La force d'une idée simple

Hommage à Claude SHANNON à l'occasion du centenaire de sa naissance

Alain Chenciner
Observatoire de Paris, IMCCE (UMR 8028), alain.chenciner@obspm.fr
& Département de mathématique, Université Paris 7

Abstract

De nombreuses fautes de frappe n'empêchent pas de reconnaître sans ambiguïté un texte pourvu que la forme altérée ressemble plus au texte initial qu'à tout autre texte admissible. Jointe à une utilisation systématique de la *loi des grands nombres* qui implique la *propriété d'équipartition asymptotique (AEP)*, cette simple remarque est à la base de la découverte par Claude Shannon de la limite $H < C$ aux performances de tout *code correcteur d'erreurs* permettant une transmission fiable d'information par un canal "bruité" (i.e. "faisant des erreurs") ainsi que de l'existence d'un code permettant d'approcher arbitrairement près de cette limite. Toutes deux de nature probabiliste, l'*entropie* H d'une source de messages et la *capacité* C d'un canal de transmission sont définies par Shannon dans l'article [Sh] qu'il publie en 1948 dans la revue des "Bell labs", l'année même où, dans les mêmes Bell labs, John Bardeen, Walter Brattain et William Shockley font la première démonstration du fonctionnement d'un transistor. Ainsi, des deux découvertes simultanées dont est né le monde d'information dans lequel nous vivons, l'une est de pure mathématique et même de la pire espèce, un théorème d'existence !

Les inventeurs du langage "Morse" avaient déjà compris l'intérêt qu'il y a, pour raccourcir le temps nécessaire à l'envoi d'un message, de représenter chaque lettre par un symbole d'autant plus court que la lettre est fréquente dans le langage utilisé, en l'occurence l'anglais : E est par exemple représenté par un point alors que Z est représenté par 4 symboles, deux tirets suivis de deux points. C'est avant la lettre une utilisation de l'*entropie* de l'anglais et une illustration de l'idée maîtresse du texte fondamental [Sh] de Shannon, à savoir que *c'est la structure statistique*[1] *d'un ensemble de messages* et non la considération d'un message isolé et encore moins de son sens (exit toute considération de sémantique) qui va tenir le rôle principal lorsqu'on cherche à transmettre une suite de symboles de façon fiable et efficace. Je laisse aux historiens (voir en particulier [Se]) le soin de décider de l'influence exacte sur Shannon, qui d'ailleurs le

[1] c'est-à-dire la fréquence relative des différents symboles ou des différentes suites de symboles composant un message.

remercie explicitement dans une note de [Sh] (partie III, note 4, page 626), des idées de Norbert Wiener sur les *séries temporelles* et les *filtres* développées dans les années quarante mais classifiées et publiées seulement en 1949 dans le livre *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*[2]. Norbert Wiener qui se considérera lui-même comme le co-fondateur de la théorie de l'information, reconnaît cependant l'influence d'Andrei Kolmogorov et en particulier de sa note *Sur l'interpolation et l'extrapolation des suites stationnaires* publiée en 1939 aux Comptes Rendus de l'Académie des Sciences de Paris. Ainsi, on ne s'étonnera pas que Kolmogorov, qui au début des années trente avait donné sa forme mathématique actuelle à la théorie des probabilités et qui au milieu des années cinquante introduira l'entropie métrique d'un système dynamique ([Ko1, Ka]), ait été l'un des premiers sinon le premier mathématicien à comprendre l'importance fondamentale du travail de Shannon[3].

Une illustration simple (et même simpliste) d'une telle structure statistique est le jeu de pile ou face : les 26 lettres de l'alphabet sont remplacées par les deux symboles 0 et 1 et l'écriture d'un message de N lettres par celle d'une suite $a_1 a_2 \cdots a_N$ obtenue à partir de N lancers d'une pièce de monnaie en posant que $a_i = 0$ ou 1 suivant qu'au i-ème lancer la pièce est retombée sur pile ou sur face. Si N est assez grand, si la pièce est bien équilibrée et si chaque lancer est indépendant des autres, la *loi des grands nombres* affirme qu'avec une probabilité d'autant plus grande que N est grand, la suite obtenue contiendra approximativement autant de 0 que de 1. Si maintenant la pièce est biaisée, avec les probabilités respectivement p et $q = 1 - p$ de retomber sur pile ou sur face, le rapport entre le nombre de 0 et le nombre de 1 dans la suite obtenue aura, dans les mêmes conditions, de très grandes chances d'être proche de p/q ; plus précisément, le nombre α de 0 et le nombre β de 1 seront de la forme $\alpha = pN + r, \beta = qN - r$ où r est un $o(N)$ lorsque N tend vers l'infini. De telles suites sont dites *typiques*. D'autre part, la probabilité du résultat du i-ème lancer étant indépendante de i, la probabilité d'obtenir α fois 0 et β fois 1 à des positions prescrites au bout de N lancers est $p^\alpha q^\beta$. Les suites typiques produites par un grand nombre N de lancers, celles donc que l'on est presque certain d'obtenir, ont donc toutes à peu près la même probabilité $\Pi_{p,q}^N$ d'être réalisées, probabilité dont le logarithme est de la forme

