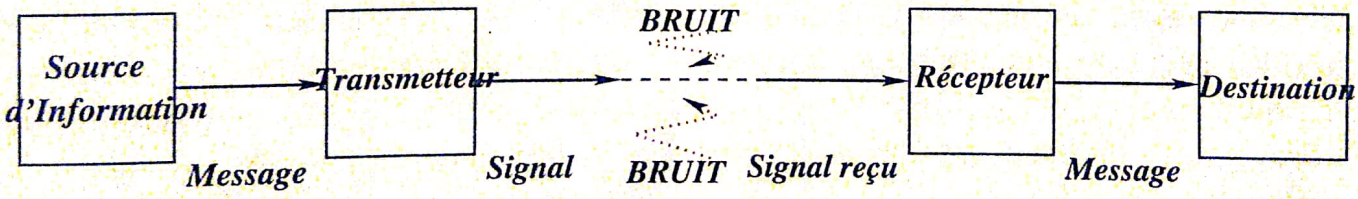
Théorie de l'information de Claude Shannon : aux fondements de la numérisation et des codes correcteurs d'erreurs (Michel Sortais).

Au moment où commence la deuxième guerre mondiale, un jeune ingénieur et mathématicien américain, Claude Shannon, se penche sur quelques questions fondamentales ayant trait à l'édition et à la transmission de messages :

1. Comment traduire *efficacement* un message vocal ou textuel en une suite de symboles binaires ? (Shannon fait partie des tout premiers chercheurs à employer le mot *'bit'* ; par traduction *efficace*, on entend une traduction où sont employés, en moyenne, un nombre minimum de bits/mot).

2. Comment rendre *fiable* la transmission d'une séquence de bits sur un canal imparfait, tout en *allongeant le moins possible* cette séquence ? Dans cette seconde question, on se figure que le message à transmettre a déjà été transformé en une séquence de bits, et que l'on est tenu de rajouter des "bits de contrôle" à la séquence d'origine parce que le canal de transmission altère certains des bits (cf schéma ci-dessous).



A l'évidence, l'étude de ces questions intéresse au plus haut point l'Etat-Major des armées américaines ; d'ailleurs Shannon participera directement à l'effort de guerre de son pays dès 1941, de par son travail au sein des laboratoires *Bell*. Mais la profondeur des réponses qu'il propose alors permettra de mettre en route la fameuse *"révolution numérique"* de la seconde moitié du XXème siècle [1] - il n'y a encore pas si longtemps, on faisait développer des photos argentiques pour garder quelques souvenirs de vacances, tandis que la Passion selon St Matthieu tenait sur huit faces de disques 33T ...

1 Entropie et Information

Dans son article fondateur paru en 1948 ([Sha48]), Claude Shannon commence par se pencher sur la question suivante :
Comment faudrait-il mesurer l'incertitude liée au résultat de telle ou telle autre expérience à caractère aléatoire ?
A titre d'exemple (simplet), considérons les trois expériences suivantes :

1°) On jette un dé équilibré à vingt faces (*icosaèdre*).

2°) On jette un dé équilibré à six faces.

3°) On jette un dé à six faces faisant apparaître la face numéro 1 avec prob. 1/2 et chacune des cinq autres faces avec prob. 1/10.

De toute évidence: le résultat de la première expérience est "plus incertain" que celui de la seconde, et le résultat de la seconde expérience est "plus incertain" que celui de la troisième. Mais *"comment convient-il de quantifier ces différents niveaux d'incertitude ?"*
Shannon fournit une réponse mémorable à cette première question : il n'y a guère qu'une seule manière de mesurer convenablement de telles incertitudes !

1.1 L'entropie comme mesure d'incertitude

Revenons à notre exemple élémentaire (comparaison de trois lancers de dés). Comme l'issue de chacun de ces lancers est incertaine, il convient de considérer que les nombres qui vont apparaître sont des variables aléatoires X_1, X_2, X_3 : ici, on a bien entendu affaire à des variables al. discrètes, la 1ère étant à valeurs dans $[\![1;20]\!]$ et les deux suivantes à valeurs dans $[\![1;6]\!]$. On est en fait à la recherche d'une fonction H qui associe à chaque var. al. discrète X un nombre $H(X) > 0$ mesurant l'incertitude liée à la réalisation de X. Dans l'exemple qui nous intéresse, on s'attend bien sûr à ce que

$$H(X_1) > H(X_2) > H(X_3) > 0$$

(le résultat du 1er lancer est plus incertain que celui du 2ème, et celui du 2ème plus incertain que celui du 3ème). On s'attend aussi à ce que $H(X)$ ne dépende pas du nom donné aux valeurs prises par X. Dans notre exemple standard, il est bien évident que l'incertitude liée au lancer du dé à 20 faces ne change pas si celles-ci sont renommées $a, b, c, d, \ldots, s, t$; de même, l'incertitude liée au lancer du 1er ou du 2ème dé à 6 faces ne varierait pas si l'on devait renommer celles-ci $\alpha, \beta, \gamma, \delta, \varepsilon, \zeta$. Notre fonction H dépend donc de la variable X seulement à travers sa loi, en sorte que $H(X_1) = H(\frac{1}{20}, \frac{1}{20}, \ldots, \frac{1}{20})$, $H(X_2) = H(\frac{1}{6}, \frac{1}{6}, \ldots, \frac{1}{6})$,

$$H(X_3) = H(\frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}) = H(\frac{1}{10}, \frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$$

