

## Entropie de Shannon, théorie de l'information, et compression de données

En 1948, tandis qu'il travaillait aux Laboratoires Bell, l'ingénieur en génie électrique Claude Shannon formalisa mathématiquement la nature statistique de "l'information manquante" dans les signaux des lignes téléphoniques. Pour ce faire, il développa le concept général d'entropie et d'information. Dans ce tutorat, nous allons étudier cette entropie (qui se trouve être la même, à une constante multiplicative près, que celle de Boltzmann) et ses liens avec la compression de données.

### 1. Entropie et probabilité: particules dans des boîtes

Soit  $N$  particules et  $k$  boîtes, considérons l'expérience suivante: on prend la première particule et on la jette dans l'une des boîtes, de façon équiprobable. On fait ensuite la même chose avec la seconde, puis la troisième, etc... A la fin, les particules ont toutes été jetées dans les  $N$  boîtes de façon équiprobable (chacune des boîtes est associée à une énergie  $E_1, E_2, \dots, E_k$ ).

- Donnez le nombre  $\mathcal{N}_N(N_1, N_2, \dots, N_k)$  d'expériences qui conduisent à une configuration dans laquelle la  $j$ -ème boîte contient  $N_j$  particules. Donnez le nombre total d'expériences. En déduire la probabilité de chaque configuration.
- Considérez la limite "thermodynamique"  $N \rightarrow \infty$ . On note le nombre d'occupation du  $i$ -ème état par  $n_i = N_i/N$ . Montrez que l'on peut écrire  $\mathcal{N}_N(Nn_1, \dots, Nn_k) \simeq e^{N\Sigma(n_1, \dots, n_k)}$ , et donnez la fonction  $\Sigma(n_1, \dots, n_k)$ . (Utilisez la formule de Stirling).
- Question subsidiaire: En utilisant le formalisme des multiplicateurs de Lagrange, calculez les nombres d'occupation  $n_i$  les plus probables sachant que (i) le nombre de particules  $\sum_{i=1}^k N_i = N$  est fixé (et très grand) et (ii) l'énergie  $\sum_{i=1}^k E_i n_i = \epsilon$  est aussi fixée. On utilisera la fonction  $\Sigma$ .

### 2. Surprise, incertitude et information

Passons maintenant à la théorie de l'information. Considérez une variable aléatoire  $X$  qui peut prendre une valeur parmi  $N$  possibles dans un alphabet  $\chi$  qui contient  $L$  lettres. La probabilité de prendre une valeur  $X = x_i \in \chi$  est  $p_i$ . Pour cet exercice, considérons par exemple un alphabet de quatre lettres  $\chi = \{A, B, C, D\}$  avec les probabilités  $p_A = 1/2, p_B = 1/4, p_C = 1/8, p_D = 1/8$ .

- Nous tirons une valeur  $X = x_i$ . Plus cette valeur est improbable, plus nous sommes surpris de la voir apparaitre. Définissons la "surprise" comme  $\log_2 1/p_i$ . Le fait que nous soyons surpris est clairement lié au fait que nous ne sommes pas certain du résultat du tirage avant qu'il n'ait lieu. Appelons, avec Shannon, la surprise moyenne "l'incertitude"  $H[X]$ . Comment s'écrit l'incertitude en général? Que vaut-elle dans notre exemple?
- Montrez que l'incertitude  $H[X]$  satisfait  $0 \leq H[X] \leq \log_2 L$ ; que la valeur 0 est atteinte si et seulement si une seule lettre a une probabilité non nulle; et que la valeur  $\log_2 L$  est atteinte si et seulement si toutes les lettres sont équiprobables.
- Une autre approche est la suivante: je mesure une valeur  $X = x_i$ , mais je la garde secrète en vous laissant le soin de la deviner. Vous ne connaissez que les probabilités  $p_i$ , ce qui peut vous donner quand même une idée sur la valeur en question. Plus  $x_i$  est probable, plus petite est l'information que je vous donnerais en vous disant ce qu'est vraiment la valeur mesurée  $x_i$ . Si l'on définit l'information qu'il vous manque comme  $\log_2 1/p_i$ , quelle est la valeur moyenne de cette information manquante en général? Que vaut-elle dans notre exemple?

