Groupe de travail pour élèves de lycée

Concentration de la mesure

par

Yann Ollivier

(Texte produit et tapé par Xavier Caruso)

Le 8 juin 2003

Table des matières

1 Mesure

1.1 Pile ou face ?

Vous connaissez certainement tous le jeu de pile ou face : il s'agit de lancer une pièce en l'air avec une certaine dextérité et de regarder de quel côté elle retombe ou, au choix, de quel côté elle est rattrapée. Ce sont ces côtés qui s'appellent l'un *pile* et l'autre *face*.

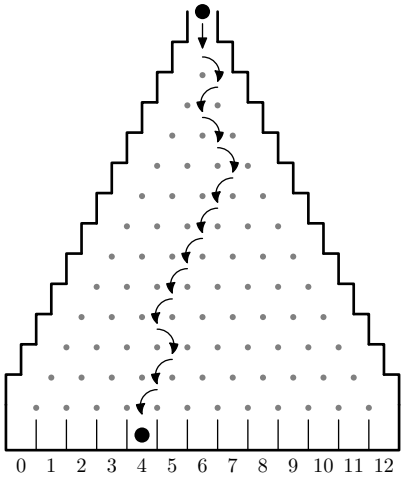
Supposons donc, qu'un jour où le ciel nous verse un jour noir plus triste que les nuits, l'on décide de jouer à ce jeu, puis de recommencer, puis de recommencer à nouveau, et ainsi de suite. Comme on est méthodique, on note de surcroît tous les résultats obtenus et on compte le nombre de fois que la pièce est retombée sur le côté *pile* après les N premiers lancers.

On obtient comme cela une suite de nombres, disons X_N, évidemment croissante et de façon tout aussi évidente, on voit que pour passer de X_{N-1} à X_N, on ajoute soit 0, soit 1 et ce de façon aléatoire et équiprobable.

Bien sûr, si l'on arrive à trouver une autre personne pour jouer au même jeu, avec une autre pièce, elle va presqu'à coup sûr ne pas écrire la même suite X_N. Donc, pour sûr, on ne peut pas calculer certainement X_N. Par contre, ce que l'on peut calculer, c'est la probabilité que X_N vaille un certain entier k compris entre 0 et N. Ce calcul n'est pas compliqué : il y a 2^N issues possibles au bout de N lancers, et parmi elles seulement C_N^k aboutissent à obtenir $X_N = k$. Ainsi[1] :

$$\mathrm{Prob}\,(X_n = k) = \frac{\mathrm{C}_N^k}{2^N}$$

Pour terminer ce paragraphe, mentionnons l'existence d'une façon plus visuelle de présenter le jeu précédent. Elle est illustrée par le dessin suivant :



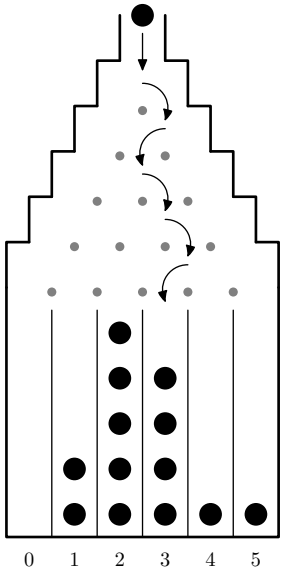
Ce dessin représente un plateau dont la forme est plus ou moins triangulaire, sur lequel sont disposés des bâtonnets. On lache une bille en haut et, à chaque étage, la bille bute contre un bâtonnet, ce qui la contraint à choisir entre tomber à gauche ou à droite. C'est notre tirage de pile ou face. Par exemple sur notre dessin, le premier bâtonnet l'a fait tomber

[1] On rappelle que C_N^k est le nombre de manières de choisir k éléments parmi N et vaut $\frac{N!}{k!(N-k)!}$ avec $n! = 1 \times 2 \times \cdots \times n$.

à droite, le second à gauche et ainsi de suite. Finalement avec N étages (ici $N = 12$), la bille est recueillie et le compartiment dans lequel elle tombe correspond directement au nombre de déplacements vers la droite qui ont été faits pendant la descente. Pour notre exemple, on voit que c'est 4. Maintenant, il suffit d'identifier les déplacements vers la droite au côté *pile* de la pièce pour voir que l'on considère exactement le même problème.

Le premier avantage évident de cette seconde formulation est qu'elle est plus agréable à mettre en place et à expérimenter et que, de surcroît, elle donne directement le résultat. Un corollaire de cela est qu'il est possible d'observer facilement les résultats d'un nombre important d'expériences : il suffit de lacher un nombre important de billes, et si l'on ne veut pas avoir à récupérer la bille en bas à chaque fois, il va suffire d'allonger les compartiments de recueil afin de permettre à plusieurs billes de s'y entasser.

Par exemple, pour $N = 5$, on peut obtenir :



Le fait que X_N ne soit pas à l'avance déterminé correspond simplement au fait qu'au final toutes les billes ne tombent pas dans la même rigole. La probabilité que l'on a calculée précédemment $\text{Prob}\,(X_N = k)$, correspond à la proportion de billes qui tombent dans la rigole k. Ainsi cette petite expérience, relativement simple à réaliser, donne un modèle empirique de la répartition.

C'est cette répartition qui va nous intéresser dans toute la suite. Ces répartitions que l'on vient d'évoquer se traduisent mathématiquement naturellement par des mesures. Il y a principalement deux sortes de mesures : les mesures discrètes et les mesures continues.

1.2 Les mesures discrètes

Une *mesure discrète* est la donnée d'une famille (finie ou non) de couples (x_i, p_i). Intuitivement, x_i est le résultat d'une expérience aléatoire (ici, ce sera toujours un réel) et p_i (noté parfois $p\,(x_i)$) est la probabilité que le résultat x_i survienne. On a ainsi les contraintes $p_i \geqslant 0$ et

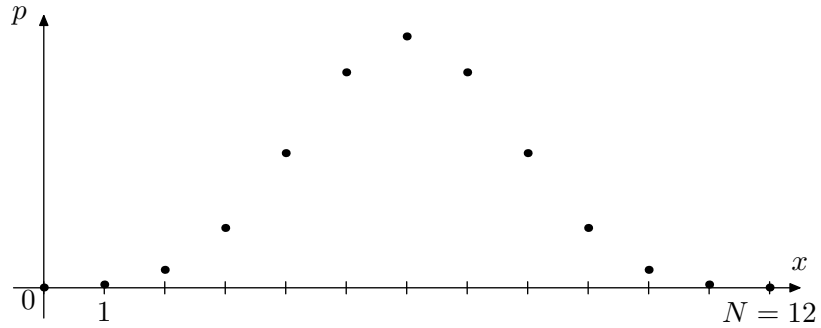
$$\sum_i p_i = 1$$

On n'a pour l'instant pas dit dans quel ensemble on prenait nos indices i. En général, on les choisit dans ce que l'on appelle l'*univers*, ensemble que l'on note traditionnellement

Ω, mais cela n'a que peu d'importance finalement. De notre côté, on ne précisera jamais cet ensemble et on espère que lorsque l'on parlera de sommation, il n'y aura pas d'ambiguïté. Bien sûr, précédemment, la somme était étendue à tous les i possibles.

Il est une façon commode de représenter ces mesures. Si les x_i sont tous des réels, on peut porter en abscisse les x_i et en ordonnée les probabilités correspondantes.

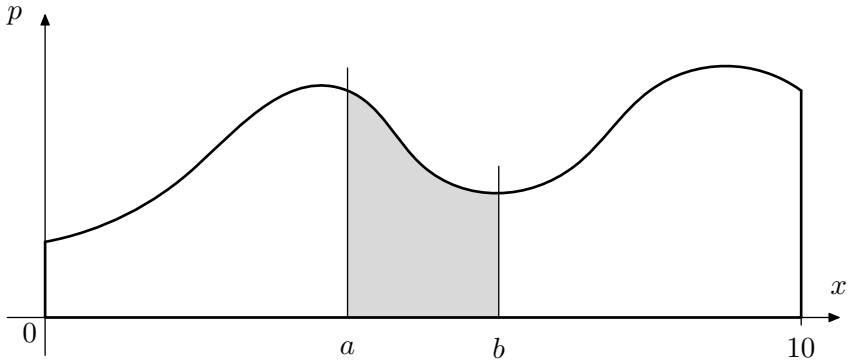
Expliquons cela sur l'exemple précédent du tirage répété de pile ou face. Les issues possibles, c'est-à-dire les x_i, sont les entiers de l'intervalle $[0, N]$. On pose donc $x_i = i$, et i décrit par la suite toujours l'ensemble $\{0, 1, \ldots, N\}$. On a déjà calculé les probabilités correspondantes et on rappelle que l'on avait trouvé $p_i = \frac{C_N^i}{2^N}$. On aboutit donc naturellement au diagramme suivant (pour $N = 12$) :



Lorsque l'on augmente N, on voit bien que ces points vont se rapprocher, probablement pour former une courbe. C'est cette courbe en fait que l'on aimerait décrire, mais celle-ci ne correspondra sans doute pas à une mesure discrète telle que l'on vient de la définir. Il nous faut donc introduire des mesures plus générales.

1.3 Les mesures continues

Cette fois-ci, l'ensemble des issues possibles (donc ce qui était avant l'ensemble des x_i) est un intervalle (ou une réunion d'intervalles) de $\mathbb{R}$, et en fait très souvent $\mathbb{R}$ tout entier. La distribution de probabilité est alors donnée par une fonction (continue) définie sur cet intervalle. Par exemple :



Il faut maintenant comprendre à quoi correspond cette fonction. Le nombre $p(x)$ ne peut bien sûr pas désigner la probabilité de tirer x : si l'on peut obtenir tous les réels entre 0 et 10 par exemple, il est légitime de penser qu'il n'y a aucune chance de tomber au hasard sur π ou sur $\sqrt{2} + e$.

Il faut plutôt interpréter le graphe de la façon suivante. Si l'on choisit un réel x strictement compris entre 0 et 10, et si l'on considère $\mathrm{d}x$ un infiniment petit[2], la probabilité de tirer un nombre compris entre x et $x + \mathrm{d}x$ vaut $p(x)\,\mathrm{d}x$.

Quelle va être la probabilité de tirer un nombre qui tombe dans l'intervalle $[a, b]$ $(a < b)$? Il faut pour cela sommer tous les $p(x)\,\mathrm{d}x$, pour x variant entre a et b, et c'est exactement l'objet d'une intégrale. Plus précisément :

$$\mathrm{Prob}([a,b]) = \int_a^b p(x)\,\mathrm{d}x$$

Finalement les contraintes que doit vérifier la fonction p sont :

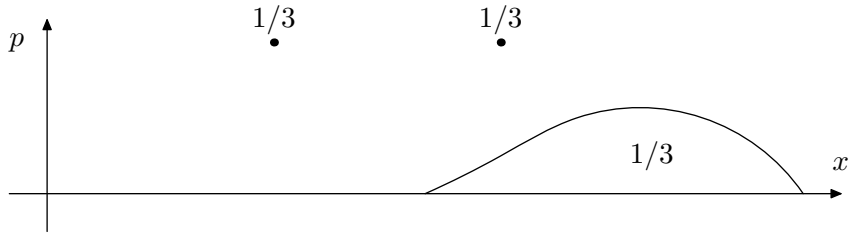
1. $p(x) \geqslant 0$, $\forall x \in \mathbb{R}$
2. $\int_{\mathbb{R}} p(x)\,\mathrm{d}x = 1$

On peut voir ce qui précède comme un jeu de fléchettes. On lance aléatoirement et de façon équiprobable une fléchette dans la partie située entre la courbe et l'axe des abscisses (délimitée en gras sur la figure), puis on regarde l'abscisse du point atteint par la fléchette et c'est ce nombre le résultat du tirage. Puisqu'il est passablement évident que la probabilité d'atteindre une zone est proportionnelle à l'aire, on justifie *a posteriori* la formule intégrale pour $\mathrm{Prob}([a,b])$. Si la courbe est plus haute à certaines abscisses, la probabilité de tirer un nombre proche des abscisses en question est également plus grande.