(la dernière égalité tient encore au fait que ces niveaux d'incertitude sont insensibles aux "réétiquetages").
Donnons-nous un domaine de définition pour H ; pour en entier $k \geq 1$, soit

$$\Sigma_k = \{(p_1, \ldots, p_k) \in [0;+\infty[^k | p_1 + p_2 + \ldots + p_k = 1\}$$

Vecteur de proba : un vecteur avec des valeurs positives dont la somme vaut 1

le simplexe fermé constitué de toutes les lois de variables discrètes prenant au plus k valeurs (Σ_2 est un segment que l'on pourra dessiner dans le plan muni d'un repère, Σ_3 un triangle que l'on pourra représenter dans l'espace). Il est maintenant naturel de définir la fonction H sur la réunion

$$\Sigma = \bigcup_{k\geq 1} \Sigma_k,$$

et l'on s'attend à ce que $H : \Sigma \longrightarrow [0; +\infty[$ soit une fonction *symétrique* ($H(p_{\sigma(1)}, \ldots, p_{\sigma(k-1)}, p_{\sigma(k)}) = H(p_1, \ldots, p_{k-1}, p_k)$ pour toute permutation σ) satisfaisant quelques "règles de bon sens" supplémentaires, à savoir les **Axiomes** ci-dessous :

- **(A1): Continuité** si $(p_1^{(n)}, \ldots, p_k^{(n)}) \underset{n\to\infty}{\longrightarrow} (p_1, \ldots, p_k)$ dans Σ_k, on aura

$$H(p_1^{(n)}, \ldots, p_k^{(n)}) \underset{n\to\infty}{\longrightarrow} H(p_1, \ldots, p_k).$$

  (cette règle de continuité se comprend bien en représentant une suite de points sur le segment Σ_2 convergeant vers un point précis, ou encore une suite convergente sur le triangle Σ_3).

- **(A2): Croissance** on a $H(1) = 0$, et pour tout $k \geq 1$:

$$H(\frac{1}{k+1}, \frac{1}{k+1}, \ldots, \frac{1}{k+1}) > H(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$$

  (plus une urne contient de bulletins, plus un tirage au sort à l'aveugle dans cette urne est incertain).

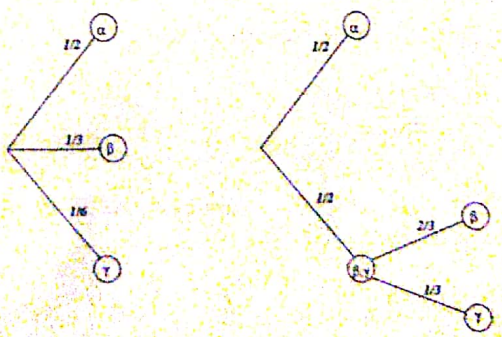
- **(A3): Conditionnement** si $(p_1, \ldots, p_k) \in [0; +\infty[^k$ et $(q_1, \ldots, q_l) \in [0; +\infty[^l$ sont tels que $\sum_{i=1}^k p_i = p$, $\sum_{j=1}^l q_j = q$, avec $p, q > 0$ et $p+q = 1$, on doit avoir

$$\begin{aligned} H(p_1, \ldots, p_{k-1}, p_k, q_1, q_2, \ldots, q_l) &= H(p, q) + p \cdot H(\tfrac{p_1}{p}, \ldots, \tfrac{p_{k-1}}{p}, \tfrac{p_k}{p}) \\ &\quad + q \cdot H(\tfrac{q_1}{q}, \tfrac{q_2}{q}, \ldots, \tfrac{q_l}{q}) \end{aligned}$$

  Ce dernier axiome semble plus difficile à justifier ou même à comprendre, illustrons-le sur un exemple élémentaire, proposé par Shannon *himself* dans son fameux article paru en 1948 :

$$\begin{aligned} H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) &= H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2} \cdot H(1) + \tfrac{1}{2} \cdot H(\tfrac{2}{3}, \tfrac{1}{3}) \\ &= H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2} H(\tfrac{2}{3}, \tfrac{1}{3}) \end{aligned}$$

  (puisque $H(1) = 0$).
  Ces identités se comprennent parfaitement bien en recourant au schéma ci-dessous :



Bien entendu, on pourra illustrer ce 3ème axiome en comparant d'autres arbres de probabilité équivalents. Une comparaison particulièrement éclairante consiste à décomposer un choix uniforme parmi $m \cdot n$ valeurs en un premier choix uniforme parmi m valeurs, suivi d'un second choix uniforme parmi n valeurs, ce qui conduit à l'identité

$$H(\frac{1}{mn}, \frac{1}{mn}, \ldots, \frac{1}{mn}) = H(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}) + H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$$