$$\log \Pi_{p,q}^N = \log(p^{pN+r} q^{qN-r}) = N(p \log p + q \log q) + o(N).$$

Les suites non typiques ayant une probabilité négligeable dès que N est assez grand, le nombre total $\mathcal{T}_{p,q}^N$ de suites typiques est environ l'inverse de cette

[2] *Cybernetics*, qui développe la représentation des phénomènes d'information et de régulation dans l'animal et la machine, paraît en 1948, l'année même de publication de l'article de Shannon.

[3] ... puis à s'en distinguer en introduisant dans les années 60, en même temps que Solomonov et Chaitin, la *théorie algorithmique de l'information* basée non sur la structure statistique d'un ensemble d'objets mais sur la longueur minimale d'un programme décrivant un objet donné prise comme caractéristique de l'information que contient ce dernier ([Ko2]).

probabilité ; autrement dit,

$$\log \mathcal{T}_{p,q}^N = N(p \log \frac{1}{p} + q \log \frac{1}{q}) + o(N).$$

La limite lorsque N tend vers l'infini,

$$H = \lim_{N\to\infty} \frac{1}{N} \log \mathcal{T}_{p,q}^N = \lim_{N\to\infty} \frac{1}{N} \log \frac{1}{\Pi_{p,q}^N} = p \log \frac{1}{p} + q \log \frac{1}{q}\,,$$

est l'*entropie de Shannon*[4]. Elle tend vers 0 lorsque le résultat du lancer est certain ($(p,q) = (1,0)$ ou $(0,1)$), c'est-à-dire dire si un lancer de la pièce ne nous apporte aucune *information* (i.e. ne nous apprend rien) ; elle est maximale (égale à 1 si les logarithmes sont en base 2) si 0 et 1 sont équiprobables ($(p,q) = (1/2, 1/2)$), auquel cas un lancer de la pièce nous donne une information maximale puisqu'aucun pronostic ne pouvait être fait a priori (voir la figure 7 de [Sh]). La propriété qui vient d'être mise en évidence, encore appelée *équipartition asymptotique de la probabilité*, est fondamentale. On peut la paraphraser ainsi : dans un jeu de pile ou face sans mémoire où p et q représentent respectivement les probabilités de pile (0) et face (1), sur les 2^N résultats possibles d'une suite de N lancers, seuls environ[5] 2^{NH} *suites typiques* ont une chance non infime d'apparaître ; de plus, ces suites typiques ont chacune approximativement la même probabilité 2^{-NH} d'être obtenue. Si par exemple $p = 0.9$ et donc $q = 0.1$, une suite typique contiendra environ 9 fois plus de 0 que de 1 et l'entropie est proche de $1/2$. Le nombre des suites typiques est donc la racine carrée du nombre total de résultats a priori possibles : si $n = 13$, seules environ 90 suites parmi les $2^{13} = 8192$ sont typiques.

Un codage économique de l'ensemble de ces suites attribuera les codes les plus courts aux suites typiques *(il faudra donc environ NH "bits" pour coder une suite (un message) de longueur N)* et des codes quelconques aux autres suites qui, en pratique, pourront être oubliées[6].

Le passage du jeu de pile ou face, c'est-à-dire d'un alphabet de deux lettres 0 et 1 muni d'une loi de probabilité (p,q), où $p \geq 0, q \geq 0, p+q = 1$, à un alphabet $A = \{A_1, \cdots, A_r\}$ de r lettres muni d'une loi de probabilité $(p_1, \cdots, p_r)$, où $p_1 \geq 0, \cdots, p_r \geq 0, \sum_{j=1}^r p_j = 1$, est immédiat : l'entropie est alors définie par

$$H(p_1, \cdots, p_r) = \sum_{j=1}^r p_j \log \frac{1}{p_j}\,\cdot$$

Cette fonction H, dont Shannon rappelle que c'est celle même (à la multiplica-

[4] précisément l'entropie par symbole, mesurée en *bits par symbole*.