Il devient ainsi évident que l'on n'a aucune chance de tomber pile sur la ligne d'abscisse a, tout autant que l'on a aucune chance de tomber pile sur celle d'abscisse b. Par contre, au vu du dessin, on a plus de chances de tomber « proche » de a que de b.

Une mesure continue classique est par exemple la mesure uniforme sur $[0, 1]$; elle correspond à la fonction constante, égale à 1, définie sur l'intervalle $[0, 1]$. Pour cette mesure, tous les réels compris entre 0 et 1 ont moralement autant de chances de sortir. Lorsque l'on demande un nombre aléatoire à un ordinateur, il le tire en général selon cette mesure.

Pour finir, disons qu'il est possible de combiner mesures discrètes et mesures continues. Un dessin est sans doute plus clair que des longues explications :



De façon générale la contrainte à imposer est :

$$\sum_i p_i + \int_{\mathbb{R}} p(x)\,\mathrm{d}x = 1$$

Finalement, si A est une partie de $\mathbb{R}$, la probabilité de tirer un élément de A est :

$$\mathrm{Prob}(A) = \sum_{i\,/\,x_i \in A} p_i + \int_A p(x)\,\mathrm{d}x$$

[2]Nous n'allons pas donner de définitions précises. Une façon de le voir est de se dire que c'est un nombre qui tend vers 0.

1.4 Convergence des mesures

Ce que l'on souhaite dire, maintenant, c'est que nos courbes correspondant aux mesures discrètes du tirage répété de pile ou face, ressemblent de plus en plus à une certaine courbe qu'il nous reste à déterminer. On a donc besoin de définir précisément ce que peut signifier « ressembler de plus en plus ». C'est l'objet de ce paragraphe.

On considère une suite de mesures (discrètes ou continues ou les deux) p_n et on veut donner une définition à la phrase « la suite (p_n) converge vers la mesure p ». La première chose qui nous vient à l'esprit est sans doute de dire que (p_n) converge vers p si pour tout réel a, les probabilités d'obtenir exactement a pour chacune des mesures p_n forment une suite qui converge vers la probabilité d'obtenir a pour la mesure p. Mais cela ne peut marcher. En effet, pour une mesure continue, la probabilité d'obtenir a est toujours nulle, et donc si l'on prenait la définition précédente, toute suite de mesures convergeant par exemple vers la mesure uniforme convergerait également vers toute mesure continue.

Pour éviter le problème d'un point isolé, on pourrait prendre des intervalles ouverts et dire qu'une suite de mesures (p_n) converge vers la mesure p si pour tout intervalle $]a,b[$, la mesure $p_n(]a,b[)$ tend vers $p(]a,b[)$. Cela n'est pas très bon non plus. Par exemple, si p_n est la mesure discrète qui donne masse 1 à $1/n$, on a envie de dire que p_n tend vers la mesure discrète p qui donne masse 1 à 0 ; mais par exemple pour $]a,b[=]0,1[$, on a $p_n(]0,1[)$ qui vaut toujours 1 alors que $p(]0,1[)=0$.

Il va donc falloir être un peu plus sioux. Dans l'essai précédent, le problème était que l'appartenance à un intervalle est un phénomène discontinu : tous les nombres $1/n$ appartiennent à $]0,1[$ alors que leur limite 0 n'y appartient pas. On aurait envie de dire que les nombres $1/n$ appartiennent « de moins en moins » à $]0,1[$ à mesure qu'ils s'approchent du bord. On peut mesurer cette appartenance par une fonction continue qui vaudrait 1 sur un intervalle comme $[0,1;0,9]$, qui vaudrait 0 en-dehors de $[-0,1;1,1]$ et qui varierait continuement de 0 à 1 sur $[-0,1;0,1]$ et sur $[0,9;1,1]$. Si φ est cette fonction de lissage, la probabilité de tomber dans $[0,1]$ « mesurée par φ » est maintenant :

$$\int_{\mathbb{R}} \varphi(x)p(x)\mathrm{d}x$$

Finalement, on dit que la suite de mesures (p_n) *converge* vers la mesure p si pour toute fonction *continue* φ (c'est la continuité qui est vraiment importante) à valeurs entre 0 et 1, les probabilités « mesurées par φ » selon les p_n forment une suite qui converge (au sens classique) vers la probabilité « mesurée par φ » pour le mesure p. Autrement dit, si

$$\int_{\mathbb{R}} \varphi(x)p_n(x)\mathrm{d}x \longrightarrow \int_{\mathbb{R}} \varphi(x)p(x)\mathrm{d}x$$

pour toute fonction φ continue à valeurs entre 0 et 1.

On laisse au lecteur le soin d'écrire les définitions similaires pour des mesures discrètes (ou des combinaisons discrètes/continues), en remplaçant les intégrales par des sommes.

Avant de passer à la suite, donnons peut-être un exemple de convergence. Regardons la mesure discrète p_n définie de la façon suivante : pour k un entier compris entre 0 et n, la probabilité de tirer $\frac{k}{n}$ est toujours $\frac{1}{n+1}$. On a déjà atteint une probabilité totale égale à 1 ; il n'est donc pas nécessaire de dire ce qui se passe pour les autres réels, ces nombres n'ont automatiquement aucune chance d'être tirés.

Il est remarquable de constater que la suite de mesures (p_n) converge vers une mesure continue, précisément vers la mesure uniforme sur $[0,1]$ définie précédemment. En effet,

prenons a et b deux réels compris entre 0 et 1 vérifiant en outre $a < b$. Pour p_n, la probabilité de tomber entre a et b est $1/(n+1)$ fois le nombre de fractions de la forme $\frac{k}{n}$ comprises entre a et b. Si l'on ne veut pas calculer ce nombre précisément, on peut toujours donner une approximation :

$$|\mathrm{Prob}_{p_n}([a,b]) - (b-a)| \leqslant \frac{1}{n+1}$$

ce qui prouve que, même sans le coup du lissage par fonctions continues, la suite des $\mathrm{Prob}_{p_n}(X \in [a,b])$ converge vers $b-a$, ce qui est bien ce que l'on voulait.

Finalement, on se rend compte en réfléchissant un peu que ce résultat n'est pas du tout surprenant et même vraiment bienvenu. De fait, on a un résultat analogue un peu plus général : on peut écrire n'importe quelle mesure continue comme limite de mesures discrètes. Prenons p une mesure continue décrite donc par une fonction p. On suppose que le support de p est l'intervalle fermé $[0,1]$ et que p y est continue (et donc bornée). Maintenant pour tout entier n, on définit les mesures p_n en copiant sur ce que l'on a fait précédemment : la probabilité selon p_n de tirer $\frac{k}{n}$ est égale à $\frac{1}{n}p\left(\frac{k}{n}\right)$. Le problème est que l'on n'obtient pas forcément ici une mesure, la condition $\sum p_i = 1$ n'étant pas obligatoirement vérifiée. Qu'à cela ne tienne, on *renormalise*, c'est-à-dire que l'on divise chacun des p_i par $\sum p_i$ de sorte que la nouvelle somme fasse bien 1. Dans ces conditions les p_n forment une suite de mesures qui converge vers la mesure d'origine p. Ainsi étant donnée une mesure continue p, il est toujours possible de la discrétiser[3].

1.5 La loi des grands nombres (v.1)

On reprend notre tirage successif à pile ou face et notre variable aléatoire X_N (nombre de « pile ») qui lui était associée. On note p_n la mesure discrète déterminée par X_N ; on rappelle que p_n est définie par :

$$\mathrm{Prob}(X_N = k) = \frac{\mathrm{C}_N^k}{2^N}$$

On se demande maintenant si p_n ne convergerait pas vers une certaine mesure, ce qui permettrait de donner une loi générale valable pour un nombre conséquent de tirages.

Ce n'est pas le cas pour des raisons bêtes : la variable X_N prend ses valeurs entre 0 et N, intervalle qui dépend de N. Ce serait mieux que toutes les variables que l'on désire comparer prennent leurs valeurs dans le même ensemble ; c'est pour cela qu'il est plus judicieux de considérer $\frac{X_N}{N}$ qui correspond à la proportion, et non au nombre, de piles obtenues.

Maintenant, regardons $\frac{X_N}{N}$. Les graphes que l'on a dessinés semblent suggérer que les valeurs obtenues par cette variable se situent autour de $\frac{1}{2}$. Et de fait, on a le théorème précis suivant :

Théorème 1 (Loi des grands nombres (v.1))**.** *Avec les notations précédentes, la suite des mesures correspondant aux variables $\frac{X_N}{N}$ converge vers la mesure $\delta_{1/2}$, où $\delta_{1/2}$ est ce que l'on appelle la* masse de Dirac *en $1/2$, cela signifiant qu'un tirage selon cette mesure répond $\frac{1}{2}$ à coup sûr.*

[3] C'est d'ailleurs ce que font implicitement les ordinateurs pour générer des nombres aléatoires.

Ceci est un exemple de convergence de mesures discrètes vers une autre mesure discrète. Sur le graphe, les mesures associées à $\frac{X_N}{N}$ forment un pic de plus en plus haut et de plus en plus étroit autour de $1/2$.

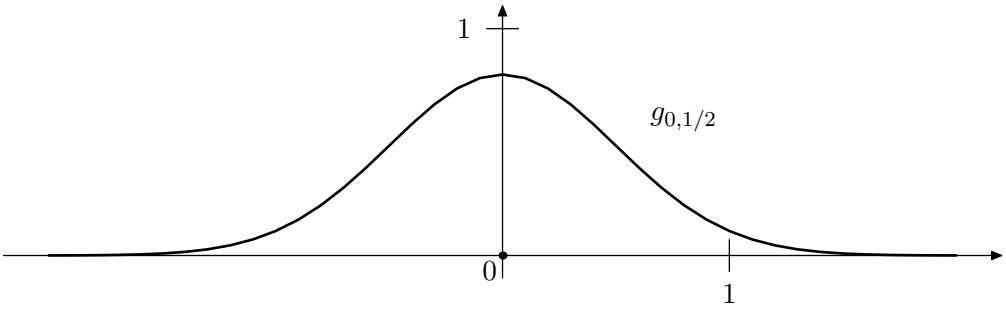
On a vu sur les dessins précédents que la courbe obtenue avait certes un pic en $\frac{N}{2}$ mais avait également une certaine forme de cloche que l'on aimerait retrouver sur la limite. En divisant par N, on a trop tassé les choses. Comme on va le voir avec le théorème suivant, la cloche a une largeur d'environ $\sqrt{N}$ et il fallait donc diviser par $\sqrt{N}$ pour pouvoir la voir jusqu'à l'infini. Mais si l'on se contente de regarder $\frac{X_N}{\sqrt{N}}$, on retombe sur le même problème que précédemment : les mesures ne prennent pas leurs valeurs dans le même intervalle. Pour tout résoudre simultanément, on regarde le quotient $\frac{X_N - N/2}{\sqrt{N}}$. On a alors le théorème suivant :

Théorème 2 (Théorème de la limite centrale (`v.1`))**.** *Avec les notations précédentes, la suite de mesures correspondant aux variables* $\frac{X_N - N/2}{\sqrt{N}}$ *converge vers la mesure continue définie sur* $\mathbb{R}$ *par la fonction :*

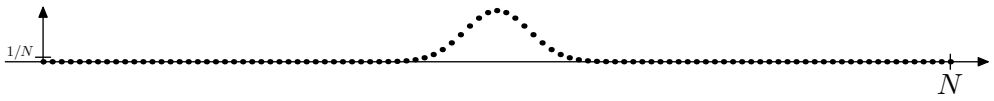
$$g_{0,1/2}(x) = \frac{1}{\sqrt{\pi/2}} \exp\left(-2x^2\right)$$

où la notation exp *désigne la fonction exponentielle classique :* $\exp(x) = e^x$.

