Théorie de l'information

KTorphée 8

Antoine Chambert-Loir

Résumé. — Petite balade aux abords de la théorie de l'information, inspiré par le cours (Chambert-Loir, 2022) que j'avais enseigné au master maths-info de l'université Paris Cité.

1. Entropie

C'est en 1948 que Claude Shannon a publié son article *La théorie mathématique de la communication*, publié en français dans (Shannon & Weaver, 2018). La définition mathématique fondamentale qu'il y propose est celle d'entropie d'une variable aléatoire.

Définition 1.1 **(Entropie).** — *Si X est une variable aléatoire discrète à valeurs dans un ensemble A, on apelle entropie de X la somme (finie ou infinie)*

$$H(X) = \sum_{a\in A} \mathbf{P}(X = a)\log(\mathbf{P}(X = a)^{-1}).$$

L'idée de cette définition est que $H(X)$ mesure l'incertitude de X, la quantité de hasard qu'elle contient. Le mot *entropie* apparaît dans de nombreux contextes scientifiques, souvent dès qu'on y trouve une expression en $x\log(x)$. Si l'idée de hasard y est souvent présente, il n'est pas tout à fait clair qu'il s'agisse à chaque fois de la même chose, et d'ailleurs, von Neumann aurait suggéré ce terme à Shannon « parce que personne ne sait bien ce qu'est l'entropie. »

Comme $\mathbf{P}(X = a)$ est un nombre entre 0 et 1, l'entropie de $H(X)$ est un nombre positif ou nul, éventuellement infini.

L'entropie est nulle si et seulement si X est certaine : il existe a tel que $\mathbf{P}(X = a) = 1$.

Si X est le lancer d'un dé à 6 faces, équilibré, : on a $A = \{1, \dots, 6\}$, et $P(X = a) = 1/6$ pour tout a, donc $H(X) = 6\dots(1/6)\log(6) = \log(6)$. Plus généralement, si X est le lancer d'un dé équilibré à n faces, alors $H(X) = \log(n)$. Pour une pièce (pile ou face), on trouve $H(X) = \log(2)$.

Si X est la somme d'un lancer de deux dés équilibrés à 6 faces, alors X prend 11 valeurs seulement, entre 2 et 12, mais les probabilités sont différentes : $\mathbf{P}(X = 2) = 1/36$ par exemple, tandis que $\mathbf{P}(X = 7) = 1/6$. On trouve $H(X) \approx 3{,}274$ alors que $\log(11) \approx 3{,}459$.

En utilisant la propriété de concavité de la fonction $x\log(x^{-1})$, on peut démontrer que si X prend n valeurs, alors $H(X) \leq \log(n)$, avec égalité si et seulement si la loi est uniforme.

La définition de l'entropie admet deux variantes.

***Définition 1.2* (Entropie conditionnelle).** — *Soit X, Y deux variables aléatoires discrètes à valeurs dans des ensembles A et B. L'*entropie conditionnelle *de X sachant Y est définie par*

$$H(X \mid Y) = \sum_{b \in B} \mathbf{P}(Y = b) H(X \mid Y = b).$$

Pour chaque valeur $b \in B$, on peut imaginer restreindre le hasard de X aux cas où Y prend la valeur b ; cela change bien sûr les probabilités et on obtient une autre variable aléatoire, « X sachant $Y = b$ », dont on calcule l'entropie ; et on prend la moyenne sur toutes les valeurs de b.

En utilisant la formule pour les probabilités conditionnelles :

$$\mathbf{P}(X = a \mid Y = b) = \frac{\mathbf{P}(X = a \ Y = b)}{\mathbf{P}(Y = b)}$$

on démontre la formule

$$H(X, Y) = H(Y) + H(X \mid Y). \tag{1.3}$$

On l'interprète ainsi : $H(X \mid Y)$ est le hasard que contiennent conjointement X et Y mais qui est dû à X. La définition de $H(X \mid Y)$ montre que c'est une quantité positive ; il y a en particulier plus de hasard dans (X, Y) que dans Y.

***Définition 1.4* (Information mutuelle).** — *Soit X, Y deux variables aléatoires discrètes à valeurs dans des ensembles A et B. L'*information mutuelle *de X sachant Y est définie par*

$$I(X, Y) = \sum_{a,b} \mathbf{P}(X = a \ \& \ Y = b) \log \left(\frac{\mathbf{P}(X = a \ \& \ Y = b)}{\mathbf{P}(X = a)\mathbf{P}(Y = b)} \right).$$

Il n'est pas tout à fait évident que c'est une quantité positive, mais cela se démontre avec des inégalités de convexité. De plus, $I(X, Y) = 0$ si et seulement si $\mathbf{P}(X = a \ \& \ Y = b) = \mathbf{P}(X = a)\mathbf{P}(Y = b)$ pour tous a, b, c'est-à-dire lorsque les variables aléatoires X et Y sont indépendantes.

