite des relations ci-dessus.

3. Par la positivité des variables, pour tout $N \in \mathbb{N}$, nous avons $X \geq S_N := \sum_{k=0}^N X_k$ et par conséquent, par la linéartité de l'espérance $\mathbb{E}(X) \geq \sum_{k=0}^N \mathbb{E}(X_k)$. La dernière inégalité est vraie pour tout $N \in \mathbb{N}$; elle sera alors vraie dans la limite $N \to \infty$, i.e. $\mathbb{E}(X) \geq \sum_{k\in\mathbb{N}} \mathbb{E}(X_k)$.
Pour conclure, il faut établir l'inégalité inverse. Soit $c \in]0, 1[$ arbitraire $\tau = \inf\{N \geq 0 : S_N \geq cX\} \in \mathbb{N} \cup \{+\infty\}$. Puisque $S_N \uparrow X$, il s'ensuit que $\tau < \infty$. Or $\{\tau < \infty\} = \cup_{N\in\mathbb{N}}\{\tau = N\}$; par conséquent, $S_\tau = S_\tau \sum_{N\in\mathbb{N}} \mathbb{1}_{\{\tau=N\}} = \sum_{N\in\mathbb{N}} S_N \mathbb{1}_{\{\tau=N\}}$. On a alors

$$\begin{aligned}c\mathbb{E}(X) \leq \mathbb{E}(S_\tau) &= \sum_{x\in S(\Omega)} x \sum_{N\in\mathbb{N}} \mathbb{P}(\tau = N, S_N = x) = \sum_{N\in\mathbb{N}} \mathbb{E}(S_N \mathbb{1}_{\{\tau=N\}})\\ &= \sum_{N\in\mathbb{N}} \sum_{k=0}^N \mathbb{E}(X_k \mathbb{1}_{\{\tau=N\}}) = \sum_{N\in\mathbb{N}} \sum_{k=0}^N \sum_{x\in X(\Omega)} x\mathbb{P}(X_k = x, \tau = N)\\ &= \sum_{n\in\mathbb{N}} \sum_{x\in X_n(\Omega)} x\mathbb{P}(X_n = x; \tau \geq n) \leq \sum_{n\in\mathbb{N}} \sum_{x\in X_n(\Omega)} x\mathbb{P}(X_n = x) = \sum_{n\in\mathbb{N}} \mathbb{E}(X_n).\end{aligned}$$

Le paramètre c étant arbitraire, en prenant la limite $c \uparrow 1$, on obtient $\mathbb{E}(X) \leq \sum_{n\in\mathbb{N}} \mathbb{E}(X_n)$.

4. On applique le résultat précédent aux variables aléatoires positives $X = Y - Y_0$ et $X_{n+1} = Y_{n+1} - Y_n$ pour $n \in \mathbb{N}$.
5. La variable aléatoire XY est une variable aléatoire discrète définie sur $(\Omega, \mathcal{F}, \mathbb{P})$. Son intégrabilité découle de l'observation

$$\begin{aligned}\sum_w |w|\mathbb{P}(XY=w) &= \sum_{w\neq 0} |w| \sum_{x\neq 0} \mathbb{P}(X=x, Y=w/x) = \sum_{x\neq 0, y\neq 0} |x||y|\mathbb{P}(X=x, Y=y)\\ &= \sum_{x\neq 0, y\neq 0} |x||y|\mathbb{P}(X=x)\mathbb{P}(Y=y) \text{ (à cause de l'indépendance)}\\ &= \mathbb{E}(|X|)\mathbb{E}(|Y|).\end{aligned}$$

La multiplicativité s'obtient en enlevant la valeur absolue dans les égalités précédentes.

□

4.1.2 Cas continu (à densité)

Dans le cas continu, lorsque la mesure de probabilité admet une densité, nous exprimons de nouveau la mesure de probabilité à l'aide de sa densité.

Définition 4.1.4. Soient $\Omega \simeq \mathbb{R}^d$, muni de sa tribu borélienne $\mathcal{F} = \mathcal{B}_d$ et $\mathbb{P}$ une probabilité sur $(\Omega, \mathcal{F})$ ayant une densité ρ par rapport à la mesure de Lebesgue. Soit X une variable aléatoire réelle définie sur $(\Omega, \mathcal{F})$.

1. On dit que X est **intégrable** (plus précisément $\mathbb{P}$-intégrable) si

$$\int_\Omega |X(\omega)|\rho(\omega)d\omega_1 \cdots d\omega_d < \infty.$$

On note alors $X \in \mathcal{L}^1 := \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.

2. Si $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, on appelle **espérance** $\mathbb{E}(X)$ de X, la quantité

$$\mathbb{E}(X) = \mathbb{E}_{\mathbb{P}}(X) := \int_{\Omega} X(\omega)\rho(\omega)d\omega_1 \cdots d\omega_d = \int_{\mathbb{X}} x\rho_X(x)dx.$$

Il faudrait bien sûr donner un sens aux intégrales (de Lebesgue) qui apparaissent dans la définition précédente. Cependant, dans tous les cas qui se présenteront dans ce cours, nous pouvons considérer que ces intégrales sont égales aux intégrales habituelles (de Riemann).

Remarque 4.1.5. Les affirmations de la proposition 4.1.3, établies dans le cas discret, restent valables dans le cas continu.

4.2 Variance et covariance

Tout comme l'espérance exprime la moyenne pondérée (barycentre) des valeurs d'une variable aléatoire, la variance mesure la dispersion de la variable autour de son espérance. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et X une variable aléatoire réelle. Si pour un $r \in \mathbb{N}^*$, la variable $X^r \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, on dit que la variable est r-intégrable et l'on note $X \in \mathcal{L}^r(\Omega, \mathcal{F}, \mathbb{P})$. Puisque $|XY| \leq X^2 + Y^2$, il s'ensuit que si $X, Y \in \mathcal{L}^2$ alors $XY \in \mathcal{L}^1$. Cette remarque permet donc de définir :

Définition 4.2.1. Soit $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ variables aléatoires réelles.

1. On appelle **variance** de X, le nombre positif $\mathsf{Var}(X) = \mathbb{E}([X - \mathbb{E}(X)]^2)$; sa racine carré est appelée **écart-type** de X.
2. On appelle **covariance** de X et Y le nombre réel

$$\mathsf{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))].$$

3. Si $\mathsf{Cov}(X, Y) = 0$ alors les variables aléatoires sont dites **non-corrélées**.
4. Si les variables aléatoires X et Y sont non-triviales (i.e. non-constantes), leur **coefficient de corrélation** est donné par

$$r(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}.$$

Théorème 4.2.2. *Soient $X, Y, (X_n)_{n\in\mathbb{N}}$ de variables aléatoires de carré intégrables et $a, b, c, d \in \mathbb{R}$. Alors,*

1. *$aX + b, cY + d \in \mathcal{L}^2$ et $\mathsf{Cov}(aX + b, cY + d) = ac\mathsf{Cov}(X, Y)$;*
2. *$\mathsf{Cov}(X, Y)^2 \leq \mathsf{Var}(X)\mathsf{Var}(Y)$;*
3. *$\mathsf{Var}(\sum_{n=1}^N X_n) = \sum_{n=1}^N \mathsf{Var}(X_n) + \sum_{1\leq m\neq n\leq N} \mathsf{Cov}(X_m, X_n)$;*
4. *si X et Y sont indépendantes, alors elles sont non-corrélées.*

Démonstration. Exercice ! □

Remarque 4.2.3. La non-corrélation n'entraîne pas nécessairement l'indépendance. (Donner un exemple).

4.3 Fonction génératrice

Définition 4.3.1. Supposons que X est une variable aléatoire sur $(\Omega, \mathcal{F}, \mathbb{P})$, à valeurs dans $\mathbb{X} = \mathbb{N}$ et l'on note ρ_X la densité discrète de $\mathbb{P}_X$. On appelle **fonction génératrice** de X (ou de $\mathbb{P}_X$) la fonction

$$G(z) = \mathbb{E}(z^X) = \sum_{x\geq 0} \mathbb{P}(X = x) z^x = \sum_{x\geq 0} \rho_X(x) z^x, z \in [0,1].$$

La série entière qui définit la fonction G converge lorsque z se trouve dans l'intervalle $[0,1]$ car $G(1) = 1$. Calculons formellement $G'(z) = \sum_{x\geq 1} \rho_X(x) x z^{x-1}$. Maintenant, rien ne garantit plus que $G'(1) < \infty$; pour que cela soit le cas, il faut que la série entière $\sum_{x\geq 1} \rho_X(x) x < \infty$. Mais $\sum_{x\geq 1} \rho_X(x) x = \mathbb{E}(X)$.

Plus généralement, la fonction génératrice contient de manière condensée plusieurs caractéristiques de la loi de X comme le montre le théorème suivant :

Théorème 4.3.2. *Soit X est une variable aléatoire sur $(\Omega, \mathcal{F}, \mathbb{P})$, à valeurs dans $\mathbb{X} = \mathbb{N}$; on note $\rho := \rho_X$ la densité discrète de $\mathbb{P}_X$ et $G := G_X$ la fonction génératrice de X. Alors*

1. *Pour tout $n \in \mathbb{N}$, on a $\rho_X(n) = \frac{1}{n!} \frac{d^n G}{dz^n}(0)$. Par conséquent, la fonction génératrice détermine de manière unique la loi de la variable aléatoire X.*
2. *$\mathbb{E}(X) < \infty$ si, et seulement si, $G'(1-)$ existe et, dans ce cas, $\mathbb{E}(X) = G'(1-) = \lim_{z\uparrow 1} G'(z)$.*
3. *$\mathsf{Var}(X) < \infty$ si, et seulement si, $G''(1-)$ existe et, dans ce cas, $\mathsf{Var}(X) = G''(1-) - G'(1-)^2 + G'(1-)$.*

Démonstration. La démonstration est triviale si la loi de la variable aléatoire a un support fini. Nous nous concentrons donc au cas où X prend une infinité de valeurs dans $\mathbb{N}$ avec probabilité strictement positive.

1. Puisque pour $z \in [0,1]$, nous avons $G(z) = \sum_{k\geq 0} \rho(k) z^k$, il s'ensuit que $G(0) = \rho(0)$. De même, pour $0 \geq z < 1$, $G'(z) = \sum_{k\geq 1} \rho(k) k z^{k-1}$ et $G'(0) = \rho(1)$. Par récurrence, pour$0 \geq z < 1$, nous avons $G^{(n)}(z) = \sum_{k\geq n} \rho(k) k(k-1)\cdots 1 z^{k-1}$ et $G'(0) = k!\rho(k)$.
2. $$\begin{aligned} \frac{G(1) - G(z)}{1-z} &= \sum_{x\geq 1} \rho(x) \frac{1 - z^x}{1-z} = \sum_{x\geq 1} \rho(x)(1 - z^x) \sum_{k\geq 0} z^k \\ &= \sum_{x\geq 1} \rho(x) \sum_{k=0}^{x-1} z^k, \text{ expression bien définie sur } [0,1[. \end{aligned}$$

   En prenant la limite

   $$\begin{aligned} \lim_{z\uparrow 1} \frac{G(1) - G(z)}{1-z} &= \lim_{z\uparrow 1} \sum_{x\geq 1} \rho(x) \sum_{k=0}^{x-1} z^k \\ &= \sup_{z<1} \sup_{N\geq 1} \sum_{x=1}^{N} \rho(x) \sum_{k=0}^{x-1} z^k \\ &= \sup_{N\geq 1} \sum_{x=1}^{N} \rho(x) \sup_{z<1} \sum_{k=0}^{x-1} z^k, \quad \text{car, par monotonie, on peut intervertir les limites} \\ &= \sup_{N\geq 1} \sum_{x=1}^{N} \rho(x) x \\ &= \sup_{z<1} \sum_{x=1}^{\infty} \rho(x) x z^{x-1}, \quad \text{car } \sup_{z<1} x z^{x-1} = x. \end{aligned}$$

   Si toutes ces expressions restent finies lorsque $z = 1$, nous avons que $\mathbb{E}X = \lim_{z\uparrow 1} G'(z) = G'(1-)$.

3. De même,
$$\lim_{z\uparrow 1}\frac{G'(1-)-G'(z)}{1-z}=G''(1-)=\sum_{x\geq 2}\rho(x)x(x-1)\in[0,\infty].$$
Si $G''(1-)<\infty$, alors, $G''(1-)=\mathbb{E}(X^2)-\mathbb{E}(X)$.

□

4.4 Fonction caractéristique

Définition 4.4.1. Soit X une variable aléatoire réelle définie sur un espace probabilisé $(\Omega,\mathcal{F},\mathbb{P})$. On appelle **fonction caractéristique** , l'application $\chi_X:\mathbb{R}\to\mathbb{C}$, définie par
$$\chi_X(t)=\mathbb{E}(\exp(itX))=\int_\Omega\exp(itX(\omega))\mathbb{P}(\mathrm{d}\omega)=\int_{\mathbb{R}}\exp(itx)\mathbb{P}_X(\mathrm{d}t),t\in\mathbb{R}.$$

On remarque immédiatement que la fonction caractéristique ne dépend pas de X mais de $\mathbb{P}_X$; elle est la transformée de Fourier de cette loi.

Proposition 4.4.2. *La fonction caractéristique possède les propriétés suivantes :*

1. *Pour tout $t\in\mathbb{R}$, on a*
$$|\chi(t)|\leq 1=\chi(0)\quad et\quad \overline{\chi(t)}=\chi(-t).$$
2. *Elle est uniformément continue sur $\mathbb{R}$, i.e.*
$$\forall\varepsilon>0,\exists\delta>0:|h|<\delta\Rightarrow\forall t\in\mathbb{R},|\chi(t+h)-\chi(t)|\leq\varepsilon.$$
3. *La fonction caractéristique est réelle si, et seulement si, elle provient d'une loi symétrique.*
4. *Si pour un entier $n\geq 1$ on a $\mathbb{E}|X|^n<\infty$, alors $\chi_X^{(r)}$ existe pour tout entier $r\leq n$ et $\lim_{t\to 0}\varepsilon_n(t)=0$.*
   — $\chi_X^{(r)}(t)=\int_{\mathbb{R}}(ix)^r\exp(itx)\mathbb{P}_X(dx)$,
   — $\mathbb{E}X^r=\frac{\chi_X^{(r)}(0)}{i^r}$,
   — $\chi_X(t)=(\sum_{r=0}^n\frac{(it)^r}{r!}\mathbb{E}X^r)+\frac{(it)^n}{n!}\varepsilon_n(t)$, *où* $\varepsilon_n(t)\leq 3\mathbb{E}|X|^n$ *et* $\lim_{t\to 0}\varepsilon_n(t)=0$.
5. *Si $\chi_X^{(2n)}(0)$ existe et est finie, alors $\mathbb{E}X^{2n}<\infty$.*
6. *Si $\mathbb{E}|X|^n<\infty$ pour tout $n\geq 1$ et $\limsup_{n\to\infty}\frac{(\mathbb{E}|X|^n)^{1/n}}{n}=\frac{1}{eR}<\infty$, alors*
$$\chi_X(t)=\sum_{n\in\mathbb{N}}\frac{(it)^n}{n!}\mathbb{E}X^n,$$
   pour tout t avec $|t|<R$.
7. *Elle définit un noyau positif, i.e. pour toutes les familles finies de réels $(t_j)_{j=1,\dots,n}$ et de complexes $(z_j)_{j=a,\dots,n}$, on a*
$$\sum_{j=1}^n\sum_{k=1}^n\chi(t_j-t_k)z_j\bar{z}_k\geq 0.$$

Démonstration. Voir [65, Théorème 1, §II.12, pp. 278–281]. □

Théorème 4.4.3 (Critère de Bochner). *Une application $f : \mathbb{R} \to \mathbb{C}$ est une fonction caractéristique si, et seulement si, elle définit un noyau positif, elle est continue en 0 et vérifie $f(0) = 1$.*

Démonstration. Voir [15, Théorème 6.5.2, page 180] ou [47, Théorème 4.2.1, pp. 71–73] par exemple. □

Comme un cas particulier d'un résultat de Cramér [18] sur la représentation des certaines fonctions par des intégrales de Fourier, on a aussi le

Théorème 4.4.4 (Critère de Cramér). *Une fonction f complexe, bornée, continue sur $\mathbb{R}$ est caractéristique, si et seulement si,*

1. *$f(0) = 1$ et*
2. *la fonction g, définie par*

$$g(x, A) = \int_0^A \int_0^A f(t-s) \exp(i(t-s)x) dt ds,$$

est réelle et positive pour tout $x \in \mathbb{R}$ et tout $A > 0$.

Démonstration. Voir [47, Théorème 4.2.3, pp. 73–75]. □

Par ailleurs la fonction caractéristique caractérise la loi dans le sens que si X_1 et X_2 ont la même loi, alors $\chi_{X_1} = \chi_{X_2}$ et nous avons une formule d'inversion décrite par le

Théorème 4.4.5 (Formule d'inversion de Lévy). *La loi $\mathbb{P}_X$ de laquelle découle la fonction caractéristique χ_X s'obtient par la formule d'inversion*

$$\mathbb{P}_X(]x_1, x_2[) + \frac{\mathbb{P}_X(\{x_1\}) + \mathbb{P}_X(\{x_2\})}{2} = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{\exp(-itx_1) - \exp(-itx_2)}{it} \chi_X(t) dt.$$

Démonstration. Voir [14, Théorème 1, §8.3, page 287] □

4.5 Théorèmes des grands nombres

Avant de présenter les deux théorèmes de grands nombres, commençons par un exemple. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, avec $\Omega = \{\omega = (\omega_1, \ldots, \omega_n), \omega_i \in \{0, 1\}\}$ et $\mathcal{F} = \mathcal{P}(\Omega)$. On probabilise l'espace par la loi non-uniforme ayant la densité discrète $\rho(\omega) = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i}$ pour un $p \in [0, 1]$ et on s'intéresse à la famille de variables aléatoires $X_i : \Omega \to \{0, 1\}$, définies par $X_i(\omega) = \omega_i$, $i = 1, \ldots, n$. On peut alors calculer la i^e marginale

$$\begin{aligned}
\mathbb{P}(X_i = 1) &= \mathbb{P}(\{\omega : \omega_i = 1\}) \\
&= p \sum_{\omega_1,\ldots,\omega_{i-1},\omega_{i+1},\ldots,\omega_n} p^{\sum_{k\neq i}\omega_k}(1-p)^{n-1-\sum_{k\neq i}\omega_k} \\
&= p(1-p)^{n-1} \sum_{\omega_1,\ldots,\omega_{i-1},\omega_{i+1},\ldots,\omega_n} \prod_{k\neq i} \left(\frac{p}{1-p}\right)^{\omega_k} \\
&= p.
\end{aligned}$$

Par conséquent, on calcule de même que $\mathbb{P}(X_i = 0) = 1 - p$. Maintenant, si on note la somme partielle $S_k = X_1 + \ldots + X_k$, pour $k = 1, \ldots, n$, il est immédiat de calculer

$\mathbb{E}(S_k) = \sum_{i=1}^{k} \mathbb{E}(X_i) = pk$ et par conséquent $\mathbb{E}\left(\frac{S_n}{n}\right) = p$: l'espérance du nombre moyen de fois où l'on a observé « face » est égale à la probabilité d'obtenir « face » lors d'une réalisation.

D'un autre côté, on ne peut pas s'attendre à ce que pour tout $\varepsilon > 0$ et tout $\omega \in \Omega$ on ait $|\frac{S_n(\omega)}{n} - p| < \varepsilon$. En effet, pour $p \in]0,1[$,

$$\begin{aligned} \mathbb{P}\left(\frac{S_n}{n} = 1\right) &= \mathbb{P}(X_1 = 1, \ldots, X_n = 1) = p^n \\ \mathbb{P}\left(\frac{S_n}{n} = 0\right) &= \mathbb{P}(X_1 = 0, \ldots, X_n = 0) = (1-p)^n. \end{aligned}$$

En choisissant $p = \frac{1}{2}$ et $\varepsilon < \frac{1}{2^n}$, on voit qu'en notant

$$A_s = \left\{\omega \in \Omega : \frac{S_n(\omega)}{n} = s\right\}, s \in [0,1],$$

A_1 est un événement avec $\mathbb{P}(A_1) = p^n \neq 0$ (par conséquent non-vide). Pour des $\omega \in A_1$, on aura $\frac{S_n(\omega)}{n} - p = 1 - p = \frac{1}{2} \geq \varepsilon$.

Nous observons cependant que lorsque n est grand, les probabilités des événements A_0 et A_1 sont exponentiellement petites en n, puisque $\mathbb{P}(A_1) = p^n$ et $\mathbb{P}(A_0) = (1-p)^n$. Il est donc naturel de s'attendre à ce que la probabilité des événements $\{\omega \in \Omega : |\frac{S_n(\omega)}{n} - p| > \varepsilon\}$ soit petite.

Proposition 4.5.1 (Inégalité de Markov)**.** *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et ξ une variable aléatoire réelle positive définie sur Ω. Alors*

$$\forall \varepsilon > 0, \mathbb{P}(\xi \geq \varepsilon) \leq \frac{\mathbb{E}\xi}{\varepsilon}.$$

Démonstration. Nous avons de manière évidente

$$\xi = \xi \mathbb{1}_{\xi < \varepsilon} + \xi \mathbb{1}_{\xi \geq \varepsilon} \geq \xi \mathbb{1}_{\xi \geq \varepsilon} \geq \varepsilon \mathbb{1}_{\xi \geq \varepsilon}.$$

Par conséquent, $\mathbb{E}(\xi) \geq \varepsilon \mathbb{P}(\xi \geq \varepsilon)$. □

Remarque 4.5.2. L'inégalité de Markov présente un intérêt uniquement si $\mathbb{E}(\xi) < \infty$.

Corollaire 4.5.3. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et ξ une variable aléatoire réelle définie sur Ω. Alors, pour tout $\varepsilon > 0$, on a*

$$\begin{aligned} \mathbb{P}(|\xi| \geq \varepsilon) &\leq \frac{\mathbb{E}(|\xi|)}{\varepsilon} \\ \mathbb{P}(|\xi| \geq \varepsilon) &\leq \frac{\mathbb{E}(\xi^2)}{\varepsilon^2} \\ \mathbb{P}(|\xi - \mathbb{E}(\xi)|) \geq \varepsilon) &\leq \frac{\mathsf{Var}(\xi)}{\varepsilon^2} \text{ (inégalité de Bienaymé-Tchebychev).} \end{aligned}$$

Remarque 4.5.4. De nouveau ces inégalités présentent un intérêt si le second membre est petit et en particulier si les espérances ou la variance qui y apparaissent sont finies. L'inégalité de Bienaymé-Tchebychev est très grossière. Dans beaucoup de situations intéressantes, on peut améliorer de manière très substantielle le majorant de la probabilité $\mathbb{P}(|\xi - \mathbb{E}(\xi)|) \geq \varepsilon)$ (cf. exercice **??**).

Lemme 4.5.5. *Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles de carré intégrables (i.e. $X_i \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R})$, pour tout $i \in \mathbb{N}$) indépendantes et identiquement distribuées. On note $m = \mathbb{E}(X_1)$ l'espérance de l'une d'entre elles, $\sigma^2 = \mathsf{Var}(X_1)$ la variance de l'une d'entre elles et, pour tout entier $k \geq 0$, $S_k = \sum_{i=1}^k X_k$. Alors*

$$\begin{aligned} \mathbb{E}(S_k) &= km \\ \mathsf{Var}(S_k) &= k\sigma^2. \end{aligned}$$

Démonstration. En utilisant la linéarité de l'espérance, nous obtenons $\mathbb{E}(S_k) = \sum_{i=1}^k \mathbb{E}(X_i) = km$. Par 4.2.2, nous avons

$$\mathsf{Var}(S_k) = \sum_{i=1}^k \mathsf{Var}(X_i) + \sum_{i=1}^k \sum_{j=1;j\neq i}^k \mathsf{Cov}(X_i, X_j) = k\sigma^2$$

car l'indépendance entraîne la non-corrélation. □

Remarque 4.5.6. Dans le lemme 4.5.5, nous utilisons implicitement le fait que l'espace $(\Omega, \mathcal{F}, \mathbb{P})$ est suffisamment grand pour contenir *dans son intégralité* la suite $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$, avec $X_k : \Omega \to \mathbb{X}$, où $(\mathbb{X}, \mathcal{X})$ est un espace d'événements. Lorsque nous nous référons aux lois d'une variable aléatoire individuelle, par exemple X_k, nous sous-entendons que cette loi est la marginale uni-dimensionnelle de la loi conjointe pour *toute* la suite :

$$\mathbb{P}(X_k \in A) = \mathbb{P}(X_1 \in \mathbb{X}, \ldots X_{k-1} \in \mathbb{X}, X_k \in A, X_{k+1} \in \mathbb{X}, X_{k+2} \in \mathbb{X}, \ldots), A \in \mathcal{X}.$$

Théorème 4.5.7 (Théorème faible des grands nombres)**.** *Soit $(X_k)_{k\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées, définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, à valeurs dans $\mathbb{X} \subseteq \mathbb{R}$, telles que $\mathbb{E}(X_1^2) < \infty$. Alors*

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}\left(\left\{\omega \in \Omega : \left|\frac{S_n(\omega)}{n} - \mathbb{E}(X_1)\right| \geq \varepsilon\right\}\right) = 0.$$

Démonstration. La condition de carrée intégrabilité $\mathbb{E}(X_1^2) < \infty$ entraîne la finitude de σ^2. Étant donné que la suite est indépendante et équidistribuée, nous utilisons le résultat obtenu dans le lemme 4.5.5 pour établir que $\mathsf{Var}(S_n) = n\sigma^2$, où $\sigma^2 = \mathsf{Var}(X_1)$. Il s'ensuit, de l'inégalité de Biaynaymé-Tchebychev, que pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}(X_1)\right| \geq \varepsilon\right) = \mathbb{P}\left(|S_n - n\mathbb{E}(X_1)| \geq n\varepsilon\right) \leq \frac{\mathsf{Var}(S_n)}{n^2\varepsilon} = \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

□

Définition 4.5.8. Soit $(\xi_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires définies sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et ξ une autre variable sur $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que la suite (ξ_n) **converge en probabilité** vers ξ, et on note $\lim_{n\to\infty} \xi_n \overset{\mathbb{P}}{=} \xi$, si

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}(|\xi_n - \xi| > \varepsilon) = 0.$$

Remarque 4.5.9. Le théorème faible des grands nombres établit que pour une suite de variables aléatoires indépendantes, identiquement distribuées et de carré intégrables, on a $\lim_{n\to\infty} \frac{\sum_{k=1}^n X_k}{n} \overset{\mathbb{P}}{=} \mathbb{E}(X_1)$. Les conditions d'applicabilité du théorème peuvent être affaiblies de différentes manières.

— Pour des variables aléatoires de carré intégrables, le théorème reste valable pour des variables qui sont seulement décorrélées au lieu d'être indépendantes ; même la condition d'équidistribution peut être affaiblie. On a alors $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n (X_k - \mathbb{E}(X_k)) \overset{\mathbb{P}}{=} 0$.

— La condition de carré-intégrabilité peut être affaiblie en une condition d'intégrabilité. La suite doit, dans ce cas, rester indépendante et équidistribuée (voir [65, Théorème 2, §III.3, pp. 325–326]). On a alors $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n X_k \overset{\mathbb{P}}{=} \mathbb{E}(X_1)$.

Lemme 4.5.10. *Soient (ζ_n) et (ξ_n) des suites de variables aléatoires sur $(\Omega, \mathcal{F}, \mathbb{P})$ et (a_n) une suite numérique. Alors*

1. $[\lim_{n\in\mathbb{N}} \zeta_n \overset{\mathbb{P}}{=} 0] \wedge [\lim_{n\in\mathbb{N}} \xi_n \overset{\mathbb{P}}{=} 0] \Longrightarrow [\lim_{n\in\mathbb{N}} (\zeta_n + \xi_n) \overset{\mathbb{P}}{=} 0]$.
2. $[\lim_{n\in\mathbb{N}} \xi_n \overset{\mathbb{P}}{=} 0] \wedge [\sup_{n\in\mathbb{N}} |a_n| < \infty] \Longrightarrow [\lim_{n\in\mathbb{N}} a_n \xi_n \overset{\mathbb{P}}{=} 0]$.

D'autres types de convergence, plus fortes que la convergence en probabilité, sont possibles pour des suites de variables aléatoires.

Définition 4.5.11. Soit $(\xi_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires définies sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et ξ une autre variable sur $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que la suite (ξ_n) **converge presque sûrement** vers ξ, et on note $\lim_{n\to\infty} \xi_n \overset{\text{p.s.}}{=} \xi$, si

$$\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)\}) = 1.$$

Théorème 4.5.12 (Théorème fort des grands nombres)**.** *Soit une suite (X_n) de variables aléatoires réelles définies sur un espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$ qui sont deux-à-deux décorrélées et vérifiant $\sup_{n\in\mathbb{N}} \mathsf{Var}(X_n) < \infty$. Alors*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n (X_k - \mathbb{E}(X_k)) \overset{p.s.}{=} 0.$$

La démonstration de ce théorème reste en dehors des exigences de ce cours. L'utilisation de l'adjectif « fort » est justifiée par le fait que l'on peut montrer que la convergence presque sûre entraîne la convergence en probabilité mais que la réciproque n'est pas vraie.

Exercice 4.5.13. (Principe d'une simulation Monte Carlo pour des échantillons indépendants). Soit $(U_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées avec la loi uniforme sur l'intervalle $[0, 1]$. Soit $f : [0, 1] \to \mathbb{R}$ une application de carré intégrable. On construit la suite des sommes partielles $S_n = \sum_{k=1}^n f(U_k)$. La suite $\frac{S_n}{n}$ converge-t-elle en probabilité et si oui vers quoi ? Application : simuler sur ordinateur avec la suite $\frac{S_n}{n}$ avec $f(x) = \frac{1}{1+x^2}$.

4.6 Théorème central limite

Les théorèmes des grands nombres affirment qu'en répétant plusieurs fois la même expérience aléatoire, la moyenne empirique des observations finit par se stabiliser à la valeur numérique de l'espérance. Le théorème central limite nous renseigne sur la dispersion de ces observations empiriques autour la valeur moyenne. On commence par rappeler deux résultats élémentaires.

Lemme 4.6.1. *Soit* $(c_n)_{n\in\mathbb{N}}$ *une suite de nombre complexes vérifiant* $\lim_{n\to\infty} c_n = c$ *pour un certain* $c \in \mathbb{C}$. *Alors* $\lim_{n\to\infty}(1+\frac{c_n}{n})^n = \exp(c)$.

Lemme 4.6.2. *On note* χ_{m,σ^2} *la fonction caractéristique de la loi normale de moyenne* m *et de variance* σ^2, *c'est-à-dire de la loi ayant une densité* $\frac{1}{\sqrt{2\pi}\sigma}\exp(-(x-m)^2/2\sigma^2)$ *par rapport à la mesure de Lebesque sur* $\mathbb{R}$. *On a alors*

$$\chi_{m,\sigma^2}(t) = \exp(itm - t^2\sigma^2/2).$$

Dans le tableau 4.1 on rappelle les densités ρ de quelques lois usuelles ainsi que leurs fonctions caractéristiques χ.

Loi de X	$\mathbb{X}$	ρ_X	χ_X
Binomiale $\mathcal{B}_{n,p}$	$\{0,\dots,n\}$	$C_n^x p^x (1-p)^{n-x}$	$(1-pe^{it})^n$
Poisson(λ)	$\mathbb{N}$	$\exp(-\lambda)\frac{\lambda^x}{x!}$	$\exp(\lambda(e^{it}-1))$
Uniforme $[a,b]$	$[a,b]\subset\mathbb{R}$	$\frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$	$\frac{1}{b-a}\frac{e^{itb}-e^{ita}}{it}$
Normale $\mathcal{N}(m,\sigma^2)$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-m)^2}{2\sigma^2})$	$\exp(itm - t^2\sigma^2/2)$
$\Gamma(p,\lambda)$	$\mathbb{R}_+$	$\frac{1}{\Gamma(p)}\lambda^p x^{p-1}\exp(-\lambda x)$	$(1-\frac{it}{\lambda})^{-p}$

TABLE 4.1 – Densités et fonctions caractéristiques pour quelques lois usuelles.

Théorème 4.6.3. *(Théorème central limite) Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de variables aléatoires indépendantes et identiquement distribuées avec* $\mathbb{E}X_1 = m$ *et* $\mathsf{Var}X_1 = \sigma^2$. *En notant* $S_n = \sum_{k=1}^n X_k$, *on a*

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{S_n - nm}{\sigma\sqrt{n}} \le x\right) = \Phi_{0,1}(x),$$

où $\Phi_{0,1}$ *désigne la fonction de répartition d'une variable aléatoire normale d'espérance* 0 *et de variance* 1.

La convergence établie par ce théorème est une convergence plus faible que la convergence en probabilité ; on l'appelle **convergence en loi.**

Définition 4.6.4. Soient $(\xi_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles sur un espace $(\Omega,\mathcal{F},\mathbb{P})$ et ξ une variable aléatoire réelle sur le même espace. On dit que la suite $(\xi_n)_{n\in\mathbb{N}}$ **converge en loi** vers ξ, et l'on note [1] $\lim_{n\to\infty}\xi_n \overset{\text{loi}}{=} \xi$, si, pour tout $x\in\mathbb{R}$,

$$\lim_{n\to\infty}\mathbb{P}(\xi_n \le x) = F_\xi(x) := \mathbb{P}(\xi \le x).$$

1. Cette convergence est aussi notée $\lim_{n\to\infty}\xi_n \overset{\text{d}}{=} \xi$, surtout dans la littérature anglo-saxonne (pour convergence in distribution).

Démonstration du théorème 4.6.3. Puisque pour toute variable aléatoire réelle X et tout couple de nombres réels a, b on $\chi_{aX+b}(t) = \chi_X(at)\exp(itb)$ (voir exercice 54, on peut sans perte de généralité, supposer que $\mathbb{E}X_1 = 0$.

On utilise l'indépendance des variables et les faits que $\chi'_{X_1}(0) = \mathbb{E}X_1 = 0$ et $\chi''_{X_1}(0) = \mathbb{E}X_1^2 = \sigma^2$ pour des variables d'espérance nulle et de variance σ^2, pour écrire

$$\mathbb{E}(\exp(it\frac{S_n}{\sigma\sqrt{n}}) = \left(\chi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n.$$

Or, en utilisant la proposition 4.4.2, on a, pour grand n, que $\chi(\frac{t}{\sigma\sqrt{n}}) = 1 - \frac{\sigma^2}{2}(\frac{t}{\sigma\sqrt{n}})^2 + o(\frac{|t|}{\sigma\sqrt{n}})^2$. Par conséquent,

$$\lim_{n\to\infty}\mathbb{E}(\exp(it\frac{S_n}{\sigma\sqrt{n}})) = \left(1 - \frac{\sigma^2}{2}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + o\left(\frac{|t|}{\sigma\sqrt{n}}\right)^2\right)^n = \exp(-t^2/2) = \chi_{\mathcal{N}(0,1)}(t), t \in \mathbb{R},$$

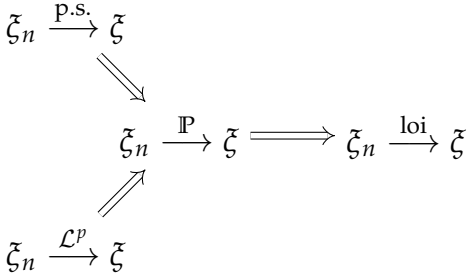
où $\chi_{\mathcal{N}(0,1)}$ est la fonction caractéristique de la loi normale centrée (i.e. d'espérance nulle) réduite (i.e. de variance 1). □

Remarque 4.6.5. La condition de carré intégrabilité ne peut pas être affaiblie. Si les variables aléatoires de la suite sont intégrables mais pas de carré intégrables (i.e. ont des "queues lourdes"), la bonne normalisation n'est plus en $\sqrt{n}$ mais des exposants déterminés par la puissance des moments fractionnaires qui existent et la convergence peut avoir lieu vers des lois stables autres que la loi normale.

Dans ce chapitre, nous avons introduit plusieurs notions de convergence : en probabilité, presque sûre, en loi. On peut introduire d'autres, par exemple les convergences en $\mathcal{L}^p$, avec $p \in [1, \infty[$ définies par

$$\lim_n \xi_n \overset{\mathcal{L}^p}{=} \xi \Leftrightarrow \lim_n \|\xi_n - \xi\|_p = 0,$$

où $\|\xi_n - \xi\|_p = [\mathbb{E}(|\xi_n - \xi|^p)]^{1/p}$. La raison d'être de toutes ces définition est qu'elles ne sont pas équivalentes ; nous avons l'hiérarchie suivante :



4.7 Exercices

Lois, espérance, variance

40. Soit α la variable aléatoire définie sur $(\Omega, \mathcal{F}, \mathbb{P})$ qui prend des valeurs dans $\mathbb{X} = \{0, \pi/2, \pi\}$ avec probabilité uniforme. On note $X = \sin\alpha$ et $Y = \cos\alpha$. Calculer

(a) $\mathbb{E}X$,
(b) $\mathbb{E}Y$,
(c) $\mathsf{Cov}(X,Y)$,
(d) $\mathbb{P}(X=1,Y=1)$.

Quelle est votre conclusion ?

41. Soient $(\xi_k)_{k=1,\dots,n}$ des variables aléatoires indépendantes sur $(\Omega,\mathcal{F},\mathbb{P})$ prenant un nombre fini de valeurs réelles. Calculer la loi de $X=\max\xi_k, k=1,\dots,n$ et de $Y=\min\{\xi_k, k=1,\dots,n\}$.
42. Soient $(\xi_k)_{k=1,\dots,n}$ des variables aléatoires indépendantes sur $(\Omega,\mathcal{F},\mathbb{P})$ à valeurs dans $\mathbb{X}=\{0,1\}$ et $(\lambda_k)_{k=1,\dots,n}$ une suite de nombres strictement positifs fixés. On sait que $\mathbb{P}(\xi_k=1)=\lambda_k\Delta$ avec Δ un petit nombre réel strictement positif.
    (a) Pouvez-vous donner une borne sur Δ qui traduit proprement la notion de « petitesse » ?
    (b) Estimer $\mathbb{P}(\xi_1+\ldots+\xi_n=1)$ en ordre Δ^2.
    (c) Estimer $\mathbb{P}(\xi_1+\ldots+\xi_n>1)$.
43. Soient X et Y deux variables aléatoires réelles sur $(\Omega,\mathcal{F},\mathbb{P})$ avec $\mathsf{Var}X>0$ et $\mathsf{Var}Y>0$.
    (a) Montrer que $|r(X,Y)|\leq 1$.
    (b) Monter que $|r(X,Y)|=1$ si et seulement si $X=aY+b$.
44. Soit ξ une variable aléatoire réelle avec $\mathbb{E}(\xi^2)<\infty$.
    (a) Montrer que $\mathbb{E}(|\xi|)<\infty$.
    (b) Montrer que $\inf_{a\in\mathbb{R}}\mathbb{E}((\xi-a)^2)$ est atteint pour une valeur $a_0\in\mathbb{R}$. Déterminer ce a_0.
    (c) Déterminer la valeur de $\inf_{a\in\mathbb{R}}\mathbb{E}((\xi-a)^2)$ en termes d'une quantité relative à la variable ξ.
45. Soit ξ une variable aléatoire réelle de loi $\mathbb{P}_\xi$ et fonction de répartition F_ξ.
    (a) Déterminer $F_{a\xi+b}$ pour $a>0$ et $b\in\mathbb{R}$.
    (b) Déterminer F_{ξ^2}.
    (c) Si $\xi^+=\max(\xi,0)$, déterminer F_{ξ^+}.
46. Soient ξ,η de variables aléatoires réelles définies sur $(\Omega,\mathcal{F},\mathbb{P})$ à valeurs dans une partie discrète $\mathbb{X}\subset\mathbb{R}$. On note $\mathbb{P}_{(\xi,\eta)}$ leur loi conjointe ; on suppose que $\mathbb{E}\xi=\mathbb{E}\eta=0$ et $\mathsf{Var}\xi=\mathsf{Var}\eta=1$. On note $r:=r(\xi,\eta)$ le coefficient de corrélation de ξ et η. Monter que
$$\mathbb{E}(\max(\xi^2,\eta^2))\leq 1+\sqrt{1-r^2}.$$
47. (Identité de Wald). Soient $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles, indépendantes et identiquement distribuées définies sur $(\Omega,\mathcal{F},\mathbb{P})$ et τ une variable aléatoire définie sur le même espace $(\Omega,\mathcal{F},\mathbb{P})$, à valeurs dans $\mathbb{N}$. Nous supposons que $\mathbb{E}(|X_1|)<\infty$, $\mathbb{E}(\tau)<\infty$ et que la variable τ est indépendante des variables aléatoires $(X_n)_{n\in\mathbb{N}}$. Montrer que la variable aléatoire
$$S_\tau=\sum_{n=1}^{\tau}X_i$$
a une espérance et que $\mathbb{E}(S_\tau)=\mathbb{E}(\tau)\mathbb{E}(X_1)$. (La somme S_τ est une somme partielle de la série de terme général X_n comportant un nombre *aléatoire* de termes).

Fonctions génératrices

48. On rappelle qu'une variable aléatoire X est dite suivre la loi binomiale $\mathcal{B}_{n,p}$ de paramètres $p \in [0,1]$ et $n \geq 1$ si $\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k}$, pour $k = 0, \ldots, n$.
    (a) Calculer la fonction génératrice $G := G_{\mathcal{B}_{n,p}}$ correspondante.
    (b) Calculer $\mathbb{E}(X)$ à l'aide de la fonction génératrice.
    (c) Calculer $\mathsf{Var}(X)$ à l'aide de la fonction génératrice.
49. On rappelle qu'une variable aléatoire X est dite suivre la loi exponentielle $\mathcal{E}_\lambda$ de paramètre $\lambda > 0$ si $\mathbb{P}(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}$, pour $k \in \mathbb{N}$.
    (a) Calculer la fonction génératrice $G := G_{\mathcal{E}_\lambda}$ correspondante.
    (b) Calculer $\mathbb{E}(X)$ à l'aide de la fonction génératrice.
    (c) Calculer $\mathsf{Var}(X)$ à l'aide de la fonction génératrice.
50. Soient X et Y deux variables aléatoires indépendantes sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{N}$. Exprimer la fonction génératrice G_{X+Y} de la somme $X + Y$, en termes de fonctions génératrices G_X et G_Y des variables individuelles.
51. En se servant du résultat de l'exercice 50, déterminer $G_{\mathcal{B}_{n,p}}$ en termes de $G_{\mathcal{B}_{1,p}}$ et comparer avec le résultat direct, obtenu en exercice 48.
52. (Extrait du CC du 16 novembre 2017).
    Soit $\mathbf{p} := (p_n)_{n\in\mathbb{N}}$ la suite définie par
$$p_n = \frac{a^n}{(1+a)^{n+1}}, \quad \text{pour} \quad n \in \mathbb{N} \quad \text{et} \quad a > 0.$$
    (a) Montrer que pour tout $a > 0$, la suite $\mathbf{p}$ est un vecteur de probabilité sur $\mathbb{N}$.
    (b) Soit X une variable aléatoire sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{N}$ dont la loi est déterminée par le vecteur de probabilité $\mathbf{p}$. Calculer la fonction génératrice $G_X(z)$ de X pour $z \in [-1, 1]$.
    (c) Calculer $G'_X(z)$ et $G''_X(z)$.
    (d) En déduire $\mathbb{E}(X)$ et $\mathsf{Var}(X)$.
53. Une somme de variables aléatoires comportant un nombre aléatoire de termes. (Extrait de l'examen du 15 décembre 2018).
    Soient $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles, indépendantes et identiquement distribuée, d'espérance nulle et de variance finie, et N une variable aléatoire à valeurs dans $\mathbb{N}$, indépendante de la suite (X_n), de loi $\nu(k) = \mathbb{P}(N = k) = \frac{1}{2^{k+1}}$ pour tout $k \in \mathbb{N}$. On note $S_N = \sum_{n=1}^N X_n$ et $\sigma^2 = \mathsf{var}(X_1)$.
    (a) Calculer la fonction génératrice $G(z) = \mathbb{E}(z^N)$ et s'en servir pour calculer $\mathbb{E}(N)$.
    (b) Montrer que $\mathbb{E}(S_N) = 0$.
    (c) Calculer $\mathsf{var}(S_N)$. (Utiliser l'indépendance !)

Fonctions caractéristiques

54. Soient X une variables aléatoire réelle et $a, b \in \mathbb{R}$. Montrer que $\chi_{aX+b}(t) = \chi_X(at)\exp(itb)$.
55. Soient $(\chi_n)_{n\in\mathbb{N}}$ une suite de fonctions caractéristiques et $(\lambda_n)_{n\in\mathbb{N}}$ une suite de variables positives avec $\sum_{n\in\mathbb{N}} \lambda_n = 1$. Monter que $\sum_{n\in\mathbb{N}} \lambda_n \chi_n$ est encore une fonction caractéristique.

56. Soit la famille $(\chi_j)_{j=1,\dots,n}$ de fonctions caractéristiques. Montrer que $\prod_{j=1}^n \chi_j$ est encore une fonction caractéristique.

Inégalité de Bienaymé-Tchebychev, théorèmes des grands nombres

57. Dans cet exercice, les variables aléatoires utilisées sont à valeurs dans des parties finies de $\mathbb{R}$. Toutes les espérances utilisées dans l'énoncé existent.
    (a) Soit ξ une variable aléatoire discrète à valeurs dans une partie de $\mathbb{R}_+$. Rappeler comment on démontre, pour tout $a > 0$, l'inégalité de Markov-Tchebychev
    $$\mathbb{P}(\xi \geq a) \leq \frac{\mathbb{E}\xi}{a}.$$
    (b) Soient $a > 0$ et $f : \mathbb{R} \to \mathbb{R}_+$ une fonction croissante telle que $f(a) > 0$. Montrer que
    $$\mathbb{P}(\xi \geq a) \leq \frac{\mathbb{E}(f(\xi))}{f(a)}.$$
    (c) En conclure que
    $$\mathbb{P}(\xi \geq a) \leq \inf_{s>0} \left[\exp(-sa)\mathbb{E}(\exp(s\xi))\right].$$
    (d) Soit $(X_i)_{i\in\mathbb{N}}$ une suite de variables aléatoires indépendantes à valeurs dans $\{0,1\}$ et de même loi, chargeant 1 avec probabilité p, avec $p \in]0,1[$. Calculer, pour un $s > 0$ arbitraire,
    $$g(s) := \mathbb{E}\left(\exp(s\sum_{i=1}^n X_i)\right)$$
    et utiliser cette formule explicite pour majorer, pour un $a \in]p,1[$, la probabilité $\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \geq a)$.
    (e) Déterminer $s_0 := \arg\min_{s>0} \exp(-nas)g(s)$.
    (f) En conclure que
    $$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq a\right) \leq \exp\left(nh(a,p)\right),$$
    où $h(a,p) := -a\log\frac{a}{p} - (1-a)\log\frac{1-a}{1-p}$.

58. Soit $(\xi_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes réelles et de même loi, vérifiant $\mathbb{P}(\xi_n = 1) = p = 1 - \mathbb{P}(\xi_n = -1)$, pour tout $n \in \mathbb{N}$. On note $S_n = \sum_{i=1}^n \xi_i$. Le but de l'exercice est d'établir, dans le cas particulier où $p = 1/2$, la formule (valable pour $p \in [0,1]$ arbitraire)
    $$\forall \varepsilon > 0,\ \mathbb{P}\left(\left|\frac{S_n}{n} - (2p-1)\right| \geq \varepsilon\right) \leq 2\exp(-\frac{1}{2}\varepsilon^2 n).$$
    Il s'agit d'une amélioration de l'inégalité de Bienaymé-Tchebychev. On verra plus loin (exercice 74) que cette inégalité peut se généraliser à des variables aléatoires bornées indépendantes non nécessairement équidistribuées.

59. Dans une circonscription d'un million d'électeurs les candidats A et B sont en lice. Parmi ces électeurs, 2000 connaissent bien le candidat A et votent unanimement pour lui. Les autres 998000 choisissent purement au hasard en lançant une pièce honnête. Minorer la probabilité pour que le candidat A gagne.
60. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées, avec $\mathbb{E}|X_1| < \infty$. On note $m = \mathbb{E}X_1$.
    (a) Calculer la fonction caractéristique de la variable aléatoire $\frac{S_n}{n}$, où $S_n = \sum_{k=1}^n X_k$.
    (b) S'en servir pour montrer que $\frac{S_n}{n} \xrightarrow{\text{loi}} m$.
61. Soit $(\Psi_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes, uniformément distribuées sur $[0, 2\pi[$, définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. Une particule se déplace aléatoirement dans le plan selon la règle suivante : lorsque à l'instant n elle est en position $\boldsymbol{\xi}_n \in \mathbb{R}^2$, à l'instant $n+1$ elle sera en une position $\boldsymbol{\xi}_{n+1} \in \mathbb{R}^2$ telle que la longueur du déplacement $\boldsymbol{\xi}_{n+1} - \boldsymbol{\xi}_n$ est une constante $r > 0$ et l'angle formé par ce déplacement et l'axe des abscisses est Ψ_{n+1}. Soit $D_n^2 = \|\boldsymbol{\xi}_n - \boldsymbol{\xi}_0\|^2$ la distance entre les positions initiale et au temps n de la particule.
    (a) Calculer $\lim_{n\to\infty} \frac{\boldsymbol{\xi}_n}{n}$.
    (b) Calculer $\mathbb{E}(D_n^2)$.
    (c) Calculer $\lim_{n\to\infty} \frac{D_n^2}{n}$.
62. Soit $(T_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées à valeurs réelles positives, bornées (donc intégrables), définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. La variable aléatoire T_n sera interprétée comme représentant la durée de vie de la n^e ampoule changée. Dès qu'une ampoule est grillée on la remplace aussitôt. Pour tout $t > 0$, on note
$$N_t = \sup\{N \geq 1 : \sum_{n=1}^{N} T_n \leq t\}$$
le nombre d'ampoules utilisées jusqu'au temps t. Montrer que $\lim_{t\to\infty} \frac{N_t}{t} \overset{\text{p.s.}}{=} \frac{1}{\mathbb{E}(T_1)}$.
63. (Extrait du contrôle du 16 novembre 2017).
    On considère un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ suffisamment grand pour pouvoir contenir simultanément toutes les variables aléatoires dont on se servira dans cet exercice.
    (a) Soient X et Y deux variables aléatoires indépendantes sur $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{X} = \{0, 1\}$. La loi de X est déterminée par le vecteur de probabilité $\mathbf{p} = (p, 1-p)$ et celle de Y par le vecteur $\mathbf{q} = (q, 1-q)$, où $0 \leq p, q \leq 1$. Exprimer la valeur $r := \mathbb{P}(X = Y)$ en fonction des p et q.
    (b) Calculer $\mathsf{Var}(\mathbb{1}_{\{X=Y\}})$.
    (c) On considère maintenant deux suites définies sur $(\Omega, \mathcal{F}, \mathbb{P})$. La première est une suite $(X_n)_{n\in\mathbb{N}}$ de copies indépendantes de X (i.e. elle est indépendante et identiquement distribuée selon la loi $\mathbf{p}$). La seconde est une suite $(Y_n)_{n\in\mathbb{N}}$ de copies indépendantes de Y (i.e. elle est indépendante et identiquement distribueé selon la loi $\mathbf{q}$). La deux suites sont en outre mutuellement indépendantes. Utiliser un résultat du cours pour montrer que la suite $\frac{1}{N}\sum_{n=1}^N \mathbb{1}_{\{X_n=Y_n\}}$ converge en probabilité vers une constante qu'il faudra déterminer.

5

Chaînes de Markov sur des espaces d'états dénombrables

Les chaînes de Markov présentent des multiples intérêts :

— De point de vue purement probabiliste, elles s'intègrent dans la hiérarchie générale des modèles probabilistes introduite par le théorème **??** ; elles font partie du niveau qui se situe juste au dessus des suites indépendantes étudiées dans le chapitre précédant.

— En théorie de l'information, elles modélisent le fonctionnement des canaux de transmission et permettent l'étude de leur fonctionnement asymptotique.

— En théorie de la complexité, elles modélisent la notion d'automate fini (et plus généralement de machine de Turing) et les méthodes développées pour leur étude servent à étudier le « problème d'arrêt » en théorie de calculabilité.

5.1 Probabilités de transition, matrices stochastiques

Soit $\mathbb{X}$ un espace dénombrable (muni de sa tribu exhaustive). On considère un système aléatoire qui peut occuper n'importe quel état de $\mathbb{X}$. À l'instant initial $n = 0$, le système occupe l'état $x_0 \in \mathbb{X}$ avec $\mathbb{P}(X_0 = x_0) = \rho_0(x_0)$. Aux instants suivants $n = 1, 2, \ldots$ le système change son état en $X_1, X_2, \ldots$ selon des probabilités conditionnelles déterminées comme suit : Supposons que jusqu'à l'instant n le système ait visité les états $x_0, x_1, \ldots, x_n$. Alors, il évoluera selon la probabilité conditionnelle vérifiant la condition

$$\rho_{n+1,\mathbf{x}\restriction_n}(y) := \mathbb{P}(X_{n+1} = y | X_0 = x_0, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = y | X_n = x_n) =: P_{x_n,y},$$

i.e. l'évolution future du système oublie le passé pour ne se souvenir que du présent. Cette condition de dépendance s'appelle **propriété faible de Markov**.

Cette évolution est un cas particulier du théorème 3.1.8. Il est évident que $\forall x, y \in \mathbb{X}$, nous avons $P_{x,y} \geq 0$ et $\sum_{z\in\mathbb{X}} P_{x,z} = 1$, i.e. chaque ligne de la matrice $P = (P_{x,y})_{x,y\in\mathbb{X}}$ est un vecteur de probabilité sur $\mathbb{X}$. Une telle matrice est appelée **matrice stochastique**.

Définition 5.1.1. Soient $\mathbb{X}$ un espace dénombrable, $\mathcal{X} = \mathcal{P}(\mathbb{X})$ sa tribu exhaustive et P une matrice stochastique sur $\mathbb{X}$. On note $\Omega = \mathbb{X}^{\mathbb{N}}$ et $\mathcal{F} = \otimes_{n\in\mathbb{N}}\mathcal{X}$ et ρ_0 un vecteur de probabilité sur $\mathcal{X}$. Soit $\mathbb{P}_{\rho_0}$ l'unique probabilité (qui existe en vertu du théorème 3.1.8) sur $(\Omega, \mathcal{F})$ avec $\rho_{n+1,\mathbf{x}\restriction_n}(y) = P_{x_n,y}$. Alors la suite de variables aléatoires $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ sur $(\Omega, \mathcal{F}, \mathbb{P}_{\rho_0})$ à valeurs dans $(\mathbb{X}, \mathcal{X})$ vérifiant

1. $\mathbb{P}_{\rho_0}(X_0 = x_0) = \rho_0(x_0)$ et
2. $\mathbb{P}(X_{n+1} = x_{n+1}|X_0 = x_0, \ldots, X_n = x_n) = P_{x_n,x_{n+1}}$

est appelée **chaîne de Markov** à espace d'états $\mathbb{X}$, matrice stochastique P et probabilité initiale ρ_0, notée $\mathsf{CM}(\mathbb{X}, P, \rho_0)$.

Le cas où $\mathbb{X}$ est non-dénombrable ne sera pas abordé dans ce cours (cf. [56], par exemple, pour une construction dans le cas où $\mathbb{X}$ est un espace mesurable).

En traduisant la construction du théorème 3.1.8 dans la situation présente, nous obtenons la loi conjointe de $(X_0, \ldots, X_n)$ comme marginale fini-dimensionnelle de la loi $\mathbb{P}_{\rho_0}$, par la formule

$$\mathbb{P}_{\rho_0}(X_0 = x_0, \ldots, X_n = x_n) = \rho_0(x_0)P_{x_0,x_1} \cdots P_{x_{n-1}x_n}.$$

À cause de la forme simplifiée qu'a le second membre, nous pouvons alors calculer la n^e marginale

$$\begin{aligned}\mathbb{P}_{\rho_0}(X_n = x_n) &= \sum_{x_0,\ldots,x_{n-1}\in\mathbb{X}} \rho_0(x_0)P_{x_0,x_1} \cdots P_{x_{n-1},x_n} \\ &= (\rho_0 P^n)(x_n),\end{aligned}$$

où le vecteur de probabilité ρ_0 est écrit sous forme de vecteur ligne. On voit donc que dans l'hiérarchie des modèles décrits par le théorème **??**, le modèle de dépendance markovienne se situe au niveau de complexité juste au dessus du modèle indépendant introduit en §3.2.

Remarque 5.1.2. Lorsque le vecteur $\rho_0(y) = \delta_{x,y}$ pour un $x \in \mathbb{X}$ donné, nous écrivons $\mathbb{P}_x$ au lieu de $\mathbb{P}_{\rho_0}$. Nous avons alors $\mathbb{P}_x(X_n = y) = P^n(x, y)$. Il ne faut pas confondre la notation $\mathbb{P}_X$, signifiant la loi d'une variable aléatoire X, avec $\mathbb{P}_x$, signifiant la loi sur l'espace des trajectoires avec démarrage initiale déterministe $\rho_0 = \varepsilon_x$. En fait, $\mathbb{P}_x$ désigne la loi conjointe $\mathbb{P}_{\mathbf{X}}$ de la suite infinie $\mathbf{X} = (X_0, X_1, \ldots)$ conditionnée à l'évènement $\{X_0 = x\}$.

Exemple 5.1.3. On dispose de deux pièces de monnaie, une honnête et une lestée, donnant « face » avec probabilité $1/3$. Si la pièce lancée au n^e jeu montre « pile » on utilise la pièce honnête pour la lancer au $(n+1)^e$ jeu ; si elle montre « face » on utilise la pièce lestée pour jouer au $(n+1)^e$ jeu. Si on note (X_n) la suite de variables aléatoires dans $\mathbb{X} = \{0, 1\}$ correspondant à ce jeu, elles constituent une chaîne de Markov, avec probabilité conditionnelle de transition

$$\mathbb{P}(X_{n+1} = y|X_n = x) = P_{xy},$$

où la matrice $P := (P_{xy})_{x,y\in\mathbb{X}}$ est égale à $P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}$. L'étude du comportement asymptotique à grand n de la probabilité $\mathbb{P}_{\rho_0}(X_n = x) = (\rho_0 P^n)(x)$ peut se faire en s'inspirant de la méthode développée à l'exercice 67.

Théorème 5.1.4. *La définition d'une chaîne de Markov est équivalente à la propriété d'indépendance du futur et du passé, conditionnellement au présent. Plus précisément, en notant pour $1 < p < n$ et $x_1, \dots, x_n \in \mathbb{X}$,*

$$\begin{aligned} A &= \{X_1 = x_1, \dots, X_{p-1} = x_{p-1}\}, \\ B &= \{X_p = x_p\}, \\ C &= \{X_{p+1} = x_{p+1}, \dots, X_n = x_n\}, \end{aligned}$$

nous avons

$$\mathbb{P}(A \cap C | B) = \mathbb{P}(A|B)\mathbb{P}(C|B).$$

Démonstration. Exercice. □

5.2 Temps d'arrêt. Propriété forte de Markov

Définition 5.2.1. Soit $(X_n) \in \mathsf{CM}(\mathbb{X}, P, \rho_0)$ définie sur un espace $(\Omega, \mathcal{F}, \mathbb{P})$. Une variable aléatoire $T : \Omega \to \mathbb{N} \cup \{+\infty\}$ est un **temps d'arrêt**, si pour tout $n \in \mathbb{N}$, l'événement $\{T = n\}$ est entièrement déterminé par la donnée des valeurs prises[1] par les variables $X_0, \dots, X_n$.

Exemple 5.2.2. Soit $(X_n)_{n\in\mathbb{N}}$ le relevée de température (en degrés Celsius avec une décimale) à l'ombre dans un lieu donné à midi du jour n (où $n = 0$ correspond au jour où des éphémérides ont commencé à exister). Soit $T_1 := \inf\{n \geq 0 : X_n = 42.2\}$ et $T_2 := \inf\{n \geq 0 : X_n$ est la température maximale jamais enregistrée en ce lieu$\}$. Évidemment T_1 est un temps d'arrêt mais pas T_2. Qu'en est-t-il de $T_3 := \inf\{n \geq 0 : X_n > \max_{k=0,\dots,n-1} X_k\}$?

Remarque 5.2.3. La notion de temps d'arrêt est importante en théorie de la complexité algorithmique. Une machine de Turing est en effet une chaîne de Markov ; une condition nécessaire pour que la machine de Turing (i.e. un programme informatique) résolve un problème algorithmique est de s'arrêter en un temps fini. La calculabilité algorithmique d'un problème est donc directement reliée à l'existence d'un temps d'arrêt fini pour la machine de Turing sensée le résoudre.

Définition 5.2.4. Soit $(X_n) \in \mathsf{CM}(\mathbb{X}, P, \rho_0)$ définie sur $(\Omega, \mathcal{F}, \mathbb{P})$. Pour tout $A \in \mathcal{X}$, on note

$$\tau_A^0 := \inf\{n \geq 0 : X_n \in A\} \in \mathbb{N} \cup \{+\infty\} \text{ et } \tau_A = \tau_A^{(1)} := \inf\{n > 0 : X_n \in A\} \in \mathbb{N} \cup \{+\infty\}$$

les **temps de premier retour** et **temps de première entrée** de la chaîne dans l'ensemble A. Lorsque $A = \{y\}$, nous simplifions la notation en τ_y^0 au lieu de $\tau_{\{y\}}^0$ (et de même pour $\tau_{\{y\}}$).

Lemme 5.2.5. *Les temps τ_A^0 et τ_A définis dans 5.2.4 sont des temps d'arrêt par rapport à la (filtration naturelle de) la chaîne.*

1. En termes plus précis, si on note $\mathcal{F}_n = \otimes_{k=0}^n \mathcal{X}$, et si $\{T = n\} \in \mathcal{F}_n$ pour tout n, on dit que T est un $(\mathcal{F}_n)$-temps d'arrêt. La suite des tribus $(\mathcal{F}_n)$ est une suite croissante, appelée **filtration naturelle** de la chaîne. Pour tout $n \in \mathbb{N}$, la tribu $\mathcal{F}_n$ est la plus petite tribu qui rend les variables $(X_0, \dots, X_n)$ (simultanément) mesurables.

Supposons que $\Omega = \sqcup_{n\in\mathbb{N}}\{T = n\}$ (et par conséquent $\mathbb{P}_{\rho_0}(T < \infty) = 1$). Par ailleurs, la chaîne arrêtée à l'instant aléatoire T vérifiera :

$$X_T(\omega) := X_{T(\omega)}(\omega) = \sum_{n\in\mathbb{N}} X_{T(\omega)}(\omega)\mathbb{1}_{\{T=n\}}(\omega) = \sum_{n\in\mathbb{N}} X_n(\omega)\mathbb{1}_{\{T=n\}}(\omega).$$

Cette remarque, nous amène à définir $\mathcal{F}_T$ comme la tribu trace

$$\mathcal{F}_T := \{F \in \mathcal{F} : \forall n \in \mathbb{N}, F \cap \{T = n\} \in \mathcal{F}_n\}.$$

Elle contient les événements qui lorsque on les restreint sur l'événement $\{T = n\}$ ne dépendent que de $(X_0, X_1, \ldots, X_n)$.

Théorème 5.2.6. *Soient $(X_n)_{n\in\mathbb{N}}$ une CM($\mathbb{X}, \mathcal{X}, \rho$), où $\mathbb{X}$ est dénombrable (fini ou infini) et T un temps d'arrêt pour (la filtration naturelle de) la chaîne. On note $X_T^{\leqslant} = (X_0, \ldots, X_T)$ le passé de la chaîne et $X_T^{>} = (X_{T+1}, X_{T+2}, \ldots)$ son futur strict (par rapport à l'instant aléatoire T). Alors, sur l'événement $\{T < \infty\}$, la loi conditionnelle conjointe du futur strict de la chaîne $X_T^{>}$ sachant le passé $X_T^{\leqslant}$ est égale à la loi conditionnelle conjointe du futur strict de la $X_T^{>}$ sachant le présent X_T ; cette propriété est appelée **propriété forte de Markov**.*

Démonstration. Utiliser la décomposition $\Omega = \{T < \infty\} = \sqcup_{n\in\mathbb{N}}\{T = n\}$, fixer $k \geq 1$ et calculer la probabilité conditionnelle

$$\mathbb{P}(X_{T+1} = y_1, \ldots X_{T+k} = y_k | T = n; X_0 = x_0, \ldots, X_n = x_n).$$

L'événement $\{T = n\}$ est déterminé par les valeurs que prennent les variables aléatoires $(X_0, \ldots, X_n)$. Lorsque nous considérons les événements $\{T = n\}$ et $\{X_0 = x_0, \ldots, X_n = x_n\}$ des deux choses l'une :

— soit ils sont compatibles et dans ce cas $\{T = n\} \cap \{X_0 = x_0, \ldots, X_n = x_n\} = \{X_0 = x_0, \ldots, X_n = x_n\}$,

— soit ils sont incompatibles et dans ce cas $\{T = n\} \cap \{X_0 = x_0, \ldots, X_n = x_n\} = \emptyset$,

Dans le premier cas, en utilisant la propriété faible de Markov, on obtient sur $\{T = n\}$,

$$\mathbb{P}(X_{n+1} = y_1, \ldots X_{n+k} = y_k | X_0 = x_0, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = y_1, \ldots X_{n+k} = y_k | X_n = x_n).$$

Dans le deuxième cas, l'ensemble par rapport auquel nous conditionnons a une probabilité nulle ; par conséquent, nous pouvons attribuer une valeur arbitraire à la probabilité conditionnelle, par exemple la valeur prise par la probabilité conditionnelle au premiers cas, à savoir $\mathbb{P}(X_{n+1} = y_1, \ldots X_{n+k} = y_k | X_n = x_n)$. □

5.3 Classification des états ; récurrence, transience

Lorsque l'espace des états $\mathbb{X}$ est fini et P a tous ses éléments de matrice strictement positifs, il est facile de voir que si la chaîne de Markov évolue sans arrêt alors elle visite tous les éléments de $\mathbb{X}$ une infinité de fois ; on dit que tous les états sont **récurrents**. Les choses sont plus compliquées lorsque $\mathbb{X}$ est dénombrable infini ou si P n'a pas tous ses éléments strictement positifs.