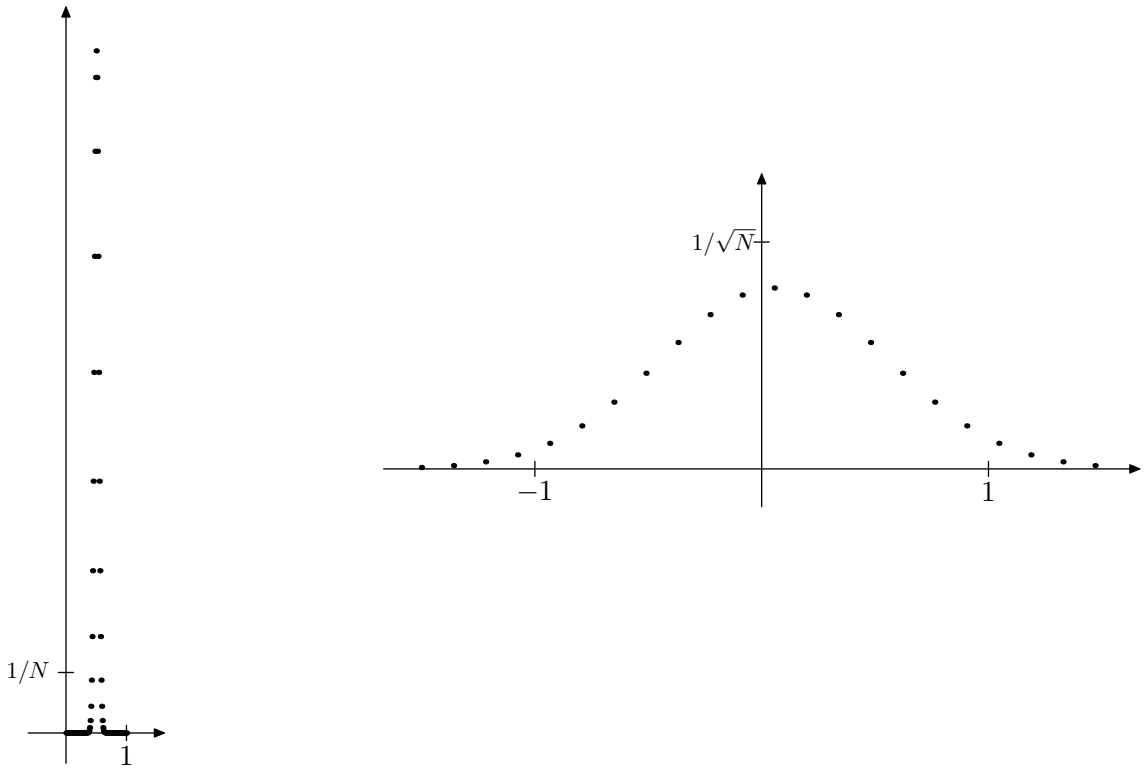
Alors la fonction $g_{0,1/2}$, ça fait cette tête-là (sic) :



À titre d'indication, voici les graphes obtenus respectivement pour les variables X_N, $\frac{X_N}{N}$ et $\frac{X_N - N/2}{\sqrt{N}}$ pour la valeur $N = 200$. Notons que seulement un point sur deux est représenté dans les figures ci-dessous. Notons également que les repères ne sont pas orthonormés mais sont au contraire choisis de façon à ce que l'on puisse comparer les courbes entre elles et avec celle tracée ci-dessus[4].



[4]Précisément, l'unité sur l'axe des ordonnées correspond à celle sur l'axe des abscisses divisée par le nombre moyen de points que l'on trouve en abscisse dans un intervalle de longueur 1.



1.6 Moyenne et variance

Les théorèmes précédents ne sont pas valables seulement pour le jeu de pile ou face : on considère une variable aléatoire X à valeurs réelles, que l'on suppose associée à une certaine mesure p. On peut voir X comme un oracle qui donne des nombres, et ce de telle façon que :

$$\mathrm{Prob}\left(X \in [a,b]\right) = \mathrm{Prob}_p\left([a,b]\right)$$

cette dernière quantité valant soit une intégrale, soit une somme, soit les deux selon que p soit discrète, continue ou les deux à la fois.

Maintenant, on suppose que l'on invoque l'oracle un grand nombre de fois et que l'on ajoute méthodiquement les réels qu'il fournit. Il faut bien se convaincre que dans le cas de pile ou face, cela revenait bien à compter le nombre de fois que la pièce était tombée sur pile. En effet, ce que faisait l'oracle alors, c'était simplement lancer une pièce, et répondre 1 si elle était tombée sur pile, et 0 sinon.

On appelle X_N le résultat obtenu après avoir ajouté les nombres fournis par l'oracle après N appels. Étant donné un réel x maintenant, on s'intéresse à la probabilité que X_N vaille x. On peut donner des formules qui calculent cette probabilité. Par exemple :

$$\mathrm{Prob}\left(X_N = x\right) = \iint\!\!\dots\!\int_D p\left(x_1\right)\dots p\left(x_n\right)\mathrm{d}x_1\dots\mathrm{d}x_n$$

l'intégration se faisant sur l'hyperplan affine D défini par l'équation $x_1 + \ldots + x_n = x$. Cette formule est valable pour une mesure continue correspondant à la fonction p définie sur tout $\mathbb{R}$.

Toutefois ce n'est pas ce qui va nous intéresser[5], on va plutôt se préoccuper comme précédemment au comportement quand N tend vers l'infini.

Pour obtenir cette loi, il nous faut présenter les notions de moyenne et de variance d'une variable aléatoire. Soit donc X une variable aléatoire.

La moyenne de X, ou ce que l'on appelle souvent l'*espérance* de X, est la quantité suivante :

– lorsque la mesure associée est discrète :

$$\mathbb{E}X = \sum_i p_i x_i$$

– lorsque la mesure associée est continue :

$$\mathbb{E}X = \int_R x\, p\,(x)\, \mathrm{d}x$$

La première formule est simplement la moyenne pondérée comme on a l'habitude de la calculer quand on veut savoir si on est premier ou deuxième ou quelle mention on risque d'avoir au bac. La seconde formule est tout aussi simplement une version continue de la première.

Il faut noter qu'il est tout à fait possible que ces sommes et ces intégrales ne convergent pas. Dans ce cas, on ne se tracasse pas, on dit simplement que la variable aléatoire X n'admet pas d'espérance. Finalement l'espérance vérifie au moins la propriété suivante :

$$\mathbb{E}X + \mathbb{E}Y = \mathbb{E}\,(X + Y)$$

valable pour toutes variables X et Y qui admettent une espérance.

Passons à la *variance*. Elle mesure les écarts que prennent les valeurs de la variable aléatoire X par rapport à la moyenne. Ainsi si X est constante (*i.e.* si l'oracle renvoie toujours la même valeur), la variance sera nulle. De façon générale, elle est définie par :

– lorsque la mesure associée est discrète :

$$\sigma^2 X = \sum_i p_i\,(x_i - \mathbb{E}X)^2$$

– lorsque la mesure associée est continue :

$$\sigma^2 X = \int_R p\,(x)\,(x - \mathbb{E}X)^2\, \mathrm{d}x$$

Évidemment, comme somme ou intégrale de nombres positifs, la variance est un nombre positif ; c'est donc tout naturellement qu'on la note σ^2 (ou parfois quand même Var). En fait, on est souvent amené à regarder la racine carrée de ce nombre que l'on appelle l'*écart type* et que l'on note évidemment σ.

Là encore, toute variable aléatoire n'admet pas une variance. Là encore, ce n'est pas grave. Plus intéressant consiste à remarquer que l'on dispose d'une expression concurrente pour le calcul de la variance. Elle est :

$$\sigma^2 X = \mathbb{E}\left(X^2\right) - (\mathbb{E}X)^2$$

[5] Réjouissez-vous... Malgré tout, si vous avez quelques notions sur les intégrales multiples, la formule précédente n'est vraiment pas difficile à établir.

la variable X^2 étant bien entendu la variable X élevée au carré (attention, élever au carré avant de faire la moyenne, ou après, ne donne pas forcément — et même presque jamais — le même résultat !).

Démontrons ce qui a été dit précédemment dans le cas où X est associé à une mesure discrète. Il s'agit donc de voir que :

$$\sum_i p_i \left(x_i - \mathbb{E}X\right)^2 = \sum_i p_i x_i^2 - \left(\sum_i p_i x_i\right)^2$$

Développons pour cela la somme de gauche. On écrit :

$$\begin{aligned}\sum_i p_i \left(x_i - \mathbb{E}X\right)^2 &= \sum_i p_i \left(x_i^2 - 2x_i\mathbb{E}X + \left(\mathbb{E}X\right)^2\right) \\ &= \sum_i p_i x_i^2 - 2\mathbb{E}X \sum_i p_i x_i + \left(\mathbb{E}X\right)^2 \sum_i p_i\end{aligned}$$

On a ici utilisé le fait selon lequel les constantes, en l'occurrence $2\mathbb{E}X$ et $\left(\mathbb{E}X\right)^2$ peuvent se mettre en facteur devant les sommes. Maintenant on reconnaît deux sommes que l'on sait calculer : précisément $\sum_i p_i x_i$ est l'espérance par définition et $\sum_i p_i$ vaut 1 par hypothèse. En réunissant finalement tout, on obtient bien l'expression voulue.

En outre, la variance vérifie d'autres propriétés sympathiques. En particulier, et peut-être de façon étonnante, on a :

$$\sigma^2\left(X + Y\right) = \sigma^2 X + \sigma^2 Y$$

Montrons cela dans le cas d'une mesure discrète. On a, d'après la formule précédente et en notant $\bar{X}$ (resp. $\bar{Y}$) l'espérance de la variable aléatoire X (resp. Y) :

$$\begin{aligned}\sigma^2\left(X + Y\right) &= \sum_i \sum_j p_i q_j \left(x_i + y_j\right)^2 - \left(\bar{X} + \bar{Y}\right)^2 \\ &= \sum_{i,j} p_i q_j \left(x_i^2 + y_j^2 + 2x_i y_j\right) - \left(\bar{X} + \bar{Y}\right)^2 \\ &= \left(\sum_j q_j\right)\left(\sum_i p_i x_i^2\right) + \left(\sum_i p_i\right)\left(\sum_j q_j y_j^2\right) + 2\left(\sum_i p_i x_i\right)\left(\sum_j q_j y_j\right) - \left(\bar{X} + \bar{Y}\right)^2\end{aligned}$$

Miraculeusement, presque tout se simplifie se simplifie dans la dernière somme, ce qui nous permet d'obtenir au final la formule annoncée.

Pour finir, un calcul simple montre que si X est une variable aléatoire qui admet une espérance, alors :

$$\sigma^2\left(\lambda X\right) = \lambda^2 \sigma^2\left(X\right)$$

pour tout réel λ. Et à ce moment, su l'on a bien suivi, on crie au scandale, car on voit directement (en prenant $X = Y$ et $\lambda = 2$) que ce n'est pas compatible avec la formule donnée précédemment censée exprimer la variance d'une somme. Mais rassurons-nous bien vite, l'égalité $\sigma^2\left(X + Y\right) = \sigma^2 X + \sigma^2 Y$ est bien valable mais seulement sous l'hypothèse affirmant que X et Y sont *indépendantes*. Cette dernière chose signifie que la probabilité pour que l'on ait à la fois $X = x$ et $Y = y$ s'exprime simplement comme le produit des probabilités des événements $X = x$ et $Y = y$. Et bien sûr, cela n'est pas vérifié dans le cas où $X = Y$.

