Chapitre 1

Introduction au concept d'information

1.1 Jeux de devinette

Dans le jeu des vingt questions, un joueur pense à un objet, et l'autre joueur tente de deviner de quel objet il s'agit en posant des questions auxquelles on peut répondre par oui ou non, par exemple : « est-ce vivant ? », ou « est-ce humain ? ». Le but est d'identifier l'objet en posant le moins de questions possible. Quelle est la meilleure stratégie pour jouer à ce jeu ?

Pour simplifier, imaginons que nous jouions à une version un peu ennuyeuse des vingt questions appelée «soixante-trois», qui consiste à deviner un nombre entre 0 et 63. Autrement dit, quel est le plus petit nombre de questions oui/non nécessaires pour identifier un entier x entre 0 et 63 ?

Intuitivement, les meilleures questions divisent successivement les 64 possibilités en ensembles de taille égale. Ainsi, six questions suffisent.

1. x est-il ≥ 32 ? *oui*
2. x est-il $\geq 32 + 16$? *non*
3. x est-il $\geq 32 + 8$? *non*
4. x est-il $\geq 32 + 4$? *non*
5. x est-il $\geq 32+2$? *oui*
6. $x = 32 + 2 + 1$? *oui*

Les réponses à ces questions, si elles sont traduites de {oui, non} à {1, 0}, donnent l'expansion binaire de x, par exemple $35 \mapsto 100011$.

Si nous supposons que toutes les valeurs de x sont également probables, alors les réponses aux questions sont indépendantes et on dit que chacune des questions nous donne un bout d'information sur le nombre qui dans la terminologie popularisée par Claude Shannon (qui en attribue l'invention à John Tukey un statisticien américain) vers fin des années 40 est aussi appelée un *bit*[1] ; l'information totale dont nous avons besoin pour deviner le nombre est ainsi de six bits. Notre stratégie de questionnement définit une manière d'encoder la variable aléatoire x sous forme binaire, le terme *bit* désigne depuis par extension chaque chiffre de la représentation binaire d'un nombre.

Jusqu'ici, l'information de Shannon a le sens suivant : elle mesure la longueur d'un mot binaire qui code x. Cependant, nous n'avons pas encore étudié les ensembles où les résultats ont des

1. Jeu de mot entre *bit*, un petit morceau, et la contraction de **bi**nary digi**t**

probabilités inégales. L'information de Shannon a-t-elle également un sens dans ce cas ?

1.2 Information contenue dans une variable aléatoire

Commençons par une discussion heuristique de ce qu'est l'information de façon générale. Si je cherche à identifier une personne, plus une caractéristique est rare, plus celle-ci me donnera d'information sur la personne. Si je vous parle d'une personne qui est dans la classe vous avez plus d'information que si je parle d'une personne à l'université, d'une personne de 1m70 par rapport à une personne de 2m10, ... Donc l'information donnée par le fait qu'une caractéristique $\mathcal{C}$ pensée comme une variable aléatoire soit égale à x sera une fonction décroissante de $P(\mathcal{C} = x)$.

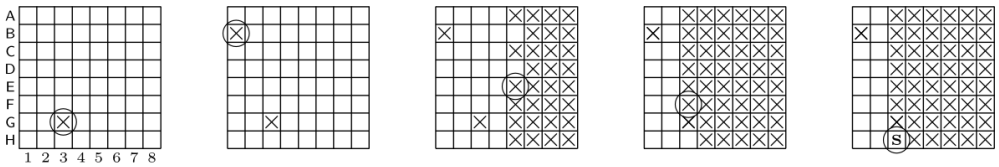
De plus, on a défini l'information des questions indépendante dans l'exemple précédent comme la somme de l'information des 6 questions. L'information est donc supposée additive. Cela motive la définition plus générale de l'information de Shannon pour une variable aléatoire discrète $\mathcal{C}$,

$$I(\mathcal{C} = x) = \log_2 1/P(\mathcal{C} = x).$$

L'unité de cette quantité sera classiquement mesurée en *bits*.

Jeu du Sous-marin Dans le jeu *Touché-Coulé*, chaque joueur cache une flotte de navires dans une mer représentée par une grille carrée. À chaque tour, un joueur tente de toucher les navires de l'adversaire en tirant sur une case de la mer de l'adversaire. La réponse à une case sélectionnée, telle que 'G3', est soit 'raté', 'touché' ou 'touché et coulé'.