Définition 5.3.1. Soit (X_n) une chaîne de Markov CM($\mathbb{X}, P, \rho_0$).

1. Le graphe dirigé ayant comme sommets l'ensemble $\mathbb{X}$ et comme arêtes les couples $(x, y) \in \mathbb{X}^2$ tels que $P_{x,y} > 0$ est appelé **graphe (des transitions)** de la chaîne de Markov.

2. Un état y est dit **accessible** depuis l'état x, s'il existe un entier $n := n(x,y) \geq 1$ tel que $P^n_{x,y} > 0$. On note $x \to y$.
3. Un état x communique avec un état y si $x \to y$ et $y \to x$.
4. Un état x est dit **essentiel** si pour tout état y tel que $x \to y$ alors $y \to x$. L'ensemble des états essentiels est noté $\mathbb{X}_e$. Leur complémentaire constitue l'ensemble des états inessentiels $\mathbb{X}_i = \mathbb{X} \setminus \mathbb{X}_e$.

Remarque 5.3.2. Il est immédiat de constater que $x \to y$ si, et seulement si, il existe une suite finie d'arêtes dirigées composables dans le graphe de P qui forment un chemin de x à y.

Proposition 5.3.3. *Soit (X_n) une chaîne de Markov $CM(\mathbb{X}, P, \rho_0)$.*

1. *La relation de communication restreinte à $\mathbb{X}_e$ est une relation d'équivalence. On note K l'ensemble de classes d'équivalence $\mathbb{X}_e/\leftrightarrow$.*
2. *La **classification des états** est la partition de l'espace en*

$$\mathbb{X} = \left(\sqcup_{[x]\in K}[x]\right) \sqcup \mathbb{X}_i,$$

où $[x] = \{y \in \mathbb{X}_e : x \leftrightarrow y\}$.

Démonstration. Exercice. □

Définition 5.3.4.
1. Un ensemble d'états $A \subseteq \mathbb{X}$ est dit **absorbant** (ou stochastiquement fermé) si $A \neq \emptyset$ et pour $x \in A \Rightarrow \sum_{y\in A} P_{x,y} = 1$.
2. Si pour un $x \in \mathbb{X}$, on $[x] = \mathbb{X}$, la chaîne est dite **irréductible**, i.e. $\forall (x,y) \in \mathbb{X}^2, \exists n := n(x,y) \geq 1 : P^n_{x,y} > 0$.
3. La chaîne est dite **fortement irréductible** si $\exists N > 0, \exists \alpha > 0 : \min_{x,y} P^N_{x,y} = \alpha > 0$.

Définition 5.3.5. Soit $(X_n) \in \mathsf{CM}(\mathbb{X}, P, \rho_0)$ définie sur $(\Omega, \mathcal{F}, \mathbb{P})$. Pour $x \in \mathbb{X}$ fixé et tout $n \geq 1$, on note $q_n = \mathbb{P}_x(\tau_x = n)$ et $q = \sum_{n\geq 1} q_n = \mathbb{P}_x(\tau_x < \infty)$. On dit qu'un état $x \in \mathbb{X}$ est

— **transient** si $q < 1$,
— **récurrent** si $q = 1$,
— **récurrent positif** si $\mathbb{E}_x(\tau_x) < \infty$.

Un état récurrent qui n'est pas récurrent positif est appelé **récurrent nul**.

Il s'avère pratique d'étendre la suite $(q_n)_{n\geq 1}$ en $q_0 = 0$ et de définir une autre suite $(r_n)_{n\geq 0}$ par $r_n = \mathbb{P}_x(X_n = x) = P^n(x,x), n \in \mathbb{N}$, avec bien sûr $r_0 = 1$.

Lemme 5.3.6. *Pour une chaîne de Markov et les suites (r_n) et (q_n) définies comme ci-dessus, on a :*

1. *Pour tout $n \geq 0$, la probabilité r_n se décompose en*

$$r_n = r_0 q_n + r_1 q_{n-1} + \ldots + r_n q_0.$$

2. *Les séries génératrices*

$$R(z) = \sum_{n\in\mathbb{N}} r_n z^n \text{ et } Q(z) = \sum_{n\in\mathbb{N}} q_n z^n$$

sont bien définies pour $z \in \mathbb{C}$ avec $|z| < 1$.

3. *Pour* $|z| < 1$, *on a* $R(z) = \frac{1}{1-Q(z)}$.

Démonstration. 1. On décompose

$$r_n = \mathbb{P}_x(X_n = x) = \sum_{k=1}^{n} \mathbb{P}_x(X_n = x | \tau_x = k)\mathbb{P}_x(\tau_x = k)$$
$$= \sum_{k=0}^{n} r_{n-k} q_k = r_0 q_n + r_1 q_{n-1} + \ldots + r_n q_0, \text{ avec la convention } q_0 \equiv 1.$$

2. Les séries définissant $R(z)$ et $Q(z)$ sont absolument convergentes pour $|z| < 1$. En effet

$$|R(z)| \leq \sum_{n\geq 0} r_n |z|^n \leq \sum_{n\geq 0} |z|^n = \frac{1}{1-|z|}, \text{ et}$$
$$|Q(z)| \leq \sum_{n\geq 0} q_n |z|^n \leq \sum_{n\geq 0} q_n = \mathbb{P}_x(\tau_x < \infty) \leq 1.$$

3. Pour $|z| < 1$:

$$R(z) = r_0 + \sum_{n\geq 1} r_n z^n = r_0 + \sum_{n\geq 1} \sum_{k=0}^{n} r_{n-k} z^{n-k} q_k z^k$$
$$= r_0 + \sum_{n\geq 0} \sum_{k=0}^{n} r_{n-k} z^{n-k} q_k z^k = r_0 + R(z)Q(z).$$

Puisque $r_0 = 1$, on conclut que $R(z) = \frac{1}{1-Q(z)}$.

□

Lemme 5.3.7. *Pour tout état* $x \in \mathbb{X}$,

$$\mathbb{P}_x(\tau_x < \infty) = \sum_{n\geq 1} q_n = Q(1).$$

Démonstration. Immédiate. □

Théorème 5.3.8. *Un état* $x \in \mathbb{X}$ *est récurrent si, et seulement si,* $\sum_{n\geq 0} \mathbb{P}_x(X_n = x) = \sum_{n\geq 0} r_n = \infty$.

Démonstration. L'état x est récurrent si $q := \sum_{n\geq 1} q_n = Q(1) = \lim_{z\in[0,1[, z\to 1} Q(z) = 1$. En utilisant l'expression pour $R(z)$, établie en lemme 5.3.6, nous obtenons

$$\lim_{z\in[0,1[, z\to 1} R(z) = \lim_{z\in[0,1[, z\to 1} \frac{1}{1-Q(z)} = \infty.$$

Supposons que $\sum_{n\geq 0} r_n < \infty$. Comme tous les termes r_n sont positifs, nous avons pour tout N, $\sum_{n=1}^{N} r_n \leq \lim_{z\in[0,1[, z\to 1} R(z) = \sum_{n\geq 0} r_n$. Or $\lim_{z\in[0,1[, z\to 1} R(z) = \infty$, il y a donc contradiction avec l'hypothèse $\sum_{n\geq 0} r_n < \infty$. □

Remarque 5.3.9. La condition $\sum_{n\geq 0} r_n = \infty$ est équivalente à $\mathbb{E}_x(\eta_x) = \infty$, où la variable aléatoire $\eta_A := \sum_{n\geq 0} \mathbb{1}_A(X_n)$ dénombre les retours de la chaîne de Markov dans A. Comme d'habitude, si $A = \{y\}$, on allège la notation en η_y au lieu de $\eta_{\{y\}}$.

Théorème 5.3.10. *Si l'état $x \in \mathbb{X}$ est récurrent alors tout état $y \in \mathbb{X}$, accessible depuis x, est aussi récurrent.*

Démonstration. Supposons x récurrent, $x \to y$ mais $y \not\to x$. L'accessibilité $x \to y$ signifie qu'il existe un entier M tel que $P^M(x,y) = \alpha > 0$; la non-accessibilité de x à partir de y signifie donc que $\mathbb{P}_x(\tau_x < \infty) \leq 1 - \alpha < 1$ en contradiction avec l'hypothèse de récurrence de x. Il existe donc un entier N tel que $P^N(y,x) = \beta > 0$. Nous avons donc pour tout entier $n \geq 0$

$$P^{M+N+n}(y,y) \geq P^N(y,x)P^n(x,x)P^M(x,y) = \beta\alpha P^n(x,x)$$

et

$$P^{M+N+n}(x,x) \geq P^M(x,y)P^n(y,y)P^N(y,x) = \alpha\beta P^n(y,y).$$

Par conséquent, les séries $\sum_{n\geq 0} P^n(x,x)$ et $\sum_{n\geq 0} P^n(y,y)$ soit elles convergent simultanément soit elles divergent simultanément. On conclut par le théorème 5.3.8. □

Corollaire 5.3.11. *Un chaîne de Markov irréductible sur un ensemble d'états $\mathbb{X}$ fini est nécessairement récurrente.*

5.4 Probabilité limite, probabilité invariante (stationnaire)

Nous avons vu qu'une chaîne de Markov dépend du passé lointain uniquement à travers sa dépendance du passé immédiat. Il est donc intuitivement clair qu'elle « oublie » la condition initiale pour adopter un mode d'évolution dicté par la seule matrice stochastique. Cette intuition est précisée par le théorème suivant.

Théorème 5.4.1 (Théorème ergodique pour une chaîne de Markov)**.** *Soit une chaîne de Markov $(X_n) \in CM(\mathbb{X}, P, \varepsilon_x)$ fortement irréductible sur un ensemble d'états $\mathbb{X}$ fini. Alors, il existe une unique mesure de probabilité μ sur $\mathbb{X}$ et $\lim_{n\to\infty} \mathbb{P}_x(X_n = y) = \mu(y)$ pour tout $y \in \mathbb{X}$. Par ailleurs, il existe de constantes $C, D > 0$ telles que $\max_{x\in\mathbb{X}} |P^n(x,y) - \mu(y)| \leq C\exp(-Dn)$ asymptotiquement (pour n suffisamment grand).*

Démonstration. L'irréductibilité forte de la chaîne sur l'ensemble *fini* $\mathbb{X}$ signifie qu'il existe un entier $N > 0$ et un réel $\alpha > 0$ tels que $\min_{x,y\in\mathbb{X}} P^N(x,y) = \alpha > 0$. Introduire, pour tout $y \in \mathbb{X}$ et tout entier $n \geq 1$, la notation

$$\begin{aligned} m_n(y) &= \min_{x\in\mathbb{X}} P^n(x,y) \\ M_n(y) &= \max_{x\in\mathbb{X}} P^n(x,y) \end{aligned}$$

On a alors

$$m_{n+1}(y) = \min_{x\in\mathbb{X}} P^{n+1}(x,y) = \min_{x\in\mathbb{X}} \sum_{z\in\mathbb{X}} P(x,z)P^n(z,y) \geq m_n(y)$$

et

$$M_{n+1}(y) = \max_{x\in\mathbb{X}} P^{n+1}(x,y) = \max_{x\in\mathbb{X}} \sum_{z\in\mathbb{X}} P(x,z)P^n(z,y) \leq M_n(y).$$

Par conséquent, les suites $(m_n(y))$ et $(M_n(y))$ sont adjacentes

$$m_1(y) \leq \ldots \leq m_n(y) \leq m_{n+1}(y) \leq \ldots \leq M_{n+1}(y) \leq M_n(y) \leq \ldots \leq M_1(y).$$

Par ailleurs, pour $x, x' \in \mathbb{X}$ arbitraires, nous avons

$$\sum_{y\in\mathbb{X}} P^N(x,y) = \sum_{y\in\mathbb{X}} P^N(x',y) = 1.$$

Il s'ensuit la cascade d'égalités

$$\begin{aligned} 0 &= \sum_{y\in\mathbb{X}} P^N(x,y) - \sum_{y\in\mathbb{X}} P^N(x',y) \\ &= \sum_{y\in\mathbb{X}}^{+} [P^N(x,y) - P^N(x',y)] + \sum_{y\in\mathbb{X}}^{-} [P^N(x,y) - P^N(x',y)] \end{aligned}$$

où $\overset{+}{\sum}$ signifie la somme sur ceux des $y \in \mathbb{X}$ pour lesquels $[P^N(x,y) - P^N(x',y)] \geq 0$ et pareillement pour $\overset{-}{\sum}$. La condition d'irréductibilité renforcée implique qu'il existe un réel $0 \leq d < 1$ tel que $\max_{x,x'} \sum_{y\in\mathbb{X}}^{+} [P^N(x,y) - P^N(x',y)] = d$. On estime maintenant la différence

$$\begin{aligned} M_N(y) - m_N(y) &= \max_{x\in\mathbb{X}} P^N(x,y) - \min_{x\in\mathbb{X}} P^N(x,y) = \max_{x,x'\in\mathbb{X}} [P^N(x,y) - P^N(x',y)] \\ &\leq \max_{x,x'\in\mathbb{X}} \sum_{z\in\mathbb{X}}^{+} [P^N(x,z) - P^N(x',z)] = d. \end{aligned}$$

On peut maintenant itérer cette différence

$$\begin{aligned} M_{N+n}(y) - m_{N+n}(y) &= \max_{x,x'\in\mathbb{X}} \sum_{z\in\mathbb{X}} [P^N(x,z) - P^N(x',z)] P^n(z,y) \\ &\leq \max_{x,x'\in\mathbb{X}} \left[\sum_{z\in\mathbb{X}}^{+} [P^N(x,z) - P^N(x',z)] M_n(y) + \sum_{z\in\mathbb{X}}^{-} [P^N(x,z) - P^N(x',z)] m_n(y) \right] \\ &= \max_{x,x'\in\mathbb{X}} \left[\sum_{z\in\mathbb{X}}^{+} [P^N(x,z) - P^N(x',z)] [M_n(y) - m_n(y)] \right] \\ &= d[M_n(y) - m_n(y)]. \end{aligned}$$

Par conséquent, $M_{kN+n}(y) - m_{kN+n}(y) \leq d^{k+1}$, pour $k \in \mathbb{N}$ arbitraire; en combinant avec la propriété d'adjacence pour les suites $(m_n(y))$ et $(M_n(y))$, nous concluons qu'elles convergent vers la même limite $\mu(y) := \lim_{n\to\infty} m_n(y) = \lim_{n\to\infty} M_n(y)$, car $d < 1$. Par ailleurs, $\max_{x\in\mathbb{X}} |P^n(x,y) - \mu(y)| \leq M_n(y) - \mu(y) \leq d^{[n/N]-1}$. L'inégalité est alors prouvée pour avec $C = \frac{1}{d}$ et $D = -\frac{\ln d}{N}$. □

Définition 5.4.2. La probabilité $\mu(y) = \lim_{n\to\infty} \mathbb{P}_x(X_n = y) = \lim_{n\to\infty} P^n(x,y)$ est appelée **probabilité limite**. Le vecteur propre gauche, π, de P associé à la valeur propre 1 est appelé **probabilité invariante (ou stationnaire)**.

Corollaire 5.4.3. *Sous les conditions du théorème 5.4.1, le vecteur (ligne) de probabilité limite μ est égale à un vecteur propre gauche, associé à la valeur propre 1, de la matrice P.*

Démonstration. De l'égalité $P^{n+1}(x,y) = \sum_{z\in\mathbb{X}} P^n(x,z)P(z,y)$, en passant à la limite $n \to \infty$, nous obtenons $\mu(y) = \sum_{z\in\mathbb{X}} \mu(z)P(z,y)$. □

On dispose d'une formulation (voir le théorème 5.4.5 ci-dessous) plus complète du théorème de convergence qui permet de s'affranchir de la l'hypothèse d'irréductibilité forte.

Définition 5.4.4. Soit $x \in \mathbb{X}$ un état vérifiant $\mathbb{P}_x(\tau_x < \infty) > 0$. On appelle **période** de x la quantité

$$d_x = \mathsf{pgcd}\{n \geq 1 : \mathbb{P}_x(X_n = x) > 0\}.$$

Lorsque $d_x = 1$, l'état est appelé apériodique.

On peut facilement montrer que la période de x est une propriété de classe, i.e. tous les états de la classe $[x]$ ont la même période.

En tant que matrice, P agit sur l'espace vectoriel $\mathbb{C}^{\mathbb{X}}$. Le **spectre** de la matrice P est défini comme l'ensemble

$$\mathsf{spec}(P) = \{\lambda \in \mathbb{C} : P - \lambda I \text{ est non inversible}\}.$$

Puisque la matrice P est stochastique, toutes les valeurs spectrales ont un module $|\lambda| \leq 1$; en outre on a toujours $1 \in \mathsf{spec}(P)$. L'ensemble $\mathsf{spec}(P) \setminus \{1\}$ est appelé **spectre périphérique** de P tandis que l'ensemble $\{\lambda \in \mathsf{spec}(P) : |\lambda| < 1\}$ est le **spectre contractant**. Notons que les spectres périphérique ou contractant peuvent être vides.

Soit P une matrice irréductible et d la période d'un état (donc de tous les états) de la chaîne. La partie du spectre non-contractant (i.e. le spectre périphérique et le singleton $\{1\}$) est égal aux racines d-èmes de l'unité. Par conséquent, le spectre périphérique est vide si, et seulement si, la chaîne est apériodique.

Théorème 5.4.5. *Soit P la matrice de transitions d'une chaîne sur un espace d'états $\mathbb{X}$ fini. On note E_λ le projecteur sur l'espace $D^\lambda(P) = \cup_{k\geq 1} \mathsf{ker}(P - \lambda I)^k$.*

1. *Il existe une constante $K_1 > 0$ telle que pour grand n,*

$$\|\frac{1}{n}\sum_{k=0}^{n-1} P^k - E_1\| \leq \frac{K_1}{n}.$$

   En outre pour toute probabilité initiale ρ (vue comme vecteur ligne), on a $\|\frac{1}{n}\sum_{k=0}^{n-1} \rho P^k - \rho E_1\| \leq \frac{K_1}{n}$.
2. *Si le spectre périphérique est vide (i.e. la chaîne est apériodique), il existe une constante $K_2 > 0$ et un paramètre $r \in]0,1[$ tels que*

$$\|P^n - E_1\| \leq K_2 r^n.$$

3. *Si la valeur propre 1 est simple (i.e. $\dim D^\lambda(P) = 1$), le projecteur E_1 définit une unique probabilité invariante π par $E_1 f = \pi(f)\mathbb{1}$, pour tout $f : \mathbb{X} \to \mathbb{C}$ (mesurable) bornée.*

La démonstration de ce résultat est simple mais longue car elle nécessite une longue série de lemmes préliminaires ; elle peut être consultée dans [56].

Un autre résultat important est le

Théorème 5.4.6. *Soit P la matrice stochastique irréductible d'une chaîne de Markov sur un espace d'état dénombrable. On a équivalence entre les affirmations suivantes :*

1. *tout état est récurrent positif,*
2. *il existe un état récurrent positif,*
3. *la matrice P possède une probabilité invariante π.*

En outre (si 3. est vérifié) alors $\mathbb{E}_x(\tau_x) = \frac{1}{\pi(x)}$.

Démonstration. Voir [53, Théorème 1.7.7, p. 37]. □

5.5 Stationnarité, réversibilité

Définition 5.5.1. Soient $\mathbb{X}$ un ensemble dénombrable (muni de sa tribu exhaustive) et $(X_n)_{n\in\mathbb{Z}}$ (i.e. le temps varie sur $\mathbb{Z}$ au lieu de $\mathbb{N}$) une famille doublement infinie de variables aléatoires à valeurs dans $\mathbb{X}$. Si pour tout $n \in \mathbb{N}$ et tout n-uplet d'instants $t_1 < t_2 < \cdots < t_n$, avec $t_i \in \mathbb{Z}$ pour $i = 1, \ldots, n$, la loi conjointe des variables aléatoires $(X_{t_1}, \ldots, X_{t_n})$ est connue, la famille $(X_n)_{n\in\mathbb{N}}$ est dite **processus stochastique** (à temps discret). Le processus est

— **stationnaire** si la loi conjointe de $(X_{T+t_1}, \ldots, X_{T+t_n})$ est la même pour tous les $T \in \mathbb{Z}$;

— **réversible** si la loi conjointe de $(X_{t_1}, \ldots, X_{t_n})$ est la même que la loi de $(X_{T-t_1}, \ldots, X_{T-t_n})$, pour tout $T \in \mathbb{Z}$;

— **markovien** si

$$\mathbb{P}(X_{t_n} = x_n | X_{t_1} = x_1, \ldots, X_{t_{n-1}} = x_{n-1}) = \mathbb{P}(X_{t_n} = x_n | X_{t_{n-1}} = x_{n-1}).$$

Si la probabilité conditionnelle $\mathbb{P}(X_{n+1} = y | X_n = x)$ ne dépend pas de n, la chaîne est dite **homogène** et cette probabilité conditionnelle définit une matrice stochastique $(P(x,y))_{x,y\in\mathbb{X}}$ par $P(x,y) := \mathbb{P}(X_{n+1} = y | X_n = x)$.

Remarque 5.5.2. Si la loi conjointe du processus, échantillonnée à un nombre arbitraire d'instants,

— est invariante aux translations temporelles, le processus est stationnaire;

— est invariante aux renversements du temps, le processus est réversible.

L'invariance ci-dessus signifie qu'il est **statistiquement** impossible de distinguer entre la loi initiale et la loi transformée (par translation ou par renversement).

Lemme 5.5.3. *Un processus réversible est stationnaire.*

Démonstration. Par réversibilité, les lois de $(X_{t_1}, \ldots, X_{t_n})$ et de $(X_{T+t_1}, \ldots, X_{T+t_n})$ sont les mêmes que la loi de $(X_{-t_1}, \ldots, X_{-t_n})$, car la loi de $(X_{T+t_1}, \ldots, X_{T+t_n})$ est égale à la loi de $(X_{S-T-t_1}, \ldots, X_{S-T-t_n})$, pour tout $S \in \mathbb{Z}$. Il suffit donc de choisir $S = T$. □

Théorème 5.5.4. *Soit (X_n) une $\mathsf{MC}(\mathbb{X}, P, \pi)$ irréductible et ayant π comme probabilité invariante. Pour un $N \in \mathbb{N}$, on pose $Y_n = X_{N-n}$ pour $n = 0, \ldots, N$. Alors (Y_n) est une chaîne de Markov (à horizon fini). Si la matrice Q est définie par*

$$\pi(y)Q_{yx} = \pi(x)P_{xy},$$

alors Q est stochastique et irréductible et admet π comme probabilité invariante (i.e. $\pi Q = \pi$). En outre, elle est la matrice des transitions de la chaîne (Y_n).

Démonstration. — La matrice Q est stochastique car

$$\sum_{x\in\mathbb{X}} Q_{yx} = \frac{1}{\pi(y)} \sum_{x\in\mathbb{X}} \pi(x)P_{xy} = \frac{\pi(y)}{\pi(y)} = 1.$$

— La probabilité π est invariante pour Q car

$$\sum_{y\in\mathbb{Y}} \pi(y)Q_{yx} = \sum_{y\in\mathbb{X}} \pi(x)P_{xy} = \pi(x).$$

— De la définition de la chaîne $(Y_n)_{n=0,\dots,N}$ découlent les égalités :

$$\begin{aligned}\mathbb{P}(Y_0 = y_0, \dots, Y_N = y_N) &= \mathbb{P}(X_N = y_0, \dots, X_0 = y_N)\\ &= \pi(y_N) P_{y_N y_{N-1}} \cdots P_{y_1 y_0}\\ &= \pi(y_0) Q_{y_0 y_1} \cdots Q_{y_{N-1} y_N}.\end{aligned}$$

L'irréductibilité de P signifie que pour tout $x, y \in \mathbb{X}$, il existe $L \in \mathbb{N}$ et une suite d'états $y_L \equiv x, y_{L-1}, \dots, y_0 \equiv y$ telle que

$$P_{y_L y_{L-1}} \cdots P_{y_1 y_0} > 0.$$

Pour $N \geq L$, il s'ensuit que

$$Q_{y_0 y_1} \cdots Q_{y_{L-1} y_L} = \frac{\pi(y_L)}{\pi(y_0)} P_{y_L y_{L-1}} \cdots P_{y_1 y_0} > 0,$$

ce qui garantit l'irréductibilité de Q.

□

Théorème 5.5.5. *Une chaîne de Markov stationnaire et irréductible est réversible si, et seulement si, la chaîne vérifie la condition de* ***bilan détaillé****, i.e. s'il existe un vecteur de probabilité* $\pi := (\pi(x))_{x \in \mathbb{X}}$ *tel que*

$$\pi(x) P(x, y) = \pi(y) P(y, x), \forall x, y \in \mathbb{X}.$$

Si un tel vecteur de probabilité existe, alors π *est la probabilité invariante de la chaîne.*

Démonstration. Supposons le processus réversible. Puisqu'il est stationnaire, sa marginale π_t à l'instant t est constante en t, i.e. $\pi(x) = \mathbb{P}(X_t = x)$. La réversibilité impose alors : $\mathbb{P}(X_t = x, X_{t+1} = y) = \mathbb{P}(X_t = y, X_{t+1} = x)$, ou

$$\pi(x) P(x, y) = \pi(y) P(y, x), \forall x, y \in \mathbb{X}.$$

Réciproquement, supposons qu'un vecteur de probabilité π, vérifiant la condition de bilan détaillé existe. Il est alors de toute évidence une probabilité invariante pour le processus. Nous aurons alors

$$\begin{aligned}\mathbb{P}(X_t = x_0, X_{t+1} = x_1, \dots, X_{t+n} = x_n) &= \pi(x_0) P(x_0, x_1) \cdots P(x_{n-1}, x_n),\\ \mathbb{P}(X_t = x_n, X_{t+1} = x_{n-1}, \dots, X_{t+n} = x_0) &= \pi(x_n) P(x_n, x_{n-1}) \cdots P(x_1, x_0).\end{aligned}$$

Mais la condition de bilan détaillé entraîne que les seconds membres de ces égalités sont identiques. On conclut alors la réversibilité du processus. □

Exemple 5.5.6 (Modèle d'urne de Paul et Tatiana Ehrenfest [22])**.** Supposons que N boules discernables (par exemple étiquetées) sont distribuées dans deux urnes notées respectivement urne 0 et urne 1. À chaque instant entier, une étiquette est choisie au hasard et la boule correspondante est transférée dans l'autre urne avec probabilité $p \in]0, 1[$ ou laissée sur place avec probabilité $1 - p$. On note X_n le nombre de boules dans l'urne 0 à l'instant n. La suite $(X_n)_{n \in \mathbb{Z}}$ constitue une chaîne de Markov sur $\mathbb{X} = \{0, \dots, N\}$ qui évolue selon la matrice stochastique

$$P(x, y) = \begin{cases} p\frac{x}{N} & \text{si } x \in \{1, \dots, N\} \text{ et } y = x - 1,\\ 1 - p & \text{si } x \in \mathbb{X} \text{ et } y = x,\\ p(1 - \frac{x}{N}) & \text{si } x \in \{0, \dots, N-1\} \text{ et } y = x + 1,\\ 0 & \text{sinon.}\end{cases}$$

Cette chaîne est fortement irréductible. Par le théorème ergodique pour les chaînes de Markov , on conclut que la probabilité limite $\pi(y) = \lim_{n\to\infty} \mathbb{P}_x(X_n = y) = \lim_{n\to\infty} P^n(x,y)$ existe et est indépendante de x; elle est l'unique mesure invariante de la chaîne. Par ailleurs, on montre facilement que la probabilité $\mu(x) = \frac{C_N^x}{2^N}, x \in \mathbb{X}$, vérifie la condition de bilan détaillé; la chaîne est donc réversible et $\pi = \mu$. Puisque μ est la probabilité d'équilibre, après un temps suffisamment long, la probabilité pour qu'une boule soit dans l'une ou l'autre urne est $1/2$. D'un autre côté, la réversibilité de la chaîne signifie que (pour T grand)

$$\mathbb{P}_\mu(X_0 = N, X_T = N/2) = \mathbb{P}_\mu(X_0 = N/2, X_T = N),$$

résultat qui peut paraître contre-intuitif de prime abord. (Nous retournerons sur ce résultat au paragraphe 7.6).

5.6 Théorème des grands nombres pour les chaînes de Markov

Nous avons établi le théorème de grands nombres pour des suites indépendantes. Ici on généralise le résultat au cas de suites stationnaires.

Soit $(X_n)_{n\in\mathbb{N}}$ une $\mathsf{CM}(\mathbb{X}, P, \varepsilon_x)$ une chaîne de Markov à valeurs dans un ensemble $\mathbb{X}$ dénombrable (muni de sa tribu exhaustive) irréductible et récurrente. On définit $\tau_x^{(0)} = 0$ et par récurrence la suite des temps successifs de retour en x :

$$\tau_x^{(r)} = \inf\{n > \tau_x^{(r-1)} : X_n = x\}, \text{ pour } r \geq 1.$$

La chaîne étant récurrente, elle revient indéfiniment en x, donc $r \to \infty$. Par ailleurs, pour tout $r \in \mathbb{N}$, $\tau_x^{(r)}$ est un temps d'arrêt pour la chaîne. Pour alléger la notation dans la suite, nous fixons x et nous écrivons simplement τ^r au lieu de $\tau_x^{(r)}$.

Soit $f : \mathbb{X} \to \mathbb{R}$ une fonction réelle. On s'intéressera à la somme partielle $S_n := S_n^f = \sum_{k=0}^n f(X_k)$; plus précisément, on veut examiner sour quelles hypothèses la $\lim_{n\to\infty} \frac{S_n}{n}$ existe et déterminer sa valeur.

Lemme 5.6.1. *Avec les notations et hypothèses précédentes, on introduit la suite $(Z_r)_{r\geq 1}$ par*

$$Z_r := Z_r^f = \sum_{k=\tau^r+1}^{\tau^{r+1}} f(X_k), r \geq 1.$$

Alors cette suite est indépendante et identiquement distribuée.

Démonstration. La propriété forte de Markov garantit que la loi conjointe du futur strict à partir de τ^r conditionnellement au passé jusqu'au temps τ^r est égale à la loi conjointe du futur strict de τ^r conditionnellement au présent. Or à l'instant τ^r, la chaîne vaut $X_{\tau^r} = x$. Par conséquent : la loi conjointe de $(X_{\tau^r+1}, X_{\tau^r+2}, \ldots)$ conditionnellement au passé est égale à la loi conjointe de $(X_1, X_2, \ldots)$ conditionnellement à $X_0 = x$. On a alors, pour tout $r \geq 1$,

$$\mathbb{P}(Z_r \in B | \text{passé}) = \mathbb{P}_x(Z_1 \in B), \text{ pour tout } B \in \mathcal{B}(\mathbb{R}).$$

□

Théorème 5.6.2. *Soient $(X_n)_{n\in\mathbb{N}}$ une CM$(\mathbb{X}, P, \cdot)$, $x \in \mathbb{X}$ un point récurrent positif et $f : \mathbb{X} \to \mathbb{R}$ telle que*

$$\mathbb{E}_x(|f(X_1)| + \ldots + |f(X_{\tau_x^1}|) < \infty.$$

Alors

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} f(X_k) = \frac{\mathbb{E}_x(f(X_1) + \ldots + f(X_{\tau_x^1})}{\mathbb{E}_x \tau_x^1}, \mathbb{P}_x - p.s.$$

Si la chaîne est irréductible, la convergence est presque sûre indépendamment de la condition initiale.

Démonstration. Pour $x \in \mathbb{X}$ et $n \in \mathbb{N}$, définir le nombre de visites de la chaîne au point x avant le temps n

$$\eta_n(x) = \sum_{k=1}^{n} \mathbb{1}_{\{x\}}(X_k) = \sup\{r \geq 1 : \tau_x^r \leq n\}.$$

Puisque la chaîne est récurrente, $\lim_{n\to\infty} \eta_n(x) = \infty$. Par ailleurs, sur l'événement $\{X_0 = x\}$, on a la décomposition évidente :

$$\frac{1}{n} \sum_{k=0}^{n} f(X_k) = \frac{1}{n} \sum_{k=0}^{\tau_x^1} f(X_k) + \frac{1}{n} \sum_{r=1}^{\eta_n(x)} Z_r - \frac{1}{n} \sum_{k=n+1}^{\tau_x^{\eta_n(x)+1}} f(X_k).$$

La première somme partielle du membre de droite, tends vers 0 lorsque $n \to \infty$ car

$$|\sum_{k=0}^{\tau_x^1} f(X_k)| \leq \sum_{k=0}^{\tau_x^1} |f(X_k)|$$

et la dernière somme est intégrable (par hypothèse), donc presque sûrement finie.

La deuxième somme partielle peut se réécrire

$$\frac{1}{n} \sum_{r=1}^{\eta_n(x)} Z_r = \frac{\eta_n(x)}{n} \frac{1}{\eta_n(x)} \sum_{r=1}^{\eta_n(x)} Z_r.$$

Comme $\lim_n \eta_n(x) = \infty$ et les variables aléatoires Z_r sont i.i.d., on aura

$$\lim_n \frac{1}{\eta_n(x)} \sum_{r=1}^{\eta_n(x)} Z_r = \mathbb{E}_x(Z_1)$$

par le théorème de grands nombres pour des suites indépendantes.

Pour conclure avec le deuxième terme, il reste à montrer que $\lim_n \frac{\eta_n(x)}{n} = \mathbb{E}_x(\tau_x^1)$. Or, la chaîne étant récurrente, on $\mathbb{P}_x(\tau_x^r < \infty) = 1$, pour tout $r \geq 1$. Par conséquent, par la propriété forte de Markov, la famille $(X_{\tau_x^r+n})_{n\in\mathbb{N}}$ est une MC$(\mathbb{X}, P, \varepsilon_x)$ et, en outre, la loi conjointe de $(X_{\tau_x^r+n})_{n\in\mathbb{N}}$ est indépendante de $(X_0, \ldots, X_{\tau_x^r-1})$. On note $\sigma_x^r := \tau_x^{r+1} - \tau_x^r$ le temps qui s'écoule entre deux retours successifs (r et $r+1$) au point x. On a

$$\sigma_x^1 + \ldots + \sigma_x^{\eta_n(x)-1} \leq n - 1 \leq n \leq \sigma_x^1 + \ldots + \sigma_x^{\eta_n(x)}.$$

En divisant ces inégalités par $\eta_n(x)$, on obtient :

$$\frac{\sigma_x^1 + \ldots + \sigma_x^{\eta_n(x)-1}}{\eta_n(x)} \leq \frac{n}{\eta_n(x)} \leq \frac{\sigma_x^1 + \ldots + \sigma_x^{\eta_n(x)}}{\eta_n(x)}.$$

Chacun des termes aux extrémités des inégalités précédentes converge (par la loi forte de grands nombres et la propriété forte de Markov) presque sûrement vers $\mathbb{E}_x(\sigma_x^1) = \mathbb{E}_x(\tau_x^1)$. Par conséquent $\lim_{n\to\infty} \frac{n}{\eta_n(x)} = \mathbb{E}_x(\tau_x^1)$.

Reste à montrer que le terme de correction $R_n = \frac{1}{n} \sum_{k=n+1}^{\tau_x^{\eta_n(x)+1}} f(X_k)$ tendra vers 0 ce qui est garanti par la condition $\mathbb{E}_x(|f(X_1)| + \ldots + |f(X_{\tau_x^1}|) < \infty$. □

5.7 Exemples d'applications algorithmiques

5.7.1 Classification de pages internet; algorithme PageRank

Lorsqu'on lance une requête sur internet, on reçoit une liste de liens vers des documents classée selon la pertinence (supposée) de chaque document. Une question importante est de comprendre comment cette liste est-elle obtenue.

En 1998, il y avait 26×10^6 de pages indexées par Google. En 2017, ce nombre se montait en 30×10^{12} pages. Il est donc naturel que les méthodes utilisées aient largement évolué au cours du temps.

Au début de la toile, les moteurs de recherche maintenait un index de mots-clefs $\boldsymbol{k} = (k_1, \dots, k_N)$. Chaque page était balayée et le nombre d'apparitions de chaque mot-clef était compté. Ainsi la page u était représentée par un vecteur sémantique $\boldsymbol{\nu}(u) = (\nu_1(u), \dots, \nu_N(u))$, où $\nu_i(u)$ est le nombre de fois que le terme k_i apparaît dans la page u. Une requête était à son tour assimilée à un vecteur de requête $\mathbf{r} = (r_1, \dots, r_N)$, où les r_i prennent la valeur 0 ou 1 selon que le terme k_i apparaissait dans la requête. Les liens vers les pages classées par valeur décroissante du produit scalaire $\langle \mathbf{r}, \boldsymbol{\nu}(u) \rangle$ étaient alors retournés.

Cette méthode de classification souffre d'un inconvénient majeur. Elle est susceptible de retourner de liens vers des pages totalement impertinentes qui ne contiennent qu'un très grand nombre de répétitions du même mot-clef.

Serguey Brin et Lawrence Page (dans [11]) ont fait une avancée significative dans l'efficacité d'indexation des pages; leur méthode porte le nom d'algorithme de PageRank. Elle est implémentée dans les moteurs de recherche Google ou Bing.

Soient $\mathbb{X}$ l'ensemble de pages de la toile et $(A_{x,y})_{x,y\in\mathbb{X}}$ la matrice d'adjacence définie par

$$A_{x,y} = \begin{cases} 1 & \text{si la page } x \text{ pointe vers la page } y \\ 0 & \text{sinon.} \end{cases}$$

On construit une matrice stochastique

$$P_{x,y} = \begin{cases} \frac{1}{d_x} & \text{si } \ A_{x,y} = 1, \\ 0 & \text{sinon,} \end{cases}$$

où d_x est le nombre de pages vers lesquelles pointe la page x. Un degré d'autorité $\beta_x = \frac{1}{|\mathbb{X}|}$ ad hoc est rattaché à la page $x \in \mathbb{X}$ et ensuite le système évolue de la manière suivante : si la page x pointe vers la page y, la portion $\beta_x P_{x,y}$ de son autorité est héritée par y. Dans l'état d'équilibre, on aura alors un vecteur de probabilité $\alpha := (\alpha_x)_{x\in\mathbb{X}}$ — l'autorité d'équilibre — qui vérifiera $\alpha = \alpha P$, i.e. il sera la probabilité stationnaire de la chaîne de Markov de matrice stochastique P.

Cette méthode présente encore l'inconvénient que rien ne garantit que P soit fortement irréductible. Brin et Page proposèrent alors de modifier P en une matrice P' dont les éléments de matrice sont $P'_{x,y} = (1-p)P_{x,y} + p\frac{1}{|\mathbb{X}|}$, pour un petit paramètre $p \in]0,1[$ qui est le « secret maison » de Google. La matrice P' est maintenant fortement irréductible et l'on peut appliquer le théorème 5.5.6 pour montrer l'existence et l'unicité de α ainsi que la convergence exponentielle des itérations successives de P' vers la probabilité invariante. Les pages retournées par le moteur de recherche sont alors classées dans l'ordre décroissant des composantes de α.

5.7.2 Algorithme de Metropolis

Sur un espace fini $(\mathbb{X}, \mathcal{X})$, on veut simuler des variables distribuées selon une certaine loi π. Dans les applications $|\mathbb{X}| \simeq 10^6 - 10^7$ et la mesure π traduit souvent en langage probabiliste des contraintes rigides non-triviales sur les x (cf. exercice 69); il est donc assez difficile de simuler directement la loi π par génération selon la loi uniforme et rejet des échantillons qui ne vérifient pas la contrainte. Il est en général plus facile de générer une chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ à valeurs dans $\mathbb{X}$ qui admet π comme probabilité stationnaire, i.e. construire une chaîne qui évolue suivant une matrice stochastique P fortement irréductible telle que $\pi P = \pi$. Étant donné que la matrice P est de taille $|\mathbb{X}| \times |\mathbb{X}| \simeq 10^{12} - 10^{14}$, il est hors de question de stocker une telle matrice et s'en servir pour générer la chaîne de Markov. Il faut concevoir un algorithme qui génère une chaîne de Markov évoluant selon P, telle que $\pi P = \pi$, sans que P soit stockée.

Soit Q une matrice stochastique irréductible et apériodique *arbitraire* sur $\mathbb{X}$.

Définition 5.7.1. Pour $x \in \mathbb{X}$, on note $\mathbb{A}_x = \{y \in \mathbb{X} : Q_{xy} > 0\}$ les états accessibles de x en une étape par la matrice stochastique Q. Pour tout $y \in \mathbb{A}_x$, $\nu_x(y) = Q_{xy}$ définit une probabilité sur $\mathbb{A}_x$ appelée **probabilité de tentative** de la transition $x \to y$.

La matrice Q sert donc à proposer des transitions $x \to y$. On peut se poser la question quel est son intérêt puisque P remplit le même rôle ? La réponse est que tandis $|\mathbb{X}| \simeq 10^6$, ! typiquement $|\mathbb{A}_x| \simeq 10$. Bien sûr si nous remplaçons P par Q nous changeons la probabilité d'équilibre du problème. On doit donc refuser certaines transitions proposées par Q.

Définition 5.7.2. Soit $A = (A_{xy})_{x,y\in\mathbb{X}} \in \mathbb{M}_{|\mathbb{X}|\times|\mathbb{X}|}(]0,1])$. Cette matrice est appelée **matrice d'acceptation**. Une transition $x \to y$, proposée par Q, sera acceptée avec probabilité A_{xy} et rejeteée (on reste alors sur place) avec probabilité $1 - A_{xy}$.

On construit ainsi la matrice P par

$$\begin{aligned} P_{xy} &= \mathbb{P}(X_{n+} = y | X_n = x) \\ &= \begin{cases} Q_{xy}A_{xy} & \text{si } y \in \mathbb{A}_x \setminus \{x\}, \\ Q_{xx} + \sum_{z\in\mathbb{A}_x\setminus\{x\}} (1 - A_{xz})Q_{xz} & \text{si } y = x, \\ 0 & \text{sinon.} \end{cases} \end{aligned}$$

Lemme 5.7.3. *La matrice P est stochastique, irréductible et apériodique.*

Démonstration. La positivité des éléments de P est évidente. Sa stochasticité s'obtient par

$$\begin{aligned} \sum_{y\in\mathbb{X}} P_{xy} &= P_{xx} + \sum_{y\in\mathbb{A}_x\setminus\{x\}} P_{xy} \\ &= Q_{xx} + \sum_{z\in\mathbb{A}_x\setminus\{x\}} (1 - A_{xz})Q_{xz} + \sum_{y\in\mathbb{A}_x\setminus\{x\}} A_{xy}Q_{xy} \\ &= \sum_{z\in\mathbb{A}_x} Q_{xz} = 1. \end{aligned}$$

Son irréductibilité et son apériodicité s'obtiennent par l'observation que pour $x, y \in \mathbb{X}$,

$$Q_{xy} > 0 \Rightarrow P_{xy} > 0.$$

□

Lemme 5.7.4. *Si la matrice d'acceptation A vérifie la condition*

$$\frac{A_{xy}}{A_{yx}} = \frac{\pi(y)Q_{yx}}{\pi(x)Q_{xy}}, \forall x \in \mathbb{X}, \forall y \in \mathbb{A}_x \setminus \{x\}$$

alors la matrice P vérifie la condition de bilan détaillé et réciproquement.

Démonstration. **(⇒)** Pour $y \in \mathbb{A}_x \setminus \{x\}$, on a

$$\begin{aligned} \pi(x)P_{xy} &= \pi(x)Q_{xy}A_{xy} \\ &= \frac{A_{yx}}{A_{xy}}\pi(y)Q_{yx}A_{xy} \\ &= \pi(y)P_{yx}. \end{aligned}$$

(⇐) L'égalité $\pi(x)P_{xy} = \pi(y)P_{yx}$ devient

$$\pi(x)Q_{xy}A_{xy} = \pi(y)Q_{yx}A_{yx}.$$

□

Corollaire 5.7.5. *Soit $F : [0,\infty] \to [0,1]$ une application telle que $\frac{F(z)}{F(1/z)} = z$ pour tout $z \in [0,\infty]$. Si*

$$A_{xy} = F\left(\frac{\pi(y)Q_{yx}}{\pi(x)Q_{xy}}\right), \forall x \in \mathbb{X}, \forall y \in \mathbb{A}_x \setminus \{x\},$$

alors P vérifie la condition de bilan détaillé.

Exemple 5.7.6. Les fonctions données par les formules $F(z) = \frac{z}{1+z}$ et $F(z) = \min(1,z)$ vérifient les hypothèses du lemme.

La construction de la matrice P à l'aide des matrices de tentatives et d'acceptation selon le procédé décrit dans ce paragraphe est connu comme **algorithme de Metropolis**. Sa mise en œuvre pour la génération d'une chaîne de Markov selon P est appelée **simulation Monte Carlo markovienne**.

5.8 Exercices

Chaînes de Markov ; matrices stochastiques

64. Soient $\mathbb{X} = \{0,1,\ldots,m\}$ un ensemble fini d'états et une suite indépendante $(U_n)_{n\geq 1}$, identiquement distribuée selon la loi uniforme sur $\{1,\ldots,m\}$. On introduit la suite $(X_n)_{n\geq 0}$ de variables aléatoires, définies par $X_0 = 0$ et $X_{n+1} = \max(X_n, U_{n+1})$, pour $n \geq 0$.
    (a) Montrer que (X_n) est une chaîne de Markov et déterminer sa matrice de transition P.
    (b) Esquisser le graphe de la matrice stochastique lorsque m est petit.
    (c) En utilisant des arguments d'estimation directe, indiquer quels états sont récurrents et quels sont transients.
    (d) Pour $m = 4$, calculer P^2.

65. Soient $(\xi_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes, identiquement distribuées sur l'ensemble $\{-1,1\}$ selon une loi μ avec $\mu(-1) = p \in [0,1]$. On construit la suite $X'_n = \sum_{k=1}^{n} \xi_k$, pour $n \in \mathbb{N}$ avec la convention que la somme sur une famille vide d'indices est nulle.
    (a) Montrer que (X'_n) est une chaîne de Markov sur l'espace des états $\mathbb{X}'$, avec matrice de transition P et probabilité initiale ρ (on précisera $\mathbb{X}'$, P' et ρ').
    (b) Montrer que la suite (X_n) définie récursivement par $X_0 = x_0 \in \mathbb{Z}$ et $X_{n+1} = X_n + \xi_{n+1}$, pour $n \geq 0$, est aussi une $\mathsf{CM}(\mathbb{X}, P, \rho)$ (préciser $\mathbb{X}$, P et ρ). La suite X_n est appelée marche aléatoire simple en dimension 1.
    (c) Montrer — sous l'hypothèse que $p \in]0,1[$ — que la marche aléatoire simple en dimension 1 est irréductible (argumenter en utilisant le graphe de la matrice de transition).
    (d) Montrer que la marche aléatoire simple en dimension 1 est récurrente si $p = 1/2$, transiente si $p \neq 1/2$.
    (e) Soient $a, b \in \mathbb{Z}$ avec $a < b$. Pour un $x \in [a,b] \cap \mathbb{Z}$, on note $h(x) = \mathbb{P}_x(\tau_a^0 < \tau_b^0)$.
        — Pour $p = 1/2$, montrer que $h(x) = \frac{1}{2}h(x-1) + \frac{1}{2}h(x+1)$ pour $x \in [a+1, b-1] \cap \mathbb{Z}$. Par un argument géométrique simple résoudre cette équation en utilisant les valeurs au bord évidentes pour $h(a)$ et $h(b)$ pour montrer que $h(x) = 1 - \frac{x-a}{b-a}$.
        — Établir la relation analogue pour $h(x)$ lorsque $p \neq 1/2$ et résoudre explicitement la récurrence.
    (f) Interpréter ces résultats pour le problème de la « ruine du joueur ».

66. (Extrait du contrôle du 16 novembre 2017).
Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov $\mathsf{CM}(\mathbb{X}, P, \cdot)$, où $\mathbb{X} = \{1,2,3,\dots,\} = \mathbb{N}_>$ et la matrice stochastique $P = (P_{x,y})_{x,y\in\mathbb{X}}$ est définie par

$$P_{x,y} = \begin{cases} \frac{x}{x+1} & \text{si } x \in \mathbb{X}, y = 1 \\ \frac{1}{x+1} & \text{si } x \in \mathbb{X}, y = x+1 \\ 0 & \text{sinon.} \end{cases}$$

    (a) Soit $\mathbf{p} := (p_x)_{x\geq 1}$ la suite définie par $p_x = \frac{1}{(e-1)x!}$. Montrer que $\mathbf{p}$ est un vecteur de probabilité sur $\mathbb{X}$.
    (b) Montrer que le vecteur de probabilité $\mathbf{p}$ est invariant pour P, i.e. le vecteur ligne $\mathbf{p}$ est vecteur propre gauche de P associé à la valeur propre 1). *Clin d'œil :* $\frac{x}{x+1} = 1 - \frac{1}{x+1}$.
    (c) Déterminer la décomposition de $\mathbb{X}$ en classes de communication.

67. (Extrait de l'examen du 7 mai 2013).
Soit $\beta \in]0,1[$. On considère la matrice

$$P := \begin{pmatrix} 1-\beta & \beta \\ \beta & 1-\beta \end{pmatrix}.$$

    (a) On note $\mathrm{spec}(P)$ l'ensemble des valeurs propres de P. Vérifier que $\mathrm{spec}(P) = \{1, 1-2\beta\}$.
    (b) Vérifier que les vecteurs $\mathbf{D}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ et $\mathbf{D}_{1-2\beta} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ sont des vecteurs propres droits associés aux valeurs propres correspondantes.

(c) Pour chaque $\lambda \in \mathsf{spec}(P)$, on note $\mathbf{d}_\lambda = \frac{\mathbf{D}_\lambda}{\|\mathbf{D}_\lambda\|_{\sup}}$ et $\delta_\lambda = \frac{\mathbf{D}_\lambda}{\|\mathbf{D}_\lambda\|_1}$, où, pour tout $\mathbf{z} \in \mathbb{R}^2$, on note $\|\mathbf{z}\|_1 = |z_1| + |z_2|$ et $\|\mathbf{z}\|_{\sup} = \max\{|z_1|, |z_2|\}$. Déterminer $\mathbf{d}_\lambda$ et δ_λ, pour $\lambda, \lambda' \in \mathsf{spec}(P)$.

(d) Pour $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ et $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ vecteurs arbitraires, on note $\mathbf{a} \otimes \mathbf{b}^t = \begin{pmatrix} a_1 b_1 & a_1 b_2 \\ a_2 b_1 & a_2 b_2 \end{pmatrix}$. Vérifier que $P = \sum_{\lambda \in \mathsf{spec}(P)} \lambda E_\lambda$, où $E_\lambda = \mathbf{d}_\lambda \otimes \delta_\lambda^t$.

(e) Montrer que $E_\lambda E_{\lambda'} = \delta_{\lambda,\lambda'} E_\lambda$.

(f) En déduire que pour un $n \geq 1$ arbitraire,

$$P^n = \begin{pmatrix} p_n & 1-p_n \\ 1-p_n & p_n \end{pmatrix}.$$

(g) On déterminera explicitement p_n. Quelle est la valeur de $\lim_{n\to\infty} p_n$?

(h) On note (X_n) la chaîne de Markov sur un espace $\mathbb{X}$ à deux états, dont la matrice de transition est $\mathbf{P}$ et la probabilité initiale ρ_0. Déterminer $\pi(y) := \lim_{n\to\infty} \mathbb{P}_{\rho_0}(X_n = y)$, pour $y \in \mathbb{X}$.

(i) Montrer que π est une probabilité invariante.

Simulations Monte Carlo, algorithme de Metropolis

68. Comment à partir d'une pièce truquée à mémoire, de matrice stochastique (des tentatives)

$$Q = \begin{pmatrix} \frac{3}{5} & \frac{2}{5} \\ \frac{1}{5} & \frac{4}{5} \end{pmatrix}$$

peut-on construire une matrice stochastique P qui admet $\pi = (\frac{1}{2}, \frac{1}{2})$ comme probabilité d'équilibre ?

69. (Marches aléatoires sans recoupement sur $\mathbb{Z}^d$, devoir du 20 octobre 2019). Une marche aléatoire simple sur $\mathbb{Z}^d$ est une chaîne de Markov (X_n) obtenue par la récurrence $X_{k+1} = X_k + \xi_{k+1}$, $k \in \mathbb{N}$, avec $X_0 = x \in \mathbb{Z}^d$ et $(\xi_k)_{k\geq 1}$ une suite de variables aléatoires indépendantes et uniformément distribuées sur $E_d = \{\pm \mathbf{e}_j, j = 1, \ldots, d\}$, où $\mathbf{e}_j, j = 1, \ldots, d$, sont les vecteurs de la base canonique de $\mathbb{Z}^d$. L'ensemble des trajectoires (de longueur N fixée) possibles de telles marches constitue l'ensemble $\mathbb{X}^{\mathrm{mas}}$

$$\mathbb{X}^{\mathrm{mas}} := \mathbb{X}^{\mathrm{mas}}_{x,N} = \{\mathbf{y} : \{0, \ldots, N\} \to \mathbb{Z}^d : [y_0 = x] \wedge [y_{k+1} - y_k \in E_d]\}.$$

On s'intéresse au sous-ensemble des trajectoires sans recoupement

$$\mathbb{X}^{\mathrm{msr}} := \mathbb{X}^{\mathrm{msr}}_{x,N} = \{\mathbf{y} \in \mathbb{X}^{\mathrm{mas}}_{x,N} : y(k) \neq y(l), 0 \leq k < l \leq N\}$$

muni de la probabilité uniforme. Ces ensembles, pour $N \simeq 10^4$, modélisent bien certains polymères tels que le poly-chloro-éthylène, mais la contrainte géométrique forte de non-recoupement ne nous permet même pas d'estimer le cardinal de $\mathbb{X}^{\mathrm{msr}}_{x,N}$ dès que N prenne de valeurs modérément grandes.

On note G_d le sous-groupe discret de transformations orthogonales de $\mathbb{R}^d$ qui laisse $\mathbb{Z}^d$ invariant. Par exemple, en dimension $d = 2$, le sous-groupe G_d devient le groupe diédral

$$G_2 = \{e, \pm\frac{\pi}{2}, \pi, \text{réflexions/axes } Ox, Oy, \text{réflexions/diagonales } D_1, D_2\}.$$

Soit ρ une probabilité sur G_d telle que

— si $g \in G_d \setminus \{e\}$, alors $\rho(g) > 0$.
— $\rho(g) = \rho(g^{-1}), \forall g \in G_d$.
Pour $k \in \{1, \ldots, N\}$ et $g \in G_d$ on introduit la transformation $(k, g) : \mathbb{X}^{\mathrm{msr}}_{x,N} \to \mathbb{X}^{\mathrm{msr}}_{x,N}$ par

$$(k, g)\mathbf{y} = \mathbf{y}' := (y_0, \ldots, y_{k-1}, y_{k-1} + g(y_k - y_{k-1}), \ldots, y_{k-1} + g(y_N - y_{k-1}), \forall \mathbf{y} \in \mathbb{X}^{\mathrm{msr}}_{x,N}.$$

Dans la suite on note simplement $\mathbb{X}^{\mathrm{mas}}$ et $\mathbb{X}^{\mathrm{msr}}$ au lieu de $\mathbb{X}^{\mathrm{mas}}_{x,N}$ et $\mathbb{X}^{\mathrm{msr}}_{x,N}$.

Algorithme de génération de marches aléatoires sans recoupement

Require: Générateur $g \in G_d$ selon ρ,
générateur unif$(\{1, \ldots, N\})$
$\mathbf{y} \in \mathbb{X}^{\mathrm{msr}}$.
Ensure: $\mathbf{z} \in \mathbb{X}^{\mathrm{msr}}$.
Choisir $g \in G_d$ selon ρ.
Choisir $k \in \{1, \ldots, N\}$ selon unif$(\{1, \ldots, N\})$.
if $(k, g)\mathbf{y} \in \mathbb{X}^{\mathrm{msr}}$ **then**
$\mathbf{z} \leftarrow (k, g)\mathbf{y}$
else
$\mathbf{z} \leftarrow \mathbf{y}$
end if

(a) Montrer que l'algorithme ci-dessus définit une matrice stochastique P sur $\mathbb{X}^{\mathrm{msr}}$. Plus précisément, montrer que pour toute paire $\mathbf{y}, \mathbf{z} \in \mathbb{X}^{\mathrm{msr}}$, on a

$$P(\mathbf{y}, \mathbf{z}) = \begin{cases} \frac{1}{N} \sum\limits_{(k,g)} \rho(g) \mathbb{1}_{\{\mathbf{z}\}}((k, g)\mathbf{y}) & \text{si } \mathbf{z} \neq \mathbf{y}, \\ \frac{1}{N} \sum\limits_{(k,g)} \rho(g) \mathbb{1}_{\{\mathbf{y}\}}((k, g)\mathbf{y}) + \frac{1}{N} \sum\limits_{(k,g)} \rho(g) \mathbb{1}_{\mathbb{X}^{\mathrm{mas}} \setminus \mathbb{X}^{\mathrm{msr}}}((k, g)\mathbf{y}) & \text{si } \mathbf{z} = \mathbf{y}. \end{cases}$$

(b) Montrer que pour tout entier $N \geq 2$, la matrice P est stochastique, bistochastique, irréductible et apériodique.
(c) Conclure que P admet la mesure uniforme sur $\mathbb{X}^{\mathrm{msr}}$ comme mesure d'équilibre.

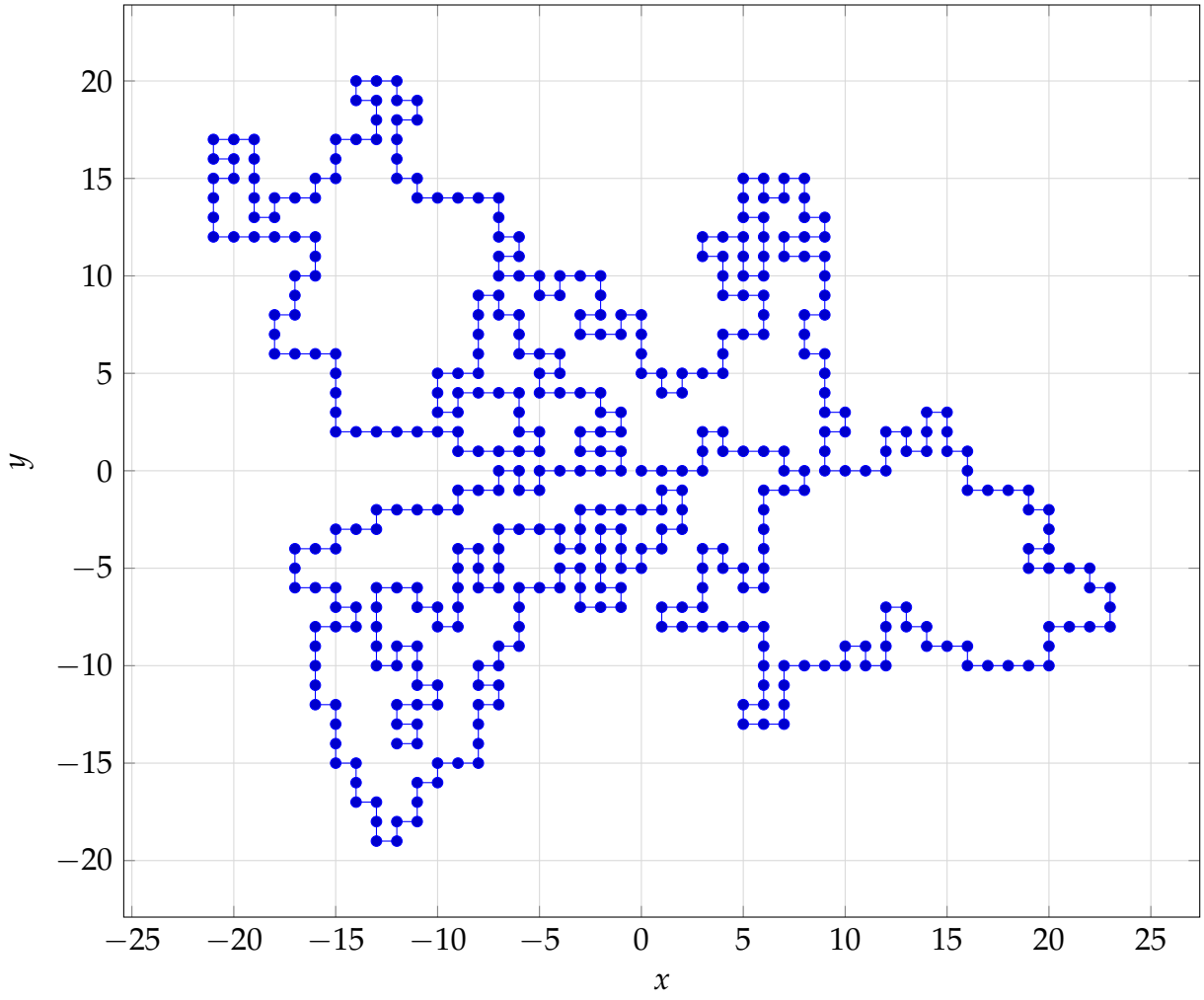


FIGURE 5.1 – Exemple de trajectoire d'une marche sans recoupement sur $\mathbb{Z}^2$ partant de l'origine et de longueur 437.

6 Notions de statistique

6.1 Motivation

L'objet de la théorie des probabilités peut se formuler succinctement comme l'étude de variables aléatoires définies sur un espace abstrait de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$ et qui prennent des valeurs dans un espace d'événements $(\mathbb{X}, \mathcal{X})$. Les questions que nous avons abordées aux chapitres précédents, concernent

— la loi $\mathbb{P}_X$ sur $(\mathbb{X}, \mathcal{X})$, image de $\mathbb{P}$ par X,
— certaines fonctionnelles définies sur l'espace de variables aléatoires à valeurs dans $\mathbb{X}$, comme l'espérance, la variance,
— le comportement asymptotique des suites de variables aléatoires $(X_n)_{n\in\mathbb{N}}$, lorsque celles-ci sont indépendantes ou ont des dépendances moins triviales, des dépendances markoviennes par exemple, etc.

En résumant, la théorie des probabilités permet de répondre à la question suivante : étant donné un certain procédé de génération de X, que peut-on dire à propos des observations qui en découlent. Par exemple, quelle est la valeur de $\mathbb{P}(X = x)$ lorsque $x \in \mathbb{X}$, ou quel est le comportement de $\sum_{k=1}^{n} f(X_k)$ à grand n, lorsque f a « des bonnes propriétés » et $(X_k)_{k\in\mathbb{N}}$ est une suite indépendante ou une suite markovienne, sont des questions étudiées par la théorie des probabilités.

En statistique mathématique, nous intéressons encore moins à l'espace $(\Omega, \mathcal{F}, \mathbb{P})$ qu'en théorie des probabilités. Nous avons un espace d'événements $(\mathbb{X}, \mathcal{X})$ et une famille spécifique $\Pi \subset \mathcal{M}_1(\mathbb{X}, \mathcal{X})$ de probabilités[1], avec $|\Pi| \geq 2$ pour éviter les trivialités. Nous disposons d'une variable aléatoire X à valeurs dans $\mathbb{X}$ dont la loi est un élément de l'ensemble Π mais nous ignorons lequel. Nous observons plusieurs réalisations (observations) de la variable aléatoire X et nous voulons, à partir des observations inférer[2] la loi qui a servi à générer les réalisations de X.

1. En dehors de quelques cas particulièrement simples, il n'est pas possible de choisir $\Pi = \mathcal{M}_1(\mathbb{X}, \mathcal{X})$ car le problème serait mal posé.

2. INFÉRER, verbe trans. A–(log.) Tirer, d'un fait ou d'une proposition donné(e), la conséquence qui en résulte. B–(p. ext.) Tirer une conclusion d'un fait ou d'un événement donnés. *Trésor de la langue française*, version en ligne (2019).

Exemple 6.1.1. On note $(\mathbb{X}, \mathcal{X}, \Pi)$ le triplet qui permet de poser le problème d'inférence.