Toutefois, on suppose que si on interroge un même oracle N fois d'affilée, toutes ses réponses sont indépendantes. Ainsi dans le cas qui nous intéresse on aura bien :

$$\sigma^2 X_N = N \cdot \sigma^2 X$$

Récapitulons pour finir les propriétés que l'on vient de prouver.

Propriété 1. *Si X et Y sont des variables aléatoires qui admettent une espérance et une variance, et si λ est un réel, alors :*
i) $\mathbb{E}(\lambda X) = \lambda \cdot \mathbb{E}(X)$
ii) $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$
iii) $\sigma^2(\lambda X) = \lambda^2 \cdot \sigma^2 X$
Si de plus X et Y sont indépendantes, on a :
iv) $\sigma^2(X+Y) = \sigma^2 X + \sigma^2 Y$

1.7 La loi des grands nombres (v.2)

On est maintenant en mesure d'énoncer des versions plus générales des théorèmes précédents. Soit X une variable aléatoire. Définissons X_N comme cela a été expliqué au début du chapitre précédent. Pour simplifier les notations, posons $m = \mathbb{E}X$ et $\sigma^2 = \sigma^2 X$.

Avant d'énoncer le théorème, remarquons qu'un simple calcul maintenant nous dit que l'espérance de $\frac{X_N}{N}$ est constante et vaut m. En fait, on a plus précisément :

Théorème 3 (Loi des grands nombres (v.2))**.** *Avec les notations précédentes, la suite des mesures correspondant aux variables $\frac{X_N}{N}$ converge vers la mesure δ_m.*

On voit, encore par un simple calcul, que la variance de $\frac{X_N}{N}$ n'est pas constante, mais que c'est celle de $\frac{X_N}{\sqrt{N}}$ qui l'est. Cela explique accessoirement pourquoi il fallait diviser par $\sqrt{N}$ dans la première partie pour garder une dispersion constante. Cependant la variable aléatoire $\frac{X_N}{\sqrt{N}}$ n'a plus une moyenne constante. Pour concilier les deux, le plus simple est de regarder le quotient $\frac{X_N - mN}{\sqrt{N}}$. On a alors le théorème suivant :

Théorème 4. *Avec les notations précédentes, la suite de mesures correspondant aux variables $\frac{X_N - mN}{\sqrt{N}}$ converge vers la mesure continue définie sur $\mathbb{R}$ par la fonction :*

$$g_{0,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

De façon générale en fait, on définit la fonction $g_{m,\sigma}$ par :

$$g_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

c'est ce que l'on appelle la *gaussienne* de moyenne m et d'écart type σ.

2 Pourquoi e^{-x^2} ? La théorie de l'information

Nous allons, dans ce chapitre, principalement présenter ce que l'on appelle la *théorie de l'information*. Elle a été développée par Shannon vers la fin de la Seconde Guerre Mondiale.

2.1 D'abord sans probabilité

On se donne pour l'instant un ensemble fini Ω et on se demande combien il nous faut d'information pour pouvoir isoler un élément parmi tous ceux de Ω. Pour être plus précis, on demande combien il faut « en gros » poser de questions fermées oui/non pour être sûr d'avoir trouvé le bon élément.

La réponse est évidemment $\log(\operatorname{Card}\Omega)$, où le symbole log désigne ici le logarithme calculé en base 2. Si ln désigne le logarithme népérien classique, on a $\log x = \frac{\ln x}{\ln 2}$. Par exemple, si Ω est de cardinal 16, il faut poser 4 questions, le principe étant bien entendu de couper à chaque fois l'ensemble en deux.

Pour tout élément $x \in \Omega$, on définit donc :

$$I(x) = \log(\operatorname{Card}\Omega)$$

c'est la *quantité d'information* véhiculée par x. Prenons maintenant A une partie de Ω et essayons de déterminer la quantité d'information véhiculée par la partie A. Sans connaître A, il fallait poser $\log(\operatorname{Card}\Omega)$ questions pour trouver un x. Si on sait maintenant en plus que x est dans A, il suffit d'en poser $\log(\operatorname{Card} A)$. A véhicule donc la quantité d'information différence. Que ce petit calcul vous convainque ou non, on définit la *quantité d'information* véhiculée par la partie A *via* la formule suivante :

$$I(A) = \log(\operatorname{Card}\Omega) - \log(\operatorname{Card} A) = -\log\left(\frac{\operatorname{Card} A}{\operatorname{Card}\Omega}\right)$$

Bien évidemment, il aurait été absurde de définir $I(A)$ comme la somme des $I(x)$ pour $x \in A$: savoir que x appartient à A donne évidemment moins d'information que le fait de savoir que x est x. La quantité d'information, comme nous l'avons déjà dit, correspond moralement au nombre de questions fermées qu'il faut poser pour repérer un élément dans un ensemble.

2.2 Et maintenant, avec

On prend encore un ensemble fini Ω, mais on suppose que tous ses éléments ne sont pas forcément équiprobables. On se demande alors la quantité d'information que véhicule un élément $x \in \Omega$. Par analogie avec ce que l'on a fait dans le paragraphe précédent, on peut dire que la *quantité d'information* véhiculée par x est :

$$I(x) = -\log p(x)$$

où $p(x)$ désigne la probabilité que x_i soit tiré. L'analogie avec le cas précédent vient de la remarque suivante : on a vu que l'information « être élément de A » est quantifiée par le nombre $I(A) = -\log\left(\frac{\operatorname{Card} A}{\operatorname{Card}\Omega}\right)$; or on peut remplacer la partie A par un seul bloc de poids $\operatorname{Card} A$, et il est alors logique que l'information « être ce bloc » soit quantifiée par le même nombre, c'est-à-dire $-\log\left(\frac{\operatorname{Card} A}{\operatorname{Card}\Omega}\right) = -\log p(A)$.

Intuitivement on peut remarquer qu'autant l'arrivée d'un événement fréquent (*i.e.* de probabilité grande) apporte peu d'information autant l'arrivée d'un événement rare en apporte beaucoup. Si vous n'êtes pas convaincu, essayez de vous rappeler de votre jeunesse lorsque vous jouiez au pendu : lorsque l'on arrive à déceler un `Z` dans un mot, on est bien plus content que lorsque l'on trouve des `E`. Pourquoi cela ? Parce que l'on obtient ainsi plus d'information, pardi. Pour un exemple concret, essayez de compléter les mots `E___E` et `Z___E`. Pour lequel pensez-vous trouver le plus vite ?

On peut définir ensuite la moyenne de toutes ces quantités d'information. C'est ce que l'on appelle l'entropie :

$$S(\Omega) = -\sum_{x\in\Omega} p(x)\log p(x)$$

Elle correspond au nombre moyen de questions qu'il faut poser pour déterminer un élément parmi ceux de Ω. Remarquons que c'est une fonction de l'ensemble Ω avec sa répartition de probabilités, et pas d'un élément x en particulier.

Au cas où l'expression précédente vous paraîtrait encore farfelue, voyons le théorème suivant :

Théorème 5. *La fonction* $(p_i) \mapsto -\sum_i p_i \log p_i$*, pour* $0 \leqslant p_i \leqslant 1$ *et* $\sum p_i = 1$*, est la seule qui soit*

- *symétrique et continue en les* p_i
- *positive*
- *telle que* $S\left(\frac{1}{2}, \frac{1}{2}\right) = 1$
- *telle que* $S(p_1, p_2, \ldots, p_n) = S(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2)\, S\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$

Il faut faire quelques remarques. La troisième condition est simplement une condition de normalisation qui fixe l'unité de la quantité d'information. Elle dit donc que pour distinguer entre deux termes équiprobables, on a besoin d'*une* (et une seule) information ; cela correspond bien à l'idée intuitive de question fermée que l'on se fait.

La dernière condition n'est pas non plus mystérieuse. Supposons que je doive envoyer un message à deux personnes différentes, mais qu'une des deux personnes n'arrive jamais à distinguer entre mes `E` et mes `F`. Il va falloir que pour chacune de ces lettres, elle me repose la question « est-ce un `E` ou un `F` ? ». Donc si p_1 désigne la probabilité d'apparition d'un `E`, et p_2 celle d'un `F`, cette personne qui lit mal devra me poser une question supplémentaire avec une fréquence p_1+p_2 des cas. Ainsi, elle reçoit pour quantité d'information, la quantité $S(p_1 + p_2, p_3, \ldots, p_n)$ (qui correspond à la quantité d'information, lettres `E` et `F` confondues) mais elle doit en redemander $S\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$ avec probabilité p_1+p_2. Finalement, l'égalité découle de ces constatations.

On peut remarquer que l'on a toujours $S \leqslant S_0$ où $S_0 = \log(\operatorname{Card}\Omega)$. C'est une simple inégalité de convexité, pas difficile à établir si l'on connaît[6]. Cela signifie qu'on a besoin de moins d'information pour deviner ce qui s'est passé dans le cas « non équiprobable » : on a plus d'information au départ lorsqu'on sait que certaines choses apparaissent plus fréquemment.

2.3 Codage de Huffman

C'est la dernière idée du paragraphe précédent que nous allons exploiter ici. Il est plus simple sans doute pour ce qui va suivre de voir les éléments de Ω comme les lettres d'un certain alphabet, par exemple de l'alphabet français, mais aussi pourquoi pas de celui qui regroupe tous les caractères `ASCII` ou n'importe quoi d'autre. Prenons pour simplifier l'alphabet français.

On n'est sûrement pas sans savoir que toutes les lettres ne sont pas équiprobables. En particulier, les `E` reviennent plus souvent que les `Z` comme nous l'avons déjà fait remarquer. Ainsi l'entropie de cet ensemble probabilisé est inférieure à $\log(26)$ et lorsque votre meilleur

[6] D'ailleurs si vous voulez vous entraîner à le faire, cela ne peut être qu'instructif.

ami choisit au hasard une lettre dans un texte et vous demande de la deviner, la meilleure stratégie n'est peut-être pas de commencer à lui demander si elle est dans la première partie de l'alphabet et à continuer ainsi, mais peut-être est-il plus judicieux de lui demander directement si c'est un `E` ? Voici ci-dessous un tableau avec les fréquences d'apparition de chaque lettre dans les textes français[7] :

Lettre	Pourcentage d'apparition
A	8,40 %
B	1,06 %
C	3,03 %
D	4,18 %
E	17,27 %
F	1,12 %
G	1,27 %
H	0,92 %
I	7,34 %

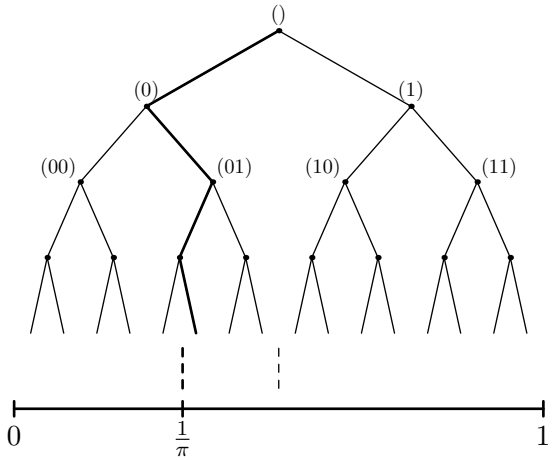
Lettre	Pourcentage d'apparition
J	0,31 %
K	0,05 %
L	6,01 %
M	2,96 %
N	7,13 %
O	5,26 %
P	3,02 %
Q	0,99 %
R	6,55 %

Lettre	Pourcentage d'apparition
S	8,08 %
T	7,07 %
U	5,75 %
V	1,32 %
W	0,04 %
X	0,45 %
Y	0,30 %
Z	0,12 %

En consultant le tableau précédent, on voit que lettres `E`, `A`, `S`, `I` et `N` constituent à elles seules pratiquement la moitié des occurrences des lettres rencontrées dans des textes français standard. Une première question intéressante serait sans doute, non pas de demander si la lettre choisie est un `E`, mais de demander si elle ne ferait pas partie des lettres précédemment énumérées.