Dans une version simplifiée de ce jeu que l'on va appeler *Sous-marin*, chaque joueur cache un seul sous-marin qui tient dans une case d'une grille de huit par huit. La Figure 1.2 montre quelques étapes de ce jeu : le cercle représente la case visée, et les $\times$ montrent les cases dans lesquelles le résultat était un raté, $\times = n$; le sous-marin est touché (résultat $\times = y$ indiqué par le symbole s) à la 49e tentative.



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

FIGURE 1.1 – Quelques étapes du jeu Sous-marin

Chaque coup effectué par un joueur définit un ensemble. Les deux résultats possibles sont y, n, correspondant à un touché et à un raté, et leurs probabilités dépendent de l'état du plateau. Au début, $P(y) = 1/64$ et $P(n) = 63/64$. Au deuxième coup, si le premier coup a raté, $P(y) = 1/63$ et $P(n) = 62/63$. Au troisième coup, si les deux premiers coups ont raté, $P(y) = 1/62$ et $P(n) = 61/62$.

L'information de Shannon acquise à partir d'un résultat $\times$ est $I(\times) = \log(1/P(\times))$. Si nous avons de la chance et touchons le sous-marin dès le premier tir, alors

$$I(\times) = I_{(1)}(y) = \log_2 1/64 = 6 \text{ bits.}$$

Maintenant, il peut sembler un peu étrange qu'un seul résultat binaire puisse transmettre six bits. Mais nous avons appris l'emplacement caché, qui aurait pu être l'un des 64 carrés ; donc, par une seule question binaire chanceuse, nous avons effectivement appris six bits.

Que se passe-t-il si le premier tir rate ? L'information de Shannon que nous obtenons de ce résultat est

$$I(\times) = I_{(1)}(n) = \log_2 \frac{64}{63} = 0,0227 \text{ bits.}$$

Est-ce que cela a du sens ? Ce n'est pas si évident. Continuons. Si notre deuxième tir rate également, le contenu informationnel de Shannon du deuxième résultat est

$$I_{(2)}(n) = \log_2 \frac{63}{62} = 0,0230 \text{ bits.}$$

Si nous ratons trente-deux fois (en tirant sur une nouvelle case à chaque fois), l'information de Shannon totale acquise est

$$\begin{aligned} \log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \ldots + \log_2 \frac{33}{32} \\ = \log_2 \frac{64}{32} = 1 \text{ bit.} \end{aligned}$$

Pourquoi ce nombre rond ? Eh bien, qu'avons-nous appris ? Nous savons maintenant que le sous-marin ne se trouve dans aucune des 32 cases sur lesquels nous avons tirés ; apprendre ce fait revient à jouer au jeu précédent, en posant comme première question : "Est-ce que x est l'un des trente-deux numéros correspondant à ces carrés sur lesquels j'ai tiré ?", et recevoir la réponse "non". Cette réponse élimine la moitié des hypothèses, nous donnant ainsi un bit d'information.

Après 48 tirs infructueux, l'information acquise est de 2 bits : l'emplacement inconnu a été réduit à un quart de l'espace hypothétique original.

Et si nous touchons le sous-marin au 49e tir, quand il reste 16 carrés ? La quantité d'information de cet événement est

$$I_{(49)}(y) = \log_2 16 = 4,0 \text{ bits.}$$

Le contenu d'information de Shannon total de tous les événements est

$$\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \ldots + \log_2 \frac{17}{16} + \log_2 16 = \log_2 64 = 6,0 \text{ bits.}$$

Ainsi, une fois que nous savons où se trouve le sous-marin, le contenu d'information Shannon total acquis est de 6 bits.

Ce résultat est valable quelle que soit le moment auquel nous touchons le sous-marin. Si nous le touchons quand il reste n carrés à choisir parmi $n+1$ – n était 16 dans l'équation précédente – alors l'information totale acquise est :

$$\begin{aligned}&\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \ldots + \log_2 \frac{n+1}{n} + \log_2 \frac{n}{1} \\ &= \log_2 \left[\frac{64}{63} \times \frac{63}{62} \times \ldots \times \frac{n+1}{n} \times n \right] = \log_2 64 = 6,0 \text{ bits.}\end{aligned}$$

Je pense que l'exemple du sous-marin permet d'affirmer que la définition de Shannon est une mesure sensée de la quantité d'information. Et le jeu de soixante-trois montre que le contenu d'information Shannon peut être intimement lié à la taille d'un fichier qui code les résultats d'une expérience aléatoire, suggérant ainsi un lien possible avec la compression de données.

