Aspects de l'entropie en mathématiques

Yann Ollivier

Ces notes sont librement inspirées d'un séminaire du laboratoire de mathématiques de l'ENS-Lyon, tenu au château de Goutelas en avril 2002, et qui portait sur ce même sujet. Je remercie très vivement les organisateurs et l'ensemble du laboratoire pour l'accueil très chaleureux qu'ils m'ont réservé à cette occasion, ainsi, bien évidemment, que les orateurs ayant inspiré ces textes.

On traitera successivement de théorie de l'information, de systèmes dynamiques, de probabilités et finalement de physique. Il est conseillé de lire les bases de la théorie de l'information avant de passer aux autres parties, qui sont indépendantes.

Table des matières

Chapitre 1

La théorie de l'information : l'origine de l'entropie

La théorie de l'information est due à Shannon (vers 1948), avec bien sûr l'influence des grands théoriciens de l'informatique (Turing, von Neumann, Wiener). À noter des convergences avec les travaux de Fisher.

Le problème est celui de la communication entre une source et un récepteur : la source émet un message que le récepteur lit. On voudrait quantifier l'« information » que contient chaque message émis. Par exemple, il est clair que si l'émetteur dit toujours la même chose, la quantité d'information apportée par une répétition supplémentaire est nulle.

Le cas le plus simple est le suivant : le récepteur attend une information de type *oui/non*, le oui et le non étant a priori aussi vraisemblables l'un que l'autre. Lorsque la source transmet soit un *oui* soit un *non*, on considère que le récepteur reçoit une unité d'information (un *bit*). Autrement dit : une unité d'information, c'est quand on a a priori un ensemble de deux possibilités, et que l'une d'elles se réalise.

L'entropie existe en version combinatoire, en version de probabilités discrètes ou encore en probabilités continues (ce dernier thème étant très proche de problématiques d'analyse). On commence par la première.

1.1 Information combinatoire

Que se passe-t-il si on a plus de possibilités ? Supposons d'abord qu'on a un ensemble de possibilités Ω, et que le message consiste à spécifier un élément de Ω. Si tous les éléments de Ω sont aussi vraisemblables a priori (on verra ci-dessous ce qui se passe si on dote Ω d'une probabilité), quelle est l'information transmise par le message « telle possibilité s'est réalisée » ?

Si Ω a deux éléments, on transmet une information d'une unité. Si Ω comporte 2^n éléments, on peut spécifier un élément de Ω en donnant n informations élémentaires (par exemple par dichotomie de type « moitié de droite / moitié de gauche » ou bien en numérotant les éléments de Ω et en donnant la décomposition en base 2). On a donc envie de dire que spécifier un élément parmi un ensemble Ω de possibilités revient à transmettre $\log_2 |\Omega|$ unités d'information. (Désormais dans cette partie, tous les logarithmes seront implicitement pris en base 2.)

À noter que la quantité d'information n'est pas une propriété intrinsèque d'un certain objet, mais une propriété de cet objet en relation avec un ensemble de possibilités dans lequel on considère qu'il se trouve : comme l'entropie en physique, la quantité d'information est une notion relative à la connaissance préalable de l'observateur du système, du récepteur du message.

À ce stade, en notant $I_\Omega(x)$ la quantité d'information de l'événement x appartenant à l'ensemble Ω, on a donc

$$I_\Omega(x) = \log |\Omega|$$

Avant d'en venir à un cadre probabilisé, regardons quelle est l'information d'une phrase telle que « l'événement réalisé appartient à un sous-ensemble A de l'ensemble Ω des possibilités ». On applique le principe intuitif suivant : si on dit que l'événement réalisé appartient à une partie A, puis qu'on spécifie ensuite de quel événement de A il s'agit, on a totalement spécifié l'événement, comme si on l'avait donné directement dès le début. Spécifier directement l'événement réalisé, c'est transmettre $\log |\Omega|$ unités d'information. Spécifier un événement en sachant déjà qu'il appartient à un sous-ensemble A, peut se faire en transmettant $\log |A|$ unités d'information. On en déduit qu'en précisant que l'événement appartient à A, on avait déjà transmis $\log |\Omega| - \log |A|$ unités d'information d'où

$$I_\Omega(A) = \log |\Omega| \,/\, |A|$$

1.2 Cadre probabiliste

Supposons maintenant que toutes les possibilités ne sont pas équiprobables mais qu'on sait que certaines, a priori, apparaîtront plus souvent que d'autres. L'idée est que les événements plus rares contiennent plus d'information (exemple : complétez les mots français de cinq lettres Z____E et E____E).

On peut s'inspirer de la version combinatoire ci-dessus : on sait que l'appartenance à une partie A dans un ensemble Ω est un événement de $\log |\Omega| \,/\, |A|$ unités d'information. Si on suppose qu'Ω est un ensemble probabilisé où tous les événements sont équiprobables, la probabilité de la partie A est

$|A| / |\Omega|$; l'information apportée par la réalisation d'un événement de A est donc $-\log p(A)$.

Si désormais (Ω, p) est un espace probabilisé, et $A \subset \Omega$ un événement, on pose donc

$$I_\Omega(A) = -\log p(A)$$

et en particulier, pour un élément x

$$I_\Omega(x) = -\log p(x)$$

Cette définition a bien la propriété qu'on attendait, à savoir que la survenue d'un événement rare contient plus d'information. Inversement, la survenue d'un événement certain (de probabilité 1) n'apporte aucune information.

La quantité d'information dépend plus de la distribution de probabilité que d'un événement x particulier. On va donc définir l'*entropie* d'une distribution de probabilité : c'est l'information moyenne qu'on obtient si on tire un élément de Ω suivant la probabilité p :

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

Revenons au modèle de l'émetteur et du récepteur : on suppose qu'à chaque instant, l'émetteur envoie une lettre a de l'alphabet avec la probabilité $p(a)$; l'entropie est alors la quantité moyenne d'information apportée par chaque nouvelle lettre transmise, ou encore, l'incertitude moyenne sur la prochaine lettre qui va arriver.

L'entropie est maximale quand toutes les possibilités sont a priori équiprobables ; s'il y a n possibilités, l'entropie est alors $\log n$. Inversement, si la mesure est concentrée en un point de probabilité 1, alors un tirage sous cette loi n'apporte aucune information car le résultat est connu d'avance, l'entropie est nulle.

L'entropie d'une loi de probabilité est ainsi une mesure de sa dispersion.

La formule ci-dessus est celle obtenue par Boltzmann (à un facteur k près, la constante de Boltzmann, qui est un changement de l'unité d'information) pour la thermodynamique : Boltzmann suppose qu'on a un système composé de N particules indiscernables, et on sait que la proportion de particules se trouvant dans l'état i est p_i. Quelle est la quantité d'information apportée par la spécification complète de l'état microscopique des particules, autrement dit, la liste qui pour chaque particules dit dans quel état elle se trouve ? Le nombre de possibilités de répartir les N particules en respectant les proportions est

$$|\Omega| = \frac{N!}{(p_1 N)!\,(p_2 N)! \ldots (p_k N)!}$$

et la quantité d'information moyenne par particule est

$$S = \frac{1}{N} \log |\Omega| \sim -\sum p_i \log p_i$$

quand N est grand. (On a simplement utilisé $\log N! \sim N \log N$.)

Si toutes les probabilités sont égales à $1/|\Omega|$, on trouve en particulier la formule célèbre $S = \log |\Omega|$ qui (avec la constante de proportionnalité k) est gravée sur la tombe de Boltzmann. Dans ce cas, l'entropie de la distribution de probabilité est bien sûr égale à la quantité d'information apportée par chaque événement particulier. (Ce qui est à l'origine d'un certain nombre de confusions, du fait que l'entropie est une notion globale définie pour une mesure de probabilité, et que parler de l'entropie d'un état particulier n'a pas de sens. En physique, l'entropie d'un état macroscopique donné est en fait l'entropie de la mesure uniforme sur tous les états microscopiques donnant cet état macroscopique.)

Tous ces raisonnements intuitifs sont justifiés par le théorème suivant, dû à Shannon, sans doute le premier théorème de théorie de l'information :

Théorème 1.1. *La famille de fonctions $S_n(p_1, p_2, \ldots, p_n) = -\sum p_i \log p_i$ est la seule vérifiant :*

- *S_n est symétrique en ses arguments, positive, continue*
- *$S_2(1/2, 1/2) = 1$*
- *$S_n(p_1, \ldots, p_n) = S_{n-1}(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2) S_2(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$*

Le dernier axiome est un axiome de regroupement, qu'on avait déjà utilisé ci-dessus pour calculer la quantité d'information apportée par une partie $A \subset \Omega$. Supposons par exemple que le message soit constitué de lettres, mais qu'on confonde les lettres E et F. Alors la quantité d'information reçue à la transmission est seulement $S_{25}(p(E) + p(F), p(A), \ldots p(Z))$ et pour reconstituer le message, il faut, dans une proportion $p(E) + p(F)$ des cas, demander une information supplémentaire qui départage entre E et F, ce qui transmet exactement $S_2(\frac{p(E)}{p(E)+p(F)}, \frac{p(F)}{p(E)+p(F)})$ unités d'information.

1.3 Entropie d'une suite de variables

Si X et Y sont deux variables aléatoires, on peut définir l'entropie du couple X, Y : $S = -\sum p_{ij} \log p_{ij}$. On peut bien sûr généraliser à un nombre quelconque de variables.

Il résulte de la définition que si X et Y sont deux variables aléatoires indépendantes, on a

$$S(X, Y) = S(X) + S(Y)$$

autrement dit, pour des variables indépendantes, l'information conjointe est égale à la somme des informations.