En outre, on remarque que l'entropie du système est à peine de $4,00$. Ainsi il devrait être possible de déterminer une lettre en seulement 4 questions et donc de coder une lettre sur 4 bits (alors qu'*a priori* $\log 26 \simeq 4,7$). On obtient ainsi un compresseur de données.

Voyons comment l'on concrétise les idées précédentes et énonçons un théorème qui précise tout cela. Identifions d'abord les réels de l'intervalle $[0,1]$ aux chemins infinis d'un arbre binaire complet, comme le montre le dessin suivant :



Expliquons ce qu'est la correspondance. Prenons un chemin dans l'arbre qui part de la racine (en haut) et qui descend infiniment. Il lui est associé un unique réel de l'intervalle $[0,1]$ dont on peut déterminer directement l'écriture en base 2 de sa partie « décimale » :

[7] Ce genre de statistiques est très utile à plein de choses : par exemple, elles permettent de reconnaître à coup sûr en quelle langue est écrit un texte, disons s'il possède plus d'une centaine de mots.

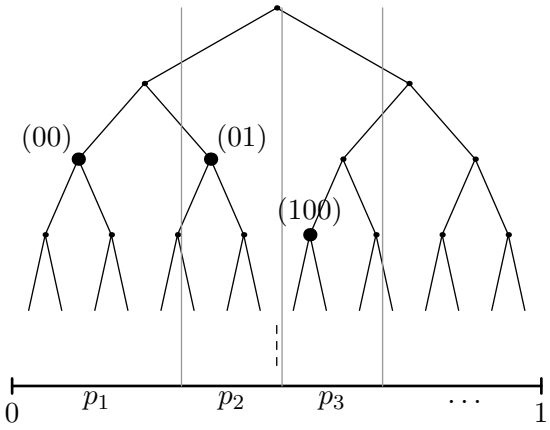
si l'on descend à gauche dans l'arbre on rajoute un 0 dans l'écriture en base 2, sinon un 1. On obtient comme cela un réel de l'intervalle $[0, 1]$.

Ainsi tous les chemins qui passent par le nœud marqué (0) correspondent à des réels compris entre 0 et 1 dont le premier chiffre après la virgule en base 2 est un 0, c'est-à-dire des réels compris entre 0 et $\frac{1}{2}$. Il est donc normal de dire que le nœud marqué (0) correspond à l'intervalle $\left[0, \frac{1}{2}\right]$. De la même façon le nœud marqué (01) correpond à l'intervalle $\left[\frac{1}{4}, \frac{1}{2}\right]$ et ainsi de suite. Descendre dans l'arbre correspond à chaque étape à couper notre intervalle en deux et à choisir un des deux sous-intervalles ; à la limite, on n'obtient plus qu'un réel. De cette façon, le chemin dont le début est dessiné en gras correspond au réel $\frac{1}{\pi}$.

Il faut par contre remarquer qu'à un réel donné peuvent correspondre deux chemins de l'arbre. Par exemple, pour obtenir $\frac{1}{2}$, on peut aller d'abord à gauche puis toujours à droite, ou d'abord à droite puis toujours à gauche. Cette ambiguïté est présente pour tous les nombres de la forme $\frac{k}{2^n}$, k et n étant des entiers. Pour tous les autres réels, il n'y a qu'un seul chemin qui leur aboutit.

On veut à présent coder par une suite (*a priori* de longueur variable) de bits chacun des éléments de Ω et ce de telle façon qu'en moyenne on utilise le moins possible de bits pour coder un caractère. Écrivons par exemple $\Omega = \{x_1, \ldots, x_n\}$ en mettant en premier les éléments les plus fréquents, de telle façon que si p_i désigne la probabilité attachée à l'élément x_i, on ait $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_n$.

Comme on l'a déjà plus ou moins évoqué précédemment, l'idée consiste à attribuer des suites courtes aux caractères fréquents et des suites plus longues aux caractères rares. Pour faire cela, on utilise la correspondance précédente :



On découpe le segment représentant $[0, 1]$ en n segments de longueurs respectives $p_1, \ldots, p_n$. On regarde ensuite le premier nœud (*i.e.* celui qui est le plus haut possible) dont l'intervalle correspondant est inclus dans $[0, p_1]$. Ici, c'est celui étiqueté (00) : on décrète alors que `00` est le code affecté à l'élément x_1. Maintenant, on barre sur l'arbre tous les nœuds qui sont soit « avant », soit « après » le nœud retenu. Ici les nœuds qui sont avant (00) sont (0) et () et les nœuds qui sont après sont ceux dont l'étiquette commence par (00). On barre également le nœud retenu.

Puis on continue : pour déterminer le code du second élément, on regarde parmi les nœuds non barrés le premier qui correspond à un intervalle inclus dans $[0, p_1 + p_2]$. Ici c'est (01) et donc `01` est le code recherché. On barre ensuite les nœuds qu'il faut barrer et on recommence la manipulation jusqu'à épuisement de Ω. Sur notre exemple `100` est le code de x_3.

Maintenant supposons que l'on veuille coder un mot dont les lettres sont prises dans l'alphabet Ω. Pour cela, on regarde une par une les lettres du mot et on les remplace par les codes déterminés précédemment. Le fait que l'on ait barré à chaque fois les nœuds « avant » et « après » dit précisément qu'il n'y aura aucune ambiguïté pour déchiffrer le code, car aucun code n'est le début d'un autre code.

Mais que gagne-t-on à coder de la façon précédente ? On gagne sur la longueur moyenne (en bits) du code d'une lettre de Ω : elle se rapproche de S, l'entropie du système. En effet, le code d'un intervalle de largeur p_i a une taille environ $-\log p_i$, et donc, avec probabilité p_i on a une longueur de code $-\log p_i$ d'où l'expression de la longueur moyenne. Le « à peu près » est obligatoire : on ne peut être précis au possible puisqu'on ne peut travailler sur des fractions de bits. L'entropie est le minimum *théorique* si l'on veut récupérer toute l'information[8].

Disons tout de suite que le type de codage que l'on vient de présenter est réellement utilisé en pratique pour la compression de données. En particulier les formats `zip` ou `gif` utilisent essentiellement des variations sur ce thème[9].

Pour terminer ce paragraphe, regardons ce que donne l'algorithme si Ω est l'ensemble des lettres de l'alphabet français probabilisé avec les statistiques précédentes. On obtient :

Lettre	Code
E	000
A	001
S	0100
I	0101
N	0110
T	0111
R	1000
L	1001
U	1010

Lettre	Code
O	1011
D	1100
C	11010
P	11011
M	11100
V	111010
G	111011
F	111100
B	1111010

Lettre	Code
Q	1111011
H	111110
X	11111100
J	11111101
Y	11111110
Z	111111110
K	1111111110
W	1111111111

Un calcul rapide nous dit que la longueur moyenne du code d'un caractère sera de $4,04$, bien inférieure à $\log 26 \simeq 4,7$, et assez proche de l'entropie du système. On gagne donc (en place) à utiliser le code précédent ; on compresse le texte à environ 86% de sa taille d'origine.

On constate *a posteriori* que l'idée que nous avions avancée au tout début est justifiée. Le premier bit du code vaut `0` pour les lettres `E`, `A`, `S`, `I`, `N` et `T`, et vaut `1` pour les autres lettres. Si l'on adopte le point de vue « questions », cela signifie exactement que si l'on veut déterminer la lettre en un nombre minimum de questions, il vaut mieux commencer par demander si la lettre ne serait pas l'une de celles précédemment citées.

2.4 Pour une mesure continue

Il est sans doute temps d'expliquer le rapport entre ce qui précède et ce que nous nous proposions d'étudier au premier chapitre. Déjà, on remarque que l'ensemble Ω muni d'une

[8] Et si les lettres successives sont vraiment indépendantes. Ce n'est pas le cas en français où, par exemple, après deux lettres *io* on a une chance importante de rencontrer la lettre n ; et donc, si l'on code par groupes de trois lettres plutôt que par lettre, on peut tirer avantage de cette information : l'entropie des groupes de trois lettres est inférieure à celle des lettres isolées.

[9] Le `mp3` ou le `jpg` utilisent des méthodes totalement différentes. Elles sont bien plus efficaces mais ne permettent en général pas de restituer *toute* l'information.

probabilité n'est autre qu'une variable aléatoire associée à une mesure discrète. Évidemment, les x_i n'étant pas des réels, cette variable aléatoire n'est pas à valeurs réelles. On ne pourra donc par exemple pas en calculer l'espérance ou la variance, mais de toute façon on n'en avait pas besoin dans ce qui précédait...

Nous allons donc pouvoir copier les définitions précédentes dans le cas d'une variable aléatoire associée à une mesure continue, disons donnée par la fonction continue p. L'entropie se calcule au moyen de la formule suivante :

$$S(X) = S(p) = -\int_{\mathbb{R}} p(x) \log p(x) \, \mathrm{d}x$$

Bien sûr, on peut généraliser la définition précédente pour une variable aléatoire « quelconque », c'est-à-dire qui est associée à une mesure à la fois discrète et continue.

Donnons des exemples. L'entropie de la masse de Dirac en un point quelconque est nulle. On peut, ici, reprendre notre analogie en termes de nombre de questions nécessaires à poser. Si vous demandez à l'oracle de vous donner un réel distribué selon la masse de Dirac en 0, bien évidemment il va toujours vous répondre 0 et je n'aurais besoin de poser aucune question pour savoir ce qu'il vous a dit.

De même, l'entropie de la mesure uniforme sur N points est $\log N$. On peut faire le calcul évidemment, mais cela doit paraître évident si l'on raisonne en termes de « nombre de questions nécessaires ». Si vous demandez à l'oracle de vous donner équiprobablement un nombre parmi N fixés à l'avance, il va falloir que je vous pose $\log N$ questions (en moyenne) pour déterminer quel est ce nombre. C'était notre point de départ.

Si l'on prend maintenant la mesure uniforme sur $[0,1]$, un simple calcul nous dit que son entropie est nulle. Qu'est-ce que cela signifie alors ? Que si vous demandez à l'oracle de vous donner au hasard un réel entre 0 et 1, il va me falloir aucune question pour le déterminer ? En fait, je pourrais poser autant de questions que je le veux, je ne pourrais jamais le déterminer exactement, il y a bien trop de réels. Il faut remplacer dans ce contexte « déterminer » par « déterminer avec une précision de 1 »[10]. Alors évidemment, après ce changement je n'ai plus aucune question à poser pour déterminer un intervalle d'amplitude 1 dans lequel est le nombre de l'oracle ; je n'ai qu'à répondre systématiquement $[0,1]$.