Mais alors, quitte à poser $h(N) = H(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N})$, on voit apparaître une propriété d'*additivité*

$$h(mn) = h(m) + h(n)$$

qui n'est pas sans rappeler notre bonne vieille fonction *logarithme* !
Réflexion faite, il n'est plus trop étonnant de voir apparaître l'*Entropie de Shannon* H comme unique solution au problème posé (trouver une fonction symétrique pour laquelle **(A1)**, **(A2)**, **(A3)** sont vérifiés) :

Théorème:

Si $H : \Sigma \longrightarrow [0; +\infty[$ est une fonction symétrique vérifiant $(A1), (A2)$ et $(A3)$, alors H est de la forme

$$H(p_1, \ldots, p_n) = -\kappa \sum_{i=1}^{n} p_i \ln p_i$$

pour une certaine constante $\kappa > 0$.

on prend $K = \frac{1}{\ln(2)}$

Réciproquement, toute fonction $H : \Sigma \longrightarrow [0; +\infty[$ de la forme donnée ci-dessus définit une fonction symétrique sur Σ vérifiant $(A1), (A2)$ et $(A3)$!

N.B. : dans la formule définissant H, la constante multiplicative κ ne joue pas du tout un rôle essentiel ; sa présence tient simplement au fait qu'il nous faut choisir une "unité d'incertitude" pour définir H de manière tout à fait univoque. En adoptant le point de vue de l'Informaticien, on pourra choisir l'expérience du "Pile ou Face équilibré" comme référence et faire en sorte que H prenne la valeur 1 pour la loi de Bernoulli correspondante, ce qui revient à fixer

$$\kappa = \frac{1}{\ln 2}$$

Autrement dit, dans toute la suite on posera

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \,,$$

avec des logarithmes pris en base 2.

$I(X,Y)$ incertitude sur X qui disparaît en connaissant Y

$= H(X) - H(X|Y)$

1.2 Entropie conditionnelle et information mutuelle

Etant données deux variables aléatoires discrètes X, Y (interdépendantes ou indépendantes), on définit l'entropie de X *conditionnellement à* $(Y = y)$ par

$$H(X|Y=y) = -\sum_x \mathbb{P}\{X=x|Y=y\} \times \log\left[\mathbb{P}\{X=x|Y=y\}\right],$$

la somme ci-dessus portant sur toutes les valeurs possibles x de X, et y étant une valeur possible (fixée) de Y. L'entropie de X *conditionnellement à* Y est ensuite définie en passant à la moyenne sur les valeurs prises par Y :

$$H(X|Y) = \sum_y \mathbb{P}\{Y=y\} \times H(X|Y=y)$$

et il s'avère qu'une telle entropie conditionnelle peut aussi s'obtenir à partir de l'entropie conjointe $H(X,Y)$ et de l'entropie marginale $H(Y)$ comme suit :

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

Vérifions-le sur un premier exemple tout à fait abordable : si X et Y sont toutes deux à valeurs dans l'alphabet fini $\mathcal{A} = \{a,b,c,d\}$ et prennent de telles valeurs conjointement suivant le tableau de prob. ci-dessous

$X \setminus Y$	a	b	c	d
a	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
b	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
c	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
d	$\frac{1}{4}$	0	0	0

Des sommations horizontales permettent de voir que

$$\mathbb{P}(X=a) = \mathbb{P}(X=b) = \mathbb{P}(X=c) = \mathbb{P}(X=d) = \frac{1}{4}$$

La v.a. X, considérée isolément, est donc distribuée uniformément sur $\mathcal{A} = \{a,b,c,d\}$, tandis que Y ne l'est pas, puisque

$$\mathbb{P}(Y=a) = \frac{1}{2} > \mathbb{P}(Y=b) = \frac{1}{4} > \mathbb{P}(Y=c) = \mathbb{P}(Y=d) = \frac{1}{8},$$

comme le montrent des sommations pratiquées verticalement dans le tableau. En pratiquant des divisions par $\mathbb{P}(Y=a) = \frac{1}{2}$ de chacun des termes situés dans la première colonne, on voit ensuite que la loi de X conditionnellement à $(Y=a)$ est donnée par

$$\mathbb{P}(X=a|Y=a) = \frac{1/8}{1/2} = \frac{1}{4},\ \mathbb{P}(X=b|Y=a) = \mathbb{P}(X=c|Y=a) = \frac{1}{8},\ \mathbb{P}(X=d|Y=a) = \frac{1}{2}$$