Remarquons que le logarithme en base 2 est naturel dans le cadre d'un codage à deux symboles. Dans la suite nous utiliserons le logarithme népérien pour la définition de l'entropie et nous verrons que pour un codage à d lettre, la quantité de normalisation $\log d$ correspondant à la base d apparaîtra régulièrement. On retrouvera cette normalisation dans le choix d'unité : bit pour la base 2, ban pour la base 10.

Sans ces renormalisations, l'unité est parfois appelée nat, pour natural unit of information. On omettra l'unité dans ce cas.

1.3 Premières propriétés de l'entropie

On peut maintenant interpréter le jeu des devinettes comme un jeu où l'on doit atteindre une certaine quantité d'information qui augmente avec chaque question en fonction de la réponse qui est une variable aléatoire. La stratégie optimale est alors de poser la question qui va maximiser la quantité d'information supplémentaire moyenne obtenue. Notre choix de question induit une réponse qui sera une variable aléatoire $\mathcal{C}$ et nous cherchons donc à maximiser à chaque étape

$$H(\mathcal{C}) = E\left[\log 1/P(\mathcal{C})\right] = -\sum_{x \in \mathcal{C}} p(x) \log p(x).$$

Où l'on prend par convention $0 \log 0 = 0$. Cette quantité apparaît aussi en physique, notamment en mécanique statistique et thermodynamique, Shannon lui a donc donné le même nom [2] : *entropie*.

Proposition 1.3.1. *Soit X une variable aléatoire discrète à valeur dans un ensemble fini indexé par $\{1, \ldots, d\}$ dont chaque élément a une probabilité $p_1, \ldots, p_d$.*

- *$H(X) \geq 0$ avec égalité si et seulement si il existe un indice i tel que $p_i = 1$.*
- *$H(X) \leq \log d$ avec égalité si et seulement si $p_i = 1/\log d$ pour tout indice i.*

Nous ferons la démonstration en exercice.

Théoreme 1.3.2 (de concentration)**.** *Pour une suite de variables aléatoires indépendantes et identiquement distribués (i.i.d) $(X_i)_{i \in \mathbb{N}}$, pour tout $\epsilon, \eta > 0$, il existe un rang à partir duquel on peut trouver un sous ensemble A des valeurs possibles pour $X_1, \ldots, X_n$ tel que*

2. d'après Shannon c'est John von Neumann qui lui aurait conseillé *"You should call it entropy for two reasons : first because that is what the formula is in statistical mechanises but second and more important, as nobody knows what entropy is, whenever you use the term you will always be at an advantage !"*

1. *pour toute suite de valeurs* $(x_1, \ldots, x_n) \in A$

$$\left| -\frac{\log P(x_1, \ldots, x_n)}{n} - H(X) \right| < \eta$$

2. $P(A) > 1 - \epsilon$.

Démonstration. Il suffit de remarquer que par indépendance des variables

$$-\frac{\log P(x_1, \ldots, x_n)}{n} = -\frac{\log P(x_1) + \cdots + \log P(x_n)}{n}.$$

Si on prend les $x_1, \ldots, x_n$ selon la loi P, cela correspond à faire la moyenne des valeurs d'une variable aléatoire $Y = -\log P(X)$ tirée n fois et dont l'espérance est par définition égales à $H(X)$. Le théorème est alors une conséquence directe de la loi faible des grands nombres. □

Loi faible des grands nombres. Pour une suite de variable aléatoire (X_i) i.i.d. telles que $E(X)$ et $Var(X) = E\left(|X - E(X)|^2\right)$ sont finis. Alors pour tout $\epsilon > 0$,

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - E(X)\right| \geq \epsilon\right) = P\left(\left|\frac{X_1 + \cdots + X_n}{n} - E(X)\right|^2 \geq \epsilon^2\right) \leq \frac{Var(X)}{n\epsilon^2}.$$

Donc $P\left(\left|\frac{X_1+\cdots+X_n}{n} - E(X)\right| \geq \epsilon\right) \to 0$ quand n tend vers l'infini. Khintchine a prouvé que l'on peut retirer l'hypothèse sur la variance finie pour avoir cette convergence.

Considérons l'ensemble des suites de n valeurs $x_1, \ldots, x_n$ possibles et ordonnons les par ordre décroissant de probabilité. Si on fixe $\lambda \in]0, 1[$, on note $N_n(\lambda)$ le plus petit entier k tel que la somme des k premiers éléments de cet ensemble, donc les plus probables, dépasse λ.