Ceci n'est pas toujours vrai : il se peut que la variable X contienne de l'information sur ce que va être Y. Par exemple, si on répète toujours deux fois de suite la même chose, l'information de la deuxième répétition est nulle : $S(X, X) = S(X)$. En fait on a toujours

$$S(X, Y) \leqslant S(X) + S(Y)$$

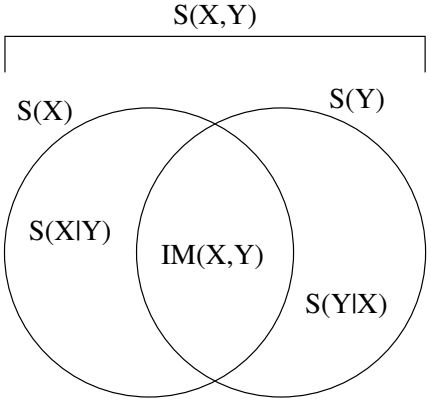
avec égalité si et seulement si X et Y sont indépendants. Ceci peut servir par exemple à définir une information partielle

$$S(Y|X) = S(X, Y) - S(X)$$

c'est la quantité d'information réellement apportée par Y si on connaît déjà X, ainsi qu'une information mutuelle

$$IM(X, Y) = S(X) + S(Y) - S(X, Y)$$

c'est la quantité d'information présente « en double » dans X et dans Y.



Revenons à une source qui émet régulièrement des signaux. Soit X_t le signal émis au temps t, c'est une variable aléatoire à valeurs dans un certain alphabet. Ces variables ne sont pas forcément indépendantes (cas d'un texte en français : chaque lettre dépend des précédentes). L'entropie de la suite infinie $X = (X_1, \ldots, X_n, \ldots)$ est définie par

$$S(X) = \lim_{n\to\infty} \frac{1}{n} S(X_1, \ldots, X_n)$$

si cette limite existe, où $S(X_1, \ldots X_n)$ est l'entropie de la distribution jointe du n-uplet $(X_1, \ldots, X_n)$.

Si les lettres successives du message sont toutes indépendantes et équidistribuées (la loi de X_n est égale à la loi de X_1 pour tout n), alors on a bien sûr

$$S(X_1, \ldots, X_n, \ldots) = \lim \frac{1}{n} S(X_1, \ldots, X_n) = S(X_1)$$

Autrement dit, dans le cas d'une source qui émet des lettres successives et indépendantes toujours avec la même loi, la quantité d'information moyenne pour chaque nouvelle lettre, est égale à l'entropie de la loi.

1.4 Entropie et codages

1.4.1 Codage sans bruit

La base des définitions ci-dessus était la constatation qu'un élément d'un ensemble de taille 2^n peut être codé par n unités d'information. C'est le premier lien entre théorie de l'information et codage.

Un *codage* sur un alphabet A est une application injective de A vers l'ensemble des mots sur un alphabet B (souvent $B = \{0, 1\}$). On dit que le codage est *instantanément décodable* s'il n'existe pas deux lettres a, a' de A telles que le code de a soit un préfixe du code de a'. Tous les codages ci-dessous seront supposés instantanément décodables.

On étend le codage aux mots sur A par concaténation des codes des lettres. La propriété d'être instantanément décodable garantit alors que le codage des mots est non ambigu.

Le problème est le plus souvent de trouver les codages les plus courts possibles. La stratégie de base consiste à attribuer des codes courts aux lettres fréquentes, et des codes longs aux lettres moins fréquentes.

Soient $X_1, \ldots, X_n, \ldots$ des variables aléatoires identiquement distribuées à valeurs dans un ensemble A. Soit C un codage de A utilisant un alphabet B. On définit la longueur moyenne du codage :

$$L(C) = \lim_n \frac{1}{n} \mathbb{E}\ell(C(X_1 \ldots X_n))$$

où ℓ représente la longueur d'un mot sur l'alphabet B. On a alors le théorème de codage de Shannon, qui identifie entropie et taux de compression maximal :

Théorème 1.2. *Soit C un codage instantanément décodable pour la suite de variables $X_1, \ldots, X_n, \ldots$. Alors*

$$L(C) \geqslant \frac{S(X_1, \ldots, X_n, \ldots)}{\log |B|}$$

De plus, on peut construire des codages ayant des longueurs arbitrairement proches de cette valeur.

La preuve qu'aucun code ne fait mieux que cette valeur est simple : c'est essentiellement le fait que si $\sum p_i = \sum q_i = 1$, alors $\sum p_i \log q_i$ est maximal quand $p_i = q_i$, et la remarque que si les ℓ_i sont les longueurs des codes d'un codage instantanément déchiffrable, on a $\sum 2^{-\ell_i} \leqslant 1$.

L'idée de la preuve que l'optimum peut être approché est la suivante : les mots de longueur n obtenus par tirage des X_i peuvent être décomposés en deux classes, une classe de mots « typiques » et une classe de mots « rares ». Les mots typiques sont environ au nombre de 2^{nS}, chacun d'entre eux étant de probabilité environ 2^{-nS} ; les mots rares ont une probabilité totale négligeable. Pour coder dans $\{0, 1\}$ un mot typique, on met un « 0 » suivi du code en base 2 du numéro du mot typique parmi les 2^{nS} mots typiques (ce qui prend nS chiffres en base 2) ; pour un mot rare, on met un « 1 » suivi simplement du code en base 2 du numéro du mot rare parmi l'ensemble des mots possibles sur l'alphabet de départ.

Un codage plus simple, proche du codage optimal, est le codage de Shannon-Fano : il consiste à attribuer à la lettre a un code en base 2 de longueur $-\log p(a)$ (arrondi à l'entier supérieur), ce qui est toujours possible : par exemple, arranger les probabilités de manière croissante sur l'intervalle $[0; 1]$, ce qui donne une partition en sous-intervalles, si un intervalle est de longueur p_i on peut trouver un nombre binaire à $-\log p_i$ chiffres qu'on lui associe, de manière à former un codage sans préfixes. Autrement dit on code les lettres plus fréquentes par des mots plus courts.

1.4.2 Codage avec bruit

On s'intéresse désormais au codage par des canaux de communication qui peuvent introduire des erreurs. Soit X le message à la source, Z le message codé transmis dans un canal (qui est une fonction aléatoire de X). Soit φ une fonction de décodage, on veut que $\varphi(Z) = X$ le plus souvent possible, mais le passage de X vers Z n'est pas forcément injectif. Le théorème suivant, dû à Fano, affirme que la probabilité minimale d'erreur est liée à l'entropie :

Théorème 1.3. *Pour toute fonction de décodage φ, la probabilité que $\varphi(Z)$ soit différent de X est supérieure ou égale à*

$$\frac{S(X|Z) - 1}{\log |X|}$$

où $|X|$ est la taille de l'alphabet de X.

Un canal de communication binaire étant donné, avec des probabilités d'erreur, il y a un arbitrage à faire entre concision du codage et probabilité de décodage correct : si on prend un codage sans aucune redondance, le décodage est très sensible à toute erreur ; si on répète trois fois chaque mot, on a de meilleures chances d'être compris.

Supposons donc qu'un émetteur fait transiter une lettre X d'un alphabet A au travers d'un canal binaire $\mathcal{C}$. Auparavant, il code X par une suite binaire Y de longueur ℓ. Le récepteur, lui, reçoit à la sortie du canal une suite binaire Z qui est une fonction aléatoire de la suite binaire Y donnée à l'entrée du canal.

On définit la capacité d'un canal $\mathcal{C}$ par $Cap(\mathcal{C}) = \sup IM(Y,Z)/S(Y)$, le sup étant pris sur toutes les lois de probabilité possibles pour le mot binaire Y. Cette quantité ne dépend que du canal.

Par exemple, imaginons que le canal transmette des 0 et des 1, mais, avec une faible probabilité p, change un 0 en 1 ou un 1 en 0. La capacité est alors $1 + p\log p + (1-p)\log(1-p) \leqslant 1$. Dans le cas où $p = \frac{1}{2}$, aucune information ne peut être tirée, la capacité est nulle.

Un canal étant donné, on peut se demander quel codage adopter. Soit A l'alphabet de départ, un codage binaire est une application $C : A \to \{0,1\}^\ell$ pour une certaine longueur ℓ (pour simplifier, on considère des codages à longueur constante). On définit la probabilité d'erreur P_e d'un codage C (avec fonction de décodage D) par $P_e(C) = \sup_{X\in A} P(D(Z) \neq X | Y = C(X))$. On définit aussi le taux de compression du codage par $R(C) = (\log |A|)/\ell$.

Shannon montre alors le théorème suivant, qui identifie la capacité du canal avec le taux de compression optimal :

Théorème 1.4. *La capacité $Cap(\mathcal{C})$ d'un canal $\mathcal{C}$ est égale au sup des nombres R tels que, pour tout $\varepsilon > 0$, il existe un alphabet A, une longueur ℓ, un codage $C : A \to \{0,1\}^\ell$ tels que $R = R(C)$ et $P_e(C) \leqslant \varepsilon$.*

La preuve de ce théorème est hautement non constructive (on choisit le codage au hasard !), et comme ci-dessus elle utilise des mots « typiques ». Ce domaine de recherche est encore très ouvert.

1.4.3 Codage continu

On s'intéresse désormais à la situation où on cherche à transmettre une quantité continue (intensité, voltage...). Cette fois-ci on a donc des lois de probabilité sur $\mathbb{R}$. Par analogie avec le cas discret, l'entropie d'une distribution de probabilité f sur $\mathbb{R}$ est égale à $-\int_{x\in\mathbb{R}} f(x)\log f(x)dx$.

On remarque que, si on multiplie par c la variable transmise par le canal, on ajoute $\log c$ à l'entropie de l'information transmise (ce qui est naturel : en multipliant par 2 on a une précision deux fois plus grande, ce qui donne un bit d'information en plus). Pour obtenir des résultats pertinents, on suppose en général que les canaux utilisés sont limités en puissance, par exemple qu'ils ne peuvent pas transmettre une variable dont la variance est supérieure à un certain seuil (sans cette hypothèse, on peut facilement transmettre une quantité infinie d'information).

Un simple calcul variationnel montre que, parmi les distributions de probabilité de variance fixée, les gaussiennes sont celles d'entropie maximale. En effet, soit f une fonction maximisant $-\int f \log f$, et calculons l'entropie d'une fonction voisine $f + \delta f$ de même variance. Comme f et $f + \delta f$ sont des mesures de probabilité, donc d'intégrale 1, on a $\int \delta f = 0$. On peut supposer que f et δf sont de moyenne nulle (par translation), et donc $\int t\delta f(t) = 0$. Alors le fait que $f + \delta f$ ait même variance que f s'écrit $\int t^2 \delta f(t) = 0$. Maintenant, la variation d'entropie est $-\delta \int f \log f = -\int \delta f \log f - \int f\delta \log f = -\int \delta f \log f - \int f\delta f/f = -\int \delta f \log f$. Pour tout δf vérifiant $\int \delta f = \int t\delta f(t) = \int t^2 \delta f(t) = 0$, on doit donc avoir $\int \delta f \log f = 0$ si f est un extrémum d'entropie. Cela implique $\log f(t) = A + Bt + Ct^2$, d'où la gaussienne.