Autrement dit, $Loi(X|Y=a)$ est donnée à travers le vecteur de probabilité $(\frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{2})$. Par voie de conséquence, on a donc

$$\begin{aligned} H(X|Y=a) &= -\left[\tfrac{1}{4}\cdot\log\tfrac{1}{4} + \tfrac{1}{8}\cdot\log\tfrac{1}{8} + \tfrac{1}{8}\cdot\log\tfrac{1}{8} + \tfrac{1}{2}\cdot\log\tfrac{1}{2}\right] \\ &= -\left[\tfrac{1}{4}\cdot(-2) + \tfrac{1}{8}\cdot(-3) + \tfrac{1}{8}\cdot(-3) + \tfrac{1}{2}\cdot(-1)\right] \\ &= \tfrac{7}{4} = 1,75 \end{aligned}$$

Des calculs analogues permettent de voir que $Loi(X|Y=b)$ est donnée à travers le vecteur de probabilité $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0)$, en sorte que

$$\begin{aligned} H(X|Y=b) &= -\left[\tfrac{1}{4}\cdot\log\tfrac{1}{4} + \tfrac{1}{2}\cdot\log\tfrac{1}{2} + \tfrac{1}{4}\cdot\log\tfrac{1}{4} + 0\right] \\ &= -\left[\tfrac{1}{4}\cdot(-2) + \tfrac{1}{2}\cdot(-1) + \tfrac{1}{4}\cdot(-2)\right] \\ &= \tfrac{3}{2} = 1,5 \end{aligned}$$

De même :

$$\begin{aligned} H(X|Y=c) &= H(X|Y=d) \\ &= -\left[\tfrac{1}{4}\cdot\log\tfrac{1}{4} + \tfrac{1}{4}\cdot\log\tfrac{1}{4} + \tfrac{1}{2}\cdot\log\tfrac{1}{2} + 0\right] \\ &= -\left[\tfrac{1}{4}\cdot(-2) + \tfrac{1}{4}\cdot(-2) + \tfrac{1}{2}\cdot(-1)\right] \\ &= \tfrac{3}{2} = 1,5 \end{aligned}$$

En effectuant une *moyenne* sur les valeurs possibles de Y, on obtient enfin

$$\begin{aligned} H(X|Y) &= \textstyle\sum_{y=a}^{d} H(X|Y=y)\cdot\mathbb{P}(Y=y) \\ &= 1,75\cdot\mathbb{P}(Y=a) + 1,5\cdot\mathbb{P}(Y=b) + 1,5\cdot\mathbb{P}(Y=c) + 1,5\cdot\mathbb{P}(Y=d) \\ &= \tfrac{7}{4}\cdot\tfrac{1}{2} + \tfrac{3}{2}\cdot\tfrac{1}{4} + \tfrac{3}{2}\cdot\tfrac{1}{8} + \tfrac{3}{2}\cdot\tfrac{1}{8} \\ &= \tfrac{13}{8} = 1,625 \end{aligned}$$

Vérifions maintenant que le calcul de la différence $H(X,Y) - H(Y)$ entre *l'entropie conjointe* $H(X,Y)$ et *l'entropie marginale* $H(Y)$ permet de retrouver $H(X|Y)$:

$$\begin{aligned} H(X,Y) &= -\textstyle\sum_{x,y}\mathbb{P}(X=x, Y=y)\cdot\log\mathbb{P}(X=x, Y=y) \\ &= -\left[\tfrac{1}{8}\cdot\log\tfrac{1}{8} + \tfrac{1}{16}\cdot\log\tfrac{1}{16} + \ldots + 0\right] \\ &= \tfrac{108}{32} = \tfrac{27}{8}, \end{aligned}$$

$$H(Y) = -\left[\frac{1}{2}\cdot\log\frac{1}{2} + \frac{1}{4}\cdot\log\frac{1}{4} + 2\cdot\frac{1}{8}\cdot\log\frac{1}{8}\right] = \frac{7}{4},$$

ainsi

$$H(X,Y) - H(Y) = \frac{27}{8} - \frac{7}{4} = \frac{13}{8} = 1,625 = H(X|Y)$$

En résumé, les entropies conditionnelles $H(X|Y)$ et $H(Y|X)$ peuvent être calculées de deux façons équivalentes, et l'équivalence en question correspond à une identité connue sous le nom de **"Règle d'enchaînement"** (***"Chain Rule"*** en anglais) :

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

On se gardera de croire que les entropies conditionnelles $H(X|Y)$ et $H(Y|X)$ vérifient une propriété de symétrie : en général, ces deux entropies conditionnelles sont bien distinctes - de même qu'il convient de bien distinguer les prob. conditionnelles $\mathbb{P}\{X=x|Y=y\}$ et $\mathbb{P}\{Y=y|X=x\}$ ...

