

THERMODYNAMIQUE – TD n° 2 :

Codage et théorie de l'information

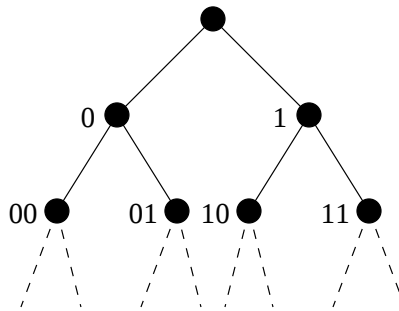
Benjamin Langlois, Swann Piatecki et Guilhem Semerjian

Mars 2012

1 Théorème de Shannon pour la compression de données

L'objectif de l'exercice est de démontrer la relation entre l'entropie statistique d'une loi de probabilité et le taux de compression que l'on peut atteindre pour une séquence de symboles tirés avec cette loi.

1. On considère un alphabet de m symboles $\{S_1, \dots, S_m\}$. On définit un code pour cet alphabet en se donnant pour chaque symbole S_i un mot binaire w_i , de ℓ_i bits. Le codage d'une suite de symboles $S_{i_1} \dots S_{i_n}$ est donné par la concaténation des mots correspondants $w_{i_1} \dots w_{i_n}$. Prenons par exemple $m = 3$, avec $w_1 = 0$, $w_2 = 10$ et $w_3 = 11$. Décoder la séquence binaire 10011010.
2. Il est commode de représenter les éléments w_i en les associant aux sommets d'un arbre binaire :



Dans l'exemple de la question précédente vous avez pu décoder la séquence, sans ambiguïté, en lisant les bits un par un. Un code possédant cette propriété est dit instantané. Comment caractériser cette propriété en terme de la représentation en arbre ?

3. En déduire que si $\{w_1, \dots, w_m\}$ est un code instantané, les longueurs $\ell_1, \dots, \ell_m$ de ses mots vérifient l'inégalité dite de Kraft,

$$\sum_{i=1}^m 2^{-\ell_i} \leq 1. \quad (1)$$

On pourra introduire dans le raisonnement la longueur $\ell_{\max} = \max_i \ell_i$.

4. La distance de Kullback-Leibler entre deux lois de probabilités $p = (p_1, \dots, p_m)$ et $q = (q_1, \dots, q_m)$ est définie par $D(p||q) = \sum_{i=1}^m p_i \log_2 \left(\frac{p_i}{q_i} \right)$. Montrer que cette quantité est positive quelles que soient p et q .
5. On suppose que les symboles $S_1, \dots, S_m$ apparaissent avec des probabilités connues $p_1, \dots, p_m$. Pour un code donné on définit la longueur moyenne de ses mots par $L = \sum_{i=1}^m p_i \ell_i$. Montrer que pour un code instantané on a nécessairement $L \geq S$, où S est l'entropie (en bits) de la loi p . On pourra pour cela calculer la distance de Kullback-Leibler entre p et la loi q définie par $q_i = 2^{-\ell_i}/z$, où z est un facteur de normalisation.
6. Si $\{\ell_1, \dots, \ell_m\}$ est un jeu de longueurs entières positives vérifiant l'inégalité de Kraft (1), montrer que l'on peut construire un code instantané $\{w_1, \dots, w_m\}$ de mots de ces longueurs. On pourra supposer $\ell_1 \leq \ell_2 \leq \dots \leq \ell_m = \ell_{\max}$ et traiter les mots dans cet ordre.

7. Montrer que le jeu de longueurs $\ell_i = \lceil \log_2 \frac{1}{p_i} \rceil$ vérifie l'inégalité de Kraft, où $\lceil x \rceil$ dénote le plus petit entier supérieur ou égal à x .
8. En déduire le théorème de Shannon : si L_* est le minimum parmi les codes instantanés de la longueur moyenne des mots, on a

$$S \leq L_* < S + 1 . \quad (2)$$

L'entropie de Shannon donne donc, à un bit près, le nombre moyen de bits nécessaires pour décrire un tirage de la loi de probabilité p .

9. Quelle propriété doivent vérifier les p_i pour qu'il existe un code instantané de longueur moyenne égale à l'entropie ? Donner des exemples de telles distributions p , en particulier dans le cas $m = 3$.

2 Codage d'évènements rares

On veut encoder une série d'informations météorologiques simplifiées : pour chaque jour, on représente par 1 l'évènement « il a plu », par 0 sinon. On suppose ces évènements indépendants d'un jour sur l'autre, avec une faible probabilité de pluie, $\frac{1}{30}$.

1. Quelle est, en bits, l'entropie de cette distribution de probabilités ?
2. Combien de bits par jour faut-il utiliser pour encoder cette information si l'on traite chaque jour l'un après l'autre ?
3. On peut améliorer ce résultat en utilisant des blocs de durée plus longue, par exemple de 30 jours. Proposer un encodage binaire de la série météorologique sur 30 jours, en se concentrant sur les (rares) jours de pluie, qui contienne toute l'information et qui soit uniquement décodable (il faut donc inclure un séparateur qui signale la fin de la série). Combien de bits par jour utilise-t-il en moyenne ?
4. Peut-on améliorer ce résultat en utilisant une durée de bloc différente ?

Pour en savoir plus sur le codage et la théorie de l'information, *Elements of Information Theory* de T. Cover et J. Thomas est un ouvrage de référence, ainsi que *Information Theory, Inference, and Learning Algorithms* de D. MacKay, disponible gratuitement sur le site <http://www.inference.phy.cam.ac.uk/mackay/itila/book.html>