1. Lorsque $\mathbb{X} = \{0,1\}$ muni de sa tribu exhaustive, on peut s'intéresser à la famille
$$\Pi = \{\mathbb{P}_\theta \in \mathcal{M}_1(\mathcal{X}) : \mathbb{P}_\theta = (1-\theta)\varepsilon_0 + \theta\varepsilon_1, \theta \in \Theta := [0,1]\}.$$
Dans ce cas simple, la famille $\Pi = (\mathbb{P}_\theta)_{\theta\in\Theta}$, avec $\Theta = [0,1]$, coïncide avec $\mathcal{M}_1(\mathbb{X}, \mathcal{X})$. On veut, à partir d'un échantillon $X_1, \dots, X_n$ inférer la valeur de θ. On peut aussi identifier $\mathbb{P}_\theta$ avec la densité discrète (par rapport à la mesure de dénombrement) f_θ qui vaut
$$f_\theta(x) = \theta^x(1-\theta)^{1-x}, x \in \mathbb{X}, \theta \in \Theta.$$
2. Lorsque $\mathbb{X} = \mathbb{R}$ et il est muni de sa tribu borélienne, on peut s'intéresser à la famille
$$\Pi = \{\mathbb{P}_\theta = \mathcal{N}(m, s^2) : \theta = (m, s^2) \in \Theta := \mathbb{R} \times \mathbb{R}_+\}$$
strictement incluse dans $\mathcal{M}_1(\mathbb{X}, \mathcal{X})$ et on veut, à partir d'un échantillon $X_1, \dots, X_n$ inférer la valeur de θ. Il est à signaler que dans cet exemple, la famille des probabilités qui composent Π est équivalente à la famille $(f_\theta)_{\theta\in\Theta}$ des fonctions positives intégrables sur $\mathbb{R}$ qui sont des densités de probabilité, i.e. pour $\theta = (m, s^2)$ et pour tout $B \in \mathcal{B}(\mathbb{R})$,
$$\mathbb{P}_\theta(B) = \int_B f_\theta(x)\mathrm{d}x, \text{ avec } f_\theta(x) = \frac{1}{\sqrt{2\pi}s}\exp\left(-\frac{(x-m)^2}{s^2}\right), x \in \mathbb{R}.$$
3. Lorsque $\mathbb{X} = \mathbb{R}$, muni de sa tribu borélienne, on peut considérer la famille des probabilités donnée par
$$\Pi = \{\mathbb{P} \in \mathcal{M}_1(\mathcal{X}) : \mathbb{P} \text{ a un densité } f, \text{ t.q. } \int_{\mathbb{R}} f''(t)^2\mathrm{d}t < \infty\}$$
strictement incluse dans $\mathcal{M}_1(\mathbb{X}, \mathcal{X})$.

Dans les deux premiers cas, la famille Π est isomorphe à un (sous)-espace de $\mathbb{R}^d$ pour un $d < \infty$. Dans le dernier cas, Π est isomorphe à un espace fonctionnel de dimension infinie (l'espace de Sobolev $W^{2,2}(\mathbb{R})$). Dans tous les cas, on peut identifier les probabilités qui composent Π avec leur densité par rapport à une mesure de référence.

Définition 6.1.2. Un **modèle statistique** est la donnée $(\mathbb{X}, \mathcal{X}, \Pi)$, où la famille des probabilités (ou de leurs densités par rapport à une mesure de référence) $\Pi \subset \mathcal{M}_1(\mathcal{X})$ est appelée **population**. Lorsque Π est isomorphe à un ensemble de paramètres) Θ, partie d'un espace de dimension finie, on parle de modèle statistique **paramétrique**, sinon de modèle statistique **non paramétrique**.

Lorsque un modèle statistique $(\mathbb{X}, \mathcal{X}, \Pi)$ est donné, nous pouvons considérer le modèle statistique produit, noté par abus de notation, $(\mathbb{X}, \mathcal{X}, \Pi)^n = (\mathbb{X}^n, \mathcal{X}^{\otimes n}, \Pi^{\otimes n})$, où
$$\Pi^{\otimes n} = \{\mathbb{P}^{\otimes n} : \mathbb{P} \in \Pi\}$$
qui correspond à n répétitions indépendantes d'une même expérience statistique. Lorsque X est une variable aléatoire dans $\mathbb{X}$ de loi $\mathbb{P} \in \Pi$ et $\mathbf{X} = (X_1, \dots, X_n)$ une suite de

n variables copies indépendantes de X, alors $\mathbf{X} \in \mathbb{X}^n$ et sa loi sera $\mathbb{P}^{\otimes n}$, i.e. le modèle statistique pour l'échantillon indépendant de taille n est entièrement déterminé par le modèle $(\mathbb{X}, \mathcal{X}, \Pi)$. Parfois, nous notons donc le modèle statistique simplement $(\mathbb{X}, \mathcal{X}, \Pi)$, même lorsqu'il s'agit d'un échantillon indépendant de taille n. Nous distinguons cependant entre les variables aléatoires X ou $\mathbf{X}$ à valeurs respectivement dans $\mathbb{X}$ ou $\mathbb{X}^n$ et leurs réalisations respectives $x = X(\omega) \in \mathbb{X}$ ou $\mathbf{x} = \mathbf{X}(\omega) \in \mathbb{X}^n$.

Définition 6.1.3. Soient un modèle statistique $(\mathbb{X}, \mathcal{X}, \Pi)$, un espace d'événements $(\mathbb{Y}, \mathcal{Y})$ et $S : \mathbb{X}^n \to \mathbb{Y}$ une applcation mesurable (une variable aléatoire sur $\mathbb{X}^n$). Alors S est appelée une **statistique**.

Une question pertinente est la suivante : puisqu'une statistique est une variable aléatoire sur $\mathbb{X}^n$ à valeurs dans $\mathbb{Y}$ pourquoi introduisons-nous un nouveau terme ? La réponse est que l'usage que nous en ferons de S engendrera une intuition différente de celle développée pour des variables aléatoires « ordinaires ».

6.2 Estimation paramétrique

Dans ce paragraphe, le modèle statistique est $(\mathbb{X}, \mathcal{X}, \Pi)$, où Θ est un sous-ensemble d'un espace $\mathbb{R}^d$ avec un certain $d < \infty$ et $\Pi := (\mathbb{P}_\theta)_{\theta \in \Theta}$, où . Notre but, est d'introduire des statistiques S qui permettront, à partir d'un échantillon généré avec une certaine probabilité $\mathbb{P}_\theta$, d'inférer θ.

6.2.1 Estimation ponctuelle

Exemple 6.2.1. Soient $\mathbb{X} = \{0, 1\}$ et $\mathbf{X}^{(n)} := (X_1, \ldots, X_n)$ des variables aléatoires indépendantes identiquement distribuées selon la loi $\mathbb{P}_\theta = \mathsf{Bernoulli}(\theta)$, avec $\theta \in \mathbb{R}_>$. La **moyenne empirique**

$$\overline{m}_n(\mathbf{X}^{(n)}) = \frac{1}{n} \sum_{k=1}^{n} X_k$$

est une statistique pour le modèle. Elle a la particularité que $\overline{m}_n(\mathbf{X}^{(n)}) \xrightarrow{\mathbb{P}_\theta} \theta$, lorsque $n \to \infty$. En outre $\mathbb{E}_\theta(\overline{m}^n) = \theta$.

Définition 6.2.2. $(\mathbb{X}, \mathcal{X}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ un modèle statistique et $\mathbf{X}^{(n)} := (X_1, \ldots, X_n)$ des variables aléatoires indépendantes sur $\mathbb{X}$ distribuées selon la loi inconnue $\mathbb{P}_\theta$, pour un certain $\theta \in \Theta \subset \mathbb{R}^d$. Un **estimateur ponctuel** (de θ) est une statistique $\hat{\theta}_n : \mathbb{X}^n \to \mathbb{R}^d$. L'estimateur est

— **cohérent** si $\hat{\theta}_n(\mathbf{X}^{(n)}) \xrightarrow{\mathbb{P}_\theta} \theta$, lorsque $n \to \infty$,
— **non-biaisé** si le biais $b_\theta(\hat{\theta}_n) := \mathbb{E}_\theta(\hat{\theta}_n) - \theta = 0$, pour tout n.

L'**erreur quadratique moyenne** d'un estimateur est définie comme

$$\mathsf{EQM}_\theta(\hat{\theta}_n) := \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2].$$

Dans l'exemple 6.2.1, la moyenne empirique est un estimateur ponctuel cohérent et non-biaisé de θ.

Lemme 6.2.3.

$$\mathit{EQM}_\theta(\hat{\theta}_n) = b_\theta(\hat{\theta}_n)^2 + \mathsf{Var}_\theta(\hat{\theta}_n).$$

Démonstration. Voir exercice 72. □

Définition 6.2.4. Un estimateur est **asymptotiquement normal** si

$$\frac{\hat{\theta}_n - \theta}{\sqrt{n}} \xrightarrow{\text{loi}} \mathcal{N}(0,1).$$

Au delà de l'estimateur de la moyenne empirique, d'autres estimateurs sont couramment utilisés, comme celui du maximum de vraisemblance.

Définition 6.2.5. Soient un modèle statistique $(\mathbb{X}, \mathcal{X}, (f_\theta)_{\theta\in\Theta})$, $\mathbf{X}^{(n)} = (X_1, \dots, X_n)$ une famille indépendante de variables aléatoires à valeurs dans $\mathbb{X}$, identiquement distribuées selon une loi de densité (discrète ou continue) f_θ. On définit les statistiques suivantes :

vraisemblance : $L_n(\theta) := L_n(\theta, \mathbf{X}^{(n)}) = \prod_{i=1}^n f_\theta(X_i)$,
log-**vraisemblance :** $l_n(\theta) = \log L_n(\theta)$,
estimateur du maximum de vraisemblance : $\hat{\theta}_n^{\text{MV}} := \arg\max L_n(\theta) = \arg\max l_n(\theta)$.

Exemple 6.2.6. Soit $\mathbf{X}^{(n)} = (X_1, \dots, X_n)$ une famille de variables aléatoires indépendantes et identiquement distribuées.

— Si X_1 suit la loi de Bernoulli(θ) avec $\theta \in [0,1]$, l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\text{MV}}$ coïncide avec la moyenne empirique $\overline{\theta}_n$.

— Si X_1 suit une loi normale de paramètres $\theta := (m, s^2)$, l'estimateur de maximum de vraisemblance est $\hat{\theta}_n^{\text{MV}} = (\overline{m}_n, \overline{v}_n)$, où l'estimateur de maximum de vraisemblance pour m coïncide avec la moyenne emprique $\hat{\theta}_{n,1} = \overline{m}_n$ et

$$\overline{v}_n = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{m}_n)^2$$

est un estimateur de la variance.

— Si X_1 suit la loi uniforme sur $[0,\theta]$, $\theta > 0$, alors l'estimateur du maximum de vraisemblance de θ est $\hat{\theta}_n^{MV} = \max(X_1, \dots, X_n)$.

Définition 6.2.7. Soit $\mathbf{X}^{(n)} = (X_1, \dots, X_n)$ une famille de variables aléatoires indépendantes et identiquement distribuées et de variance finie. La statistique

$$\overline{V}_n(\mathbf{X}^{(n)}) = \frac{1}{n-1}\sum_{k=1}^n (X_k - \overline{m}_n(\mathbf{X}^{(n)}))^2$$

est appelée **variance empirique**.

Remarque 6.2.8. Il ne faut pas confondre $\overline{v}_n$ avec $\overline{V}_n$ (ils diffèrent en la normalisation) et ne deviennent égaux qu'asymptotiquement. Il est montré (voir exercice 70) que la variance empirique est un estimateur non-biaisé de la variance.

6.2.2 Estimation d'intervalles de confiance

Définition 6.2.9. Soit $(\mathbb{X}, \mathcal{X}, (\mathbb{P}_\theta)_{\theta\in\Theta}$ un modèle statistique, avec $\Theta \subset \mathbb{R}$. Un **intervalle de confiance** à niveau d'erreur $\alpha \in [0,1]$ est un intervalle (aléatoire) de la forme $C_n = [a_n, b_n]$, où $a_n := a_n(X_1, \dots, X_n)$ et $b_n := b_n(X_1, \dots, X_n)$ sont deux statistiques telles que

$$\mathbb{P}(\theta \in C_n) \geq 1 - \alpha.$$

Remarque 6.2.10. Un intervalle de confiance n'est pas une affirmation probabiliste à propos de θ car θ n'est pas aléatoire ! La signification de l'affirmation est que si on répétait l'expérience indéfiniment, i.e. on observait $\mathbf{X}^{(i)} = (X_1^{(i)}, \dots, X_n^{(i)})$ pour $i \in \mathbb{N}$, la suite d'intervalles aléatoires $C_n^{(i)}$, déterminés par les statistiques $a_n^{(i)}$ et $b_n^{(i)}$, contiendraient la valeur déterministe (mais inconnue) θ dans $100(1-\alpha)\%$ des cas.

Définition 6.2.11. Une variable aléatoire Z à valeurs dans $\mathbb{R}_{\geq}$ dont la loi a comme densité $\mathbb{P}(Z \in \mathrm{d}z) = f_k(z)\mathrm{d}z$, avec

$$f_k(z) = \frac{z^{\frac{k}{2}-1}e^{-\frac{z}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})},$$

est dite suivre la **loi de χ^2_k avec k degrés de liberté**. La fonction de répartition de cette loi est

$$F_k(z) = \frac{\gamma(k/2, z/2)}{\Gamma(k/2)},$$

où γ est la loi gamma incomplète (F_k est tabulée).

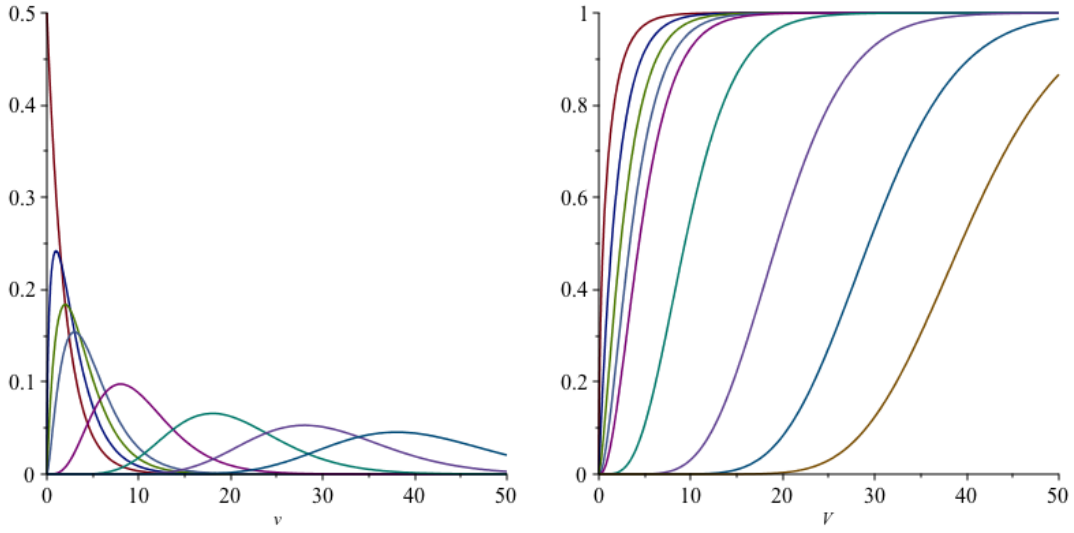


FIGURE 6.1 – La densité f_k (figure de gauche) et la fonction de répartition F_k (figure de droite) respectivement pour $k = 2, 3, 4, 5, 10, 20, 30, 40$ et $k = 1, 2, 3, 4, 5, 10, 20, 30, 40$ degrés de liberté. La densité pour $k = 1$ n'est pas présentée car elle diverge en 0.

Lemme 6.2.12. *Soient k variables $Y_1, \dots, Y_k$ indépendantes et identiquement distribuées selon la loi $\mathcal{N}(0,1)$. Alors la variable $Z = \sum_{i=1}^k Y_i^2$ suit la loi de χ^2_k avec k degrés de liberté.*

Théorème 6.2.13. *Soit $\mathbf{X}^{(n)} = (X_1, \dots, X_n)$ une famille de variables aléatoires indépendantes et identiquement distribuées selon la loi $\mathcal{N}(m, \sigma^2)$. Alors $(n-1)\overline{V}_n/\sigma^2$, où $\overline{V}_n$ est la variance empirique (définie en 6.2.7), suit la loi du χ^2_{n-1}.*

Exemple 6.2.14. Lors d'une expérience, les valeurs suivantes ont été obtenues lors de 8 réalisations indépendantes d'une variable aléatoire X : $4.4, 4.7, 4.8, 4.5, 4.4, 4.2, 4.2, 4.0$. La moyenne empirique est $\overline{m}_8 = 4.4$ et la variance empirique $\overline{V}_8 = 0.0714$. On cherche

à déterminer un intervalle de confiance à seuil d'erreur $\alpha = 0.1$ pour la variance σ^2 de la loi qui a servi pour générer l'échantillon. On sait que $7\overline{V}_8/\sigma^2$ suit la loi de χ^2_7. À partir des tables, on détermine les valeurs $z_-(\alpha)$ et $z_+(\alpha)$ telles que la variable Z suivant la loi de χ^2_7, vérifie

$$\mathbb{P}(Z \leq z_-(\alpha)) = \mathbb{P}(Z \geq z_+(\alpha)) = \alpha/2.$$

Par conséquent, l'intervalle de confiance à niveau d'erreur α sera $C = [\frac{7\overline{V}_8}{z_+(\alpha)}, \frac{7\overline{V}_8}{z_-(\alpha)}]$ et on aura que $\mathbb{P}(\sigma^2 \in C) = 1 - \alpha$. Pour $\alpha = 0.1$, on détermine à partir des tables de F_7 que $z_-(\alpha) = 2.17$ et $z_+(\alpha) = 14.1$, ce qui conduit à l'estimation $\mathbb{P}(0.19 \leq \sigma \leq 0.48) = 0.9$.

6.2.3 Tests d'hypothèses

Théorème de Neyman-Pearson

On illustre le problème dans une situation simple. Soit $(\mathbb{X}, \mathcal{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un modèle statistique, avec $\Theta = \{\theta_0, \theta_1\} \equiv \{0, 1\}$ et $\mathbb{X} \subset \mathbb{R}$ discret. Pour alléger la notation, on écrit f_0 et f_1 pour les densités de probabilité de la population. On observe un échantillon $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ généré selon l'une des lois ; on doit décider s'il a été engendré par une loi de densité f_0 ou f_1. (En général, on note f_θ la densité de la loi $\mathbb{P}_\theta$). On appelle traditionnellement

— H_0 l'hypothèse **nulle** : « l'échantillon a été engendré par la loi de densité f_0 » que l'on test contre

— H_1 l'hypothèse dite **alternative** : « l'échantillon n'a pas été engendré par f_0 » (dans le cas présent ultrasimplifié, H_1 correspond bien sûr à la génération par f_1).

Définition 6.2.15. Un **test d'hypothèses** est la donnée du triplet $(\mathbf{X}^{(n)}, (H_0, H_1), R)$ où $R \subseteq \mathbb{R}^n$ est la **région critique de rejet de** H_0 : si $\mathbf{X}^{(n)} \in R$, alors on accepte H_1 sinon, on accepte H_0.

— On appelle [3] **erreur de type I** la quantité

$$\alpha := \alpha(R) = \mathbb{P}(\mathbf{X}^{(n)} \in R) = \mathbb{P}(H_1 \text{ acceptée} | H_0 \text{ prévaut})$$

— et **erreur de type II** la quantité

$$\beta := \beta(R) = \mathbb{P}(\mathbf{X}^{(n)} \in R^c) = \mathbb{P}(H_0 \text{ acceptée} | H_1 \text{ prévaut}).$$

Le paramètre α est aussi appelé **taille de la région critique** tandis que le paramètre $1 - \beta$ est appelé **puissance**.

On souhaiterait minimiser la probabilité de prendre une décision erronée :

$$\begin{aligned}\gamma &= \mathbb{P}(H_1 \text{ acceptée et } H_0 \text{ prévaut}) + \mathbb{P}(H_0 \text{ acceptée et } H_1 \text{ prévaut}) \\ &= \alpha\text{« }\mathbb{P}\text{ »}(H_0 \text{ prévaut}) + \beta\text{« }\mathbb{P}\text{ »}(H_1 \text{ prévaut})\end{aligned}$$

mais ce problème est mal posé car il n'y a pas de moyen objectif de déterminer la probabilité notée « $\mathbb{P}$ » ci-dessus. Cependant, même si « $\mathbb{P}$ » est une quantité subjective, on

3. Il serait plus cohérent de les appeler erreurs d'ordre 0 et 1 respectivement mais, pour des raisons historiques, on garde la nomenclature type I et II.

exige que « $\mathbb{P}$ »(H_0 prévaut) + « $\mathbb{P}$ »(H_1 prévaut) = 1, par conséquent, $\gamma \leq \max(\alpha, \beta)$. La théorie des tests d'hypothèses consiste donc à construire des régions critiques de taille α fixée qui minimisent β. Le théorème de Neyman-Pearson 6.2.16 nous donne une méthode de construction explicite de la région d'acceptation de l'hypothèse H_1.

Pour $\mathbb{P}_\theta$, de densité f_θ, $\theta \in \Theta \simeq \{0,1\}$ et pour un n-uplet $\mathbf{x} = (x_1, \ldots, x_n)$, la **fonction de vraisemblance** prend la forme $L_\theta(\mathbf{x}) := f_\theta(x_1) \cdots f_\theta(x_n)$. On remarque que l'erreur de type I, correspondant à une région critique de rejet R, est

$$\alpha(R) = \mathbb{P}_0((X_1, \ldots, X_n) \in R) = \int_R f_0(x_1) \cdots f_0(x_n) \mathrm{d}^n x = \int_R L_0(\mathbf{x}) \mathrm{d}\mathbf{x},$$

tandis que l'erreur de type II est déterminée par complémentation de

$$1 - \beta(R) = \mathbb{P}_1((X_1, \ldots, X_n) \in R) = \int_R f_1(x_1) \cdots f_1(x_n) \mathrm{d}^n x = \int_R L_1(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

Théorème 6.2.16 (Neyman-Pearson). *Soit $((X_1, \ldots, X_n), (H_0, H_1), R)$ un test d'hypothèses où, pour tout $\delta > 0$, on pose $R := R(\delta) = \{\mathbf{x} \in \mathbb{R}^n : \frac{L_0(\mathbf{x})}{L_1(\mathbf{x})} \leq \delta\}$ pour une région critique et $\alpha := \alpha(R)$ pour sa taille. Toute autre région critique S ayant la même taille, i.e. $\alpha(S) = \alpha(R)$, vérifiera $\beta(S) \geq \beta(R)$. En d'autres termes, la région R est la région critique optimale à taille fixée α.*

Démonstration. On aura démontré l'optimalité de R si pour toute région critique S avec $\alpha(S) = \alpha(R)$ on aura montré $1 - \beta(S) \leq 1 - \beta(R)$ ou, ce qui est équivalent, $\int_S L_1(\mathbf{x}) \mathrm{d}\mathbf{x} \leq \int_R L_1(\mathbf{x}) \mathrm{d}\mathbf{x}$. Or, on a

$$\alpha(R) = \alpha(S) \Leftrightarrow \int_R L_0(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_S L_0(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

En décomposant

$$R = (R \cap S) \sqcup (R \cap S^c),$$
$$S = (S \cap R) \sqcup (S \cap R^c),$$

nous avons

$$\int_{R \cap S} L_0(\mathbf{x}) \mathrm{d}\mathbf{x} + \int_{R \cap S^c} L_0(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{R \cap S} L_0(\mathbf{x}) \mathrm{d}\mathbf{x} + \int_{S \cap R^c} L_0(\mathbf{x}) \mathrm{d}\mathbf{x}$$

ou

$$\int_{R \cap S^c} L_0(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{S \cap R^c} L_0(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

Par ailleurs, sur R, nous avons $L_1 \geq \frac{L_0}{\delta}$ tandis que sur R^c, nous avons $L_1 < \frac{L_0}{\delta}$. On conclut que

$$\int_{R \cap S^c} L_1(\mathbf{x}) \mathrm{d}\mathbf{x} \geq \int_{R \cap S^c} \frac{L_0(\mathbf{x})}{\delta} \mathrm{d}\mathbf{x} = \int_{S \cap R^c} \frac{L_0(\mathbf{x})}{\delta} \mathrm{d}\mathbf{x} \geq \int_{S \cap R^c} L_1(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

En comparant les expressions aux deux extrémités des inégalités précédentes, on a finalement

$$\int_R L_1(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{R \cap S^c} L_1(\mathbf{x}) \mathrm{d}\mathbf{x} + \int_{R \cap S} L_1(\mathbf{x}) \mathrm{d}\mathbf{x} \geq \int_{S \cap R^c} L_1(\mathbf{x}) \mathrm{d}\mathbf{x} + \int_{R \cap S} L_1(\mathbf{x}) = \int_S L_1(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

□

Test χ_2 d'ajustement de Pearson

Dans l'approche ci-dessus, nous avons utilisé la statistique de vraisemblance. Une autre statistique couramment utilisée est celle de Pearson, notée Q_n, dont la définition est donnée dans l'énoncé du théorème 6.2.17 ci-dessous, communément appelée « test du χ^2 ». Nous donnons quelques détails supplémentaires sur ce test dans un cadre un peu plus général.

Supposons qu'à partir du modèle stochastique primaire $(\mathbb{X}, \mathcal{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$, on agrège les observations en $k+1$ classes, c'est-à-dire, on définit une partition $\mathbb{X} = \sqcup_{i\in\mathbb{K}}\mathbb{X}_i$ où $\mathbb{K} \simeq \{0,\dots,k\}$ et on introduit la statistique $\boldsymbol{\nu} : \mathbf{X}^n \to \mathbb{N}^{\mathbb{K}}$ par

$$\boldsymbol{\nu}(\mathbf{X}^{(n)}) = \Big(\nu_0(\mathbf{X}^{(n)}),\dots,\nu_k(\mathbf{X}^{(n)})\Big),$$

où, pour tout $j \in \mathbb{K}$, on pose $\nu_j := \nu_j(\mathbf{X}^{(n)}) = \sum_{i=1}^n \mathbb{1}_{\mathbb{X}_j}(X_i)$. Il est évident que ν_j/n représente la fréquence empirique avec laquelle la classe $j \in \mathbb{K}$ est chargée par l'échantillon. Par ailleurs, chaque $\mathbb{P}_\theta$ de la population induit un vecteur de probabilité $\mathbf{p} := \mathbf{p}_\theta \in \mathsf{VP}_{\mathbb{K}}$, où

$$\mathsf{VP}_{\mathbb{K}} = \{\mathbf{p} \in \mathbb{R}_>^{k+1} : \sum_{i=0}^k p(i) = 1\}.$$

i.e. on suppose que pour tout $i \in \mathbb{K}$, on a $p_i > 0$. Par le théorème de grands nombres

$$\lim_{n\to\infty} \frac{\nu_i(\mathbf{X}^{(n)})}{n} = \mathbb{P}_\theta(\mathbb{X}_i) = p_i.$$

Ce résultat permet d'affirmer que composante par composante, ν_i/n est un estimateur de p_i. Or les paramètres $(p_i)_{i\in\mathbb{K}}$ ne sont pas libres car elles vérifient l'égalité $\sum_i p_i = 1$. Nous voulons cependant une approche qui permet de former un estimateur simultané pour toutes les composantes du vecteur de probabilité. Pour cela nous utiliserons l'estimateur du maximum de vraisemblance. Pour tout vecteur de probabilité $\mathbf{q} \in \mathsf{VP}_{\mathbb{K}}$, on introduit la fonction de vraisemblance

$$L_n(\mathbf{q}) := L_n(\mathbf{q}, \mathbf{X}^{(n)}) = \prod_{i\in\mathbb{K}} q_i^{\nu_i(\mathbf{X}^{(n)})}$$

et on cherche le maximum de vraisemblance (voir exercice 76). On constate que le maximum de vraisemblance est atteint pour $\hat{\mathbf{q}} := \hat{\mathbf{q}}_n = \frac{\boldsymbol{\nu}(\mathbb{X}^{(n)})}{n}$.

Nous voulons tester l'hypothèse nulle $H_0 : \mathbf{p} = \hat{\mathbf{p}}$ (où $\hat{\mathbf{p}} = \boldsymbol{\nu}/n$) contre l'hypothèse alternative $H_1 : \mathbf{p} \neq \hat{\mathbf{p}}$.

Il convient donc de considérer le rapport des vraisemblances

$$R_n = \frac{L_n(\hat{\mathbf{p}})}{L_n(\mathbf{p})}.$$

Par un simple calcul, on constate que

$$\log R_n = n \sum_{i\in\mathbb{K}} \hat{p}_i \log \frac{\hat{p}_i}{p_i}.$$

Or la fonction $f : \mathbb{R}_\geq \to \mathbb{R}$ donnée par la formule $f(x) = x \log \frac{x}{x_0}$, où $x_0 > 0$ et $0 \log 0 = 0$, admet comme développement $f(x) = (x - x_0) + \frac{1}{2x_0}(x - x_0)^2 + o(|x - x_0|^3)$. Par conséquent $\log R_n = \sum_{i\in\mathbb{K}} \frac{(\nu_i - np_i)^2}{np_i} + o(|\hat{p}_i - p_i|^3)$.

Théorème 6.2.17. *Sous l'hypothèse H_0, lorsque $n \to \infty$, la statistique*

$$Q_n = \sum_{i \in \mathbb{K}} \frac{(\nu_i - np_i)^2}{np_i}$$

converge en loi vers χ^2_k.

Par conséquent, nous pouvons exprimer la condition de rejet de l'hypothèse H_0 à niveau d'erreur α si $Q_n > \chi^2_{k,\alpha}$, où $\chi^2_{k,\alpha}$ est le quantile supérieur de la loi χ^2_k, i.e. si Z est une variable aléatoire distribuée selon la loi du χ^2_k,

$$\mathbb{P}(Z > \chi^2_{k,\alpha}) = \alpha.$$

6.3 Exercices

Estimation ponctuelle

70. Montrer que la **variance empirique**, définie par :

$$\overline{V}_n(\mathbf{X}^{(n)}) := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{m}_n(\mathbf{X}^{(n)}) \right)^2,$$

est un estimateur non-biaisé de la variance.

71. Soit $(\mathbb{X}, \mathcal{X}, \Pi)$ un modèle statistique adéquat. Déterminer l'estimateur du maximum de vraisemblance lorsque $\Pi = (\mathbb{P}_\theta)_{\theta \in \Theta}$ avec
    (a) $\Theta = \mathbb{R}_>$ et $\mathbb{P}_\theta = \mathcal{B}(N, \theta)$,
    (b) $\Theta = \mathbb{R} \times \mathbb{R}_>$ et $\mathbb{P}_\theta = \mathcal{N}(\theta_1, \theta_2^2)$.

72. Démontrer le lemme établissant l'égalité

$$\mathsf{EQM}_\theta(\hat{\theta}_n) = b_\theta(\hat{\theta}_n)^2 + \mathsf{Var}_\theta(\hat{\theta}_n).$$

73. Soit la population $\Pi = \mathsf{unif}([0, \theta])$ avec $\theta \in \mathbb{R}_>$. Considérer les statistiques
    (a) $S_1(\mathbf{x}) = \frac{2}{n} \sum_{i=1}^{n} x_i$ et
    (b) $S_2(\mathbf{x}) = \max\{x_1, \ldots, x_n\}$.
    (c) $S_3(\mathbf{x})$ l'estimateur de maximum de vraisemblance.

    Définissent-elles des estimateurs de θ ? Si oui, ces estimateurs sont-ils non-biaisés ?

Intervalles de confiance

74. Le but de cet exercice est de démontrer une amélioration — connue sous le nom d'inégalité de Hoeffding — de l'inégalité obtenue en exercice 58.

    Inégalité de Hoeffding : Soit $(X_i)_{i \in \mathbb{N}}$ une suite de variables aléatoires indépendantes centrées telles que, pour tout $i \in \mathbb{N}$, nous avons $X_i \in [a_i, b_i]$ (pour deux réels $a_i < b_i$). Soit $\varepsilon > 0$. Alors

$$\mathbb{P}(\sum_{i=1}^{n} X_i \geq \varepsilon) \leq \inf_{t>0} \left(\exp(-\varepsilon t) \prod_{i=1}^{n} \exp\left(\frac{t^2 (b_i - a_i)^2}{8} \right) \right).$$

(a) Utiliser l'inégalité de Markov pour montrer que, pour tout $t > 0$,

$$\mathbb{P}(\sum_{i=1}^{n} X_i \geq \varepsilon) \leq \exp(-t\varepsilon) \prod_{i=1}^{n} \mathbb{E} \exp(tX_i).$$

(b) Montrer que l'on peut écrire, pour tout $i \in \mathbb{N}$, la variable aléatoire X_i comme une combinaison convexe aléatoire, i.e.

$$[a_i, b_i] \ni X_i = (1 - \xi_i)a_i + \xi_i b_i$$

pour une variable aléatoire ξ_i que l'on exprimera en fonction de X_i.

(c) Utiliser la convexité de $t \mapsto \exp(tx)$, pour $x \in \mathbb{R}$ arbitraire fixé, pour majorer $\exp(tX_i)$ en se servant de la décomposition convexe de X_i, établie à la question précédente.

(d) Utiliser le centrage des variables aléatoires ($\mathbb{E}X_i = 0$), pour exprimer $\mathbb{E} \exp(tX_i) = \exp(g(u_i))$, où $u_i = t(b_i - a_1)$ et $g(u) = -c_i u_i + \log(1 - c_i + c_i \exp(u_i))$ et $c_i = -\frac{a_i}{b_i - a_i}$.

(e) Calculer $g(0), g'(0)$ et montrer que pour tout $v > 0$, $g''(v) \leq 1/4$.

(f) On rappelle que pour toute fonction $g \in C^2(\mathbb{R})$, et pour tout $u > 0$, il existe $v \in [0, u]$ tel que $g(u) = g(0) + ug'(0) + \frac{u^2}{2} g''(v)$. S'en servir pour conclure que $\mathbb{E} \exp(tX_i) \leq \exp(g(u_i)) \leq \exp(\frac{t^2}{8}(b_i - a_i)^2)$.

75. Soit $(X_i)_{i \in \mathbb{N}}$ une suite indépendante et identiquement distribuée selon la loi $\mathsf{Bernoulli}(\theta)$, avec $\theta \in [0, 1]$.

(a) Soient $\alpha \in]0, 1[$ et $\varepsilon_n = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$. Noter $\overline{m}_n$ la moyenne empirique et

$$C_n =]\overline{m}_n - \varepsilon_n, \overline{m}_n + \varepsilon_n[\cap [0, 1]$$

la troncature éventuelle de l'intervalle aléatoire de longueur $2\varepsilon_n$ et de centre aléatoire $\hat{\theta}_n$ induite par son intersection avec l'intervalle $[0, 1]$. Montrer que

$$\mathbb{P}_\theta(C_n \ni \theta) \geq 1 - \alpha.$$

(b) Effectuer l'expérience suivante sur ordinateur. Fixer $\alpha = 0.05, \theta = 0.4$ et générer un échantillon $(X_i)_{i=1,\ldots,10000}$.

(c) Déterminer pour tout $j = 1, \ldots, 10000$, les bords de l'intervalle C_j.

(d) Calculer la suite $K_n = \sum_{i=1}^{n} \mathbb{1}_{C_i}(\theta)$ pour $n = 1, \ldots, 10000$ et présenter les couples (n, K_n) et $(n, |C_n|)$ pour $n = 1, \ldots, 10000$.

Tests d'hypothèses

76. Maximiser la fonction de vraisemblance $L_n(\mathbf{p}_\theta, \mathbf{X}) = \prod_{j=0}^{k} p_\theta(j)^{\nu_j(\mathbf{X})}$ (correspondant au modèle statistique agrégé en $k + 1$ classes) sur l'ensemble VP_{k+1} en introduisant un multiplicateur de Lagrange pour tenir compte de la contrainte $\sum_{j=0}^{k} p(j) = 1$. Conclure que l'estimateur du maximum de vraisemblance est $\hat{\mathbf{p}}_n = \frac{\boldsymbol{\nu}}{n}$.

77. Utiliser un générateur de nombres aléatoires sur votre ordinateur et engendrer un échantillon de taille 1000. Séparer l'intervalle $[0, 1]$ en 10 classes équiprobables et tester par la méthode de χ^2 si l'hypothèse H_0 : le générateur est uniforme (contre l'alternative H_1 : le générateur n'est pas uniforme) doit être rejetée.

Deuxième partie

Théorie de l'information

7 Quantification de l'information

Dans le chapitre 1 nous avons présenté les arguments qui nous amènent à mesurer la quantité d'information en bits. Dans ce chapitre, nous allons formaliser précisément les propriétés que doit avoir la quantité d'information, démontrer que la seule fonction qui possède ces propriétés est la fonction introduite en chapitre 1 et donner trois interprétations différentes de cette notion.

Une fois la notion d'information clarifiée, nous introduirons la notion de registre de stockage et de ses états (informationnels) comme une abstraction mathématique d'un dispositif physique dans lequel l'information puisse être stockée et nous allons décrire son état informationnel comme une représentation abstraite de son contenu à un instant particulier.

7.1 Postulats d'une quantité d'incertitude, entropie

Soit X une variable aléatoire définie sur un espace $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans un ensemble (alphabet) fini $\mathbb{X}$ (muni de sa tribu exhaustive) et de loi $\mathbb{P}_X$ (décrite par son vecteur de probabilité). Si $\mathsf{card}\mathbb{X} = M$, nous pouvons toujours identifier l'alphabet $\mathbb{X}$ avec $\{x_1, \ldots, x_M\} \simeq \{1, \ldots, M\} \simeq \{0, \ldots, M-1\}$. La loi de X sera caractérisée par la donnée de $\mathbb{P}_X(\{x\}) = \mathbb{P}(X = x)$, pour $x \in \mathbb{X}$, codée dans le vecteur de probabilité $\mathbf{p} = (p(x))_{x\in\mathbb{X}}$, i.e.

$$\mathbb{P}_X(\{x\}) = \mathbb{P}(X = x) = p(x).$$

Dans le cadre défini ci-dessus, nous introduisons deux quantités.

Une fonction h : quantifiant la réduction de l'incertitude d'un événement.

Considérons l'événement $\{X = x\}$, pour un $x \in \mathbb{X}$. Avant que l'expérience décrite par la variable aléatoire X soit réalisée, nous avons une incertitude maximale ; lorsque la valeur prise par X nous est révélée, l'incertitude s'annule. La réduction de l'incertitude sera donc exactement l'incertitude a priori. Il est intuitivement clair que l'incertitude a priori de $\{X = x\}$ est une fonction de $p(x)$. En effet, si nous savons que la loi de X est concentrée sur x, cette incertitude est nulle. Par contre, elle doit être maximale lorsque la loi de X est l'équidistribution

sur $\mathbb{X}$. Nous définissons donc, pour un événement $A \in \mathcal{X}$ donné de probabilité $\mathbb{P}_X(A) = p$, la fonction h par

$$]0,1] \ni p \mapsto h(p) \in \mathbb{R}.$$

La fonction H : l'espérance de h.

La fonction h pouvant varier beaucoup avec p, il est plus raisonnable d'associer à l'**incertitude d'une variable aléatoire** X dont la loi est décrite par le vecteur $\mathbf{p} = (p(x))_{x\in\mathbb{X}}$ la quantité $H(\mathbf{p}) = \sum_{x\in\mathbb{X}} p(x)h(p(x))$ qui représente l'incertitude moyenne de l'observation. Les variables aléatoires étant en bijection avec les vecteurs de probabilités qui décrivent leurs lois, nous pouvons identifier la quantité $H(\mathbf{p})$ avec $H(X)$ et nous pouvons (provisoirement) noter $H_M(X)$ où $H_M(\mathbf{p})$ cette espérance lorsque $|\mathbb{X}| = M$.

Supposons que la variable aléatoire X soit uniformément distribuée dans $\mathbb{X}$, i.e. $p(x) = \frac{1}{M}$, pour tout $x = \mathbb{X}$, où $|\mathbb{X}| = M$, et notons $f(M) = H_M((\frac{1}{M}, \ldots, \frac{1}{M}))$ la valeur de l'incertitude moyenne associée dans ce cas d'équidistribution. L'incertitude inhérente à une expérience consistant à choisir entre les valeurs prise par une pièce honnête est plus petite que l'incertitude inhérente d'un tirage du loto, i.e. $f(2) < f(1.3 \times 10^7)$. Il est donc intuitivement clair d'exiger comme premier postulat :

Postulat 7.1.1 (Postulat de monotonie)**.** *La fonction $f : \mathbb{N} \to \mathbb{R}_+$ définie par $f(M) := H_M((\frac{1}{M}, \ldots, \frac{1}{M}))$ est une fonction strictement croissante de son argument.*

Considérons maintenant deux variables aléatoires indépendantes X et Y définies sur le même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ uniformément distribuées respectivement dans $\mathbb{X}$ et $\mathbb{Y}$, avec $|\mathbb{X}| = L$ et $|\mathbb{Y}| = M$. L'expérience composite fait intervenir la variable aléatoire (X, Y) à valeurs dans $\mathbb{X} \times \mathbb{Y}$, dont le cardinal est $L \times M$. Si la valeur prise par X nous est révélée, l'incertitude de Y n'est pas affectée car Y est indépendante de X. Cependant, l'incertitude totale $f(LM)$ est réduite de l'incertitude $f(L)$ relative à X. Ceci nous amène donc au deuxième

Postulat 7.1.2 (Postulat d'extensivité)**.** *Pour tout $L, M \geq 1$, nous avons $f(LM) = f(L) + f(M)$.*

Nous allons maintenant relaxer la condition de distribution uniforme. Supposons que $\mathbf{p}$ est un vecteur de probabilité arbitraire sur $\mathbb{X}$ (de cardinal $|\mathbb{X}| = M$) et que X prend des valeurs dans $\mathbb{X}$ avec probabilité $\mathbb{P}(X = x) = p(x)$. Considérons une partition $\mathbb{X} = \mathbb{X}_1 \sqcup \mathbb{X}_2$ et notons $q_i = \sum_{x\in\mathbb{X}_i} p(x)$, avec $|\mathbb{X}_i| = M_i, i = 1, 2$.

Nous allons considérer l'expérience effectuée en deux étapes : lors de la première étape, nous nous intéressons à l'événement $X \in \mathbb{X}_i$, avec $\mathbb{P}(X \in \mathbb{X}_i) = q_i, i = 1, 2$ et lors de la deuxième étape nous examinons si la variable aléatoire prend une valeur précise sachant qu'elle est dans l'un ou l'autre des groupes $\mathbb{X}_1$ ou $\mathbb{X}_2$. Plus précisément, nous calculons

$$\begin{aligned}
\mathbb{P}(X = x) &= \mathbb{P}(X = x|X \in \mathbb{X}_1)\mathbb{P}(X \in \mathbb{X}_1) + \mathbb{P}(X = x|X \in \mathbb{X}_2)\mathbb{P}(X \in \mathbb{X}_2) \\
&= \frac{p(x)}{q_1} q_1 \mathbb{1}_{\mathbb{X}_1}(x) + \frac{p(x)}{q_2} q_2 \mathbb{1}_{\mathbb{X}_2}(x) \\
&= p(x).
\end{aligned}$$

Nous formulons alors le troisième postulat de la quantité d'incertitude

Postulat 7.1.3 (Postulat de regroupement)**.** *Avec les mêmes notations que ci-dessus,*

$$H_M(\mathbf{p}) = q_1 H_{M_1}\left(\frac{\mathbf{p}\restriction_{\mathbb{X}_1}}{q_1}\right) + q_2 H_{M_2}\left(\frac{\mathbf{p}\restriction_{\mathbb{X}_2}}{q_2}\right) + H_2((q_1, q_2)).$$

Finalement, nous introduisons le

Postulat 7.1.4 (Postulat de continuité)**.** *La fonction $H_2(p, 1-p)$ est continue en $p \in [0,1]$.*

Théorème 7.1.5. *L'unique fonction (à une constante multiplicative près) qui vérifie les postulats de monotonie, d'extensivité, de regroupement et de continuité est la fonction*

$$PV_M \ni \mathbf{p} \mapsto H_M(\mathbf{p}) = -C \sum_{x \in \mathbb{X}} p(x) \log p(x),$$

où le logarithme est en une base $b > 1$ arbitraire, où PV_M désigne l'ensemble de vecteurs de probabilité en dimension M.

Démonstration. Il n'est pas difficile de montrer que la fonction H_M donnée dans l'énoncé du théorème vérifie les 4 postulats précités. Nous devons montrer la réciproque. Nous observons que pour tout $a, b > 1$ et tout $x > 0$, on a $\log_a x = \log_a b \log_b x$. Par conséquent, nous pouvons travailler dans une base arbitraire de logarithmes car le changement de base peut être absorbé dans la constante C. La démonstration se fera en plusieurs étapes :

1. Pour tous les entiers $M, k \geq 1$ nous avons, par le postulat d'extensivité,
$$f(M^k) = f(M \cdot M^{k-1}) = f(M) + f(M^{k-1}),$$
ce qui, en itérant, mène à $f(M^k) = k f(M)$.
2. Nous allons montrer que pour $M \geq 1$ entier, il existe $C > 0$ telle que $f(M) = C \log M$. Pour $M = 1$, nous avons $f(1) = f(1 \cdot 1) = f(1) + f(1)$, donc $f(1) = 0$ qui vérifie donc l'égalité $f(M) = C \log M$. Supposons-la vraie pour $M > 1$ entier. Pour tout entier $r \geq 1$, il existe entier $k \geq 0$ tel que $M^k \leq 2^r \leq M^{k+1}$ (où l'une de deux inégalités d'encadrement est stricte). Nous avons
$$\begin{aligned} f(M^k) \leq f(2^r) \leq f(M^{k+1}) \quad &\text{(par le postulat de monotonie)} \\ k f(M) \leq r f(2) \leq (k+1) f(M) \quad &\text{(par le postulat d'extensivité)} \end{aligned}$$
ce qui entraîne l'encadrement $\frac{k}{r} \leq \frac{f(2)}{f(M)} \leq \frac{k+1}{r}$. Par ailleurs, la fonction log (pour une base $b > 1$ est strictement croissante. Des inégalités $M^k \leq 2^r \leq M^{k+1}$, nous avons donc aussi les encadrements $\frac{k}{r} \leq \frac{\log 2}{\log M} \leq \frac{k+1}{r}$. Par conséquent,
$$\left| \frac{\log 2}{\log M} - \frac{f(2)}{f(M)} \right| < \frac{1}{r}.$$
Puisque M est fixé tandis que r est arbitraire, nous avons
$$\frac{\log 2}{\log M} = \frac{f(2)}{f(M)} \Rightarrow f(M) = C \log M, \text{ où } C = \frac{f(2)}{\log 2}.$$
Par ailleurs $f(1) = 0$ tandis que f est strictement croissante, donc $f(2) > 0$; par conséquent $C > 0$.

3. Si $p \in \mathbb{Q}^+$, nous démontrerons que $H((p,1-p)) = -C[p\log p + (1-p)\log(1-p)]$. En effet, si nous écrivons $p = \frac{r}{s}$ avec des entiers $r, s \geq 1$, nous obtenons, en utilisant le postulat de regroupement,

$$f(s) = H_s\left((\underbrace{\frac{1}{s},\ldots,\frac{1}{s}}_{r},\underbrace{\frac{1}{s},\ldots,\frac{1}{s}}_{s-r})\right) = H_2\left((\frac{r}{s},\frac{s-r}{s})\right) + \frac{r}{s}f(r) + \frac{s-r}{s}f(s-r).$$

Nous avons donc, en utilisant ce que nous avons démontré en 2, que

$$f(s) = C\log s = H_2((p,1-p)) + Cp\log r + C(1-p)\log(s-r).$$

En résolvant, nous obtenons, pour $p \in \mathbb{Q}^+$, que

$$H_2((p,1-p)) = -C[p\log p + (1-p)\log(1-p)].$$

4. Nous utilisons maintenant le postulat de continuité pour établir que pour tout $p \in]0,1[$, nous avons $H((p,1-p)) = -C[p\log p + (1-p)\log(1-p)]$. Ceci est une conséquence immédiate du fait que tout réel $p \in]0,1[$ peut être approximé par une suite $(p_n)_n$ de rationnels $p_n \in]0,1[$ pour tout $n \in \mathbb{N}$.
5. Il reste à établir la formule pour H_M dans le cas général $\mathbf{p} = (p_1,\ldots,p_M)$. La formule $H_M(\mathbf{p}) = -C\sum_{i=1}^M p_i\log p_i$ est vraie pour $M = 1$ et $M = 2$. Pour $M > 2$, nous supposons la formule vraie jusqu'à l'ordre $M-1$ et nous utilisons le postulat de regroupement pour l'établir dans le cas M. En effet, en notant $q = p_1 + \ldots + p_{M-1}$ et en utilisant le postulat de regroupement, nous obtenons

$$\begin{aligned} H_M(\mathbf{p}) &= H_2((q,p_M)) + qH_{M-1}((\frac{p_1}{q},\ldots,\frac{p_{M-1}}{q})) + p_M H_1((1)) \\ &= -C\left[q\log q + p_M\log(p_M) + q\sum_{i=1}^{M-1}\frac{p_i}{q}\log\frac{p_i}{q}\right] \\ &= -C\sum_{i=1}^M p_i\log p_i. \end{aligned}$$

□

Remarque 7.1.6. Sauf indication contraire, nous utiliserons des logarithmes en base 2 et la constante de normalisation $C = 1$. La fonction H mesure alors la réduction de l'incertitude lorsque la valeur du tirage d'une pièce honnête nous est révélée en une unité appelée **bit**. D'autres choix sont possibles, par exemple exprimer le logarithme en base e et poser $C = 1$; l'unité correspondante est alors appelée **nat**. En Physique, on utilise la constante $C = 1.3806488 \times 10^{-23}$ J/K — appelée constante de Boltzmann — et on exprime le logarithme en base e ; l'unité de H est alors exprimée en J/K (Joules par degré Kelvin) et la quantité H est alors appelée **entropie**. Dans la suite de ce cours, log représentera le logarithme binaire ; nous utiliserons la notation ln pour le logarithme neperien.

Remarque 7.1.7. L'incertitude associée à l'ignorance des valeurs d'une variable aléatoire X avant que l'expérience soit effectuée et la réalisation nous soit relevée, ne dépend pas des valeurs possibles de X mais uniquement des probabilités avec lesquelles

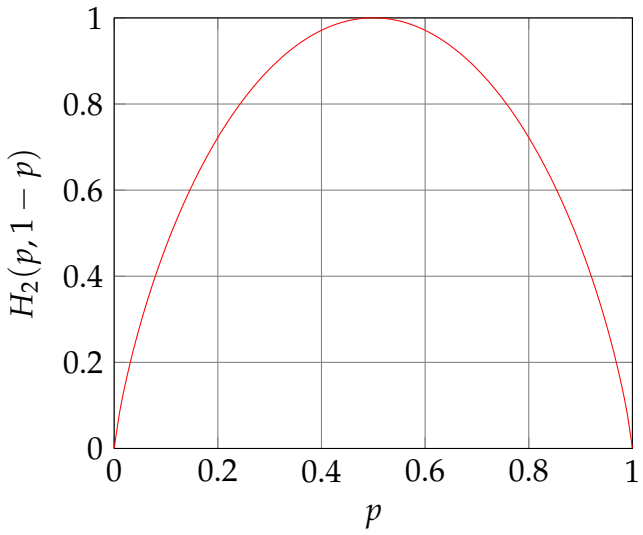


FIGURE 7.1 – La fonction H joue un rôle capital dans toute la suite de ce cours. Il est donc utile de garder à l'esprit le comportement qualitatif de $H_2(p, 1-p)$ en fonction de $p \in [0,1]$.

ces variables sont prises. Si par exemple $X \in \mathbb{X} = \{-1, 1\}$ correspond à une expérience dans laquelle vous perdez un euro si la pièce honnête tombe sur « pile » ou bien $X \in \mathbb{Y} = \{\text{exécution}, \text{libération}\}$ correspond à l'expérience où un prisonnier est exécuté si la pièce honnête tombe sur « pile », les deux expériences ont exactement la même incertitude même si les conséquences sont autrement graves pour le perdant.

Nous écrirons dans la suite indifféremment $H(X)$ ou $H(\mathbf{p})$ pour signifier l'entropie associée à la variable aléatoire X ou à sa loi.

Remarque 7.1.8. Dans ce qui précède, nous avons indiqué H_M pour signifier l'entropie pour une variable aléatoire prenant M valeurs distinctes. Dans la suite nous omettrons cette dépendance et nous considérerons la fonction H définie sur $\mathsf{PV} = \cup_{M=1}^{\infty} \mathsf{PV}_M$ où

$$\mathsf{PV}_M = \{\mathbf{p} := (p_1, \dots, p_M) \in \mathbb{R}_+^M : \sum_{i=1}^{M} p_i = 1\}.$$

En fait, H est une collection dénombrable de fonctions $(H_M)_{M \geq 1}$. Lorsque nous appliquons H sur un vecteur de probabilité $\mathbf{p} \in \mathsf{PV}$, en réalité nous appliquons H_M sur $\mathbf{p} \in \mathsf{PV}_M$ pour un certain M. Le résultat est alors calculé par application de la fonction H_M. Finalement, nous simplifierons la notation pour écrire $H(p_1, \dots, p_M)$ au lieu de $H_M((p_1, \dots, p_M))$.

Définition 7.1.9. Soit $\mathbf{p} \in \mathsf{PV}$ un vecteur de probabilité. La quantité

$$H(\mathbf{p}) := H_{\dim(\mathbf{p})}(\mathbf{p}) = -\sum_{i=1}^{\dim(\mathbf{p})} p_i \log p_i,$$

est appelée **entropie** ou (quantité d')**information** associée au vecteur $\mathbf{p}$.

La formule de la définition 7.1.9 est souvent attribuée à Shannon [1][62] dans pratiquement tous les livres consacrés à la théorie de l'information. Sans vouloir minimiser

1. Claude Elwood Shannon, Michigan 1916 – Massachussets 2001, ingénieur électrique et mathématicien américain, fondateur de la théorie de l'information.

l'immense apport de Shannon en théorie de l'information, il est cependant nécessaire, afin de rétablir la vérité historique, de rappeler que si l'article de Shannon a effectivement donné pour la première fois les propriétés mathématiques de l'information, elles se trouvent être identiques à celles d'une autre quantité — physique — découverte un siècle plus tôt (1856) par Rudolph Julius Emmanuel Clausius [2] et formalisée par Ludwig Eduard Boltzmann [3] [9] (cf. figure 7.2) et Josiah Willard Gibbs [4] [30], connue sous le nom d'**entropie** [5].

L'influence de la théorie des probabilités et de la thermodynamique est explorée dans des textes classiques [1, 17, 28, 41, 50, 48] (mentionnons aussi l'article de revue [46] ainsi que dans le livre récent [38]), textes desquels ces notes sont librement inspirées. Plusieurs travaux récents (cf. [5, 27, 54]) empruntent le chemin inverse en examinant les conséquences qu'ont en thermodynamique les notions introduites en théorie de l'information.

7.2 Trois interprétations de l'entropie

7.2.1 H est une espérance (qui nous fait vieillir !)

Soit $\mathbf{p} \in \mathsf{PV}_M$ un vecteur de probabilité. On définit une variable aléatoire $\xi = -\log p(X)$ sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ prenant de valeurs dans l'alphabet

$$\mathbb{Y} = \{-\log p_1, \dots, -\log p_M\}$$

avec loi $\mathbb{P}(Y = -\log p_i) = p_i$, pour $i = 1, \dots, M$. Alors

$$\mathbb{E}\xi = \mathbb{E}\log p(X) = -\sum_{i=1}^{M} p_i \log p_i = H(\mathbf{p}).$$

L'entropie est donc égale à l'espérance de la variable aléatoire ξ, définie ci-dessus.

7.2.2 H est le nombre moyen de questions nécessaires pour déterminer la valeur que prend une variable aléatoire

Nous commençons par l'exemple d'une variable aléatoire X prenant de valeurs dans $\mathbb{X} = \{x_1, \dots, x_5\}$ selon un vecteur de probabilité est $\mathbf{p} = (0.3, 0.2, 0.2, 0.15, 0.15)$.

2. Physicien allemand, Koszalin 1822 – Bonn 1888, connu pour ses contributions en thermodynamique.

3. Physicien autrichien, Vienne 1844 – Trieste 1906. Fondateur de la physique statistique et défenseur de la théorie atomique qui à son temps n'était pas encore acceptée comme théorie physique. Il a par ailleurs introduit l'hypothèse ergodique, idée selon laquelle sous certaines conditions on peut remplacer des moyennes temporelles de suites stationnaires par des moyennes spatiales, une des idées les plus fructueuses en physique mathématique ; elle a donné naissance à toute une sous-discipline mathématique connue aujourd'hui sous le nom de **théorie ergodique**. Dans son livre *Vorlesungen über Gastheorie* [9], il a introduit la fonction H (cf. définition 7.1.9 et facsimilé de la page 41 de ce livre, en figure 7.2) pour expliquer l'irréversibilité macroscopique de la physique régie microscopiquement par des équations différentielles réversibles.

4. Physicien théoricien, chimiste et mathématicien américain, 1839 – 1903. Il a posé les fondements mathématiques de la thermodynamique chimique et de la mécanique statistique.

5. John von Neumann a signalé par ailleurs à Shannon que la quantité d'information introduite par ce dernier était déjà connue sous le nom d'entropie par Boltzmann et c'est von Neumann qu'a suggéré à Shannon de garder le nom d'entropie.

[Gleich. 35] § 6. Math. Bedeutung der Grösse *H*. 41

andere Permutation möglich. Viel wahrscheinlicher schon wäre der Fall, dass die Hälfte der Moleküle eine bestimmte, bestimmt gerichtete, die andere Hälfte eine andere, wieder für alle gleiche und gleichgerichtete Geschwindigkeit hätten. Dann wäre die Hälfte der Geschwindigkeitspunkte in einer, die andere Hälfte in einer zweiten Zelle; es wäre also:

$$Z = \frac{n!}{\left(\frac{n}{2}\right)!\left(\frac{n}{2}\right)!} \text{ u. s. w.}$$

Da nun die Anzahl der Moleküle eine überaus grosse ist, so sind $n_1\,\omega$, $n_2\,\omega$ u. s. w. ebenfalls als sehr grosse Zahlen zu betrachten.

Wir wollen die Annäherungsformel:

$$p! = \sqrt{2\,p\,\pi}\left(\frac{p}{e}\right)^p$$

benützen, wobei e die Basis der natürlichen Logarithmen und p eine beliebige grosse Zahl ist.[1])

Bezeichnen wir daher wieder mit l den natürlichen Logarithmus, so folgt:

$$l[(n_1\,\omega)!] = (n_1\,\omega + \tfrac{1}{2})\,l\,n_1 + n_1\,\omega\,(l\,\omega - 1) + \tfrac{1}{2}(l\,\omega + l\,2\,\pi).$$

Vernachlässigt man hier $\frac{1}{2}$ gegen die sehr grosse Zahl $n_1\,\omega$ und bildet den analogen Ausdruck für $(n_2\,\omega)!$, $(n_3\,\omega)!$ u. s. f., so ergibt sich:

$$l\,Z = -\,\omega\,(n_1\,l\,n_1 + n_2\,l\,n_2 \ldots) + C,$$

wobei

$$C = l(n!) - n\,(l\,\omega - 1) - \frac{\zeta}{2}(l\,\omega + l\,2\,\pi)$$

für alle Geschwindigkeitsvertheilungen denselben Werth hat, also als Constante zu betrachten ist. Denn wir fragen ja bloss nach der relativen Wahrscheinlichkeit der Eintheilung der verschiedenen Geschwindigkeitspunkte unserer Moleküle in unsere Zellen ω, wobei selbstverständlich die Zelleneintheilung, daher auch die Grösse einer Zelle ω, die Anzahl der Zellen ζ und die Gesammtzahl n der Moleküle und deren gesammte lebendige Kraft als unveränderlich gegeben betrachtet werden müssen. Die wahrscheinlichste Eintheilung der Geschwindig-

[1]) Siehe Schlömilch, Comp. der höh. Analysis. Bd. 1. S. 437. 3. Aufl.

FIGURE 7.2 – Facsimilé de la page 41 du livre de Boltzmann *Vorlesungen über Gastheorie* [9], où une définition mathématique de la fonction entropie, **identique** à la définition 7.1.9, est donnée pour la première fois, quoiqu'obtenue par une méthode différente de celle utilisée par Shannon. La formule correspondante est encadrée en rouge dans le texte original de Boltzmann qui est reproduit dans la figure ci-dessus (le logarithme népérien est noté l). Ce livre a été traduit en français [10].

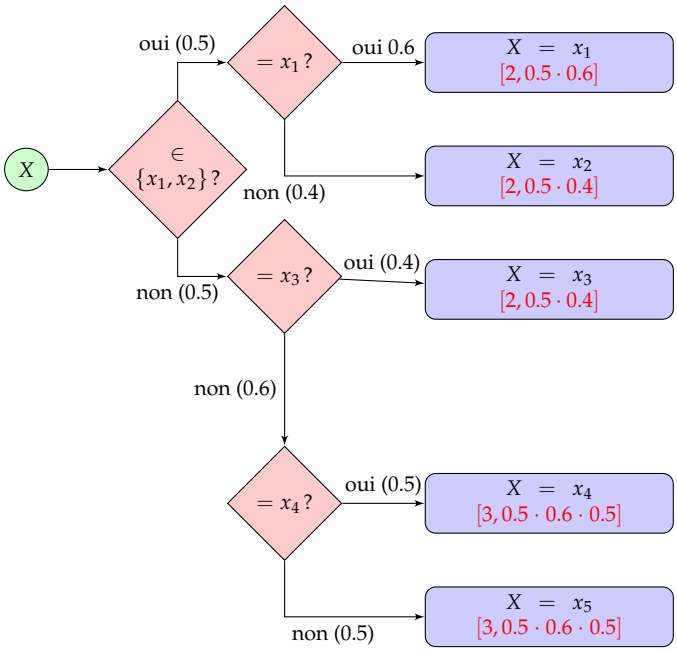


FIGURE 7.3 – Le diagramme logique permettant de déterminer la valeur de X. Les étiquettes entre parenthèse sur les arrêtes du diagramme désignent les probabilités conditionnelles de l'arbre de décisions sachant à quel embranchement on se trouve. Par exemple, l'étiquette en vert sur le diagramme ci-dessus, correspond à la probabilité conditionnelle $\mathbb{P}(X \neq x_3 | X \notin \{x_1, x_2\})$. Les étiquettes de la forme $[N, p]$ entre crochets désignent les nombres de questions posées N et la probabilité de la feuille de l'arbre p.

Le diagramme logique — donné en figure 7.3 — permet de déterminer les valeurs de la variable aléatoire X en posant un certain nombre de questions admettant comme réponse soit « oui » soit « non ».

Le díagramme correspond à un arbre de décision ; chaque feuille de l'arbre (détermination non ambiguë de la valeur de X) correspond à un chemin sur l'arbre muni d'une probabilité. Par ailleurs chaque chemin qui mène de la racine à une feuille de l'arbre correspond à un certain nombre de questions posées, noté N, (c'est le nombre de losanges rencontrés le long du chemin). Calculons $\mathbb{E}(N)$ pour l'exemple ci-dessus :

$$\begin{aligned} \mathbb{E}(N) &= 2 \cdot [0.5 \cdot 0.6 + 0.5 \cdot 0.4 + 0.5 \cdot 0.4] + 3 \cdot [0.5 \cdot 0.6 \cdot 0.5 + 0.5 \cdot 0.6 \cdot 0.5] \\ &= 2 \cdot [0.3 + 0.2 + 0.2] + 3 \cdot [0.15 + 0.15] \\ &= 2.3. \end{aligned}$$

Par ailleurs, nous pouvons calculer l'entropie de $\mathbf{p}$:

$$H(\mathbf{p}) = -0.3 \log 0.3 - 0.4 \log 0.2 - 0.3 \log 0.15 = 2.27.$$

La proximité des valeurs numériques de $\mathbb{E}(N)$ et de $H(\mathbf{p})$ est très frappante. Nous verrons dans les chapitres suivants qu'*il n'existe pas de schéma* permettant de déterminer la valeur de X en posant des questions binaires dont le nombre moyen serait plus petit que $H(\mathbf{p})$.

Le vecteur de probabilité $\mathbf{p}$ dans l'exemple ci-dessus est tel que son entropie $H(\mathbf{p})$ et l'espérance du nombre de questions $\mathbb{E}(N)$ sont très proches. Ceci n'est pas toujours le cas. Considérons en effet l'exemple d'une variable aléatoire à valeurs dans $\mathbb{X} = \{x_1, x_2\}$ dont la loi est décrite par le vecteur de probabilité $\mathbf{p} = (0.7, 0.3)$. On

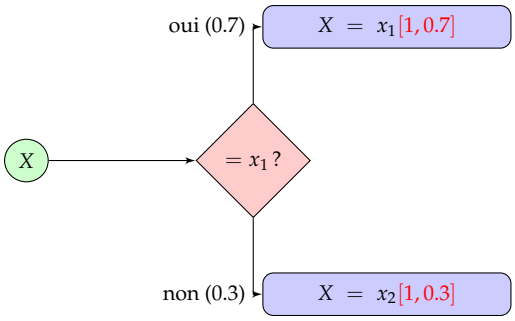


FIGURE 7.4 – Le diagramme logique permettant de déterminer la valeur de X avec les mêmes conventions de notation que pour la figure 7.3.

observe que dans ce cas $H(\mathbf{p}) = 0.88$ tandis que $\mathbb{E}(N) = 1$ et il semble que nous ne pouvons pas faire mieux que cela. En fait, il est possible de faire mieux. Supposons en effet qu'au lieu de considérer une variable aléatoire X, nous considérons une paire $Z = (X_1, X_2)$ de deux copies indépendantes de X, i.e. la variable aléatoire Z prend de valeurs dans $\{(x_1, x_1), (x_1, x_2), (x_2, x_1), (x_2, x_2)\}$; sa loi sera décrite par le vecteur de probabilité $\mathbf{q} = (0.49, 0.21, 0.21, 0.09)$. On peut construire facilement un schéma de décision des valeurs que prend la variable aléatoire Z avec $\mathbb{E}(N) = 1.81$. (Construisez ce schéma). Cependant, ce schéma permet de déterminer une paire indépendante de variables aléatoires X, donc l'espérance du nombre de questions par variable aléatoire est 0.905, plus proche de l'entropie que la valeur 1 déterminée précédemment. Nous pouvons continuer ce procédé avec des triplets, quadruplets, etc. de variables aléatoires ; chaque fois nous obtiendrons une espérance du nombre de questions par variable qui s'approche de $H(\mathbf{p})$. L'exercice 86e montre que l'espérance du nombre de questions par variable converge en décroissant vers l'entropie de la loi de l'une d'entre elles mais il n'existe pas de schéma permettant de faire mieux que la barrière de l'entropie.