En revanche, la règle d'enchaînement permet de faire apparaître une symétrie intéressante mettant en jeu la notion d'*Information*. Dans la situation présente, il semble naturel de définir l'*Information portée par Y sur X* comme *quantité d'incertitude sur X disparaissant au moment où l'on prend connaissance de la valeur prise par Y* ; plus mathématiquement, il convient donc de définir cette information $I(Y,X)$ par l'identité :

$$I(Y,X) = H(X) - H(X|Y)$$

Mais la règle d'enchaînement nous montre que

$$H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)\,,$$

en sorte que

$$I(X,Y) = I(Y,X) = H(X) + H(Y) - H(X,Y)\,,$$

remarquable propriété de l'*"information mutuelle"* !
Ainsi a-t-on, dans le cas des variables X, Y examinées précédemment à titre d'exemple :

$$H(X) = -4 \times \frac{1}{4} \cdot \log \frac{1}{4} = \log 4 = 2$$

puis

$$I(X,Y) = H(X) - H(X|Y) = 2 - \frac{13}{8} = \frac{3}{8} = 0,375$$

Un bon exercice complémentaire consiste à vérifier que $H(Y|X) = \frac{11}{8}$, en sorte que

$$I(Y,X) = H(Y) - H(Y|X) = \frac{7}{4} - \frac{11}{8} = \frac{3}{8} = I(X,Y)$$

Rappelons qu'un calcul précédent nous avait donné $H(X,Y) = \frac{108}{32} = \frac{27}{8}$; ceci nous permet à présent d'établir que

$$H(X) + H(Y) - H(X,Y) = 2 + \frac{7}{4} - \frac{27}{8} = \frac{30}{8} - \frac{27}{8} = \frac{3}{8} = I(X,Y) = I(Y,X)$$

Ainsi, dans l'exemple que nous venons de traiter, l'information portée par Y sur la variable uniforme X est plutôt faible : 0,375 bit seulement, alors que X a une entropie s'élevant à 2 bits ...

Plus généralement, l'information mutuelle $I(X,Y)$ fournit un bel outil de mesure de l'interdépendance entre les variables X et Y puisque, comme nous le verrons dans un chapitre ultérieur *(Les canaux bruités et leurs capacités)* :

$$I(X,Y) = 0 \iff X \text{ et } Y \text{ sont indépendantes}$$

1.3 Quelques premiers exercices sur les notions d'*Entropie* et d'*Information*

1.3.1 Calculs d'Entropie en basse dimension

1. Représenter avec soin la fonction f définie sur $[0;1]$ par $f(x) = H(x;(1-x))$, en vérifiant que f est concave, à valeurs ≥ 0, nulle seulement en $x = 0$ et $x = 1$, que son graphe admet $x = \frac{1}{2}$ pour axe de symétrie, et qu'elle atteint son maximum de manière unique en $x = \frac{1}{2}$ avec : $f(\frac{1}{2}) = 1$ (conformément à notre choix d'unité).

2. Comment représenter graphiquement la fonction de 2 variables g définie par
$$g(x,y) = H(x;y;(1-x-y)) \ ,$$
où $x, y \geq 0$ sont tels que $(x+y) \leq 1$?
Quelles sont les propriétés remarquables de cette fonction g ?
Où atteint-elle son maximum, et quel est celui-ci ?

1.3.2 Comparaison d'entropies

DJ WhatsUpDude doit se produire au cours d'une nuit de folie, à bord d'une péniche amarrée sur les quais de Seine. Ses prestations comprennent la prononciation, toutes les 20 secondes, d'un mot choisi entièrement au hasard au sein de son trésor lexical de 10'000 mots ; en tout, il sera amené à prononcer un millier de mots durant la nuit, avec de possibles répétitions.
Vérifier que l'incertitude liée à la séquence de mots qu'il aura prononcée durant la nuit est *bien moindre* que celle correspondant à la contemplation d'une image télévisée sur un écran comportant 500 lignes et 600 colonnes de pixels, chaque pixel pouvant prendre au hasard l'une des 16 valeurs chromatiques envisageables.

1.3.3 Entropies comparées de X et de $Y = \varphi(X)$

Comment se comparent les entropies de X et de $Y = \varphi(X)$ si

1. $Y = 2^X$?
2. $Y = \cos X$??

A quelle condition sur la fonction φ peut-on affirmer que X et $Y = \varphi(X)$ ont *même entropie* ?
Qu'en est-il de l'information mutuelle $I(X;Y)$?

1.3.4 Autour des axiomes (A1)-(A2)-(A3)

1. Vérifier que pour chaque entier $n \geq 1$:
$$H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) = \log n \, ,$$
et que H atteint là son maximum sur Σ_n.

2. Montrer que la fonction d'Entropie H définie par Shannon vérifie bien l'axiome de conditionnement **(A3)**.

1.3.5 Entropies associées aux lois classiques

1. Pour $X \hookrightarrow \mathcal{B}(n;p)$, comment varie $H(X)$ en fonction de p ?
   (On pourra mener les calculs explicitement pour de gentilles valeurs de n et p, e.g. $n = 4$ puis $p = 0,25$, $p' = 0,5$, $p'' = 0,55$).
2. Pour $Y \hookrightarrow \mathcal{H}(N;n;p)$, avec $Np \geq n$ et $N(1-p) \geq n$, comment se comparent les entropies $H(X)$ et $H(Y)$? (Ici encore, on pourra mener quelques calculs explicites à titre d'exemple et pour de gentilles valeurs de N, n et p, e.g. $N = 10$, $n = 4$ puis $p = 0,4$, $p' = 0,5$, $p'' = 0,6$).
3. Comment calculer l'entropie d'une variable de Poisson ? Quelle devrait être la nature de la dépendance en λ (paramètre poissonien) de cette entropie ?
4. Comment calculer l'entropie d'une variable géométrique ? Quelle devrait être la nature de la dépendance en p (paramètre de loi géométrique) de cette entropie ?