[5] rigoureusement, c'est le logarithme de ce nombre qui est "équivalent à NH lorsque N tend vers l'infini."

[6] Ce résultat est connu sous le nom de *Théorème du codage de source* (Source coding theorem) ou *Théorème du codage en l'absence de bruit* (Noiseless channel coding theorem) ou simplement *Premier théorème de Shannon*. On en trouvera des illustrations dans [P]

tion près par une célèbre constante) du fameux *théorème H* de Boltzmann[7], s'annule dans les cas de certitude (une seule des probabilités p_j égale à 1, les autres égales à 0) et atteint son maximum $\log r$ en cas d'équiprobabilité ($p_1 = \cdots = p_r = 1/r$).

Nous retenons de ceci qu'étant donné un alphabet $A = \{A_1, \cdots, A_r\}$ muni d'une loi de probabilité $(p_1, \cdots, p_r)$, parmi les $2^{N \log r}$ messages formés de N lettres tirées au hasard et de façon indépendante suivant cette loi de probabilité, seuls $2^{NH(p_1,\cdots,p_r)}$ messages "typiques" ont, dès que N est assez grand, une chance non infime d'être rencontrés et de plus, le probabilité de chacun de ces messages typiques est à peu près la même, égale à $2^{-NH(p_1,\cdots,p_r)}$. Les messages non typiques peuvent ainsi être oubliés.

On peut à ce point remarquer que la définition que donne Shannon de l'entropie se réduit, *si l'on ne prend en compte que les messages typiques*, au quotient par la longueur N des messages du logarithme du nombre total de ceux-ci, c'est-à-dire au quotient par la longueur N des messages du logarithme de l'inverse de la probabilité de chacun, ... qui n'est autre que la définition que Ralph Hartley avait donné 20 ans auparavant de l'information fournie par le choix d'un message parmi un ensemble de messages supposés équiprobables !

Shannon considère également le cas plus réaliste[8] de sources de messages markoviennes dans lesquelles la probabilité d'une nouvelle lettre dépend de la lettre qui précède. Les formules sont plus compliquées mais la théorie est la même.

La deuxième idée introduite par Shannon est son *diagramme schématique d'un système général de communication* représenté dans la figure 1 au tout début de [Sh].

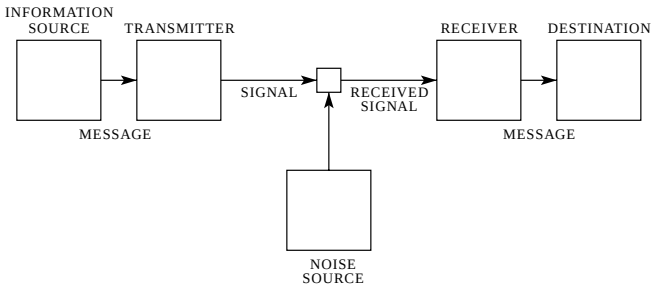


Fig. 1—Schematic diagram of a general communication system.

[7] où elle quantifie (asymptotiquement via la formule de Stirling) le degré d'ignorance (précisément le log du nombre) des micro-états correspondant à un macro-état donné (voir [Ba]). D'après [J], c'est Gibbs qui, le premier, écrit explicitement la formule probabiliste ci-dessus. La relation de la notion d'information à l'entropie thermodynamique et au second principe, proposée dès 1929 par Leo Szilard pour expliquer l'expérience de pensée du *démon de Maxwell (1867)*, est l'objet du livre *Science and Information Theory* que Léon Brillouin publie en 1956. Le thème en est l'identification entre information liée à la structure d'un système physique et *néguentropie*, c'est-à-dire une quantité qui se soustrait à l'entropie totale du système. À ce sujet, voir dans [Col] la passionnante conférence où Sergio Ciliberto, illustrant le *principe de Landauer* qui, en 1961, précisait la nature physique de l'information, montre comment les expérimentateurs "dansent aujourd'hui avec le démon de Maxwell".

[8] éloquemment illustré au début de [Sh] par l'exemple de la langue anglaise.