### 3. Entropie de Shannon et questions OUI/NON

Nous voulons maintenant répondre à la question: à quel point est-il difficile de deviner la valeur d'une variable aléatoire?

- a) J'ai tiré au hasard l'une des lettres A,B,C,D avec les probabilités précédentes. Vous avez le droit de poser des questions, du type "Cette lettre est-elle dans l'ensemble  $\{x_1, x_2, \dots\}$ ?", auxquelles je répondrai par oui ou par non. Votre but est de trouver le plus vite possible la lettre tirée. Combien de questions devez vous poser en moyenne avant de découvrir la lettre que j'ai tirée? Comparez ce nombre avec l'entropie de l'alphabet.
- b) Considérons un nouvel alphabet de  $2^b$  lettres équiprobables (avec  $b$  entier). Combien des questions est-il nécessaire de poser en moyenne? Comparez avec l'entropie du système.
- c) (Théorème de Shannon (1948)) Considérons enfin un alphabet général avec  $A$  lettres  $x_i$ ,  $i = 1 \dots A$  avec probabilités  $p_i$ . Comment généraliser la stratégie précédente? Combien des questions devez vous poser en moyenne pour découvrir la lettre tirée? Pour simplifier, on fera l'hypothèse qu'il soit toujours possible de partager les lettres en deux groupes tels que la probabilité de chacun des deux groupes est  $1/2$ .

### 4. Compression de données

Vous avez un fichier avec  $N$  valeurs aléatoires  $x_i \in \chi$ , l'alphabet  $\chi$  contenant  $L$  lettres. Ce fichier prend une place de  $N \log_2 L$  bits dans votre disque dur.

- a) Adoptons la stratégie suivante, en utilisant les questions OUI/NON. On écrit un fichier où, au lieu de garder chaque lettre, on garde, pour chaque lettre, les réponses aux questions de l'exemple précédent avec OUI= 1 et NON=0. Quelle est la taille de ce nouveau fichier? Quel est le facteur de compression (rapport entre la taille du fichier avant et après compression)?
- b) Comparez ce résultat avec les performances de votre programme de compression favori (zip, rar ...). Pour cela, utiliser les fichiers ci-joints, qui sont des suites de  $10^4$  caractères Ascii <sup>1</sup> tirés au hasard avec les alphabet suivant:
  - (i) *uniform.txt*: caractères tirés au hasard parmi L'ENSEMBLE des 256 caractères possibles.
  - (ii) *half.txt*: caractères tirés au hasard parmi LA MOITIE (soit 128) caractères possibles.
  - (iii) *abcd.txt*: caractères tirés au hasard parmi ABCD.
  - (iv) *abcd2.txt*: caractères tirés au hasard parmi ABCD avec les probabilités de l'exercice 2.

### 5. L'entropie de l'anglais

- a) Les fréquences des lettres dans l'anglais sont disponibles sur:  
[http://en.wikipedia.org/wiki/Letter\\_frequencies](http://en.wikipedia.org/wiki/Letter_frequencies). Calculez le rapport de compression de l'anglais (négligez les espaces, majuscules et autres caractères spéciaux.)
- b) Comparez vos résultats aux meilleurs compresseurs disponibles:  
[http://en.wikipedia.org/wiki/Hutter\\_Prize](http://en.wikipedia.org/wiki/Hutter_Prize). D'où vient la différence observée?
- c) Estimez (et commentez vos résultats) l'entropie de l'anglais grâce à l'expérience numérique:  
<http://www.math.psu.edu/dlitttle/java/informationtheory/entropy/index.html>

---

<sup>1</sup>On rappelle que les caractères Ascii sont codés sur 8 bits et prennent 256 valeurs possibles, voir <http://fr.wikipedia.org/wiki/ASCII>