Intéressons-nous au cas où on cherche à transmettre une information X à travers un canal limité en puissance, la limite étant p (autrement dit l'espérance de X^2 doit être inférieure à p). Mais ce canal est bruité ; plus exactement, on a un ennemi qui a accès à ce canal et qui peut transmettre du bruit Y (indépendant de X), la puissance du bruit transmis étant inférieure à p'. Le récepteur reçoit $X + Y$, qui fait perdre de l'information par rapport à X. Ce qui intéresse le transmetteur est de maximiser $IM(X + Y, X)$ à Y donné, tandis que l'ennemi cherche à minimiser cette quantité à X donné (si chacun connaît la stratégie appliquée par l'autre).

Théorème 1.5. *Il y a un équilibre de Nash à ce jeu, qui vérifie :*

$$\inf_Y \sup_X IM(X+Y,X) = \sup_X \inf_Y IM(X+Y,X) = \frac{1}{2}\log(1+\frac{p}{p'})$$

et cet équilibre consiste pour chacun à utiliser des variables gaussiennes de variances p et p'.

À noter que même si $p' > p$, il reste encore quelque chose du message initial.

La preuve utilise des inégalités fines. Par analogie avec le fait que les variances de variables indépendantes s'ajoutent, on définit la puissance-entropie

$N(X)$ d'une variable X comme la variance qu'aurait une gaussienne de même entropie que X. (On vérifie qu'en dimension d, on a $N = \exp(2S/d)/2\pi e$.)

On a alors l'inégalité de puissance-entropie de Shannon pour des variables indépendantes : $N(X+Y) \geqslant N(X) + N(Y)$. Pour conserver l'information on doit monter en puissance... Autre forme équivalente : pour $0 < \lambda < 1$, si X et Y sont des variables aléatoires indépendantes, alors $S(\sqrt{\lambda}\, X + \sqrt{1-\lambda}\, Y) \geqslant \lambda S(X) + (1-\lambda) S(Y)$. Ceci peut servir à montrer par exemple que si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes identiquement distribuées, alors $(X_1 + \cdots + X_n)/\sqrt{n}$ ressemble à une gaussienne (au moins si n est une puissance de 2)...

Ces inégalités n'ont pas été rigoureusement démontrées par Shannon (elles le sont désormais). Elles sont à rapprocher, par exemple, de l'inégalité de Brunn-Minkowski, ou encore à des problèmes de constantes optimales dans l'inégalité de convolution de Young $\|f * g\|_{L^r} \leqslant C_{pqr} \|f\|_{L^p} \|g\|_{L^q}$ pour $1 + 1/r = 1/p + 1/q$...

Le sujet est donc loin d'être clos.

Chapitre 2

Les entropies de systèmes dynamiques

Un système dynamique est une application d'un espace dans lui-même, que l'on itère. On s'intéresse à des propriétés telles que l'existence de points fixes, périodiques, la caractérisation des orbites denses, la recherche de quantités invariantes, la divergence d'orbites partant de points proches, etc. On considère généralement que l'espace a une structure supplémentaire : une structure topologique, ou bien, en théorie ergodique, une mesure de probabilité.

L'idée de la définition de l'entropie d'un système dynamique est la suivante : on considère que la position initiale du système n'est pas connue avec une précision infinie, mais que le comportement qu'on va observer en itérant le système va nous renseigner de mieux en mieux sur le point dont on est parti (par exemple, à chaque étape, on sait dire si on se trouve dans la moitié droite ou gauche de l'espace ; dans beaucoup de cas, cette information sur l'ensemble de la trajectoire permet de reconstituer le point de départ). La quantité moyenne d'information qu'on gagne à chaque itération est l'entropie du système.

On traite successivement l'entropie dans les cadres ergodique et topologique.

2.1 L'entropie ergodique

Soit X un espace doté d'une mesure μ de masse 1. Un système dynamique ergodique sur X est alors une application mesurable $T : X \to X$ préservant la mesure, c'est-à-dire que pour toute partie $A \subset X$ (mesurable), on a $\mu(T^{-1}A) = \mu(A)$. L'application T n'est pas nécessairement inversible.

Quelques exemples :

— Une rotation sur le cercle préserve la mesure d'angle $d\theta$.
— Soit $A \in SL_2(\mathbb{Z})$ une matrice 2×2 à coefficients entiers, de déterminant 1. Elle agit sur $\mathbb{R}^2$ et l'action se factorise sur le tore $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$, et cette action préserve la mesure de Lebesgue.
— Le décalage de Bernoulli agit par décalage sur les mots infinis sur un alphabet $\{a_1, \ldots, a_m\}$: si $x = (x_0, x_1, x_2, \ldots) \in \{a_1, \ldots, a_m\}^{\mathbb{N}}$, on pose $B(x) = (x_1, x_2, \ldots)$. Cette action préserve les mesures produits où on assigne à la lettre a_i une probabilité p_i, et à un mot le produit des probabilités de ses lettres.

On dit que T est *ergodique* si toute partie de X invariante par T est de mesure soit 0, soit 1 (si ce n'est pas le cas, on décompose).

On dit que deux systèmes ergodiques $T : (X, \mu) \to (X, \mu)$ et $T' : (X', \mu') \to (X', \mu')$ sont *mesurablement équivalents* s'il existe une bijection mesurable $\varphi : X \to X'$ (modulo des ensembles de mesure nulle dans X et X') qui envoie la mesure μ sur μ' et qui envoie l'action sur l'action, i.e. $T' \circ \varphi = \varphi \circ T$.

Un des principaux buts de la théorie est d'essayer de classer les systèmes ergodiques à équivalence près.

2.1.1 Motivation : les invariants spectraux

Von Neumann a défini toute une classe d'invariants de systèmes ergodiques : les invariants spectraux. On peut évidemment se demander si ces invariants suffisent à déterminer la dynamique à équivalence près. C'étaient les seuls invariants connus jusqu'à l'introduction de l'entropie ergodique par Kolmogorov.

L'idée est de faire agir T sur des espaces de fonctions sur X en envoyant une fonction f sur $f \circ T$. En particulier, ceci définit un opérateur $U_T : L^2(X) \to L^2(X)$, et comme T conserve la mesure, cet opérateur est une isométrie de $L^2(X)$.

Les propriétés de cet opérateur permettent de capturer une partie du comportement du système. Par exemple, le fait que T soit ergodique est équivalent au fait que $\dim \mathrm{Ker}(U_T - \mathrm{Id}) = 1$. Le fait que T soit mélangeant (i.e. pour toutes parties $A, B \subset X$, on a $\lim \mu(A \cap T^{-n}B) = \mu(A)\mu(B)$) est équivalent au fait que pour toutes fonctions $f, g \in L^2(X)$, on a $\lim \int f.U_T^n g = \int f \,.\, \int g$.

Dans le cas d'une matrice de $SL_2(\mathbb{Z})$ agissant sur le tore $\mathbb{T}^2$, la base de Fourier de L^2 permet de calculer explicitement l'opérateur. Par exemple si $A \in SL_2(\mathbb{Z})$, la transformation induite est ergodique si et seulement si le spectre de A ne contient pas de racine de l'unité, ou encore si et seulement si toutes les orbites de la transposée ${}^tA$ agissant sur $\mathbb{Z}^2$ sont infinies. On voit

alors sur la base de Fourier, dans cette situation, que toutes les matrices A ergodiques seront spectralement équivalentes. Sont-elles mesurablement équivalentes ?

Pour λ valeur propre de U_T, on note $H_\lambda = \mathrm{Ker}(U_T - \lambda\,\mathrm{Id})$.

Une application ergodique T vérifie alors les propriétés suivantes. Si $f \in H_\lambda$, alors le module de f est constant. De plus, pour tout λ on a $\dim H_\lambda = 1$. Enfin, l'ensemble des valeurs propres de T est un sous-groupe dénombrable du cercle unité de $\mathbb{C}$. (De plus, on peut montrer que tout sous-groupe dénombrable du cercle peut être obtenu ainsi.)

On dit que T est *à spectre purement atomique* si les H_λ engendrent (au sens L^2) l'espace $L^2(X)$.

Von Neumann a démontré que pour des opérateurs à spectre purement atomique, la dynamique est caractérisée par les invariants spectraux :

Théorème 2.1. *Deux applications ergodiques définissant des opérateurs à spectre purement atomique sont mesurablement équivalentes si et seulement si ces opérateurs ont les mêmes valeurs propres.*

À ce stade on ne sait toujours pas si les actions linéaires sur le tore sont mesurablement équivalentes. L'entropie ergodique permet de répondre à cette question.

2.1.2 Définition de l'entropie ergodique

Pour définir l'entropie ergodique, on se donne une partition $\mathcal{P}$ (non triviale) de X. On regarde dans quelle partie de la partition tombent les itérés $T^n x$ d'un point de départ x. L'idée est que cette suite de parties fournit de l'information sur le point x ; l'entropie est alors la quantité d'information moyenne que chaque itération de T apporte.

La suite des parties dans lesquelles tombe $T^n x$ constitue donc une sorte de code de x.

Soit $\mathcal{P}_n(x)$ l'ensemble des points $y \in X$ tels que pour tout $k \leqslant n$, les points $T^k x$ et $T^k y$ sont dans la même partie de la partition $\mathcal{P}$.

On dit que T est *fortement ergodique* si tous les itérés T^n sont ergodiques (une application ergodique n'est pas nécessairement fortement ergodique).

On montre alors facilement que si T est fortement ergodique, alors pour tout x, la mesure $\mu(\mathcal{P}_\infty(x))$ est nulle. Autrement dit, le codage code bien. En effet, soit $Y = \mathcal{P}_\infty(x)$ et supposons $\mu(Y) > 0$. Par le théorème de récurrence de Poincaré, il existe un n tel que $\mu(Y \cap T^{-n}Y) > 0$. Soit donc $y \in Y \cap T^{-n}Y$, alors le code de y est périodique de période n. En particulier, le code de x

est périodique. Donc $Y \subset T^{-n}Y$. Comme T conserve la mesure, on a donc $Y = T^{-n}Y$ ce qui contredit l'ergodicité de T^n.