Si l'on ne considère plus la mesure uniforme sur $[0,1]$ mais sur l'intervalle $[a,b]$, on peut calculer l'entropie correspondante et on trouve $\log(b-a)$. On s'attend bien à ce que l'on pensait : pour déterminer, avec une amplitude de 1, un nombre dans un intervalle de longueur $b-a$, il faut poser en moyenne $\log(b-a)$ questions. On note quand même que si $b-a<1$, l'entropie est négative. Ce n'est pas absurde : si l'on doit déterminer un intervalle d'amplitude 1 dans lequel se trouve le réel x et que l'on sait par ailleurs que x est compris entre 0 et $\frac{1}{2}$, on a déjà trop d'informations. Il faut donc en donner, c'est-à-dire poser un nombre négatif de questions, pour rétablir la situation.

2.5 La loi de l'emmerdement maximal

Théorème 6. *Fixons m et σ deux paramètres. Parmi toutes les variables aléatoires X associées à des mesures continues sur tout $\mathbb{R}$ et vérifiant $\mathbb{E}X = m$ et $\sigma^2 X = \sigma^2$, c'est la gaussienne associée à la fonction $g_{m,\sigma}$ qui réalise le maximum de l'entropie.*

[10] En fait l'intégrale à calculer n'était pas $-\int_{\mathbb{R}} p(x) \log p(x) \, \mathrm{d}x$ mais plutôt $-\int_{\mathbb{R}} p(x) \, \mathrm{d}x \log(p(x) \, \mathrm{d}x)$. Seulement pour tout sens raisonnable que l'on peut donner à cette dernière intégrale, elle vaudra toujours $+\infty$ (ce qui rend donc bien l'idée selon laquelle je ne pourrai jamais déterminer le réel donné par l'oracle). On choisit donc de fixer la valeur de $\mathrm{d}x$ à l'intérieur du log, ce qui fixe par ailleurs une certaine précision. Si on fixe $\mathrm{d}x = 1$, on fixe la précision à 1.

Combiné avec le théorème de la limite centrale, on obtient la loi de l'emmerdement maximal : les mesures que l'on obtient naturellement en itérant une expérience (dans le but de l'étudier) sont celles qui maximisent l'entropie, c'est-à-dire celles qui minimisent la quantité d'information *a priori*.

Ne restons pas sur cette note triste et donnons plutôt les grandes lignes de la démonstration de ce théorème. On suppose donc que p est une fonction qui maximise l'entropie. Alors pour toute fonction δp qui est telle que $p+\delta p$ vérifie encore les hypothèses du théorème, on doit avoir $S(p+\delta p) \leqslant S(p)$. Commençons par calculer $S(p+\delta p)$. Par définition on a :

$$S(p+\delta p) = -\int_{\mathbb{R}} (p+\delta p)\log(p+\delta p)$$

Dans l'écriture précédente, on a omis la variable x et de fait la quantité $\mathrm{d}x$, mais il ne faut pas oublier que p et δp sont bien des fonctions, et que ce sont à peu de choses près elles que l'on intègre.

Maintenant, si δp est suffisamment petit, on va avoir :

$$\log(p(x)+\delta p(x)) = \frac{1}{\ln 2}\left(\ln p(x) + \frac{\delta p(x)}{p(x)}\right) + O\left((\delta p(x))^2\right)$$

le terme $O((\delta p(x))^2)$ réunit en fait tous les autres termes dans lesquels on peut mettre $(\delta p(x))^2$ en facteur.

Nous allons écrire cette dernière égalité plus simplement sous la forme :

$$\log(p+\delta p) = \frac{1}{\ln 2}\left(\ln p + \frac{\delta p}{p}\right) + O\left(\delta^2 p\right)$$

Ainsi, après remplacement, on obtient :

$$S(p+\delta p) = S(p) - \int_{\mathbb{R}} \delta p\left(\frac{1}{\ln 2} + \ln p\right) + O\left(\delta^2 p\right)$$

Ce qu'il faut comprendre maintenant, c'est que tous les termes qui sont cachés dans le $O\left(\delta^2 p\right)$ sont, quand δp est petit, tout petits devant l'intégrale $\int_{\mathbb{R}} \delta p\,(1+\ln p)$. Ainsi si l'on veut que $S(p+\delta p) \leqslant S(p)$, il faut que l'intégrale précédente soit toujours positive. Mais ce que l'on voit en outre, c'est que si l'on change δp en $-\delta p$ (si δp vérifiait les hypothèses qu'il fallait, alors on vérifie que $-\delta p$ aussi), cette dernière intégrale change de signe. Donc si l'on veut qu'elle soit toujours positive, on veut en fait qu'elle soit toujours nulle. Une condition que l'on doit vérifier est donc finalement :

$$\int_{\mathbb{R}} \delta p\left(\frac{1}{\ln 2} + \ln p\right) = 0 \tag{1}$$

pour toute fonction δp vérifiant les bonnes hypothèses. Voyons maintenant quelles sont ces hypothèses.

On veut tout d'abord que $(p+\delta p)$ soit encore une fonction qui corresponde à une mesure, ce qui implique $\int_{\mathbb{R}} p+\delta p = 1$, ou encore $\int_{\mathbb{R}} \delta p = 0$. On veut en outre que l'espérance de $p+\delta p$ soit m et que sa variance soit σ^2, comme pour p. On vérifie que cela implique

respectivement $\int_{\mathbb{R}} x\delta p\,(x)\,\mathrm{d}x = 0$ et $\int_{\mathbb{R}} x^2\delta p\,(x)\,\mathrm{d}x = 0$. Finalement on peut regrouper ce qui précède en disant que l'on doit se restreindre aux fonctions δp qui vérifient

$$\int_{\mathbb{R}} \left(A + Bx + Cx^2\right) \delta p\,(x) = 0 \tag{2}$$

et ce pour tous réels A, B et C.

Si l'on récapitule, on cherche donc une fonction p qui soit telle que toute fonction δp vérifiant (2) vérifie aussi (1). Bien sûr si l'on choisit p de telle sorte que

$$\frac{1}{\ln 2} + \ln p\,(x) = A + Bx + Cx^2$$

cela va marcher. En fait, on peut montrer que ce sont les seules solutions. Et donc finalement que p s'écrit :

$$p\,(x) = \exp\left(A' + Bx + Cx^2\right)$$

ce qui veut dire que p est une gaussienne.

Il ne reste plus qu'à déterminer A', B et C *via* les conditions de moyenne et variance. On retombe bien sûr sur $g_{m,\sigma}$ puisque c'est la seule gaussienne avec ces caractéristiques.

3 Concentration

3.1 Pour le cube

Revenons au théorème de la limite centrale (`v.1`). On peut le reformuler de plusieurs façons et notamment :

Théorème 7 (Théorème de la limite centrale (`v.1`)). *Soit $f : \{0,1\}^N \to \mathbb{R}$ la fonction définie par :*

$$f\,(x_1, \ldots, x_N) = \frac{1}{N} Card\,\{i \,/\, x_i = 0\}$$

La fonction f est alors constante à $\frac{1}{\sqrt{N}}$ près. Cela signifie plus exactement que si ε est un réel strictement positif, on a l'inégalité :

$$Prob\left(\left|f - \frac{1}{2}\right| \geqslant \varepsilon\right) \leqslant k \exp\left(-2N\varepsilon^2\right)$$

où k est une constante.

(Notons que la variance σ^2 vaut ici 1/4, d'où le facteur 2 dans l'exposant). Cette majoration ne découle en fait peut-être pas directement du précédent théorème que l'on avait énoncé. En effet, ce dernier résultat ne donnait des informations que pour N grand, et pas pour tout entier N. Si l'on veut une formulation plus faible donc mais qui soit une conséquence directe du théorème de la limite centrale donné auparavant, il faut dire que sous les hypothèses du théorème précédent, on a la majoration suivante :

$$\mathrm{Prob}\left(\left|f - \frac{1}{2}\right| \geqslant \frac{\lambda}{\sqrt{N}}\right) \leqslant k \exp\left(-2\lambda^2\right)$$

mais valable non pas pour tout N mais seulement pour N suffisamment grand.

Il est remarquable de constater que le théorème précédent s'étend à des fonctions bien plus générales que la fonction f « proportion de zéros » définie ci-dessus (avec néanmoins une constante un chouia moins bonne). C'est ce que précise le théorème suivant :

Théorème 8. *Soit $f : \{0,1\}^N \to \mathbb{R}$ vérifiant l'hypothèse suivante :*

$$\forall (x_1, \ldots, x_n) \in \{0,1\}^n, \forall i \quad |f(x_1, \ldots, x_N) - f(x_1, \ldots, x_{i-1}, 1 - x_i, x_{i+1}, \ldots, x_N)| \leqslant \frac{1}{N}$$

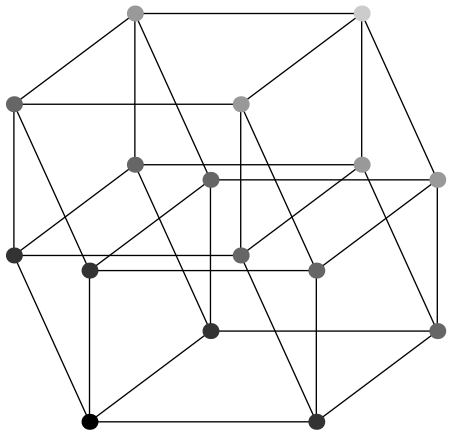
Alors f est constante à $\frac{1}{\sqrt{N}}$ près. Autrement dit, il existe une constante M telle que pour tout réel ε strictement positif, on ait l'inégalité :

$$Prob\left(|f - M| \geqslant \varepsilon\right) \leqslant 2\exp\left(-N\varepsilon^2\right)$$

Faisons quelques commentaires. Déjà ce théorème date des années 1990. Voyons ensuite l'hypothèse. Elle dit que la fonction f ne doit pas trop bouger lorsque l'on change une coordonnée (chaque variable influe sur le résultat d'au plus $1/N$), mais elle permet quand même des écarts importants : $f(0, \ldots, 0)$ et $f(1, \ldots, 1)$ peuvent être distants de 1. Si on a une fonction f qui varie plus, il suffit de renormaliser en considérant la fonction f/k, avec k assez grand, pour appliquer le théorème — on obtient alors évidemment un résultat k fois moins bon.

La conclusion, quant à elle, ne dit pas autre chose que f ne s'écarte pas trop de la valeur moyenne M, les écarts étant du même ordre de grandeur que ceux enregistrés lors d'un jeu de pile ou face répété.

Ce qu'il faut dire également, c'est qu'ainsi formulé, ce théorème n'est plus un théorème de probabilité, mais bien un théorème de géométrie qui donne des renseignements sur la forme du cube discret (*i.e.* l'ensemble $\{0,1\}^N$). Plus précisément, esquissons un dessin pour $N = 4$:



Ce qui est représenté ci-dessus est un cube en dimension 4, ce que l'on appelle généralement un *hypercube*. La fonction f est donc définie sur les sommets de cet hypercube. Les niveaux de gris représentent les valeurs de cette fonction : plus la couleur est foncée, plus le nombre est grand. L'hypothèse nous dit que l'on n'a pas le droit de trop modifier l'intensité de la couleur lorsque l'on se déplace sur une arête. La conclusion, quant à elle, nous dit que les sommets sont essentiellement tous de la même couleur[11]. Cela signifie donc qu'il y a beaucoup de contraintes et donc beaucoup de segments qui relient tous ces sommets. Voici l'information géométrique cachée derrière ce théorème.

[11] Évidemment, cela ne se voit pas très bien pour des valeurs de N trop petites : 4 n'est pas encore très grand.

3.2 Pour la sphère

On a un théorème analogue pour la sphère de dimension N (et de rayon fixé à 1). Dans un premier temps, disons que cette sphère est définie comme étant l'ensemble :

$$\mathbb{S}^N = \left\{(x_0, \ldots, x_N) \in \mathbb{R}^{n+1} \,/ \quad x_0^2 + \ldots + x_N^2 = 1\right\}$$

Il faut également dire quelle notion remplace celle de probabilité présente jusqu'alors : c'est naturellement celle de volume, ou de surface si l'on préfère. Par exemple pour $N = 2$, dans la sphère $\mathbb{S}^2$ (incluse dans l'espace), la probabilité d'une partie de la sphère est sa surface divisée par la surface totale de la sphère. En dimension plus grande, on remplace la surface par la surface N-dimensionnelle.