1.3.6 Entropie conditionnelle et information mutuelle 1er exemple

Etant données $X : \Omega \longrightarrow \{a,b\}$ et $Y : \Omega \longrightarrow \{\alpha,\beta\}$ ayant une loi conjointe donnée à travers le tableau ci-dessous :

$X \setminus Y$	α	β	
a	0	$\frac{3}{4}$	$3/4$
b	$\frac{1}{8}$	$\frac{1}{8}$	$1/4$
	$\frac{1}{8}$	$7/8$	

Calculer

1. L'entropie conjointe $H(X,Y)$.
2. Les entropies marginales $H(X)$ et $H(Y)$.
3. Les entropies conditionnelles $H(X|Y=\alpha)$ et $H(X|Y=\beta)$.
4. Les entropies conditionnelles $H(X|Y)$ et $H(Y|X)$.
5. L'information mutuelle $I(X,Y) = I(Y,X)$.

$H(X|Y=\alpha) =$
$H\left(\frac{P(X=a) \cap P(X=\alpha)}{P(X=\alpha)}\right.$
$P(X=b) \cap P(X=\alpha$
$P(X=$

1.3.7 Valeurs extrêmes de l'information mutuelle

1. Donner un exemple de couple (X,Y) pour lequel $I(X,Y) = I(Y,X) = 0$.
2. Donner un exemple de couple (X,Y) tel que $I(X,Y) = I(Y,X) = H(X)$.

1.3.8 Calculs d'informations

On considère une population humaine constituée d'hommes et de femmes à parts égales. Au sein de cette population, 6% des femmes sont "grandes" (plus de $1,85\,m$) tandis que chez les hommes, cette proportion monte à 20%.

1. Quelle est la probabilité qu'un individu choisi au hasard dans cette population soit un homme *sachant qu'il est* grand (plus de $1,85\,m$) ?
2. Quelle est la quantité d'information apportée par la stature d'un individu ("grand"/"petit") relativement à son sexe (homme/femme) ?

1.3.9 Entropie conditionnelle et information mutuelle

On considère deux var. al. $X : \Omega \longrightarrow \{a, b, c\}$ et $Y : \Omega \longrightarrow \{\alpha, \beta, \gamma, \delta\}$ ayant une loi conjointe donnée à travers le tableau ci-dessous :

$X \setminus Y$	α	β	γ	δ
a	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
b	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{32}$	$\frac{1}{32}$
c	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

Calculer

1. L'entropie conjointe $H(X, Y)$.
2. Les entropies marginales $H(X)$ et $H(Y)$.
3. Les entropies conditionnelles $H(X|Y)$ et $H(Y|X)$.
4. L'information mutuelle $I(X, Y) = I(Y, X)$.

Peut-on affirmer que les variables X et Y sont *indépendantes* ??

1.3.10 Application au jeu *"MASTERMIND"*

Rappelons que dans le jeu *"MASTERMIND"*, un premier joueur (J_1) est censé préparer, de manière cachée, une configuration précise $X = (C_1; C_2; C_3; C_4)$ de quatre couleurs choisies parmi les six couleurs à disposition (répétitions possibles!). Le deuxième joueur (J_2) cherche alors à déterminer précisément quelle est cette configuration ; il en propose une première au hasard, soit $X' = (C'_1; C'_2; C'_3; C'_4)$, puis J_1 répond à J_2 à travers un couple d'entiers $X'' = (m_1; n_1)$ lui indiquant

- le nombre m_1 de couleurs correctement devinées et correctement placées
- le nombre n_1 de couleurs correctement devinées mais mal placées

dans X'.

1. Evaluer $H(X)$, $H(X')$, $H(X, X')$ puis $H(X, X', X'')$.
2. Que vaut $I(X, X')$?
3. Evaluer l'entropie conditionnelle $H\left(X|X' = (V, V, V, V), X'' = (2; 0)\right)$, et comparer cette entropie conditionnelle avec l'entropie originale $H(X)$.

1.3.11 Application en écologie animale

On suppose qu'une population de corbeaux est constituée à parts égales de mâles et de femelles, que 60% de ces oiseaux sont noirs, les autres étant gris, mais que les corbeaux mâles sont trois fois plus susceptibles d'être noirs que ne le sont les corbeaux femelles. Alice observe un corbeau noir au loin. Quelle est la prob. que celui-ci soit un mâle ?
Quelle est l'information portée par la couleur d'un corbeau sur son sexe ?