En utilisant encore la formule pour les probabilités conditionnelles, on démontre la formule

$$H(X, Y) + I(X, Y) = H(X) + H(Y). \tag{1.5}$$

Autrement dit : Le hasard contenu dans le couple (X, Y) est plus petit que la somme des deux hasards contenus dans X et Y, la différence étant l'information mutuelle.

Il y a une caractérisation abstraite de l'entropie en tant que fonction de la loi de la variable aléatoire X, c'est-à-dire des probabilités $\mathbf{P}(X = a)$: c'est l'unique fonction qui vérifie les conditions suivantes :

– Elle est continue en les probabilités $\mathbf{P}(X = a)$;

– Elle est additive pour des variables aléatoires indépendantes, et, plus généralement, vérifie la relation d'addivité pour l'entropie conditionnelle.

– Elle est normalisée par $H(X) = \log(2)$ pour un tirage à pile ou face.

2. Codage

2.1. — Introduisons maintenant l'application que Shannon avait en vue, la théorie mathématique de la communication. Dans le modèle de communication qu'il propose, Shannon voit cinq étapes :

- La source (le chanteur, l'écrivaine, votre grand-mère…) ;
- L'émetteur (la station de radio, de télévision, le téléphone…) ;
- Le canal (l'air, la fibre optique, le cable en cuivre…) ;
- Le récepteur (votre poste de radio, de téléphision, votre téléphone) ;
- La destination (ici, vous, mais la situation est parfois symétrique).

Un message, issu d'une source, est produit par l'émetteur, transmis via un canal, reconstruit par le récepteur et reçu par le destinataire. Shannon ajoute une difficulté dans la troisième étape, la présence de *bruit* qui pourrait corrompre le message transmis.

La question de base est : à quelle vitesse le système émetteur–canal–récepteur peut-il transmettre l'information, et avec quelle fiabilité ? La réponse de Shannon est : en l'absence de bruit, aussi vite que ne le permet l'entropie de la source, considérée comme une variable aléatoire, mais pas plus vite ; et en présence de bruit, il propose une notion de *capacité d'un canal* qui est exactement le facteur limitant (indépendamment de la source).

Commençons par le cas sans bruit.

2.2. — La première remarque qu'il faut justifier est la modélisation probabiliste de cette histoire. Du point de vue du destinataire, le message qu'ielle va recevoir est inconnu et on part du principe qu'il est aléatoire. Le message est naturellement conçu comme une suite de symboles élémentaires, des lettres, des valeurs de son, etc. Ces symboles élémentaires forment une suite de variables aléatoires, ce qu'on appelle un processus stochastique.

L'hypothèse de base est que le message obéit à une certaine loi de probabilité ; par exemple, en français, toutes les lettres n'ont pas la même fréquence, de même que les digrammes, trigrammes. La conséquence est que ce processus stochastique est un peu compliqué, les valeurs successives ne sont pas indépendantes les unes des autres, mais on fera comme si ce ne posait pas de problème. On peut toujours prétendre qu'on agglomère un grand nombre de valeurs consécutives.

2.3. — L'idée initiale de Shannon est que l'entropie de X est liée à la quantité d'information qu'elle contient. Pour justifier cette idée, il propose une *interprétation statistique* de l'entropie qui est fondamentale dans son approche.

Imaginons une variable aléatoire discrète X dont on répète indéfiniment le tirage, de façon indépendante, et qu'on collecte toutes ces valeurs. Le théorème de Shannon (*asymptotic equirepartition property*) est que tout se passe comme si l'on procédait à une répartition de tirages uniformes dans un ensemble de cardinal $\exp(H(X))$. Pour donner sens à ceci, on fixe un nombre réel $\varepsilon > 0$, supposé petit, et on regarde un grand nombre n de tirages et on ne garde que l'ensemble A_ε^n des suites $(a_1, \dots, a_n)$ dont la probabilité vérifie

$$\exp(-nH(X) - n\varepsilon) \leq \mathbf{P}(X_1 = a_1, \dots, X_n = a_n) \leq \exp(-nH(X) + n\varepsilon).$$