Enfin, pour deux points sur la sphère on peut bien sûr parler de leur distance $\text{dist}_{\mathbb{S}^N}$, égale à la longueur du plus court chemin *contenu dans la sphère* joignant les deux points.

Passons à l'énoncé de notre théorème :

Théorème 9. *Soit $f : \mathbb{S}^N \to \mathbb{R}$ une fonction vérifiant*

$$|f(x) - f(y)| \leqslant dist_{\mathbb{S}^N}(x, y)$$

Alors il existe une constante M qui soit telle que pour tout réel ε strictement positif :

$$\frac{Vol\left(\left\{x \in \mathbb{S}^N \,/\, |f(x) - M| \leqslant \varepsilon\right\}\right)}{Vol\,\mathbb{S}^N} \leqslant 2\exp\left(-N\varepsilon^2/2\right)$$

L'hypothèse $|f(x) - f(y)| \leqslant \text{dist}_{\mathbb{S}^N}(x, y)$ est bien nécessaire : si la fonction peut varier n'importe comment, on ne risque pas de contrôler ses variations. Cette hypothèse dit que la fonction change peu quand on bouge un peu sur la sphère.

Là encore, ce théorème est un théorème de géométrie qui donne des informations sur la forme de la sphère en grande dimension. Prenons pour illustrer cela l'exemple de la fonction *hauteur* définie par :

$$h(x_0, \ldots, x_N) = x_0$$

C'est une projection orthogonale donc même si l'on ne sait pas comment calculer la longueur sur une sphère, on se doute bien qu'elle ne peut pas agrandir les longueurs ; elle vérifie donc, par un argument d'autorité certes, l'hypothèse du théorème précédent. Par symétrie du problème, il n'est pas non plus bien difficile de se convaincre que la constante M qui apparaît est forcément nulle.

Ainsi si l'on prend ε de l'ordre de $\frac{1}{\sqrt{N}}$, on voit que l'essentiel du volume de la sphère de dimension N est compris entre les plans d'équations $X_0 = -\frac{1}{\sqrt{N}}$ et $X_0 = \frac{1}{\sqrt{N}}$. La sphère est donc aplatie dans ce sens. Cela n'est peut-être pas surprenant, mais pensez maintenant que cela est vrai dans *toutes* les directions de l'espace[12], c'est bien plus troublant semblerait-il : la sphère de dimension N se situerait principalement dans la boule de rayon $\frac{1}{\sqrt{N}}$! En contrepartie, si vous ne croyez pas à cette affirmation peut-être gratuite, libre à vous de faire les calculs de volume qui s'imposent pour vérifier.

[12] Il s'agit bien de *toutes* les directions, pas seulement celles parallèles aux axes de coordonnées par exemple.

3.3 Les espaces concentrés

En vertu de deux théorèmes énoncés dans les paragraphes précédents, on dit souvent que le cube discret et la sphère sont des *espaces concentrés*.

Il existe d'autres exemples d'espaces concentrés, c'est-à-dire d'autres espaces pour lesquels on dispose de théorèmes analogues. En particulier, les convexes de grande dimension sont des espaces concentrés ; le théorème est alors exactement le même à la constante près dans l'exposant, cette constante dépendant en fait simplement du rayon de courbure maximal de l'espace (plus l'espace est courbé, c'est-à-dire en quelque sorte recroquevillé sur lui-même, meilleure est la concentration).

3.4 Quelques idées de démonstration

Nous allons dans ce dernier paragraphe expliquer plus ou moins en détail comment on prouve les théorèmes 8 et 9. L'idée de base est identique pour les deux théorèmes : le phénomène de concentration se voit comme la conséquence d'une inégalité d'isopérimétrie.

Isopérimétrie et concentration

L'inégalité isopérimétrique dans le plan est bien connue : parmi les figures de surface donnée, c'est le disque qui a le plus petit périmètre. Sa généralisation en dimension supérieure est claire : parmi toutes les parties de $\mathbb{R}^N$ dont le volume est fixé, c'est la boule qui admet la plus petite aire (*i.e.* volume en dimension $N-1$). Autrement dit, si une partie A a un volume égal à celui de la boule de rayon R, alors la frontière de A a une aire plus grande que celle de la sphère de même rayon.

On peut imaginer sans trop de problèmes un équivalent pour l'espace $\mathbb{S}^N$ (parmi les parties de $\mathbb{S}^N$, ce sont les « calottes » qui minimisent le volume du bord, à volume interne fixé), mais on a peut-être plus de mal à se rendre compte de ce que cela pourrait être pour l'espace discret $\{0,1\}^N$.

Déjà, on est en droit de se demander quels sont les équivalents de volume et périmètre dans le cube discret. Définir un volume d'une partie A n'est pas le plus compliqué : on compte le nombre de points présents dans A, ce cardinal convient parfaitement comme volume.

Que doit-on faire pour le périmètre ? On a besoin au préalable de la donnée d'une distance sur le cube discret. C'est la *distance de Hamming* qui va être retenue, seule distance sur le cube relativement naturelle au demeurant. Si $x = (x_1, \dots, x_N)$ et $y = (y_1, \dots, y_N)$ sont deux N-uplets composés de 0 et de 1, on définit la distance entre x et y comme $1/N$ fois le nombre d'indices i tels que $x_i \neq y_i$, et on note ce nombre $d(x,y)$ (on divise par N pour que la distance maximale soit 1).

Maintenant si A est une partie de $\{0,1\}^N$ dont on veut définir le périmètre, on commence par regarder les ensembles suivants :

$$A^t = \left\{ x \in \{0,1\}^N \ / \quad d(x,A) \geqslant t \right\}$$

t étant ici un paramètre et $d(x,A)$ désignant la distance de l'élément x à la partie A, c'est-à-dire le plus courte distance séparant x et un élément de A. Moralement, A^t est une forme « régulière » dont le contour ressemble fortement à celui de A (pensez peut-être au cas de $\mathbb{R}^N$ pour mieux visualiser les choses). Il est donc raisonnable de penser que, si périmètre d'une partie de A on peut définir, ce périmètre est relié d'une façon ou d'une

autre aux volumes de A^t (pour t petit). Ce que l'on appelle *périmètre* de A dans la suite de ce paragraphe[13], c'est la fonction $t \mapsto \mathrm{Vol}\left(A^t\right)$.

Les définitions précédentes se généralisent directement en fait à tout espace Ω sur lequel on est capable de définir une distance, un volume, et aussi pour lequel le volume total de Ω est fini[14]. Quitte à tout renormaliser (*i.e.* multiplier tous les volumes par une même constante, en l'occurrence $\frac{1}{\mathrm{Vol}(\Omega)}$), on peut supposer non pas que le volume de Ω est fini, mais qu'il vaut 1.

On énonce alors un théorème général sur les espaces précédents :

Théorème 10. *Soit Ω un espace comme décrit précédemment, vérifiant en outre l'hypothèse suivante : il existe un réel σ tel que pour toute partie $A \subset \Omega$ de volume supérieur à $\frac{1}{2}$, on ait*

$$\frac{Vol\left(A^t\right)}{Vol\,\Omega} \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Alors, pour toute fonction $f : \Omega \to \mathbb{R}$ vérifiant $|f(x) - f(y)| \leqslant d(x,y)$, il existe une constante M telle que :

$$Vol\left(\{x \in \Omega \,/\, |f(x) - M| \geqslant t\}\right) \leqslant 4\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Avant de nous lancer dans la démonstration de ce théorème, faisons un commentaire. L'hypothèse dit que toute partie de volume suffisamment grand a un périmètre petit : c'est exactement l'inégalité isopérimétrique. Le paramètre σ est à interpréter comme une « taille caractéristique » de l'espace. La conclusion, quant à elle, est en fait exactement celle des théorèmes 8 et 9 : c'est un phénomène de concentration. Essentiellement, le théorème précédent est une formulation rigoureuse de l'implication « isopérimétrie $\Rightarrow$ concentration ». La réciproque est en fait également vraie ; il n'est pas difficile par exemple de se convaincre que si A est une partie de Ω, la fonction $x \mapsto d(x,A)$ vérifie l'hypothèse du phénomène de concentration et c'est ainsi que l'on retrouve l'inégalité isopérimétrique.

Ce théorème fait donc rentrer dans un cadre général les deux cas du cube et de la sphère. C'est en ce sens qu'en grande dimension, le cube discret et la sphère se ressemblent.

Passons maintenant à la preuve du théorème 10. On commence par prendre une fonction $f : \Omega \to \mathbb{R}$ vérifiant $|f(x) - f(y)| \leqslant d(x,y)$ pour tous x et y dans Ω. Le nombre M est défini comme *une* médiane de f, c'est-à-dire un réel vérifiant $\mathrm{Vol}\,\Omega_+ \geqslant \frac{1}{2}$ et $\mathrm{Vol}\,\Omega_- \geqslant \frac{1}{2}$ où par définition :

$$\Omega_+ = \{x \in \Omega \,/\, f(x) \geqslant M\} \qquad \text{et} \qquad \Omega_- = \{x \in \Omega \,/\, f(x) \leqslant M\}$$

Il n'est pas difficile de se convaincre qu'un tel réel existe : la fonction qui à t associe $\mathrm{Vol}\left(\{x \in \Omega \,/\, f(x) \leqslant t\}\right)$ est croissante et on choisit pour M la plus petite valeur pour laquelle elle dépasse $\frac{1}{2}$.

Si l'on applique l'hypothèse d'isopérimétrie à Ω_- on obtient $\mathrm{Vol}\left(\Omega_-^t\right) \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$. Mais l'hypothèse faite sur la fonction f cette fois-ci nous dit que si $f(x) \geqslant M + t$ alors $d(x,\Omega_-) \geqslant t$. Ainsi $\{x \in \Omega \,/\, f(x) \geqslant M + t\}$ est inclus dans Ω_-^t. On en déduit :

$$\mathrm{Vol}\left(\{x \in \Omega \,/\, f(x) \geqslant M + t\}\right) \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

[13] Attention, ce n'est sans doute pas une définition officielle.

[14] Cela ne marche donc par pour $\mathbb{R}^N$.

De même en considérant Ω_+, on trouve :

$$\mathrm{Vol}\left(\{x \in \Omega \,/\, f(x) \leqslant M - t\}\right) \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Finalement en combinant ces deux dernières inégalités, on démontre le théorème.

Le cube discret vérifie l'inégalité isopérimétrique

Pour terminer la preuve du théorème 8, il ne reste plus qu'à montrer que le cube discret vérifie une inégalité isopérimétrique. En fait on prouve successivement les deux lemmes suivants. On rappelle que l'on munit le cube discret d'une notion de volume : si A est une partie de $\{0,1\}^N$, on pose $\mathrm{Vol}(A) = \frac{\mathrm{Card}\, A}{2^N}$.

On utilise dans ce paragraphe une variante de la distance de Hamming, qui augmente de 1 à chaque fois qu'on change une composante. C'est N fois la distance de Hamming définie ci-dessus (cette convention de fixer le diamètre à N simplifie certains calculs). Pour appliquer la conclusion du lemme 2 comme hypothèse du théorème 10, afin de démontrer le théorème 8, il ne faudra pas perdre ce détail de vue.