On sous-estimerait facilement aujourd'hui la nouveauté d'une telle conceptualisation du problème, directement issue d'un article écrit pendant la guerre sur la cryptographie[9] : les messages issus d'une *source* sont envoyés (c'est le *codage*) dans un transmetteur (l'*entrée*), puis transmis à travers un *canal* – nom général désignant aussi bien un télégraphe qu'un téléphone, un réseau Wifi, un tam-tam ou tout autre moyen de communication – juqu'à un récepteur (la *sortie*) et enfin interprétés (le *décodage*). De plus, une source de bruit interfère avec les messages transitant par le canal : en effet, comme tout intermédiaire matériel, le canal ne peut être parfait et la transmission des messages est nécessairement entachée d'une certaine proportion d'erreurs.

À la source et au canal sont associés des alphabets $A = \{A_1, \cdots, A_r\}$ pour la source, $X = \{X_1, \cdots X_a\}$ pour l'entrée (i.e. pour les messages codés que l'on veut envoyer) et $Y = \{Y_1, \cdots, Y_b\}$ pour la sortie (i.e. pour les messages reçus qu'il faudra décoder). Pour la prise en compte des erreurs, Shannon se place là encore sur le terrain probabiliste, ne retenant comme seule caractéristique du canal, en dehors des alphabets d'entrée X et de sortie Y, que les *probabilités conditionnelles* $p(y_1 \cdots y_N | x_1 \cdots x_N)$, symbole qui se lit "probabilité de $y_1 \cdots y_N$ si $x_1 \cdots x_N$" et désigne la probabilité que le message $y_1 \cdots y_N$ soit reçu si le message $x_1 \cdots x_N$ est envoyé[10]. Faisant l'hypothèse simplificatrice que les lettres successivement envoyées sont perturbées par le bruit indépendamment les unes des autres (i.e. seule la i-ème lettre envoyée x_i influe sur la i-ème lettre reçue y_i), ces probabilités conditionnelles vérifient

$$p(y_1 \cdots y_N | x_1 \cdots x_N) = \Pi_{i=1}^N p(y_i | x_i).$$

Elles sont donc parfaitement définies par les probabilités $p(Y_k|X_j)$ de recevoir la lettre Y_k si la lettre X_j est envoyée.
À cette donnée sont attachées des entropies : en effet, envoyer la lettre $x = X_j$ définit sur l'alphabet de sortie Y une loi de probabilité $y \mapsto p(y|x)$ ainsi donc qu'une entropie

$$H(Y|x) = \sum_{y \in Y} p(y|x) \log \frac{1}{p(y|x)} = \sum_{k=1}^{b} p(Y_k|X_j) \log \frac{1}{p(Y_k|X_j)} \cdot$$

Shannon introduit à ce point une nouvelle idée fondamentale : supposant l'alphabet d'entrée $X = \{X_1, \cdots, X_a\}$ muni d'une loi de probabilité $p(x)$:

$$p_1 = p(X_1), \cdots, p_a = p(X_a),$$

il définit l'*entropie condtionnelle* $H(Y|X)$ comme la moyenne[11] des entropies $H(Y|x)$:

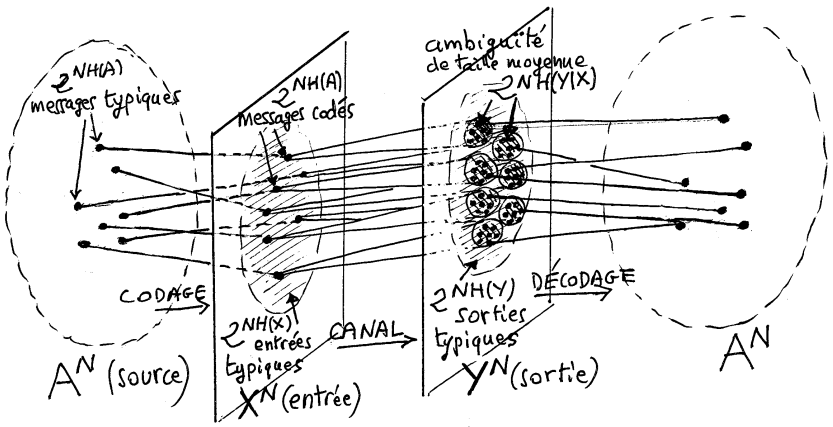
[9] *A Mathematical Theory of Cryptography* écrit en 1945 mais classifié jusqu'en 1949, date à laquelle il est publié sous le nouveau titre *Communication Theory of Secrecy Systems*.