On voit donc que $\mu(\mathcal{P}_n(x))$ tend vers 0. En fait cette quantité tend exponentiellement vite vers 0, et l'exposant est précisément lié à l'entropie ergodique de T. On sait par la théorie de l'information que donner le code de x dans la partition jusqu'à l'étape n, c'est donner une quantité d'information $-\log \mu(\mathcal{P}_n(x))$. Ceci est précisé par le théorème-définition suivant, énoncé d'abord par Shannon :

Théorème 2.2. *Soit $T : X \to X$ une application préservant la mesure μ. Soit $\mathcal{P}$ une partition de X telle que*

$$-\sum_{P\in\mathcal{P}} \mu(P)\log\mu(P) < \infty$$

Alors la limite

$$h(T,\mathcal{P},x) = -\lim_{n\to\infty} \frac{1}{n}\log\mu(\mathcal{P}_n(x))$$

existe pour μ-presque tout x, et cette quantité, en tant que fonction de x, converge dans $L^1(X)$ vers une fonction T-invariante.

En particulier, si T est ergodique, l'entropie ne dépend pas de x. Sinon, on moyenne en posant $h(T,\mathcal{P}) = \int_X h(T,\mathcal{P},x)d\mu(x)$. Ensuite, on remarque que l'entropie augmente lorsqu'on raffine la partition, on pose donc :

$$h(T) = \sup_{\mathcal{P}} h(T,\mathcal{P})$$

c'est l'entropie ergodique de T.

Par construction, c'est un invariant d'équivalence mesurable.

Pour $A \in SL_2(\mathbb{Z})$ agissant sur le tore, on peut montrer que cette entropie est égale au log du module de la plus grande valeur propre. En particulier, toutes les matrices ergodiques ne sont pas mesurablement équivalentes.

Pour le décalage de Bernoulli, sur un alphabet $\{a_1,\dots,a_m\}$, considérons la partition $\mathcal{P} = \{P^k\}$ où $P^k = \{(x_0,x_1,\dots), x_0 = a_k\}$. On peut raffiner cette partition par le décalage, cela revient à fixer les n premières lettres, on obtient ainsi des partitions arbitrairement fines. L'entropie de ces partitions se calcule facilement : on a $\mathcal{P}_n(x) = \{(y_0,y_1,\dots), x_i = y_i \text{ pour tout } i \leqslant n\}$. Si p_k est la probabilité d'occurrence de la lettre k, la mesure de l'ensemble $\mathcal{P}_n(x)$ vaut alors simplement $\prod_{i=0}^n p_{x_i}$. On a donc $-\frac{1}{n}\log\mu(\mathcal{P}_n(x)) = -\frac{1}{n}\sum \log p_{x_i}$. Or, pour presque tout x, la proportion des x_i qui sont égaux à la lettre a_k est, d'après la loi des grands nombres, p_k. La quantité $-\lim_n \frac{1}{n}\sum \log p_{x_i}$ vaut donc, pour μ-presque tout x :

$$-\sum p_k \log p_k$$

qui est ainsi l'entropie du décalage de Bernoulli de probabilités (p_k). On retrouve donc la vieille formule de Boltzmann...

En fait, un théorème difficile d'Ornstein affirme que deux décalages de Bernoulli (même sur des alphabets n'ayant pas le même nombre de lettres !) sont mesurablement équivalents si et seulement s'ils ont la même entropie. Ce théorème, combiné à un autre de Katznelson qui affirme que toute application de $SL_2(\mathbb{Z})$ agissant sur le tore est mesurablement équivalente à un décalage de Bernoulli (indexé par $\mathbb{Z}$), permet de traiter aussi le cas du tore.

2.2 L'entropie topologique

2.2.1 Définitions

On se place désormais dans un cadre métrique plutôt que mesuré. Soit donc (X, dist) un espace métrique compact, et $f : X \to X$ une application continue. La théorie de l'entropie topologique que l'on développe alors est due à Adler, Konheim, McAndrew.

L'idée est là encore qu'on ne peut séparer les points qu'avec une certaine précision, et qu'on espère que l'observation des trajectoires des points par f nous renseignera sur leur position initiale.

Soit donc $\varepsilon > 0$. On dit que deux points $x, y \in X$ sont ε-séparés si $\mathrm{dist}(x, y) > \varepsilon$. On dit que x et y sont ε-séparés en temps n s'il existe un $k \leqslant n$ tel que $\mathrm{dist}(f^k x, f^k y) > \varepsilon$. Ceci amène naturellement à définir la distance

$$\mathrm{dist}_n(x, y) = \max_{0 \leqslant k \leqslant n} \mathrm{dist}(f^k x, f^k y)$$

Plus n est grand, plus on sépare de points. Soit $H(n, \varepsilon)$ le nombre maximum de points d'une famille de points deux à deux (n, ε)-séparés. Combinatoirement, identifier l'un de ces points est donner une information $\log H(n, \varepsilon)$. On définit l'*entropie topologique* de f par

$$h_{top}(f, \varepsilon) = \overline{\lim_{n \to \infty}} \frac{1}{n} \log H(n, \varepsilon)$$

et

$$h_{top}(f) = \lim_{\varepsilon \to 0} h_{top}(f, \varepsilon)$$

(À noter que $h_{top}(f, \varepsilon)$ croît quand ε décroît, cette limite est donc bien définie.)

On aurait pu donner une variante de cette définition en posant pour $H(n, \varepsilon)$ le nombre minimal de boules de rayon ε pour dist_n recouvrant tout X. On trouve la même entropie.

Une autre manière de voir est de considérer le graphe $\Gamma_k = \{(x, fx, \ldots, f^k x)\} \subset X^{k+1}$ et de compter le nombre de pavés de côté ε nécessaires pour le recouvrir.

Comme X est compact, deux métriques quelconques donnant la même topologie sont uniformément équivalentes. Cela implique que l'entropie définie ci-dessus ne dépend pas de la métrique choisie, d'où son qualificatif de topologique.

L'entropie topologique est liée à l'entropie ergodique définie plus haut :

$$h_{top}(f) = \sup\{h(f, \mu), \mu \text{ mesure de probabilité } f\text{-invariante sur } X\}$$

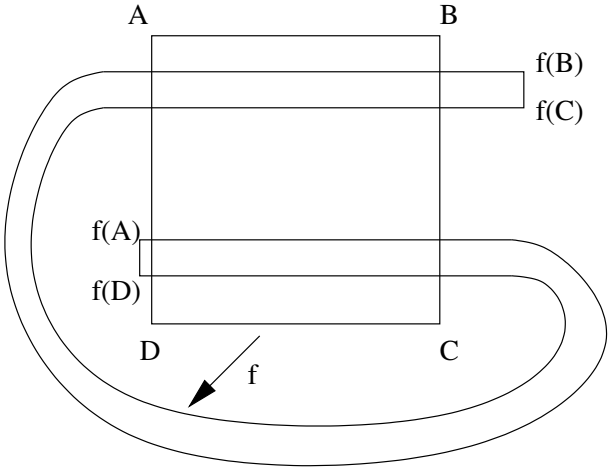
On peut de plus montrer que si f est un difféomorphisme C^∞ d'une variété, ce sup est atteint.

Premier exemple : l'entropie d'une rotation du cercle est nulle, comme celle de toute isométrie.

L'entropie de l'application du cercle unité de $\mathbb{C}$ définie par $z \mapsto z^2$ est égale à $\log 2$: en effet on a $\text{dist}_n = 2^n \,\text{dist}$, on sépare deux fois mieux les points à chaque itération. De manière plus générale, l'entropie de $z \mapsto z^k$ sur le cercle est égale à $\log k$.

Soit la matrice $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ agissant sur le tore $\mathbb{T}^2$. On a une valeur propre dilatante $\frac{3+\sqrt{5}}{2}$, et une étude locale montre que l'entropie est égale au log de cette valeur.

De manière générale, si un système dynamique (ou l'un de ses itérés) possède une figure topologiquement équivalente à un « fer à cheval », c'est-à-dire un carré dont l'image par f l'intersecte deux fois (dans la bonne direction), alors l'entropie topologique sera strictement positive.



En effet dans cette situation, on a un ensemble limite composé d'une infinité de bandes dans le carré, et spécifier un point sur une bande demande de spécifier, pour chaque étape, si on choisit la partie haute ou la partie basse.

Inversement, un théorème de Katok affirme que si f est un difféomorphisme C^∞ d'une surface compacte, d'entropie strictement positive, alors f ou l'un de ses itérés possède un fer à cheval.

2.2.2 Entropie topologique et entropie algébrique

Dans cette situation, on peut définir d'autres invariants à l'aide de l'idée d'entropie. L'un d'eux est l'entropie algébrique.

Soient X une variété compacte lisse et f une application C^∞. Elle induit un morphisme sur le groupe fondamental de X, soit $f_* : \pi_1(X) \to \pi_1(X)$. (Mettons pour simplifier qu'il existe un point périodique, qu'on prend comme point-base du π_1.)

Le groupe fondamental $\pi_1(X)$ est engendré par une partie génératrice $S = \{a_1, \ldots, a_k\}$ (on prend S symétrique, c'est-à-dire que S contient les inverses de ses éléments), ces éléments vérifiant certaines relations. Alors, tout élément du π_1 peut être écrit comme un produit d'éléments de S. On définit la longueur $\ell(x)$ d'un élément $x \in \pi_1(X)$ comme le nombre minimal d'éléments de S qu'il faut pour l'écrire.

On pose alors

$$h_{\pi_1}(f) = \overline{\lim_{n\to\infty}} \frac{1}{n} \log \max_{1\leqslant i\leqslant k} \ell(f_*^n a_k)$$

c'est l'*entropie algébrique* de f.

L'entropie ne dépend pas du système S de générateurs choisi. Cela se voit en regardant la longueur des éléments d'une nouvelle partie génératrice par rapport à l'ancienne, et en utilisant la relation $\ell(xy) \leqslant \ell(x) + \ell(y)$.