Lemme 1. *Si A est une partie de $\{0,1\}^N$ et si λ est un réel strictement positif, on a :*

$$\frac{1}{2^N}\sum_{x\in\{0,1\}^N} \exp\left(\lambda d(x,A)\right) \leqslant \frac{1}{Vol(A)}\exp\left(\frac{N\lambda^2}{4}\right)$$

Lemme 2. *Soit A une partie de $\{0,1\}^N$ et soit t un réel strictement positif. Alors, avec les notations introduites précédemment :*

$$Vol\left(A^t\right) \leqslant \frac{1}{Vol(A)}\exp\left(-\frac{t^2}{N}\right)$$

On remarque, dans un premier temps, que le lemme 2 implique directement l'inégalité isopérimétrique dans le cube discret. Nous laissons le soin au lecteur sceptique d'écrire tous les détails.

Maintenant, légèrement plus subtil est de voir que le lemme 2 est une conséquence du lemme 1. Prenons pour cela une partie A de $\{0,1\}^N$ et λ et t deux réels strictement positifs. Dans ces conditions, dire que $d(x,A) \geqslant t$ équivaut à dire que $\exp\left(\lambda d(x,A)\right) \geqslant \exp(\lambda t)$. En sommant ces inégalités sur le lieu où elles sont vraies, on obtient :

$$\exp(\lambda t)\,\mathrm{Vol}\left(A^t\right) = \frac{1}{2^N}\sum_{x\in A^t}\exp(\lambda t) \leqslant \frac{1}{2^N}\sum_{x\in A^t}\exp\left(\lambda d(x,A)\right)$$

Cette dernière quantité est inférieure, d'après le lemme 1, à $\frac{1}{\mathrm{Vol}(A)}\exp\left(\frac{N\lambda^2}{4}\right)$. En regroupant tout, on obtient :

$$\mathrm{Vol}\left(A^t\right) \leqslant \frac{1}{\mathrm{Vol}(A)}\exp\left(\frac{N\lambda^2}{4} - \lambda t\right)$$

On cherche alors le λ qui minimise la quantité de droite : c'est $\lambda = \frac{2t}{N}$ et pour ce λ on obtient l'inégalité annoncée.

Il ne reste plus qu'à prouver le lemme 1. Cela se fait par récurrence sur la dimension. L'initialisation ne pose pas de problème. Voyons l'hérédité. Pour A une partie de $\{0,1\}^{N+1}$, on est amené à estimer la somme :

$$S = \frac{1}{2^{N+1}} \sum_{x \in \{0,1\}^{N+1}} \exp\left(\lambda d\left(x, A\right)\right)$$

Bien sûr pour cela, on privilégie certains indices et se ramène ainsi à des dimensions inférieures. Précisément, on définit les ensembles :

$$\begin{aligned} B_0 &= \left\{ y \in \{0,1\}^N \ / \ (0,y) \in A \right\} \\ B_1 &= \left\{ y \in \{0,1\}^N \ / \ (1,y) \in A \right\} \end{aligned}$$

On définit finalement $B = B_0 \cup B_1$. Les parties B_0 et B_1 sont des tranches, et B une projection. Maintenant, on remarque que si on écrit $x = (\alpha, y)$, on a :

$$d\left(x, A\right) \leqslant \min\left(d\left(y, B_\alpha\right), d\left(y, B\right) + 1\right)$$

et puis :

$$\begin{aligned} S &\leqslant \frac{1}{2^{N+1}} \sum_{\alpha=0}^{1} \sum_{y \in \{0,1\}^N} \exp\left(\lambda d\left(x, A\right)\right) \\ &\leqslant \frac{1}{2^{N+1}} \sum_{\alpha=0}^{1} \sum_{y \in \{0,1\}^N} \min\left(\exp \lambda d\left(y, B_\alpha\right), e^\lambda \exp\left(\lambda d\left(y, B\right)\right)\right) \\ &\leqslant \frac{1}{2^{N+1}} \sum_{\alpha=0}^{1} \min\left(\sum_{y \in \{0,1\}^N} \exp\left(\lambda d\left(y, B_\alpha\right)\right), e^\lambda \sum_{y \in \{0,1\}^N} \exp\left(\lambda d\left(y, B\right)\right) \right) \end{aligned}$$

Grâce à l'hypothèse de récurrence, on peut majorer les deux termes qui apparaissent dans le min. On obtient comme cela :

$$\begin{aligned} S &\leqslant \frac{1}{2} \sum_{\alpha=0}^{1} \min\left(\frac{1}{\mathrm{Vol}\left(B_\alpha\right)} \exp\left(\frac{N\lambda^2}{4}\right), \frac{e^\lambda}{\mathrm{Vol}\left(B\right)} \exp\left(\frac{N\lambda^2}{4}\right) \right) \\ &= \frac{1}{\mathrm{Vol}\left(B\right)} \exp\left(\frac{N\lambda^2}{4}\right) \frac{1}{2} \sum_{\alpha=0}^{1} \min\left(\frac{\mathrm{Vol}\left(B\right)}{\mathrm{Vol}\left(B_\alpha\right)}, e^\lambda \right) \end{aligned}$$

Les nombres $\frac{\mathrm{Vol}(B)}{\mathrm{Vol}(B_0)}$, $\frac{\mathrm{Vol}(B)}{\mathrm{Vol}(B_1)}$ et e^λ sont tous les trois plus grands que 1. En étudiant successivement leur position relative, et en utilisant l'inégalité $\exp\left(\frac{\lambda^2}{4}\right) \geqslant \sqrt{\frac{1}{2}\left(e^\lambda + e^{-\lambda}\right)}$, on montre que l'on a dans tous les cas :

$$\frac{1}{2} \sum_{\alpha=0}^{1} \min\left(\frac{\mathrm{Vol}\left(B\right)}{\mathrm{Vol}\left(B_\alpha\right)}, e^\lambda \right) \leqslant \frac{\exp\left(\frac{\lambda^2}{4}\right)}{\frac{\mathrm{Vol}(B_0)}{\mathrm{Vol}(B)} + \frac{\mathrm{Vol}(B_1)}{\mathrm{Vol}(B)}} = \frac{\mathrm{Vol}\left(B\right)}{\mathrm{Vol}\left(A\right)} \cdot \exp\left(\frac{\lambda^2}{4}\right)$$

la dernière égalité venant du fait que $\mathrm{Vol}\left(B_0\right) + \mathrm{Vol}\left(B_1\right) = \mathrm{Vol}\left(A\right)$ car A s'écrit par définition comme l'union des ensembles disjoints $\{0\} \times B_0$ et $\{1\} \times B_1$. On voit finalement que tout cela mis bout à bout permet de conclure.

Cette dernière preuve se généralise en fait directement au cas d'un espace de la forme X^N, X n'étant donc pas forcément $\{0,1\}$, mais simplement un espace sur lequel une distance et une mesure de masse totale 1 sont données. Pour définir la distance entre $(x_1,\dots,x_N)$ et $(y_1,\dots,y_N)$, on additionne simplement toutes les distances entre x_i et y_i. Une légère difficulté intervient lorsque l'on a à définir une notion de volume sur X^N, cepndant elle est contourne facilement en procédant par récurrence.

Tout cela permet de parler d'inégalité isopérimétrique sur X^N et les lemmes 1 et 2 restent vrais dans cette plus grande généralité, la démonstration étant en tout point semblable, les sommes devant être changées en intégrales pour le plaisir de tous.

Annexe : En physique

La notion d'entropie

La notion d'entropie n'est pas apparue dans les mathématiques, mais plutôt en physique *via* la thermodynamique. On dit souvent qu'elle est censée décrire le désordre d'un système macroscopique, la chose fondamentale étant, en physique, que sous certaines hypothèses (système fermé...) ce désordre ne peut qu'augmenter lorsqu'on relâche une contrainte, par exemple lorsqu'on mélange deux liquides auparavant séparés par la paroi d'un récipient, ou lorsqu'on lâche une assiette.

La notion mathématique que nous avons évoquée correspond en fait exactement à cette mesure du désordre des physiciens. Un système macroscopique est souvent décrit par plusieurs grandeurs directement mesurables, comme la pression, la température, les gaz présents, la forme de l'enceinte, *etc.* Toutefois à l'échelle microscopique, tout un paquet d'états distincts correspondent à ces données. Par exemple, dans une boîte cubique vide (*i.e.* remplie simplement d'air) et fermée, on sait bien que les molécules d'oxygène ou d'azote par exemple ne restent pas gentiment à leur place ; plutôt elles se déplacent (d'autant plus vite que la température est élevée), s'entrechoquent. Bref, toute une vie.

Lorsqu'on fait face à un système physique observé à notre échelle, le nombre de possibilités pour le détail de l'état microscopique est donc très grand. L'entropie physique du système est alors, à une constante multiplicative k près fixant l'unité de mesure de l'entropie en physique, l'entropie correspondant à ce nombre de possibilités (chacune étant considérée comme équiprobable) : $S = k \log(\operatorname{Card} \Omega)$. Le nombre k est ce que l'on appelle la constante de Boltzmann.

Pour finir, disons que l'on comprend bien à ce niveau que l'entropie mesure le désordre du système. Un système désordonné est un système qui a tendance à visiter un grand nombre d'états. De fait, pour savoir dans quel état précis il se trouve à un instant donné, il faudra l'interroger plus longuement, et l'entropie est plus grande.

Les espaces concentrés

Le théorème 9 peut s'interpréter physiquement d'une façon agréable. Si l'on reprend ce que l'on disait précédemment, en général les états possibles sont toujours décrits par l'ensemble des positions et des vecteurs vitesse de chacune des molécules qui constituent le système.

Ces données sont toutefois soumises à certaines contraintes. En particulier, le centre de gravité de toutes les molécules doit avoir un mouvement rectiligne uniforme, et donc si le système est borné en taille, ce centre de gravité doit rester immobile. En outre, l'énergie cinétique totale des particules doit être conservée. Ces deux lois définissent une sphère dans un espace de grande dimension.

Le théorème 9 nous dit que si l'on considère une fonction f qui ne bouge pas trop lorsqu'une molécule ne change quasiment ni de position, ni de vitesse, cette fonction est quasiment constante. La pression est un exemple de telle fonction. En appliquant le résultat, on voit ainsi que les fluctuations de pression dans un système macroscopique quelconque sont de l'ordre de 10^{-12} (inverse de la racine carrée de la constante de Boltzmann), ce qui est, ma foi, fort raisonnable.

Bibliographie commentée

L'absence quasi-totale de probabilités dans les cursus scolaires fait qu'il y a peu d'ouvrages introductifs sur le sujet, hormis des cours de licence ou CAPES. Mentionnons quand même [1].

Ce livre, écrit sur un ton humoristique, rassemble des applications variées du calcul des probabilités à la vie courante : la sincérité des sondages, la surréservation des places d'avion, le stationnement des voitures, le débit d'une file, l'État casino, le loto et ses variantes, la bourse ou la vie.

Pour un livre donnant un traitement mathématique précis des bases des probabilités, en démarrant à un niveau relativement élémentaire (cours, exercices), on pourra consulter [2]. Pour un cours de probabilités plus avancé, on pourra consulter [3].

L'ouvrage de référence, bien écrit, pour la théorie de l'information est [4].

La concentration de la mesure est un sujet très récent. Le seul livre de référence, très technique, est [5].

Références

[1] C. Bouzitat, G. Pagès, *En passant par hasard*, Vuibert, 1999

[2] D. Foata, A. Fuchs, *Calcul des probabilités*, Dunod

[3] A.N. Shiryaev, *Probability*, Graduate Texts in Mathematics, **95**, Springer, 1984

[4] M. Cover, J.A. Thomas, *Elements of information theory*, Wiley, 1991

[5] M. Ledoux, *The concentration of measure phenomenon* American Mathematical Society, 2001