[10] La suite $x_1 \cdots x_N$ est une suite d'éléments de X, c'est-à-dire un élément de X^N ; autrement dit, elle est de la forme $X_{j_1} \cdots X_{j_N}$. De même, la suite $y_1 \cdots y_N$, qui est un élément de Y^N, est de la forme $Y_{k_1} \cdots Y_{k_N}$.

[11] dans le langage des probabilités l'*espérance de la variable aléatoire* $x \mapsto H(Y|x)$.

$$H(Y|X) = \sum_{x\in X} p(x)H(Y|x) = \sum_{j=1}^{a} p_j H(Y|X_j)$$
$$= \sum_{j=1}^{a}\sum_{k=1}^{b} p(X_j)p(Y_k|X_j)\log\frac{1}{p(Y_k|X_j)}\cdot$$

Il montre alors que sous les hypothèses d'indépendance qui ont été faites, un long message $x_1\cdots x_N$ envoyé dans le canal aura *en moyenne* une égale probabilité de donner à la réception l'un parmi $2^{NH(Y|X)}$ messages $y_1\cdots y_N$. La figure 10 de [Sh], qui ne retient que les messages typiques, résume parfaitement la situation : les probabilités conditionnelles y sont représentées comme un cône dont le sommet $x_1\cdots x_N$ appartient à l'ensemble X^N des messages (et même à celui des messages typiques) de N lettres écrits dans l'alphabet d'entrée X et dont la base, contenue dans l'ensemble Y^N des mots de N lettres $y_1\cdots y_N$ écrits dans l'alphabet de sortie Y, est formée des $y_1\cdots y_N$ typiques[12] pouvant être reçus.
Le problème du codage, une fois cette moyenne effectuée, y apparaît comme un problème d'empilement d'oranges dans une boîte de taille fixée, ce que tente de représenter la figure ci-dessous :



Étant donnée une source de messages A et un canal $(X \to Y)$, comment coder les $2^{NH(A)}$ messages typiques de façon à ce que chacun des sous-ensembles de taille (moyenne) $2^{NH(Y|X)}$ pouvant provenir de chacun des $2^{NH(A)}$ messages codés envoyés soient disjoints, sachant que l'ensemble est contenu dans la boîte des messages typiques dans Y^N, de contenance $2^{NH(Y)}$? Une condition nécessaire est évidemment que

$$2^{NH(A)} \times 2^{NH(Y|X)} < 2^{NH(Y)}, \quad \text{c'est-à-dire} \quad H(A) < H(Y) - H(Y|X).$$

[12] pour donner un sens à cette dernière asertion, il faut remarquer que la donnée de la loi de probabilité $p(x)$ sur X et des probabilités conditionnelles $p(y|x)$ caractérisant le canal définit sur Y une loi de probabilité $q(y) = \sum_{x\in X} p(x)p(y|x)$ et donc une entropie $H(Y)$.

Le même raisonnement peut être tenu en échangeant les rôles de l'entrée et de la sortie, c'est-à-dire en raisonnant sur les probabilités conditionnelles $p(x|y)$ qu'ayant reçu y on ait envoyé x : les "oranges" représentant l'ambiguïté moyenne (en anglais "equivocation") sont maintenant du côté de l'entrée et représentent pour chaque message reçu $y_1 \cdots y_N$ le nombre moyen, de l'ordre de $2^{NH(X|Y)}$, de messages susceptibles d'avoir été envoyés. Interprétée en termes de distance au centre des oranges, la condition de disjonction, qui est maintenant

$$2^{NH(A)} \times 2^{NH(X|Y)} < 2^{NH(X)}, \quad \text{c'est-à-dire} \quad H(A) < H(X) - H(X|Y),$$

n'est pas sans rapport avec le fait que de nombreuses fautes de frappe n'empêchent pas de reconnaître sans ambiguïté un texte pourvu que la forme altérée ressemble plus au texte initial qu'à tout autre texte admissible.