7.2.3 H est le rapport des logarithmes du volume des configurations typiques sur celui de toutes les configurations

Soit X une variable aléatoire prenant de valeurs[6] dans un alphabet fini $\mathbb{A}$, dont la loi est décrite par le vecteur de probabilité $\mathbf{p} := (p_a)_{a\in\mathbb{A}}$. Pour tout $n \in \mathbb{N}$, nous considérons la variable aléatoire $\mathbf{X} := \mathbf{X}^{(n)} := (X_1, \dots, X_n)$, composée de n copies indépendantes de la variable aléatoire X et qui prend des valeurs dans $\mathbb{A}^n$. Pour tout mot de n lettres $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_n) \in \mathbb{A}^n$ et toute lettre $a \in \mathbb{A}$, nous notons $\nu_a(\boldsymbol{\alpha}) := \nu_a^{(n)}(\boldsymbol{\alpha}) = \sum_{i=1}^n \mathbb{1}_{\{a\}}(\alpha_k)$. À cause de l'indépendance des variables aléatoires $(X_i)_{i=1,\dots,n}$, il est évident que

$$\begin{aligned} \mathbb{P}\left(\mathbf{X}^{(n)} = \boldsymbol{\alpha}\right) &= \prod_{a\in\mathbb{A}} p_a^{\nu_a(\boldsymbol{\alpha})}, \quad \forall \boldsymbol{\alpha} \in \mathbb{A}^n \\ \mathbb{P}\left(\nu_a^{(n)}(\mathbf{X}^{(n)}) = l\right) &= C_n^l p_a^l (1-p_a)^{n-l}, \quad \forall a \in \mathbb{A},\ l = 0, \dots, n. \end{aligned}$$

Remarque 7.2.1. Il est nettement plus aisé de travailler avec des suites infinies $\boldsymbol{\alpha} \in \mathbb{A}^{\mathbb{N}}$ ou avec des suites aléatoires infinies $\mathbf{X} = (X_1, X_2, \dots)$ et de définir $\nu_a^{(n)}(\boldsymbol{\alpha}) =$

6. Si $M = \text{card}\mathbb{A}$, sans perte de généralité, nous pouvons supposer que $\mathbb{A} \simeq \{1, \dots, M\}$ et que $\mathbf{p} = (p_i)_{i=1,\dots,M}$.

$\sum_{k=1}^{n} \mathbb{1}_{\{a\}}(\alpha_k)$ par la somme tronquée aux n premières lettres de la suite. De même, nous pouvons utiliser la notation $\boldsymbol{\alpha}\restriction_n$ ou $\mathbf{X}\restriction_n$ pour signifier la troncature des suites aux n premières lettres. On peut aussi remarquer que pour tout $n \in \mathbb{N}_>$ et $\boldsymbol{\alpha} \in \mathbb{A}^{\mathbb{N}}$, on a : $\frac{1}{n}\boldsymbol{\nu}^{(n)}(\boldsymbol{\alpha}) \in \mathsf{PV}_{\mathbb{A}}$; ce vecteur de probabilité est appelé **type** de la suite $\boldsymbol{\alpha}$.

Nous calculons aisément que $\mathbb{E}(\nu_a^{(n)}(\mathbf{X})) = np_a$, pour tout $a \in \mathbb{A}$. Il est donc intuitivement clair pourquoi nous définissons :

Définition 7.2.2. Soient un entier $n \geq 1$, un alphabet fini $\mathbb{A}$, un vecteur de probabilité $\mathbf{p} \in \mathsf{PV}_{\mathsf{card}\mathbb{A}}$ et un entier $K > 0$. Une suite $\boldsymbol{\alpha} \in \mathbb{A}^n$ est dite **typique** (plus précisément $(n, \mathbf{p}, K)$-typique) si

$$\forall a \in \mathbb{A}, \left| \frac{\nu_a^{(n)}(\boldsymbol{\alpha}) - np_a}{\sqrt{np_a(1-p_a)}} \right| < K,$$

sinon, elle est appelée **atypique**. L'ensemble

$$\mathbb{T}_{n,\mathbf{p},K} := \{\boldsymbol{\alpha} \in \mathbb{A}^n : \boldsymbol{\alpha} \text{ est } (n, \mathbf{p}, K)\text{-typique}\} \subset \mathbb{A}^n$$

désigne l'ensemble de suites $(n, \mathbf{p}, K)$-typiques.

Remarque 7.2.3. Dans une suite $\boldsymbol{\alpha}$, typique pour un vecteur de probabilité $\mathbf{p}$, nous avons : $|\frac{\nu_a^{(n)}(\boldsymbol{\alpha})}{n} - p_a| < \frac{K}{\sqrt{p_a(1-p_a)}}\frac{1}{\sqrt{n}} = \mathcal{O}(n^{-1/2})$, pour toute lettre $a \in \mathbb{A}$. *Il faut souligner que les suites typiques dépendent du vecteur de probabilité* $\mathbf{p}$ *mais ne sont pas elles-mêmes aléatoires.* Il s'agit tout simplement d'une famille de mots à n lettres sur l'alphabet $\mathbb{A}$ dont les lettres apparaissent avec une densité pré-fixée (définie par $\mathbf{p}$).

Théorème 7.2.4. *Soient* $\varepsilon \in]0,1[$ *et* $K > \lceil \sqrt{\frac{\mathsf{card}\mathbb{A}}{\varepsilon}} \rceil$. *Pour tout* $n \geq K$,

1. $\mathbb{P}(\mathbf{X}\restriction_n \notin \mathbb{T}_{n,\mathbf{p},K}) < \varepsilon$;
2. $\exists c > 0$ *tel que* $\forall \boldsymbol{\alpha} \in \mathbb{T}_{n,\mathbf{p},K}$, *nous avons*

$$2^{-nH(\mathbf{p})-c\sqrt{n}} \leq \mathbb{P}(\mathbf{X}\restriction_n = \boldsymbol{\alpha}) \leq 2^{-nH(\mathbf{p})+c\sqrt{n}};$$

3. $\mathsf{card}(\mathbb{T}_{n,\mathbf{p},K}) = 2^{n(H(\mathbf{p})+\delta_n)}$, *avec* $\lim_{n\to\infty} \delta_n = 0$.

Démonstration. 1. Pour alléger la notation, écrivons $\mathbf{X}$ au lieu de $\mathbf{X}\restriction_n$.

$$\begin{aligned} \mathbb{P}(\mathbf{X} \notin \mathbb{T}_{n,\mathbf{p},K}) &= \mathbb{P}\left(\bigcup_{a\in\mathbb{A}} \left\{ \left| \frac{\nu_a^{(n)}(\mathbf{X}) - np_a}{\sqrt{np_a(1-p_a)}} \right| \geq K \right\} \right) \\ &\leq \sum_{a\in\mathbb{A}} \mathbb{P}\left(\left| \frac{\nu_a^{(n)}(\mathbf{X}) - np_a}{\sqrt{np_a(1-p_a)}} \right| \geq K \right). \end{aligned}$$

L'inégalité de Bienaymé-Tchebychev nous permet d'estimer les termes individuels dans la dernière expression par

$$\mathbb{P}\left(\left| \frac{\nu_a^{(n)}(\mathbf{X}) - np_a}{\sqrt{np_a(1-p_a)}} \right| \geq K \right) \leq \frac{\mathbb{E}(|\nu_a^{(n)}(\mathbf{X}) - np_a|^2)}{K^2 np_a(1-p_a)} = \frac{1}{K^2} < \frac{\varepsilon}{\mathsf{card}\mathbb{A}},$$

ce qui nous permet de conclure que $\mathbb{P}(\mathbf{X} \notin \mathbb{T}_{n,\mathbf{p},K}) \leq \frac{\varepsilon}{\mathsf{card}\mathbb{A}} \times \mathsf{card}\mathbb{A} = \varepsilon$.

2. Pour toute $\boldsymbol{\alpha} \in \mathbb{A}^n$, nous avons à cause de l'indépendance des variables aléatoires (X_i) que
$$-\log \mathbb{P}(\mathbf{X} = \boldsymbol{\alpha}) = -\sum_{a \in \mathbb{A}} \nu_a^{(n)}(\boldsymbol{\alpha}) \log p_a.$$
Or, pour $\boldsymbol{\alpha} \in \mathbb{T}_{n,\mathbf{p},K}$, nous avons, pour tout $a \in \mathbb{A}$,
$$np_a - K\sqrt{np_a(1-p_a)} \leq \nu_a^{(n)}(\boldsymbol{\alpha}) \leq np_a + K\sqrt{np_a(1-p_a)}.$$
Remplaçant les encadrements de $\nu_a^{(n)}(\boldsymbol{\alpha})$ obtenus dans la dernière formule dans la pénultième et en définissant $c := -K\sum_{a\in\mathbb{A}} \sqrt{p_a(1-p_a)}\log p_a$, nous concluons.
3. Par l'affirmation 1 du théorème, nous avons
$$1 - \varepsilon \leq \mathbb{P}(\mathbf{X} \in \mathbb{T}_{n,\mathbf{p},K}) \leq 1.$$
À partir de l'affirmation 2, en combinant avec la relation précédente, nous obtenons
$$1 - \varepsilon < \mathbb{P}(\mathbf{X} \in \mathbb{T}_{n,\mathbf{p},K}) = \sum_{\boldsymbol{\alpha} \in \mathbb{T}_{n,\mathbf{p},K}} \mathbb{P}(\mathbf{X} = \boldsymbol{\alpha}) \leq 2^{-nH(\mathbf{p})+c\sqrt{n}} \mathsf{card}(\mathbb{T}_{n,\mathbf{p},K}),$$
permettant de minorer $\mathsf{card}(\mathbb{T}_{n,\mathbf{p},K}) \geq 2^{nH(\mathbf{p})-c\sqrt{n}+\log(1-\varepsilon)}$. Par ailleurs,
$$\mathsf{card}(\mathbb{T}_{n,\mathbf{p},K}) 2^{-nH(\mathbf{p})-c\sqrt{n}} \leq \sum_{\boldsymbol{\alpha} \in \mathbb{T}_{n,\mathbf{p},K}} \mathbb{P}(\mathbf{X} = \boldsymbol{\alpha}) \leq 1,$$
permettant de majorer $\mathsf{card}(\mathbb{T}_{n,\mathbf{p},K}) \leq 2^{nH(\mathbf{p})+c\sqrt{n}}$. En combinant les deux bornes nous obtenons
$$2^{n(H(\mathbf{p})-\delta'_n)} \leq \mathsf{card}(\mathbb{T}_{n,\mathbf{p},K}) \leq 2^{n(H(\mathbf{p})+\delta_n)},$$
où, $\delta_n = \frac{c}{\sqrt{n}} \to 0$ et $\delta'_n = \frac{c}{\sqrt{n}} - \frac{\log(1-\varepsilon)}{n} \to 0$.

□

Remarque 7.2.5. Le théorème précédent établit les faits importants suivants. L'affirmation 1, établit que si $\mathbf{p}$ n'est pas l'équidistribution sur $\mathbb{A}$, parmi les $[\mathsf{card}(\mathbb{A})]^n$ suites possibles une *proportion exponentiellement petite* (de cardinal $2^{nH(\mathbf{p})\pm c\sqrt{n}}$) supporte pratiquement toute la masse de $\mathbb{P}$. L'affirmation 2 établit que dans ce petit ensemble de configurations typiques, toutes ont essentiellement la même probabilité, i.e. il y a équidistribution sur l'ensemble de configurations typiques.

7.3 Propriétés de la fonction entropie, entropie relative

Il est évident que $H(\mathbf{p}) \geq 0$ car $-p_a \log p_A \geq 0$ pour tout $a \in \mathbb{A}$.

Lemme 7.3.1. *Soient $\mathbf{p}, \mathbf{q} \in PV_{\mathsf{card}\mathbb{A}}$, deux vecteurs de probabilité arbitraires. Alors*
$$-\sum_{a\in\mathbb{A}} p_a \log p_a \leq -\sum_{a\in\mathbb{A}} p_a \log q_a.$$

Démonstration. La fonction $t \mapsto \ln t$ est concave sur $\mathbb{R}_+$, i.e. le graphe de cette fonction est en dessous de la tangente à n'importe quel point du graphe. Nous avons $\ln 1 = 0$ et $(\ln t)'|_{t=1} = \frac{1}{t}|_{t=1} = 1$, par conséquent, la tangente qui passe par le graphe de la fonction ln au point 1 a comme équation $t - 1$. La concavité devient donc $\ln t \leq t - 1$ pour tout $t > 0$, avec égalité si et seulement si $t = 1$. Nous aurons donc $\ln \frac{q_a}{p_a} \leq \frac{q_a}{p_a} - 1$ (avec égalité si et seulement si $q_a = p_a$). Nous aurons donc

$$\sum_{a\in\mathbb{A}} p_a \ln \frac{q_a}{p_a} \leq \sum_{a\in\mathbb{A}} p_a(\frac{q_a}{p_a} - 1) = \sum_{a\in\mathbb{A}} (p_a - q_a) = 0.$$

□

Théorème 7.3.2. *Pour tout* $\mathbf{p} \in PV_{\mathsf{card}\mathbb{A}}$,

$$H(\mathbf{p}) \leq \log \mathsf{card}\mathbb{A},$$

avec égalité si et seulement si $p_a = \frac{1}{\mathsf{card}\mathbb{A}}$ *pour tout* $a \in \mathbb{A}$.

Démonstration. Appliquer le lemme au vecteur de probabilité $\mathbf{q}$ uniforme sur $\mathbb{A}$. Nous aurons $H(\mathbf{p}) \leq \log \mathsf{card}\mathbb{A}$ et la borne sera saturée pour $\mathbf{p} = \mathbf{q}$. □

Pour deux vecteurs de probabilité $\mathbf{p}$ et $\mathbf{q}$ sur le même alphabet $\mathbb{A}$, on dit que $\mathbf{p}$ est **absolument continu** par rapport à $\mathbf{q}$, et l'on note $\mathbf{p} \ll \mathbf{q}$, si $p_a = 0$ pour tout $a \in \mathbb{A}$ tel que $q_a = 0$, i.e. si $q_a = 0$ implique que $p_a = 0$.

Définition 7.3.3. Soient $\mathbf{p}$ et $\mathbf{q}$ deux vecteurs de probabilité sur le même espace $\mathbb{A}$. On appelle **entropie relative** ou **contraste de Kullback-Leibler** de $\mathbf{p}$ par rapport à $\mathbf{q}$ la quantité

$$D(\mathbf{p}\|\mathbf{q}) := \begin{cases} \sum_{a\in\mathbb{A}} p_a \log\left(\frac{p_a}{q_a}\right) & \text{si } \mathbf{p} \ll \mathbf{q} \\ +\infty & \text{sinon.} \end{cases}$$

L'exercice 82 établit les propriétés importantes de D. En particulier, il permet de montrer que $D(\mathbf{p}\|\mathbf{q}) \geq 0$, pour $\mathbf{p}$ et $\mathbf{q}$ arbitraires.

Remarque 7.3.4. La fonction D n'est pas symétrique en ses arguments. Cependant, plus la valeur $D(\mathbf{p}\|\mathbf{q})$ est grande, plus il est facile de discriminer entre les probabilités $\mathbf{p}$ et $\mathbf{q}$. De cette remarque découle la grande utilité de D en statistique. La définition 7.3.5 et la proposition 7.3.6 illustrent le pouvoir discriminant de D.

Définition 7.3.5. Soient $(\mathbb{X}, \mathbf{p})$ et $(\mathbb{Y}, \mathbf{q})$ deux espaces probabilisés finis (munis de leurs tribus exhaustives) avec $\mathbb{X} \subseteq \mathbb{Y}$. On dit que $(\mathbb{Y}, \mathbf{q})$ est une **fragmentation** de $(\mathbb{X}, \mathbf{p})$ (ou que $(\mathbb{X}, \mathbf{p})$ est une **coalescence** de $(\mathbb{Y}, \mathbf{q})$) si on peut partitioner $\mathbb{Y} = \sqcup_{x\in\mathbb{X}} \mathbb{Y}_x$, de sorte que pour tout $x \in \mathbb{X}$, on a $p(x) = \sum_{y\in\mathbb{Y}_x} q(y)$.

Proposition 7.3.6. *Pour* $i = 0, 1$, *on suppose que* $(\mathbb{Y}, \mathbf{q}_i)$ *sont des fragmentations de* $(\mathbb{X}, \mathbf{p}_i)$. *Alors*

$$D(\mathbf{q}_0\|\mathbf{q}_1) \geq D(\mathbf{p}_0\|\mathbf{p}_1),$$

en d'autres termes, la fragmentation augmente le contraste de Kullback-Leibler.

Démonstration. Sans perte de généralité, on suppose que $\mathbf{q}_0 \ll \mathbf{q}_1$. On a

$$\begin{aligned} D(\mathbf{q}_0\|\mathbf{q}_1) - D(\mathbf{p}_0\|\mathbf{p}_1) &= \sum_{y\in\mathbb{Y}} q_0(y)\log\frac{q_0(y)}{q_1(y)} - \sum_{x\in\mathbb{X}} p_0(x)\log\frac{p_0(x)}{p_1(x)} \\ &= \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}_x}\left(q_0(y)\log\frac{q_0(y)}{q_1(y)} - q_0(y)\log\frac{p_0(x)}{p_1(x)}\right) \\ &= \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}_x} q_0(y)\log\frac{q_0(y)p_1(x)}{q_1(y)p_0(x)} \\ &\geq \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}_x}\left(q_0(y) - q_0(y)\frac{q_1(y)p_0(x)}{q_0(y)p_1(x)}\right) \ (\text{car } \log t \geq 1-\frac{1}{t}) \\ &= 0. \end{aligned}$$

□

7.4 Entropie des évolutions markoviennes

Soit $(X_t)_{t\in\mathbb{N}}$ une chaîne de Markov irréductible et apériodique sur un espace dénombrable $\mathbb{X}$ (muni de sa tribu exhaustive $\mathcal{X}$) et de matrice stochastique P. On note π sa probabilité d'équilibre, vérifiant $\pi = \pi P$ et $\mu_n(y) := \mathbb{P}_\rho(X_t = y)$ pour une probabilité initiale $\rho \in \mathcal{M}_1(\mathbb{X})$ fixée.

Supposons que $f : \mathbb{R}_+ \to \mathbb{R}$ soit une fonction mesurable strictement concave et, pour $n \in \mathbb{N}$,

$$F_n = \sum_{y\in\mathbb{X}} \pi(y) f\left(\frac{\mu_n(y)}{\pi(y)}\right).$$

Théorème 7.4.1. *Sous le conditions ci-dessus, la suite des fonctions F_n est strictement croissante en n.*

Démonstration. La chaîne étant irréductible et apériodique, sa mesure invariante π charge nécessairement tous les états $y \in \mathbb{X}$. Le rapport $u_n(y) := \frac{\mu_n(y)}{\pi(y)}$ est donc bien défini pour tout $n \in \mathbb{N}$ et tout $y \in \mathbb{Y}$. En notant, comme d'habitude par $\hat{P}(x,y) :=$

$\frac{\pi(y)P(y,x)}{\pi(x)}$ le noyau de la chaîne renversée dans le temps, nous avons

$$\begin{aligned}
u_{n+1}(x) &= \frac{\mu_{n+1}(y)}{\pi(y)} \\
&= \sum_{z\in\mathbb{X}} \mu_n(z)\frac{P(z,x)}{\pi(x)} \\
&= \sum_{z\in\mathbb{X}} \frac{\mu_n(z)}{\pi(z)}\pi(z)\frac{P(z,x)}{\pi(x)} \\
&= \sum_{z\in\mathbb{X}} \hat{P}(x,z)u_n(z).
\end{aligned}$$

$$\begin{aligned}
f(u_{n+1}(x)) &= f\left(\sum_{z\in\mathbb{X}} \hat{P}(x,z)u_n(z)\right) \\
&> \sum_{z\in\mathbb{X}} \hat{P}(x,z)f(u_n(z)) \text{ (à cause de la stricte concavité de } f).
\end{aligned}$$

$$\begin{aligned}
F_{n+1} &= \sum_{x\in\mathbb{X}} \pi(x)f(u_{n+1}(x)) \\
&> \sum_{x,z\in\mathbb{X}} \pi(x)\frac{\pi(z)P(z,x)}{\pi(x)}f(u_n(z)) \\
&= \sum_{z\in\mathbb{X}} \pi(z)f(u_n(z)) \text{ (car } \sum_x P(z,x) = 1\text{).} \\
&= F_n.
\end{aligned}$$

□

Corollaire 7.4.2. *Sous les mêmes conditions, le contraste de Kullback-Leibler entre μ_n et π est strictement décroissant en n. Il s'annule lorsque μ_n devient égale à la probabilité d'équilibre π.*

Démonstration. La fonction $f := t \mapsto -t\log t$ est strictement concave, par conséquent, pour cette fonction f,

$$F_n = \sum_{x\in\mathbb{X}} \pi(x)f\left(\frac{\mu_n(x)}{\pi(x)}\right) = -\sum_{x\in\mathbb{X}} \pi(x)\frac{\mu_n(x)}{\pi(x)}\log\frac{\mu_n(x)}{\pi(x)} = -D(\mu_n\|\pi).$$

Puisque $D(\mu_n\|\pi) = -F_n$, le contraste de Kullback-Leibler est strictement décroissante en n et, par ailleurs, minoré par 0. Donc la suite $(D(\mu_n\|\pi))_n$ admet une limite. On en déduit que $\lim_{n\to\infty} D(\mu_n\|\pi) = 0$ lorsque μ_n converge vers la mesure d'équilibre π. □

En d'autres termes, le contraste de Kullback-Leibler entre la mesure μ_n — qui garde la trace de la condition initiale — et la mesure d'équilibre π s'amenuise avec le temps. Lorsque le contraste s'annule, la chaîne a atteint son état d'équilibre. Si l'initialisation se fait avec π (au lieu de la mesure arbitraire ρ), on a stationnarité de la chaîne, i.e. $\mu_n = \pi$ pour tout n et, dans ce cas, $D(\mu_n\|\pi) = 0$.

7.5 Couples de variables aléatoires

7.5.1 Entropie conjointe

Si X et Y sont deux variables aléatoires définies sur le même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et prennent leurs valeurs dans les alphabets finis respectivement $\mathbb{X}$ et $\mathbb{Y}$, la loi du couple (X, Y) est la loi conjointe déterminée par le vecteur $\boldsymbol{\kappa}$ de probabilité conjointe $\kappa(x, y) := \mathbb{P}(X = x; Y = y)$, pour $x \in \mathbb{X}$ et $y \in \mathbb{Y}$. Il est donc naturel de définir

Définition 7.5.1. Soient X et Y deux variables aléatoires définies sur le même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans les alphabets finis $\mathbb{X}$ et $\mathbb{Y}$. On note $\boldsymbol{\kappa}$ le vecteur de probabilité de leur loi conjointe. Nous définissons l'**entropie conjointe**

$$H(X, Y) := H(\boldsymbol{\kappa}) = -\sum_{(x,y)\in\mathbb{X}\times\mathbb{Y}} \kappa(x, y) \log \kappa(x, y) = -\mathbb{E}(\log \kappa(X, Y)).$$

De manière similaire, si $\mathbf{X} = (X_1, \ldots, X_n)$ est un vecteur aléatoire de loi conjointe décrite par le vecteur $\boldsymbol{\kappa}$, i.e. pour $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{X}_1 \times \cdots \times \mathbb{X}_n$ on a $\kappa(\mathbf{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$, l'entropie conjointe est définie par

$$H(X_1, \ldots, X_n) = -\sum_{\mathbf{x}\in\mathbb{X}_1\times\cdots\times\mathbb{X}_n} \mathbb{P}(\mathbf{X} = \mathbf{x}) \log \mathbb{P}(\mathbf{X} = \mathbf{x}) = -\mathbb{E}(\log \kappa(\mathbf{X})).$$

Théorème 7.5.2. *L'entropie est une quantité sous-additive, i.e.*

$$H(X, Y) \leq H(X) + H(Y),$$

avec égalité si, et seulement si, les variables aléatoires X et Y sont indépendantes.

Démonstration. Soit $\boldsymbol{\kappa}$ le vecteur de probabilité conjointe. Notons $\mu(x) = \sum_{y\in\mathbb{Y}} \kappa(x, y)$ le vecteur de probabilité de la première marginale, $\nu(y) = \sum_{x\in\mathbb{X}} \kappa(x, y)$ le vecteur de probabilité de la deuxième marginale et $q(x, y) = \mu(x)\nu(y)$ le vecteur de probabilité de la loi conjointe de deux variables aléatoires Z_1 et Z_2 indépendantes ayant comme lois respectives μ et ν. Nous avons

$$\begin{aligned}
H(X) &= -\sum_{x\in\mathbb{X}} \mu(x) \log \mu(x) = -\sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} \kappa(x, y) \log \mu(x) \\
H(Y) &= -\sum_{y\in\mathbb{Y}} \nu(y) \log \nu(y) = -\sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} \kappa(x, y) \log \nu(y) \\
H(X) + H(Y) &= -\sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} \kappa(x, y) \log[\mu(x)\nu(y)] = -\sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} \kappa(x, y) \log q(x, y).
\end{aligned}$$

Par le lemme 7.3.1 nous obtenons

$$H(X, Y) = -\sum_{x\in\mathbb{X}, y\in\mathbb{Y}} \kappa(x, y) \log \kappa(x, y) \leq -\sum_{x\in\mathbb{X}, y\in\mathbb{Y}} \kappa(x, y) \log q(x, y) = H(X) + H(Y),$$

avec égalité si, et seulement si, $\kappa(x, y) = q(x, y) = \mu(x)\nu(y)$, pour tout $x \in \mathbb{X}$ et $y \in \mathbb{Y}$, i.e. si les variables aléatoires X et Y sont indépendantes. □

Corollaire 7.5.3.

1. $H(X_1, \ldots, X_n) \leq \sum_{k=1}^{n} H(X_k)$ *avec égalité si, et seulement si, les variables aléatoires $(X_k)_{k=1,\ldots,n}$ sont indépendantes.*
2. $H(X_1, \ldots, X_m; Y_1, \ldots, Y_n) \leq H(X_1, \ldots, X_m) + H(Y_1, \ldots, Y_n)$ *avec égalité si, et seulement si, les vecteurs aléatoires $\mathbf{X} = (X_1, \ldots, X_m)$ et $\mathbf{Y} = (Y_1, \ldots, Y_n)$ sont indépendants.*

7.5.2 Entropie conditionnelle

Si X et Y sont deux variables définies sur le même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs respectivement dans $\mathbb{X}$ et $\mathbb{Y}$, alors la connaissance que $X = x$ permet de déterminer la loi conditionnelle de $\mathbb{P}(Y \in B | X = x)$, pour $B \subseteq \mathbb{Y}$ et $x \in \mathbb{X}$. Étant donné que la probabilité conditionnelle est une probabilité, il s'ensuit que nous pouvons définir son entropie par

$$H(Y|X = x) = -\sum_{y \in \mathbb{Y}} \mathbb{P}(Y = y | X = x) \log \mathbb{P}(Y = y | X = x).$$

Définition 7.5.4. L'**entropie conditionnelle** de Y sachant X est définie par

$$H(Y|X) = \sum_{x \in \mathbb{X}} H(Y|X = x) \mathbb{P}(X = x).$$

Remarque 7.5.5. L'expression de l'entropie conditionnelle n'est pas symétrique par rapport aux deux variables. Nous avons en effet

$$\begin{aligned} H(Y|X) &= -\sum_{x \in \mathbb{X}} \mathbb{P}(X = x) \sum_{y \in \mathbb{Y}} \mathbb{P}(Y = y | X = x) \log \mathbb{P}(Y = y | X = x) \\ &= -\sum_{(x,y) \in \mathbb{X} \times \mathbb{Y}} \mathbb{P}(X = x, Y = y) [\log \mathbb{P}(Y = y; X = x) - \log \mathbb{P}(X = x)] \\ &= H(X, Y) - H(X). \end{aligned}$$

Nous obtenons une expression plus symétrique en écrivant

$$H(X, Y) = H(Y|X) + H(X) = H(X|Y) + H(Y).$$

La signification de l'entropie conditionnelle $H(Y|X)$ découle immédiatement des formules précédentes : elle correspond à l'incertitude de la loi conjointe du couple (X, Y) *réduite* de l'incertitude de la variable aléatoire X car la valeur de X nous est révélée.

Théorème 7.5.6. *$H(Y|X) \leq H(Y)$ avec égalité si, et seulement si, Y et X sont indépendantes.*

Démonstration. Par l'égalité établie en remarque 7.5.5 et par la propriété de sous-additivité de l'entropie (théorème 7.5.2), nous avons

$$H(X, Y) \overset{7.5.5}{=} H(Y|X) + H(X) \overset{7.5.2}{\leq} H(Y) + H(X),$$

avec égalité si, et seulement si, les variables sont indépendantes. □

7.5.3 Information mutuelle

Une autre quantité importante est la différence $H(X) - H(X|Y)$; l'entropie $H(X)$ correspond à l'incertitude *a priori* que nous avons sur X (sur la base de sa loi), l'entropie conditionnelle correspond à l'incertitude sur X sachant que Y a été observée. La différence ci-dessus contient donc l'information sur X que véhicule l'observation de Y.

Définition 7.5.7. On appelle **information mutuelle** de X et Y, la quantité

$$I(X : Y) := H(X) - H(X|Y).$$

Remarque 7.5.8. Le théorème 7.5.6 garantit que $I(X:Y) \geq 0$, avec égalité si, et seulement si, les variables aléatoires X et Y sont indépendantes ; dans le cas d'indépendance, l'observation de Y ne nous apprend rien sur X, par conséquent, si nous soustrayons de l'incertitude de X l'incertitude conditionnelle, l'incertitude résiduelle est nulle. Par ailleurs, des égalités

$$H(X,Y) = H(Y|X) + H(X) = H(X|Y) + H(Y),$$

nous obtenons la symétrie de l'information mutuelle :

$$\begin{aligned} I(X:Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= I(Y:X). \end{aligned}$$

7.6 Registres de stockage de l'information

En informatique nous devons stocker, accéder et traiter l'information ; nous avons donc besoin d'un dispositif physique dont les états physiques peuvent être associés à des états logiques (abstraits). Un tel dispositif doit

— posséder suffisamment d'états physiques dans lesquels le système peut être préparé,
— son état courant doit être déterminé de manière non ambiguë et
— son état doit être transformable dans n'importe quel autre état légitime en exerçant sur le dispositif une action spécifique.

Des dispositifs physiques avec ces propriétés sont appelés **registres** et doivent être pensés comme les cellules mémoire d'un ordinateur.

Définition 7.6.1. Soit $\mathbb{A}$ un alphabet fini, l'**ensemble des états logiques**, $(\mathbb{X}, \mathcal{X})$ an espace mesurable separable[7], vérifiant $|\mathbb{X}| \geq |\mathbb{A}|$, et $\theta > 0$ le **seuil de discernabilité**. L'ensemble des mesures de probabilité $\mathcal{M}_1(\mathbb{X})$ correspond à l'**ensemble d'états physiques**. Pour un seuil $\theta > 0$ donné, un **registre** est une application abstraitei $R : \mathbb{A} \to \mathcal{P}(\mathcal{M}_1(\mathbb{X}, \mathcal{X}))$ telle que pour tout $a, b \in \mathbb{A}$, avec $a \neq b$,

— $R(a) \cap R(b) = \emptyset$ et
— $\forall \mu \in R(a), \forall \nu \in R(b) : D(\mu \| \nu) \geq \theta$.

En d'autres termes, les mesures de probabilité dans $R(a)$ et dans $R(b)$, associées à des lettres différentes $a, b \in \mathbb{A}$ peuvent être discriminées efficacement. En termes opérationnels, un registre est un dispositif physique qui associe à chaque état logique a une réalisation physique du système, préparé dans un état physique $\rho \in R(a)$.

Puisque $\mathbb{A}$ est un alphabet fini arbitraire, la définition précédente s'étend trivialement à des n-registres, pour $n \geq 1$, vus comme applications $R : \mathbb{A}^n \to \mathcal{P}(\mathcal{M}_1(\mathbb{X}))$, où $|\mathbb{X}| \geq |\mathbb{A}|^n$.

Remarque 7.6.2. Dans la définition 7.6.1, nous pouvons remplacer le contraste de Kullback-Leibler $D(\mu\|\nu)$ par le distance de variation totale $\|\mu - \nu\|_1$ qui, contrairemen à D, a l'avantage d'être une vraie distance quoique moins discriminante que D. Dorénavant, lorsque $\mathrm{card}\mathbb{A} > 2$, nous utiliserons toujours la distance de variation totale ; pour $\mathrm{card}\mathbb{A} = 2$, nous utilisons le contraste D ou la variation totale.

7. I.e. sa tribu $\mathcal{X}$ est dénombrablement engendrée et contient les singletons.

Exemple 7.6.3. (Un registre idéal vu comme un objet mathématique abstrait). Supposons $\mathbb{A} = \mathbb{X} = \{0,1\}$, avec $\mathbb{X}$ équipé de sa tribu exhaustive. Les mesures de probabilité extrémales sont les mesures qui n'admettent pas de décomposition convexe non-triviale en d'autres mesures de probabilité. Il est aussi évident que $\{\varepsilon_x, x \in \mathbb{X}\} \subseteq \partial_e \mathcal{M}_1(\mathbb{X}, \mathcal{X})$. Cette inclusion devient une égalité si l'espace $(\mathbb{X}, \mathcal{X})$ est séparable, ce qui est effectivement la cas dans la situation présente.

De façon abstraite (i.e. en négligeant les contraintes de la réalisation physique), un registre idéal[8] est l'application définie par la formule $x \mapsto R(x) := \{\varepsilon_x\}$ pour tout $x \in \mathbb{X}$. Lorsque $x \neq y$,

$$D(\varepsilon_x \| \varepsilon_y) = +\infty,$$

par conséquent les lettres x et y sont réalisées par deux mesures de probabilité uniquement définies et parfaitement discernables.

Exemple 7.6.4. (Un registre idéal vu comme un objet physique). Sous la même notation que dans l'exemple 7.6.3, i.e. $\mathbb{A} = \mathbb{X} = \{0,1\}$, associer avec l'état logique '0', une pièce totalement biaisée avec deux cotés « face » et avec l'état logique '1', une pièce totalement biaisée avec deux cotés « pile ». Ce système remplit les conditions pour qu'il soit considéré comme registre ; il implémente physiquement la notion mathématique abstraite de registre. Cependant, un tel système est totalement inutile pratiquement car le codage des états logiques se fait en dur dans la construction de la pièce qui empêche la transformation de 0 en 1 par l'action d'une opération raisonnable sur le système. Bien sûr, il est toujours possible de faire fondre la pièce pour en frapper une nouvelle avec un côté « pile » et un côté « face », une opération qui n'est pas très pratique pour stocker, extraire et traiter l'information sur un ordinateur !

La discernabilité parfaite est une exigence trop forte pour qu'elle soit physiquement implémentable. Pour cette raison, la définition 7.6.1 assouplit la condition de discernabilité parfaite en la condition de discernabilité au dessus d'un certai seuil $\theta > 0$, en déclarant deux mesures μ et ν dans $\mathcal{M}_1(\mathbb{X})$ comme discernées (au seuil $\theta \in]0, \infty[$) si $\theta \leq D(\mu \| \nu)$. Or, si le contraste n'est pas infini, les deux mesures ne peuvent pas être mutuellement singulières ; par conséquent, elles ne peuvent pas être des masses de Dirac masses supportées par des lettres différentes. Il est donc nécessaire de considérer pas seulement des mesures extrémales mais aussi des mesures générales dans $\mathcal{M}_1(\mathbb{X}) = \mathsf{co}(\{\varepsilon_x, x \in \mathbb{X}\})$.

Une difficulté supplémentaire surgit lorsqu'on relaxe la condition de stricte discernabilité. Tandis que l'ensemble des probabilités extrémales $\partial_e \mathcal{M}(\mathbb{X})$ est en bijection avec l'ensemble des états logiques $\mathbb{X}$, l'ensemble $\mathcal{M}_1(\mathbb{X})$ est beaucoup plus grand que $\mathbb{X}$. En outre, puisque la discrimination est faite à l'aide du seuil $\theta > 0$, l'image $R(\mathbb{A}) = \sqcup_{a \in \mathbb{A}} R(a)$ ne couvre pas $\mathcal{M}_1(\mathbb{X})$. Puisqu'alors $\mathcal{M}_1(\mathbb{X}) \setminus R(\mathbb{A}) \neq \emptyset$, il existe d'états physiques qui ne correspondent pas à des états logiques ; lorsque le système est préparé dans l'un de ces états, son état logique reste **indéterminé**. Moyennant l'extension de l'alphabet $\mathbb{A}$ en $\mathbb{A}_u$ par adjonction d'une nouvelle lettre $\mathbb{A}_u := \mathbb{A} \sqcup \{\partial\}$, l'application R peut être conçue comme l'application de l'alphabet étendu dans les classes d'équivalence des mesures de probabilité sur $\mathbb{X}$, engendrée par la partition $\mathcal{M}_1(\mathbb{X}) = \sqcup_{a \in \mathbb{A}_u} R(a)$, où $R(\partial) = \mathcal{M}_1(\mathbb{X}) \setminus R(\mathbb{A})$.

Le contraste de Kullback-Leibler n'est pas de maniement facile pour discriminer le codage de différentes lettres pour des alphabets à trois lettres ou plus. La distance

8. I.e. qui discerne parfaitement les états physiques p.

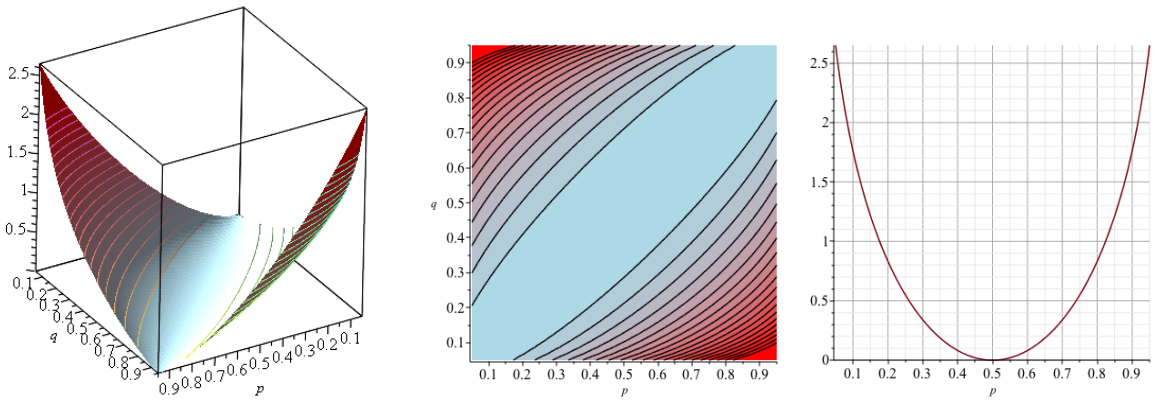


FIGURE 7.5 – **Gauche :** $D(\mathbf{p}\|\mathbf{q})$ pour $\mathbf{p} = (p, 1-p)$, $\mathbf{q} = (q, 1-q)$, comme fonction de $p, q \in]0.05, 0.95[$. Les intersections avec les plans horizontaux à hauteur $\theta \in \{0.1, 0.2, \ldots, 1.9, 2.0\}$ sont représentées par les courbes rouges sur la surface. **Milieu :** Projection de la surface précédente sur me plan horizontal (à hauteur 0) ; elle permet de visualiser les courbes de niveau à hauteur θ. Si nous fixons le seuil de discrimination à $\theta = 0.1$, la région interne (« lac » coloré en bleu clair) correspond à la classe d'équivalence des états physiques $R(\partial)$ provenant de l'état logique indéterminé ∂. La région gauche supérieure (au delà de la rive nord-ouest du lac) correspond à la classe d'équivalence $R(1)$ des états physiques provenant de l'état logique 1, la région droite inférieure (au delà de la rive sud-est du lac) à la classe d'équivalence $R(0)$ provenant de l'état logique 0. **Droite :** Intersection de la surface descrvant $D(\mathbf{p}\|\mathbf{q})$ avec le plan vertical $q = 1 - p$. Nous remarquons que quand $p = 1/2$, les vecteurs $\mathbf{p} = (p, 1-p)$ et $\mathbf{q} = (1-p, p)$ coïncident ; par conséquent leur contraste s'annule et les vecteurs deviennent indiscernables. Les probabilités peuvent être discriminées efficacement lorsque leur contraste excède un certain seuil, ex. 0.1.

de variation totale, malgré son pouvoir discriminant inférieur à celui du contraste, est plus adaptée pour des tels alphabets ; la figure 7.6 donne un exemple dans le cas $\mathbb{A} = \mathbb{X} = \{a, b, c\}$.

De nos jours, les registres de mémoire sont implémentés par des transistors. Plus précisément, un registre peut être réalisé par un système physique bistable composé de deux jonctions bipolaires montées en opposition. Il n'est pas le but de ce texte de présenter tous les détails de la technologie des transistors [9]. Il est par contre important de souligner que la définition de registre donnée ici et la bijection de l'alphabet logique étendu avec les classes d'équivalence $(R(a))_{a \in \mathbb{A}_u}$ décrit parfaitement les registres à transistors.

7.7 Irreversibilité et principe de Landauer

Dans ce paragraphe, nous allons présenter un résultat général — le principe de Landauer — qui régit le fonctionnement de registres mémoire. La formulation initiale de ce principe [44] est faite de manière assez informelle et les arguments utilisés pour sa justification font appel à des principes thermodynamiques. Ces approches ne sont pas très rigoureuses car elle font intervenir des arguments thermodynamiques pour décrire des systèmes composés d'un seul atome.

9. Le lecteur intéressé peut consulter les livres [25, 34] pour des détails techniques.

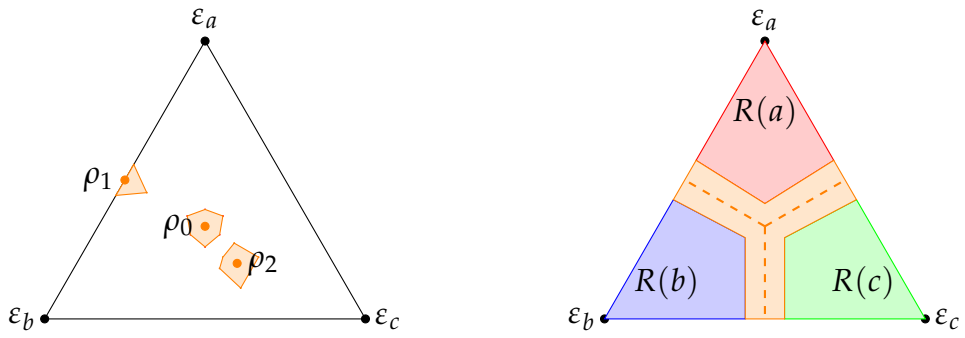


FIGURE 7.6 – **Gauche :** L'ensemble convexe des états $\mathcal{M}_1(\{a,b,c\})$ et exemples de boules (pour la distance de variation totale) de rayon $\delta = 0.1$ autour des états $\rho_0 = \frac{1}{3}(\varepsilon_a + \varepsilon_b + \varepsilon_c)$, $\rho_1 = \frac{1}{2}(\varepsilon_a + \varepsilon_b)$ et $\rho_2 = 0.2\varepsilon_a + 0.3\varepsilon_b + 0.5\varepsilon_c$. **Droite :** Partition de l'ensemble des états en régions disjointes $R(a), R(b), R(c)$ codant les lettres $a, b, c \in \mathbb{X}$ et région tampon (orange) d'épaisseur 0.1, correspondant à l'état indéterminé ∂. Cette région est obtenue comme enveloppe convexe de toutes les boules de rayon $\delta = 0.1$ centrées autour des états entre ρ_0 et $\frac{1}{2}(\varepsilon_a + \varepsilon_b)$, entre ρ_0 et $\frac{1}{2}(\varepsilon_b + \varepsilon_c)$ et entre ρ_0 et $\frac{1}{2}(\varepsilon_c + \varepsilon_a)$. Les régions $R(a), R(b)$ et $R(c)$ sont des ensembles disjoints provenant de la complémentation de la région tampon ; elles correspondent aux classes d'équivalence $[\varepsilon_a], [\varepsilon_b]$ et $[\varepsilon_c]$ d'états qui peuvent sans ambiguïté être associés (à seuil de discernabilité that can be $\theta = 0.1$) aux lettres a, b et c respectivement.

Le principe de Landauer constitue un aspect de l'irréversibilité inhérente à tout calcul sur ordinateur [10] Dans la suite, nous esquissons l'idée de démonstration du principe de Landauer en faisant appel à des notions de physique statistique [11].

Un modèle de registre plus réaliste — quoique encore naïf — que celui de l'exemple 7.6.4 est un système physique bistable possédant deux états distincts quasi-stables. Par exemple, considérer une petite masse sphérique qui évolue dans le potentiel à double puit esquissée en figure figure 7.7. L'état logique 0 est codé dans l'état physique x_0 et l'état logique 1 dans x_1.

Considérer l'opération logique REMISEÀZERO (RÀZ) définie comme $0 \mapsto 0$ et $1 \mapsto 0$. Naïvement, on peut penser que lorsque l'état initial du registre est x_0, l'opération RÀZ est synonyme à l'operation NERIENFAIRE entraînant un coût énergétique nul. Lorsque l'état initial est x_1, une force dépendante du temps doit être exercée sur la masse. Durant la première partie du mouvement, la force doit agir dans le même sens que le mouvement en fournissant le travail nécessaire pour élever la masse ; après avoir atteint le maximum local du potentiel, une force freinant la masse doit être exercée pour qu'elle arrive à x_0 à vitesse nulle. Puisque la force freinante agit en sens opposé au mouvement, de travail mécanique est produit égal en quantité au travail consommé en première étape. Nous pouvons donc conclure que l'opération RÀZ s'effectue à coût énergétique nul.

10. Il ne faut pas oublier que la notion d'entropie a été introduite pour expliquer l'irréversibilité de certains systèmes physiques macroscopiques.

11. La mécanique statistique est la théorie physique qui explique la phénoménologie de la thermodynamique.

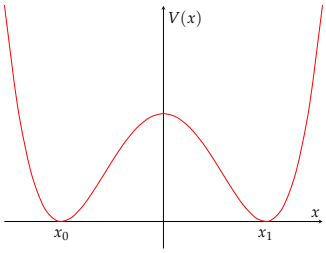


FIGURE 7.7 – Un double puits de potentiel avec deux positions d'équilirbre x_0 et x_1. Une masse sphérique avec vitesse initiale null placée en x_0 ou x_1 restera là pour toujours et peut être considérée comme codant les états logiques 0 ou 1. Quand la masse, initialement en x_0, est poussée vers x_1 avec une force suffisante pour surmonter la barrière de potentiel, et ensuite ralentie durant sa descente pour arriver à x_1 avec une vitesse nulle, l'évolution du système correspondra à un changement de l'état physique de celui codant le bit 0 à celui codant 1 à coût énergétique nul.

Cependant, ce schéma ne reflète pas le fonctionnement réaliste d'un registre mémoire d'un ordinateur. En réalité la *même routine est appliquée au registre* pour implémenter l'opération RÀZ, indépendamment de l'état initial du registre.

En outre, comme mentionné en §7.6, le registre peut se trouver dans un état physique qui n'est pas un état pur correspondant à x_0 ou x_1 mais à un état mélange. La description d'un registre simple fournie par la figure 7.7 est assez rudimentaire et naïve puisqu'elle ne représente les états logiques 0 ou 1 que par les états purs $R(0) = \{(1,0)\}$ et $R(1) = \{(0,1)\}$. Mais, comme déjà expliqué, les mémoires d'un ordinateur sont implémentées par des circuits physiques de transistors ; il est donc totalement illusoire de penser qu'elles correspondent à des strictes masses de Dirac. Il est important d'avoir des systèmes capables de représenter n'importe quel état physique $\mathbf{p} = (p, 1-p)$, pour un $p \in [0,1]$ arbitraire.

Un réel arbitraire $p \in [0,1]$ peut être approché par des rationnels. Soient r, N des entiers tels que $\frac{r}{N} \leq p < \frac{r+1}{N}$ et supposons qu'une urne à deux compartiments contient and N boules indiscernables. L'ensemble de configurations possibles des N boules dans les deux compartiments est isomorphe à

$$\mathbb{B} = \{\mathbf{b} = [b_1, \ldots, b_N], b_k \in \mathbb{A} = \{0,1\}\}.$$

Puisque l'ordre des lettres dans le mot $\mathbf{b}$ est sans importance, $\mathbf{b}$ est bijectivement représenté par le vecteur d'occupation :

$$\mathbb{B} \ni \mathbf{b} \mapsto \boldsymbol{\rho}_{\mathbf{b}} \in \mathcal{M}_1(\mathbb{A}),$$

où

$$\rho_{\mathbf{b}}(a) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}_{\{a\}}(b_i).$$

Autrement dit, le mot $\mathbf{b} = [\underbrace{0 \cdots 0}_{r} \underbrace{1 \cdots 1}_{N-r}]$ engendre le vecteur $\boldsymbol{\rho}_{\mathbf{b}} = (\frac{r}{N}, \frac{N-r}{N})$. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité (suffisamment grand). Nous introduisons le variable aléatoire $U : \Omega \to \{1, \ldots, N\}$ et une famille paramétrique de variables aléatoire $X : \mathbb{B} \times \Omega \to \mathbb{A}$, où U est uniformement distribuée sur $\{1, \ldots, N\}$ et $X(\mathbf{b}, \omega) = b_{U(\omega)}$. Lorsqu'il y a r boules dans le compartiment gauche et $N - r$ boules dans le compartiment droit,

$$\mathbb{P}(X = 0) = \sum_{k=1}^{N} \mathbb{P}(X = 0 | U = k) \mathbb{P}(U = k) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}_{\{0\}}(b_k) = \rho_{\mathbf{b}}(0).$$

L'ensemble des états physiques représentables est alors $\{\rho_{\mathbf{b}}, \mathbf{b} \in \mathbb{B}\} \subset \mathcal{M}_1(\mathbb{A})$.

Exemple 7.7.1. Instancier le modèle précédent dans le cas $N = 2$ (deux boules dans une urne à 2 compartiments) engendrant les états physiques $(1,0)$, $(1/2,1/2)$ et $(1,1)$; ils peuvent être réalisés respectivement par les configurations des boules $[00]$, $[01]$ et $[11]$. De toute évidence, nous associons l'état logique 0 avec la configuration physique $\rho_{[00]}$, l'état logique 1 avec l'état physique $\rho_{[11]}$ et l'état logique ∂ avec la configuration $\rho_{[01]}$.

Instancier maintenant les considérations précédentes dans le cas de N suffisamment grand et d'un r arbitraire $r \in \{0, \ldots, N\}$. Un tel système modélise un registre dans l'état $(p, 1-p)$ avec $p = \frac{r}{N} \in [0,1]$. Réaliser physiquement ce registre comme un récipient de volume V avec une paroi externe adiabatique [12] et une paroi adiabatique séparant le récipient en deux compartiments ; sur cette paroi séparatrice il existe un trou qui peut être ouvert ou fermé. Initialement le trou est fermé et le compartiment gauche contient r atomes d'un gaz parfait et le droit $N - r$ atomes du même gaz (voir (a) de la figure 7.8). Le recipient est maintenu à la température T. Selon la loi des gaz parfaits, le pression P dans chaque compartiment est donnée par les formules

$$P_{\text{gauche}} \frac{V}{2} = rkT = pNkT, \quad P_{\text{droite}} \frac{V}{2} = (N-r)kT = (1-p)NkT,$$

où $k := k_B = 1.380\,649 \times 10^{-23}\,\text{J/K}$ est la constante de Boltzmann.

L'opération logique RÀZ peut maintenant être implémentée par le processus physique décrit dans la légende de la figure 7.8.

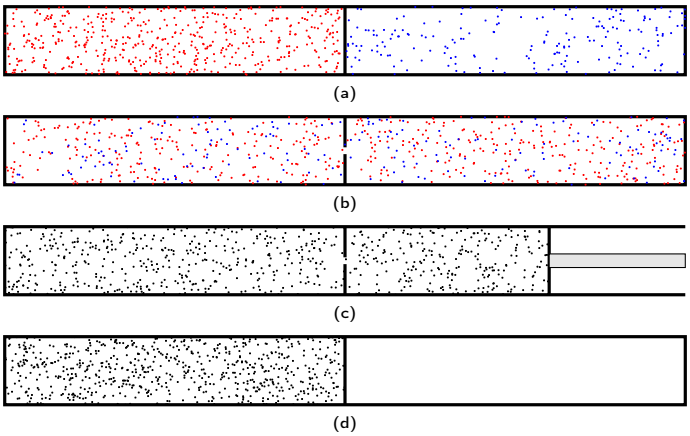


FIGURE 7.8 – **(a)** État initial du système $(p, 1-p)$, avec $p = r/N$. **(b)** Le trou dans la paroi interne est ouverte à coût énergétique zéro. Très rapidement après l'ouverture du troi, le gaz occupe désormais uniformément la totalité du volume. Les couleurs différentes pour les atomes sont uniquement pour des raisons d'illustration pour rappeler la provenance de chaque atome individuel. Penser cependant que les atomes sont indiscernables. Par conséquent la couleur ne sera pas montrée dans les étapes suivantes. **(c)** La paroir à l'extrémité droite est en réalité un piston qui comprime le gaz de façon isotherme jusqu'à la moitié du volume. **(d)** Le trou est de nouveau fermé et le piston retourné à sa position initiale à l'extrémité droite du récipient. Maintenant, tous les atomes se trouvent dans le compartiment gauche et le registre est dans l'état pur $(1,0)$.

12. ADIABATIQUE, adj. et subst. fém. Qui se produit sans qu'il y ait échange de chaleur avec l'extérieur.

Nous sommes maintenant en mesure de faire quelques calculs élémentaires. Écrire $V = LA$ pour le volume, où L est la longueur du récipient et A l'aire de sa section transverse. Considérer la situation décrite en (c) lorsque l'abscisse courante du piston est $x \in [L/2, L]$. La force instantanée $F(x)$ qui doit être exercée sur le piston pour équilibrer la force exercée sur le piston par la pression est $F(x) = \frac{NkT}{x}$ (la force pointe vers la gauche), donc le travail total nécessaire pour comprimer le gaz du volume V à $V/2$ est

$$W_{(b)\to(d)} = Nk_BT \int_L^{L/2} \frac{\mathrm{d}x}{x} = -Nk_BT \ln 2.$$

Le signe négatif signifie que le travail doit être dépensé pour changer l'état du système de l'état (b) à (d). Par ailleurs, si l'on associe une variable aléatoire 0 ou 1 à chaque atome selon qu'il se trouve à gauche ou à droite de la position courante du piston, l'entropie totale[13] de cette famille, lorsque le piston est en position (b) est $NH(1/2, 1/2) = N\text{bits} = Nk_B \ln 2$ J/K, tandis que $NH(1, 0) = 0$ bits $= 0$ J/K. Par conséquent, le coût énergétique *par atome* pour changer l'état initial du registre de l'état arbitraire $(p, 1-p)$ à l'état $(1, 0)$ est $\frac{T}{N}\Delta S$ où $\Delta S = S(1, 0) - S(1/2, 1/2)$.

L'opération logique RÀZ correspond à l'effacement[14] du contenu informationnel initial du registre. Ce que monte le calcul heuristique précédent est l'égalité

$$\Delta S = \frac{W}{k_BT \ln 2}.$$

Plus spécificement, « effcer » un bit d'information coûte de l'énergie. Dans le calcul précédent, uniquement des transformations adiabatiques sont considérées. Si l'hypothèse d'adiabaticité est assouplie, au lieu de l'égalité nous avons l'inégalité

$$\Delta S \leq \beta \Delta Q,$$

où Q est la chaleur ajoutée au système et le symbole β est habituellement utilisé en mécanique statistique pour désigner la quantité $\beta = \frac{1}{k_BT}$ (nous pouvons absorber $\ln 2$ en une redéfinition de k_B.)

L'évolution des systèmes physiques isolés est réversible. Ceci signifie qu'il n'existe pas d'expérience physique réalisable qui permet de distinguer un film d'évolution d'un sytème projeté à l'endroit du même film projeté à l'envers.

L'entropie est étroitement liée à l'irréversibilité puisque selon le second principe de thermodynamique, dans sa version microscopique formulée par Carathéodory [13], stipule que l'entropie d'un système **isolé**

— est une fonction croissante du temps ; pour des tels systèmes, l'énergie est préservée et ils évoluent spontanément vers des états d'équilibre thermodynamique, i.e. des états à entropie maximale ;

13. Il faut se rappeler que l'entropie informationnelle est mesurée en bits et correspond à $k = 1$ et logarithme binaire tandis que l'entropie thermodynamique est mesurée en J/K and correspond à $k = 1.380\,649 \times 10^{-23}$ J/K and logarithme néperien.

14. Dans les langages de programmation anciennes, lorsque une réservation de cellules mémoire pour un tableau était faite, il était nécessaire d'appliquer l'opération RÀZ. Une erreur fréquente commise par des programmeurs inexpérimentés des années '70 était d'omettre cette opération et commencer à référencer les cellules de ce tableau directement ; or, les cellules mémoire pouvaient contenir de bits « aléatoires », reliques des programmes précédents. Les compilateurs modernes évitent ce piège en initialisant à 0 presque systématiquement toutes les cellules mémoires lorsqu'elles sont référencées pour la première fois.

— peut rester constante **uniquement** pour des évolutions réversibles isolées.

Remarque 7.7.2. Une évolution d'un système physique (classique) avec espace des états **S** est un noyau stochastique agissant à gauche sur des états (≡ probabilités) en les transformant à des nouvelles probabilités. Soit H la fonction entropie telle que définie dans le cas fini (i.e. card$\mathbb{X} < \infty$) au théorème **??**). Dans ce cas, les noyaux stochastiques sont des matrices stochastiques ; ils correspondent à des évolutions réversibles si le noyau est une matrice stochastique déterministe inversible, i.e. représente une permutation on $\mathbb{X}$. L'invariance de H aux permutations est inhérente à sa definition.

Lorsqu'un système physique évolue dans le temps, son état change vers un nouvel état. Il est
— expérimentalement observé en thermodynamique,
— postulé en physique théorique,
— démontré en mécanique statistique et en théorie des probabilités,
que l'entropie est une fonction croissante dans le temps ; elle peut rester constante si, et seulement si, l'évolution est réversible (une telle évolution correspond à celle d'un système isolé). Cette remarque a des conséquences importantes en théorie de l'information car le traitement, la transmission et l'extraction de l'information sont des procédés physiques, malgré que ce fait n'est jamais souligné dans les livres consacrés à la théorie de l'information où tout est décrit en termes du formalisme abstrait des portes logiques.

L'effacement des bits est de toute évidence une opération non réversible. Une question naturelle se pose : pourquoi effacer des bits, gaspillant [15] ainsi de l'énergie électrique ou mécanique précieuse en la transformant en chaleur (une forme d'énergie dégradée) ?

Nous avons déjà évoqué (cf. note en bas de la page **??**) la nécessité d'effacer les reliques des calculs précédents de cellules mémoire. On pourrait alors dire « trop de bruit pour rien ». Mais réfléchissons un peu ! Sur un ordinateur classique, les opération logiques sont exécutées sur des portes logiques (AND, OR, etc.). Ces portes on 2 entrées et une sortie ; elles sont donc nécessairement irréversibles. Un bit est perdu (effacé) chaque fois qu'une telle porte est appelée, i.e. des milliards de fois par seconde sur un ordinateur moderne !

Le principe de Landauer [44, 4] : Tout gain informationnel (décroissance de l'entropie) induit par des opérations logiquement irréversibles **doit impérativement** être compensé par une augmentation de l'entropie des degrés de liberté non-informationnels de l'ordinateur et de son environnement (registres, alimentation électrique et autres parties non génératrices d'information) au moins équivalent à la information acquise. I.e. pour acquérir 1 bit d'information, l'entropie de l'environnement doit croître, **d'au moins** 0.957×10^{-23} J/K. La dissipation d'énergie résultant de l'effacement d'un bit d'information a été expérimentalement vérifié sur un système microscopique pour la première fois 2012 [6] et confirmé en 2016 [33].

15. La thermodynamique introduit une hiérarchie parmi les différentes formes d'énergie. Les formes d'énergie les plus nobles sont celle qui peuvent facilement être converties en d'autres formes avec un facteur de conversion élevé. Par exemple, on peut convertir de l'énergie électrique en énergie mécanique avec un facteur de l'ordre de 95% dans un moteur électrique. Ceci constitue la raison pour laquelle les trains électriques sont si fréquents. À l'inverse, une machine à vapeur peut servir à transformer l'énergie chimique contenue dans du charbon en chaleur et ensuite en énergie mécanique, avec un facteur de conversion total de l'ordre de 1%. C'est la raison pour laquelle on ne rencontre plus des locomotives à vapeur que dans les musées technologiques.

Penser que la valeur $0.957 \times 10^{-23}\,\mathrm{J/K}$, correspondant à l'accroissement d'entropie pour l'acquisition d'un bit d'information, est le **minimum théorique absolu**. Un ordinateur classique de tous les jours accroît l'entropie de l'environnement par des valeurs dépassant de plusieurs ordres de grandeur cette borne. Typiquement, l'accroissement de l'entropie est couplé avec la consommation d'énergie électrique qui est fournie à l'ordinateur par son alimentation et qui est convertie en chaleur, dissipée finalement dans l'environnement. Ce phénomène peut facilement s'observer lorsque un ordinateur portable exécute un calcul compliqué ; il devient de plus en plus chaud et son ventilateur se met à fonctionner de plus en plus rapidement.

Le coût énergétique de la technologie d'information et du calcul (ICT pour Information and computer technology) constitue une part importante de l'énergie consommée sur Terre. Comme reporté dans [26], le cout d'ICT en 2012 était approximativement de 3.3 HJ, soit le 4.7% de l'électricité mondialement produite [16].

7.8 Exercices

Entropie, mesure de l'information

78. Les habitants d'un village sont divisés en deux parties. Une partie A contient des individus qui disent la vérité avec probabilité $1/2$, mentent avec probabilité $3/10$ et refusent de répondre avec probabilité $2/10$. La partie B contient des individus dont les probabilités pour chaque type de comportement sont respectivement $3/10$, $1/2$ et $2/10$. Soit $p \in [0,1]$ la probabilité qu'un habitant du village choisi au hasard appartienne au groupe A. On note $i := i(p)$ l'information sur son comportement vis-à-vis des questions posées qui est véhiculé par son appartenance à un groupe donné. Calculer $p_0 = \arg\max i(p)$ et $i(p_0)$.
79. Soient $(a_i)_{i=1,\dots,n}$ des nombres positifs vérifiant $\sum_{i=1}^n a_i = 1$ et $(x_i)_{i=1,\dots,n}$ des nombres strictement positifs. Établissez l'inégalité

$$x_1^{a_1} \cdots x_n^{a_n} \leq \sum_{i=1}^n a_i x_i.$$

80. Soit $\mathbf{p} = (p_1, \dots, p_n)$ un vecteur de probabilité (i.e. ayant des coordonnées $p_i \geq 0$ et vérifiant la condition de normalisation $\sum_{i=1}^n p_i = 1$). Supposons que $p_1 > p_2$ et notons $\delta \in]0, \frac{p_2 - p_1}{2}]$. On considère le nouveau vecteur de probabilité $\mathbf{p}'$ avec $p_1' = p_1 - \delta$, $p_2' = p_2 + \delta$ et $p_i' = p_i$ pour $i = 3, \dots, n$. Les entropies $H(\mathbf{p}')$ et $H(\mathbf{p})$ se comparent-elles et si oui comment ?
81. Soient $A = (A_{ij})_{i,j=1,\dots n}$ une matrice doublement stochastique (i.e. une matrice dont tous les termes sont positifs et dont chaque ligne et chaque colonne a une somme normalisée à 1), et $\mathbf{p} = (p_1, \dots, p_n)$ un vecteur (ligne) de probabilité.

    (a) Montrer que le vecteur (ligne) $\mathbf{p}' = \mathbf{p}A$ est un vecteur de probabilité.
    (b) Comparer les entropies $H(\mathbf{p})$ et $H(\mathbf{p}')$.

16. Selon les données publiées par l'International Energy Agency, la consommation totale d'élecricité est environ de 68.4 HJ et représente 13% du total de l'énergie produite (ca. 540 HJ.) Le symbole H désigne le préfixe hexa, représentant le facteur 10^{18}.