Le théorème de Shannon est que

$$\exp(nH(X) - n\varepsilon) \leq \mathrm{Card}(A_\varepsilon^n) \leq \exp(nH(X) + n\varepsilon)$$

et

$$\mathbf{P}(\complement A_\varepsilon^n) < c/n\varepsilon^2.$$

2.4. — Que faire donc ? C'est très simple en fait ! Choisir $\varepsilon > 0$ petit et vérifier qu'on peut négliger ε dans tout ce qui suit, pourvu que n soit assez grand. On a donc un grand message, de n symboles, et cet ensemble A_ε^n de cardinal en gros $\exp(nH(X))$. On numérote naïvement ces $\exp(nH(X))$ valeurs et, au lieu de transmettre $(a_1, \dots, a_n)$, on transmet les chiffres du numéro qui lui a été donné, par exemple en base 2. Il se passe qu'un nombre entre 1 et N a en gros $\log_2(N)$ chiffres ; et donc il suffit de transmettre $\log_2(N)$ bits. Comme $N = \exp(nH(X))$, cela signifie $nH(X)/\log(2)$ bits. Et s'il faut $nH(X)/\log(2)$ bits pour transmettre n valeurs de la variable X, il en faut donc en moyenne $H(X)/\log(2)$ pour en transmettre une seule.

Bien sûr, il faut aussi transmettre $(a_1, \dots, a_n)$ lorsque cette suite n'est pas dans l'ensemble A_ε^n ; dans ce cas, on la transmet naïvement. Cela prend plus de temps mais ça arrive peu souvent.

En renversant le paradigme, on voit que si on prend des logarithmes en base 2, $H(X)$ est la quantité moyenne de questions oui/non qu'il faut poser pour connaître la valeur de X. Dans l'exemple du jeu des 20 questions, c'est le résultat des n parties qu'on n'a pas encore jouées ; pour un texte qu'on veut transmettre, c'est les n premiers symboles.

2.5. — Si on revient au processus stochastique initial, on voit également que ce qui compte, c'est l'entropie de la variable aléatoire $(X_1, \dots, X_n)$ divisée par n. Lorsque n tend vers l'infini, la limite de cette expression est appelée taux d'entropie du processus stochastique X.

2.6. — Le procédé proposé par Shannon est tout sauf praticable. Huffmann (1951) en a proposé un qui est très facile à programmer, et est en plus optimal. La méthode consiste à écrire les symboles avec leurs fréquence et à construire peu à peu un arbre dont ces symboles sont les feuilles. On regroupe les deux symboles de fréquences les plus faibles en un seul symbole, de fréquence somme ; et on recommence. Il faut faire un dessin pour visualiser cet arbre en train de se créer. À la fin, il y a un seul symbole, de fréquence 1 ; c'est la racine de l'arbre de Huffman.

Pour envoyer un symbole, il suffit de coder le chemin (suite de droite/gauche, 0/1) qui va de la racine à ce symbole. Par construction de l'arbre, ce chemin est moins long lorsque ce symbole est fréquent et on démontre que c'est un procédé optimal.

2.7. — Lorsque le canal est bruité, les choses sont plus compliquées et Shannon propose un modèle mathématique du bruit. L'idée est que le canal transforme chaque symbole en un autre, de façon aléatoire, et surtout indépendante (il dit que le canal est sans mémoire).

Dans le cas le plus simple, d'un canal binaire sans mémoire, il y a deux symboles, 0 et 1, qui ont une probabilité p d'être échangés, et donc $1 - p$ d'être préservés. Si le bruit est faible, p est petit. S'il est élevé, p est grand. Mais il faut voir que le cas le pire est $p = 1/2$. Car si, à

l'extrême, $p = 1$, alors le canal échange systématiquement les valeurs, et il suffit de défaire l'échange.

La modélisation de Shannon d'un canal bruité, sans mémoire, est ainsi celle d'une famille de lois : lorsqu'il reçoit a, le canal transmet b avec une probabilité $p(b \mid a)$.

***Définition 2.8* (Capacité d'un canal). —** *La capacité $I(C)$ d'un tel canal est la borne supérieure Dans ce contexte, Shannon considère où X et Y sont des variables aléatoires telles que $\mathbf{P}(Y = b \mid X = a) = p(b \mid a)$ pour tous a, b.*

Shannon interprète la formule

$$I(X, Y) = H(X) - H(X \mid Y)$$

comme disant que $I(X, Y)$ est la quantité d'information qu'il faut pour connaître X sachant qu'on connaît Y.