Il se trouve que les conditions de disjonction à la sortie ou à l'entrée coïncident et fournissent à Shannon sa définition de l'*information mutuelle* :

$$\begin{aligned} I(Y,X) &= H(Y) - H(Y|X) = H(X) - H(X|Y) = I(X,Y) \\ &= \sum_{j=1}^{a} \sum_{k=1}^{b} \pi(X_j, Y_k) \log \frac{\pi(X_j, Y_k)}{p(X_j) q(Y_k)} \cdot \end{aligned}$$

Le numérateur $\pi(X_j, Y_k) = p(X_j)p(Y_k|X_j)$ du logarithme est la probabilité de l'évènement conjoint "avoir envoyé X_j et reçu Y_k" alors que le dénominateur $p(X_j)q(Y_k)$ est cette même probabilité conjointe dans le cas où l'entrée et la sortie seraient indépendantes (i.e. n'auraient aucune relation de causalité). Cette formule est interprétée aujourd'hui comme l'*entropie relative de Kullback-Leibler*[13] de ces deux lois de probabilités sur l'espace produit $X \times Y$.

Notons le rôle crucial de la moyenne dans la définition de $H(Y|X)$: alors que l'entropie conditionnelle $H(Y|X)$ est toujours inférieure ou égale à $H(Y)$, autrement dit que l'information mutuelle est toujours positive ou nulle[14], ne s'annulant que si X et Y sont indépendants, il se peut que pour un x donné, $H(Y|x)$ soit supérieure à $H(Y)$, par exemple lorsque la loi de probabilité $p(y|x)$ rend équiprobables tous les éléments y de Y et donc maximale l'entropie $H(Y|x)$.

D'autres notions d'information avaient précédé celle de Shannon, en particulier celle, locale dans l'espace des densités de probabilités, définie par Ronald Fisher au milieu des années vingt comme outil d'estimation statistique de la pertinence d'un modèle. C'est d'ailleurs à cette problématique que se rattache l'entropie de Kullback-Leibler que nous venons d'évoquer.

Il ne reste plus qu'à définir la capacité C du canal comme le sup des informations mutuelles pour toutes les lois de probabilité sur l'alphabet X, ce qui conduit à la célèbre condition [15]

$$H(A) < C : \quad \text{entropie de la source inférieure à la capacité du canal,}$$

[13] introduite au début des années cinquante par deux cryptanalystes américains Solomon Kullback et Richard Leibler qui travaillaient pour la NSA.

[14] C'est l'*inégalité de Shannon*, conséquence de la convexité de la fonction $x \mapsto \log \frac{1}{x}$.

[15] les quantités sont exprimées en bits par symbole ou bien, multipliées par le nombre de symboles transmis par seconde, en bits par seconde.

qui rassemble le *théorème du codage de source* et le *théorème du codage en présence de bruit* (Noisy channel coding theorem).

Il est remarquable que cette condition nécessaire soit également suffisante pour qu'il existe un code permettant de transmettre des messages avec une probabilité d'erreur arbitrairement petite pourvu que les messages soient assez longs (N assez grand).

L'idée de la démonstration, donnée par Shannon, précisée et généralisée aux sources ergodiques par ses successeurs, Robert Fano, Amiel Feinstein, Brockway MacMillan, Leo Breiman, Alexandre Khinchine (voir [Kh, Bi]), ... et dont le meilleur exposé se trouve dans le livre [CT], est un remarquable calcul de moyenne des probabilités d'erreurs sur un ensemble de codes *choisis au hasard*. En voici une esquisse : fixons la longueur N des messages provenant de la source ainsi qu'un *canal stationnaire sans mémoire*, c'est-à-dire un ensemble d'entrées[16] X^N muni de probabilités $p(x_1 \cdots x_N) = \Pi_{i=1}^N p(x_i)$, un ensemble de sorties Y^N et des probabilités conditionnelles $p(y_1 y_2 \cdots y_N | x_1 x_2 \cdots x_N) = \Pi_{i=1}^N p(y_i|x_i)$, donc des probabilités $q(y_1 \cdots y_N) = \Pi_{i=1}^N q(y_i) = \Pi_{i=1}^N (\sum_{x \in X} p(x) p(y_i|x))$ sur Y^N. Nous ne nous intéresserons qu'aux messages typiques issus de la source c'est-à-dire, en négligeant les epsilons, à $2^{NH(A)}$ messages que l'on peut considérer comme équiprobables. Un code est un plongement quelconque dans X^N de ces $2^{NH(A)}$ messages, c'est-à-dire un ensemble de $2^{NH(A)}$ messages codés formant les lignes d'une matrice

$$\mathcal{C} = \begin{pmatrix} x_1(1) & x_2(1) & \cdots & x_N(1) \\ \cdots & \cdots & \cdots & \cdots \\ x_1(2^{NH(A)}) & x_2(2^{NH(A)}) & \cdots & x_n(2^{NH(A)}) \end{pmatrix}.$$