Propriétés de différentes variantes d'entropie et relations entre elles

82. (a) Montrer l'inégalité de Jensen (dans un cas simple), à savoir que si f est une fonction convexe dérivable et $(a_n)_{i\in I}$ une famille finie de nombre positifs tels que $\sum_{i\in I} a_i = 1$, alors
$$\sum_{i\in I} a_i f(t_i) \geq f(\sum_{i\in I} a_i t_i).$$
(b) Soient $(a_i)_{i=1,\dots,n}$ une famille de nombres positifs et $(b_i)_{i=1,\dots,n}$ une famille de nombres strictement positifs. Montrer que
$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{n} \log \frac{\sum_{j=1}^{n} a_j}{\sum_{k=1}^{n} b_k}.$$
(c) Montrer que $D(\mathbf{p}\|\mathbf{q})$ est convexe en le couple $(\mathbf{p},\mathbf{q})$, i.e. si $(\mathbf{p},\mathbf{q})$ et $(\mathbf{p}',\mathbf{q}')$ sont deux couples de vecteurs de probabilité et $\lambda \in [0,1]$ alors
$$D(\lambda\mathbf{p} + (1-\lambda)\mathbf{p}'\|\lambda\mathbf{q} + (1-\lambda)\mathbf{q}') \leq \lambda D(\mathbf{p}\|\mathbf{q}) + (1-\lambda)D(\mathbf{p}'\|\mathbf{q}').$$

83. En se servant du résultat précédent pour $\mathbf{q} = \mathbf{q}' = \mathbf{u}$, où $\mathbf{u}$ est le vecteur de probabilité de la loi uniforme, montrer que $H(\mathbf{p})$ est concave en $\mathbf{p}$.
84. (Extrait de l'examen du 19 décembre 2013).
Soient X_1 et X_2 deux variables aléatoires sur $(\Omega,\mathcal{F},\mathbb{P})$ prenant des valeurs dans le même espace $\mathbb{X}$ avec des lois décrites par les vecteurs de probabilité $\mathbf{p}$ et $\mathbf{p}'$ respectivement. Soit Y une variable aléatoire sur le même espace $(\Omega,\mathcal{F},\mathbb{P})$ prenant de valeurs dans $\mathbb{Y} = \{1,2\}$ selon le vecteur de probabilité $(\lambda, 1-\lambda)$, avec $\lambda \in [0,1]$; on note $Z = X_Y$ la variable aléatoire à valeurs dans $\mathbb{X}$.
(a) Calculer $\mathbb{P}(Z = x)$.
(b) En se servant du fait que le conditionnement réduit l'incertitude, i.e. $H(Z) \geq H(Z|Y)$, établir le résultat de concavité de $H(\mathbf{p})$ par rapport à $\mathbf{p}$ avec une méthode alternative à celle utilisée dans l'exercice précédent.

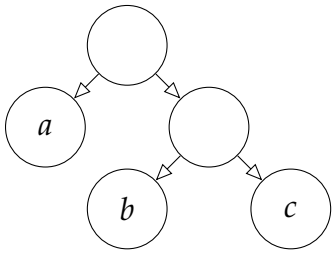
Compléments sur l'entropie

85. (Extrait de l'examen du 20 décembre 2017).
Le but de cet exercice est de montrer le
Théorème : *Soient $\mathbb{X}$ un ensemble fini, β un paramètre positif et $U : \mathbb{X} \to \mathbb{R}_+$ une variable aléatoire réelle positive sur $\mathbb{X}$. Pour une mesure de probabilité arbitraire $\nu \in \mathcal{M}_1(\mathbb{X})$ sur $\mathbb{X}$, on note $\nu U := \sum_{x\in\mathbb{X}} \nu(x)U(x)$, l'espérance [17] de U sous ν. Alors,*
(a) il existe une mesure de probabilité μ_β sur $\mathbb{X}$ qui sature le $\sup_{\nu\in\mathcal{M}_1(\mathbb{X})} (H(\nu) - \beta\nu U)$, où $H(\nu)$ désigne l'entropie de ν,
(b) $\mu_\beta(x) = \frac{\exp(-\beta U(x))}{\mathcal{Z}(\beta)}$, pour tout $x \in \mathbb{X}$, où $\mathcal{Z}(\beta) = \sum_{y\in\mathbb{X}} \exp(-\beta U(y))$ est un facteur de normalisation.
(a) Utiliser la concavité de la fonction log pour montrer que pour tout $\nu \in \mathcal{M}_1(\mathbb{X})$, on a $H(\nu) - \beta\nu U \leq \log \mathcal{Z}(\beta)$.
(b) Calculer $H(\mu_\beta) - \beta\mu_\beta U$.

17. Le vecteur de probabilité ν est considéré comme un vecteur ligne, la variable aléatoire U comme un vecteur colonne ; pour ν fixé, l'espérance de U est la forme linéaire νU.

86. Simulation d'une loi arbitraire avec une pièce honnête. (Extrait de l'examen du 15 décembre 2014). On dispose d'une pièce honnête (prenant de valeurs dans $\mathbb{B} = \{0,1\}$ selon le vecteur de probabilité $(1/2, 1/2)$). Les lancers successifs de la pièce sont modélisés par des suites aléatoires de longueur arbitraire de bits indépendants, c'est-à-dire par des mots $\xi \in \mathbb{B}^+$ (on rappelle que $\mathbb{B}^+ = \cup_{n=1}^{\infty} \mathbb{B}^n$). On veut simuler une variable aléatoire X à valeurs dans un ensemble fini $\mathbb{X}$ dont la loi est décrite par un vecteur de probabilité $\mathbf{p}$ arbitraire. Autrement dit, nous voulons exprimer X comme fonction de certains mots, choisis d'une certaine manière, parmi les mots de $\mathbb{B}^+$, de façon que la loi de X soit déterminée par les probabilités des mots choisis.

    (a) Commencer par l'ensemble à trois éléments $\mathbb{X} = \{a, b, c\}$ et $\mathbf{p} = (1/2, 1/4, 1/4)$. Placer les lettres a, b, c comme des feuilles d'un arbre binaire complet (i.e. dont chaque nœud a 0 ou 2 descendants) comme dans la figure adjacente. En associant le bit 0 aux arêtes gauches et le bit 1 aux arêtes droites, on constate que l'ensemble de feuilles $\mathbb{F} = \{0, 10, 11\}$ de l'arbre se surjecte sur $\mathbb{X}$. On note $F : \mathbb{F} \to \mathbb{X}$ cette surjection (dans ce cas particulier, il s'agit d'une bijection). Expliciter l'algorithme de génération de X.



    (b) Dans le cas particulier de la question précédente, estimer le nombre moyen de fois qu'il faudra lancer la pièce pour réaliser X et comparer ce résultat avec l'entropie $H(X)$.
    (c) Considérer maintenant l'ensemble $\mathbb{X} = \{a, b\}$ et le vecteur de probabilité $\mathbf{p} = (2/3, 1/3)$. *Suggestion : Observer que* $\sum_{k=0}^{\infty} \frac{1}{2^{2k+1}} = \frac{2}{3}$ *et utiliser cette égalité pour donner les représentations binaires des nombres* $2/3$ *et* $1/3$*; se servir de cette représentation pour déterminer l'ensemble* $\mathbb{F}$ *de feuilles et la surjection* $F : \mathbb{F} \to \mathbb{X}$.
    (d) Estimer le nombre moyen de lancers nécessaires pour simuler X.
    (e) Pouvez-vous proposer une méthode générale pour un ensemble fini $\mathbb{X}$ arbitraire muni d'un vecteur de probabilité arbitraire $\mathbf{p}$?

8
Sources et leur codage

Nous disposons d'une source qui génère une suite de symboles que nous supposons qu'ils puissent être transmis sans erreur, i.e. à travers un canal de transmission parfait (il restitue à sa sortie un message identique à celui qu'il reçoit en entrée). Notre préoccupation donc est uniquement d'optimiser le codage de la source en tenant compte de la probabilité avec laquelle les lettres constitutives du message-source sont produites.

8.1 Sources

Définition 8.1.1. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé adéquat.

1. Une **source discrète** est un dispositif qui émet une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires à valeurs dans un ensemble fini $\mathbb{X}$. L'ensemble $\mathbb{X}$ constitue les symboles émis par la source. La source est totalement déterminée par la donnée $((\Omega, \mathcal{F}, \mathbb{P}), \mathbb{X})$ que nous pouvons abréger en $(\mathbb{X}, \mathbb{P})$.
2. Une source discrète $(\mathbb{X}, \mathbb{P})$ est **stationnaire** si pour tout $N \in \mathbb{N}$, tout $n \geq 1$ et tout n-uplet $(x_1, \ldots, x_n) \in \mathbb{X}^n$, on a
$$\mathbb{P}(X_{N+1} = x_1, \ldots, X_{N+n} = x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$
3. Une source discrète $(\mathbb{X}, \mathbb{P})$ est **markovienne** si la suite $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov $\mathsf{MC}(\mathbb{X}, P, \cdot)$.
4. Une source discrète $(\mathbb{X}, \mathbb{P})$ est **sans mémoire** si la suite $(X_n)_{n\in\mathbb{N}}$ est une suite indépendante.
5. Une source discrète $(\mathbb{X}, \mathbb{P})$ sans mémoire est **homogène** si la suite $(X_n)_{n\in\mathbb{N}}$ est une suite indépendante et identiquement distribuée avec $\mathbb{P}(X_n = x) = p(x)$, pour tout $x \in \mathbb{X}$ et tout $n \in \mathbb{N}$, où $\mathbf{p} = (p(x))_{x\in\mathbb{X}} \in \mathsf{PV}_{\mathbb{X}}$ est le vecteur de probabilité de la loi de X_1.

Dans ce chapitre nous nous limiterons au cas de sources discrètes homogènes sans mémoire, c'est-à-dire que les messages produits peuvent être considérés comme des

réalisations d'une suite de variables aléatoires à valeurs dans $\mathbb{X}$ indépendantes et identiquement distribuées [1] selon la loi décrite par le vecteur de probabilité $\mathbf{p}$. De telles sources seront notées dans la suite $(\mathbb{X}, \mathbf{p})$. Sans perte de généralité, on peut supposer que $\mathbb{X} = \operatorname{supp} \mathbf{p}$.

8.2 Codes uniquement décodables

Nous nous intéressons à une représentation de chaque symbole émis par la source dans une alphabet fini $\mathbb{A}$.

Définition 8.2.1. Soient $(\mathbb{X}, \mathbb{P})$ une source discrète arbitraire et $\mathbb{A}$ un alphabet fini avec $\operatorname{card}\mathbb{A} \geq 2$.

- Un **codage** des symboles émis par la source est une application (le code) $C : \mathbb{X} \to \mathbb{A}^+$.
- On appelle **glossaire du code** l'ensemble $C(\mathbb{X}) \subset \mathbb{A}^+$.
- On note $|C(x)|$ la **longueur** du mot $\alpha \in \mathbb{A}^+$ codant le symbole x. Lorsque tous les symboles $x \in \mathbb{X}$ sont représentés par des mots de $\mathbb{A}^n$ (avec un certain $n \geq 1$), le code est dit de longueur fixe n, sinon il est de longueur variable.
- Un **code par blocs** (plus précisément m-blocs, avec $m \geq 1$), est un codage de l'alphabet étendu $\mathbb{X}^m$, c'est-à-dire une application $C : \mathbb{X}^m \to \mathbb{A}^+$.
- Un **code étendu par concaténation** est une application $\tilde{C} : \mathbb{X}^+ \to \mathbb{A}^+$, définie pour tout $\xi \in \mathbb{X}^+$ par
$$\tilde{C}(\xi) := C(\xi_1) \cdots C(\xi_{|\xi|}).$$
Dans la suite nous omettrons le signe distinctif $\tilde{C}$ pour l'extension du code par concaténation et nous noterons indifféremment C le code primaire ou son extension.

Exemple 8.2.2. Le code ASCII [2] permet le codage de l'ensemble $\mathbb{X}$ contenant les 95 caractères imprimables et les 33 caractères de contrôle non-imprimables en mots de 7 bits. La colonne « Char » de la figure 8.1 représente les éléments de $\mathbb{X}$ tandis que la colonne « Decimal » représente la valeur décimale du code binaire sur $\mathbb{A}^7$, où $\mathbb{A} := \{0, 1\}$.

Ce codage fut développé aux débuts de l'ère informatique pour coder l'anglais. Aujourd'hui il est connu sous l'appelation ISO/CEI-646. Ce code fut étendu en différentes variantes de longueur fixe de 8-bits pour permettre le codage des principales langues européennes. La variante ISO/CEI-8859-1 (aussi connue sous le nom de latin-1) permet le codage de plusieurs langues d'Europe occidentale (dont le français).

Exemple 8.2.3. La standard UNICODE [16] est un codage $C : \mathbb{X} \to \mathbb{A}^{32}$ de longueur fixe de 32 bits où $\mathbb{X}$ est l'ensemble de lettres utilisées par tous les alphabets des langues principales du monde [3]. Il comporte en tout 1114112 symboles différents, donc $C(\mathbb{X})$ est un petit sous-ensemble de $\mathbb{A}^{32}$ qui a un cardinal $|\mathbb{A}^{32}| = 2^{32} = 4294967296$. Le

1. Il faut cependant garder à l'esprit que plusieurs des résultats obtenus dans ce chapitre se généralisent au cas des sources markoviennes.

2. American standard code for information interchange.

3. Y compris les alphabets latin, grec, cyrillique, arménien, hébreux, arabe, syriaque, thaana, devanagari, bengali, gurmukhi, gujarati, oriya, tamil, telugu, kannada, malayalam, sinhala, thaï, laotien, tibétain, myanmar, géorgien, hangul jamo, éthiopien, cherokee, aborigène canadien, syllabique, ogham, runique, khmer, philippinais, limbu, tai le, tai lue, mongole, phonétique, extensions du latin et du grec,

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

FIGURE 8.1 – Le code ISO/CEI-646 $C_{128} : \mathbb{X} \to \mathbb{A}^7$ de longueur fixe de 7 bits. Par exemple la lettre $z \in \mathbb{X}$ est codée en $C_{128}(z) = \langle 122 \rangle_{10} = \langle 7\mathrm{A} \rangle_{16} = \langle 1111010 \rangle_2 \in \mathbb{A}^7$. (Source de la figure : domaine public).

codage UTF-8, connu aussi comme ISO/CEI-10646, du standard UNICODE est un code $C : \mathbb{X} \to \{0,1\}^{8m}$ de longueur variable en $8m$ bits où $m \in \{1,2,3,4\}$, assurant la compatibilité descendante avec le code ASCII.

Exemple 8.2.4.

1. Soient $\mathbb{X} = \{1,2,3,4\}$, $\mathbb{A} = \{0,1\}$, $\mathbf{p} = (1/2, 1/4, 1/8, 1/8)$ et C le code

x	$C(x)$
1	0
2	10
3	110
4	111

   Nous calculons $\mathbb{E}(|C(X)| = 1.75$; il se trouve que pour cet exemple, l'entropie a aussi la valeur numérique $H(\mathbf{p}) = 1.75$.
2. Soient $\mathbb{X} = \{1,2,3\}$, $\mathbb{A} = \{0,1\}$, $\mathbf{p} = (1/3, 1/3, 1/3)$ et C le code

x	$C(x)$
1	0
2	10
3	11

   Nous calculons $\mathbb{E}(|C(X)| = 1.66$ tandis que la valeur numérique de l'entropie est $H(\mathbf{p}) = 1.58$.

buginais, balinais, linéaire A et B, italique, gothique, ugaritique, chypriote, phénicien, kharostht, systèmes de numérotation anciens, cunéiforme, ainsi que les parties constitutives d'idéogrammes chinois, japonais et coréens, des symboles mathématiques courants, etc.

Pour les deux codes de l'exemple précédent 8.2.4, nous observons que si $\alpha \in \mathbb{A}^*$ est un mot, image par C d'un mot $\xi \in \mathbb{X}^+$, nous pouvons de manière unique décoder α pour obtenir le mot duquel il est construit. Par exemple, pour le premier codage, si $\alpha = 0110111100110$, nous constatons aisément que $\alpha = C(134213)$. De la même façon, pour le deuxième exemple, le mot $\alpha = 0100011 = C(12113)$. Cette possibilité de décoder de manière unique est une caractéristique essentielle du code.

Définition 8.2.5. Le code C est dit

1. **non-singulier** si $C : \mathbb{X} \to \mathbb{A}^+$ est injective, i.e. $x \neq x' \Rightarrow C(x) \neq C(x')$,
2. **uniquement décodable** si son extension à $\mathbb{X}^+$ est injective,
3. **instantané** [4] si aucun mot $C(x), x \in \mathbb{X}$ du glossaire n'est préfixe d'un autre mot du glossaire.

Les familles de différents types de codes — sous familles de la famille de codes arbitraires $\mathcal{C}$ — sont notées respectivement $\mathcal{C}_{\text{inst}}$, $\mathcal{C}_{\text{ud}}$, $\mathcal{C}_{\text{ns}}$. Elles vérifient les inclusions suivantes

$$\mathcal{C}_{\text{inst}} \subseteq \mathcal{C}_{\text{ud}} \subseteq \mathcal{C}_{\text{ns}} \subseteq \mathcal{C}.$$

Exemple 8.2.6. Parmi les codes suivants,

x	$C_1(x)$	$C_2(x)$	$C_3(x)$	$C_4(x)$
1	0	0	10	0
2	0	010	00	10
3	1	01	11	110
4	1	10	110	111

C_1 est singulier car il n'est pas injectif, C_2 est non-singulier mais non uniquement décodable car $C_2^{-1}(010) = \{31, 2, 14\}$, C_3 est uniquement décodable mais non instantané car le mot de code 11 est préfixe du mot de code 110, C_4 est instantané.

Le qualificatif instantané pour le codes dont les mots de code ne sont pas préfixes d'autres mots de code provient du fait qu'un mot α nous est donné, son décodage se fait au fur et à mesure de sa lecture ; la fin de chaque mot de code le composant est déterminée uniquement par l'information accumulée par la lecture du mot de gauche à droite ; il n'est pas nécessaire de connaître les lettres ultérieures pour décoder. Par exemple, le code C sur $\mathbb{X} = \{0, 1\}$ défini par $C(0) = 0, C(1) = \underbrace{0 \ldots 0}_{n} 1$ est uniquement décodable ; cependant, si le mot $\underbrace{0 \ldots 0}_{n+1} 1$ est envoyé, nous devons attendre la réception des tous les $n + 2$ symboles pour savoir que le mot source était 01. Il n'est donc pas instantané (comme son caractère non-préfixe le laisse entendre).

Remarque 8.2.7. L'unique décodabilité d'un code C impose l'injectivité de l'extension de C sur $\mathbb{X}^+$. Ceci implique, qu'il existe alors une fonction partielle $\Delta : \mathbb{A}^+ \to \mathbb{X}^+$ qui est l'inverse de C lorsque cette dernière application est vue comme $C : \mathbb{X}^+ \to C(\mathbb{X}^+)$, i.e.

$$\forall \xi \in \mathbb{X}^+, \Delta(C(\xi)) = \xi.$$

4. Dans la littérature, les codes instantanés sont souvent appelés « codes préfixes », terminologie pour le moins absurde car ils sont précisément des « codes **non**-préfixes ».

Il est donc important de pouvoir décider si un code est uniquement décodable. Or si décider de l'instantanéité du code est facile car il suffit de comparer tous les couples (α, β) des mots de $\{C(x), x \in \mathbb{X}\}$ et vérifier que $\alpha \not\prec \beta$ et $\beta \not\prec \alpha$, décider si un code C est uniquement décodable est plus difficile car il faut s'assurer de l'injectivité de C sur $\mathbb{X}^+$.

On peut établir un critère (condition nécessaire et suffisante) d'unique décodabilité d'un code arbitraire à l'aide d'une procédure itérative. On note $S_0 = C(\mathbb{X})$ l'ensemble de mots du code (vu comme suffixes du mot vide). Pour tout $n \geq 1$, on construit itérativement la suite des suffixes relatifs possibles

$$S_n = \{\beta \in \mathbb{A}^+ : (\alpha\beta = \gamma) \vee (\gamma\beta = \alpha), \text{ pour } \alpha \in S_0, \gamma \in S_{n-1}\}.$$

On note ensuite $S^\infty = \cup_{n\geq 1} S_n$. Le critère est donné par le

Théorème 8.2.8. *[Sardinas-Patterson (1953)] Un code $C : \mathbb{X} \to \mathbb{A}^+$ est uniquement décodable si, et seulement si,*

$$S^\infty \cap S_0 = \emptyset.$$

Démonstration. À rédiger ultérieurement. □

Remarque 8.2.9. L'utilité du critère précédent peut paraître limitée car son application nécessite la détermination de la suite infinie de suffixes relatifs $(S_n)_{n\geq 1}$. Il n'en est rien. Notons en effet $L = \max_{x\in\mathbb{X}} |C(x)|$ la longueur maximale des mots du code. Il s'ensuit que tous les mots de $\beta \in S_n$ vérifient $|\beta| \leq L$ et, par conséquent, tous les ensembles de suffixes S_n sont finis. La suite (S_n) est donc nécessairement finalement périodique (i.e. il existe deux entiers $N, p \geq 1$ tels que pour tout $m \geq N : S_m = S_{m+p}$). Une instance particulière de périodicité finale est le cas $S_N = \emptyset$ pour un certain $N \geq 1$; il est alors évident que $S_m = \emptyset$ pour tout $m \geq N$. Cette remarque garantit que l'algorithme 8.2.10 s'arrête en un temps fini.

Pour $B, C \subseteq \mathbb{A}^+$, on note $B^{-1}C = \{\gamma \in \mathbb{A}^* : \beta\gamma \in C, \beta \in B\}$. Le symbole ε est toujours réservé pour le mot vide.

Algorithme 8.2.10. Sardinas-Patterson

Require: $C : \mathbb{X} \to \mathbb{A}^+$
Ensure: $\mathbb{1}_{\mathcal{C}_{\text{ud}}}(C)$
  $S_0 \leftarrow C(\mathbb{X})$
  $S_1 \leftarrow S_0^{-1}S_0 \setminus \{\varepsilon\}$
  $n \leftarrow 1$
  while $\varepsilon \notin S_n$ et $\forall m < n : S_n \neq S_m$ **do**
    $n \leftarrow n + 1$
    $S_n \leftarrow S_0^{-1}S_{n-1} \cup S_{n-1}^{-1}S_0$
  end while
  $\mathbb{1}_{\mathcal{C}_{\text{ud}}}(C) \leftarrow 1 - \mathbb{1}_{S_n}(\varepsilon)$

8.3 Théorème de Shannon sur le codage sans bruit

8.3.1 Inégalité de Kraft

Définition 8.3.1. Soit $\mathbb{A}$ un alphabet avec $A = \mathsf{card}\mathbb{A}$. On dit qu'une famille I d'entiers $(l_i)_{i\in I}$, $l_i \geq 1$ pour tout $i \in I$, vérifie l'**inégalité de Kraft** si

$$\sum_{i\in I} A^{-l_i} \leq 1.$$

Théorème 8.3.2. *(Kraft [43]).*

1. *Si $C : \mathbb{X} \to \mathbb{A}^+$ est un code instantané, alors la famille des longueurs des mots du code $(|C(x)|)_{x\in\mathbb{X}}$ doit vérifier l'inégalité de Kraft.*
2. *Réciproquement, si $(l_x)_{x\in\mathbb{X}}$ est une famille d'entiers ≥ 1 vérifiant l'inégalité de Kraft, alors il existe un code instantané $C : \mathbb{X} \to \mathbb{A}^+$ tel que $|C(x)| = l_x, \forall x \in \mathbb{X}$.*

Démonstration. Il est élémentaire de montrer que — pour tout entier $L \geq 1$ — l'arbre enraciné $\mathbb{T}$ possédant exactement L générations et dont chaque nœud possède A descendants est en bijection avec $\cup_{l=0}^{L}\mathbb{A}^l$.

(⇒) Supposons que nous ayons rangé l'ensemble de symboles $\mathbb{X}$ de sorte que la condition

$$x \preceq y \Longrightarrow l_x := |C(x_i)| \leq |C(y)| =: l_y$$

soit vérifiée. Notons $\alpha_x = C(x)$, pour $x \in \mathbb{X}$, le mot qui code le symbole x et $L = \max\{|C(x)|, x \in \mathbb{X}\}$. Puisque le mot α_x ne peut pas être préfixe d'aucun autre mot du code, lorsque nous le plaçons sur un nœud de l'arbre $\mathbb{T}$, nous devons exclure tout le sous-arbre émanant de α_x, par conséquence exclure A^{L-l_x} feuilles de la génération L. Nous recommençons ensuite avec le mot $\alpha_{x'} = C(x'), x \preceq x'$ qui sera placé sur une autre branche de l'arbre ; son placement exclura $A^{L-l_{x'}}$ nouvelles feuilles de la génération L. Pour que cette opération puisse être répétée avec tous les mots du code, il faut que le nombre total de feuilles exclues soit majoré par le nombre total de feuilles de la génération L, i.e.

$$\sum_{x\in\mathbb{X}} A^{L-|C(x)|} \leq A^L \Longrightarrow \sum_{x\in\mathbb{X}} A^{-|C(x)|} \leq 1.$$

(⇐) Sans perte de généralité, nous pouvons supposer que les x soient rangées en $x_1, \ldots, x_{\mathsf{card}\mathbb{X}}$ de sorte que $l_1 \leq l_2 \leq \ldots \leq l_{\mathsf{card}\mathbb{X}}$, où $l_i = l_{x_1}$. Nous choisissons un nœud arbitraire de la l_1^e génération de $\mathbb{T}$ et nous assignons le mot qui correspond au nœud choisi à $C(x_1)$; nous excluons simultanément le sous-arbre enraciné à $\alpha_1 = C(x_1)$ ce qui revient à exclure A^{L-l_1} feuilles de la génération L. Il est évident que

$$\sum_{x\in\mathbb{X}} A^{-l_x} \leq 1 \Rightarrow \sum_{x\in\mathbb{X}} A^{L-l_x} \leq A^L \Rightarrow A^{L-l_1} < A^L;$$

Il y a donc au moins une feuille qui subsiste après l'exclusion de ce sous-arbre et donc toute la branche qui la relie à la génération l_1. Nous allons ensuite choisir un nœud arbitraire de la génération l_2 pour assigner le mot $C(x_2)$. Nous savons que ceci est possible car nous venons de montrer qu'il y a au moins une branche

reliant la génération l_1 avec une feuille de la génération L ; cette branche rencontrera nécessairement un nœud de la génération l_2. Nous recommençons donc l'argument :

$$\sum_{x\in\mathbb{X}} A^{-l_x} \leq 1 \Rightarrow \sum_{x\in\mathbb{X}} A^{L-l_x} \leq A^L \Rightarrow A^{L-l_1} + A^{L-l_2} < A^L;$$

il subsiste donc au moins une feuille de la génération L après effacement des deux sous-arbres précédents. L'inégalité de Kraft vérifiée par les (l_x) garantit que card$\mathbb{X}$ mots puissent être placés sans qu'aucun ne soit préfixe d'un autre.

□

8.3.2 Codes optimaux

L'inégalité de Kraft est une condition nécessaire et suffisante pour l'existence d'un code instantané avec longueurs des mots (l_x). Cependant, toute permutation des (l_x) est encore une famille de longueurs acceptables pour des mots d'un code. Si les l_x ne sont pas constantes et la probabilité sous-jacente sur $\mathbb{X}$ n'est pas l'uniforme, parmi ce deux codes, il y a nécessairement un qui est meilleur que l'autre. Il est donc évident qu'il ne suffit pas de vérifier l'inégalité de Kraft pour construire un code optimal. Il faut plutôt minimiser $\mathbb{E}|C(X)| = \sum_{x\in\mathbb{X}} p(x)l_x$ sous la contrainte $\sum_{x\in\mathbb{X}} A^{-l_x} \leq 1$. Nous utilisons la méthode traditionnelle des multiplicateurs de Lagrange pour trouver la solution qui sature la contrainte, i.e. impose $\sum_x A^{-l_x} - 1 = 0$; nous considérons par conséquent la fonctionnelle

$$J(\boldsymbol{\ell}) = \sum_x p(x)l_x + \lambda(\sum_x A^{-l_x} - 1),$$

où $\boldsymbol{\ell} = (l_x)_{x\in\mathbb{X}}$. En dérivant, nous obtenons

$$\forall x, \frac{\partial J}{\partial l_x} = p(x) - \lambda A^{-l_x} \log A = 0,$$

équation qui admet comme solution $A^{-l_x^*} = \frac{p(x)}{\lambda \log A}$. La saturation de la contrainte permet de déterminer la constante $\lambda = 1/\log A$ et par conséquent, exprimer la solution sous la forme : $l_x^* = -\log_A p(x)$. Si les l_x^* étaient tous des entiers, on pourrait alors construire un code instantané C^* qui aurait une longueur moyenne de $\mathbb{E}|C^*(X)| = -\sum_x p(x)\log_A p(x) = H_A(X)$. Cependant, les l_x^* ne sont pas nécessairement des entiers. Nous avons donc le

Théorème 8.3.3 ((Shannon [62]).). *Pour tout code instantané $C : \mathbb{X} \to \mathbb{A}^+$, avec $A =$ card$\mathbb{A}$, nous avons la minoration*

$$\mathbb{E}|C(X)| \geq H_A(X),$$

avec égalité si, et seulement si, $A^{-|C(x)|} = p(x), \forall x$.

Démonstration. Introduisons un nouveau vecteur de probabilité $r(x) = \frac{A^{-|C(x)|}}{\sum_y A^{-|C(y)|}}$ sur

$\mathbb{X}$. Nous aurons

$$\begin{aligned}\mathbb{E}|C(X)| - H_A(X) &= \sum_x p(x)|C(x)| + \sum_x p(x)\log_A p(x)\\ &= -\sum_x p(x)\log_A A^{-|C(x)|} + \sum_x p(x)\log_A p(x)\\ &= \sum_x p(x)\log_A \frac{p(x)}{r(x)} - \log(\sum_y A^{-|C(y)|})\\ &= D(\mathbf{p}\|\mathbf{r}) + \log\frac{1}{\sum_y A^{-|C(y)|}}\\ &\geq 0,\end{aligned}$$

car l'entropie relative D est positive et $\sum_y A^{-|C(y)|} \leq 1$. Nous aurons égalité si $D(\mathbf{p}\|\mathbf{r}) = 0$ et $\sum_y A^{-|C(y)|} = 1$. □

Théorème 8.3.4. *Soit X une variable aléatoire de loi décrite par le vecteur de probabilité $\mathbf{p}$. Il existe un code instantané C, tel que*

$$H_A(\mathbf{p}) \leq \mathbb{E}|C(X)| < H_A(\mathbf{p}) + 1.$$

Démonstration. En général, nous ne pouvons pas nous attendre à ce que les valeurs $l_x^* = -\log_A p(x)$ qui minimisent la fonctionnelle d'optimisation J — introduite en début du paragraphe — soient entières. Cependant, chaque intervalle $[-\log_A p(x), -\log_A p(x) + 1[$ contient nécessairement un entier l_x. La famille des entiers $(l_x)_{x\in\mathbb{X}}$ vérifie l'inégalité de Kraft car $A^{-l_x} \leq A^{-l_x^*}$, pour tout x. Par conséquent, il existe un code instantané C qui admet cette famille comme famille de longueurs de mots du code. De l'encadrement $l_x^* \leq l_x \leq l_x^* + 1$ découle immédiatement l'inégalité $H_A(\mathbf{p}) \leq \mathbb{E}|C(X)| < H_A(\mathbf{p}) + 1$. □

Nous avons établi l'inégalité de Kraft pour des codes $C \in \mathcal{C}_{\text{inst}}$. Or la classe $\mathcal{C}_{\text{ud}}$ est plus grande que $C \in \mathcal{C}_{\text{inst}}$. Nous pouvons donc légitimement nous interroger si en optimisant $\mathbb{E}|C(X)|$ sur la famille des codes uniquement décodables nous pouvons améliorer la borne de l'entropie. Le théorème suivant répond négativement à cette question.

Théorème 8.3.5. *(McMillan [51]).*

1. *Tout code $C \in \mathcal{C}_{ud}$ vérifie $\sum_x A^{-|C(x)|} \leq 1$.*
2. *Réciproquement, pour toute famille d'entiers (l_x) vérifiant l'inégalité de Kraft, il existe un code $C \in \mathcal{C}_{ud}$ tel que $|C(x)| = l_x, \forall x$.*

Démonstration. On utilise le même symbole C pour noter le code primaire et son extension sur l'alphabet $\mathbb{X}^k$ de blocs de taille k.

1. On note

$$\begin{aligned}\mathbb{X}_j^k &= \{\xi \in \mathbb{X}^k : |C(\xi)| = j\};\\ \mathbb{B}_{i_1,\ldots,i_k} &= \{\alpha \in \mathbb{A}^+ : \alpha = \alpha_1\cdots\alpha_k, |\alpha_1| = i_1, \ldots, |\alpha_k| = i_k\}.\end{aligned}$$

Introduisons les symboles $L = \max_x |C(x)|$ et $\nu_l = \text{card}\mathbb{X}_l^1$. Pour tout $k \geq 1$:

$$\left(\sum_x A^{-|C(x)|}\right)^k = \left(\sum_{l=1}^{L}\sum_{x\in\mathbb{X}_l^1} A^{-l}\right)^k = \left(\sum_{l=1}^{L}\nu_l A^{-l}\right)^k = \sum_{l=k}^{kL} N_l A^{-l},$$

où

$$N_l = \sum_{\substack{m_1,\dots,m_k \\ m_1+\dots+m_k=l}} \nu_{m_1} \cdots \nu_{m_k} = \mathsf{card}\mathbb{X}_l^k.$$

L'unique décodabilité du code C implique que pour chaque mot $\beta \in \mathbb{B}_{i_1,\dots,i_k}$ avec $i_1 + \dots + i_k = l$, il existe *au plus* un mot $\xi \in \mathbb{X}_l^k$ tel que $C(\xi) = \beta$. Il est évident par ailleurs que

$$\bigcup_{\substack{i_1,\dots,i_k \\ i_1+\dots+i_k=l}} \mathbb{B}_{i_1,\dots,i_k} = \mathbb{A}^l.$$

Ceci entraîne la majoration

$$N_l = \mathsf{card}\mathbb{X}_l^k \leq \mathsf{card}\left(\bigcup_{\substack{i_1,\dots,i_k \\ i_1+\dots+i_k=l}} \mathbb{B}_{i_1,\dots,i_k} \right) = \mathsf{card}\mathbb{A}^l = A^l.$$

En remplaçant, on conclut que

$$\left(\sum_x A^{-|C(x)|} \right)^k \leq \sum_{l=k}^{kL} A^l A^{-l} = kL - k + 1 \leq kL,$$

ce qui entraîne que $\sum_x A^{-|C(x)|} \leq k^{1/k} L^{1/k}$. Cette majoration étant vraie pour tout $k \geq 1$, restera vraie en passant à la limite $k \to \infty$, établissant ainsi l'inégalité de Kraft pour les longueurs $|C(x)|$ du code C.

2. Par le théorème 8.3.2, nous savons que si l'inégalité de Kraft est vérifiée pour une famille d'entiers (l_x), il existe un code instantané C qui admet cette famille d'entiers comme famille des longueurs. Or tout code instantané est uniquement décodable.

□

Définition 8.3.6. 1. Un code C est dit $\mathcal{K}$-**optimal** pour une classe particulière $\mathcal{K} \subset \mathcal{C}$ si pour tout code $C' \in \mathcal{K}$ on a

$$\mathbb{E}|C(X)| \leq \mathbb{E}|C'(X)|.$$

2. Un code $C \in \mathcal{C}_{\text{ud}}$ qui sature la borne de Shannon, i.e. qui vérifie $\mathbb{E}|C(X)| = H_A(X)$, est dit **absolument optimal**.

Remarque 8.3.7. Nous avons vu en §7.2.2 qu'une interprétation de l'entropie est le nombre moyen de questions à réponse binaire que nous devons poser pour déterminer la valeur d'une variable aléatoire X. En marquant 0 chaque fois que la réponse est « non » et 1 chaque fois que la réponse est « oui », lorsque nous parcourons l'arbre de décision, nous obtenons le code optimal C (cf. l'exemple de la figure 7.3).

En général nous ne pouvons pas espérer qu'un code primaire $C : \mathbb{X} \to \mathbb{A}^+$ soit absolument optimal ; la saturation de la borne de Shannon pourra s'obtenir uniquement pour des codes étendus sur $\mathbb{X}^k$, lorsque $k \to \infty$. Il est donc utile d'avoir des résultats permettant d'ordonner de manière abstraite les codes par leur optimalité.

Lemme 8.3.8. *Un code C optimal pour la classe $\mathcal{C}_{inst}$ est optimal pour la classe $\mathcal{C}_{ud}$.*

Démonstration. Supposons qu'il existe un code $C_2 \in \mathcal{C}_{\text{ud}}$ meilleur que C. Nous aurons alors $\mathbb{E}|C_2(X)| < \mathbb{E}|C(X)|$. Notons $l_2(x) = |C_2(x)|$ pour $x \in \mathbb{X}$; le théorème de McMillan 8.3.5 implique que les longueurs $(l_2(x))$ vérifient l'inégalité de Kraft. Par le théorème de Kraft 8.3.2, nous savons qu'il existe un code $C_1 \in \mathcal{C}_{\text{inst}}$ qui admet a $|C_1(x)| = l_2(x)$. Nous avons donc trouvé un code instantané C_1 strictement meilleur que C ce qui contredit l'optimalité de ce dernier. □

8.3.3 Algorithme de Huffman pour la construction de codes optimaux

Notons $M = \text{card}\mathbb{X}$ et supposons que $|\mathbb{A}| = 2$; par conséquent nous pouvons, sans perte de généralité, identifier $\mathbb{X} \simeq \{1, \ldots, M\}$ et $\mathbb{A} = \{0, 1\}$. Si $C : \mathbb{X} \to \mathbb{A}^+$ est un code, nous notons $(l_i)_{i=1,\ldots,M}$ la famille des longueurs de ses mots. Si $\mathbf{p} \in \text{PV}_M$ est un vecteur de probabilité, nous supposons que l'identification de $\mathbb{X}$ avec $\{1, \ldots, M\}$ s'est faite en imposant : $p_1 \geq \ldots \geq p_M$ et que si $p_i = \ldots = p_{i+s}$, pour un $s \geq 1$, alors $l_i \leq \ldots \leq l_{i+s}$. Avec ces conventions en vigueur, nous avons le

Lemme 8.3.9. *Sous les conditions et conventions précédentes, si C est optimal dans $\mathcal{C}_{inst}$, alors*

1. *$p_i > p_k \Rightarrow l_i \leq l_k$,*
2. *les deux symboles les moins probables, $M-1$ et M, ont de longueurs de code identiques $l_{M-1} = l_M$,*
3. *parmi tous les mots de code de longueur maximale, l_M, il en existe deux dont les $l_M - 1$ premières lettres coïncident (ils diffèrent uniquement par leur dernière lettre).*

Démonstration.

1. Si $p_i > p_k$ et $l_i > l_k$, on peut construire un code instantané C', en échangeant i et k :
$$C'(j) = \begin{cases} C(j) & \text{si } j \neq i, k, \\ C(k) & \text{si } j = i, \\ C(i) & \text{si } j = k. \end{cases}$$
Nous avons alors
$$\begin{aligned} \mathbb{E}|C'(X)| - \mathbb{E}|C(X)| &= p_i l_k + p_k l_i - p_i l_i - p_k l_k \\ &= (p_i - p_k)(l_k - l_i) \\ &< 0. \end{aligned}$$
Donc C' est meilleur que C en contradiction avec l'optimalité supposée de ce dernier.
2. Nous avons toujours $l_{M-1} \leq l_M$ car
— si $p_{M-1} > p_M$ alors $l_{M-1} \leq l_M$ par 1;
— si $p_{M-1} = p_M$ alors $l_{M-1} \leq l_M$ par l'hypothèse de rangement.
Étant donné que le code C est instantané les mots $C(M-1)$ et $C(M)$ correspondent à des feuilles de branches différentes de l'arbre de mots du code. Si $l_{M-1} < l_M$, on peut effacer la dernière lettre de $C(M)$; ce nouveau mot de longueur $l_M - 1$ correspondra encore à un mot de code instantané car ce nouveau mot sera encore sur une branche différente que $C(M-1)$. Par ailleurs, ce nouveau code est strictement meilleur que C. Nous pouvons répéter cet argument jusqu'à ce que nous obtenions un code avec $l_{M-1} = l_M$, auquel cas nous devons arrêter cette procédure car si nous enlevons encore une lettre nous obtiendrons un préfixe de $C(M-1)$.

3. S'il existe deux mots de code de longueur maximale l_M qui ne coïncident pas sur les $l_M - 1$ premières lettres, nous pouvons effacer la dernière lettre de chacun d'eux pour obtenir un code qui reste instantané et qui est meilleur que C.

□

Lemme 8.3.10. *Supposons que $C : \mathbb{X} \to \mathbb{A}^+$ est un code instantané, $\mathbb{A} = \{0,1\}$ et $\mathbf{p}$ un vecteur de probabilité. Nous utilisons les mêmes conventions que celles en vigueur pour le lemme 8.3.9 quant à l'ordre de $\mathbb{X} = \{1,\dots,M\}$. Introduire*[5] *$\mathbb{Y} \simeq \{1,\dots,M-2, y_{M-1,M}\}$ avec un vecteur de probabilité $\mathbf{q} = (p_1,\dots,p_{M-2},p_{M-1}+p_M)$. Supposons que $C_2 : \mathbb{Y} \to \mathbb{A}^+$ est un code instantané optimal. Alors, le code $C_1 : \mathbb{X} \to \mathbb{A}^+$, défini par*

$$C_1(i) = \begin{cases} C_2(i) & \text{si } i = 1,\dots,M-2, \\ C_2(y_{M-1,M})0 & \text{si } i = M-1, \\ C_2(y_{M-1,M})1 & \text{si } i = M, \end{cases}$$

est un code optimal.

Démonstration. Supposons qu'au contraire C_1 ne soit pas optimal; il existe donc un code $C_1' : \mathbb{X} \to \mathbb{A}^+$ qui lui est strictement meilleur. Par l'affirmation 2 du lemme 8.3.9, on a $|C_1'(M-1)| = |C_1'(M)|$. Par l'affirmation 3 de ce même lemme, il existe $\alpha \in \mathbb{A}^+$ et $a,b \in \mathbb{A}$ tels que $C_1'(M-1) = \alpha a$ et $C_1'(M) = \alpha b$. En recombinant les symboles $M-1$ et M en le symbole $y_{M-1,M}$, on construit un code $C_2' : \mathbb{Y} \to \mathbb{A}^+$ par

$$C_2'(i) = \begin{cases} C_1'(i) & \text{si } i = 1,\dots,M-2, \\ \alpha & \text{si } i = y_{M-1,M}. \end{cases}$$

Nous avons

$$\begin{aligned} \mathbb{E}|C_2'(X)| &= \sum_{i=1}^{M-2} p_i |C_1'(i)| + (p_{M-1}+p_M)|\alpha| \\ &= \sum_{i=1}^{M} p_i |C_1'(i)| - p_{M-1}|C_1'(M-1)| - p_M|C_1'(M)| + (p_{M-1}+p_M)(|C_1'(M-1)|-1) \\ &= \mathbb{E}|C_1'(X)| - (p_{M-1}+p_M), \quad \text{car } |C_1'(M-1)| = |C_1'(M)|, \\ &< \mathbb{E}|C_1(X)| - (p_{M-1}+p_M), \quad \text{car } C_1' \text{ est supposé strictement meilleur que } C_1, \\ &= \mathbb{E}|C_2(X)|, \end{aligned}$$

contredisant ainsi l'optimalité de C_2. □

Définition 8.3.11. Un arbre enraciné T est un **arbre binaire** si
- soit T est vide,
- soit T peut être écrit récursivement comme un triplet (r, T_g, T_d), où r est un nœud (la racine de l'arbre) et T_g et T_d sont deux sous arbres (respectivement gauche et droit).

Une collection d'arbres est appelée une **forêt**.

Remarque 8.3.12. 1. Soit $\mathbb{X}$ un ensemble de symboles. Un symbole $x \in \mathbb{X}$ peut être considéré comme l'arbre binaire ponctuel $(x, \emptyset, \emptyset)$. L'ensemble $\mathbb{X}$ peut donc être vu comme une forêt d'arbres ponctuels.

5. L'ensemble $\mathbb{Y}$ n'est donc pas nécessairement ordonné selon la convention en vigueur pour $\mathbb{X}$.

2. Un arbre binaire est dit de type $0-2$ si chaque sommet de l'arbre a 0 ou 2 descendants. De manière équivalente, un arbre est de type $0-2$ si dans sa décomposition récursive (r, T_g, T_d) les deux sous-arbres sont soit tous les deux vides, soit tous les deux non-vides.
3. Si $S = (u, S_g, S_d)$ et $T = (v, T_g, T_d)$ sont deux arbres enracinés binaires arbitraires, $S \circ V = (z, S, T)$ est un nouvel arbre enracine binaire ayant les deux arbres précédents comme sous-arbres gauche et droit. Il est souligné que z est un nouveau nœud qui constitue la racine du nouvel arbre.

Lemme 8.3.13. *Soient* $\mathbf{p} \in PV_{\mathbb{X}}$ *et* $\mathbb{A} = \{0,1\}$. *Alors il existe un code* $C : \mathbb{X} \to \mathbb{A}^+$ *instantané optimal qui vérifie les propriétés 1–3 du lemme 8.3.9.*

Nous pouvons maintenant présenter l'algorithme de Huffman [35] de construction de codes optimaux sous forme de pseudo-code.

Algorithme 8.3.14. Huffman

Require: Vecteur de probabilité $\mathbf{p}$.
Ensure: Forêt $\mathcal{F} = \{T_1, T_2\}$ composée de deux arbres binaires non vides T_1 et T_2.

```
M ← dim p
i ← 1
F ← ∅
repeat
    t_i ← (i, ∅, ∅)
    w(t_i) ← p(i)
    F ← F ∪ {t_i}
    i ← i + 1
until i > M;
repeat
    T_1 ← arg min_{t∈F} w(t)
    T_2 ← arg min_{t∈F\T_1} w(t)
    T ← T_1 ∘ T_2
    F ← F \ {T_1, T_2}
    F ← F ∪ {T}
    w(T) ← w(T_1) + w(T_2)
until card F = 2.
```

Théorème 8.3.15. *Le code de Huffman, consistant à l'attribuer la lettre 0 à tout sous-arbre gauche et la lettre 1 à tout sous-arbre droit dans l'algorithme précédent, est un code instantané optimal.*

Exemple 8.3.16.

8.3.4 Examen critique du code de Huffman

Le code de Huffman nous fournit un codage pour un alphabet de source $\mathbb{X}$ et un vecteur de probabilité $\mathbf{p}$ donnés. Lorsque les probabilités $p(x)$, pour $x \in \mathbb{X}$ sont des puissances négatives de 2, le code de Huffman est optimal (sature la borne d'entropie) car les nombres $-\log p(x)$ sont des entiers pour tout $x \in \mathbb{X}$ est déterminent la longueur des mots de code. En général, les probabilités $p(x)$ ne sont pas des puissances négatives de 2 et, dans ce cas, le code de Huffman ne fournit des codes optimaux qu'asymptotiquement pour des alphabets étendus $\mathbb{X}^k$ en des blocs de taille k grande. Dans ce cas,

la taille de la table du code, i.e. du vecteur $(C(\mathbf{x}))_{\mathbf{x}\in\mathbb{X}^k}$ à $|\mathbb{X}|^k$ composantes, croît exponentiellement avec la taille k des blocs. Or pour pouvoir décoder les messages, et étant donné que le code de Huffman n'est pas uniquement déterminé par $\mathbf{p}$, la connaissance de la table du code est nécessaire par le destinataire. Ce qui implique que la table doit aussi être transmise, ce qui dégrade les qualités du codage de Huffman.

Un autre inconvénient du code est que sa table dépend très fortement du vecteur $\mathbf{p}$. Si par exemple on veut transmettre du texte bilingue français/anglais, le vecteur de probabilité variera selon que la portion codée est en français ou en anglais. Or le code de Huffman n'est pas adapté à cette situation.

Pour pallier ces inconvénients, d'autres codages on été introduits. Les codes arithmétiques, dont un exemple est présenté en §8.4.1, nécessite toujours la connaissance du vecteur de probabilité mais cette information est suffisante pour coder et décoder le message sans qu'une table complète pour des blocs de grande taille soit nécessaire. Les codes par dictionnaire, dont un exemple est présenté en §8.4.2, ne nécessitent même pas la connaissance du vecteur de probabilité et sont donc adaptés pour des codages de textes plurilingues.

8.4 Autres types de codes

8.4.1 Le code arithmétique de Shannon-Fano-Elias

Dans le folklore de la théorie de l'information, le code présenté dans ce paragraphe porte le nom de code de Shannon, Fano et Elias (FSE). Cependant, selon [55], la première trace écrite de ce type de codage se trouve dans le livre [36][6]. Ce code est un précurseur du standard JPEG.

Afin de présenter précisément cet algorithme on introduit la

Notation 8.4.1. On note $A = |\mathbb{A}|$ (on identifie $\mathbb{A} = \{0, \dots A-1\}$) et on définit deux applications num et seq comme suit :

— L'application num est définie par la formule

$$\mathbb{A}^+ \ni \beta = b_1 \cdots b_l \mapsto \mathsf{num}(\beta) = \sum_{i=1}^{l} \frac{b_i}{A^i} \in [0,1[.$$

— Pour un $x \in [0,1[$, on définit seq comme le pseudo-inverse de num par

$$[0,1[\ni x \mapsto \mathsf{seq}(x) = [\beta] \in \mathbb{A}^{\mathbb{N}}/\sim,$$

où deux suites $\beta = b_1 \cdots b_n(A-1)(A-1)\cdots$ et $\gamma = b_1 \cdots (b_n+1)00\cdots$, pour $0 \le b_n \le A-2$, sont équivalentes — et on note $\beta \sim \gamma$ — car $\mathsf{num}(\beta) = \mathsf{num}(\gamma)$. Pour pallier cet inconvénient, on peut choisir systématiquement comme représentant de la classe d'équivalence la suite qui se termine par une suite infinie de zéros.

Par ailleurs, une suite de $\mathbb{A}^{\mathbb{N}}$ qui se termine par une suite infinie de zéros, sera identifiée avec une suite de $\mathbb{A}^L$ où L est sa dernière lettre non nulle.

On considère maintenant un alphabet d'entrée $\mathbb{X}$ de cardinal $M = \mathsf{card}\mathbb{X} \ge 1$ et une probabilité de génération de lettres déterminée par le vecteur de probabilité $\mathbf{p} \in \mathbb{R}_+^M$

6. L'auteur n'a pas pu accéder à la référence [36].

(avec $\sum_{x\in\mathbb{X}} p(x) = 1$). Le symbole précis qui représente chaque lettre d'entrée ne jouera aucun rôle ; on identifie donc $\mathbb{X}$ avec l'ensemble $\{0,\dots,M-1\}$. Cette identification induit un ordre naturel sur le lettres de l'alphabet. Nous pouvons donc définir la fonction de répartition d'une variable aléatoire X prenant des valeurs dans $\mathbb{X}$ selon la loi décrite par $\mathbf{p}$:

$$F(z) = \mathbb{P}(X \le z) = \sum_{y\in\mathbb{X}, y\le x} p(y),$$
$$F(z-) = \mathbb{P}(X < z) = \sum_{y\in\mathbb{X}, y< x} p(y),$$

pour tout $z \in \mathbb{R}$.

Exemple 8.4.2. Soit l'alphabet de la source $\mathbb{X} = \{a,b,c\} \simeq \{0,1,2\}$ et un symbole $X \in \mathbb{X}$ émis avec $\mathbb{P}(X = x) = p(x)$., avec $p(0) = 0.2$, $p(1) = 0.5$ et $p(2) = 0.3$.

La fonction de répartition pour cet exemple est donnée dans la figure 8.2.

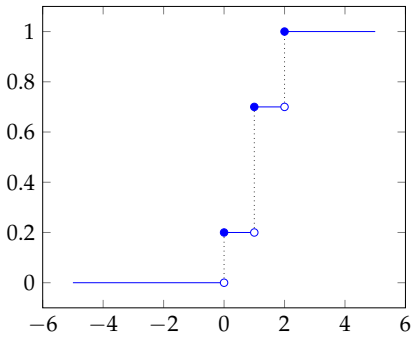


FIGURE 8.2 – Fonction de répartition pour l'exemple 8.4.2.

Derrière cet algorithme se trouvent quelqes idées simples :
— Chaque lettre $x \in \mathbb{X}$ va être associée, non pas à un mot de $\mathbb{A}^+$ mais à un sous-intervalle I_x de $[0,1[$, où $I_x = [F(x-), F(x)[= [F(x-), F(x-) + p(x)[$, de longueur $p(x)$. Le tableau 8.1 indique cette représentation pour l'exemple 8.4.2.
— Par construction les intervalles sont disjoints.

x	$p(x)$	$F(x-)$	I_x
$a \simeq 0$	0.2	0.0	$[0.0, 0.2[$
$b \simeq 1$	0.5	0.2	$[0.2, 0.7[$
$c \simeq 2$	0.3	0.7	$[0.7, 1.0[$

TABLE 8.1 – L'association $x \mapsto I_x$, $x \in \mathbb{X}$, pour l'exemple 8.4.2.

— Maintenant, si on veut coder un mot de 2 lettres, par exemple $\xi = bc$, nous commençons par lire la première lettre b. L'intervalle qui représentera ξ sera donc un sous-intervalle de $I_b = [0.2, 0.7[$. Nous devons ensuite subdiviser l'intervalle I_b en sous-intervalles semi-ouverts contigus de I_b qui auront des rapports de longueurs $|I_{bx}|/|I_b| = p(x)$, pour $x \in \mathbb{X}$. Comme la deuxième lettre est c, on conclut que $|I_{bc}| = 0.5 \times 0.3 = 0.15$ et son extrémité gauche sera au point $0.2 + 0.5 \times (0.2 + 0.5) = 0.55$.

— Plus généralement, pour coder un mot $\xi = x_1 \cdots x_K \in \mathbb{X}^N$, on commence par déterminer l'intervalle $I_{x_1} = [g_1, g_1 + l_1[$, où $g_1 = F(x_1-)$, de longueur $l_1 = p(x_1)$ et ensuite procéder récursivement en définissant, pour $1 < k \leq N$:

$$\begin{aligned} g_k &= g_{k-1} + F(x_k-)l_{k-1} \\ l_k &= p(x_k)l_{k-1} \\ I_{x_1 \ldots x_k} &= [g_k, g_k + l_k[. \end{aligned}$$

On peut même commencer la récurrence ci-dessus à $k = 0$, en associant l'intervalle $I_\varepsilon = [0, 1[$ à l'unique mot de longueur 0, à savoir le mot vide.

— Maintenant, on doit transmettre l'information qui permet au récipiendaire du message de déterminer l'intervalle I_ξ. Or, on ne peut pas transmettre un intervalle mais une suite finie de bits. Supposons que nous avons déterminé $I_\xi = [g, g + l[$. Étant donné qu'il existe un unique I_ξ correspondant à un $\xi \in \mathbb{X}^N$ donné, tout $y \in I_\xi$ peut servir à représenter I_ξ, par exemple $y = g + l/2$. Cependant, on ne peut pas transmettre y non plus car, en général il s'agit d'un réel, représenté par une suite (en principe infinie) $\beta = \mathsf{seq}(y) \in \mathbb{A}^{\mathbb{N}}$.

— Supposons que nous tronquons la suite $\beta = \mathsf{seq}(y)$ à L digits (L sera déterminé dans la suite). On dispose alors d'une suite tronquée $\hat{\beta} = \beta\restriction_L \in \mathbb{A}^L$ dont la valeur numérique $\mathsf{num}(\hat{\beta}) = \sum_{k=1}^{L} \frac{\beta_i}{A^i} = \hat{y} \in [0, 1[$. On a alors $y - \hat{y} = \sum_{i \geq L+1} \frac{\beta_i}{A^i} \leq (A-1) \sum_{i \geq L+1} \frac{1}{A^i} = \frac{1}{A^L}$. En choisissant $\frac{1}{A^L} < \frac{l}{2}$, i.e.

$$L > -\log_A(l/2) = \lceil -\log_A(l/2) \rceil,$$

on est sûr que l'approximation $\hat{y}$ de y est encore dans l'intervalle I_ξ ; elle peut donc servir à représenter I_ξ.

— Il s'ensuit que nous pouvons coder le mot $\xi \in \mathbb{X}^N$ en posant $C(\xi) = \hat{\beta}\restriction_L$.

— Il est à souligner que le nombre de bits $L := L_\xi$ nécessaire dépend du mot $\xi = x_1 \cdots x_N$ à travers la dépendance de la longueur $l := l_\xi$ de l'intervalle I_ξ. En effet, $l_\xi = p(x_1) \cdots p(x_N)$. Par conséquent $L_\xi = \lceil -\sum_{k=1}^{N} \log_A p(x_k) + \log_A 2 \rceil$.

— Le décodage s'effectue en suivant le cheminement logique inverse.

On est alors en mesure de présenter le codage et le décodage SFE sous forme algorithmique.

Algorithme 8.4.3. Codage SFE

Require: Vecteur de probabilité **p** (donc fonction de répartition F),
mot à coder $\xi = x_1 \cdots x_{|\xi|} \in \mathbb{X}^+$.
Ensure: N longueur du mot ξ et $\beta = C(\xi) \in \mathbb{A}^+$.
$N \leftarrow |\xi|$
$k \leftarrow 0$
$g_k \leftarrow 0$
$l_k \leftarrow 1$
while $k < N$ **do**
  $k \leftarrow k + 1$
  $g_k \leftarrow g_{k-1} + F(x_k-)l_{k-1}$
  $l_k \leftarrow p(x_k)l_{k-1}$
end while
$L \leftarrow \lceil -\log_A(l_N/2) \rceil$
$\beta \leftarrow \mathsf{seq}(g_N + \frac{l_N}{2})\restriction_L \in \mathbb{A}^L$

Algorithme 8.4.4. Décodage SFE

Require: $\hat{\beta}$ qui finit par des 0, N et $\rho(\Leftrightarrow F)$.
Ensure: $\xi \in \mathbb{X}^N$.
  $k \leftarrow 1$
  $r_1 \leftarrow \mathsf{num}(\hat{\beta})$
  $x_1 \leftarrow \arg\max_y \mathbb{1}_{[F(y-),F(y)[}(r_1)$
  $\xi \leftarrow \xi x_k$
  while $k < N$ **do**
    $k \leftarrow k+1$
    $r_k \leftarrow \frac{r_{k-1}-F(x_{k-1}-)}{\rho(x_{k-1})}$
    $x_k \leftarrow \arg\max_y \mathbb{1}_{[F(y-),F(y)[}(r_k)$
    $\xi \leftarrow \xi x_k$
  end while

Théorème 8.4.5. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires, à valeurs dans un alphabet $\mathbb{X}$, indépendantes et identiquement distribuées. On note $X^{(N)} = X_1 \cdots X_N \in \mathbb{X}^N$ le N-uplet de N copies indépendantes de la même variable aléatoire X distribuée selon la loi commune et $C : \mathbb{X} \to \mathbb{A}^+$ le code de Shannon-Fano-Elias (et son extension sur $\mathbb{X}^+$). Alors*

$$H_A(X) + \frac{\log_A 2}{N} \leq \frac{\mathbb{E}|C(X^{(N)})|}{N} \leq H_A(X) + \frac{1+\log_A 2}{N}.$$

Démonstration. On a vu que $L_{X^{(N)}} = \lceil -\sum_{k=1}^N \log_A p(X_k) + \log_A 2 \rceil$, par conséquent

$$-\sum_{k=1}^N \log_A p(x_k) + \frac{\log_A 2}{N} \leq |C(X^{(N)})| = L_{X^{(N)}} \leq -\sum_{k=1}^N \log_A p(x_k) + \log_A 2 + 1.$$

En multipliant tous le membres de l'inégalité par $\prod_{n=1}^n p(x_n)$ et en intégrant sur les mots, on obtient

$$H_A(X_1,\ldots,X_N) + \log_A 2 \leq \mathbb{E}|C(X^{(N)})| \leq H_A(X_1,\ldots,X_N) + \log_A 2 + 1.$$

On conclut par l'additivité $H_A(X_1,\ldots,X_N) = NH_A(X)$ de l'entropie dans le cas indépendant. □

Une généralisation de cet algorithme qui consiste à utiliser deux ordres différents pour indexer les lettres de l'alphabet est utilisée dans [37] pour proposer un algorithme de cryptage.

8.4.2 Codage universel et algorithme de Lempel-Ziv

Les défauts des codages précédents sont corrigés par le codage de Lempel et Ziv [74, 73] et ses variantes ultérieures. Nous présentons l'algorithme connu sous l'acronyme LZ78 qui est largement utilisé en informatique, par exemple par l'utilitaire `gzip` en système d'exploitation linux/unix.

L'algorithme effectue un premier passage sur le texte à coder et construit un dictionnaire. Le premier mot du dictionnaire est le mot vide et ensuite des bouts de mots sont ajoutés dans le dictionnaire chaque fois qu'un nouveau mot est rencontré dans l'analyse du texte. Le deuxième passage se fait après la construction de la totalité du

dictionnaire pour déterminer sa taille. Avant de présenter l'algorithme commençons par un exemple commenté. Supposons que nous voulons coder le mot $\xi \in \mathbb{X}^+$, ou $\mathbb{X} = \{0,1\}$ et $\xi = 1011010100010$, de longueur $L = |\xi| = 13$. Le dictionnaire en position 0 contient toujours le mot vide, noté ε. En analysant α, nous rencontrons comme premier sous-mot, non encore contenu dans le dictionnaire la lettre 1, identifiée avec le mot $\varepsilon 1$. Nous codons ce mot en indiquant le couple $(\texttt{adr}, \texttt{suffixe})$, où `suffixe` est la lettre que nous devons ajouter comme suffixe au mot qui se trouve en position `adr` dans le dictionnaire ; en l'occurrence, nous codons $(0,1)$, signifiant « ajouter au mot en position 0 dans le dictionnaire le suffixe 1. Ainsi le dictionnaire contient maintenant le mot ε en position 0 et le mot 1 en position 1. Ensuite dans l'analyse de ξ nous rencontrons la lettre 0 qui est identifiée au mot $\varepsilon 0$. Le dictionnaire contiendra alors en position 2 le mot 0, codé $(0,0)$. Ceci signifie que le mot est obtenu en ajoutant comme suffixe au mot en position 0 la lettre 0, etc. En continuant ainsi l'analyse de ξ, nous remplissons les dictionnaire comme indiqué sur la table 8.2.

adr	mot	code
0	ε	
1	1	(0,1)
2	0	(0,1)
3	11	(1,1)
4	01	(2, 1)
5	010	(4, 0)
6	00	(2, 0
7	10	(1,0)

TABLE 8.2 – La construction du dictionnaire selon LZ78 ; elle correspond à l'analyse du mot $\xi = \cdot 0 \cdot 1 \cdot 11 \cdot 01 \cdot 010 \cdot 00 \cdot 10$; le symbole $\cdot$ sert à visualiser les positions de césure. Le dictionnaire contient 8 entrées —numérotées de 0 à 7 — nécessitant donc 3 bits pour être codées. Le code brut correspondant à ξ est $C(\xi) = (0,1)(0,0)(1,1)(2,1)(4,1)(2,0)(1,0)$. En codant la position de chaque préfixe dans le dictionnaire par un mot de 3 bits, on obtient pour le code final $C(\xi) = 000100000010101100101000010$ de longueur $|C(\xi)| = 28$ bits.

La structure de données du dictionnaire D sera donc une liste ordonnée. Si $L = (L_1, \ldots, L_N)$ est une liste ordonnée d'objets $L_i \in \mathbb{B}$, pour $i = 1, \ldots, N$ et $\mathbb{B}$ un certain ensemble, nous définissons l'opération « annexer en dernière position » ADP définie par $\mathsf{ADP}(L, \beta) := (L_1, \ldots, L_N, \beta)$ pour $\beta \in \mathbb{B}$. Nous pouvons maintenant donner la définition précise de l'algorithme.

Algorithme 8.4.6. LZ78

Require: Mot $\alpha \in \mathbb{A}^+$, fonctions ADP et adr .
Ensure: Dictionnaire D (i.e. liste ordonnée de mots de $\mathbb{A}^+$) et code $C(\alpha)$ (i.e. liste ordonnée de couples $(\mathtt{adr}, \mathtt{suf})$).

```
β ← ε
D ← (ε)
C ← ()
K ← |α|
k ← 1
drapeau_fin ← 0
while k ≤ K et drapeau_fin = 0 do
  a ← α_k
  if βa ∈ D then
    if k = K then
      C ← ADP(C, (adr(β|_{k−1}), a))
      drapeau_fin ← 1
    else
      β ← βa
    end if
  else
    D ← ADP(D, βa)
    C ← ADP(C, (adr(β), a))
    β ← ε
  end if
  k ← k + 1
end while
```

On peut remarquer que le code LZ78 en aucun moment ne fait appel explicitement au vecteur de probabilité de l'alphabet. Par ailleurs, il semble que pour le mot court de l'exemple ci-dessus, on a $|C(\xi)| > |\xi|$, i.e. un code qui est plus plus long que le mot sur lequel il s'applique. Il est cependant remarquable que l'on peut montrer le

Théorème 8.4.7. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoire indépendantes et identiquement distribuées à valeurs dans $\mathbb{X}$. Pour tout $n \in \mathbb{N}$, on note $\mathbf{X} := \mathbf{X}^{(n)} = (X_1 X_2 \cdots X_n) \in \mathbb{X}^n$. On a alors, en désignant par C le code LZ78,*

$$\lim_{n\to\infty} \frac{|C(\mathbf{X}^{(n)})|}{n} = H(X_1).$$

De manière encore plus remarquable, on peut étendre ce résultat à des suites de variables aléatoires stationnaires (cf. définition 5.5.1).