La borne supérieure dans la définition de la capacité $I(C)$ signifie que la source a la possibilité d'adapter la façon dont elle envoie un message au canal donné. Pour donner un exemple caricatural : si un symbole était transmis de façon absolument imprédictible, il ne faudrait pas l'utiliser.

Pour un canal binaire comme précédemment, on trouve $I(C) = 1 - h(p)$, qui est minimal, égal à 0, lorsque $p = 1/2$ et maximal égal à 1 lorsque $p = 0$ ou $p = 1$.

En général, il est impossible de calculer explicitement $I(C)$, mais on peut toujours, pour des valeurs données, résoudre numériquement le problème d'optimisation.

2.9. — Ce que peut, et veut faire, l'émetteur, c'est choisir un *code* qui transforme le message initial en un mot X de longueur n dans un ensemble de symboles M ; ce mot est ensuite transmis par le canal, reçu en un mot Y par le récepteur, et décodé. On peut alors calculer la *probabilité d'erreur de transmission*. Le second théorème de Shannon est que l'on peut s'arranger pour que cette probabilité d'erreur soit aussi petite que l'on veut et qu'en même temps le taux de transmission du code,

$$\frac{\log(\mathrm{Card}(M))}{n},$$

soit au plus égal à $I(C)$. À l'inverse, si l'on veut un taux de transmission plus grand que $I(C)$, la probabilité d'erreur sera proche de 1.

La démonstration est assez délicate, elle est surtout paradigmatique d'un procédé de démonstration qui a été popularisé dans les années 50 par Erdős sous le nom de *méthode probabiliste*. Elle consiste à regarder tous les codes possibles, pour chacun d'entre eux, regarder le décodage le moins idiot, et calculer la probabilité globale d'erreur. Shannon démontre qu'elle est petite ; c'est donc qu'il y a *un* code pour laquelle elle est petite. C'est donc une démonstration non constructive qui n'aide absolument pas un émetteur donné à résoudre son problème de transmission.

2.10. — Ce théorème a donné l'impulsion nécessaire à la théorie des codes détecteurs/correcteurs d'erreur dont des exemples courants sont la clé de contrôle du numéro de sécurité sociale, ou celui des numéros de banque. La présence de ce code de contrôle

augmente la longueur du numéro mais diminue la probabilité qu'une erreur de transmission ne soit pas détectée.

La théorie des codes a vu une grande activité, par des méthodes d'algèbre (codes de Hamming, années 60) ou de géométrie algébrique (codes de Goppa, années 80), mais aucun de ces codes n'approchait les bornes promises par Shannon. Il a fallu l'invention (vers 2000) des turbocodes par Claude Berrou pour les atteindre.

3. Échantillonage

3.1. — Dernière étape de cette excursion en théorie de l'information, un troisième théorème de Shannon — également attribué à Nyquist — et que Shannon qualifiait de « *common knowledge in the communication art.* »

Échantillonner un signal, c'est le mesurer à intervalles (a priori) réguliers, en vue de le reconstituer ensuite. C'est donc d'un signal qui dépend d'une grandeur continue dont il est question ici, par exemple la pression de l'air sur la membrane d'un microphone en fonction du temps, la couleur ou la luminosité d'une zone d'image. Après tout, le cinéma n'est rien d'autre qu'un échantillonage d'images tous les $\frac{1}{24}$ s.

Ce principe d'échantillonages est nécessaire au traitement numérique des signaux : autant on peut transmettre un signal variable de façon analogique, autant le traitement numérique importe de sélectionner des valeurs numériques. Lorsqu'on lit 44.1 kHz sur une pochette de CD, cela signifie que le signal est échantillonné 44 mille fois par seconde. Et lorsqu'on lit 320 Kbps sur la spécification du MP3, cela signifie qu'on aura quantifié ces valeurs numériques c'est-à-dire en gardé une précision finie, de sorte à disposer de 320 mille bits d'information par seconde.

Bien évidemment, tout échantillonage provoque la perte de l'information entre deux instants successifs, de sorte qu'il est nécessaire de savoir que le signal ne varie pas trop vite par rapport à la fréquence d'échantillonage.

L'exemple de base est un signal sinusoïdal, $\varphi(t) = \sin(\omega t)$, de fréquence $\omega/2\pi$; si on l'échantillonne à la même fréquence, on récupère toutes les valeurs $\varphi(n2\pi/\omega) = \sin(2\pi n)$=0, pour $n \in \mathbf{Z}$, c'est-à-dire identiquement 0. Le même problème se produit si l'on double la fréquence d'échantillonage. Cette limite était bien connue : il ne faut pas échantillloner à une fréquence inférieure au double de la « fréquence naturelle du signal ».