1.3.12 Entropie conditionnelle et règle d'enchaînement

1. Un couple de bits (B, B') est engendré aléatoirement comme suit :
   - Tout d'abord, B prend la valeur 0 avec prob. p
   - Ensuite, si $B = 0$, B' prend la valeur 0 avec prob. q, tandis que si $B = 1$, B' prend la valeur 0 avec prob. r.

   Utiliser la notion d'entropie conditionnelle et la règle d'enchaînement afin d'évaluer $H(B, B')$.
   Vérifier ensuite que ce calcul de $H(B, B')$ est correct en se référant au tableau croisé des prob. conjointes de B et B'.

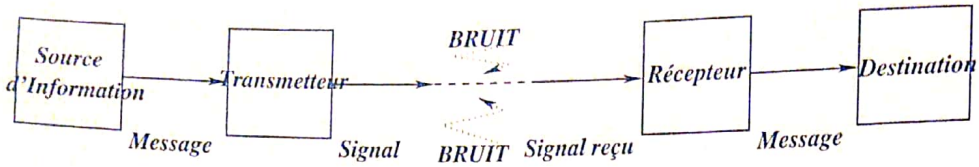
2. Une variable W à valeurs dans $\mathcal{A} = \{a, b, c\}$ se comporte de la façon suivante :
   - W prend la valeur a avec prob. p
   - si W ne prend pas la valeur a, c'est avec prob. q qu'elle prend la valeur b et avec prob. $(1 - q)$ qu'elle prend la valeur c.

   Utiliser la notion d'entropie conditionnelle et la règle d'enchaînement afin d'évaluer $H(W)$.
   Vérifier ensuite que ce calcul de $H(W)$ est correct en se référant aux probabilités $\mathbb{P}\{W = a\}$, $\mathbb{P}\{W = b\}$, $\mathbb{P}\{W = c\}$.

2 Construction de codes-source et 1er Théorème de Shannon

2.1 Présentation du problème, premières définitions

Revenons au *schéma de communication* proposé par Shannon :



et intéressons-nous au "1er maillon" de ce schéma (e.g. *transformation d'une suite de caractères latins, ou encore d'une suite d'idéogrammes chinois (etc) en une suite de "bits"*).

Deux questions apparaissent de façon assez évidente:

1°) *Combien de bits seront nécessaires* à la transcription (codage) d'un texte donné (ou d'une image, ou d'un fichier audio, etc...) ?

2°) *Comment* coder nos données de façon *économique*, l'idée étant d'utiliser, *en moyenne*, un nombre minimum de bits/caractère latin (ou par idéogramme, etc) ?

Le 1er Théorème de Shannon répond de manière tout à fait précise et satisfaisante à la 1ère question en faisant jouer un rôle central à l'entropie H rencontrée précédemment.
Quant à la seconde question, elle trouve sa réponse dans la mise en oeuvre de l'*Algorithme de Huffman*.

Bien entendu, la réponse à la 1ère Question dépend fortement des *propriétés statistiques* des données que l'on entend coder en binaire.
Exemple très simple: on suppose que le texte à coder ne comporte que des caractères latins choisis parmi $\{a, b, c, d\}$.
Comment associer à chacun de ces caractères une suite de bits de façon "économique"?
$\rightsquigarrow$ si chacun des quatre caractères apparaît avec la même fréquence $(\frac{1}{4})$, on pourra par exemple utiliser le codage

$$a \mapsto 00, \qquad b \mapsto 01, \qquad c \mapsto 10, \qquad d \mapsto 11,$$

on utilise alors 2 bits/caractère, et il n'y a pas moyen de faire mieux (en moyenne).
$\rightsquigarrow$ si, par contre, les fréquences d'apparition respectives de ces 4 caractères sont $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ et $\frac{1}{8}$, le codage

$$a \mapsto 0, \qquad b \mapsto 10, \qquad c \mapsto 110, \qquad d \mapsto 111,$$

convient lui aussi, et utilise seulement 1,75 bits/caractère (en moyenne)!

Pour être valide un code doit un code dans lequel aucun des codes n'est le début d'une autre (préfixe).

Notations et définitions:

- Une **source** $\mathcal{W} = (X_1, X_2, \ldots, X_k, \ldots)$ est simplement une suite de variables aléatoires ; si ces v.a. sont *indépendantes* et *identiquement distribuées* (*i.i.d.*), on parlera de **source sans mémoire**. On parlera aussi de **source discrète** si toutes ces v.a. sont discrètes.

- Dans le cas discret, si toutes les v.a. considérées sont à valeurs dans un même ensemble fini $\mathcal{V} = \{v_1, v_2, \ldots, v_m\}$, on dira que $\mathcal{V}$ est l'**alphabet-source**. Dans le cas discret sans mémoire, on pourra considérer l'**extension** d'ordre n associée à $\mathcal{W}$:

$$\mathcal{W}^{(n)} = (X_1^{(n)}, X_2^{(n)}, \ldots, X_k^{(n)}, \ldots),$$

où $X_1^{(n)} = (X_1, X_2, \ldots, X_n), X_2^{(n)} = (X_{n+1}, X_{n+2}, \ldots, X_{2n})$... L'alphabet-source devient alors $\mathcal{V}^n$, et chacun des n-uplets $(v_1, \ldots, v_n)$ apparaît avec une probabilité égale au produit des prob. individuelles d'apparition des v_i (par *indépendance* !).