Exemple 8.4.8. La notion de processus stationnaire a été introduite dans la définition 5.5.1. Il s'ensuit qu'une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $(\mathbb{X}, \mathcal{X})$ est **stationnaire** si, pour tout $k \in \mathbb{N}$, les lois conjointes $\mathbb{P}(X_n \in B_0, \ldots X_{n+k} \in B_k)$, pour $B_0, \ldots, B_k \in \mathcal{X}$, ne dépendent pas de n. Les suites stationnaires généralisent les suites i.i.d.

1. Toute suite i.i.d. est stationnaire. Il existe cependant des suites stationnaires qui ne sont pas i.i.d. comme le montrent les cas ci-dessous.

2. Soient $(\xi_n)_{n\in\mathbb{N}}$ une suite variables aléatoires réelles i.i.d. intégrables et $(\alpha_n)_{n\in\mathbb{N}}$ une suite réelle sommable. La suite $(X_n)_{n\in\mathbb{N}}$ définie, pour tout $n \in \mathbb{N}$, par $X_n = \sum_{m=0}^{\infty} \alpha_m \xi_{n+m}$ est stationnaire.
3. Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov $(\mathbb{X}, P, \pi)$, où π est la probabilité invariante de la chaîne. Alors $(X_n)_{n\in\mathbb{N}}$ est stationnaire.

Définition 8.4.9. Soit $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ une suite aléatoire arbitraire.

1. Le **taux de croissance de l'entropie conjointe (ou entropie spécifique ou taux de production d'entropie)** de $\mathbf{X}$ est
$$h(\mathbf{X}) = \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n} \text{ (lorsque cette limite existe).}$$
2. Le **taux de production d'entropie conditionnelle** est
$$\hat{h}(\mathbf{X}) = \lim_{n\to\infty} H(X_n | X_{n-1} \cdots X_1) \text{ (lorsque cette limite existe).}$$

Exemple 8.4.10.
1. Entropie produite par un « singe dactylographe » : Soit $\mathbb{X}$ un alphabet arbitraire fini. Si on tape au hasard sur une machine à écrire à alphabet $\mathbb{X}$, on produit un texte $\mathbf{X} := X_1 X_2 X_3 \cdots$ avec (X_k) suite i.i.d. uniformément distribuée sur $\mathbb{X}$, alors $h(\mathbf{X}) = H(X_1) = \log |\mathbb{X}|$.
2. Si on assouplit la condition d'identique distribution, il n'est plus garanti que la limite existe.

Théorème 8.4.11. *Soient $\mathbf{X} = (X_k)_{k\in\mathbb{N}}$ une suite stationnaire sur un alphabet fini et $\hat{h}_n(\mathbf{X}) = H(X_n | X_{n-1} \cdots X_1)$. Alors la suite $(\hat{h}(\mathbf{X}))_n$ est décroissante et possède une limite $\hat{h}(\mathbf{X})$.*

Démonstration. Cf. exercice 96. □

Théorème 8.4.12. *Soit $\mathbf{X} = (X_k)_{k\in\mathbb{N}}$ une suite stationnaire sur un alphabet fini. Alors les limites $h(\mathbf{X})$ et $\hat{h}(\mathbf{X})$ existent et*
$$h(\mathbf{X}) = \hat{h}(\mathbf{X}).$$

Démonstration. Cf. exercice 97. □

On peut aussi montrer le

Théorème 8.4.13. *Soit $\mathbf{X} := (X_n)_{n\in\mathbb{N}}$ une chaîne de Markov $(\mathbb{X}, P, \pi)$, où π est la probabilité invariante de la chaîne. Alors, si C désigne le code LZ78,*
$$\lim_{n\to\infty} \frac{|C(X_1 \cdots X_n)|}{n} = \hat{h}(\mathbf{X}).$$

La démonstration (voir [7, pp. 184–189] par exemple) de ce théorème est en dehors des exigences de ce cours. On peut même démontrer un résultat analogue pour toute suite stationnaire ergodique — pas nécessairement une chaîne de Markov possédant une probabilité invariante — (voir [17, pp. 319–326]). Ce résultat s'inscrit dans le cadre plus général de la relation entre la complexité de Kolmogorov — notion que sera effleurée dans le cours de Complexité [58] — et l'entropie.

8.5 Exercices

Codes instantanés, codes uniquement décodables

87. Soit $C : \{\mathtt{a}, \mathtt{b}\} \to \{0,1\}^*$ le code défini par

x	$C(x)$
a	1
b	101

(a) Le code C est-il instantané ?
(b) Est-il uniquement décodable ?

88. Déterminer si les codes suivants — définis sur des ensembles à 7 ou 8 éléments selon le cas — sont instantanés ; sinon sont-ils uniquement décodables.

x_1	abc	0101	00	00
x_2	abcd	0001	112	11
x_3	e	0110	0110	0101
x_4	dba	1100	0112	111
x_5	bace	00011	100	1010
x_6	ceac	00110	201	100100
x_7	ceab	11110	212	0110
x_8	eabd	101011	22	

89. Soient $\mathbb{X}$ un ensemble de cardinal $M = 4$ et $\mathbf{p} = (1/2, 1/4, 1/8, 1/8)$ un vecteur de probabilité sur $\mathbb{X}$. On considère les quatre codes $C_k, k = 1, \ldots, 4$, définis ci-dessous.

X	$C_1(X)$	$C_2(X)$	$C_3(X)$	$C_4(X)$
a	0	00	0	0
b	10	01	1	01
c	110	10	00	011
d	111	11	11	111

(a) Calculer $H(\mathbf{p})$.
(b) Pour $k = 1, \ldots, 4$, calculer $\mathbb{E}|C_k(X)|$ et comparer avec $H(\mathbf{p})$.
(c) Pour $k = 1, \ldots, 4$, déterminer si C_k est un code instantané et sinon s'il est uniquement décodable.

Codes de Huffman

90. Soient $\mathbb{X}$ un ensemble de cardinal M et $\mathbf{p}$ un vecteur de probabilité sur $\mathbb{X}$. Calculer un code de Huffman et comparer l'espérance de sa longueur avec l'entropie dans les cas suivants :

(a) $M = 5$ et $\mathbf{p} = (0.25, 0.25, 0.2, 0.15, 0.15)$.
(b) $M = 6$ et $\mathbf{p} = (0.3, 0.25, 0.2, 0.1, 0.1, 0.05)$.

91. Soient $\mathbb{X} = \{0,1\}$, $\mathbf{p} = (0.9, 0.1)$ et $\mathbf{p}' = (0.6, 0.4)$ deux vecteurs de probabilité et K un entier strictement positif. On note $X^K = (X_1, \ldots, X_K)$, où $(X_i)_{i=1,\ldots,K}$ sont des copies indépendantes de la même variable aléatoire sur $\mathbb{X}$.

(a) Calculer $H(X^K)$ lorsque X_1 est distribuée selon $\mathbf{p}$ et selon $\mathbf{p}'$.
(b) Calculer un code de Huffman C pour $K = 2, \ldots, 4$ dans les deux cas $\mathbf{p}$ et $\mathbf{p}'$.

(c) Calculer $\mathbb{E}_{\mathbf{p}}|C(X^K)|$ et $\mathbb{E}_{\mathbf{p}'}|C(X^K)|$ pour $K = 2, \ldots, 4$.

92. Déterminer un vecteur de probabilité $\mathbf{p}$ sur $\mathbb{X} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$ qui donne lieu à deux codes de Huffman différents. Pour chacun de ces codes calculer l'espérance de sa longueur et comparer la avec $H(\mathbf{p})$.

Compléments sur les codes

93. Soient X une variable aléatoire à valeurs dans $\mathbb{X} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$ distribuée selon une loi décrite par le vecteur de probabilité $\mathbf{p} = (1/2, 1/4, 1/8, 1/8)$ et $C : \mathbb{X} \to \mathbb{A}^+$, avec $\mathbb{A} = \{0, 1\}$, un code de Huffman. On note avec le même symbole — C — l'extension du code sur $\mathbb{X}^+$ et on considère un mot $\xi \in \mathbb{X}^+$ de longueur $L = |\xi|$. On désigne par $\alpha = C(\xi) \in \mathbb{A}^+$ la suite codant ξ. Quelle est la probabilité qu'un bit au hasard de cette suite prenne la valeur 1 lorsque $L \to \infty$?
94. Construire le dictionnaire de chaînes-préfixes du codage du mot

$$\alpha = \texttt{the fat cat sat on the mat.}$$

par l'algorithme de Lempel-Ziv 78.
95. Coder et décoder selon LZ78 le mot

$$\beta = 00121212102101210122101$$

96. Soit $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ une suite stationnaire sur un ensemble fini.
    (a) Montrer que pour tout $n \geq 1$, on a : $H(X_{n+1}|X_n \cdots X_1) \leq H(X_{n+1}|X_n \cdots X_2)$.
    (b) Conclure que $\lim_{n\to\infty} H(X_n|X_{n-1} \cdots X_1)$ existe.
97. Soit $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ une suite stationnaire sur un ensemble fini.
    (a) Vérifier que
    $$\frac{1}{n}H(X_1 \cdots X_n) = \frac{1}{n}\sum_{k=1}^{n} H(X_k|X_{k-1} \cdots X_1).$$
    (b) Utiliser le résultat obtenu en exercice 96 pour conclure.

9
Canaux bruités sans mémoire

La notion de canal est très générale. Initialement elle a été introduite pour signifier la transmission d'un signal codant de l'information à travers un milieu et décrire les altérations que subit l'information à cause du bruit. Cependant, on parle aujourd'hui de canal pour modéliser une transformation *arbitraire* que subit un mot codé dans un alphabet fini. Les codes que nous avons vus au chapitre précédent sont des cas particuliers des canaux. Ce qui va nous intéresser dans ce chapitre n'est pas de coder efficacement la source mais de décrire précisément l'action du canal, c'est-à-dire les perturbations dues au bruit.

9.1 Modélisation markovienne

On suppose que le canal reçoit en entrée un mot aléatoire $\mathbf{X} \in \mathbb{X}^+$, où $\mathbb{X}$ est un alphabet fini et transmet en sortie un mot aléatoire $\mathbf{Y} \in \mathbb{Y}^+$. La probabilité conditionnelle $\mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{x} = \mathbf{x})$ est appelée probabilité de transmission du canal. Dans la suite, nous supposons que les entrées et les sorties ont la même longueur, i.e. $|\mathbf{x}| = |\mathbf{y}|$ et que les symboles qui arrivent en entrée sont émis selon une source indépendante. Cette hypothèse d'indépendance peut être relaxée au prix d'une complication dans les formules et dans la modélisation de la source.

Définition 9.1.1. Soit $(p_n)_{n\in\mathbb{N}}$ une suite d'applications, définies pour tout $n \in \mathbb{N}$, par

$$\mathbb{X}^n \times \mathbb{Y}^n \ni (\mathbf{x}, \mathbf{y}) \to p_n(\mathbf{x}, \mathbf{y}) := \mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{x} = \mathbf{x}) \in [0, 1].$$

1. La triplet $(\mathbb{X}, \mathbb{Y}, (p_n)_{n\in\mathbb{N}})$ est appelé **canal discret**.
2. Le canal est dit **sans mémoire**, s'il existe une matrice stochastique $P : \mathbb{X} \times \mathbb{Y} \to [0, 1]$ telle que pour tout $n \in \mathbb{N}$ et tous $\mathbf{x} \in \mathbb{X}^n$ et $\mathbf{y} \in \mathbb{Y}^n$, la probabilité conditionnelle s'écrit $p_n(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^n P(x, y)$. Un canal sans mémoire sera alors identifié avec le triplet $(\mathbb{X}, \mathbb{Y}, P)$, où P est une matrice stochastique $|\mathbb{X}| \times |\mathbb{Y}|$.

Dorénavant, nous ne considérons que de canaux sans mémoire.

Exemple 9.1.2. Considérer le canal avec $\mathbb{X} = \mathbb{Y} = \{0,1\}$ et $P = \begin{pmatrix} 1-e_0 & e_0 \\ e_1 & 1-e_1 \end{pmatrix}$, avec $e_0, e_1 \in [0,1]$. Il s'agit d'un canal discret sans mémoire. On calcule

$$\mathbb{P}(\mathbf{Y} = 110 | \mathbf{X} = 010) = P(0,1)P(1,1)P(0,0) = e_0(1-e_1)(1-e_0).$$

Si $e_0 = e_1 = 0$ ce canal est parfait car $P = I$. Dans un certain sens, qui sera précisé en §9.2 mais qui est intuitivement clair, le cas avec $e_0 = e_1 = 1$ correspond aussi à un canal parfait, tandis que le cas $e_0 = e_1 = 1/2$ à un canal inutile.

Remarque 9.1.3. La notion de canal est une notion très générale. Nous avons étudié en chapitre 8 le codage de la source. Si $\mathbb{X} = \{a,b,c,d\}$ et $\mathbb{A} = \{0,1\}$, nous avons vu que $C : \mathbb{X} \to \mathbb{A}^+$ avec $C(a) = 0$, $C(b) = 10$, $C(c) = 110$ et $C(d) = 111$ est un code instantané. Ce code peut aussi être vu comme un canal avec $\mathbb{Y} = C(\mathbb{X}) = \{0, 10, 110, 111\}$ et P une matrice stochastique déterministe, correspondant à un canal parfait. Plus généralement toute fonction $f : \mathbb{X} \to \mathbb{Y}$ *est équivalente* à une matrice stochastique déterministe $|\mathbb{X}| \times |\mathbb{Y}|$, notée K_f, dont les éléments de matrice sont définis par

$$\mathbb{X} \times \mathbb{Y} \ni (x,y) \mapsto K_f(x,y) = \delta_{f(x)}(y).$$

C'est-à-dire une fonction $f : \mathbb{X} \to \mathbb{Y}$ est le canal $(\mathbb{X}, \mathbb{Y}, K_f)$.

Nous avons établi en chapitre 7 les propriétés de l'entropie et les relations entre entropie conjointe, entropie conditionnelle et information mutuelle. Rappelons aussi que si la loi (le vecteur de probabilité) de la source $\mu \in \mathsf{PV}_{\mathbb{X}}$ et la matrice stochastique P du canal sont données, nous pouvons calculer

- l'entropie de la source $H(X)$,
- la loi conjointe $\kappa(x,y) = \mathbb{P}(X = x, Y = y) = \mu(x)P(x,y)$ (et par conséquent l'entropie conjointe $H(X,Y)$),
- la loi de sortie $\nu(y) = \sum_{x \in \mathbb{X}} \kappa(x,y) = \sum_{x \in \mathbb{X}} \mu(x)P(x,y)$ (et par conséquent l'entropie de la sortie $H(Y)$),
- les entropies conditionnelles $H(X|Y)$ et $H(Y|X)$ et l'information mutuelle $I(X:Y)$.

La figure 9.1 rappelle schématiquement les relations qui existent entre ces différentes quantités. Il va de soi que toutes ces quantités sont fonctions du couple (μ, P). Ceci nous amène, lorsque la loi de l'entrée est fixée, à une notation plus précise du canal.

Notation 9.1.4. Un canal discret sans mémoire est le quadruplet $(\mathbb{X}, \mu, \mathbb{Y}, P)$, où $\mathbb{X}, \mathbb{Y}$ sont les alphabets d'entrée er de sortie, μ la loi de l'entrée et P la matrice de transmission. On garde cependant la notation sous forme de triplet $(\mathbb{X}, \mathbb{Y}, P)$ lorsque nous ne voulons pas préciser d'emblée la loi d'entrée (par exemple lorsque nous voulons étudier le comportement du canal pour une famille de lois d'entrée (cf. définition de capacité 9.3.1 plus loin).

9.2 Classification des canaux

9.2.1 Canaux sans perte

Les **canaux sans perte** sont les canaux caractérisés par la propriété $H(X|Y) = 0$; cette propriété signifie que si nous connaissons la sortie, il ne reste plus aucune incertitude quant à la valeur en entrée, autrement dit, l'entrée est totalement déterminée par

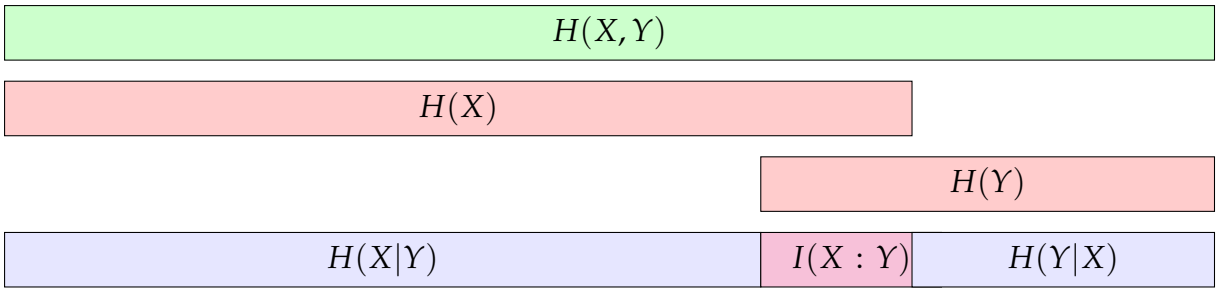


FIGURE 9.1 – Représentation schématique des valeurs de différentes entropies. Toutes ces quantités sont fonctions de la loi de la source μ et de la matrice stochastique P du canal.

la sortie. De manière équivalente, $I(X:Y) = H(X) - H(X|Y) = H(X)$, c'est-à-dire que l'information mutuelle entre l'entrée et la sortie coïncide avec toute l'information de l'entrée.

Il est instructif de visualiser un tel canal. Supposons que $\mathrm{card}\mathbb{X} = M$ (par conséquent $\mathbb{X} \simeq \{x_1, \dots, x_M\}$) et que $\mathbb{Y}$ soit partitionné en M parties $(B_i)_{i=1,\dots,N}$ non-vides (i.e. $\mathbb{Y} = \sqcup_{i=1}^{M} B_i$). Le canal est décrit par une matrice stochastique $(P(x,y))_{x\in\mathbb{X}, y\in\mathbb{Y}}$ telle que, pour tout $i \in \{1,\dots,M\}$ on ait

$$\mathbb{P}(Y \in B_i | X = x_i) = \sum_{y\in B_i} P(x_i, y) = P(x_i, B_i) = 1.$$

Mais alors, en notant μ la loi de la source,

$$\begin{aligned}
\mathbb{P}(X = x_i; Y \in B_i) &= \mu(x_i);\\
\mathbb{P}(X = x_i | Y \in B_i) &= \frac{\mu(x_i)}{\mathbb{P}(Y \in B_i)} = \frac{\mu(x_i)}{\sum_j \mathbb{P}(Y \in B_i | X_j = x_j)\mathbb{P}(X = x_j)}\\
&= \frac{\mu(x_i)}{\sum_j \delta_{i,j}\mu(x_j)} = \frac{\mu(x_i)}{\mu(x_i)} = 1.
\end{aligned}$$

9.2.2 Canaux déterministes

Les **canaux déterministes** sont les canaux dont la matrice stochastique P est déterministe, i.e. dans chaque ligne il existe un unique élément égal à 1 ($\forall x \in \mathbb{X}, \exists! y := y_x \in \mathbb{Y}, P(x, y_x) = 1$) tandis que tous les autres éléments de matrice sont nuls. Si μ est la loi d'entrée, on calcule la loi conjointe d'entrée-sortie

$$\mathbb{X} \times \mathbb{Y} \ni (x,y) \mapsto \kappa(x,y) = \mu(x)P(x,y) = \mu(x)\mathbb{1}_{\{y_x\}}(y).$$

Un calcul élémentaire donne $H(X,Y) = H(\kappa) = H(X)$ et par conséquent $H(Y|X) = H(X,Y) - H(X) = 0$ et $I(X:Y) = H(Y) - H(Y|X) = H(Y)$.

Exemple 9.2.1. Supposons que nous disposions d'un jeu de 52 cartes. Soit $X \in \mathbb{X} = \{1,\dots,10,\mathrm{J},\mathrm{Q},\mathrm{K}\} \times \{\heartsuit, \diamondsuit, \spadesuit, \clubsuit\}$ le vecteur aléatoire du couple (valeur, enseigne) d'une carte tirée au hasard et $Y \in \mathbb{Y} = \{\heartsuit, \diamondsuit, \spadesuit, \clubsuit\}$ la variable aléatoire « enseigne » de la carte qui est transmise. Si la loi de X est la loi uniforme sur $\mathbb{X}$, alors $I(X:Y) = H(Y) = \log 4 = 2$.

9.2.3 Canaux sans bruit

Un **canal sans bruit** est un canal qui est à la fois sans perte ($H(X|Y) = 0$, i.e. l'entrée est déterminée par la sortie) et déterministe ($H(Y|X) = 0$, i.e. la sortie est déterminée par l'entrée). Par conséquent, $I(X : Y) = H(X) = H(Y)$.

9.2.4 Canaux inutiles

Un **canal inutile** est un canal qui, pour toute loi d'entrée $\mu \in \mathcal{M}_1(\mathbb{X})$, vérifie $I(X : Y) = 0$. On a alors

$$0 = I(X : Y) = H(X) - H(X|Y) = 0 \Rightarrow H(X) = H(X|Y),$$

c'est-à-dire que les variables X et Y d'entrée-sortie sont indépendantes ou, dit autrement, l'observation de la sortie ne nous apprend rien sur l'entrée. Une autre manifestation de l'indépendance de variables X et Y est que toutes les lignes de la matrice stochastique P sont égales entre elles (et bien-sûr correspondent à un vecteur de probabilité sur $\mathbb{Y}$).

9.2.5 Canaux symétriques

Définition 9.2.2. Soient $\mathbb{S}_n$ le groupe des permutations sur n objets et $(\mathbb{X}, \mathbb{Y}, P)$ un canal sans mémoire. Supposons qu'il existe un vecteur de probabilité $\mathbf{p} \in \mathcal{M}_1(\mathbb{Y})$ et un vecteur dans $\mathbf{z} \in [0,1]^{|\mathbb{X}|}$, tels que

1. $\forall x \in \mathbb{X}, \exists \sigma_x \in \mathbb{S}_{|\mathbb{Y}|} : \forall y \in \mathbb{Y}, P(x,y) = p(\sigma_x y)$ et
2. $\forall y \in \mathbb{X}, \exists \sigma_y \in \mathbb{S}_{|\mathbb{X}|} : \forall x \in \mathbb{X}, P(x,y) = z(\sigma_y x)$.

Alors le canal est dit **canal symétrique**.

Exemple 9.2.3. Les canaux (entre les espaces d'entrée et sortie idoines) dont les matrices de transmission sont

$$P_1 = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \text{ et } P_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \end{pmatrix}$$

sont des canaux symétriques. Effectivement, dans les deux cas, chaque ligne est la permutation d'un même vecteur de probabilité et chaque colonne la permutation d'un même vecteur de $[0,1]^{|\mathbb{X}|}$.

Calculons l'entropie conditionnelle :

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathbb{X}} \mu(x) H(Y|X = x) \\ &= -\sum_{x \in \mathbb{X}} \mu(x) \sum_{y \in \mathbb{Y}} \mathbb{P}(Y = y|X = x) \log \mathbb{P}(Y = y|X = x) \\ &= -\sum_{x \in \mathbb{X}} \mu(x) \sum_{y \in \mathbb{Y}} P(x,y) \log P(x,y) \\ &= -\sum_{x \in \mathbb{X}} \mu(x) \sum_{y \in \mathbb{Y}} p(\sigma_x y) \log p(\sigma_x y) \\ &= H(\mathbf{p}). \end{aligned}$$

On constate donc que l'entropie conditionnelle dépend uniquement des caractéristiques du canal (à travers le vecteur de probabilité $\mathbf{p}$) et elle est indépendante de la loi d'entrée μ. L'information mutuelle sera donnée par $I(X:Y) = H(Y) - H(Y|X) = H(Y) - H(\mathbf{p})$ et il est facile de voir que

$$\mathsf{cap}(P) := \sup_{\mu \in \mathcal{M}_1(\mathbb{X})} I(X:Y) = \log \mathsf{card}\mathbb{X} - H(\mathbf{p}).$$

En particulier pour un canal symétrique binaire, $\mathsf{cap}(P) = \sup_{\mu \in \mathcal{M}_1(\mathbb{X})} I(X:Y) = 1 - H(\mathbf{p})$ s'annule pour $\mathbf{p} = (1/2, 1/2)$ signifiant que — dans ce cas — le canal binaire est alors inutile.

9.3 Capacité du canal, propriétés de la capacité

Définition 9.3.1. Pour un canal fixé (i.e. une matrice de transmission fixée P), on appelle **capacité** la quantité

$$\mathsf{cap} := \mathsf{cap}(P) = \sup_{\mu \in \mathcal{M}_1(\mathbb{X})} I(X:Y).$$

Remarque 9.3.2. La signification de la définition 9.3.1 n'est pas encore très claire. Elle le deviendra dans le §9.5. Signalons pour l'instant les deux résultats qui seront montrés dans la suite :

1. il est possible de transmettre de l'information avec un taux d'erreur arbitrairement petit à tout taux de transmission R (cf. définition 9.5.2) $R < \mathsf{cap}$,
2. dès que le taux de transmission R dépasse la capacité cap, la transmission n'est plus fiable.

Ceci signifie que chaque canal se comporte comme un « tuyau » à travers lequel on peut faire passer un débit maximal de fluide.

Proposition 9.3.3. *Soient un canal sans bruit* $(\mathbb{X}, \mathbb{Y}, P)$ *et* cap *sa capacité.*

1. $\mathsf{cap} \geq 0$.
2. $\mathsf{cap} \leq \log \mathsf{card}\mathbb{X}$.
3. $\mathsf{cap} \leq \log \mathsf{card}\mathbb{Y}$.

Démonstration.

1. Comme $I(X:Y) \geq 0$ (cf. remarque 7.5.8), la positivité de cap en découle immédiatement.
2. On a
$$\mathsf{cap} = \sup_{\mu \in \mathcal{M}_1(\mathbb{X})} I(X:Y) = \sup_{\mu \in \mathcal{M}_1(\mathbb{X})} (H(X) - H(X|Y)) \leq \sup_{\mu \in \mathcal{M}_1(\mathbb{X})} H(X) \leq \log \mathsf{card}\mathbb{X}.$$
3. On a
$$\mathsf{cap} = \sup_{\mu \in \mathcal{M}_1(\mathbb{X})} I(X:Y) = \sup_{\mu \in \mathcal{M}_1(\mathbb{X})} (H(Y) - H(Y|X)) \leq \sup_{\mu \in \mathcal{M}_1(\mathbb{X})} H(Y) \leq \log \mathsf{card}\mathbb{Y}.$$

□

Le calcul de la capacité du canal est un problème difficile. Une formule fermée pour cap peut être trouvée uniquement dans quelques cas simples, comme les canaux symétriques, mais nous ne connaissons de formule dans le cas général.

9.4 Un exemple illustratif simple

Exemple 9.4.1. Soit un canal $(\mathbb{X}, \mathbb{Y}, P)$ avec $\mathbb{X} = \{x_1, x_2, x_3\}$, $\mathbb{Y} = \{y_1, y_2, y_3\}$ et

$$P_{xy} := \mathbb{P}(Y = y | X = x) = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.3 & 0.5 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}.$$

On doit fixer une règle de décision raisonnable $\Delta : \mathbb{Y} \to \mathbb{X}$ permettant de deviner le symbole émis $x \in \mathbb{X}$ lorsque $y \in \mathbb{Y}$ est reçu. Cette décision dépend évidemment de P mais aussi de la probabilité d'émission μ des symboles de la source. Supposons que cette dernière est la probabilité uniforme. Il paraît plausible qu'il faille définir $y_1 \mapsto x_1$ et $y_3 \mapsto x_2$ mais quelle est la valeur plausible pour y_2 ? Toutes le valeurs de x semblent plausibles. Si le canal était plus compliqué aurait-on une manière systématique pour définir la règle de décision Δ ?

Définition 9.4.2. La règle de décision

$$\Delta(y) \in \arg\max_{z \in \mathbb{X}} \mathbb{P}(X = z | Y = y), y \in \mathbb{Y}$$

s'appelle **règle du maximum de vraisemblance**.

Remarquons qu'en général, $\mathbb{P}(X = z | Y = y) \neq P_{zy}$! Pour calculer cette probabilité conditionnelle, nous devons utiliser la formule de Bayes :

$$\hat{P}^t_{zy} = \hat{P}_{yz} = \mathbb{P}(X = z | Y = y) = \frac{\mathbb{P}(X = z, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mu(z) P_{zy}}{\sum_{w \in \mathbb{X}} \mu(w) P_{wy}}.$$

Si μ est la loi uniforme sur $\mathbb{X}$, cette formule se réduit à

$$\hat{P}^t_{zy} = \hat{P}_{yz} = \mathbb{P}(X = z | Y = y) = \frac{P_{zy}}{\sum_{w \in \mathbb{X}} P_{wy}}.$$

Dans l'exemple ci-dessus (avec probabilité uniforme pour la source), elle devient donc

$$\hat{P}^t = \begin{pmatrix} 0.5 & 0.33333\ldots & 0.181818\ldots \\ 0.2 & 0.33333\ldots & 0.454545\ldots \\ 0.3 & 0.33333\ldots & 0.363636\ldots \end{pmatrix}.$$

On constate que la solution au problème de maximisation de vraisemblance est

y	$\arg\max_{z \in \mathbb{X}} \hat{P}_{zy}$
y_1	$\{x_1\}$
y_2	$\{x_1, x_2, x_3\}$
y_3	$\{x_2\}$

Tout élément x^* dans l'ensemble $\arg\max_{z \in \mathbb{X}} \mathbb{P}(X = z | Y = y)$, pour $y \in \mathbb{Y}$, peut servir à définir $\Delta(y) := \Delta_{\mathrm{MV}}(y) = x^*$.

D'autres règles de décision sont possibles. Par exemple si $\mathbb{X} = \{0, 1\} \simeq \{000, 111\}$ et $\mathbb{Y} = \{0, 1\}^3$, on peut définir, pour tout $\mathbf{y} = (y_1, y_2, y_3) \in \mathbb{Y}$, la **règle de la majorité**, donnée par la formule

$$\Delta(\mathbf{y}) := \Delta_{\mathrm{Maj}}(\mathbf{y}) = \begin{cases} 000 & \text{si } y_1 + y_2 + y_3 \leq 1, \\ 111 & \text{si } y_1 + y_2 + y_3 > 1, \end{cases}$$

où $\mathbf{y} = y_1y_2y_3$. On peut même définir des **règles de décision stochastiques**, i.e. décrites par des noyaux stochastiques non-déterministes $K_\Delta : \mathbb{Y} \times \mathbb{X} \to [0,1]$, signifiant qu'à chaque symbole $y \in \mathbb{Y}$ reçu, on assigne un symbole $x \in \mathbb{X}$ avec probabilité $K_\Delta(y,x)$. Dans l'exemple 9.4.1, on peut définir une règle de décision stochastique décrite par le noyau stochastique

$$K_\Delta := \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1 & 0 \end{pmatrix}.$$

Mais ceci n'est pas le seul choix. *Tout noyau stochastique* $K_\Delta : \mathbb{Y} \times \mathbb{X} \to [0,1]$ correspondrait à une règle de décision pas nécessairement plausible mais acceptable du point de vue de la théorie. Noter aussi que les règles de décision du maximum de vraisemblance et de la majorité (ainsi que toute autre règle de décision décrite par *une fonction* $\Delta : \mathbb{Y} \to \mathbb{X}$) sont aussi décrites par des noyaux stochastiques déterministes $K_\Delta : \mathbb{Y} \times \mathbb{X} \to \{0,1\}$, définis par $K_\Delta(y,x) = \mathbb{1}_{\{\Delta(y)\}}(x)$.

On arrive ainsi à la définition suivante :

Définition 9.4.3. Une **règle de décision** est un noyau stochastique

$$K_\Delta : \mathbb{Y} \times \mathbb{X} \to [0,1]$$

qui assigne à tout symbole $y \in \mathbb{Y}$ reçu par le canal, un symbole émis $x \in \mathbb{X}$ avec probabilité $K_\Delta(y,x)$.

9.5 Le théorème fondamental de la transmission

9.5.1 Codage du canal bruité

Nous disposons d'un canal sans mémoire $(\mathbb{X}, \mathbb{Y}, P)$ à travers lequel allons transmettre des mots de longueur fixe n. Nous avons deux méthodes équivalents pour décrire le processus de transmission :

— soit avec le canal $(\mathbb{X}, \mathbb{Y}, P)$ et considérer la transmission des *vecteurs aléatoires* $\mathbf{X} = X_1 \cdots X_n \in \mathbb{X}^n$ vers des *vecteurs aléatoires* $\mathbf{Y} = Y_1 \cdots Y_n \in \mathbb{Y}^n$ selon la matrice de transmission P, c'est-à-dire (cf. exemple 9.1.2)

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} P(x_i, y_i) =: Q_n(\mathbf{x}, \mathbf{y}),$$

— soit étendre le canal en $(\mathbb{X}^n, \mathbb{Y}^n, Q_n)$, où Q_n est la matrice de transmission entre $\mathbb{X}^n$ et $\mathbb{Y}^n$ définie dans la ligne précédente, et considérer la transmission des *variables aléatoires* $\mathbf{X} \in \mathbb{X}^n$ vers des *variables aléatoires* $\mathbf{Y} \in \mathbb{Y}^n$, selon la matrice de transmission Q_n, c'est-à-dire

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) =: Q_n(\mathbf{x}, \mathbf{y}),$$

Le choix de l'une ou l'autre approche est une pure question de commodité d'écriture.

Cependant, les ensembles (de mots) $\mathbb{X}^n$ et $\mathbb{Y}^n$ qui apparaissent dans la description précédente, ne sont pas les objets qui nous intéressent réellement ; nous sommes plutôt intéressés à un ensemble fini de messages $\mathbb{M}$ qui sont codés en des mots de $\mathbb{X}^n$ et aux messages de $\mathbb{M}$ qui sont obtenus en décodant les mots de $\mathbb{Y}^n$. Plus précisément nous disposons des ensembles finis $\mathbb{M}$, $\mathbb{X}^n$ et $\mathbb{Y}^n$ et de deux applications

— une de codage $\mathbb{M} \ni m \mapsto \mathbf{C}(m) \in \mathbb{X}^n$ et
— une de décodage [1] — supposée déterministe dans ce paragraphe (cf. exercice 102 pour un exemple de décodage stochastique) — $\mathbb{Y}^n \ni \mathbf{y} \mapsto \Delta(\mathbf{y}) \in \mathbb{M}$.

La figure 9.2 résume schématiquement ces actions prises séparément tandis que la figure 9.3 schématise la séquence de ces actions.

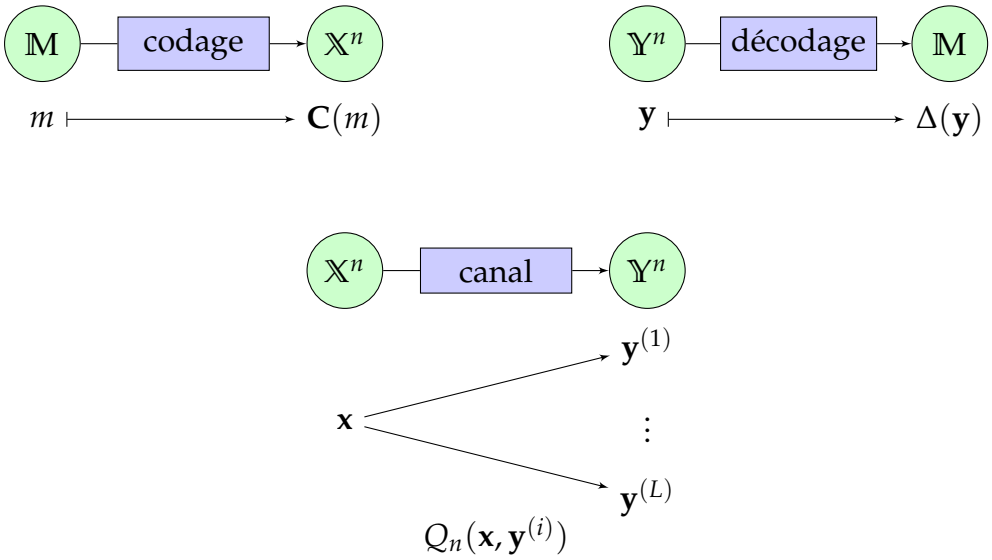


FIGURE 9.2 – Résumé des actions de codage, transmission et décodage, considérées séparément.

Notez que l'action du canal transforme toute entrée $\mathbf{x} \in \mathbb{X}^n$ en une variable aléatoire $\mathbf{Y} \in \mathbb{X}^n$ de loi conditionnelle (à l'entrée) $Q_n(\mathbf{x}, \cdot)$, i.e. pour tout $A \subset \mathbb{Y}^n$, $\mathbb{P}(\mathbf{Y} \in A | \mathbf{X} = \mathbf{x}) = \sum_{\mathbf{y} \in A} Q_n(\mathbf{x}, \mathbf{y})$.

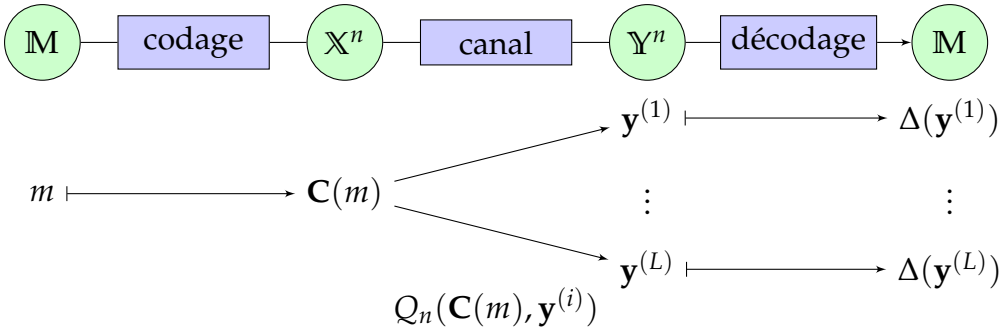


FIGURE 9.3 – Résumé des actions de codage, transmission et décodage, considérées séquentiellement.

L'intercalation du canal entraîne la probabilisation naturelle de l'ensemble $\mathbb{Y}^n$ et, de ce fait, la probabilisation naturelle de l'ensemble $\mathbb{M}$ de sortie. Par conséquent, la présentation devient plus symétrique si nous probabilisons d'emblée l'ensemble $\mathbb{M}$ (d'entrée) avec une probabilité *a priori* μ, la probabilisation de l'espace $\mathbb{M}$ (de sortie) se fera alors par une probabilité *a posteriori* $\nu := \nu^\mu$ que nous déterminons ci-dessous. L'opération de codage $\mathbf{C}$ est une fonction ; elle sera donc équivalente à une matrice stochastique déterministe $|\mathbb{M}| \times |\mathbb{X}|^n$, notée $K_{\mathbf{C}}$, dont les éléments de matrice sont définis par

$$\mathbb{M} \times \mathbb{Y}^n \ni (m, \mathbf{x}) \mapsto K_{\mathbf{C}}(m, \mathbf{x}) = \delta_{\mathbf{C}(m), \mathbf{x}}.$$

1. On pourrait raisonnablement la noter D mais le symbole D est déjà utilisé pour le contraste de Kullback-Leibler.

I.e. le codage est le canal déterministe $(\mathbb{M}, \mathbb{X}^n, K_{\mathbf{C}})$ (cf. remarque 9.1.3). L'opération de décodage Δ est aussi une fonction ; elle est donc équivalente au canal déterministe $(\mathbb{Y}^n, \mathbb{M}, K_\Delta)$, où $K_\Delta(\mathbf{y}, m) = \delta_{\Delta(\mathbf{y}),m}$. La séquence codage-transmission-décodage équivaut donc à la multiplication de 3 matrices stochastiques $K_{\mathbf{C}} Q_n K_\Delta$ (vérifier que les dimensions des matrices permettent leur multiplication dans cet ordre). Ce canal composé transformera tout vecteur de probabilité (ligne) a priori $\mu \in \mathcal{M}_1(\mathbb{M})$ en un vecteur de probabilité (ligne) a posteriori $\nu^\mu = \mu K_{\mathbf{C}} Q_n K_\Delta \in \mathcal{M}_1(\mathbb{M})$. En particulier, il transformera l'entrée déterministe $M = m$ (correspondant au vecteur de probabilité $\mu = \delta_m$) en la variable aléatoire $M' \in \mathbb{M}$ de loi ν, c'est-à-dire,

$$\begin{aligned}\nu^{\delta_m}(m') &= \mathbb{P}_m(M' = m') = \mathbb{P}(M' = m' | M = m) \\ &= \sum_{v \in \mathbb{M}} \mathbb{P}(M' = m' | M = v)\delta_m(v) = K_{\mathbf{C}} Q_n K_\Delta(m, m') \\ &= \sum_{\mathbf{x} \in \mathbb{X}^n} \sum_{\mathbf{y} \in \mathbb{Y}^n} K_{\mathbf{C}}(m, \mathbf{x}) Q_n(\mathbf{x}, \mathbf{y}) K_\Delta(\mathbf{y}, m') = \sum_{\mathbf{y} \in \mathbb{Y}^n} Q_n(\mathbf{C}(m), \mathbf{y})\delta_{\Delta(\mathbf{y}),m'}.\end{aligned}$$

Plus généralement, lorsque le message d'entrée est lui même considéré comme résultant d'un choix aléatoire selon une loi μ non-déterministe, nous obtenons

$$\begin{aligned}\nu^\mu(m') = \mathbb{P}_\mu(M' = m') &= \sum_{m \in \mathbb{M}} \mathbb{P}(M' = m' | M = m)\mu(m) \\ &= \sum_{m \in \mathbb{M}} \sum_{\mathbf{y} \in \mathbb{Y}^n} \mu(m) Q_n(\mathbf{C}(m), \mathbf{y})\delta_{\Delta(\mathbf{y}),m'}.\end{aligned}$$

9.5.2 Probabilité d'erreur de transmission

La forme explicite de la loi ν^μ sur $\mathbb{M}$ (obtenue dans le paragraphe précédent pour une canal $(\mathbb{X}^n, \mathbb{Y}^n, Q_n)$) qui décrit le résultat de décodage nous permet d'estimer la probabilité d'erreur de transmission. On parle d'erreur de transmission lorsque le résultat m' que nous obtenons à la fin du traitement diffère du message initial m. La probabilité de réalisation d'une telle erreur est, pour tout $m \in \mathbb{M}$,

$$\begin{aligned}e(m) := e^{(n)}(m) = \mathbb{P}_{\delta_m}(M' \neq m) &= \sum_{m' \neq m} \sum_{\mathbf{y} \in \mathbb{Y}^n} Q_n(\mathbf{C}(m), \mathbf{y})\delta_{\Delta(\mathbf{y}),m'} \\ &= \sum_{\mathbf{y} \in \mathbb{Y}^n} Q_n(\mathbf{C}(m), \mathbf{y}) \mathbb{1}_{\mathbb{M} \setminus \{m\}}(\Delta(\mathbf{y})).\end{aligned}$$

La qualité de transmission est quantifiée soit par l'**erreur maximale**

$$e_{\max} := e^{(n)}_{\max} = \max_{m \in \mathbb{M}} e^{(n)}(m),$$

soit par l'**erreur moyenne**

$$\bar{e} := \bar{e}^{(n)} = \sum_{m \in \mathbb{M}} \mu(m) e^{(n)}(m).$$

Comme $e^{(n)}(m) \leq e^{(n)}_{\max}$ pour tout $m \in \mathbb{M}$, il est évident que $\bar{e}^{(n)} \leq e^{(n)}_{\max}$. Cependant, un des résultats importants du théorème fondamental de la transmission sera que lorsque $n \to \infty$, les deux erreurs sont du même ordre.

Définition 9.5.1. Un $[n,k]$**-code (par blocs)** (avec k et n entiers supérieurs à 1) pour un canal discret sans mémoire $(\mathbb{X},\mathbb{Y},P)$ est la donnée
— d'un ensemble de messages $\mathbb{M}$ avec $\mathsf{card}\mathbb{M}=k$,
— d'un codage $\mathbf{C}:\mathbb{M}\to\mathbb{X}^n$ de taille fixe n,
— d'un décodage $\Delta:\mathbb{Y}^n\to\mathbb{M}$.
On note ce code $\mathcal{K}$, ou plus précisément $\mathcal{K}(n,k)$ (ou simplment $[n,k]$) si on veut préciser ses paramètres. L'ensemble $\mathbf{C}(\mathbb{M})\subseteq\mathbb{X}^n$ est appelé **glossaire du code** $\mathcal{K}$.

Le codage $\mathbf{C}$ est toujours supposé non-singulier, i.e. injectif. En le considérant comme une application $\mathbf{C}:\mathbb{M}\to\mathbf{C}(\mathbb{M})$, il devient aussi surjectif, donc bijectif. Étant donné que $\mathbf{C}$ est bijectif entre $\mathbb{M}$ et son glossaire, on peut identifier $\mathbb{M}$ avec le glossaire et considérer $\mathbb{M}$ comme isomorphe à une partie spécifique de $\mathbb{X}^n$.

Il est évident qu'une fois le code $\mathcal{K}$ choisi, les erreurs $e^{(n)}_{\max}$ et $\overline{e}^{(n)}$ définies plus haut, dépendent de $\mathcal{K}$; on précise cette dépendance en écrivant $e_{\max}[\mathcal{K}(k,n)]$ ou $\overline{e}[\mathcal{K}(k,n)]$.

Définition 9.5.2. Soit $\mathcal{K}$ un $[n,k]$-code par blocs.

1. Son **taux de transmission** R est défini par
$$R:=R[\mathcal{K}]=\frac{\log_{|\mathbb{X}|}k}{n}.$$
2. Un taux de transmission R est **atteignable** s'il existe une suite $(\mathcal{K}_\ell)_{\ell\in\mathbb{N}}$ de $[n_\ell,k_\ell]$-codes par blocs, tels que
$$\lim_{l\to\infty}\frac{\log_{|\mathbb{X}|}k_\ell}{n_\ell}\to R \text{ et } \lim_{l\to\infty}e_{\max}[\mathcal{K}_\ell]=0.$$

Exemple 9.5.3. Soit $\mathbb{X}=\{0,1\}$ l'alphabet binaire.

1. On considère le $(3,2)$-code par blocs de répétition de taille 3 avec $\mathbb{M}=\{0,1\}$, défini par $\mathbf{C}(b)=bbb$, pour $b\in\mathbb{M}$, i.e. de glossaire $\mathbf{C}(\mathbb{M}):=\{000,111\}$. Il a un taux de transmission $R=\frac{\log_2 2}{3}=\frac{1}{3}$.
2. On considère le $(3,4)$-code par blocs avec bit de parité de taille 3 avec $\mathbb{M}=\{00,01,10,11\}$, découlant, pour tout message $ab\in\mathbb{M}$, du codage $\mathbf{C}(ab)=abc\in\mathbb{X}^3$ avec $c=a+b \mod 2$, i.e. ayant comme glossaire $\mathbf{C}(\mathbb{M})=\{000,011,101,110\}$. Il a un taux de transmission $R=\frac{\log_2 4}{3}=\frac{2}{3}$.

9.5.3 Le théorème fondamental de la transmission

Le théorème fondamental de la transmission établit que la capacité est le supremum des taux de transmission atteignables.

Théorème 9.5.4. *Soit $(\mathbb{X},\mathbb{Y},P)$ un canal sans mémoire de capacité $\mathsf{cap}:=\mathsf{cap}(P)$.*

1. *Pour tout $R<\mathsf{cap}$, il existe une suite infinie $(\mathcal{K}_\ell)_{\ell\in\mathbb{N}}$ de $[n_\ell,k_\ell]$-codes, ayant des taux de transmission $R_\ell:=\frac{\log_{|\mathbb{X}|}k_\ell}{n_\ell}$ vérifiant $R_\ell\to R$, tels que $\lim_{\ell\to\infty}\overline{e}[\mathcal{K}_\ell]=0$.*
2. *Réciproquement, pour tout $R>\mathsf{cap}$ et toute suite $(\mathcal{K}_\ell)_{\ell\in\mathbb{N}}$ de $[n_\ell,k_\ell]$-codes avec des blocs de taille croissante (i.e. $n_1<n_2<n_3<\ldots$) et des taux de transmission vérifiant $R_\ell\geq R$, on a $\lim_{\ell\to\infty}\overline{e}[\mathcal{K}_\ell]=1$.*

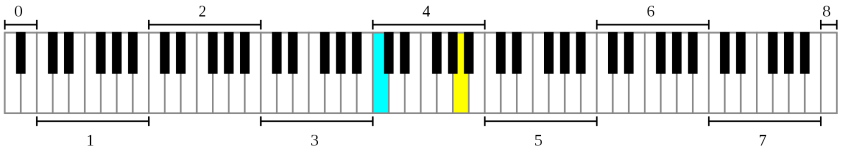


FIGURE 9.4 – Le touches blanches correspondent aux notes de base : do, re, mi, fa, sol, la, si. La touche représentée en cyan sur la figure correspond au « do » de base, la touche représentée en jaune sur la figure au « la » et la gamme se répète périodiquement sur le clavier. *Source de la figure :* Wikipédia ; distribuée sous licence CC BY-SA 3.0.

La démonstration de ce théorème est longue. Elle est basée sur la notion de suites conjointement typiques et est esquissée aux exercices 105 et 106. Elle fait usage du fait que si $I(X : Y) > 0$, en prenant des blocs suffisamment longs, on peut distiller un code et en raréfiant l'ensemble de messages à coder (i.e. en ne considérant qu'un sous-ensemble suffisamment épars de messages) on peut transmettre avec une erreur arbitrairement petite. L'idée est illustrée dans l'exemple 9.5.5.

Exemple 9.5.5. (*Das ‚falschtemperierte' Klavier* ou le problème du « pianiste non-doué »). Les touches des notes do-re-mi-fa-sol-la-si se répètent périodiquement sur le clavier du piano (cf. figure 9.4). On peut donc les supposer enroulés sur un cercle de façon que le si se trouve à proximité immédiate du do. (C'est le cas en réalité, le si d'une octave se trouve effectivement juste avant le do de l'octave supérieure). Un pianiste mal préparé, lorsqu'il lit une note dans la partition, frappe tantôt cette note tantôt la suivante avec probabilité $1/2$. Ainsi, lorsqu'il lit mi, il frappe mi avec probabilité $1/2$ et fa avec probabilité $1/2$. Le système est donc décrit comme un canal bruité (c'est le pianiste qui cause le bruit ...) avec $\mathbb{M} = \mathbb{X} = \mathbb{Y} = \{\text{do, re, mi, fa, sol, la, si}\}$ et $C = \Delta = \mathrm{id}$. La matrice de transmission du canal est, à cause de la périodisation,

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 1/2 \end{pmatrix}.$$

On calcule immédiatement que $H(Y|X) = H(\mathbf{p})$, où $\mathbf{p} = (1/2, 1/2, 0, 0, 0, 0, 0)$, c'est-à-dire $H(Y|X) = \log 2 = 1$. On a donc $I(X : Y) = H(Y) - H(Y|X) = H(Y) - 1$. Par ailleurs, le vecteur de probabilité uniforme $\mu(x) = 1/7$, pour tout $x \in \mathbb{X}$ est invariant, i.e. $\mu P = \mu$ ce qui signifie que la loi uniforme est un choix possible comme loi de Y. Mais la loi uniforme sature la borne de l'entropie : $H(Y) = \log 7$. On a donc calculé la capacité de ce pianiste : $\mathsf{cap} = \sup_\mu (H(Y) - 1) = \log 7 - 1 = \log \frac{7}{2}$. (Si le pianiste était doué, on aurait $\mathsf{cap} = \log 7$).

Si le compositeur n'utilisait que les notes $\mathbb{M} = \{\text{do, mi, sol}\}$, par exemple, (c'est-à-dire il raréfiait suffisamment l'ensemble des messages pour que les signaux transmis appartiennent à des parties disjointes de $\mathbb{Y}$), on pourrait décoder l'œuvre, telle qu'interprétée par le pianiste, sans erreur[2]. Par contre, si le pianiste était doué, on pourrait décoder sans erreur sans raréfaction de $\mathbb{M}$ car alors la capacité serait $\log 7 = \log \mathrm{card}\mathbb{M}$.

2. Le théorème ne dit rien sur ... la musicalité de l'interprétation. Le titre de cet exemple est un jeu de mots faisant référence à l'œuvre *Das wohltemperierte Klavier* de J.-S. Bach, connu en français sous le titre *Le clavier bien tempéré*. *Falschtemperierte* pourrait se traduire par *mal tempéré*.

L'idée intuitive sous-tendant la démonstration du théorème 9.5.4 est que tout canal, pour des blocs de taille n suffisamment grande, se comporte comme l'exemple 9.5.5.

9.5.4 Démonstration du théorème 9.5.4

Typicité conjointe

Définition 9.5.6. Soient $\mathbb{X}$ et $\mathbb{Y}$ deux alphabets finis. On considère un couple de variables aléatoires (X, Y) à valeurs dans $\mathbb{X} \times \mathbb{Y}$ de loi conjointe κ, i.e. $\mathbb{P}(X = x, Y = y) = \kappa(x, y)$, pour $x \in \mathbb{X}$ et $y \in \mathbb{Y}$. On note μ la première marginale de κ et ν la seconde, i.e.

$$\mu(x) = \sum_{y \in \mathbb{Y}} \kappa(x, y), x \in \mathbb{X} \qquad \text{et} \qquad \nu(y) = \sum_{x \in \mathbb{X}} \kappa(x, y), y \in \mathbb{Y}.$$

L'ensemble des configurations **conjointement typiques** (pour la probabilité κ) à niveau $\varepsilon > 0$ et taille $n \in \mathbb{N}_>$ est l'ensemble

$$\begin{aligned} \mathbb{J} := \mathbb{J}^{\kappa,\varepsilon,n} = \{(\mathbf{x}, \mathbf{y}) \in \mathbb{X}^n \times \mathbb{Y}^n \ : \ & |-\frac{1}{n}\sum_{i=1}^n \log \mu(x_i) - H(\mu)| < \varepsilon, \\ & |-\frac{1}{n}\sum_{i=1}^n \log \nu(x_i) - H(\nu)| < \varepsilon, \\ & |-\frac{1}{n}\sum_{i=1}^n \log \kappa(x_i, y_i) - H(\kappa)| < \varepsilon\}. \end{aligned}$$

Il s'avère pratique dans la suite de décomposer l'ensemble $\mathbb{J} = A_1 \cap A_2 \cap A_3$, où

$$A_1 := A_{1,\kappa,\varepsilon,n} = \left\{ \mathbf{x} \in \mathbb{X}^n : |-\frac{1}{n}\sum_{i=1}^n \log \mu(x_i) - H(\mu)| < \varepsilon \right\},$$

$$A_2 := A_{2,\kappa,\varepsilon,n} = \left\{ \mathbf{y} \in \mathbb{Y}^n : |-\frac{1}{n}\sum_{i=1}^n \log \nu(y_i) - H(\nu)| < \varepsilon \right\},$$

$$A_3 := A_{3,\kappa,\varepsilon,n} = \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{X}^n \times \mathbb{Y}^n : |-\frac{1}{n}\sum_{i=1}^n \log \kappa(x_i, y_i) - H(\kappa)| < \varepsilon \right\},$$

et d'introduire les abbreviations

$$\mu_n(\mathbf{x}) := \prod_{i=1}^n \mu(x_i), \ \text{ pour } \ \mathbf{x} \in \mathbb{X}^n,$$

$$\nu_n(\mathbf{y}) := \prod_{i=1}^n \nu(y_i), \ \text{ pour } \ \mathbf{y} \in \mathbb{Y}^n,$$

$$\kappa_n(\mathbf{x}, \mathbf{y}) := \prod_{i=1}^n \kappa(x_i, y_i) \ \text{ pour } \ \mathbf{y} \in \mathbb{X}^n \times \mathbb{Y}^n.$$

Théorème 9.5.7. *Soit $(X_i, Y_i)_{i \in \mathbb{N}}$ une suite de couples aléatoires à valeurs dans $\mathbb{X} \times \mathbb{Y}$ indépendants et identiquement distribués selon la loi conjointe κ, i.e. $\forall i \in \mathbb{N}$,*

$$\mathbb{P}((X_i, Y_i) = (x, y)) = \mathbb{P}((X_1, Y_1) = x, y) = \kappa(x, y).$$

On note $\mathbf{X}\restriction_n$ (resp. $\mathbf{X}\restriction_n$) la troncature des suites infinies aux n premiers termes, i.e. $\mathbf{X}\restriction_n = (X_1, X_2, \ldots, X_n)$ (resp. $\mathbf{Y}\restriction_n = (Y_1, Y_2, \ldots, Y_n)$). Alors, pour $\varepsilon \in]0, 1[$ (suffisamment petit)

1. $\lim_{n\to\infty} \mathbb{P}\left((\mathbf{X}\restriction_n, \mathbf{Y}\restriction_n) \in \mathbb{J}^{\kappa,\varepsilon,n}\right) \overset{\mathbb{P}}{=} 1$.
2. $(1-\varepsilon)2^{n(H(\kappa)-\varepsilon)} \leq |\mathbb{J}^{\kappa,\varepsilon,n}| \leq 2^{n(H(\kappa)+\varepsilon)}$.
3. *Soient* $\tilde{\mathbf{X}} = (\tilde{X}_i)_{i\in\mathbb{N}}$ *et* $\tilde{\mathbf{Y}} = (\tilde{Y}_i)_{i\in\mathbb{N}}$ *des suites aléatoires indépendantes et mutuellement indépendantes, à valeurs respectivement dans* $\mathbb{X}$ *et* $\mathbb{Y}$*, distribués selon les lois* $\mathbb{P}(\tilde{X}_1 = x) = \mu(x)$ *et* $\mathbb{P}(\tilde{Y}_1 = y) = \nu(y)$*, i.e. la loi du couple est donnée par* $\mathbb{P}\left((\tilde{\mathbf{X}}\restriction_n, \tilde{\mathbf{Y}}\restriction_n) = (x,y)\right) = \mu(x)\nu(x)$*. Alors*

$$(1-\varepsilon)2^{-nI(X_1:Y_1)+3\varepsilon)} \leq \mathbb{P}\left((\tilde{\mathbf{X}}\restriction_n, \tilde{\mathbf{Y}}\restriction_n) \in \mathbb{J}^{\kappa,\varepsilon,n}\right) \leq 2^{-nI(X_1:Y_1)-3\varepsilon)},$$

où $I(X_1 : Y_1)$ *désigne l'information mutuelle entre les variables aléatoires* X_1 *et* Y_1*.*

Démonstration. Voir exercice 105. □

Démonstration de l'affirmation directe

Le but de ce paragraphe est de démontrer la première affirmation du théorème 9.5.4 (dans la formulation simplifée 9.5.10 ci-dessous).

Remarque 9.5.8. Le code avec des tailles de blocks n pourrait coder jusqu'à $|\mathbb{X}|^n$ messages. Ce que nous dit le théorème fondamental de la transmission est que si $\mathsf{cap} < 1$ on peut, en raréfiant l'ensemble de message à un cardinal $k = |\mathbb{X}|^{nR}$ avec $R < \mathsf{cap}$, transmettre les messages sans erreur. C'est précisément ce résultat que est illustré par l'exemple 9.5.5 et c'est le résultat que nous allons démontrer dans la suite.

Dans la suite on raréfiera l'ensemble de messages $\mathbb{M}$ en $\mathbb{M}_k \subset \mathbb{M}$ avec $|\mathbb{M}_k| = k < |\mathbb{M}|$. Le paramètre k sera déterminé plus tard. Sans perte de généralité, on identifie l'ensemble $\mathbb{M}_k$ avec $\{1,\dots,k\}$.

En utilisant un code $\mathbf{C} : \mathbb{M} \to \mathbb{X}^n$, nous avons déjà établi que la probabilité d'erreur de transmission d'un message spécifique $m \in \mathbb{M}$ est donnée par

$$e(m) = \mathbb{P}(\hat{M} \neq m) = \sum_{\mathbf{y}\in\mathbb{Y}^n} Q_n(\mathbf{C}(m),\mathbf{y})\mathbb{1}_{\mathbb{M}\setminus\{m\}}(\hat{M}(\mathbf{y}))$$

où $\hat{M}$ est un estimateur (souvent du maximum de vraisemblance) du message d'entrée lorsque le sortie est déterminée par le canal. Nous allons généraliser ce modèle de plusieurs façons.

— Étant donné que nous raréfions l'ensemble de messages, nous allons appliquer la formule de probabilité d'erreur sur les messages de $\mathbb{M}_k$ au lieu de $\mathbb{M}$, i.e.

$$e(m) = \mathbb{P}(\hat{M} \neq m) = \sum_{\mathbf{y}\in\mathbb{Y}^n} Q_n(\mathbf{C}(m),\mathbf{y})\mathbb{1}_{\mathbb{M}_k\setminus\{m\}}(\hat{M}(\mathbf{y})), \text{ pour } m \in \mathbb{M}_k. \quad (9.1)$$

— Connaître le code $\mathbf{C}$ c'est connaître son **glossaire** $\mathcal{G}$, i.e. la liste ordonnée $\mathcal{G} = (\mathbf{C}(m))_{m\in\mathbb{M}_k} = (\mathbf{C}(1),\dots,\mathbf{C}(k))$. Dans la suite, nous allons considérer des **codes aléatoires**, i.e. nous allons introduire un aléa supplémentaire et, pour chaque $m \in \mathbb{M}_k$, associer à $\mathbf{C}(m)$ à un mot aléatoire $\mathbf{X}^{(m)} \in \mathbb{X}^n$ de longueur n sur l'alphabet $\mathbb{X}$. La loi des variables aléatoires $\mathbf{X}^{(m)}$ est donnée par μ_n, i.e.

$$\mathbb{P}(\mathbf{X}^{(m)} = \mathbf{x}^{(m)}) = \mu_n(\mathbf{x}^{(m)}) = \mu(x_1^{(m)})\cdots\mu(x_n^{(m)})$$

et les variables $\mathbf{X}^{(l)}$ et $\mathbf{X}^{(m)}$ codant des messages différents $l \neq m$ sont indépendantes. Par conséquent, le choix du code sera déterminé par la loi

$$\mathbb{P}\left(\mathcal{G} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)})\right) = \prod_{m=1}^{k} \mu_n(\mathbf{x}^{(m)}) = \prod_{m=1}^{k} \prod_{i=1}^{n} \mu(x_i^{(m)}). \tag{9.2}$$

— Supposons que l'on s'intéresse à la probabilité d'erreur lors de la transmission du message spécifique $m \in \mathbb{M}_k$.
L'estimateur $\hat{M}$ que nous allons utiliser ne sera pas celui de maximum de vraisemblance mais celui de **typicité conjointe**. L'estimateur $\hat{M}$ de typicité conjointe nécessite l'information pas seulement de la variable $\mathbf{Y}^{(m)}$ mais aussi de tous les mots potentiellement contenus dans le glossaire. $\hat{M}$ sera m si les deux conditions suivantes sont simultanément vérifées :
— $(\mathbf{X}^{(m)}, \mathbf{Y}^{(m)}) \in \mathbb{J}^{\kappa,\varepsilon,n}$ et
— il n'existe aucun autre message $l \in \mathbb{M}_k$, tel que $(\mathbf{X}^{(l)}, \mathbf{Y}^{(m)}) \in \mathbb{J}^{\kappa,\varepsilon,n}$.
Si l'une des conditions est violée, on introduit un nouveau symbole $\partial \notin \mathbb{M}_k$ et on retourne ce symbole comme estimation du message d'entrée (il contribuera toujours à l'erreur). Ainsi, pour un mot $\mathbf{Y}$ arbitraire,

$$\hat{M}(\mathbf{Y}) = \begin{cases} r & \text{si } (\mathbf{X}^{(r)}, \mathbf{Y}) \in \mathbb{J} \text{ et } \forall l \neq r, (\mathbf{X}^{(l)}, \mathbf{Y}) \notin \mathbb{J}, \\ \partial & \text{si } \forall r, (\mathbf{X}^{(r)}, \mathbf{Y}) \notin \mathbb{J} \text{ ou } \exists (l,r), l \neq r \text{ avec } (\mathbf{X}^{(l)}, \mathbf{Y}) \in \mathbb{J}, (\mathbf{X}^{(r)}, \mathbf{Y}) \in \mathbb{J}. \end{cases} \tag{9.3}$$

Remarque 9.5.9. Dans la définition de l'estimateur $\hat{M}$, on reconnaît l'idée véhiculée par l'exemple 9.5.5 du polycopié : on peut décoder sans erreur si l'espace de messages $\mathbb{M}_k$ est tellement raréfié que le code $\mathbf{C} : \mathbb{M}_k \to \mathbf{C}(\mathbb{M}_k)$, avec $\mathbf{C}(\mathbb{M}_k) \subset \mathbb{Y}^n$, peut-être rendu (approximativement) bijective.

Le schéma du canal et de son codage est donné en figure 1.

