Semaine 8 : Série d'exercices sur les signaux et l'entropie [Solutions]

1 Questions-test

a) AAAAAAHH et HAHAHAHA : L'entropie du premier mot est égale à $(3/4)\log_2(4/3)+(1/4)\log_2(4) = 2-(3/4)\log_2(3) = 0.811$; celle du deuxième est plus grande : $(1/2)\log_2(2)+(1/2)\log_2(2) = 1$.

Certains suivant la « définition » (approximative) de l'entropie comme « nombre moyen de questions que l'on doit poser pour deviner une lettre dans une séquence » pourraient être ammenés à penser que les deux cas sont les mêmes : après tout, il faudrait poser **1** question dans les deux cas : « est-ce A ? ».

Mais imaginez que le fait de poser une question vous coûte quelque chose (et le fait de trouver vous rapporte). Alors dans le cas `AAAAAAHH`, certains joueurs voudront maximiser leur gain en essayant de proposer '`A`' **SANS** poser de question. (Cela ne fait évidemment pas de sens pour le jeu `HAHAHAHA`).

Dans ce contexte (certains joueurs pariant même **sans** poser de question), vous pourrez alors avoir que le « nombre moyen de questions que l'on doit poser pour deviner une lettre dans une séquence » soit plus petit que 1 et retrouvez une cohérence avec le résultat obtenu par la formule mathématique [1].

b) ABBA et BEBE : L'entropie est égale à 1 dans les deux cas.

c) CALC et CALCUL : L'entropie du premier mot est $1/2\log_2(2)+(1/2)\log_2(4) = 1.5$; celle du deuxième est plus grande : $(2/3)\log_2(3)+(1/3)\log_2(6) = \log_2(3)+1/3 = 1.918$,

d) MEDITERRANNEE et MEDETERRENNEE : L'entropie du premier mot est clairement plus grande : les apparitions des lettres sont *moins* équilibrées ; vous pouvez considérer avoir le même alphabet de huit lettres (A, D, E, I, M, N, R, T) dans les deux cas avec des nombres d'apparitions nuls dans le second cas.

e) EPFL et EEPPFFLL : L'entropie est égale à 2 dans les deux cas.

2 Quelques pâtisseries

En comptant simplement le nombre d'apparitions des lettres :

a. TRESSE AU BEURRE ($n = 8$)

E	R	_	S	U	A	B	T
4	3	2	2	2	1	1	1

b. PAIN AU CHOCOLAT ($n = 11$)

A	_	C	O	H	I	L	N	P	T	U
3	2	2	2	1	1	1	1	1	1	1

c. CROISSANT FOURRE ($n = 12$)

1. Pour être totalement rigoureux, il faut que la probabilité de ne pas vouloir poser une question soit $1 + p\log(p) + (1-p)\log(1-p)$ où p est la probabilité de 'H'

R	O	S	_	A	C	E	F	I	N	T	U
3	2	2	2	1	1	1	1	1	1	1	1

d. CHOUX A LA CREME ($n = 11$)

_	A	C	E	H	L	M	O	R	U	X
3	2	2	2	1	1	1	1	1	1	1

e. GATEAUX MILANAIS ($n = 12$)

A	I	_	E	G	L	M	N	S	T	U	X
4	2	1	1	1	1	1	1	1	1	1	1

on voit alors facilement que

$$H(a) < H(b) = H(d) < H(c)$$

$H(a) < H(b)$: moins de lettres et plus d'écart à l'équi-probabilité

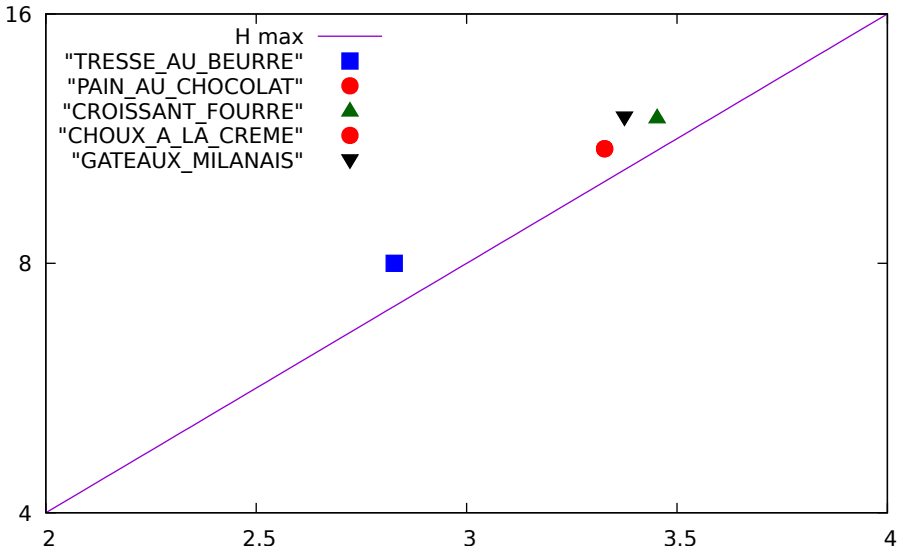
$H(b) = H(d)$: même nombre de lettres et même distribution de probabilités