Corollaire 1.3.3. *Quand n tend vers ∞,*

$$\frac{\log N_n(\lambda)}{n} \longrightarrow H(X).$$

Démonstration. On dit qu'une suite $x_1, \ldots, x_n$ est standard si sa probabilité vérifie l'inégalité 1 du théorème précédent, ou de manière équivalent, si

$$e^{-n(H(X)+\eta)} < P(x_1, \ldots, x_n) < e^{-n(H(X)-\eta)}$$

Par le théorème 2, si on choisit $\epsilon < \lambda$ et $\epsilon < 1 - \lambda$, la somme des probabilités des suites standard est supérieure à λ si n est assez grand. Si on ordonne les probabilité des suites $x_1, \ldots, x_n$,

$$\underbrace{\overbrace{p_1, \ \cdots, \ p_i}^{\geq e^{-n(H(X)-\eta)}}, \overbrace{p_{i+1}, \ \cdots, \ p_{N_n(\lambda)}}^{\text{suites standard}}}_{>\lambda}, \ p_{N_n(\lambda)+1}, \ \cdots, \ p_j, \ \overbrace{p_{j+1}, \ \cdots, \ p_{|A|^n}}^{\leq e^{-n(H(X)+\eta)}}$$

On remarque que $N_n(\lambda)$ tombe parmi les suites standards. En effet, la somme des probabilité jusqu'à p_j contient la somme des probabilité des suites standard donc est supérieure à λ. Ainsi

$N_n(\lambda) \leq j$. Et la somme jusqu'à i est inférieure à ϵ donc $N_n(\lambda) > i$.

Donc pour tout $k \leq N_n(\lambda)$, $p_k > e^{-n(H(X)+\eta)}$ et $p_1 + \cdots + p_{N_n(\lambda)} > N_n(\lambda) \cdot e^{-n(H(X)+\eta)}$. Par ailleurs la somme $p_1 + \cdots + p_{N_n(\lambda)-1} < \lambda$ et $p_{N_n(\lambda)} < e^{-n(H(X)-\eta)}$ donc

$$\lambda < p_1 + \cdots + p_{N_n(\lambda)} < \lambda + e^{-n(H(X)-\eta)}.$$

Finalement, pour tout η

$$N_n(\lambda)e^{-n(H(X)+\eta)} \leq \lambda + e^{-n(H(X)-\eta)}$$

donc

$$\frac{\log N_n(\lambda)}{n} - H(X) - \eta \leq \frac{\log \lambda}{n} + \frac{1}{n} \cdot \log\left(1 + \frac{1}{\lambda} \cdot e^{-n(H(X)-\eta)}\right)$$

d'où

$$\limsup_{n\to\infty} \frac{\log N_n(\lambda)}{n} \leq H(X) + \eta.$$

Maintenant la somme des probabilités $p_1 + \cdots + p_i$ est plus petite que le ϵ choisi en début de preuve, ce qui implique que la somme des probabilités $p_{i+1} + \cdots + p_{N_n(\lambda)} \geq \lambda - \epsilon$. D'autre part, pour les suites standard sélectionnées la probabilité est plus petite que $e^{-n(H(X)-\eta)}$ on a donc

$$\lambda - \epsilon \leq (N_n(\lambda) - i) \cdot e^{-n(H(X)-\eta)} \leq N_n(\lambda)e^{-n(H(X)-\eta)}$$

Donc

$$\frac{\log N_n(\lambda)}{n} \geq H(X) - \eta + \frac{\log(\lambda - \epsilon)}{n}$$

d'où

$$\liminf_{n\to\infty} \frac{\log N_n(\lambda)}{n} \geq H\left(X\right) - \eta.$$

□

Ces résultats nous disent donc que pour un grand nombre de tirages indépendants n d'une variable aléatoire, tout se passe comme si avec une probabilité arbitrairement proche de 1, les résultats étaient tirés de façon équiprobable dans un ensemble de $e^{nH(X)}$ valeurs.