1) On munit l'ensemble de ces codes de la loi de probabilité

$$Pr(\mathcal{C}) = \prod_{w=1}^{2^{NH(A)}} \prod_{i=1}^{N} p\big(x_i(w)\big),$$

ce qui implique que, choisissant au hasard un tel code, la probabilité est grande que tous les messages codés appartiennent au sous-ensemble typique de X^N et, étant donné la définition de la probabilité sur la sortie Y^N, que tous les messages transmis appartiennent à l'ensemble typique de Y^N. Ici aussi, nous ne nous préoccuperons pas des epsilons et oublierons les codes improbables n'ayant pas cette propriété, ce qui nous laisse avec environ $2^{NH(X)}$ entrées et $2^{NH(Y)}$ sorties.

2) Le code choisi et les caractéristiques du canal sont supposés connus de l'envoyeur et du destinataire.

3) Un message reçu $y_1 \cdots y_N$ est décodé comme provenant de w si ce dernier est l'unique message source tel que la paire $(x(w), y)$ soit *conjointement typique*,

[16] choisir pour les messages codés la même longueur N que celle des messages source n'est pas important mais simplifie les notations. En pratique, une "lettre" de l'alphabet X pourra être un "mot" formé par exemple de 0's et de 1's.

ce qui signifie d'une part que $x(w)$ et y sont typiques, respectivement dans X^N et Y^N, et que le couple est typique dans $X^N \times Y^N$ muni de la loi de probabilité $p(x_1 \cdots x_N, y_1 \cdots y_N) = p(x_1 \cdots x_N)p(y_1 \cdots y_N | x_1 \cdots x_N)$. Dans tous les autres cas, il est interprété comme une erreur. Un calcul simple montre alors que, calculée à la fois sur l'ensemble des messages et l'ensemble des codes, l'espérance de la probabilité d'erreur tend vers 0 lorsque N tend vers l'infini dès que $H(A) < I(X,Y)$, ce qui montre l'existence lorsque cette condition est remplie d'un code ayant une probabilité d'erreur moyenne arbitrairement petite. La conclusion s'ensuit facilement

Heuristiquement, le fait que le codage se fasse au hasard implique la non corrélation (au sens probabiliste) de 2 messages reçus provenant de 2 messages codés distincts, ce qui rend improbable leur coïncidence dès qu'il y a suffisamment de place à la sortie (i.e. dès qu'il y a effectivement la place d'empiler toutes les oranges).

Remarquons pour finir que ce n'est pas la condition $H(A) < C$ qui est la plus souvent citée comme représentant le résultat principal de Shannon mais son avatar "continu" pour un bruit gaussien

$$C = W \log \left(1 + \frac{P}{N}\right)$$

décrit dans le chapitre IV de [Sh] (W est la largeur de bande passante du canal et P/N le rapport signal sur bruit). Or cette formule apparaît – mais sans l'appareil théorique qui fait toute la force du travail de Shannon – dans un certain nombre de travaux contemporains, tels ceux de Norbert Wiener, William G. Tuller, H. Sullivan mentionnés par Shannon, mais aussi Stanford Goldman, Charles W. Earp, et deux ingénieurs français André G. Clavier et Jacques Laplume (voir [FR] et la conférence d'Olivier Rioul dans [Col]).

Restée longtemps théorique[17], la limite de Shannon est aujourd'hui pratiquement atteinte par les turbocodes développés dans les années 1990 par Claude Berrou et son équipe : l'art de la construction d'un code correcteur consiste en la recherche de la façon la plus efficace d'introduire la *redondance* minimale qui permette de décoder les messages avec une très faible probabilité d'erreur. Dans [Be], Claude Berrou explique qu'entrelaçant deux petits "codes convolutifs", les turbocodes ont une certaine analogie avec les grilles de mots croisés : c'est en reprenant successivement et plusieurs fois définitions horizontales et définitions verticales que l'on finit par lever l'ambiguïté[18].

Il est un cas cependant où le code le plus simple s'approche de la borne optimale : dans une courte note de 1949 intitulée *A case of efficient coding*

[17] Shannon remarque explicitement que construire un code proche de l'optimal en suivant la démonstration d'existence est impraticable en raison de la taille de N.