$H(d) < H(c)$: moins de lettres et plus d'écart à l'équi-probabilité (presque la même distribution mais la lettre ajoutée dans (c) contribue à uniformiser la distribution (un 1 de plus))

La place de la séquence (e) est quant à elle plus difficile à déterminer. Est est clairment en dessous de (c) (car même nombre de lettres mais moins uniforme) mais il nous semble difficile de la placer par rapport à (d).
En calculant les entropies explicitement (à l'aide d'une machine à calculer ou de `http://www.shannonentropy.netmark.pl/`, ou, encore mieux, votre propre programme C++ [cf exercice 5 de la série 8]), on trouve finalement que

$$H(a) < H(b) = H(d) < H(e) < H(c)$$

Si l'on utilise un graphique similaire au cours ($H(X)$ en fonction de n avec une échelle logarithmique de n), voici ce que l'on obtient :



3 Quelques pièces de monnaie

a) $\log_3(N)$. On utiliserait la même définition mais avec un $\log_3$ au lieu d'un $\log_2$, ce qui au final ne fait que multiplier par une constante puisque $\log_3(x) = \log_2(x)/\log_2(3)$.

On utilise ici cette *entropie ternaire* du fait que la balance divise le problème en 3 à chaque pesée (au lieu de 2 dans le cas du cours avec des question oui/non).

b) Appelons les 3 pièces A, B et C. Si on pose A à gauche et B à droite, on a 3 possibilités :
- Si la balance penche à gauche, alors on sait que la pièce plus légère est B.
- Si la balance penche à droite, alors on sait que la pièce plus légère est A.
- Si la balance reste stable, alors on sait que la pièce plus légère est C.

Donc une seule pesée suffit.

Le lien avec l'entropie et les codes de Shannon-Fano est que $1{=}\log_3(3) = \log_3(1/(1/3))$, les 3 situations (A plus légère, B plus légère, C plus légère) étant a priori équiprobable (probabilité 1/3).

c) L'entropie ternaire ($\log_3$) vaut dans ce cas $2 = \log_3(9)$.

Voyons maintenant comment réaliser ces deux pesées. Désignons les 9 pièces par les lettres A à I. Une d'entre elles est plus légère : il y a donc 9 possibilités en tout. Pour diviser par 3 l'ensemble des possibilités, nous effectuons la pesée suivante :

ABC–DEF

(signifiant que l'on place les pièces A, B et C à gauche et D, E et F à droite de la balance ; les trois autres pièces restant sur la table.)

- Si la balance penche à gauche (ce qu'on note ABC>DEF : le poids de ABC est plus grand que celui de DEF), alors on sait que la pièce plus légère est dans DEF.
- Si la balance penche à droite (ce qu'on note ABC<DEF), alors on sait que la pièce plus légère est dans ABC.
- Si la balance reste stable (ce qu'on note ABC=DEF), alors on sait que la pièce plus légère est G, H ou I (restées sur la table).

Ainsi, on a bien réduit par 3 l'ensemble des possibilités. Il suffit ensuite de répéter l'opération avec les trois pièces suspectes comme à la question **b)**.

d) Dans ce contexte le nombre de cas possibles double car on ne sait pas si la fausse pièce est plus légère ou plus lourde.

Pour 9 pièces, il y a donc 18 cas possibles. On obtient comme entropie ternaire : $\log_3(18)$ qui donne une valeur entre 2 et 3 ($\log_3(18) = \log_3(2 \times 9) = 2 + \log_3(2)$, compris entre 2.5 et 3 puisque $\sqrt{3} < 2 < 3$). Donc cette fois il sera nécessaire de faire 3 pesées (dans certains cas).

On procède comme avant en séparant les 9 pièces en 3 groupes de 3 :

Si ABC=DEF la fausse pièce est l'une des 3 restantes (G, H, I). On compare alors G et H :

Si équilibre la fausse pièce est I. Il suffit de la comparer à une vraie pièce pour savoir si la fausse est plus lourde ou plus légère.

Si ça penche vers G , c'est que G est la fausse pièce (plus lourde) ou que H est la fausse pièce, plus légère. En comparant G avec une autre pièce en une pesée finale, on déterminera laquelle est fausse (et comment).

Si ça penche vers H , on applique la même méthode.

Si ABC>DEF , on compare alors ABC à GHI.

Si équilibre : c'est que la pièce fausse est dans DEF et qu'elle est plus légère que les autres. Il suffit de comparer D et E pour déterminer, en une pesée, laquelle parmi D, E ou F est plus légère.

Si ABC>GHI , c'est que la pièce fausse est dans ABC et est plus lourde. On procède comme dans le cas précédent.

La balance ne peut pas pencher vers GHI

Si la balance penche vers DEF , on est dans le cas symétrique du précédent, et la méthode est la même.