1.4 Lien avec la compression des données

On modélise une source (discrète) de données par une suite de v.a. $(X_i)_{i\in\mathbb{N}}$ à valeurs dans un ensemble fini $\mathcal{A}$ appelé alphabet de la source. Si les X_i sont i.i.d. la source est dite sans mémoire. En général on préfère penser qu'une source génère un ou plusieurs mots finis aléatoires de longueur quelconque. C'est à dire un élément de

$$\mathcal{A}^* := \bigcup_{n\in\mathbb{N}} \mathcal{A}^n.$$

Un question centrale en théorie de l'information est celle de l'étude des codages de ces mots. C'est à dire des applications injectives $C : \mathcal{A}^* \to \mathcal{B}^*$. ;

L'idée essentielle de la compression est d'observer que si certaines suite de lettres apparaissent plus fréquemment que d'autres, on pourra réduire la taille moyenne des mots en privilégiant des

Lettre	Code	Longueur	Fréquence
E	•	1	12,49%
T	–	3	9,28%
A	•–	4	8,04%
O	– – –	9	7,64%
I	••	2	7,57%
N	–•	4	7,23%
S	• • •	3	6,51%
R	• – •	5	6,28%
H	• • ••	4	5,05%
L	• – ••	6	4,07%
D	– • •	5	3,82%
C	– • –•	8	3,34%
U	• • –	5	2,73%
M	––	6	2,51%
F	• • –•	6	2,40%
P	• – –•	8	2,14%
G	– – •	7	1,87%
W	• – –	7	1,68%
Y	– • ––	10	1,66%
B	– • ••	6	1,48%
V	• • •–	6	1,05%
K	– • –	7	0,54%
X	– • •–	8	0,23%
J	• – –––	10	0,16%
Q	– – •–	10	0,12%
Z	– – ••	8	0,09%

TABLE 1.1 – Tableau du Code Morse

suites courtes pour coder les suites les plus fréquentes et des suites plus longues pour les moins fréquentes.

On définit un *taux de compression* de la façon suivante ; si on a un codage des suites de longueur n dont la longueur est notée pour chaque suite $\ell(x_1, \ldots, x_n)$, le taux de compression moyen pour la longueur n est

$$\tau_n = \frac{\sum P(x_1, \ldots, x_n) \cdot \ell(x_1, \ldots, x_n)}{n}.$$

Théoreme 1.4.1. *Le taux de compression optimal tend vers $H(X)/\log|\mathcal{B}|$ quand $n \to \infty$.*

Démonstration.

Borne inférieure Pour $H' < H(X)$ et un code quelconque pour les mots de longueur n, on dit qu'une suite est spéciale si $\ell(x_1, \ldots, x_n) < nH'/\log|\mathcal{B}|$. Le nombre de séquence spéciales est nécessairement inférieur au nombre de mots d'une telle longueur

$$|\mathcal{B}|^{nH'/\log|\mathcal{B}|} = \exp\left(\log|\mathcal{B}| \cdot nH'/\log|\mathcal{B}|\right) = e^{nH'}.$$

Soit λ_n la probabilité de l'ensemble de telles séquences. Par le théorème précédent, pour tout ϵ et tout η, il faut au moins $\epsilon \cdot e^{n(H(X)-\eta)}$ séquences pour que la somme des probabilités $\lambda_n \geq 2\epsilon$. Ce qui est bien plus que $e^{nH'}$ si l'on choisit η assez petit. Donc pour tout ϵ, $\lambda_n < 2\epsilon$ et pour toute séquence non spéciale $\ell(x_1, \ldots, x_n) \geq nH'/\log|\mathcal{B}|$ donc $\tau_n \geq (1-2\epsilon)H'/\log|\mathcal{B}|$ à partir d'un certain rang.

Borne supérieure On cherche à construire un code de taux $< (H(X)+\delta)/\log|\mathcal{B}|$ pour tout $\delta > 0$. Soit $\lambda \in]0,1[$, par le résultat précédent, pour tout $\eta > 0$, si n est assez grand, $N_n(\lambda) \leq e^{n(H(X)+\eta)}$. Or cette dernière quantité est exactement le nombre de séquence de longueur $n(H(X)+\eta)/\log|\mathcal{B}|$. On code donc ces éléments par ces séquences de longueur réduite et on garde les autres tels quels de longueur n. Alors le taux de compression est majoré par

$$\frac{H(X)+\eta}{\log|\mathcal{B}|} + (1-\lambda) < \frac{H(X)+\delta}{\log|\mathcal{B}|}$$

quitte à prendre λ proche de 1 et η proche de 0. □

Shannon estime que l'entropie propre à la langue anglaise semble expérimentalement s'approcher de $0,7$ ban ou $1,6$ nat. Autrement dit, on pourrait compresser un texte d'environ $1,6/\log 26 \approx 50\%$ sans perdre d'information. En pratique on peut enlever quasiment toutes les voyelles sans perdre d'information, c'est déjà ce qui est fait en arabe et en hébreu. En fait il admet ne pas prendre en compte la structure plus globale de l'anglais (notamment grammaticale) dans son estimation, elle est donc certainement inférieure.