[18] C'est bien l'effet "turbo" qui consiste en une réutilisation d'une partie de l'énergie cinétique contenue dans les gaz d'échappement d'un moteur pour faire tourner un compresseur servant à augmenter l'apport d'oxygène dans la chambre de combustion et donc également la puissance.

for a very noisy channel, Shannon montre en effet que le code consistant en la répétition un grand nombre K de fois de chaque lettre du message source (précisément, $A = \{0,1\}, X = Y = \{0,1\}^K$), assorti d'un décodage à la majorité dans chaque groupe de K symboles reçus, est proche d'être optimal lorsque la probabilité d'erreur est proche de $1/2$. Au contraire, si cette probabilité est faible, un tel codage n'est plus du tout optimal.

Il n'aura pas, je pense, échappé aux lecteurs que sans de tels codes correcteurs la majorité des instruments que nous utilisons tous les jours n'existeraient tout simplement pas. Une évocation de l'histoire et un état des lieux des recherches actuelles[19] se trouve dans les conférences du colloque organisé à Paris pour célébrer le centenaire de la naissance de Claude Shannon [Col]. Bel exemple d'un théorème de pure mathématique dont la prédiction (il existe un code ...) se réalise cinquante années après son énoncé.

Merci à Qiaoling Wei pour de nombreuses discussions lors de la préparation d'une conférence sur Shannon au Musée des Sciences et Techniques de Pékin, à Nathan Hara pour son intérêt constant et ses questions provocantes, à Frédéric Barbaresco pour la référence [FR], à Daniel Bennequin pour la référence [J], des précisions sur l'information de Fisher et ses amicales corrections, à Marc Serrero pour la référence [Ba] et de non moins amicales remarques. Merci enfin à Sébastien Gouezel pour sa lecture critique et constructive d'une première version.

References

[Sh] C. Shannon, *A Mathematical theory of Communication*, The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948

[Ba] Claude Balian, *Du microscopique au macroscopique*, Cours de l'Ecole Polytechnique, Ellipse 1982. Traduction anglaise sensiblement modifiée *From Microphysics to Macrophysics*, Springer 1991

[Be] Claude Berrou, *Les turbocodes*

`http://www.futura-sciences.com/magazines/high-tech/infos/dossiers/d/telecoms-turbocodes-366/`

[Bi] P. Billingsley, *Ergodic Theory and Information*, John Wiley & sons, 1965

[Col] Colloque Shannon 100, IHP, 26 au 28 octobre 2016

`https://www.youtube.com/playlist?list=PL9kd4mpdvWcDMCJ-SP72HV6Bme6CSqk_k`

[19] à l'importante exception près de tout ce qui concerne l'*information quantique* où l'existence de l'*intrication* (entanglement) joue un rôle fondamental : en particulier, l'*entropie de Von Neumann* remplaçant celle de Shannon, l'entropie conditionnelle peut être négative !

[CT] T.C. Cover & J.A. Thomas, *Elements of Information Theory*, Wiley 1991

[FR] P. Flandrin & O. Rioul, *Laplume sous le masque*, Académie des Sciences, Octobre 2016 `http://www.academie-sciences.fr/fr/Evolution-des-disciplines-et-histoire-des-decouvertes/laplume-sous-le-masque-patrick-flandrin-et-olivier-rioul.html`

[J] E.T. Jaynes, *Gibbs vs Boltzmann Entropies*, American Journal of Physics, Vol. 33, N^{0}3, 391-398, May 1965

[Ka] A. Katok, *Fifty Years of Entropy in Dynamics: 1958-2007*, Journal of Modern Dynamics, Volume 2, $n^0$4 (2007) 545-596

[Kh] A. Khinchine, *Mathematical Foundations of Information Theory*, Dover 1957

[Ko1] A. N. Kolmogorov, *A new metric invariant of transitive dynamical systems and automorphisms of Lebesgue spaces*, Dokl. Akad. Nauk SSSR 119,5 (1958), 861-864.

[Ko2] A. N . Kolmogorov, *Combinatorial foundations of information theory and the calculus of probabilities*, Uspekhi Mat. Nauk 1983, qui est la rédaction de la conférence faite à Nice en 1970 au Congrès international des mathématiciens.

[P] G.Peyré, *Claude Shannon et la compression des données*, Images des maths, septembre 2016 `http://images.math.cnrs.fr/Claude-Shannon-et-la-compression-des-donnees.html`

[Se] J. Segal, *Le Zéro et le Un, Histoire de la notion scientifique d'information au 20ème siècle*, Syllepse 2003