4 Entropie et mots de passe

a) Pour une séquence de n lettres, l'entropie est comprise entre 0 et $\log_2(n)$.

b) $H(XX) = H(X)$

a. c'est normal du point de vue du jeu de choisir une lettre au hasard dans le mot : le jeu reste exactement le même avec XX qu'avec X.

b. Par contre du point de vue du choix d'un mot de passe cela semble faux : c'est plus difficile de deviner XX que de deviner X.

Il est donc important, lorsqu'on définit l'entropie, de bien savoir de quel « jeu » on parle : c'est moins l'entropie d'une séquence en tant que telle que celle de la façon dont on l'utilise (choix d'une lettre parmis n ou choix d'une séquence en tant que telle parmi toutes les séquences possibles).

c) Pour des séquences de longueur L écrites avec un alphabet de n lettres différentes :

a. l'entropie telle que définie en cours est comprise entre 0 et le $\log_2$ du nombre de lettres différentes ; si $L < n$ ce nombre est au plus L et si $L > n$ alors ce nombre est au plus n et donc la borne maximale pour l'entropie telle que vu en cours est $\min(\log_2(L), \log_2(n))$;

b. il y a n^L séquences possibles

c. la définition reste la même, *sauf que* les p_j ne sont plus des probabilités de lettres, mais des probabilités de *séquences de lettres* !

d)

a. Si toutes les séquences ont la même probabilité, cette probabilité est alors $1/n^L$ et l'on retrouve, comme dans le cours, que l'entropie est maximale et égale au log du nombre de possibles, c'est-à-dire dans ce cas $\log(n^L)$.

La « complexité » correspond donc à l'entropie dans le cas où toutes les séquences sont équiprobables (ce qui n'est bien sûr pas le cas « dans la vraie vie » : mots connus (dictionnaires), dates diverses, etc.).

b. $2\,L$: il suffit de doubler la longueur L d'un mot de passe pour doubler sa « complexité », sans changer l'alphabet utilisé.

c. n^2 : il faut *mettre au carré* la taille de l'alphabet utilisé pour doubler la « complexité » d'un mot de passe, sans changer sa longueur.

5 Est-ce que l'entropie augmente ou diminue ?

a) La réponse est non. Voici un contre-exemple : considérons la séquence ABB, dont l'entropie vaut $\frac{2}{3}\log_2(3/2) + \frac{1}{3}\log_2(3) = \log_2(3) - \frac{2}{3} \simeq 0.92$. Si on ajoute la lettre A à cette séquence, on obtient ABBA, dont l'entropie vaut $\frac{1}{2}\log_2(2) + \frac{1}{2}\log_2(2) = 1$, qui est donc plus grande que 0.92.

b) Ici, la réponse est oui, et demande donc une preuve qui montre que c'est toujours le cas pour n'importe quelle séquence. Allons-y ! (accrochez-vous...)

Soit n le nombres de lettres *différentes* apparaissant dans la séquence de longueur N, et appelons N_i le nombre d'apparitions de la lettre i dans la séquence. Selon la définition vue en cours, l'entropie de la séquence d'origine s'écrit

$$H = \sum_{i=1}^{n} \frac{N_i}{N} \log_2\left(\frac{N}{N_i}\right)$$

Ceci peut se réécrire

$$H = \sum_{i=1}^{n} \frac{N_i}{N} \left(\log_2(N) - \log_2(N_i)\right) = \log_2(N) - \sum_{i=1}^{n} \frac{N_i}{N} \log_2(N_i) \tag{1}$$

car $\sum_{i=1}^{n} N_i = N$.

Qu'en est-il de l'entropie H' de la séquence avec une lettre supplémentaire ? Remarquer tout d'abord que cette lettre supplémentaire n'apparaît qu'une seule fois dans la nouvelle séquence. Et donc

$$H' = \sum_{i=1}^{n} \frac{N_i}{N+1} \log_2\left(\frac{N+1}{N_i}\right) + \frac{1}{N+1} \log_2(N+1)$$

A nouveau, ceci peut se réécrire, en utilisant que $\sum_{i=1}^{n} N_i + 1 = N + 1$:

$$\begin{aligned} H' &= \sum_{i=1}^{n} \frac{N_i}{N+1} \left(\log_2(N+1) - \log_2(N_i)\right) + \frac{1}{N+1} \log_2(N+1) = \log_2(N+1) - \sum_{i=1}^{n} \frac{N_i}{N+1} \log_2(N_i) \\ &= \log_2(N+1) - \frac{N}{N+1} \sum_{i=1}^{n} \frac{N_i}{N} \log_2(N_i) \end{aligned}$$

En utilisant l'égalité (1), on trouve que

$$H' = \log_2(N+1) - \frac{N}{N+1}\left(\log_2(N) - H\right) = \log_2(N+1) - \log_2(N) + H + \frac{1}{N+1}(\log_2(N) - H)$$

Vu que $\log_2(N+1) > \log_2(N)$ et que $H \leq \log_2(N)$ (quelle que soit la séquence d'origine), on conclut que $H' > H$. Ouf !