On note K le vecteur de probabilité conjointe des variables $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(k)}, \mathbf{Y}^{(m)}$ pour un $m \in \mathbb{M}_k$ (cf. figure 9.5) et $\mathbb{P}_m$ la mesure de probabilité conjointe de toutes les variables aléatoires lorsque nous commençons par le message m :

$$\mathbb{P}_m(\mathbf{X}^{(1)} = \mathbf{x}^{(1)}, \dots, \mathbf{X}^{(k)} = \mathbf{x}^{(k)}; \mathbb{Y}^{(m)} = \mathbf{y}^{(m)}) = K(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)}; \mathbf{y}^{(m)}) \tag{9.4}$$

$$= \kappa_n(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}) \prod_{\substack{l=1 \\ l \neq m}}^{k} \mu_n(\mathbf{x}^{(l)}) \tag{9.5}$$

où $\kappa_n(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}) = \mu_n(\mathbf{x}^{(m)}) Q_n(\mathbf{x}^{(m)}, \mathbf{y}^{(m)})$ et Q_n est la matrice de transition pour le canal avec des entrées et sorties de blocks de taille n.

Théorème 9.5.10 (Version simplifiée du théorème fondamental)**.** *Soit $(\mathbb{X}, \mathbb{Y}, P)$ un canal sans mémoire, de capacité $\mathsf{cap} := \mathsf{cap}(P)$. Tout $[k,n]$-code de taux de transmission $R = \frac{\log_{|\mathbb{X}|} k}{n} < \mathsf{cap}$ est (asymptotiquement à grand n) atteignable.*

Démonstration. Voir exercice 106 □

La démonstration de la réciproque se base sur des arguments similaires à ceux développpés dans les exercices 105 et 106. La réciproque faible peut être consultée dans [60, pp. 98–101] ; la réciproque forte dans [60, pp. 101–102].

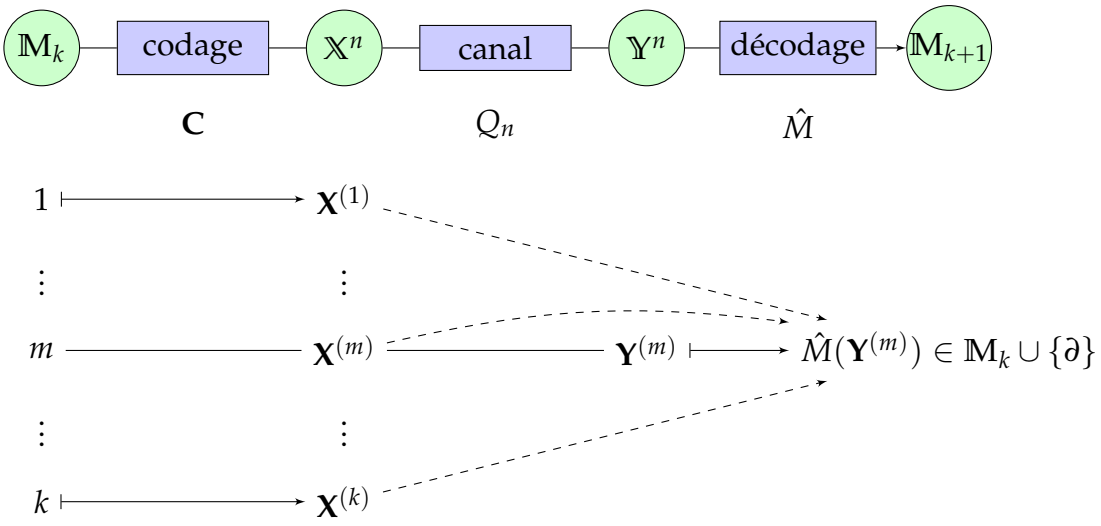


FIGURE 9.5 – Résumé des actions de codage du message $m \in \mathbb{M}_k$, de sa transmission et de son décodage. Les flèches pointillées signifient qu'une information concernant les $\mathbf{X}^{(l)}$ est utilisée lors du décodage. Il faut cependant comprendre que cette information est juste une astuce théorique pour démontrer le théorème ; on ne sous-entend pas que $\hat{M}$ a effectivement besoin de cette information contenue dans la réalisation précise des variables $\mathbf{X}^{(l)}$ pour effectuer le décodage.

9.6 Exercices

Codage du canal

98. Pour trois variables aléatoires X, W, Y discrètes arbitraires, établir les relations suivantes :
    (a) $H(W,Y|X) \leq H(W|X) + H(Y|X)$.
    (b) $H(W,Y|X) = H(W|X) + H(Y|X,W)$.
    (c) $H(Y|X,W) \leq H(Y|X)$.
99. Soient les canaux $\mathcal{K}_1 = (\mathbb{X}, \pi, \mathbb{W}, P)$ et $\mathcal{K}_2 = (\mathbb{W}, \rho, \mathbb{Y}, Q)$ où $\mathbb{X}, \mathbb{W}$ et $\mathbb{Y}$ sont des alphabets d'entrée ou de sortie et P et Q des matrices de transmission. On construit le canal $\mathcal{K} = (\mathbb{X}, \pi, \mathbb{Y}, PQ)$ en mettant les canaux $\mathcal{K}_1$ et $\mathcal{K}_2$ en cascade.
    (a) Comparer les probabilités conditionnelles $\mathbb{P}(Y = y|X = x, W = w)$ et $\mathbb{P}(Y = y|W = w)$.
    (b) Montrer que la capacité $C_{\mathcal{K}}$ du canal composé ne peut pas excéder la plus petite des capacités $C_{\mathcal{K}_1}$ et $C_{\mathcal{K}_2}$.
    (c) Commenter ce dernier résultat.
100. Soit $\mathcal{K} = (\mathbb{X}, \mathbb{Y}, P)$ le canal avec alphabets d'entrée et de sortie $\mathbb{X} = \mathbb{Y}$, ayant $\mathrm{card}\mathbb{X} = \mathrm{card}\mathbb{Y} = 27$. On suppose que ces ensembles sont ordonnés ; on peut par conséquent les identifier avec $\mathbb{Z}_{27} = \{0, \ldots, 26\}$ d'une part et avec l'alphabet latin augmenté du symbole blanc d'autre part à travers la bijection $0 \to \text{␣}$, $1 \to$ a, …, $26 \to$ z. La périodisation signifie que le symbole ␣$\equiv 0$ succède à z $\equiv 26$ et précède a $\equiv 1$. Lorsque l'entrée du canal X est une lettre $x \in \mathbb{X}$, la sortie est une variable aléatoire Y uniformément distribuée dans $\{x - 1 \mod 27, x, x + 1 \mod 27\} \subset \mathbb{Y}$.
    Suggestion : Pour cet exercice vous pouvez vous inspirez de l'exemple du « clavier mal tempéré » traité en cours.

(a) Déterminer la matrice de transmission $P := (P(x,y))_{x,y\in\mathbb{Z}_{27}}$ du canal, définie par
$$\mathbb{P}(Y = y|X = x).$$
(b) Calculer $H(Y|X)$. *Suggestion :* l'entropie s'exprime facilement soit en trits — i.e. en $\log_3$ — soit en $\log_{27}$.
(c) Le vecteur de probabilité $\mu \in \mathcal{M}_1(\mathbb{X})$ uniforme, i.e. $\mu(x) = 1/27$ pour tout $x \in \mathbb{X}$, est-il invariant par P ?
(d) Pour une entrée du canal distribuée selon μ, calculer $H(Y)$ où Y est la variable de sortie.
(e) Déterminer la capacité $\mathsf{cap}(\mathcal{K})$ du canal.
(f) Soit $\mathbb{M} = \mathbb{X}$ l'ensemble de messages que nous envisageons de faire transiter par le canal et $C : \mathbb{M} \to \mathbb{X}$ le codage « identité » $C = \mathbb{1}$. Quelle est la dilution minimale de l'ensemble $\mathbb{M}$ pour que le théorème fondamental de la transmission nous garantisse le possibilité de transmission sans erreur ?

101. Soient $\mathbb{X}$ un espace fini, X une variable aléatoire à valeurs dans $\mathbb{X}$, $g : \mathbb{X} \to \mathbb{W}$ une application arbitraire et Y une variable aléatoire à valeurs dans $\mathbb{Y}$ de loi conjointe κ avec X. Monter par deux méthodes différentes que $H(Y|g(X)) \geq H(Y|X)$.
102. **(Codage du canal et décision optimale)** – Extrait de l'examen du 19 décembre 2013.

Soit un canal discret sans mémoire ayant un alphabet d'entrée $\mathbb{X}$, un alphabet de sortie $\mathbb{Y}$, une matrice stochastique de transmission P et une loi des symboles d'entrée déterminée par le vecteur de probabilité π. Lorsque un symbole $y \in \mathbb{Y}$ est transmis, on le décode par un schéma de décision qui peut être soit une fonction déterministe $d : \mathbb{Y} \to \mathbb{X}$ soit une variable aléatoire D définie sur $\mathbb{Y}$ à valeurs dans $\mathbb{X}$.

(a) Déterminer les vecteurs de probabilité κ et ρ définissant respectivement la loi conjointe et la loi des symboles de sortie.
(b) Pour une règle de décodage $d : \mathbb{Y} \to \mathbb{X}$, on note respectivement $\mathsf{DC}(d)$ et $\mathsf{DE}(d)$ les probabilités de décodage correct et erroné. Montrer que $\mathsf{DC}(d) = \sum_{y\in\mathbb{Y}} \rho(y)\mathbb{P}(X = d(y)|Y = y)$. Que vaudra alors $\mathsf{DE}(d)$?
(c) On appelle **décodage déterministe optimal** la règle de décision définie pour tout $y \in \mathbb{Y}$ par la formule $d_o(y) = x_y$, où x_y maximise la probabilité conditionnelle $\mathbb{P}(X = x|Y = y)$, c'est-à-dire :
$$x_y = \arg\max_x \mathbb{P}(X = x|Y = y).$$
Montrer que la règle de décision ainsi définie est optimale, c'est-à-dire, pour toute autre règle déterministe d', on $\mathsf{DC}(d') \leq \mathsf{DC}(d_o)$.
(d) Soit D une règle de décision stochastique arbitraire correspondant à la probabilité conditionnelle $Q_{yx} = \mathbb{P}(D(y) = x|Y = y)$. Montrer que $\mathsf{DC}(D) \leq \mathsf{DC}(d_o)$.
(e) Dorénavant $\mathbb{X} = \{x_1, x_2, x_3\}$, $\mathbb{Y} = \{y_1, y_2, y_3\}$, $\pi = (1/2, 1/4, 1/4)$ et $P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \end{pmatrix}$, calculer ρ.

(f) Soit $d : \mathbb{Y} \to \mathbb{X}$ la règle déterministe de décodage donnée par la formule $d(y_1) = x_1; d(y_2) = d(y_3) = x_3$. Calculer la valeur numérique de la probabilité de décodage correct.

103. (Extrait de l'examen du 20 décembre 2017).

Soit $\mathcal{K} = (\mathbb{X}, \mathbb{Y}, P, \mu)$ un canal avec alphabets d'entrée et de sortie finis, notés respectivement $\mathbb{X}$ et $\mathbb{Y}$, matrice stochastique P et probabilité des symboles de la source μ. Nous écrirons $\mu(x)$ au lieu de $\mu(\{x\})$, i.e. nous identifierons — comme d'habitude — la mesure de probabilité μ avec le vecteur (ligne) de probabilité $(\mu(x))_{x\in\mathbb{X}}$. Déterminer, pour des variables d'entrée X et de sortie Y et pour des symboles $x \in \mathbb{X}$ et $y \in \mathbb{Y}$:

(a) La mesure de probabilité conjointe $\kappa(x, y) := \mathbb{P}(X = x, Y = y)$.
(b) La mesure de probabilité de sortie ν
(c) La probabilité $Q(y, x) := \mathbb{P}(X = x|Y = y)$ que l'entrée soit x, sachant que la sortie observée est y.
(d) On fixe maintenant les alphabets d'entrée et de sortie $\mathbb{X} = \mathbb{Y} = \{0, 1\}$ (canal binaire) et la loi [3] de la source $\mu = (1/4, 3/4)$.
   i. Si le canal est symétrique avec taux d'erreur $f = 1/16$,
      — Déterminer la matrice stochastique P du canal.
      — Determiner le vecteur probabilité ν de sortie.
      — Calculer $Q(1, 0)$ et $Q(1, 1)$.
   ii. Si le canal est « en Z » avec taux d'erreur $f = 1/16$, i.e. sa matrice stochastique est $P = \begin{pmatrix} 1 & 0 \\ 1/16 & 15/16 \end{pmatrix}$, calculer
      — Le vecteur de probabilité de sortie ν.
      — Les probabilités $Q(1, 0)$ et $Q(1, 1)$.

104. **(« Somme » de deux canaux).** Soient K_i, avec $i = 1, 2$ deux canaux avec alphabets d'entrée $\mathbb{X}_i$, alphabets de sortie $\mathbb{Y}_i$ et matrices de transmission P_i. On note $\mathbb{X} = \mathbb{X}_1 \boxplus \mathbb{X}_2$ (si les alphabets $\mathbb{X}_1$ et $\mathbb{X}_2$ sont distincts alors $\mathbb{X}_1 \boxplus \mathbb{X}_2 = \mathbb{X}_1 \sqcup \mathbb{X}_2$; s'ils ne sont pas distincts, on commence par distinguer artificiellement les éléments de $\mathbb{X}_1$ et de $\mathbb{X}_2$ avant de prendre leur réunion). Il en va de même de $\mathbb{Y} = \mathbb{Y}_1 \boxplus \mathbb{Y}_2$. Finalement la matrice de transmission du canal « somme » est la matrice bloc $P = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix}$.

(a) Soient X une variable aléatoire à valeurs dans $\mathbb{X}$ dont la loi est décrite par le vecteur de probabilité π et Y une variable aléatoire à valeurs dans $\mathbb{Y}$, dont la loi est déterminée par le canal. On note κ la loi conjointe du couple (X, Y). Soit $p = \sum_{x\in\mathbb{X}_1} \pi(x)$ (donc $1 - p = \sum_{x\in\mathbb{X}_2} \pi(x)$). Pour $x \in \mathbb{X}_1$ on note $\rho_1(x) = \frac{\pi(x)}{p}$ et pour $x \in \mathbb{X}_2$ on note $\rho_2(x) = \frac{\pi(x)}{1-p}$. Calculer $H(\pi)$
(b) Montrer que

$$H(X|Y) = -\,p \sum_{x\in\mathbb{X}_1, y\in\mathbb{Y}_1} \rho_1(x) P_1(x, y) \log \mathbb{P}(X = x|Y = y) \\ -\,(1-p) \sum_{x\in\mathbb{X}_2, y\in\mathbb{Y}_2} \rho_2(x) P_2(x, y) \log \mathbb{P}(X = x|Y = y).$$

3. Les valeurs numériques données pour les probabilités de la source et de l'erreur permettent de déterminer toutes les autres probabilités comme des rationnels calculables facilement à la main.

(c) On considère des variables aléatoires X_1 et X_2 respectivement à valeurs dans $\mathbb{X}_1$ et $\mathbb{X}_2$ et de lois ρ_1 et ρ_2 ; on note Y_1 et Y_2 les variables aléatoires obtenues par restriction de Y sur $\mathbb{Y}_1$ et $\mathbb{Y}_2$. Montrer que

$$H(X|Y) = pH(X_1|Y_1) + (1-p)H(X_2|Y_2)$$

et conclure que

$$C(p) := \sup_{\pi:\sum_{x\in\mathbb{X}_1}\pi(x)=p} I(X:Y) = H(p,1-p) + pC_1 + (1-p)C_2.$$

(d) Montrer que la valeur de p qui maximise $C(p)$ est $p = \frac{2^{C_1}}{2^{C_1}+2^{C_2}}$.

(e) En conclure que la capacité du canal « somme » vérifie $2^C = 2^{C_1} + 2^{C_2}$).

105. **(Typicité conjointe).** On utilise la notation introduite dans le paragraphe sur la typicité conjointe.

(a) Montrer que

$$\forall \mathbf{x} \in A_{1,\kappa,\varepsilon,n} : \quad 2^{-n(H(\mu)+\varepsilon)} \le \mu_n(\mathbf{x}) \le 2^{-n(H(\mu)-\varepsilon)}.$$

Le deux encadrements qui suivent se démontrent de la même façon. Il n'est pas demandé de les (re)démontrer ; vous aurez cependant à les utiliser ultérieurement.

$$\forall \mathbf{y} \in A_{2,\kappa,\varepsilon,n} : \quad 2^{-n(H(\nu)+\varepsilon)} \le \nu_n(\mathbf{y}) \le 2^{-n(H(\nu)-\varepsilon)},$$

et

$$\forall (\mathbf{x},\mathbf{y}) \in A_{3,\kappa,\varepsilon,n} : \quad 2^{-n(H(\kappa)+\varepsilon)} \le \kappa_n(\mathbf{x},\mathbf{y}) \le 2^{-n(H(\kappa)-\varepsilon)}.$$

(b) Montrer que pour $\forall \varepsilon > 0$,

$$\begin{aligned}
\exists n_1 := n_1(\varepsilon) \ge 1 \ &: \ \forall n \ge n_1, \ \mathbb{P}(\mathbf{X}\restriction_n \in A^c_{1,\kappa,\varepsilon,n}) \le \frac{\varepsilon}{3},\\
\exists n_2 := n_2(\varepsilon) \ge 1 \ &: \ \forall n \ge n_2, \ \mathbb{P}(\mathbf{Y}\restriction_n \in A^c_{2,\kappa,\varepsilon,n}) \le \frac{\varepsilon}{3},\\
\exists n_3 := n_3(\varepsilon) \ge 1 \ &: \ \forall n \ge n_3, \ \mathbb{P}((\mathbf{X}\restriction_n, \mathbf{Y}\restriction_n) \in A^c_{3,\kappa,\varepsilon,n}) \le \frac{\varepsilon}{3}.
\end{aligned}$$

(c) Conclure que pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left((\mathbf{X}\restriction_n, \mathbf{Y}\restriction_n) \in \mathbb{J}^{\kappa,\varepsilon,n}\right) = 1.$$

(d) En utilisant la minoration de $\kappa_n(\mathbf{x},\mathbf{y})$ pour $(\mathbf{x},\mathbf{y}) \in A_{3,\kappa,\varepsilon,n}$ — établie en question 105a — montrer que $|\mathbb{J}^{\kappa,\varepsilon,n}| \le 2^{n(H(\kappa)+\varepsilon)}$.

(e) Montrer que $\mathbb{P}(\tilde{\mathbf{X}}\restriction_n, \tilde{\mathbf{Y}}\restriction_n) \in \mathbb{J}^{\kappa,\varepsilon,n}) \le 2^{-n(I(X_1:Y_1)-3\varepsilon)}$.

(f) Montrer que $|\mathbb{J}^{\kappa,\varepsilon,n}| \ge (1-\varepsilon)2^{n(H(\kappa)-\varepsilon)}$.

(g) Conclure que

$$\mathbb{P}\left((\tilde{\mathbf{X}}\restriction_n, \tilde{\mathbf{Y}}\restriction_n) \in \mathbb{J}^{\kappa,\varepsilon,n}\right) \ge (1-\varepsilon)2^{-nI(X_1:Y_1)+3\varepsilon)}.$$

106. **(Théorème fondamental de la transmission).**

(a) Montrer que la loi conjointe du couple $(\mathbf{X}^{(m)}, \mathbf{Y}^{(m)})$ est κ_n et celle du couple $(\mathbf{X}^{(m)}, \mathbf{Y}^{(m)})$, avec $l \ne m$, est $\mu_n \otimes \nu_n$, où,

$$\nu_n(\mathbf{z}) = \sum_{\mathbf{w}\in\mathbb{X}^n} \kappa_n(\mathbf{w},\mathbf{z}) = \sum_{\mathbf{w}\in\mathbb{X}^n} \mu_n(\mathbf{w})Q_n(\mathbf{w},\mathbf{z}), \ \mathbf{z} \in \mathbb{Y}^n.$$

(b) Que peut-on conclure sur la dépendance des variables aléatoires composant le couple $(\mathbf{X}^{(m)}, \mathbf{Y}^{(m)})$? Même question pour le variables du couple $(\mathbf{X}^{(l)}, \mathbf{Y}^{(m)})$, avec $l \neq m$.

(c) Partant de l'expression évidente, valable pour un $m \in \mathbb{M}_k$ fixé, pour la probabilité d'erreur :

$$e(m) = \mathbb{P}(\hat{M} \neq m) = \sum_{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(k)} \in \mathbb{X}^n} \mathbb{P}(\hat{M} \neq m | \mathcal{G} = (\mathbf{x}^{(1)},\dots,\mathbf{x}^{(k)}))\mathbb{P}(\mathcal{G} = (\mathbf{x}^{(1)},\dots,\mathbf{x}^{(k)})), \tag{9.6}$$

montrer que

$$e(m) \leq \sum_{\substack{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(k)} \in \mathbb{X}^n \\ \mathbf{y}^{(m)} \in \mathbb{Y}^n}} K(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(k)};\mathbf{y}^{(m)}) \left(\mathbb{1}_{\mathbb{J}^c}(\mathbf{x}^{(m)},\mathbf{y}^{(m)}) + \sum_{\substack{l=1 \\ l\neq m}}^{n} \mathbb{1}_{\mathbb{J}}(\mathbf{x}^{(l)},\mathbf{y}^{(m)}) \right). \tag{9.7}$$

(d) Utiliser le résultat, obtenu en question 106b — à savoir que la loi conjointe du couple $(\mathbf{X}^{(m)}, \mathbf{Y}^{(m)})$ est κ_n — et le résultat obtenu en question 105f pour majorer la probabilité

$$\mathbb{P}((\mathbf{X}^{(m)}, \mathbf{Y}^{(m)}) \in (\mathbb{J}^{\kappa,\varepsilon,n})^c) < \varepsilon.$$

asymptotiquement à grand n.

(e) Utiliser les résultats, obtenus aux questions 106b et 105e pour majorer la probabilité

$$\mathbb{P}((\mathbf{X}^{(l)}, \mathbf{Y}^{(m)}) \in \mathbb{J}^{\kappa,\varepsilon,n})$$

asymptotiquement à grand n.

(f) Conclure que

$$e(m) \leq \varepsilon + (k-1)2^{-n(I(X_1:Y_1)-3\varepsilon)}$$

où le couple (X_1, Y_1) est distribué selon κ.

(g) Choisir maintenant $k = 2^{nR}$. Montrer que tant que $R < \mathsf{cap}$, la transmission peut se faire avec une erreur négligeable.

10 Chiffrement

Desiderata de la cryptographie militaire
Il faut bien distinguer entre un système d'écriture chiffrée, imaginé pour un échange momentané de lettres entre quelques personnes isolées, et une méthode de cryptographie destinée à régler pour un temps illimité la correspondance des différents chefs d'armée entre eux. Ceux-ci, en effet, ne peuvent à leur gré et à un moment donné, modifier leurs conventions ; de plus, ils ne doivent jamais garder sur eux aucun objet ou écrit qui soit de nature à éclairer l'ennemi sur le sens des dépêches sécrètes qui pourraient tomber entre ses mains.
Un grand nombre de combinaisons ingénieuses peuvent répondre au but qu'on veut atteindre dans le premier cas ; dans le second, il faut un système remplissant certaines conditions exceptionnelles, conditions que je résumerai sous les six chefs suivants :

1. Le système doit être matériellement, sinon mathématiquement, indéchiffrable ;
2. Il faut qu'il n'exige pas le secret, et qu'il puisse sans inconvénient tomber entre les mains de l'ennemi ;
3. La clef doit pouvoir en être communiquée et retenue sans le secours de notes écrites, et être changée ou modifiée au gré des correspondants ;
4. Il faut qu'il soit applicable à la correspondance télégraphique ;
5. Il faut qu'il soit portatif, et que son maniement ou son fonctionnement n'exige pas le concours de plusieurs personnes ;
6. Enfin, il est nécessaire, vu les circonstances qui en commandent l'application, que le système soit d'usage facile, ne demandant ni tension d'esprit, ni la connaissance d'une longue série de règles à observer.

Augustus Kerchoffs : *La cryptographie militaire* ; Journal des sciences militaires (1883) [40, page 12].

Dans l'extrait ci-dessus, Augustus Kerchoffs énonçait, en 1883, avec une surprenante modernité, les conditions que doit remplir un bon système de cryptage. On pourrait utiliser ce passage — presque mot pour mot — pour formuler les exigences cryptographiques d'aujourd'hui.

10.1 Sécurité des communications

Dans ce chapitre, nous nous intéressons aux méthodes de protection de l'information basées sur le partage d'une clé secrète entre les partenaires qui souhaitent communiquer. Ceci se réalise par le partage d'une clé aléatoire privée qui est utilisée de manière symétrique entre les partenaires, contrairement aux méthodes asymétriques à clé publique (comme RSA par exemple).

10.1.1 Les exigences du chiffrement

Le cryptage est conçu (voir [21] ou [31, 32] par exemple) pour assurer :

— La **confidentialité** des messages transmis, dans le sens que seul le récipiendaire légal du cryptogramme (possesseur de la clé) peut le décoder facilement. D'autres personnes peuvent intercepter le cryptogramme mais, ne possédant pas la clé, ne peuvent pas le décoder (dans un temps raisonnable . . .).

— La **signature** du message par son expéditeur qui permet à toute personne de vérifier efficacement la signature de l'expéditeur sur un document spécifique assurant ainsi les fonctions d'**intégrité** [1] et de **non-rétractation** [2].

— L'**authentification** qui permet à l'expéditeur de s'identifier auprès du destinataire légitime en lui prouvant qu'il connaît un secret que lui seul et le destinataire connaissent.

La discussion à propos des signatures irréfutables a été initiée durant les siècles précédents par le besoin qu'ont différentes parties de s'engager par des propositions ou de déclarations qu'elles font. Mais ces discussions ont surtout porté sur l'irréfutabilité de signatures manuscrites. En transposant (et en renforçant) ces exigences aux communications numérisées, un schéma de signature numérique irréfutable doit vérifier les conditions suivantes :

— chaque utilisateur peut efficacement apposer sa signature sur tout document de son choix,

— tout utilisateur peut efficacement vérifier si une chaîne de caractère est la signature d'un autre utilisateur spécifique apposée sur un document spécifique et

— il est impossible d'apposer des signatures de tierces personnes sur des documents qu'elles n'ont pas effectivement signés.

Le schéma analogue pour l'authentification doit vérifier les conditions suivantes :

— chacun des deux partenaires peut générer efficacement un sceau d'authentification *(authentication tag)* sur tout document de son choix ;

— chacun des deux partenaires peut vérifier efficacement si un mot est le sceau d'authentification d'un message donné ;

— nul adversaire ne peut produire efficacement des sceaux d'authentification sur des messages non émis par les deux partenaires.

Il est évident qu'une signature électronique assure l'authentification mais, en général, l'authentification ne suffit pas comme signature comme le résume le tableau suivant :

Action	Clé de vérification	Possibilité de vérification
Signature	Connue de tout le monde (y compris l'adversaire)	Par tout le monde
Authentification	Connue uniquement des partenaires légitimes	Par les partenaires légitimes

10.1.2 Les niveaux de sécurité

Un critère essentiel de la qualité d'un chiffrement est la difficulté que l'on rencontre pour le casser. Le cryptage (sans que la partie adverse connaisse la clé de cryptage)

1. Elle garantit à tout destinataire potentiel qu'aucun intrus malicieux n'ait apporté des modifications volontaires au message.

2. Dans le sens que l'expéditeur ne puisse pas prétendre qu'il n'a pas envoyé le message ou qu'il a envoyé un autre message.

offre un certain niveau de sécurité. On parle de

— **sécurité calculatoire** (ou algorithmique) si la partie adverse (sans connaissance de la clé) ne peut pas déchiffrer le code avec la technologie actuelle dans un temps de calcul raisonnable. Par exemple, le chiffrement asymétrique basé sur la difficulté (conjecturée) de factoriser un grand entier composé en facteurs premiers est encore considéré comme algorithmiquement sûr pour des opérations courantes (voir la légende du tableau 10.1).

n	$\mathcal{O}(\exp(n))$	$\mathcal{O}(\exp(n^{1/3}(\log n)^{2/3}))$	$\mathcal{O}(n^3)$
100	1.26×10^{21}s $= 4.01 \times 10^{13}$a	3.13s $= 9.93 \times 10^{-8}$a	1×10^{-3}s $= 3.17 \times 10^{-11}$a
500	3.27×10^{141}s $= 1.31 \times 10^{134}$a	6.74×10^{10}s $=$ 2139a	0.125s $= 3.96 \times 10^{-9}$a
1000	1.07×10^{292}s $= 3.39 \times 10^{284}$a	6.42×10^{17}s $= 2.03 \times 10^{10}$a	1s $= 3.17 \times 10^{-8}$a

TABLE 10.1 – Une estimation grossière de l'ordre de grandeur du temps nécessaire pour factoriser un entier à n bits (avec $n = 100, 500, 1000$), sous l'hypothèse d'exécution du programme de factorisation sur un ordinateur hypothétique faisant une opération par nanoseconde, comme fonction de la complexité temporelle de l'algorithme. Lorsque le protocole RSA a été proposée [59] en 1978, le meilleur algorithme avait une complexité temporelle de $\mathcal{O}(\exp(n))$. De nos jours, le meilleur algorithme [45] a une complexité $\mathcal{O}(\exp(n^{1/3}\log^{2/3} n))$. Si un ordinateur quantique se construit, l'algorithme de Shor [66] a une complexité $\mathcal{O}(n^3)$. Pour mémoire : « âge de l'univers » $= 1.377 \times 10^{10}$ a.

— **sécurité inconditionnelle (ou informationnelle)** si l'interception du cryptogramme ne nous apprend rien sur le message qui l'a produit. Par conséquent, un code offre une sécurité inconditionnelle si une partie adverse, même possédant une puissance de calcul illimitée, ne pourra pas le casser.
Le terme de sécurité inconditionnelle a été introduit par Diffie et Hellman d'où est tirée la définition suivante [20, page 646] :

> Unconditionally secure systems [. . .] belong to that portion of information theory, called the Shannon theory, which is concerned with optimal performance obtainable with unlimited computation. *Unconditional security results from the existence of multiple meaningful solutions to a cryptogram.* [. . .] A computationally secure cryptogram, in contrast, contains sufficient information to uniquely determine the plaintext and the key. Its security resides solely in the cost of computing them. The only unconditionally secure system in common use is the one-time pad, in which the plaintext is combined with a randomly chosen key of the same length.

Comme on verra dans les paragraphes suivants 10.2 et 10.3, la mise en place d'un code inconditionnellement sûr exige

— la génération d'une suite de même longueur que le message à coder de nombres *purement* aléatoires ,
— l'utilisation de la suite une seule fois ou *one-time pad* et
— la transmission sûre de la clé de cryptage.

Durant plus d'un siècle, ces trois exigences étaient impossibles à réaliser dans la pratique. C'est pourquoi, les méthodes de cryptographie symétrique à clé privée ont été abandonnées au profit du cryptage asymétrique à clé publique (RSA par exemple). L'avènement de la technologie quantique en cryptographie change la donne car aujourd'hui on dispose des méthodes rapides, fiables et efficaces remplissant ces exigences.

Des solutions quantiques sont déjà proposées dans un stade pré-industriel et à un coût raisonnable (cf. cours [57]).

Dans ce chapitre, nous nous intéressons uniquement aux cryptage et à l'authentification offrant une sécurité informationnelle (inconditionnelle).

10.2 Le chiffrement comme code

Un système cryptographique est défini [64] comme une famille paramétrique $(T_k)_{k\in\mathbb{K}}$ de transformations $T_k : \mathbb{M} \to \mathbb{Y}$, indexée par une famille $\mathbb{K}$ des clés de cryptage k. Pour chaque k fixé, la fonction T_k agit sur un ensemble fini $\mathbb{M}$ de messages m pour produire un cryptogramme $y \in \mathbb{Y}$. Restreinte sur les $y \in T_k(\mathbb{M})$, l'application $T_k : \mathbb{M} \to T_k(\mathbb{M})$ est supposée inversible. Dans tous les systèmes cryptologiques, il est supposé qu'il est très difficile de calculer l'inverse si la clé est inconnue.

On peut donc décrire les opérations de cryptage et de décryptage comme des opérations de codage et de décodage vues aux chapitres précédents. Plus précisément :

Définition 10.2.1. Soient $\mathbb{M}$ un ensemble fini de messages et $\mathbb{K}$ un ensemble fini de clés. Un **code de chiffrement** est une application

$$C : \mathbb{M} \times \mathbb{K} \to \mathbb{Y},$$

où $\mathbb{Y}$ est un ensemble de cryptogrammes. Un **code de déchiffrement** est une règle de décision, i.e. une application partielle Δ — définie sur une partie de $\mathbb{Y}$ (le domaine de Δ) —

$$\Delta : \mathbb{Y} \times \mathbb{K} \to \mathbb{M},$$

telle que $\Delta(C(m,k),k) = m$ pour tout message $m \in \mathbb{M}$ et toute clé $k \in \mathbb{K}$.

10.2.1 Code de Vernam *(one-time pad)*

Comme mentionné dans l'article de revue de Bellovin ([3]), en 1882, Frank Miller [3] avait proposé une méthode de chiffrement, appelée *one-time pad*, qui permettait de chiffrer des messages en les combinant avec des clés de même longueur. Le 13 septembre 1918, Gilbert Vernam dépose aux États-Unis une demande de brevet pour un dispositif, appelé « Secret signaling system » de chiffrement, selon la méthode de one-time-pad. La demande est acceptée et Vernam devient détenteur du brevet US Patent 1310719 le 22 juillet 1919 [4].

La méthode introduite par Vernam porte aujourd'hui le nom de **code de Vernam** et sa description peut être trouvée dans [71]. La méthode est censée réponde aux exigences de Kerchoffs.

Soit $\mathbb{A}$ un alphabet fini, *identifié* au groupe additif $\mathbb{Z}_{|\mathbb{A}|} := \{0,\dots,|\mathbb{A}|-1\}$ des entiers modulo $|\mathbb{A}|$. Les messages que nous enverrons constituent une partie spécifique $\mathbb{M} \subseteq \mathbb{X} := \mathbb{A}^+$, i.e. ils sont des mots finis sur cet alphabet. Les fréquences d'apparition de lettres sont arbitraires ; $\mathbb{M}$ peut par conséquent correspondre à une partie d'un langage naturel, du français par exemple.

3. L'auteur de ces notes n'a pas pu consulter le texte original de la contribution de Frank Miller, *Telegraphic code to insure privacy and secrecy in the transmission of telegrams.* C.M. Cornwell (1882). L'information qu'il dispose sur cet article est celle rapportée dans [3].

4. Par la suite il améliore son invention en accédant à d'autres brevets US Patent 1416765, US Patent 1584749 et US Patent 1613686.

Supposons que l'on veuille chiffrer le message $\mathbf{m} \in \mathbb{M}$, avec $N := |\mathbf{m}|$. On suppose que nous disposons d'une clé aléatoire $\mathbf{K} \in \mathbb{K} := \mathbb{A}^N$, distribuée selon la loi uniforme sur $\mathbb{K}$, i.e. $\mathbb{P}(\mathbf{K} = \mathbf{k}) = \frac{1}{|\mathbb{K}|} = \frac{1}{|\mathbb{A}|^N}$. Le chiffrement se fait par addition lettre par lettre modulo $|\mathbb{A}|$ des lettres de $\mathbf{m}$ et de $\mathbf{k}$, i.e. $C(\mathbf{m}, \mathbf{k}) = \mathbf{y}$, avec $y_i = m_i + k_i \mod |\mathbb{A}|$. Plus précisément, on a

Algorithme 10.2.2. Chiffrement de Vernam

Require: Message initial $\mathbf{m} \in \mathbb{M}$ et UNIF($\mathbb{A}$).
Ensure: Clé de chiffrement $\mathbf{k} \in \mathbb{A}^+$ et message chiffré $\mathbf{y} \in \mathbb{A}^+$, avec $|\mathbf{k}| = |\mathbf{y}| = |\mathbf{m}|$.
  $N \leftarrow |\mathbf{m}|$.
  for $i \in \{1, \dots, N\}$ **do**
    Générer selon UNIF($\mathbb{A}$) la i^{e} lettre $k_i \in \mathbb{A}$ de la clé de cryptage.
    $y_i \leftarrow m_i + k_i \mod |\mathbb{A}|$.
  end for
  $\mathbf{k} \leftarrow k_1 \cdots k_N$.
  $\mathbf{y} \leftarrow y_1 \cdots y_N$.

On constate que l'algorithme de Vernam génère une clé aléatoire de même longueur que le message à chiffrer et que cette clé est utilisée pour ce seul message. Le récipiendaire du message chiffré $\mathbf{y}$, s'il connaît la clé $\mathbf{k}$, exécute l'algorithme suivant :

Algorithme 10.2.3. Déchiffrement de Vernam

Require: Message chiffré $\mathbf{y} \in \mathbb{A}^+$ clé $\mathbf{k} \in \mathbb{A}^+$ avec $|\mathbf{y}| = |\mathbf{k}|$.
Ensure: Message déchiffré $\mathbf{m}' \in \mathbb{A}^{|\mathbf{y}|}$.
  $N \leftarrow |\mathbf{y}|$.
  for $i \in \{1, \dots, N\}$ **do**
    $m'_i = y_i - k_i \mod |\mathbb{A}|$.
  end for
  $\mathbf{m}' = m'_1 \cdots m'_N$.

Lemme 10.2.4. *Si $\mathbf{m}'$ est obtenu par l'application de l'algorithme 10.2.3 sur le cryptgramme $\mathbf{y}$ produit par l'algorithme 10.2.2, alors*

$$\mathbb{P}(\mathbf{M}' = \mathbf{m}' | \mathbf{M} = \mathbf{m}) = \begin{cases} 1 & \text{si } \mathbf{m}' = \mathbf{m} \\ 0 & \text{sinon.} \end{cases}$$

Théorème 10.2.5 (Shannon [64])**.** *Si la clé est*

1. *de longueur $N = |\mathbf{m}|$,*
2. *utilisée une seule fois et*
3. *connue uniquement des deux partenaires légitimes,*

alors l'algorithme de Vernam (10.2.2 et 10.2.3) est parfaitement sûr, i.e. il offre une sécurité inconditionnelle.

Démonstration: Les variables aléatoires $\mathbf{M} \in \mathbb{M}$, $\mathbf{K} \in \mathbb{K}$ et $\mathbf{Y} \in \mathbb{A}^N$ représentent respectivement le message d'entrée, la clé et le message chiffré. Pour des $\mathbf{m}, \mathbf{k}$ fixés, on

définit $\mathbf{y} := T_{\mathbf{k}}(\mathbf{m}) = \mathbf{m} \oplus \mathbf{k}$, où $\oplus$ désigne l'addition lettre par lettre modulo $|\mathbb{A}|$. Cette transformation induit un noyau stochastique déterministe

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{K} = \mathbf{k}, \mathbf{M} = \mathbf{m}) := P_{\mathbf{k}}(\mathbf{m}, \mathbf{y}) = \begin{cases} 1 & \text{si } \mathbf{y} = T_{\mathbf{k}}(\mathbf{m}) = \mathbf{m} \oplus \mathbf{k} \\ 0 & \text{sinon.} \end{cases}$$

Il est évident, par la construction de la clé, que $\mathbb{P}(\mathbf{K} = \mathbf{k}, \mathbf{M} = \mathbf{m}) = \mathbb{P}(\mathbf{M} = \mathbf{m})\mathbb{P}(\mathbf{K} = \mathbf{k})$, car le choix du mot aléatoire $\mathbf{K}$ ne dépend pas de $\mathbf{M}$. En outre,

$$\begin{aligned} \mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{M} = \mathbf{m}) &= \frac{\mathbb{P}(\mathbf{M} = \mathbf{m}, \mathbf{Y} = \mathbf{y})}{\mathbb{P}(\mathbf{M} = \mathbf{m})} = \sum_{\mathbf{k} \in \mathbb{K}} \frac{\mathbb{P}(\mathbf{M} = \mathbf{m}, \mathbf{K} = \mathbf{k}, \mathbf{Y} = \mathbf{y})}{\mathbb{P}(\mathbf{M} = \mathbf{m})} \\ &= \sum_{\mathbf{k} \in \mathbb{K}} \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{M} = \mathbf{m}, \mathbf{K} = \mathbf{k})}{\mathbb{P}(\mathbf{M} = \mathbf{m})} \mathbb{P}(\mathbf{M} = \mathbf{m}, \mathbf{K} = \mathbf{k}) \\ &= \sum_{\mathbf{k} \in \mathbb{K}} \mathbb{P}(\mathbf{K} = \mathbf{k}) \delta_{\mathbf{y}, \mathbf{m} \oplus \mathbf{k}} = \mathbb{P}(\mathbf{K} = \mathbf{y} \ominus \mathbf{m}) = \frac{1}{|\mathbb{A}|^N}, \forall \mathbf{m} \in \mathbb{M}. \end{aligned}$$

Le cryptogramme est donc équidistribué sur $\mathbb{A}^N$, indépendamment du message. On calcule de même

$$\begin{aligned} \mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{Y} = \mathbf{y}) &= \frac{\mathbb{P}(\mathbf{m} = \mathbf{m}, \mathbf{Y} = \mathbf{y})}{\mathbb{P}(\mathbf{Y} = \mathbf{y})} = \sum_{\mathbf{k} \in \mathbb{K}} \frac{\mathbb{P}(\mathbf{m} = \mathbf{m}, \mathbf{K} = \mathbf{k}, \mathbf{Y} = \mathbf{y})}{\mathbb{P}(\mathbf{Y} = \mathbf{y})} \\ &= \sum_{\mathbf{k} \in \mathbb{K}} \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{M} = \mathbf{m}, \mathbf{K} = \mathbf{k})}{\mathbb{P}(\mathbf{Y} = \mathbf{y})} \mathbb{P}(\mathbf{M} = \mathbf{m}, \mathbf{K} = \mathbf{k}) \\ &= \sum_{\mathbf{k} \in \mathbb{K}} \frac{\mathbb{P}(\mathbf{K} = \mathbf{k})\mathbb{P}(\mathbf{M} = \mathbf{m})}{\mathbb{P}(\mathbf{Y} = \mathbf{y})} \delta_{\mathbf{y}, \mathbf{m} \oplus \mathbf{k}} = \frac{\mathbb{P}(\mathbf{K} = \mathbf{y} \ominus \mathbf{m})}{\mathbb{P}(\mathbf{Y} = \mathbf{y})} \mathbb{P}(\mathbf{M} = \mathbf{m}) \\ &= \mathbb{P}(\mathbf{M} = \mathbf{m}). \end{aligned}$$

On conclut que $\mathbf{Y}$ et $\mathbf{M}$ sont indépendants. De ce fait découle la sécurité inconditionnelle du code de Vernam. □

Plusieurs formulations (équivalentes) de la sécurité inconditionnelle peuvent être données :

— Le message est indépendant du cryptogramme, i.e. $\mathbb{P}(\mathbf{M} = \mathbf{m} | \mathbf{Y} = \mathbf{y}) = \mathbb{P}(\mathbf{M} = \mathbf{m})$;
— L'information mutuelle entre message et cryptogramme s'annule, i.e. $I(\mathbf{M} : \mathbf{Y}) = 0$;
— Tous le messages ont la même probabilité de produire un même cryptogramme donné, i.e. pour tous $\mathbf{k} \in \mathbb{K}$ et $\mathbf{m}, \mathbf{m}' \in \mathbb{M}$,

$$\mathbb{P}(C(\mathbf{M}, \mathbf{K}) = \mathbf{y} | \mathbf{M} = \mathbf{m}, \mathbf{K} = \mathbf{k}) = \mathbb{P}(C(\mathbf{M}, \mathbf{K}) = \mathbf{y} | \mathbf{M} = \mathbf{m}', \mathbf{K} = \mathbf{k}).$$

Exemple 10.2.6. Le résultat précédent implique qu'en absence de connaissance de la clé, le seul moyen d'attaquer le code de Vernam est la recherche exhaustive des clés dans $\mathbb{K}$. Maintenant, supposons que le message initial est

$$\mathbf{m} = \texttt{wewonthebattlebutwedefinitelylostthewar}$$

et que par une clé $\mathbf{k}$ il est codé en $\mathbf{y}$. Un intrus qui ne connaît pas la clé $\mathbf{k}$ doit examiner toutes les clés $\mathbf{k}' \in \mathbb{K}$ (étant donné que $\mathbf{m} = 39$, il en existe $26^{39} = 1.53 \times 10^{55}$ telles

clés) et calculer tous les message $\mathbf{m}' = \mathbf{y} \ominus \mathbf{k}'$. La plupart de messages ainsi obtenus n'auront aucun sens. Mais il en existe qui ont un sens totalement différent, par exemple parmi ces messages il trouvera bien sûr $\mathbf{m}$ mais aussi

$$\mathbf{m}' = \texttt{overwhelminglyvictoriousovertheevilaxis}$$

avec $|\mathbf{m}| = |\mathbf{m}'| = 39$. Ce sont précisément
— l'uniformisation de la probabilité sur l'espace des message et
— la dégénérescence de la règle de décision
qui confèrent au code de Vernam son indéchiffrabilité.

10.2.2 Le rôle essentiel des nombres aléatoires

Insistons encore une fois sur l'importance d'un générateur qui permet de produire des clés longues (aussi longues que le message à coder) et vraiment aléatoires. Il s'agit d'une tache difficile — *stricto sensu* impossible, comme nous verrons dans le cours de « Complexité » [58] — en algorithmique classique.

La quête d'algorithmes générant de nombres pseudo-aléatoires qui miment, autant que faire se peut, les propriétés des nombres vraiment aléatoires est même source de manipulations malveillantes. Rappelons en effet, quelques faits troublants.

En 2007, le *National Institute of Standards and Technology (NIST)* [5] a émis une recommandation [6] où il préconisait l'utilisation de 4 générateurs de nombres aléatoires. Parmi ces quatre générateurs, le *dual elliptic curve deterministic random bit generator (Dual_EC_DRBG)* était floué. Malgré les alertes de la communauté scientifique internationale, le NIST, pas seulement continuait sa préconisation mais militait activement pour que cet algorithme devienne un standard ISO [7]. Au lieu de cryptographie, on est en présence d'une **affaire de ... cleptographie.**

Après que l'affaire des révélations d'Edward Snowden éclate (en 2013), le NIST est obligé d'admettre que l'algorithme est floué et publie, en 2015, une recommandation révisée [8] d'où il retire l'algorithme incriminé. L'*American Mathematical Society (AMS)* initie la publication dans ses *Notices* d'une série d'articles sur le thème « Mathematicians discuss the Snowden revelation ». Dans cette série, apparaît un article [9] par Michael Wertheimer. Notons que l'AMS se sent obligée d'apposer la précision « At the time of the writing of this piece Michael Wertheimer was the Director of Research at the NSA ; he recently retired from that position » sur cet article.

10.3 Authentification

L'**authentification** d'un message consiste à fournir au destinataire légitime l'évidence que le message reçu émane bien de l'expéditeur, même en présence d'un adversaire qui peut
— envoyer au destinataire des messages frauduleux de sa propre facture (non émis par l'expéditeur), — on parle alors d'**usurpation d'identité** (*impersonation* en anglais) — et/ou

5. Organisme dépendant du gouvernement des États-Unis dont les missions sont décrites à ce lien.
6. NIST Special Publication 800-90 : Recommendation for Random Number Generation Using Deterministic Random Bit Generators (2007).
7. Organisation internationale de normalisation, reconnue par 162 pays dont les missions sont décrites sur ce lien.
8. NIST Special Publication 800-90A Revision 1: Recommendation for Random Number Generation Using Deterministic Random Bit Generators (2015).
9. Intitulé « Encryption and the NSA role in international standards ».

— intercepter les messages expédiés en les substituant par des messages frauduleux générés par lui — on parle alors de **substitution**.

Dans ce paragraphe nous nous intéressons à une authentification informationnelle (inconditionnelle) [10] c'est-à-dire, nous supposons que la partie adverse connaît tous les algorithmes utilisés par les partenaires légitimes ; elle ignore uniquement la clé secrète qu'ils ont utilisée. Nous suivons les exposés [68, 67, 49]. Contrairement au cas de cryptage, où le théorème de Shannon fournit comme définition du parfait chiffrement l'indépendance entre les messages émis et chiffré, la situation est plus subtile dans le cas de l'authentification. On ne connaît pas de définition d'authentification parfaite : on peut rendre la **probabilité de fraude** (*deceit probability* en anglais) arbitrairement petite — en utilisant une clé d'authentification suffisamment longue — sans jamais l'annuler.

10.3.1 Illustration du problème et notation

On considère deux partenaires A et B qui partagent une clé sécrète aléatoire $K \in \mathbb{K}$, où $\mathbb{K}$ est un ensemble fini.

1. A veut envoyer un message X choisi dans un ensemble fini $\mathbb{X}$ (typiquement $\mathbb{X} = \mathbb{A}^m$ pour un certain alphabet fini $\mathbb{A}$ et m un entier strictement positif).
2. Pour le message X, A génère une clé $K \in \mathbb{K}$ ($\mathbb{K}$ un ensemble fini de clés) pour l'authentifier. Les variables X et K sont supposées distribuées selon une loi conjointe $\kappa(x,k) = \mathbb{P}(X = x, K = k)$, avec $x \in \mathbb{X}$ et $k \in \mathbb{K}$.
3. Pour authentifier le message (et éventuellement le crypter) A dispose d'une famille — indexée par $\mathbb{K}$ — de noyaux stochastiques (déterministes ou aléatoires) $M := (M_k)_{k\in\mathbb{K}}$ de $\mathbb{X}$ à un autre ensemble $\mathbb{Y}$ (typiquement $\mathbb{Y} = \mathbb{A}^n$ avec $n > m$) et s'en sert pour préparer un message [11] $Y \in \mathbb{Y}$ destiné à être envoyé à B, i.e.

$$\mathbb{P}(Y = y | X = x; K = k) = M_k(x,y), x \in \mathbb{X}, y \in \mathbb{Y}, k \in \mathbb{K}.$$

Il s'ensuit que la loi conjointe des variables X, K, Y est

$$\begin{aligned}\mathbb{P}(X = x, K = k, Y = y) &= \mathbb{P}(Y = y | X = x, K = k)\mathbb{P}(X = x, K = k)\\ &= \kappa(x,k)M_k(x,y), x \in \mathbb{X}, k \in \mathbb{K}, y \in \mathbb{Y}.\end{aligned}$$

Les lois marginales se calculent de manière élémentaire :

$$\begin{aligned}\mu(k,y) = \mathbb{P}(K = k, Y = y) &= \sum_{x\in\mathbb{X}} \mathbb{P}(X = x, K = k, Y = y)\\ &= \sum_{x\in\mathbb{X}} \kappa(x,k)M_k(x,y);\\ \rho(k) = \mathbb{P}(K = k) &= \sum_{x\in\mathbb{X}} \kappa(x,k);\\ \nu(y) = \mathbb{P}(Y = y) &= \sum_{x\in\mathbb{X}}\sum_{k\in\mathbb{K}} \kappa(x,k)M_k(x,y).\end{aligned}$$

10.3.2 Minoration de la probabilité de fraude (cas de substitution)

La partie adverse observe Y et tente de deviner la clé K. Si elle parvient, elle pourra coder n'importe quel message $X \in \mathbb{X}$ de manière authentifiable par B et donc le substi-

10. Par opposition à une authentification algorithmique (calculatoire).
11. Ce message peut être le cryptogramme de X concaténé avec son sceau d'identification.

tuer avec succés au message Y initialement envoyé à B. En résumé, l'adversaire tente d'envoyer un message $\tilde{Y} \in \mathbb{Y}$ qui serait

— interprété par B comme un message authentique et

— décodé comme un message particulier voulu par l'adversaire.

L'adversaire réussira son coup s'il devine correctement la clé car alors il pourra coder n'importe quel message et le faire admettre comme authentique. La probabilité moyenne de réussite pour l'adversaire est donnée, pour une clé k, par $\sum_{y\in\mathbb{Y}} \nu(y)\,\mathbb{P}(K = k|Y = y)$. Pour se prémunir dans tous les scénarios, on doit considérer le maximum sur les clés, donnée par la quantité

$$\overline{\beta}^{\text{sub}}_{\max} := \sum_{y\in\mathbb{Y}} \nu(y) \max_{\ell\in\mathbb{K}} \mathbb{P}(K = \ell|Y = y).$$

Théorème 10.3.1. *La probabilité moyenne $\beta := \overline{\beta}^{sub}_{\max}$ que l'adversaire réussisse sa tentative de fraude par substitution est*

$$\beta \geq 2^{-H(K|Y)}.$$

Démonstration. Nous avons

$$\begin{aligned}
H(K|Y) &= \sum_{y\in\mathbb{Y}} H(K|Y = y)\nu(y) \\
&= \sum_{k\in\mathbb{K},y\in\mathbb{Y}} \nu(y)\mathbb{P}(K = k|Y = y)\,(-\log(\mathbb{P}(K = k|Y = y))) \\
&\geq -\sum_{k\in\mathbb{K},y\in\mathbb{Y}} \mu(k,y)\log(\max_{\ell\in\mathbb{K}} \mathbb{P}(K = \ell|Y = y)) \\
&= -\sum_{k\in\mathbb{K},y\in\mathbb{Y}} \nu(y)\log(\max_{\ell\in\mathbb{K}} \mathbb{P}(K = \ell|Y = y)) \\
&\geq -\log\left(\sum_{y\in\mathbb{Y}} \nu(y) \max_{\ell\in\mathbb{K}} \mathbb{P}(K = \ell|Y = y)\right) \quad \text{par l'inégalité de Jensen.}
\end{aligned}$$

□

10.3.3 Minoration de la probabilité de fraude (cas d'usurpation d'identité)

La partie adverse — qui tente d'usurper l'identité de A mais ignore la clé utilisée — génère un message $\tilde{Y} \in \mathbb{Y}$ selon une loi (arbitraire) $\tilde{q}$, i.e. $\mathbb{P}(\tilde{Y} = y) = \tilde{q}(y)$ et l'envoie de manière intempestive à B.

Le problème mathématique se pose donc comme un problème de test d'hypothèses statistiques ; on note H_0 l'hypothèse nulle signifiant que le message reçu par B est un message authentique et H_1 l'hypothèse alternative signifiant que le message reçu est frauduleux.

Pour décider quel est le cas qui prévaut, le partenaire B dispose d'une famille — indexée par $\mathbb{K}$ — de règles de décision $(\Delta_k)_{k\in\mathbb{K}}$, i.e. des applications $\Delta_k : \mathbb{Y} \to \mathbb{D} := \{0, 1\}$, où $\mathbb{D}$ désigne l'espace de décisions ; les valeurs 0 ou 1 prises par Δ_k signifient respectivement que H_0 ou H_1 sont acceptées, i.e. que le message est respectivement jugé « authentique » ou « frauduleux ». Pour chaque k, la règle de décision permet donc de

partitioner l'ensemble $\mathbb{Y}$ en deux sous-ensembles disjoints $\mathbb{Y}_k(0) := \{y \in \mathbb{Y} : \Delta_k(y) = 0\}$ et $\mathbb{Y}_k(1) := \{y \in \mathbb{Y} : \Delta_k(y) = 1\}$.

Le partenaire B reçoit donc un message

$$\hat{Y} = \begin{cases} Y & \text{si l'hypothèse } H_0 \text{ prévaut,} \\ \tilde{Y} & \text{si l'hypothèse } H_1 \text{ prévaut} \end{cases}$$

et doit se servir de sa règle de décision pour départager les deux cas. Il est alors évident que, sur l'évènement $\{\hat{Y} = y\}$,

$$\mathbb{P}(K = k, \Delta_k(y) = 1 | H_0) = \mu(k, y),$$
$$\mathbb{P}(K = k, \Delta_k(y) = 0 | H_1) = \rho(k)\tilde{q}(y).$$

Sous les conditions et notations du paragraphe précédent, on définit

$$\alpha = \sum_{z \in \mathbb{Y}_k(1)} \mu(k, z),$$
$$\beta = \sum_{z \in \mathbb{Y}_k(0)} \rho(k)\tilde{q}(z).$$

On voit que α représente l'erreur de première espèce (juger le message non authentique tandis qu'il l'est) et β l'erreur de deuxième espèce (juger le message authentique tandis qu'il ne l'est pas) ; β représente donc la probabilité que l'adversaire réussisse son coup.

On introduit deux vecteurs de probabilité $\mathbf{p}_0$ et $\mathbf{p}_1$ sur l'espace des décisions $\mathbb{D}$, définis par

$$\mathbf{p}_0 = (1 - \alpha, \alpha) \text{ et } \mathbf{p}_1 = (\beta, 1 - \beta),$$

où α, β désignent les erreurs de type I et II respectivement. Par ailleurs, on introduit deux vecteurs de probabilité $\mathbf{q}_0$ et $\mathbf{q}_1$ sur $\mathbb{Y}$, définis pour chaque $k \in \mathbb{K}$ fixé, par

$$\mathbf{q}_0 = \mu(k, \cdot) \text{ et } \mathbf{q}_1 = \rho(k)\tilde{q}(\cdot).$$

Lemme 10.3.2. *Les espaces probabilisés $(\mathbb{Y}, \mathbf{q}_0)$ et $(\mathbb{Y}, \mathbf{q}_1)$ sont des fragmentations (cf. définition 7.3.5) des espaces probabilisés $(\mathbb{D}, \mathbf{p}_0)$ et $(\mathbb{D}, \mathbf{p}_1)$.*

Démonstration. Par simple vérification. □

Théorème 10.3.3. *Sous les conventions et avec les notations précédentes, on définit*

$$d(\alpha, \beta) := D((1 - \alpha, \alpha) \| (\beta, 1 - \beta)) = (1 - \alpha) \log \frac{1 - \alpha}{\beta} + \alpha \log \frac{\alpha}{1 - \beta}.$$

Nous avons alors

$$d(\alpha, \beta) \leq D(\mu \| \rho \otimes \tilde{q}).$$

En particulier, $\beta \geq 2^{-I(K:Y)}$.

Démonstration. La majoration $D((1 - \alpha, \alpha) \| (\beta, 1 - \beta)) \leq D(\mu \| \rho \otimes \tilde{q})$ est une conséquence directe du lemme 10.3.2 et de la proposition 7.3.6 (qui établit que la fragmentation augmente le contraste). Cette majoration est valable pour tout vecteur de probabilité $\tilde{q}$ sur $\mathbb{Y}$. On peut donc optimiser le second membre pour obtenir

$$d(\alpha, \beta) \leq \inf_{q' \in \mathcal{M}_1(\mathbb{Y})} D(\mu \| \rho \otimes q') \leq D(\mu \| \rho \otimes \nu),$$

car ν n'est pas nécessairement optimal. Or, $D(\mu\|\rho\otimes\nu) = I(K:Y)$, car

$$\begin{aligned} I(K:Y) &= H(K) + H(Y) - H(K,Y) \\ &= -\sum_{k\in\mathbb{K}} \rho(k)\log\rho(k) - \sum_{y\in\mathbb{Y}} \nu(y)\log\nu(y) + \sum_{(k,y)\in\mathbb{K}\times\mathbb{Y}} \mu(k,y)\log\mu(k,y) \\ &= \sum_{(k,y)\in\mathbb{K}\times\mathbb{Y}} \mu(k,y)\log\frac{\mu(k,y)}{\rho(k)\nu(y)} \\ &= D(\mu\|\rho\otimes\nu). \end{aligned}$$

En particulier, $d(0,\beta) = -\log\beta \leq I(K:Y)$, par conséquent la probabilité d'usurpation d'identité sera minorée par

$$\beta \geq 2^{-I(K:Y)}.$$

□

Remarque 10.3.4. Nous constatons que si K et Y sont de variables aléatoires indépendantes, les probabilités de fraude (par substitution ou par usurpation d'identité) sont égales à 1. Contrairement donc au cryptage, où l'indépendance entre K et Y est la garante de l'indéchiffrabilité du code, ici un certain degré de dépendance est nécessaire rien que pour obtenir une minoration non-triviale de la probabilité de fraude. Il faut encore montrer que cette minoration est essentiellement atteinte, ce qui a été prouvé dans [72] à l'aide des fonctions de hachage universelles.

10.4 Qu'est-ce la cryptographie post-quantique ?

Nous avons déjà évoqué dans ce cours les notions de cryptographie quantique et de calcul quantique. Récemment un nouveau terme est apparu : la cryptographie post-quantique dont la sémantique ne reflète pas très clairement son objet d'étude. Une clarification est donc nécessaire car tous ces termes mélangent des aspects algorithmiques et des aspects de réalisation physique de dispositifs de communication. En nous limitant à la partie algorithmique nous distinguons :

Le calcul quantique : appelé *quantum computing* en anglais, englobe les algorithmes qui exploitent les spécificités de la physique quantique pour réduire la complexité algorithmique du problème.

La cryptographie quantique : décrit l'ensemble de protocoles qui exploitent les spécificités de la physique quantique pour communiquer en offrant un chiffrement inconditionnel de l'information. Ces protocoles et leurs implémentation physique sont basés sur la distribution inviolable d'une clé privée.

La cryptographie post-quantique : désigne des algorithmes cryptographiques classiques à clé publique dont la complexité algorithmique ne permettrait pas leur déchiffrement (en un temps raisonnable) par l'avènement éventuel d'un ordinateur quantique. Elle ne doit pas être confondue avec la cryptographie quantique qui est une cryptographie exploitant les phénomènes quantiques pour parvenir à ses fins.

Récemment, une compétition internationale [12] est lancée par le NIST sur la recherche algorithmique post-quantique « pour solliciter, évaluer et standardiser un ou plusieurs

12. *To solicit, evaluate, and standardize one or more quantum-resistant public-key cryptographic algorithms.* https://csrc.nist.gov/projects/post-quantum-cryptography

algorithmes cryptographiques à clé publique, incassables par des ordinateurs quantiques ».

10.5 Exercices

11 Codes correcteurs d'erreur

Or rappelle qu'un système de communication est décrit par la figure 9.3. Le théorème fondamental de transmission 9.5.4 garantit l'existence d'un codage permettant la communication sans erreur à travers un canal $\mathcal{K} = (\mathbb{X}, \mathbb{Y}, P)$ pourvu que le taux de transmission $R < \mathsf{cap}(P) = \sup_{\mu \in \mathfrak{M}_1(\mathbb{X})} I(X : Y)$, où μ est la loi de la variable X à l'entrée du canal. Ce résultat s'obtient en imposant une rarefaction des messages à transmettre, i.e. $|\mathbb{M}| < |\mathbb{X}^n|$, qui a comme conséquence d'utiliser un codage redondant pour chaque message. Cependant, ce résultat est existentiel. Dans ce chapitre nous allons construire des codes effectifs dans le cas où $\mathbb{M} \simeq \mathbb{X}^l$, avec $l < n$. Une telle raréfaction induit un taux de transmission

$$R = \frac{\log |\mathbb{X}^l|}{\log |\mathbb{X}^n|} = \frac{l}{n}.$$

En contrepartie, la redondance qu'elle impose permet la correction des erreurs commises.

11.1 Structure algébrique des codes

Très souvent dans les applications, les alphabets d'entrée et de sortie du canal vérifient $\mathbb{X} = \mathbb{Y} = \{0, 1\}$. Dans ce cas, l'alphabet $\{0, 1\}$, muni de l'addition modulo 2 — notée $+$ — coïncide avec le groupe abélien $\mathbb{Z}_2$ des congruences modulo 2 ; il se trouve que le groupe $(\mathbb{Z}_2, +)$ muni d'une multiplication est aussi un corps. Munir les alphabets d'une structure algébrique présente plusieurs avantages comme on le verra dans ce chapitre. En particulier, cela permet d'identifier les mots sur l'alphabet à des vecteurs dans un espace vectoriel.

Des alphabets plus généraux sont aussi possibles. Pour tout entier $q = p^e$, où p est premier et $e > 0$, $\mathbb{F}_q$ désigne le corps fini à q éléments. Si $e = 1$, alors, $\mathbb{F}_q = \mathbb{Z}_q$ (i.e. $\mathbb{Z}_q$ est un corps) mais si $e > 1$, alors, le groupe abélien $\mathbb{Z}_q$ n'est qu'un anneau mais il n'est plus un corps car $p^e = 0$ et p est un diviseur de 0.

Pour $p > 1$ premier et $q = p^e$, soit $f(x)$ un polynôme primitif à coefficients dans $\mathbb{Z}_p$, i.e. un polynôme $f(x) = \sum_{i=0}^{e-1} a_i x^i$ de degré e avec $(a_i)_{i=0,\dots,e-1}$ une famille d'éléments de $\mathbb{Z}_p$ tel que f ait une racine α qui engendre tout le corps $\mathbb{F}_q$, i.e. $\mathbb{F}_q := \{0, 1, \alpha, \alpha^2, \dots, \alpha^{q-2}\}$.

Exemple 11.1.1. Puisque $q = 4 = 2^2 = p^2$, le corps $\mathbb{F}_4$ ne coïncide pas avec $\mathbb{Z}_4$. Le polynôme $f(x) = x^2 + x + 1$ n'a pas de racines dans $\mathbb{Z}_2$ car $f(0) = f(1) = 1$. Si α est une racine de f, en adjoignant les éléments de la forme $a + b\alpha$, $a, b \in \mathbb{Z}_2$, on voit que le corps fini $\mathbb{F}_4$ a 4 éléments : $\mathbb{F}_4 = \{0, 1, \alpha, \alpha^2\}$. On vérifie immédiatement que $\alpha^2 = 1 + \alpha$, i.e. α et $1 + \alpha$ sont l'inverse l'un de l'autre.