De même, cette entropie est invariante par automorphisme intérieur du π_1 (conjugaison par un certain élément), ce qui implique que cette définition ne dépend pas du point-base choisi.

Un théorème de Manning précise le rapport entre entropie topologique et entropie algébrique :

Théorème 2.3. *Soit f une application C^∞ sur une variété lisse. Alors*

$$h_{top}(f) \geqslant h_{\pi_1}(f)$$

A priori, l'action sur le groupe fondamental ne capture donc qu'une partie de la complexité de la dynamique.

On peut travailler sur l'homologie comme sur le groupe fondamental. L'application f définit un opérateur sur l'homologie $f_* : H_*(X, \mathbb{Z}) \to H_*(X, \mathbb{Z})$.

L'analogue de l'entropie algébrique est alors le log du rayon spectral $\rho(f_*) = \overline{\lim} \|f_*^n\|^{1/n}$ (comparer avec le cas d'une application linéaire sur le tore), et on a un analogue, dû à Yomdin, du théorème de Manning :

Théorème 2.4. *Soit $f : X \to X$ un difféomorphisme C^∞ d'une variété lisse, et soit ρ le rayon spectral de f_*, alors*

$$h_{top}(f) \geqslant \log \rho = \lim \frac{1}{n} \log \|f_*^n\|$$

Si X est de dimension n, on peut simplement restreindre f_* à l'homologie en degré n de X, soit $H_n(X, \mathbb{Z}) \simeq \mathbb{Z}$. Le rayon spectral correspondant est alors simplement le degré topologique d de f, et on a le théorème suivant :

Théorème 2.5. *Soit $f : X \to X$ une application C^1 sur une variété lisse, de degré d. Alors*

$$h_{top}(f) \geqslant \log d$$

Attention, l'hypothèse de régularité C^1 est nécessaire ! Par exemple sur $\mathbb{C}$, si on considère l'application donnée en coordonnées polaires par $\rho e^{i\theta} \mapsto \frac{1}{2}\rho e^{2i\theta}$, son degré est 2, mais toutes les orbites tendent vers 0 donc l'entropie est nulle. Bien sûr, cette application n'est pas C^1 en 0...

L'idée de la preuve est la suivante : on prend un point et on regarde l'ensemble de ses préimages par f au temps n, il y en a d^n pour un point typique. L'hypothèse de régularité C^1 intervient pour dire que tous ces points sont bien séparés (par exemple s'il n'y a pas de point critique, le jacobien est uniformément minoré).

2.2.3 Dynamique holomorphe

L'entropie algébrique ne capture donc en général qu'une partie de la complexité d'un système. On peut se demander dans quels cas on a égalité.

Soit f une application polynomiale du plan complexe complété par un point à l'infini : $f : \mathbb{CP}^1 \to \mathbb{CP}^1$. Par exemple, $f : z \mapsto z^2 + c$... Soit d le degré du polynôme, c'est aussi le degré topologique de f et on sait donc que $h_{top}(f) \geqslant \log d$. Un théorème de Gromov pose l'égalité :

Théorème 2.6. *Soit f une application polynomiale de $\mathbb{CP}^1$ dans lui-même, de degré d. Alors*

$$h_{top}(f) = \log d$$

L'idée de la preuve est de regarder le graphe $\Gamma_k = \{(x, fx, f^2x, \dots f^kx)\} \subset (\mathbb{CP}^1)^{k+1}$. On cherche à évaluer le nombre $H'(k, \varepsilon)$ de pavés de taille ε qu'il

faut pour le recouvrir. L'aire de Γ_k est supérieure à ce nombre fois la « densité minimale » de Γ_k dans un pavé de taille ε, c'est-à-dire la plus petite surface qu'on peut y mettre. Or l'aire de Γ_k est calculable par des moyens homologiques, et la densité minimale se trouve être négligeable, ce qui permet d'arriver au résultat.

Ce théorème se généralise à toute variété kählerienne compacte.

Là encore, le sujet est loin d'être clos.

Chapitre 3

L'entropie en probabilités

En physique, l'entropie d'un système hamiltonien à l'énergie E est définie comme le log du volume de l'espace des phases qui est à l'énergie E :

$$S(E) = k \log \mathrm{vol}\{(p,q),\ E \leqslant E(p,q) \leqslant E + \delta E\}$$

On veut définir un analogue en théorie des probabilités. L'idée est que, comme en physique statistique, on va regarder une variable X_N dépendant d'un grand nombre N d'événements élémentaires. La constante de Boltzmann ci-dessus, k, est égale à la constante des gaz parfaits divisée par le nombre de particules impliquées (le nombre d'Avogadro), on est donc tenté de remplacer cette constante par $1/N$. Le volume s'interprète tout naturellement en terme de probabilité, et on pose :

$$H(E) = -\frac{1}{N} \log \mathbb{P}(E \leqslant X_N \leqslant E + \delta E)$$

où on écrit un signe $-$ parce qu'une probabilité est inférieure à 1.

L'idée est qu'on arrive souvent à évaluer l'entropie d'un événement au moyen de la théorie de l'information. Cela fournit alors directement une évaluation de la probabilité d'un événement : presque par définition, un événement apportant une quantité d'information H a une probabilité $\exp -NH$.

On commence par donner l'exemple le plus simple d'une telle situation, avant d'expliquer en termes d'entropie les théorèmes plus généraux.

3.1 Le théorème de Sanov discret

On considère un alphabet fini à n lettres $\Sigma = \{a_1, \ldots, a_n\}$. On se donne une loi de probabilité μ sur cet alphabet et on tire, de manière indépendante, une suite de lettres $X_1, X_2, \ldots, X_N, \ldots$ selon cette loi. La proportion de X_i

qui sont égales à une lettre a_k est, d'après la loi des grands nombres, $\mu(a_k)$. Ce qui nous intéresse est le comportement asymptotique de la probabilité que, sur les N premières lettres, cette proportion ait une valeur très différente, mettons $\nu(a_k)$. Autrement dit : quelle est la probabilité qu'un dé non pipé sorte des « six » un quart du temps ? (Ou : si un dé prétendu non pipé sort des « six » un quart du temps, que doit-on conclure ?)

On définit donc la mesure empirique des X_i par

$$L_N(a_k) = \frac{1}{N}\#\{i \leqslant N, X_i = a_k\}$$

L_N est une variable aléatoire dont la valeur est une mesure de probabilité sur Σ. (À noter que n'importe quelle mesure sur Σ ne peut pas être une mesure empirique : les fréquences doivent être des multiples de $1/N$.)

Ce qui nous intéresse est d'évaluer la probabilité que L_N soit proche d'une certaine mesure ν sur Σ. Pour cela, on va évaluer la quantité d'information H fournie par l'événement « la mesure empirique est ν », et la réponse sera alors : la probabilité est $\exp(-NH)$.

3.1.1 Entropie relative de mesures

On rappelle qu'en théorie de l'information, l'occurrence d'un événement x qui était de probabilité $\mu(x)$ apporte une information

$$I_\mu(x) = -\log \mu(x)$$

et que l'entropie d'une mesure de probabilité μ est l'espérance de la quantité d'information obtenue en tirant un élément selon μ :

$$H(\mu) = \mathbb{E}_\mu I_\mu = -\sum \mu_i \log \mu_i$$

Spécifier un élément d'un ensemble, sachant que cet élément allait être tiré selon la loi μ, apporte donc en moyenne une information $H(\mu)$.

Maintenant, on peut se demander quelle information (par rapport à μ) est apportée par l'affirmation suivante : « en fait, l'élément va être tiré selon une autre loi ν ». Cela apporte assurément une information : par exemple, si ν est concentrée en un point x, cela revient à donner directement x ce qui apporte une information $I_\mu(x)$. On définit l'information relative :

$$I_{\nu|\mu}(x) = I_\mu(x) - I_\nu(x)$$

et l'entropie relative

$$H(\nu|\mu) = \mathbb{E}_\nu I_{\nu|\mu} = \sum \nu(x) \log \frac{\nu(x)}{\mu(x)} = \sum \frac{\nu(x)}{\mu(x)} \log \frac{\nu(x)}{\mu(x)}\ \mu(x)$$

On montre que $H(\nu|\mu) \geqslant 0$, d'où le choix des signes (c'est essentiellement la convexité de $x \mapsto x \log x$).

L'interprétation est la suivante : si on tire un élément sous la loi ν, l'information moyenne qui sera au final obtenue sera $\mathbb{E}_\nu I_\mu$, par rapport à μ. Or effectuer un tirage selon une loi ν ne fait apparaître, dans l'absolu, qu'une information $H(\nu)$. C'est donc qu'en sachant que l'élément allait être tiré selon ν, on possédait dès le départ une information $\mathbb{E}_\nu I_\mu - H(\nu) = H(\nu|\mu)$, par rapport à la loi μ.

Moralement, cette quantité d'information peut servir à définir une distance sur l'espace des mesures de probabilité sur un ensemble (mais elle n'est pas symétrique).

3.1.2 Grandes déviations pour la mesure empirique

Revenons à la loi de la mesure empirique L_N de variables aléatoires X_i tirées dans Σ selon la loi μ. Un raisonnement intuitif, à ce point, permettrait d'obtenir le résultat. En effet, si la loi empirique est L_N, c'est comme si on avait tiré N fois de suite les X_i selon la loi L_N. Ceci apporte une information $NH(L_N|\mu)$. Un événement d'information H ayant probabilité $\exp(-H)$, on en conclut que la probabilité que la mesure empirique L_N soit égale à une certaine loi ν se comporte comme $\exp(-NH(\nu|\mu))$.

Cela se passe presque ainsi. Soit donc ν une loi sur Σ. Soit $x_1, x_2, \ldots, x_N$ une suite de lettres de Σ, telle que la proportion de x_i égaux à une lettre $a_k \in \Sigma$ soit $\nu(a_k)$. Calculons la probabilité (sous μ) que $X_i = x_i$. Cette probabilité est $\prod_i \mu(x_i) = \prod_k \mu(a_k)^{N\nu(a_k)}$, et ou encore $\exp(N \sum \nu(a_k) \log \mu(a_k))$, soit encore

$$\mathbb{P}_\mu(x_1, x_2, \ldots, x_N) = \exp\ -N(H(\nu) + H(\nu|\mu))$$

Pour évaluer la probabilité que la fréquence empirique soit ν, il reste donc à multiplier cette quantité par le nombre de suites $x_1, \ldots, x_N$ telles que la proportion des x_i égaux à la letrte a_k soit ν. Pour cela, on suppose bien sûr que ν est réalisable comme une telle fréquence, i.e. que les valeurs de ν sont multiples de $1/N$.