3.2. — C'est la théorie de Fourier qui permet de préciser ce qu'est cette fréquence. Au 18e s., les mathématiciens (Euler, d'Alembert, etc.) ont pris l'habitude de représenter certaines fonctions périodiques comme une superposition de signaux sinusoïdaux et dans sa *Théorie mathématique de la chaleur*, Fourier a systématisé le procédé en concevant qu'il soit possible pour à peu près toute fonction. C'est la théorie des séries de Fourier.

Il y a une théorie analogue pour les fonctions non périodiques, fournie par la transformation de Fourier. Les preuves sont un peu plus délicates mais on peut toujours imaginer qu'on modifie une fonction en la rendant périodique à partir de ce qu'elle vaut sur un grand intervalle $[-T; T]$. Alors, la théorie des séries de Fourier s'applique. Et on fait tendre T vers l'infini.

En formule, définissons la transformée de Fourier d'une fonction φ par la formule

$$\widehat{\varphi}(\omega) = \int_{-\infty}^{\infty} \varphi(t) e^{-2\pi i \omega t}\, dt.$$

Il faut des conditions pour que l'intégrale existe, bien sûr, et ce n'est pas trop le lieu pour insister sur ces conditions. Le théorème d'inversion de Fourier dit qu'alors

$$\varphi(t) = \int_{-\infty}^{\infty} \widehat{\varphi}(\omega) e^{2\pi i t \omega}\, d\omega.$$

3.3. — Sous l'hypothèse que $\widehat{\varphi}(\omega) = 0$ pour $|\omega| \geq W$, le théorème de Nyquist-Shannon fournit la formule

$$\varphi(t) = \sum_{n \in \mathbf{Z}} \varphi(n/2W) \frac{\operatorname{sinc}(t - n/2W)}{t - n/2W}.$$

Autrement dit : en échantillonant φ à fréquence $2W$, on reconstruit sa valeur en tout point.

Une façon de faire la démonstration consiste à remplacer $\widehat{\varphi}$ par la fonction $2W$-périodique S qui coïncide avec $\widehat{\varphi}$ sur $[-W; W]$. On calcule alors les coefficients de Fourier de S : ce sont précisément les valeurs de φ à la fréquence d'échantillonage double : $c_n(S) = \varphi(-n/2W)/2W$. En écrivant que S est somme de sa série de Fourier, on a donc

$$S(\omega) = \frac{1}{2W} \sum_{n \in \mathbf{Z}} \varphi(n/2W) e^{-2\pi i n \omega / 2W}.$$

Par construction, on a $S(\omega) = \widehat{\varphi}(\omega)$ si $|\omega| \leq W$, et on a $\widehat{\varphi}(\omega) = 0$ sinon. Il reste à calculer $\varphi(t)$ par la formule d'inversion de Fourier :

$$\varphi(t) = \int_{-\infty}^{\infty} \widehat{\varphi}(\omega) e^{2\pi i t \omega}\, d\omega = \int_{-W}^{W} S(\omega) e^{2\pi i t \omega}\, d\omega = \frac{1}{2W} \sum_{n \in \mathbf{Z}} \varphi(n/2W) \int_{-W}^{W} e^{-2\pi i n \omega / 2W} e^{2\pi i t \omega}\, d\omega,$$

ce qui fournit la formule voulue.

3.4. — Il reste une remarque à faire pour comprendre les limites d'un tel théorème. La plupart des signaux ne sont pas infinis, et s'ils l'étaient, il ne pourrait être question de l'échantillonner pour toutes les valeurs $n/2W$; par exemple, parce qu'on ne sait rien du futur ! Donc en pratique, un signal a une durée finie, et les fonctions φ qui nous intéressent sont nulles hors d'un intervalle $[-T; T]$. Le souci immédiat est qu'alors, on peut démontrer que $\widehat{\varphi}$ ne vérifie *jamais* la condition du théorème.

C'est un cas très simple d'un *théorème d'incertitude* à la Heisenberg : si un signal est localisé en temps, il ne peut pas l'être en fréquence, et inversement.

Références

A. Chambert-Loir (2022), *Théorie de l'information : trois théorèmes de Claude Shannon*, Nano, Calvage & Mounet, Paris.

C. E. Shannon & W. Weaver (2018), *La Théorie Mathématique de La Communication*, Cassini.