Il faut cependant garder à l'esprit que tous les alphabets $\mathbb{X}$ (à un nombre $|\mathbb{X}|$ arbitraire d'éléments) ne peuvent pas être mis en correspondance avec des corps finis. Pour cela, il faut que $|\mathbb{X}| = p^e$ avec p premier et $e \in \mathbb{N}$. Nous avons en effet le

Théorème 11.1.2. *Soit $\mathbb{F}$ un corps fini. Alors sa caractéristique*[1] *est un premier p et $\mathbb{F}$ peut être considéré comme un espace vectoriel sur $\mathbb{Z}_p$. La dimension $e = \dim_{\mathbb{Z}_p} \mathbb{F}$ de cet espace vectoriel est le degré de l'extension $(\mathbb{F} : \mathbb{Z}_p)$, i.e. $|\mathbb{F}| = p^e$.*

Démonstration. Voir, par exemple, [70, Théorème 12.15, p. 224]. □

L'identification de l'alphabet $\mathbb{X}$ avec un corps fini $\mathbb{F}_q$ s'étend à une identification entre l'ensemble des mots de longueur n et le $\mathbb{F}_q$-espace vectoriel de dimension n. Plus précisément, il existe deux bijections $\mathbf{V} : \mathbb{X}^n \to \mathbb{F}_q^n$ et $\mathbf{W} : \mathbb{F}_q^n \to \mathbb{X}^n$ ($\mathbf{V}$ pour « vecteur » et $\mathbf{W}$ pour « word ») définies par

$$\mathbb{X}^n \ni \xi = x_1 \cdots x_n \mapsto \mathbf{V}(\xi) = \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{F}_q^n, \quad x_1, \dots, x_n \in \mathbb{F}_q$$

et

$$\mathbb{F}_q^n \ni \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \mapsto \mathbf{W}(\mathbf{x}) = \xi = x_1 \cdots x_n \in \mathbb{X}^n$$

qui sont l'inverse l'une de l'autre : $\mathbf{W}^{-1} = \mathbf{V}$ et $\mathbf{V}^{-1} = \mathbf{W}$.

11.2 Structure géométrique des codes

11.2.1 Métrisation de Hamming

Définition 11.2.1. Soit $\mathbb{X}$ un alphabet fini avec $q = |\mathbb{X}|$ puissance d'un premier, identifié avec le corps fini $\mathbb{F}_q$. Pour $n \in \mathbb{N}_>$, on identifie l'ensemble de mots $\mathbb{X}^n$ avec le $\mathbb{F}_q$-espace vectoriel $\mathbb{F}_q^n$. Pour deux mots $\xi = x_1 \cdots x_n$ et $\xi' = x'_1 \cdots x'_n$ on définit leur **distance de Hamming**[2] par

$$\mathrm{d}(\xi, \xi') := \mathrm{d}_H(\xi, \xi') = \sum_{i=1}^{n} (1 - \delta_{x_i, x'_i}).$$

1. On rappelle que la caractéristique d'un corps unifère est le plus petit entier n tel $n \cdot 1 = 0$.
2. Vérifier que d_H est bien une distance.

Cette distance induit une distance, aussi notée d_H, sur $\mathbb{F}_q^n$ par $\mathrm{d}_H(\mathbf{x},\mathbf{x}') := \mathrm{d}_H(\mathbf{W}(\mathbf{x}),\mathbf{W}(\mathbf{x}'))$. L'espace $(\mathbb{X}^n,\mathrm{d}_H)$ devient un espace métrique isomorphe à l'espace vectoriel métrisé $(\mathbb{F}_q^n,\mathrm{d}_H)$.

Le **poids de Hamming**, $w(\xi)$, d'un mot $\xi \in \mathbb{X}^n$ est défini par

$$w(\xi) = \mathrm{d}_H(\xi, 0^n) = \mathrm{d}_H(\mathbf{V}(\xi),\mathbf{0}).$$

Exemple 11.2.2. Pour $q = 2$, on identifie $\mathbb{X} = \mathbb{F}_2 = \{0,1\}$. Pour $n = 3$, on calcule $\mathrm{d}_H(000,111) = \mathrm{d}_H(\mathbf{0},\mathbf{1}) = 3 = w(111)$ (à ne pas confondre donc avec la distance euclidienne $\mathrm{d}(\mathbf{0},\mathbf{1}) = \sqrt{3}$).

Dans tout ce qui suit, les alphabets d'entrée et de sortie du canal sont identiques, isomorphes à $\mathbb{F}_q$ ($\mathbb{X} = \mathbb{Y} \simeq \mathbb{F}_q$), avec q puissance d'un premier (le plus souvent $q = 2$ et $\mathbb{X} = \mathbb{Y} = \{0,1\}$). S'il n'y a pas d'ambiguïté, nous ne ferons plus la distinction entre ces espaces. Un **code** est une application $C : \mathbb{X}^l \to \mathbb{X}^n$ avec $n \geq l$.

Définition 11.2.3.

1. Un code $C : \mathbb{X}^l \to \mathbb{X}^n$ est dit **linéaire** si C est une application linéaire. Comme tout code peut être identifié à son glossaire $\mathcal{G} := \mathcal{G}(C) = C(\mathbb{X}^l)$, un code linéaire est identifié au sous-espace vectoriel $C(\mathbb{F}_q^l)$ de $\mathbb{F}_q^n$.
2. La **dimension** du code linéaire C est la dimension du sous-espace vectoriel engendré par son glossaire $k = \dim C(\mathbb{F}_q^l)$.
3. La **distance minimale**, d, du code est la distance de Hamming minimale entre deux mots distincts du glossaire

$$d = \min_{\substack{\mathbf{x},\mathbf{x}'\in\mathcal{G}\\ \mathbf{x}\neq\mathbf{x}'}} \mathrm{d}(\mathbf{x},\mathbf{x}').$$

4. On note (n,q^l), ou plus précisément (n,q^l,d), le code linéaire sur l'alphabet $\mathbb{F}_q$ agissant sur des blocs de taille l, produisant des blocs de taille n et à distance minimale d. Une autre notation[3], qui privilégie la dimension du glossaire $k = \dim \mathcal{G}$, est $[n,k]$, ou plus précisément $[n,k,d]$.
5. Si des bases sont fixées dans $\mathbb{F}_q^l$ et $\mathbb{F}_q^n$, le code linéaire $C : \mathbb{X}^l \to \mathbb{X}^n$ est représenté par une matrice de $\mathbb{M}_{l,n}(\mathbb{F}_q)$ dépendante des bases choisies.

11.2.2 Matrice génératrice

On cherche une écriture systématique pour la matrice représentant le code. Soit $[n,k,d]$ un code linéaire C et G la matrice qui le représente. Si $k < l$, tous les vecteurs de $\mathbb{F}_q^l$ ne donneront pas d'images indépendantes. On peut donc se limiter à représenter C par une matrice G de taille $k \times n$ correspondant à une application linéaire injective $\mathbb{F}_q^k \to \mathcal{G}(C)$ définie par

$$\mathbf{x}^t G = \mathbf{y}^t,$$

où $(\cdot)^t$ désigne le vecteur transposé, $\mathbf{x} \in \mathbb{F}_q^k$ et $\mathbf{y} \in \mathbb{F}_q^n$. Cette matrice est appelée **génératrice** du code.

Dans la suite, on admettra, sans perte de généralité, que $l = k$.

3. Noter la différence de notation (n,q^l,d) et $[n,k,d]$.

Proposition 11.2.4. *Soit C un $[n,k]$-code linéaire sur $\mathbb{F}_q$ et $(\mathbf{e}_1,\dots,\mathbf{e}_k)$ une base arbitraire du glossaire $\mathcal{G}$. La matrice génératrice de C est*

$$G = \begin{bmatrix} \mathbf{e}_1^t \\ \vdots \\ \mathbf{e}_k^t \end{bmatrix} \in \mathfrak{M}_{k,n}.$$

On peut par des opérations sur les lignes de G, i.e. par un changement de base, la ramener à sa **forme systématique**

$$G := G(C) = [I_k|A],$$

où I_k est la matrice identité $k \times k$ et $A \in \mathfrak{M}_{k,n-k}(\mathbb{F}_q)$.

Exemple 11.2.5. Soit $\mathbb{X} = \{0,1\}$ (donc $q = |\mathbb{X}| = 2$).

1. Le $[n,1,n]$ code linéaire dont la matrice génératrice est $G = [1|\underbrace{1\cdots 1}_{n-1}]$ est le **code à répétition** $\mathcal{R}_n$. Le glossaire du code est $\mathcal{G} := \mathcal{G}(\mathcal{R}_n) = \{0^n, 1^n\}$.
2. Le $[k+1,k,k-1]$ code linéaire dont la matrice génératrice (sous sa forme systématique) est donnée par $G = [I_k|A]$ avec

$$A = \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathfrak{M}_{k,1},$$

est le **code de test de parité** $\mathcal{P}_n$ avec $n = k+1$. Par exemple si $k = 2$, les entrées possibles sont $\{00,01,10,11\}$ et le glossaire du code $\mathcal{P}_3$ est $\mathcal{G} := \mathcal{G}(\mathcal{P}_3) = \{000,011,101,110\}$ avec $d = 2$.
3. Le **code de Hamming** $\textsc{Ham}(7,4)$ est un code linéaire $[7,4,3]$ qui a comme matrice génératrice (sous sa forme systématique) la matrice $G = [I_4|A]$, avec

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \in \mathfrak{M}_{4,3}.$$

La matrice A effectue un test de parité sur les sous-ensembles différents à 3 (des 4) bits d'entrée, i.e. le glossaire est

$$\mathcal{G} := \mathcal{G}(\textsc{Ham}(7,4)) = \{\mathbf{y} \in \mathbb{F}_2^7 : \mathbf{x}^t G = \mathbf{y}^t, \mathbf{x} \in \mathbb{F}_2^4\}$$

et les mots qui le composent vérifient $y_i = x_i$, pour $i = 1,\dots,4$, et

$$\begin{aligned} y_5 &= x_1 + x_2 + x_3 \\ y_6 &= x_2 + x_3 + x_4 \\ y_7 &= x_1 + x_3 + x_4. \end{aligned}$$

11.2.3 Matrice de contrôle de parité

Dans les exemples précédents, nous avons examiné de $[n,k]$-codes avec k et n petits. Il était alors simple de construire explicitement le glossaire du code. Lorsque k et n sont grands, il n'est pas efficace de générer d'emblée tout le glossaire du code. Il est donc intéressant de disposer d'un algorithme efficace pour vérifier si un vecteur donné de $\mathbb{F}_q^n$ fait partie du glossaire du code.

Définition 11.2.6. Soit C un $[n,k]$-code sur un corps $\mathbb{F}$. Une matrice $H := H(C)$ à n colonnes sur $\mathbb{F}$ est dite **matrice de contrôle de parité** si

$$\mathbf{g} \in \mathcal{G}(C) \iff H\mathbf{g} = \mathbf{0},$$

i.e. $\mathcal{G}(C) = \ker H(C)$.

Remarque 11.2.7. La notation est particulièrement malencontreuse (mais standard) comme H désigne aussi la fonction entropie. La signification sera cependant claire par le contexte ; à défaut, elle sera explicitement précisée.

Le nombre de lignes de H n'est pas précisé dans la définition précédente. Cependant, lorsque $l = k = \operatorname{rg} G(C) = \dim \mathcal{G}$, alors nous avons la

Proposition 11.2.8. *Soit C un $[n,k]$-code et $G = [I_k|A] \in \mathfrak{M}_{k,n}(\mathbb{F})$ sa matrice génératrice sous sa forme systématique, avec $A \in \mathfrak{M}_{k,n-k}(\mathbb{F})$. Alors la forme systématique de la matrice de contrôle de parité est*

$$H = [-A^t|I_{n-k}] \in \mathfrak{M}_{n-k,n}(\mathbb{F}).$$

Démonstration. $\mathbf{g} \in \mathcal{G}(C)$ si, et seulement si, il existe $\mathbf{x} \in \mathbb{F}^k$ tel que $\mathbf{x}^t G = \mathbf{g}^t$; par conséquent, $\mathbf{g} = G^t\mathbf{x}$. On aura alors $H\mathbf{g} = HG^t\mathbf{x} = 0$, pour tout $\mathbf{x} \in \mathbb{F}^k$. Ceci entraîne que $HG^t = 0$. Or

$$HG^t = [-A^t|I_{n-k}] \left[\begin{array}{c} I_k \\ \hline A^t \end{array}\right] = -A^tI_k + I_{n-k}A^t = -A^t + A^t = 0.$$

□

Exemple 11.2.9. Pour des alphabets d'entrée et de sortie binaires, identifiés à $\mathbb{F}_2$, la matrices H de test de parité

— pour le code à répétition $\mathcal{R}_n$ est

$$H(\mathcal{R}_n) = [B|I_{n-1}] \in \mathfrak{M}_{n-1,n}, \text{ avec } B = \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathfrak{M}_{n-1,1} = G(\mathcal{P}_n);$$

— pour le code de test de parité $\mathcal{P}_n$ est

$$H(\mathcal{P}_n) = G(\mathcal{R}_n);$$

— pour le code $\textsc{Ham}(7,4)$ est

$$H(\textsc{Ham}(7,4)) = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

Définition 11.2.10. Soit C un $[n,k]$ code sur $\mathbb{F}_q$ dont le glossaire est $\mathcal{G}$. Le **code dual** est défini par son glossaire

$$\mathcal{G}^\perp = \{\mathbf{x} \in \mathbb{F}_q^n : \mathbf{x}\cdot\mathbf{g} = \mathbf{x}^t\mathbf{g} := \sum_{j=0}^{n-1} x_j g_j \mod q = 0, \forall \mathbf{g} \in \mathcal{G}\},$$

i.e. le code dual $C^\perp$ de C est le $[n, n-k]$-code dont le glossaire $\mathcal{G}(G^\perp) = \mathcal{G}^\perp(G)$ est l'orthogonal du glossaire du code C.

11.3 Décodage

11.3.1 Maximum de vraisemblance

Le problème que nous voulons résoudre est illustré par l'exemple suivant.

Exemple 11.3.1. Un message $\mathbf{m} \in \mathbb{F}_2^k$ est codé en un message $\mathbf{x} := C(\mathbf{m}) \in \mathbb{F}_2^n$, avec $n \geq k$. Le message $\mathbf{x}$ est ensuite transmis à travers un canal symétrique binaire qui change avec probabilité $p < 1/2$ un bit en son opposé. Le message transmis est alors une variable aléatoire $\mathbf{Y}$ à valeurs dans $\mathbb{F}_2^n$ de loi conditionnelle

$$\begin{aligned} Q_n(\mathbf{x},\mathbf{y}) &= \mathbb{P}(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x}) = p^{\mathrm{d}(\mathbf{x},\mathbf{y})}(1-p)^{n-\mathrm{d}(\mathbf{x},\mathbf{y})} \\ &= (1-p)^n \left(\frac{p}{1-p}\right)^{\mathrm{d}(\mathbf{x},\mathbf{y})}. \end{aligned}$$

— Peut-on détecter qu'un message transmis $\mathbf{y}$ contient des erreurs ?
— Si oui, peut-on choisir un message $\hat{\mathbf{y}}$ qui corrige l(es) erreur(s) de transmission ?

Le théorème de Neyman-Pearson (théorème 6.2.16, page 73) établit l'optimalité de la règle de décision du maximum de vraisemblance ; l'exercice 102 (page 144) donne un exemple concret de l'application de ce résultat au problème de décodage. En applicant ce résultat ici, on voit que le message qui estime « le mieux » le message transmis est donné par la règle du maximum de vraisemblance $\hat{\mathbf{y}} = \Delta_{MV}(\mathbf{x})$, où

$$\begin{aligned} \Delta(\mathbf{x}) &:= \arg\max_{\mathbf{z}\in\mathbb{F}_2^n} Q_n(\mathbf{x},\mathbf{z}) \\ &= \arg\max_{\mathbf{z}\in\mathbb{F}_2^n} \left(\frac{p}{1-p}\right)^{\mathrm{d}(\mathbf{x},\mathbf{z})} \\ &= \arg\min_{\mathbf{z}\in\mathbb{F}_2^n} \mathrm{d}(\mathbf{x},\mathbf{z}) \in \mathbb{F}_2^n \ (\text{car } \ p<1/2). \end{aligned}$$

La règle probabiliste du maximum de vraisemblance est donc équivalente à un résultat purement géométrique de minimisation de la distance de Hamming entre des mots du code appelé **décodage par le plus proche voisin**.

Définition 11.3.2. Soient $\mathbf{x} \in \mathbb{F}_q^n$ (pour q puissance d'un premier) et $r > 0$. La **boule de Hamming**[4] est donnée par

$$B_{n,r}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{F}_q^n : \mathrm{d}(\mathbf{x},\mathbf{y}) \leq r\} \subseteq \mathbb{F}_q^n.$$

4. Dans la littérature, cette boule est désignée par le **terme erroné** de *sphère* de Hamming !

Une boule de Hamming est centrée sur chaque mot $\mathbf{g} \in \mathcal{G}$ du glossaire. Tout $\mathbf{y} \in B_{n,r}(\mathbf{g})$ sera donc décodé par l'estimation $\hat{\mathbf{y}} = \arg\min_{\mathbf{z}\in\mathbb{F}_2^n} \mathrm{d}(\mathbf{g},\mathbf{z}) = \mathbf{g} \in \mathcal{G}$. On voit donc immédiatement les contraintes d'un bon décodage par maximum de vraisemblance :

1. Le rayon r de la boule doit être aussi grand que possible pour que le maximum des mots de $\mathbb{F}_q^n$ se trouvent dans une boule de Hamming.
2. Le rayon r doit être suffisamment petit pour que les boules de la famille $\{B_{n,r}(\mathbf{g}), \mathbf{g} \in \mathcal{G}\}$ soient disjointes de sorte que chaque mot dans la boule puisse être décodé sans ambiguïté. Cette dernière condition sera vérifiée dès que $r < \frac{1}{2}d$, où d est la distance minimale du code.

Exemple 11.3.3. Soit $C = \mathcal{R}_3$ le code à répétition sur $\mathbb{F}_2$. Il s'agit d'un $[3,1,3]$-code. L'estimateur du plus proche voisin corrigera une erreur si $r \leq 1$. Le glossaire de ce code est $\mathcal{G} = \{000, 111\}$ et les boules de Hamming de rayon 1 sont

$$B_{3,1}(000) = \{000, 001, 010, 100\} \text{ et } B_{3,1}(111) = \{111, 011, 101, 110\}.$$

On constate que $B_{3,1}(000) \sqcup B_{3,1}(111) = \mathbb{F}_2^3$. Tout mot de $\mathbb{F}_2^3$ peut donc être décodé sans ambiguïté et le code corrige toute erreur qui modifie la valeur d'un bit.

Lemme 11.3.4. *Un code linéaire C sur le corps $\mathbb{F}$, avec distance minimale d, permet de corriger jusqu'à $t = \lfloor \frac{d-1}{2} \rfloor$ erreurs.*

Démonstration. Supposons que $d \geq 2t+1$, que le mot $\mathbf{g} \in \mathcal{G}$ est envoyé et qu'un bruit additif $\mathbf{e}$ corrompt le message, i.e. le message reçu est $\mathbf{y} = \mathbf{g} + \mathbf{e} \in \mathbb{F}^n$. Supposons que le poids de Hamming du bruit vérifie $w(\mathbf{e}) \leq t$, donc

$$\mathrm{d}(\mathbf{g}, \mathbf{y}) = \mathrm{d}(\mathbf{0}, \mathbf{y} - \mathbf{g}) = w(\mathbf{e}) \leq t.$$

Pour tout $\mathbf{g}' \in \mathcal{G}$, avec $\mathbf{g}' \neq \mathbf{g}$, on a

$$2t+1 \leq d \leq \mathrm{d}(\mathbf{g}, \mathbf{g}') \leq \mathrm{d}(\mathbf{g}, \mathbf{y}) + \mathrm{d}(\mathbf{y}, \mathbf{g}').$$

On aura donc,

$$\mathrm{d}(\mathbf{y}, \mathbf{g}') \geq \mathrm{d}(\mathbf{g}, \mathbf{g}') - \mathrm{d}(\mathbf{g}, \mathbf{y}) \geq 2t+1-t = t+1 > \mathrm{d}(\mathbf{g}, \mathbf{y}).$$

Le mot $\mathbf{y}$ sera donc décodé par l'estimation $\hat{\mathbf{y}} = \Delta(\mathbf{y}) = \mathbf{g}$ qui est le mot effectivement envoyé (avec toutes les erreurs de transmission corrigées).

Réciproquement, supposons que $d < 2t+1$. Il s'ensuit que $d \leq 2t$. Par définition de d, il existe deux mots différents $\mathbf{g}$ et $\mathbf{g}'$ du glossaire différent exactement sur d positions, i.e. $d = \mathrm{d}(\mathbf{g}, \mathbf{g}')$. Il existe alors $\mathbf{x} \in \mathbb{F}^n$ avec $\mathrm{d}(\mathbf{g}, \mathbf{x}) \leq t$ et $\mathrm{d}(\mathbf{g}', \mathbf{x}) \leq t$. Par exemple, en modifiant $\mathbf{g}$ en $\lfloor \frac{d}{2} \rfloor$ positions, on obtient un tel $\mathbf{x}$. Or l'estimation $\hat{\mathbf{x}}$ sera ambiguë car tant $\mathbf{g}$ que $\mathbf{g}'$ sont à distance inférieure à d de $\mathbf{x}$. □

Pour un $[n,k]$-code sur $\mathbb{F}_q$, les k qits sont appelés **qits d'information** tandis que les $n-k$ qits restants, **qits de contrôle**.

Théorème 11.3.5 (Empilement de boules de Hamming). *Soit C un $[n,k,d]$-code linéaire sur $\mathbb{F}_q$ permettant de corriger t erreurs ($t \leq \lfloor \frac{d-1}{2} \rfloor$). Alors,*

$$V_q(n,t) := \sum_{r=0}^{t} C_n^r (q-1)^r \leq q^{n-k}.$$

Démonstration. Pour chaque mot du glossaire, $\mathbf{g} \in \mathcal{G}$, il existe une boule de Hamming, $B_{n,t}(\mathbf{g})$, centrée sur $\mathbf{g}$ et de rayon t. On a

$$\begin{aligned} |B_{n,t}(\mathbf{g})| &= |\{\mathbf{x} \in \mathbb{F}_q^n : \mathrm{d}(\mathbf{x}, \mathbf{g}) \leq t\}| \\ &= |\sqcup_{r=0}^{t} \{\mathbf{x} \in \mathbb{F}_q^n : \mathrm{d}(\mathbf{x}, \mathbf{g}) = r\}| \\ &= \sum_{r=0}^{t} |\{\mathbf{x} \in \mathbb{F}_q^n : \mathrm{d}(\mathbf{x}, \mathbf{g}) = r\}|. \end{aligned}$$

Or, pour chaque $\mathbf{g}$, les mots dans $\{\mathbf{x} \in \mathbb{F}_q^n : \mathrm{d}(\mathbf{x}, \mathbf{g}) = r\}$ diffèrent de $\mathbf{g}$ en exactement r indices. Si i est un tel indice, x_i peut prendre n'importe quel valeur $x_i \in \mathbb{F}_q \setminus \{g_i\}$. Par conséquent, le volume $V_q(n, r) = |\{\mathbf{x} \in \mathbb{F}_q^n : \mathrm{d}(\mathbf{x}, \mathbf{g}) = r\}| = C_n^r (q-1)^r$.

On a donc

$$|B_{n,t}(\mathbf{g})| = \sum_{r=0}^{t} C_n^r (q-1)^r.$$

Or, $\dim \mathcal{G} = k$, donc il existe q^k mots différents dans $\mathcal{G}$. Les boules $B_{n,t}(\mathbf{g})$ centrées sur les mots du glossaire sont disjointes car leur rayon est inférieur à $d/2$. Comme elles sont toutes contenues dans $\mathbb{F}_q^n$, on a de manière évidente,

$$\left| \bigsqcup_{\mathbf{g} \in \mathcal{G}} B_{n,t}(\mathbf{g}) \right| = \sum_{\mathbf{g} \in \mathcal{G}} |B_{n,t}(\mathbf{g})| = q^k \sum_{r=0}^{t} C_n^r (q-1)^r \leq q^n.$$

□

Ce dernier résultat peut-être lu dans l'autre sens.

Corollaire 11.3.6. *Soit C code sur $\mathbb{F}_q$ ayant un glossaire $\mathcal{G}$ de dimension k (i.e. k qits d'information). Pour que le code puisse corriger t erreurs, il faut qu'il dispose de $n-k$ qits de contrôle, où*

$$n - k \geq \log_q \left(\sum_{r=0}^{t} C_n^r (q-1)^r \right).$$

Corollaire 11.3.7. *Soit C un $[n,k]$-code sur $\mathbb{F}_2$ permettant de corriger t erreurs. Asymptotiquement, lorsque $\lim_{n,t\to\infty}$, avec $\lim_{n\to\infty} \frac{t}{n} = x \in [0,1]$, le taux de transmission $R = \lim_{n\to\infty} k/n$ du code vérifie*

$$1 - R \geq H_2(x),$$

où $H_2(x) = -x \log_2 x - (1-x)\log_2(1-x)$ est l'entropie binaire.

Démonstration. La borne d'empilement, établie en théorème 11.3.5, garantit que

$$2^n \geq 2^k \sum_{r=0}^{t} C_n^r \geq 2^k C_n^t.$$

On conclut en utilisant l'approximation de Stirling pour le factoriel : $n! \asymp (n/e)^n \sqrt{2\pi n}$.

□

Les deux corollaires précédents, montrent qu'afin de maximiser le taux de transmission, il faut choisir $M = |\mathcal{G}|$ aussi grand que possible pourvu que la borne d'empilement de Hamming soit vérifiée ; ils établissent une borne supérieure sur $M = q^k$. Plus précisément, on s'intéresse à un (n, M)-code sur $\mathbb{F}_q$ de glossaire $\mathcal{G}$, avec $M := M_n = |\mathcal{G}|$ et distance minimale $d := d_n = \mathrm{d}(\mathcal{G})$. On note

$$A_q(n, d) := \max\{|\mathcal{G}| : \mathcal{G} \subseteq \mathbb{F}_q^n, \mathrm{d}(\mathcal{G}) = d\}.$$

Le théorème 11.3.5 montre que

$$A_q(n, d) \left(\sum_{r=0}^{t} C_n^r (q-1)^r \right) \leq q^n,$$

où $t = \lfloor (d-1)/2 \rfloor$. La borne de Gilbert-Varhsamov, établie en théorème 11.3.8, donne une borne inférieure.

Théorème 11.3.8 (Borne de Gilbert-Varshamov). *Soient $q \geq 2$ et $n \geq d \geq 1$. Alors*

$$A_q(n, d) \left(\sum_{r=0}^{d-1} C_n^r (q-1)^r \right) \geq q^n.$$

Démonstration. Parmi tous le codes ayant q, n et d fixés, soit $\mathcal{G}$ le (glossaire du) code qui maximise la taille du glossaire, i.e. $M = |\mathcal{G}| = A_q(n, d)$. Les boules de Hamming $B_{d-1}(\mathbf{g}) = \{\mathbf{x} \in \mathbb{F}_q^n | \mathrm{d}(\mathbf{g}, \mathbf{x}) \leq d-1\}$, où $\mathbf{g} \in \mathcal{G}$, doivent couvrir $\mathbb{F}_q^n$. En effet, si $\mathbf{y} \in \mathbb{F}_q^n$ n'appartient pas à la boule $B_{d-1}(\mathbf{g})$, cela signifie que $\mathrm{d}(\mathbf{g}, \mathbf{y}) \geq d$; si $\mathbf{y}$ n'appartenait à aucune des boules cela signifierait que $\mathrm{d}(\mathbf{y}, \mathbf{g}) \geq d$ pour tout $\mathbf{g} \in \mathcal{G}$. Par conséquent, le code dont le glossaire serait $\mathcal{G}' = \mathcal{G} \cup \{\mathbf{y}\}$ aurait les mêmes paramètres q, n, d mais avec $|\mathcal{G}'| > |\mathcal{G}|$; ceci contredirait la maximalité de $\mathcal{G}$. Maintenant,

$$|B_{d-1}(\mathbf{g})| = \sum_{r=0}^{d-1} C_n^r (q-1)^r,$$

tandis que la réunion (pas nécessairement disjointe) de ces boules couvre $\mathbb{F}_q^n$. □

Dans la plupart de cas, la différence entre les bornes supérieure et inférieure de $A_q(n, d)$ est grande et sa valeur précise est en général difficile à calculer. On peut cependant délimiter une région entre les deux bornes asymptotiques (lorsque $n \to \infty$) en remarquant que si, $\lim_{n\to\infty} \frac{d-1}{n} = \lim_{n\to\infty} \frac{2t}{n} = z \in [0, 1]$, alors, dans le cas $q = 2$, on a l'encadrement

$$1 - H(z/2) \leq R \leq 1 - H(z),\ z \in [0, 1],$$

où H désigne la fonction entropie.

Définition 11.3.9. Un $[n, k, d]$-code C sur $\mathbb{F}_q$ pour lequel la borne du théorème 11.3.5 est saturée, i.e.

$$\sum_{r=0}^{t} C_n^r (q-1)^r = q^{n-k},$$

est dit **parfait**.

Exemple 11.3.10. Sur $\mathbb{F}_2$, les codes $\mathcal{R}_n$ avec $n \in 2\mathbb{N} + 1$ et $\textsc{Ham}(7, 4)$ sont parfaits.

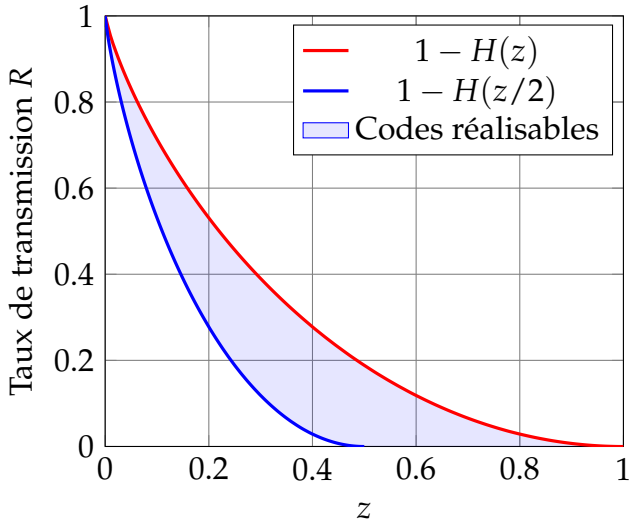


FIGURE 11.1 – Bornes supérieure et inférieure asymptotiques du taux de transmission en fonction du rapport $z = \lim_{n\to\infty} \frac{d}{n}$.

11.3.2 Décodage par partition

On se place de nouveau dans le cas d'un $[n,k]$-code linéaire C sur $\mathbb{F}_q$. Une autre méthode systématique de décodage consiste en la construction d'une partition de l'espace d'arrivée (qui induit une relation d'équivalence) et la construction d'un représentant dominant pour chaque classe d'équivalence. Plus précisément, on note $\mathbb{X} = \mathbb{F}_q^k$, $\mathbb{Y} = \mathbb{F}_q^n$ (avec $n > k$), les espace d'entrée et de sortie du code, $\mathcal{G} = \{\mathbf{g}_0, \dots \mathbf{g}_{K-1}\}$ son glossaire avec $|\mathcal{G}| = K := q^k$ et $L = q^{n-k}$.

Algorithme 11.3.11. Partition de l'espace de codage

Require: $k, n, q, \mathcal{G}$
Ensure: Partition $\mathbb{Y} = \sqcup_{r=0}^{L-1} \mathbb{Y}_r$; suite des représentants dominants $\mathbf{y}_r \in \mathbb{Y}_r$ pour $r = 0, \dots, L-1$
$K \leftarrow q^k$
$L \leftarrow q^{n-k}$
$\mathbb{Y}_0 \leftarrow \mathcal{G}$
$\mathbf{y}_0 \leftarrow \mathbf{0}$
for $r \leftarrow 1, \dots, L-1$ **do**
  $\mathbf{y}_r \leftarrow \arg\min_{\mathbf{z} \in \sqcup_{s=0}^{r-1} \mathbb{Y}_s} w(\mathbf{z})$
  $\mathbb{Y}_r \leftarrow \mathbf{y}_r + \mathcal{G}$
end for

Il est évident que la numérotation des classes n'est pas unique car plusieurs éléments peuvent avoir le même poids. Cependant, pour chaque décomposition fixée, la famille $(\mathbb{Y}_r)_{r=0,\dots,L-1}$ est une partition de $\mathbb{Y}$; il s'ensuit que (pour cet ordre)

$$\forall \mathbf{y} \in \mathbb{Y}, \exists! r \in \{0, \dots, L-1\} \quad \text{et} \quad \exists! s \in \{0, \dots, K-1\} : \mathbf{y} = \mathbf{y}_r + \mathbf{g}_s.$$

On appelle **indice pivot** de la partition, et on le note r_p,

$$r_p = \max\{r \in \{0, \dots, L-1\} : w(\mathbf{y}_r) \leq t = \lfloor \frac{d-1}{2} \rfloor.$$

Lemme 11.3.12. *Soit $(\mathbb{Y}_r)_{r=0,\dots,L-1}$ une partition de $\mathbb{Y}$ pour un $[n,k]$-code linéaire sur $\mathbb{F}_q$ et $\mathcal{G} = \{\mathbf{g}_0, \dots, \mathbf{g}_{K-1}\}$ son glossaire. On note $\mathbf{y}_r$ le représentant dominant de la classe $\mathbb{Y}_r$. Soit $\mathbf{y} \in \mathbb{Y}$.*

1. *Si $\mathbf{y} = \mathbf{y}_r + \mathbf{g}_s$ pour un $r \in \{0, \dots, L-1\}$ et $s \in \{0, \dots, K-1\}$, alors $\mathbf{g}_s$ est à distance au plus $w(\mathbf{y}_r)$ de $\mathbf{y}$ dans $\mathcal{G}$.*
2. *Si $r \leq r_p$, où r_p est l'indice pivot de la partition (i.e. $w(\mathbf{y}_r) \leq t$), alors $\mathbf{g}_s$ est l'unique élément de $\mathcal{G}$ à distance au plus $w(\mathbf{y}_r)$ de $\mathbf{y}$. On pourra donc décoder, sans erreur, $\mathbf{y}$ par $\hat{\mathbf{y}} = \mathbf{g}_s$.*

Démonstration. 1. Supposons que $\mathbf{g}_s$ ne soit pas à distance au plus $w(\mathbf{y}_r)$ de $\mathbf{y}$. Il existe donc un $\mathbf{g}_{s'} \in \mathcal{G}$ tel que $\mathrm{d}(\mathbf{y}, \mathbf{g}_{s'}) < \mathrm{d}(\mathbf{y}, \mathbf{g}_s)$. Puisque pour tout $\mathbf{x}, \mathbf{x}' \in \mathbb{F}_q^n$, nous avons $\mathrm{d}(\mathbf{x}, \mathbf{x}') = w(\mathbf{x} - \mathbf{x}')$, il s'ensuit que $w(\mathbf{y} - \mathbf{g}_{s'}) < w(\mathbf{y} - \mathbf{g}_s) = w(\mathbf{y}_r)$. Par ailleurs, $\mathbf{y} - \mathbf{g}_{s'} = \mathbf{y}_r + (\mathbf{g}_{s'} - \mathbf{g}_s) \in \mathbf{y}_r + \mathcal{G} = \mathbb{Y}_r$. Ceci contredit l'hypothèse que $\mathbf{y}_r$ est un représentant dominant de la classe de congruence $\mathbb{Y}_r$.

2. Maintenant $r < r_p$, i.e. $w(\mathbf{y}_r) \leq t$. Supposons qu'il existe $\mathbf{g}_{s'} \in \mathcal{G}$ tel que $\mathrm{d}(\mathbf{y}, \mathbf{g}_{s'}) < \mathrm{d}(\mathbf{y}, \mathbf{g}_s)$. On aura alors

$$\begin{aligned} \mathrm{d}(\mathbf{g}_s, \mathbf{g}_{s'}) &\geq d \\ &> 2t \geq 2\mathrm{d}(\mathbf{y}, \mathbf{g}_s) = \mathrm{d}(\mathbf{y}, \mathbf{g}_s) + \mathrm{d}(\mathbf{y}, \mathbf{g}_s) \\ &\geq \mathrm{d}(\mathbf{y}, \mathbf{g}_s) + \mathrm{d}(\mathbf{y}, \mathbf{g}_{s'}) \ (\text{ car } \mathrm{d}(\mathbf{y}, \mathbf{g}_{s'}) < \mathrm{d}(\mathbf{y}, \mathbf{g}_s)) \\ &\geq \mathrm{d}(\mathbf{g}_s, \mathbf{g}_{s'}) \ (\text{par l'inégalité triangulaire}). \end{aligned}$$

Ceci mène à l'inégalité impossible $\mathrm{d}(\mathbf{g}_s, \mathbf{g}_{s'}) > \mathrm{d}(\mathbf{g}_s, \mathbf{g}_{s'})$.

□

Exemple 11.3.13. (Partition de l'espace de codage pour $\mathcal{R}_4$ sur $\mathbb{F}_2$). Il s'agit d'un $[4,1,4]$-code sur $\{0,1\}$ qui permet de corriger $t \leq \lfloor \frac{d-1}{2} \rfloor = 1$ erreur. Ici $K = 2^4 = 16$ et $L = 2^3 = 8$.

r	$\mathbf{y}_r$	$w(\mathbf{y}_r)$	$\mathbb{Y}_r$
0	0000	0	$\{0000, 1111\}$
1	0001	1	$\{0001, 1110\}$
2	0010	1	$\{0010, 1101\}$
3	0100	1	$\{0100, 1011\}$
4	1000	1	$\{1000, 0111\}$
5	0011	2	$\{0011, 1100\}$
6	0101	2	$\{0101, 1010\}$
7	0110	2	$\{0110, 1001\}$

TABLE 11.1 – Par abus de notation, on écrit $\mathbf{y} = \xi$ au lieu de $\mathbf{y} = \mathbf{V}(\xi)$. Pour $\xi \in \{0,1\}^4$, le trait discontinu sépare les classes dont le représentant dominant a un poids inférieur ou égal à t des autres. Le dernier indice avant le trait est dit **indice pivot**, noté r_p ; dans cet exemple $r_p = 4$. Supposons que 1111 est transmis mais le bruit le corrompt en 0111. En cherchant dans le tableau précédent, on constate que $\mathbf{y} = 0111 \in \mathbb{Y}_4$ et $\mathbf{y} = \mathbf{y}_4 + \mathbf{g}_1$. Par conséquent, $\mathbf{g}_1 = 1111$ est à distance $1 = w(\mathbf{y}_4)$ de $\mathbf{y}$ dans $\mathcal{G}$. Comme $4 \leq r_p$, 1111 est l'unique élément de $\mathcal{G}$ à distance 1 de $\mathbf{y}$ et on le décode sans erreur $\Delta(0111) = 1111$. Si $\mathbf{y} = 0101$ est reçu, on constate que $\mathbf{y} \in \mathbb{Y}_6$. Alors 0000 est à distance 2 de $\mathbf{y}$ dans $\mathcal{G}$; mais comme $6 > r_p$, rien ne garantit que 0000 est l'unique élément de $\mathcal{G}$ à distance 2 de 0101. En effet 1111 l'est aussi.

11.3.3 Décodage par syndrome

Cette méthode est basée sur la partition de l'espace de codage introduite au §11.3.2 en systématisant la recherche de la classe de congruence dans laquelle appartient chaque message $\mathbf{y} \in \mathbb{F}^n$ reçu. Pour cela on calcule un indicateur associé à $\mathbf{y}$, son **syndrome**[5] $s(\mathbf{y})$. Le calcul du syndrome permet d'abréger l'algorithme de décodage.

Définition 11.3.14. Soient C un $[n,k]$-code linéaire sur $\mathbb{F}_q$, $\mathcal{G}$ son glossaire et H sa matrice de contrôle de parité. Pour tout $\mathbf{y} \in \mathbb{Y} = \mathbb{F}_q^n$, on appelle **syndrome** de $\mathbf{y}$, le vecteur

$$\mathbf{s} := \mathbf{s}(\mathbf{y}) = H\mathbf{y} \in \mathbb{F}_q^{n-k}.$$

De toute évidence, $\mathbf{s}(\mathbf{g}) = \mathbf{0}$ pour tout $\mathbf{g} \in \mathcal{G}$.

Lemme 11.3.15. *Soient C un $[n,k]$-code linéaire sur $\mathbb{F}_q$ et $\mathcal{G}$ son glossaire. Deux vecteurs $\mathbf{y}, \mathbf{y}' \in \mathbb{Y} = \mathbb{F}_q^n$ sont dans la même classe de congruence par rapport à $\mathcal{G}$ si, et seulement si, il ont le même syndrome.*

Démonstration.

$$\mathbf{y} + \mathcal{G} = \mathbf{y}' + \mathcal{G} \iff \mathbf{y} - \mathbf{y}' \in \mathcal{G} \iff \mathbf{s}(\mathbf{y}) - \mathbf{s}(\mathbf{y}') = 0 \iff \mathbf{s}(\mathbf{y}) = \mathbf{s}(\mathbf{y}').$$

□

Corollaire 11.3.16. *Le syndrome d'un vecteur $\mathbf{y} \in \mathbb{Y}$ reçu ne dépend pas de $\mathbf{y}$ mais uniquement du bruit additif qui l'a corrompu.*

Démonstration. Le vecteur reçu est obtenu par $\mathbf{y} = \mathbf{x}G + \mathbf{b}$, où $\mathbf{b}$ est le bruit additif. Or $\mathbf{x}G \in \mathcal{G}$; par conséquent $\mathbf{s}(\mathbf{y}) = H(\mathbf{g} + \mathbf{b}) = H\mathbf{b}$, car $\mathbf{g} = \mathbf{x}G \in \mathcal{G}$. □

Ce résultat permet de systématiser le décodage comme le montre l'algorithme 11.3.17 ci-dessous.

Algorithme 11.3.17. Décodage par syndrome

Require: $k, n, q, \mathcal{G}$, matrice H du code, partition $(\mathbb{Y}_r)$, suite de représentants dominants $(\mathbf{y}_r)$, message à décoder $\mathbf{y}$.
Ensure: $\hat{\mathbf{y}} = \Delta(\mathbf{y})$.

```
s ← Hy
r ← 0
s_r ← Hy_r
while s_r ≠ s do
    r ← r + 1
    s_r ← Hy_r
end while
ŷ ← y − y_r
```

Exemple 11.3.18. (Décodage par syndrome de $\mathcal{R}_4$ sur $\mathbb{F}_2$). On utilise le tableau introduit dans l'exemple 11.3.13, auquel on ajoute une colonne avec le calcul du syndrome.

5. SYNDROME, subst. masc. (Méd. Pathol.) Ensemble de signes, de symptômes, de modifications morphologiques, fonctionnelles ou biochimiques de l'organisme, d'apparence parfois disparate mais formant une entité reconnaissable qui, sans présager obligatoirement des causes de ces manifestations, permettent d'orienter le diagnostic. *Trésor de la langue française*, version en ligne.

r	$\mathbf{y}_r$	$w(\mathbf{y}_r)$	$\mathbf{s}(\mathbf{y}_2)$	$\mathbb{Y}_r$
0	0000	0	000	$\{0000, 1111\}$
1	0001	1	001	$\{0001, 1110\}$
2	0010	1	010	$\{0010, 1101\}$
3	0100	1	100	$\{0100, 1011\}$
4	1000	1	111	$\{1000, 0111\}$
5	0011	2	011	$\{0011, 1100\}$
6	0101	2	101	$\{0101, 1010\}$
7	0110	2	110	$\{0110, 1001\}$

TABLE 11.2 – Supposons que 1111 est transmis mais le bruit le corrompt en 0111 ayant $\mathbf{s}(0111) = 111$ correspondant à la classe $r = 4$. On calcule alors $\Delta(0111) = 0111 - 1000 = 1111$.

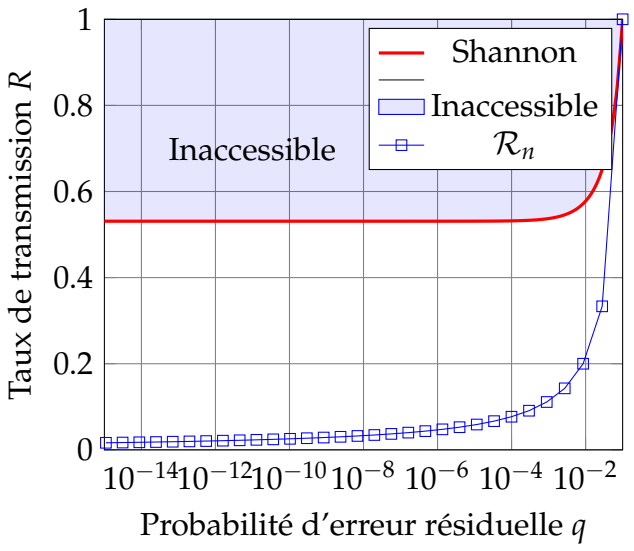


FIGURE 11.2 – Pour un canal symétrique binaire, avec taux d'erreur $p = 0.1$, la ligne rouge représente la frontière de Shannon établie en théorème 9.5.4, délimitant la région inaccessible de la région où des codes existent. Les points bleus représentent la suite des taux de transmission en fonction de la probabilité d'erreur résiduelle pour la famille de codes à répétition $\mathcal{R}_n$, avec $n = 1, 3, 5, \ldots, 61$.

11.4 Exercices

Codes à répétition

107. Dans cet exercice, on considère un canal binaire symétrique ayant une probabilité d'erreur p. On note $\mathcal{R}_n$ le code à $n \in \mathbb{N}_>$ répétitions pour chaque bit et $p_b(\mathcal{R}_n, p)$ la probabilité de décoder de manière erronée un bit par la méthode de vote majoritaire.

    (a) Calculer $p_b(\mathcal{R}_3, 0.1)$.
    (b) Calculer $p_b(R_n, p)$, pour $n \in \mathbb{N}_>$.
    (c) Pour $p = 0.1$ quel est le terme dominant dans l'expression de $p_b(\mathcal{R}_n, p)$.
    (d) À partir de quelle valeur de n, on obtient une valeur de $p_b(\mathcal{R}_n, 0.1) \leq 10^{-15}$?
    (e) Quel est le taux de transmission pour $\mathcal{R}_n$?
    (f) Placer les couples $(\mathcal{R}_n, \log p_b(\mathcal{R}_n, p))$ pour $n \in \{1, 3, 5 \ldots, 61\}$ sur un gra-

phique. Qu'observez-vous ?

108. Soit $\mathcal{R}_n$ un code à répétition sur $\mathbb{F}_q$. Montrer que
   (a) si $q = 2$ et n est impair alors $\mathcal{R}_n$ est parfait,
   (b) si $q > 2$ ou n est pair, alors $\mathcal{R}_n$ est imparfait.

Propriétés géométriques des codes

109. Soit C un code de glossaire $\mathcal{G}$ et de distance minimale d. Montrer que $d = \min\{w(\mathbf{g}), \mathbf{g} \in \mathcal{G}, \mathbf{g} \neq \mathbf{0}\}$.
110. Soit une matrice $J \in \mathfrak{M}_{n,n}(\{-1,1\})$ — vue comme l'empilement de n vecteurs lignes $\mathbf{r}_i, i = 1, \ldots, n$ — telle que $\mathbf{r}_i \cdot \mathbf{r}_j = 0$ si $i \neq j$. Une telle matrice est dite **de Hadamard**. On note HAD_n, l'ensemble de matrices de Hadamard [6] d'ordre n.
   (a) Montrer que $JJ^t = nI_n$ et conclure que $\det J = n^{n/2}$.
   (b) Montrer que si $J \in \text{HAD}_n$, alors $K = \begin{bmatrix} J & J \\ J & -J \end{bmatrix} \in \text{HAD}_{2n}$. En conclure qu'il existe une matrice dans HAD_{2^m}, pour tout $m \in \mathbb{N}$.
   (c) Si $J \in \text{HAD}_n$, montrer que n est nécessairement pair [7].
   (d) À partir des vecteurs lignes d'une matrice $J \in \text{HAD}_n$, on construit les $2n$ vecteurs $\mathbf{r}_1, -\mathbf{r}_1, \ldots, \mathbf{r}_n, -\mathbf{r}_n$ et on transforme chaque composante -1 en 0. Ces vecteurs (leurs transposés) engendrent alors un sous-espace vectoriel de $\mathbb{F}_2^n$ (i.e. constituent le glossaire d'un code, appelé code de Hadamard). Montrer que la distance minimale de ce code est $d = n/2$.
111. Soit C un $[n,k]$-code sur $\mathbb{F}_2$ de glossaire $\mathcal{G}$. Monter que

$$\sum_{\mathbf{g} \in \mathcal{G}} (-1)^{\mathbf{x}^t \mathbf{g}} = \begin{cases} |\mathcal{G}| & \text{si } \mathbf{x} \in \mathcal{G}^\perp, \\ 0 & \text{sinon.} \end{cases}$$

Codes de Hamming

112. Décoder par le code de Hamming $\text{HAM}(7,4)$ les messages suivants :
   (a) $\alpha = 1101011$,
   (b) $\alpha = 0110110$,
   (c) $\alpha = 0100111$,
   (d) $\alpha = 1111111$.
113. Calculer toutes les chaînes de bruit $\mathbf{b} \in \{0,1\}^7$ qui donnent un syndrome nul pour $\text{HAM}(7,4)$.
114. Pour le code $\text{HAM}(7,4)$ et le canal binaire symétrique avec $p = 0.1$,
   (a) déterminer la probabilité qu'un block de 7 bits ne soit pas décodé correctement,
   (b) en déduire la probabilité du taux d'erreur par bit, dans le cas où le poids du bruit est exactement $\|\mathbf{b}\| = 2$.

6. Hadamard a montré que le déterminant de toute matrice $J \in \mathfrak{M}_{n,n}([-1,1])$ avec des lignes deux-à-deux orthogonales est majoré par $n^{n/2}$; cette borne est saturée si, et seulement si, $J_{i,j} = \pm 1$.

7. On peut montrer que si J est une matrice de Hadamard d'ordre $n > 2$, alors n est divisible par 4. L'affirmation qu'il existe une matrice de Hadamard pour tout ordre n divisible par 4 reste une conjecture encore ouverte !

115. La matrice de contrôle de parité du code $\textsc{Ham}(15,11)$ est donnée par

$$H := \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Décoder le mot $\alpha = 000010000011001$.

116. Les codes $\textsc{Ham}(7,4)$ et $\textsc{Ham}(15,11)$ des exercices précédents n'ont rien de spécial. Ils sont membres d'une famille infinie de codes parfaits $\textsc{Ham}(n,k)$ pour *certaines* valeurs de n et k.

(a) Pour un code binaire parfait qui corrige une erreur, on pose $t = 1$ et $q = 2$ dans la borne d'empilement des boules de Hamming. Déterminer la relation qui doit relier k et n dans ce cas.

(b) Noter $c = n - k$ le nombre de bits de contrôle et déterminer les valeurs possibles de $n := n(c)$ et $k := k(c)$, pour $c \in \mathbb{N}_>$. Les quelques premiers jeux de valeurs sont donnés dans le tableau suivant où l'on reconnaît $\textsc{Ham}(7,4)$ et $\textsc{Ham}(15,11)$ comme des codes correspondant à $c = 3$ et 4.

c	1	2	3	4	5	…
n	1	3	7	15	31	…
k	0	1	4	11	26	…

(c) En notant $\mathcal{H}_c$, $c \geq 1$, le code de Hamming $\textsc{Ham}(n(c), k(c))$, déterminer son taux de transmission asymptotique $R = \lim_{c\to\infty} \frac{k(c)}{n(c)}$.

Bibliographie

[1] Robert B. Ash. *Information theory*. Dover Publications Inc., New York, 1990. Corrected reprint of the 1965 original. 84

[2] Philippe Barbe and Michel Ledoux. *Probabilité*. EDP Sciences, Les Ulis, 2007. Deuxième édition, revue et corrigée. 9, 15

[3] Steven M. Bellovin. Frank Miller: Inventor of the one-time pad. *Cryptologia*, 35, 07 2011. doi:10.1080/01611194.2011.583711. 152

[4] Charles H. Bennett. Notes on Landauer's principle, reversible computation and Maxwell's demon. *Studies in History and Philosophy of Modern Physics*, 34:501–510, 2003. 102

[5] Manabendra Nath Bera, Arnau Riera, Maciej Lewenstein, Zahra Baghali Khanian, and Andreas Winter. Thermodynamics as a consequence of information conservation. *Quantum*, 3:121, February 2019. URL: https://doi.org/10.22331/q-2019-02-14-121, doi:10.22331/q-2019-02-14-121. 84

[6] Antoine Bérut, Artak Arakelyan, Artyom Petrosyan, Sergio Ciliberto, Raoul Dillenschneider, and Eric Lutz. Experimental verification of Landauer's principle linking information and thermodynamics. *Nature*, 483, 3 2012. URL: http://gen.lib.rus.ec/scimag/index.php?s=10.1038/nature10872, doi:10.1038/nature10872. 102

[7] Rabi N. Bhattacharya and Edward C. Waymire. *Stochastic processes with applications*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1990. A Wiley-Interscience Publication. 125

[8] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication. 9, 13

[9] Ludwig Boltzmann. *Vorlesungen über Gastheorie, 1. Theil*. Verlag von Johann Ambrosius Barth, Leipzig, 1896. 84, 85

[10] Ludwig Boltzmann. *Leçons sur la théorie cinétique des gaz*. Traduit de l'original allemand par A. Galloti. Gauthiers-Villars, Paris, 1902. Ré-imprimé par les Éditions Jacques Gabay, Paris (1987). 85

[11] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, April 1998. URL: http:

//dx.doi.org/10.1016/S0169-7552(98)00110-X, doi:10.1016/S0169-7552(98)00110-X. 60

[12] Allen Broughton and Barthel W. Huff. A comment on unions of sigma-fields. *Amer. Math. Monthly*, 84(7):553–554, 1977. URL: http://dx.doi.org/10.2307/2320022, doi:10.2307/2320022. 19

[13] C. Carathéodory. Untersuchungen über die Grundlagen der Thermodynamik. *Mathematische Annalen*, 67(3):355–386, Sep 1909. URL: https://doi.org/10.1007/BF01450409, doi:10.1007/BF01450409. 101

[14] Yuan Shih Chow and Henry Teicher. *Probability theory*. Springer Texts in Statistics. Springer-Verlag, New York, third edition, 1997. Independence, interchangeability, martingales. URL: http://dx.doi.org/10.1007/978-1-4612-1950-7, doi:10.1007/978-1-4612-1950-7. 37

[15] Kai Lai Chung. *A course in probability theory*. Academic Press, Inc., San Diego, CA, third edition, 2001. 37

[16] The Unicode Consortium. *Unicode Standard, Version 5.0, The (5th Edition)*. Addison-Wesley Professional, 5 edition, 2006. 108

[17] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. 84, 125

[18] Harald Cramér. On the representation of a function by certain Fourier integrals. *Trans. Amer. Math. Soc.*, 46:191–201, 1939. URL: https://doi.org/10.2307/1989919, doi:10.2307/1989919. 37

[19] Persi Diaconis, Susan Holmes, and Richard Montgomery. Dynamical bias in the coin toss. *SIAM Rev.*, 49(2):211–235, 2007. URL: http://dx.doi.org/10.1137/S0036144504446436, doi:10.1137/S0036144504446436. 17

[20] Whitfield Diffie and Martin E. Hellman. New directions in cryptography. *IEEE Trans. Information Theory*, IT-22(6):644–654, 1976. URL: https://doi.org/10.1109/tit.1976.1055638, doi:10.1109/tit.1976.1055638. 151

[21] Jean-Guillaume Dumas, Jean-Louis Roch, Éric Tannier, and Sébastien Varette. *Foundations of Coding: Compression, Encryption, Error Correction*. Wiley, 2015. doi:10.1002/9781119005940. 150

[22] Paul Ehrenfest and Tatiana Ehrenfest. Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. *Physikalische Zeitschrift*, 8:311–314, 1907. 57

[23] Peter Elias. Error-free coding. *Information Theory, Transactions of the IRE Professional Group on*, 4(4):29–37, 1954. doi:10.1109/TIT.1954.1057464. 2

[24] Peter Elias. The noisy channel coding theorem for erasure channels. *Amer. Math. Monthly*, 81(8):853–862, 1974. URL: http://dx.doi.org/10.2307/2319442, doi:10.2307/2319442. 2

[25] Thomas L. Floyd. *Electronic devices*. What's New in Trades & Technology. Pearson College Division, 10th edition, 2017. 10th edition. 97

[26] Erol Gelenbe and Yves Caseau. The impact of information technology on energy consumption and carbon emissions. *Ubiquity*, 2015(June), June 2015. URL: https://doi.org/10.1145/2755977, doi:10.1145/2755977. 103

[27] J. Gemmer, M. Michel, and G. Mahler. *Quantum thermodynamics*, volume 784 of *Lecture Notes in Physics*. Springer-Verlag, Berlin, second edition, 2009. Emergence of thermodynamic behavior within composite quantum systems. 84

[28] Hans-Otto Georgii. Probabilistic aspects of entropy. In *Entropy*, Princeton Ser. Appl. Math., pages 37–54. Princeton Univ. Press, Princeton, NJ, 2003. 84

[29] Hans-Otto Georgii. *Stochastik*. de Gruyter Lehrbuch. [de Gruyter Textbook]. Walter de Gruyter & Co., Berlin, expanded edition, 2009. Einführung in die Wahrscheinlichkeitstheorie und Statistik. [Introduction to probability and statistics]. 9, 14

[30] J. Willard Gibbs. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundation of thermodynamics*. Dover publications Inc., New York, 1960. 84

[31] Oded Goldreich. *Foundations of cryptography. I*. Cambridge University Press, Cambridge, 2001. Basic tools. URL: http://dx.doi.org/10.1017/CBO9780511546891, doi:10.1017/CBO9780511546891. 150

[32] Oded Goldreich. *Foundations of cryptography. II*. Cambridge University Press, Cambridge, 2004. Basic Applications. URL: http://dx.doi.org/10.1017/CBO9780511721656.002, doi:10.1017/CBO9780511721656.002. 150

[33] Jeongmin Hong, Brian Lambson, Scott Dhuey, and Jeffrey Bokor. Experimental test of Landauer's principle in single-bit operations on nanomagnetic memory bits. *Science advances*, 2(3):e1501492–e1501492, 2016. doi:10.1126/sciadv.1501492. 102

[34] Paul Horowitz and Winfield Hill. *The art of electronics*. Cambridge Unversity Press, Cambridge, 2015. 97

[35] David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40, 1952. 118

[36] F. Jelinek. *Probabilistic Information Theory*. McGraw-Hill, N. Y., 1968. 119

[37] R.S. Katti and A. Ghosh. Security using Shannon-Fano-Elias codes. In *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*, pages 2689–2692, 2009. doi:10.1109/ISCAS.2009.5118356. 122

[38] Mark Kelbert and Yuri Suhov. *Information Theory and Coding by Example*. Cambridge University Press, Cambridge, 2013. 84

[39] Joseph B. Keller. The probability of heads. *Amer. Math. Monthly*, 93(3):191–197, 1986. URL: http://dx.doi.org/10.2307/2323340, doi:10.2307/2323340. 17

[40] Auguste Kerchoffs. La cryptographie militaire. *Journal des sciences militaires*, pages 5–38, 1883. 149

[41] Aleksandr Yakovlevich Khinchin. *Mathematical foundations of information theory*. Dover Publications Inc., New York, N. Y., 1957. Translated by R. A. Silverman and M. D. Friedman. 2, 84

[42] Andrej Nikolaevich Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer-Verlag, Berlin, 1977. Reprint of the 1933 original. 3, 9

[43] Leon G. Kraft. *A device for quantizing, grouping, and coding amplitude-modulated pulses*. PhD thesis, Massachusetts Institute of Technology, 1949. URL: http://hdl.handle.net/1721.1/12390. 112

[44] R. Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, July 1961. doi:10.1147/rd.53.0183. 97, 102

[45] A. K. Lenstra and H. W. Lenstra, Jr., editors. *The development of the number field sieve*, volume 1554 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1993. URL: http://dx.doi.org/10.1007/BFb0091534, doi:10.1007/BFb0091534. 151

[46] Annick Lesne. Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Math. Structures Comput. Sci.*, 24(3):e240311, 63, 2014. URL: https://doi.org/10.1017/S0960129512000783, doi:10.1017/S0960129512000783. 84

[47] Eugene Lukacs. *Characteristic functions*. Hafner Publishing Co., New York, 1970. Second edition, revised and enlarged. 37

[48] David J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, New York, 2003. 84

[49] Ueli M. Maurer. Authentication theory and hypothesis testing. *IEEE Trans. Inform. Theory*, 46(4):1350–1356, 2000. URL: https://doi.org/10.1109/18.850674, doi:10.1109/18.850674. 156

[50] R. J. McEliece. *The theory of information and coding*, volume 86 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, student edition, 2004. With a foreword by Mark Kac. 84

[51] Brockway McMillan. The basic theorems of information theory. *Ann. Math. Statistics*, 24:196–219, 1953. 114

[52] Jacques Neveu. *Bases mathématiques du calcul des probabilités*. Préface de R. Fortet. Deuxième édition, revue et corrigée. Masson et Cie, Éditeurs, Paris, 1970. 9

[53] J. R. Norris. *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original. 55

[54] Juan M. R. Parrondo, Jordan M. Horowitz, and Takahiro Sagawa. Thermodynamics of information. *Nature Physics*, 11:131 EP –, 02 2015. URL: https://doi.org/10.1038/nphys3230. 84

[55] William A. Pearlman and Amir Said. *Digital signal compression : principles and practice*. Cambridge University Press, 2011. 119

[56] Dimitri Petritis. Markov chains on measurable spaces, 2015. Premiminary draft of lecture notes taught at the University of Rennes 1. URL: http://perso.univ-rennes1.fr/dimitri.petritis/enseignement/markov/2_pdfsam_markov.pdf. 48, 55

[57] Dimitri Petritis. Mathematical foundations of quantum mechanics, 2018. Notes de cours pour le master de cryptographie - version préliminaire, Université de Rennes 1. URL: http://perso.univ-rennes1.fr/dimitri.petritis/enseignement/ptin/ptin.pdf. 152

[58] Dimitri Petritis. Théorie de la complexité, 2018. Notes de cours pour le master de cryptographie - version préliminaire, Université de Rennes 1. URL: http://perso.univ-rennes1.fr/dimitri.petritis/enseignement/lcm1/lcm1.pdf. 125, 155

[59] R. L. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Comm. ACM*, 21(2):120–126, 1978. URL: http://dx.doi.org/10.1145/359340.359342, doi:10.1145/359340.359342. 151

[60] Steven Roman. *Coding and information theory*, volume 134 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1992. 142

[61] Y. A. Rozanov. *Probability theory*. Dover Publications Inc., New York, english edition, 1977. A concise course, Translated from the Russian and edited by Richard A. Silverman. 9

[62] Claude E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948. 83, 113

[63] Claude E. Shannon. Communication in the presence of noise. *Proc. I.R.E.*, 37:10–21, 1949. 2

[64] Claude E. Shannon. Communication theory of secrecy systems. *Bell System Tech. J.*, 28:656–715, 1949. Claude Shannon's report, originally issued as a classified document entitled "A Mathematical Theory of Cryptography", Memorandum MM 45-110-02, September 1, 1945, was formally published in 1949 as "Communication Theory of Secrecy Systems" in Bell System Technical Journal, 28 (1949) 656–715. The original form of the paper, is nowadays available only as a low quality scanned version. Recently a version has been retyped by Jiejun Kong and made available to the community. 2, 152, 153

[65] Albert Nikolaevich Shiryayev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1984. Translated from the Russian by R. P. Boas. 9, 36, 40

[66] Peter W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM J. Comput.*, 26(5):1484–1509, 1997. 151

[67] G. J. Simmons. A survey of information authentication. In *Contemporary cryptology*, pages 379–419. IEEE, New York, 1992. 156

[68] Gustavus J. Simmons. Authentication theory/coding theory. In George Robert Blakley and David Chaum, editors, *Advances in Cryptology*, pages 411–431, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg. 156

[69] Yakov Grigorevich Sinai. *Probability theory*. Springer Textbook. Springer-Verlag, Berlin, 1992. An introductory course, Translated from the Russian and with a preface by D. Haughton. 9

[70] Karlheinz Spindler. *Abstract algebra with applications. Vol. II*. Marcel Dekker Inc., New York, 1994. Rings and fields. 162

[71] G. S. Vernam. Cipher printing telegraph systems for secret wire and radio telegraphic communications. *Transactions of the American Institute of Electrical Engineers*, XLV:295–301, Jan 1926. doi:10.1109/T-AIEE.1926.5061224. 152

[72] Mark N. Wegman and J. Lawrence Carter. New hash functions and their use in authentication and set equality. *J. Comput. System Sci.*, 22(3):265–279, 1981. Special issue dedicated to Michael Machtey. URL: https://doi.org/10.1016/0022-0000(81)90033-7, doi:10.1016/0022-0000(81)90033-7. 159

[73] Jacob Ziv. Coding theorems for individual sequences. *IEEE Trans. Inform. Theory*, 24(4):405–412, 1978. 122

[74] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, 24(5):530–536, 1978. URL: http://dx.doi.org/10.1109/TIT.1978.1055934, doi:10.1109/TIT.1978.1055934. 122