Ce nombre vaut $N!/\prod_k (N\nu(a_k))!$, qui, par un calcul très simple (essentiellement, celui de Boltzmann), vaut environ $\exp(NH(\nu))$ quand N est grand (à un facteur polynomial en N près), ce qui est bien naturel quand on sait que spécifier une suite particulière parmi l'ensemble des suites de fréquence empirique ν, fournit une information $NH(\nu)$.

Conclusion : si ν est réalisable comme fréquence d'une suite à N termes, alors la probabilité, sous μ, que la fréquence empirique L_N soit égale à ν est donc environ $\exp(NH(\nu)) \exp(-N(H(\nu) + H(\nu|\mu)))$, soit

$$\mathbb{P}_\mu(L_N = \nu) \approx \exp(-NH(\nu|\mu))$$

Pour se débarrasser des problèmes de lois réalisables ou non, on va plutôt calculer la probabilité que L_N tombe dans un petit ensemble autour de ν. On énonce alors le théorème de Sanov :

Théorème 3.1. *Soit A une partie de l'ensemble des mesures de probabilité sur Σ, et soit $\overset{\circ}{A}$ son intérieur. La probabilité que la mesure empirique L_N d'une suite de variables indépendantes tirées dans Σ avec la loi μ, appartienne à A, vérifie :*

$$-\inf_{\nu\in\overset{\circ}{A}} H(\nu|\mu) \leqslant \underline{\lim}_{N\to\infty} \frac{1}{N}\log \mathbb{P}_\mu(L_N\in A) \leqslant \overline{\lim}_{N\to\infty} \frac{1}{N}\log \mathbb{P}_\mu(L_N\in A) \leqslant -\inf_{\nu\in A} H(\nu|\mu)$$

Autrement dit, c'est la mesure la plus « proche » de μ au sens de la distance $H(\nu|\mu)$ qui contrôle le taux de décroissance de cette probabilité.

3.2 Le principe des grandes déviations

Le principe des grandes déviations est une généralisation de la situation précédente. En particulier, on ne demande plus forcément l'indépendance. On considère donc une suite de mesures de probabilité μ_N sur un espace X régulier (par exemple, métrisable avec sa tribu borélienne). On comprend que la mesure μ_N dépend de N « événements de base ». L'information ne croît pas forcément linéairement en N, on considère donc une suite de nombres a_N qui jauge cette croissance.

On considère une fonction $I : X \to [0;\infty]$ candidate à être la fonction entropie des μ_N. On suppose en général que I est semi-continue inférieurement (c'est-à-dire que les $I^{-1}([0;A])$ sont fermés), et on qualifie cette fonction de « bonne fonction de taux » si les $I^{-1}([0;A])$ sont compacts.

On dit alors que la famille μ_N satisfait le principe des grandes déviations pour la bonne fonction de taux I, si pour tout fermé $F\subset X$, on a

$$\overline{\lim}_{N\to\infty} \frac{1}{a_N}\log \mu_N(F) \leqslant -\inf_F I$$

et pour tout ouvert $\mathcal{O}\subset X$, on a

$$\underline{\lim}_{N\to\infty} \frac{1}{a_N}\log \mu_N(\mathcal{O}) \geqslant -\inf_{\mathcal{O}} I$$

Le principe de grandes déviations est donc analogue à l'existence d'une entropie.

3.3 Les gaussiennes

Si on a des variables satisfaisant un principe de grandes déviations, et que la fonction d'entropie I admet un minimum (qui vaut alors forcément 0, la probabilité totale étant 1) et est régulière, il est tentant de développer I à l'ordre 2 au voisinage de ce minimum, pour trouver que la renormalisation en $1/\sqrt{N}$ plutôt qu'en $1/N$, autour de la moyenne, donne une gaussienne... Bien sûr, on aurait aussi pu développer la probabilité à l'ordre 2 au voisinage de son maximum, on aurait trouvé que localement la probabilité se comportait comme une parabole osculant la gaussienne ci-dessus à l'ordre 2. Pour que la probabilité ressemble vraiment à une gaussienne, il faut donc vérifier que le développement de I est valable (par exemple, il suffit que la dérivée troisième soit contrôlée).

Alors, si I a un unique minimum au point z, on peut vérifier que $I(z + a/\sqrt{N}) \sim I''(z)\, a^2/2N$ et que la probabilité correspondante est donc $\approx e^{-NI} \approx e^{-I''(z)a^2/2}$, autrement dit qu'on a une gaussienne de variance $1/I''(z)$.

On avait vu en théorie de l'information que les gaussiennes maximisaient l'entropie à variance donnée, c'est exactement le phénomène qu'on retrouve ici : notre estimation de probabilités provient d'une maximisation d'entropie, et on renormalise à l'ordre deux au voisinage du maximum. Une fois de plus, les gaussiennes trouvent leur origine dans une quantité d'information...

3.4 Le théorème de Gärtner-Ellis

3.4.1 Première généralisation

À ce stade, on peut donner une première généralisation : plutôt que de s'intéresser à la mesure empirique des X_i, on peut considérer une fonction quelconque $f : \Sigma \to \mathbb{R}^d$, et s'intéresser à sa moyenne empirique $\widehat{f} = \frac{1}{N} \sum f(X_i)$. Si on prend $d = |\Sigma|$ et qu'on prend $f(a_k)$ égal au k-ième vecteur d'une base de $\mathbb{R}^d$, on retrouve bien évidemment le cas précédent.

Si $\widehat{f} = y$, cela signifie que la fréquence empirique L_N des X_i vérifie $\mathbb{E}_{L_N} f = y$, par définition. On est donc tenté de dire que la probabilité que $\widehat{f} = y$ est la somme, pour toutes les mesures ν sur Σ satisfaisant $\mathbb{E}_\nu f = y$, de la probabilité que $L_N = \nu$. Cette probabilité, comme ci-dessus, est asymptotiquement $\exp(-NH(\nu|\mu))$.

Quand N est grand, seule la contribution du meilleur ν (c'est-à-dire celui minimisant la « distance » $H(\nu|\mu)$) compte, les autres devenant négligeables. Posons donc, pour $y \in \mathbb{R}^d$:

$$I(y) = \inf\{H(\nu|\mu), \nu \text{ mesure de probabilité sur } \Sigma \text{ telle que } \mathbb{E}_\nu f = y\}$$

c'est la quantité d'information contenue dans l'événement $\widehat{f} = y$. On peut alors énoncer le théorème suivant :

Théorème 3.2. *Soit $A \subset \mathbb{R}^d$, d'intérieur $\overset{\circ}{A}$. La probabilité que la moyenne empirique $\widehat{f}$ tombe dans A vérifie*

$$-\inf_{\overset{\circ}{A}} I \leqslant \underline{\lim}_{N\to\infty} \frac{1}{N} \log \mathbb{P}_\mu(\widehat{f} \in A) \leqslant \overline{\lim}_{N\to\infty} \frac{1}{N} \log \mathbb{P}_\mu(\widehat{f} \in A) \leqslant -\inf_{A} I$$

3.4.2 Cas général

On va désormais illustrer ce principe dans un cas un peu plus général que le théorème de Sanov. Soient X_i des variables aléatoires à valeurs dans $\mathbb{R}^d$, éventuellement non indépendantes, ni identiquement distribuées. On considère la moyenne empirique $S_N = \frac{1}{N}\sum_{i=1}^N X_i$. Soit μ_N la loi de S_N. On va montrer que sous certaines hypothèses, μ_N satisfait un principe de grandes déviations, pour une fonction de taux à déterminer.

Comme précédemment, on a envie de dire que si $S_N = y$, cela signifie qu'en fait, les X_i ont collectivement une distribution empirique ν qui est de moyenne y, i.e. $\int_{t\in\mathbb{R}^d} t\, d\nu(t) = y$.

On voudrait alors dire que la probabilité d'une telle situation est $\exp -H(\nu|\mu_N)$, ou plutôt $\exp - \inf H(\nu|\mu_N)$, l'inf étant pris sur toutes les mesures ν satisfaisant la contrainte d'être de moyenne y : asymptotiquement, les contributions des mesures ne réalisant pas l'inf sont négligeables.

Comme les X_i ne sont pas indépendantes, on va plutôt travailler avec la loi jointe ρ_N du N-uplet (X_i) dans $(\mathbb{R}^d)^N$. On cherche maintenant des lois ν sur $(\mathbb{R}^d)^N$ soumises à la contrainte que $\sum_{i=1}^N \int_{t\in\mathbb{R}^d} t\, d\nu_i = Ny$, où ν_i est la loi de la i-ième composante de ν : la somme des moyennes sur chaque composante doit être égale à Ny. Parmi celles-ci on cherche celle qui a l'entropie minimale par rapport à la mesure ρ_N.

Ici intervient la remarque fondamentale suivante : *à moyenne fixée, les distributions qui minimisent l'entropie sont les distributions exponentielles* (ou maxwelliennes) de la forme $d\nu(x) = e^{\lambda.x}/Z$, où $Z = \int e^{\lambda.x}$ est la constante de normalisation, appelée fonction de partition par les physiciens. Ceci se démontre par un calcul variationnel simple, identique à celui qui montre qu'à variance fixées, ce sont les gaussiennes.

Soit $E : (\mathbb{R}^d)^N \to \mathbb{R}^d$ l'application « somme des composantes ».

Pour minimiser l'entropie par rapport à la mesure ρ_N, il est donc suffisant de chercher parmi les mesures de la forme $d\nu(x) = \frac{1}{Z} e^{\lambda.Ex}\, d\rho_N(x)$ où λ est un élément de $\mathbb{R}^d$, le produit $\lambda.Ex$ étant un produit scalaire. Cet élément λ est à déterminer de sorte que la moyenne $\mathbb{E}E = E\int_{x\in(\mathbb{R}^d)^n} x\, d\nu$ soit égale à Ny.

Ce qui nous intéresse est l'entropie de la distribution. Or pour les distributions exponentielles, il y a une relation simple entre entropie et moyenne. La moyenne de la distribution $e^{\lambda.x}/Z$ est $\mathbb{E}E = \frac{1}{Z}\int xe^{\lambda.x}$ et son entropie est $\frac{1}{Z}\int e^{\lambda.x}\log(e^{\lambda.x}/Z) = \mathbb{E}E - \log Z$.

On voit donc qu'une distribution exponentielle de moyenne $\mathbb{E}_\nu E = Ny$ a une entropie $\lambda Ny - \log Z$.

Reste quand même à déterminer λ. Là encore la forme exponentielle de la loi de probabilité joue : la dérivée de $\log Z$ par rapport à λ est précisément l'espérance de la distribution exponentielle. En effet, on a

$$\frac{d}{d\lambda}\log Z = \frac{1}{Z}\frac{d}{d\lambda}\int e^{\lambda.x} = \frac{1}{Z}\int xe^{\lambda.x}$$

Le λ recherché vérifie donc $\frac{d}{d\lambda}(\lambda Ny - \log Z) = Ny - \mathbb{E}E = 0$, autrement dit le λ recherché est un extrémum de $\lambda Ny - \log Z$. C'est en fait un maximum car $\log Z$ est une fonction convexe de λ.

Dans le principe de grandes déviations $\mathbb{P}\left(\frac{1}{N}\sum X_i = y\right) \approx e^{-NI(y)}$, on doit donc poser :

$$I(y) = \sup_{\lambda\in\mathbb{R}^d} \lambda.y - \frac{1}{N}\log Z(\lambda)$$

où

$$Z(\lambda) = \int_{(t_1,\ldots,t_N)\in(\mathbb{R}^d)^N} \exp\left(\lambda\sum t_i\right)\, d\rho_N(t_1,\ldots,t_N)$$

On a donc réussi, grâce à la remarque que les minima d'entropie sont obtenus pour les distributions exponentielles, à donner une recette de calcul de l'entropie de l'événement « la moyenne est égale à y ».

Ceci nous amène donc à énoncer le théorème de Gärtner-Ellis. Cependant, il faut faire attention à l'énoncé : par exemple, nos raisonnements ci-dessus étaient à N fixé ; il faut donc que $I(y)$ converge quand $N \to \infty$, vers une certaine valeur, ce qui ne se produit que si les X_i n'ont pas des distributions trop sauvages.

De plus, lorsque la limite $\frac{1}{N}\log Z(\lambda)$ n'est pas différentiable, il n'y a pas forcément de λ donnant une exponentielle de moyenne y pour tout y, ce qui n'empêche pas que $\sup_{\lambda\in\mathbb{R}^d} \lambda.y - \frac{1}{N}\log Z(\lambda)$ ait une certaine valeur. Un même λ peut ainsi maximiser $\lambda.y - \frac{1}{N}\log Z(\lambda)$ pour plusieurs y. Disons que $y \in \mathbb{R}^d$ est un point exposé si le λ maximisant cette quantité ne maximise pas aussi cette quantité pour un autre y', cela revient à dire que y est exposé s'il existe un λ tel que pour tout $y' \neq y$, on a $\lambda.y - \sup_{\lambda'}(\lambda'.y - \frac{1}{N}\log Z(\lambda')) > \lambda.y' - \sup_{\lambda'}(\lambda'.y' - \frac{1}{N}\log Z(\lambda'))$. Les points exposés sont ceux pour lesquels le fait que λ maximise l'entropie implique bien que l'espérance de la distribution exponentielle de paramètre λ vaut y.

L'énoncé est alors le suivant. Il se place dans un cadre un peu plus général où on ne considère pas forcément une somme de variables aléatoires ; de plus, il se peut que la bonne renormalisation ne soit pas N mais a_N où a_N est une suite tendant vers l'infini.

Théorème 3.3. *Soit (μ_N) une suite de lois de probabilité sur $\mathbb{R}^d$ et soit a_N une suite tendant vers l'infini. Pour $\lambda \in \mathbb{R}^d$, on pose*

$$Z_N(\lambda) = \mathbb{E}_{\mu_N} e^{a_N \lambda.t}$$

et on suppose que la limite

$$\Lambda(\lambda) = \lim_N \frac{1}{a_N} \log Z_N(\lambda)$$

existe et est finie pour λ dans un voisinage de 0. Pour $y \in \mathbb{R}^d$, soit

$$I(y) = \sup_{\lambda \in \mathbb{R}^d} \lambda.y - \Lambda(\lambda)$$

et soit $\mathcal{P}$ l'ensemble des points y exposés. Alors, si A est une partie de $\mathbb{R}^n$, d'adhérence $\overline{A}$ et d'intérieur $\overset{\circ}{A}$, on a

$$\overline{\lim_N} \frac{1}{a_N} \log \mu_N(A) \leqslant - \inf_{\overline{A}} I$$

et

$$\underline{\lim_N} \frac{1}{a_N} \log \mu_N(A) \geqslant - \inf_{\overset{\circ}{A} \cap \mathcal{P}} I$$

Reconnaissons que sans explication par la théorie de l'information, l'énoncé pourrait rester mystérieux.

Là encore, le sujet n'est pas clos : on peut chercher à montrer qu'un principe des grandes déviations est satisfait dans des contextes plus généraux (par exemple des chaînes de Markov), vouloir obtenir des bornes explicites plutôt que des relations asymptotiques, montrer que les grandes déviations sont contrôlées uniformément pour un grand nombre de fonctions-tests de la variable étudiée, ou encore étudier les innombrables et délicates applications à la physique statistique...

Chapitre 4

La constance de l'entropie

Ce texte entend expliquer la croyance populaire selon laquelle l'entropie augmenterait, qui semble incompatible avec la réversibilité des lois de la physique classique.

On analysera les situations classiques suivantes : une goutte d'encre versée dans un verre d'eau, un gaz initialement confiné dans une moitié de boîte et se répandant dans toute la boîte, ou encore une assiette qui tombe.

On essaiera d'expliquer comment le sens commun est capable de rejeter un film à l'envers de ces situations comme incompatible avec les lois de la physique, voire impossible.

4.1 Entropie microscopique

On considère un système physique $\mathcal{S}$ pouvant présenter un certain nombre d'états. On a deux options de modélisation : la modélisation discrète, où $\mathcal{S}$ est un ensemble fini d'états $s_1, \ldots, s_n$; la modélisation continue, où $\mathcal{S}$ est une partie (compacte) de l'espace $\mathbb{R}^n$.

On suppose qu'on a un opérateur d'évolution temporelle φ_t sur le système. Dans le premier cas, pour t un nombre entier (positif ou négatif), φ_t est une fonction de $\mathcal{S}$ dans $\mathcal{S}$. La réversibilité des lois de la physique signifie que φ_t est une bijection, ce que nous supposerons. Dans le cas continu, on suppose qu'on a une mesure de volume sur l'espace des états (en mécanique classique elle est issue de la forme symplectique $dp \wedge dq$), et que φ_t préserve le volume : c'est l'analogue continu de la condition d'irréversibilité.

Notons qu'on a les relations $\varphi_t \varphi_{t'} = \varphi_{t+t'}$. Dans le cas discret, φ est entièrement déterminé à partir de φ_1 ; si le temps est choisi continu, à partir de son générateur infinitésimal pour $t \to 0$.

L'*entropie microscopique d'un état* d'un système discret est définie comme

le logarithme du nombre d'états de ce système accessibles à partir de cet état par l'opérateur φ_t pour tout t. L'*entropie microscopique d'un état* d'un système continu est définie comme le logarithme du volume de (la fermeture topologique de) l'ensemble des états accessibles par l'opérateur φ_t.

Comme l'ensemble des états accessibles par φ_t ne change pas si on applique φ_t, les entropies microscopiques sont trivialement indépendantes du temps : *l'entropie microscopique est constante.*

Par contre, si on change le système, l'entropie peut varier : si on a un gaz dans une moitié de boîte, avec une paroi au milieu de la boîte, et qu'on retire subitement la paroi, le nombre d'états accessibles au système augmente d'un seul coup : *lors du relâchement d'une contrainte, l'entropie microscopique augmente.* À noter que cette augmentation se produit au moment où on enlève la paroi, même si le gaz ne s'est pas encore répandu ; c'est l'opérateur d'évolution φ qui a changé.

L'entropie microscopique, qui tient compte de l'ensemble des états accessibles au système, n'est pas très satisfaisante (en particulier son changement subit lorsqu'on enlève la paroi, alors même que le gaz ne s'est pas encore répandu). Cependant, cette définition est largement suffisante pour une grande part de la physique statistique, fondée sur l'étude de la manière dont l'entropie d'un système dépend de son énergie.

4.2 Entropie macroscopique

L'entropie macroscopique, historiquement définie avant la précédente, ne possède pas cet inconvénient : elle ne dépend que de ce qu'on voit du système.

On suppose que les états de notre système sont répartis en un certain nombre de classes appelés *états macroscopiques* ou *macro-états*, deux états microscopiques appartenant au même macro-état s'ils ne peuvent pas être distingués par l'expérimentateur (éventuellement en utilisant certains appareils de mesure).

Par exemple, pour notre boîte de gaz, on peut définir les macro-états caractérisés par le nombre de particules de gaz dans les moitiés droite et gauche de la boîte. On peut aussi faire des distinctions plus fines si cela nous chaut.

Alors, l'*entropie macroscopique* d'un état microscopique est le logarithme du nombre d'états microscopiques correspondant au même macro-état (ou logarithme du volume, dans le cas continu). C'est ainsi une fonction de chaque macro-état.

L'entropie macroscopique globale du système peut être définie comme la moyenne des entropies macroscopiques des états microscopiques, autrement dit $\sum n_i \log n_i$ où n_i est le nombre d'états microscopiques

dans le macro-état i. On peut aussi normaliser en $\sum n_i/n \log n_i/n$ qui est la formule classique. Ceci est une fonction globale qui ne dépend que du système, mais ni du temps ni d'un état observé ; sa constance ou sa croissance n'a donc pas de sens.

Si un des macro-états correspond à un nombre d'états microscopiques beaucoup plus grand que les autres, alors son entropie macroscopique est très proche de l'entropie microscopique du système. Ainsi ces deux définitions recouvrent souvent des notions comparables. Dans notre exemple de la boîte, de tels états sont ceux où les nombres de particules de gaz à droite et à gauche sont presque égaux. Une proportion presque totale des états microscopiques sont tels, et donc, le logarithme du macro-état correspondant est presque égal au logarithme du nombre d'états total du système.

Notons qu'au niveau macroscopique, la dynamique n'est plus définie que de manière probabiliste : si on a un macro-état, il manque de l'information, et il n'est pas sûr que tous les états microscopiques futurs issus du macro-état présent appartiennent au même macro-état futur.

Théorème. Si on choisit au hasard un état microscopique du système (uniformément dans le cas discret, selon notre mesure de volume dans le cas continu), et qu'on laisse évoluer le système une unité de temps, alors la variation moyenne de l'entropie macroscopique sera nulle. Autrement dit : *l'entropie macroscopique est constante en moyenne.*

Cependant :

Théorème. Si on choisit au hasard un macro-état, avec égale probabilité pour tous les macro-états, et qu'on laisse évoluer le système une unité de temps, alors la variation moyenne de l'entropie macroscopique sera positive. Autrement dit : *l'entropie macroscopique ne peut qu'augmenter.*

(Les démonstrations de ces théorèmes, peu difficiles au moins dans le cas discret, sont omises. Toute fonction des états microscopiques vérifierait le premier qui n'est qu'une conséquence de la bijectivité. Toute fonction croissante du nombre d'états microscopiques dans un macro-état vérifierait le second.)

Illustrons ceci dans le cas de la goutte d'encre dans l'eau. Soit 1 le nombre d'état où la goutte n'est pas diluée dans l'eau, et soit $N \ggg 1$ le nombre d'états où la goutte apparaît diluée. Alors, si on pose côte à côte un verre avec goutte non diluée et un verre avec goutte diluée, la goutte non diluée va se diluer, celle qui était diluée va tranquillement le rester, et l'entropie aura augmenté.

Par contre, si on place sur une très grande table N verres avec une goutte bien mélangée et 1 verre avec une goutte non diluée, en moyenne un des verres contiendra précisément la configuration microscopique des vitesses des particules d'encre requise pour qu'en une unité de temps, ces particules se regroupent pour former une belle goutte d'encre (une telle configuration existe forcément : il suffit d'inverser l'opérateur d'évolution à partir de l'état où la goutte est formée). Par conséquent, dans le verre où elle était présente la

goutte d'encre disparaît, mais elle apparaît en moyenne dans un autre verre. L'entropie est constante en moyenne.

De même, à long terme, si on lâche une goutte d'encre dans un verre, elle commence par se diluer ; puis, au bout d'un nombre immense d'années, elle se reforme pour se diluer à nouveau immédiatement, et le cycle recommence pour les siècles des siècles. *À long terme, l'entropie reste constante.*

Le fait est qu'en pratique, on part plus souvent d'états à entropie plus faible que la moyenne ; par exemple on lâche l'encre au milieu du verre, ou encore, on place le gaz dans une moitié de boîte au moyen d'une paroi. Si on part souvent d'états à faible entropie, on a de bonnes chances que l'entropie augmente...

4.3 La flèche du temps et la réversibilité

En voyant un film de scènes quotidiennes, on est parfaitement capable de dire si le film est passé à l'envers : on voit très couramment tomber des assiettes, mais moins souvent des morceaux d'assiette se soulever du sol pour bondir sur la table en formant une assiette complète. En revanche, un film de molécules de gaz s'entrechoquant, s'il pouvait être tourné, ne montrerait aucune dissymétrie temporelle. L'irréversibilité observée macroscopiquement est souvent attribuée à la croissance de l'entropie.

Comment, en pratique, reconnaissons-nous un film passé à l'envers ? Le raisonnement ne peut être purement déductif, puisque le film passé à l'envers est entièrement compatible avec toutes les lois de la physique, qui sont symétriques par renversement du temps (en physique classique ; les violations de symétrie par les particules élémentaires ne nous importent pas). Simplement, ce film apparaît comme extrêmement improbable. On mène en fait un raisonnement inductif, qui s'appuie sur le principe suivant, qui est un cas particulier d'une attitude beaucoup plus générale.

Principe méthodologique. On peut supposer, devant un certain macro-état et sans autre information, qu'on a affaire avec probabilité égale à un quelconque des états microscopiques correspondants (dans le cas continu, que l'état microscopique est réparti aléatoirement selon notre mesure de volume).

Reprenons nos N mélanges d'eau et d'encre de tout à l'heure. On nous montre un film où, dans un verre contenant de l'encre et de l'eau, toute l'encre se regroupe subitement et miraculeusement pour former une seule goutte, au centre du verre. On suppose que le mélange initial d'encre et d'eau était, au hasard, l'un des N tels mélanges possibles. On sait qu'un exactement de ces mélanges donne une goutte bien formée au bout d'une unité de temps, tandis que tous les autres restent aussi insondables. Deux hypothèses : le film

est à l'endroit, ou bien il est à l'envers. Dans la première, le déroulement du film nous dit que nous nous trouvions juste dans la configuration donnant une goutte, ce qui était a priori le cas avec probabilité $1/N$, le nombre N étant immense. Dans le cas où le film est à l'envers, la goutte se dilue et cela n'a rien d'anormal. Un raisonnement inductif amène donc à penser que l'hypothèse que le film n'est pas truqué est à rejeter avec probabilité $1 - 1/N$, autant dire 1 vu la taille de N.

La même interprétation vaut pour les assiettes sauteuses.

4.4 La thermodynamique classique

La thermodynamique classique, avec son second principe affirmant que l'entropie augmente, est pourtant tout à fait adaptée à la description d'un certain nombre de situations, tant en physique que dans la vie courante ; elle explique en particulier pourquoi, dans la réalité, les assiettes se brisent plus facilement qu'elles ne se forment spontanément (bien que si l'on considère l'univers dans son ensemble, plus d'assiettes aient été créées par l'action des lois de la physique que brisées...).

Une des formulations du second principe, une des plus anciennes historiquement est qu'*on ne peut pas transformer la chaleur en travail dans un système où la température est uniforme.* Chaleur et travail sont deux formes d'énergie ; la distinction entre les deux est que le travail est de l'énergie exploitable macroscopiquement, alors que la chaleur est de l'énergie non structurée d'agitation des molécules, qu'on ne peut pas exploiter (noter que la limite entre les deux dépend de la technologie disponible, en particulier pour l'exploitation de l'énergie chimique ou électrique).

Plus exactement, on peut transformer de la chaleur en travail, mais uniquement celle qu'on aurait préalablement produite en utilisant du travail (par exemple en brûlant un combustible), et en passant par un accroissement temporaire de la température d'une partie du système. On ne peut pas se contenter de refroidir un objet en pompant sa chaleur pour travailler.

Dans le cas des assiettes sauteuses, c'est l'énergie thermique du sol et de l'air, celle qui fait que l'assiette fait du bruit en tombant, qui serait récupérée pour faire s'envoler l'assiette. La chaleur ne peut pas être changée en travail.

Et pourtant :

Recette pour transformer de la chaleur en travail à température ambiante. Il faut : un gaz à température ambiante ; une boîte ; un piston relié à un dispositif capable d'utiliser son énergie de déplacement pour accomplir une tâche souhaitée. Temps de cuisson : inconnu. Mettre le gaz dans la boîte. Attendre. Au bout d'un nombre d'années à peu près égal au N de tout à

l'heure, au moment où toutes les molécules de gaz se placent temporairement dans la moitié gauche de la boîte, placer immédiatement le piston au milieu de la boîte. La différence de pression pousse le piston, et l'énergie de ce déplacement est prête à consommer. Le gaz est refroidi d'autant.

Bien évidemment, cette recette n'est pas applicable en pratique étant donné la taille de N. On peut donc formuler ainsi le second principe de la thermodynamique : *en pratique*, on ne peut pas convertir de la chaleur en travail dans un système où la température est uniforme.

Ce principe est alors simplement une conséquence de la définition des états microscopiques correspondant à un même macro-état : par hypothèse, ce sont ceux qu'on ne peut pas distinguer, et a fortiori pas manipuler pour en extraire l'énergie.

On a vu que l'irréversibilité apparente à l'échelle macroscopique est le reflet des conditions particulières dans lesquelles on se place (on part plus souvent d'état de faible entropie) et du fait que l'homme ne peut pas spécifier toutes les caractéristiques microscopiques d'un état ; nul n'est besoin d'invoquer une dissymétrie fondamentale du temps, une loi macroscopique spécifique, ou encore l'indétermination de la mécanique quantique (dont les lois d'évolution sont d'ailleurs tout aussi réversibles que les lois classiques).

Nous n'avons nulle part utilisé la définition classique de l'entropie qui dit que sa variation est égale à la variation de chaleur divisée par la température. Notre cadre de départ était beaucoup trop général pour qu'une telle relation ait même un sens. Cependant, dans un certain nombre de systèmes physiques, les deux définitions coïncident ; nous ne donnerons pas la preuve de ce résultat découvert par Boltzmann, et qu'il fit graver sur sa tombe.

Références

Tous les résultats décrits ici sont standard et sont traités dans les ouvrages de référence sur ces sujets.

La théorie de l'information a été pour la première fois introduite par Shannon dans *The mathematical theory of communication*, Bell System Tech. J. 27 (1948). L'ouvrage de référence est *Elements of information theory*, par Cover et Thomas (Wiley, 1991). Pour les rapports avec l'analyse, on peut consulter le chapitre 10 de *Sur les inégalités de Sobolev*, par Ané, Blachère, Chafaï et al. (Soc. Math. Fr. 2000).

Pour l'entropie en systèmes dynamiques, on pourra consulter l'ouvrage de référence de Katok et Hasselblatt, *Introduction to the modern theory of dynamical systems*, Cambridge University Press (1995).

Pour tout savoir des grandes déviations, l'ouvrage de référence est celui de Dembo et Zeitouni, *Large deviations techniques and applications*, Springer (1998).