- L'entropie de la source discrète $\mathcal{W} = (X_1, X_2, \ldots, X_k, \ldots)$ est définie à travers l'identité

$$H(\mathcal{W}) = \lim_{N \to +\infty} \frac{1}{N} H(X_1, X_2, \ldots, X_N),$$

pour autant que la limite ci-dessus existe, bien entendu !
Dans le cas où $\mathcal{W}$ est une source *discrète et sans mémoire*, on obtient tout simplement $H(\mathcal{W}) = H(X_1)$. En effet, dans ce cas précis :

$$\begin{aligned} H(\mathcal{W}) &= \lim_{N \to +\infty} \tfrac{1}{N} H(X_1, X_2, \ldots, X_N) \\ &= \lim_{N \to +\infty} \tfrac{1}{N} \left[H(X_1) + H(X_2) + \ldots + H(X_N) \right] \\ &= \lim_{N \to +\infty} \tfrac{1}{N} \left[N \cdot H(X_1) \right] \\ &= H(X_1) \end{aligned}$$

(La 2ème égalité ci-dessus provient de ce que les variables de la sources sont *indépendantes*, et la 3ème provient de ce qu'elles sont *identiquement distribuées*).
Toujours dans le contexte où $\mathcal{W}$ est discrète et sans mémoire, on pourra remarquer que :

$$H(\mathcal{W}^{(n)}) = H(X_1^{(n)}) = H(X_1, X_2, \ldots, X_n) = nH(X_1) = nH(\mathcal{W})$$

- Si $\mathcal{S} = \{\sigma_1, \ldots, \sigma_d\}$ est un second ensemble de *symboles*, on appellera **code-source** toute application *injective* $\varphi : \mathcal{V} \longrightarrow \mathcal{S}^*$, où $\mathcal{S}^*$ désigne l'ensemble des *suites finies* à valeurs dans $\mathcal{S}$.
On gardera à l'esprit, en priorité, le cas standard où $d = 2$ et $\mathcal{S} = \{0, 1\}$ (*symboles binaires*).

- Le code-source $\varphi : \mathcal{V} \longrightarrow \mathcal{S}^*$ sera dit **uniquement déchiffrable** si son extension naturelle par *concaténation* $\varphi^* : \mathcal{V}^* \longrightarrow \mathcal{S}^*$ est elle aussi *injective*. On parlera encore de **code préfixe** (ou de **code instantané**) si la condition suivante (condition *plus forte* que la précédente) est satisfaite

$$\forall (v, v') \in \mathcal{V}^2, \quad v \neq v' \Longrightarrow \varphi(v) \text{ n'est pas un préfixe de } \varphi(v')$$

Remarquons que dans l'exemple élémentaire évoqué précédemment, on avait affaire à deux codes préfixes ; de tels codes présentent l'avantage considérable de permettre des décodages "instantanés" (*online decoding*), mot par mot, sans qu'il soit nécessaire de commencer par lire l'intégralité de la séquence de symboles reçue.
Existe-t-il des codes uniquement déchiffrables qui ne sont pas instantanés ? Assurément, comme le montre l'exemple de codage binaire des 4 valeurs a, b, c, d à travers les configurations binaires 0, 01, 011, 0111.
(Plus simplement, le codage binaire des valeurs a/b à travers les configurations 0/01 fournit un exemple on ne peut plus élémentaire de code-source uniquement déchiffrable qui n'est pas préfixe).
Nous verrons toutefois qu'il est toujours possible de ***construire***, à partir d'une source discrète et sans mémoire $\mathcal{W}$, un code-source uniquement déchiffrable à la fois ***compact*** (dépensant *en moyenne* un nombre *minimum* de bits par valeur-source) et ***préfixe*** ! Cette construction de code-source compact et préfixe fait l'objet de l'algorithme de *Huffman*.

2.2 Inégalités classiques et 1er Théorème de Shannon

Venons-en tout d'abord à deux inégalités célèbres, qui vont nous permettre de commencer à discerner ce qui est réalisable et ce qui ne l'est pas.

Lemme : deux inégalités fondamentales
Considérons l'alphabet-source $\mathcal{V} = \{v_1, v_2, \ldots, v_m\}$, ainsi qu'un ensemble fini $\mathcal{S}$ constitué des d symboles $\sigma_1, \sigma_2, \ldots, \sigma_d$. Alors:

1. *Inégalité de Kraft:* il existe un code *préfixe* $\varphi : \mathcal{V} \longrightarrow \mathcal{S}^*$ faisant apparaître des mots (éléments de $\mathcal{S}^*$) de longueurs $|\varphi(v_1)| = l_1, |\varphi(v_2)| = l_2, \ldots, |\varphi(v_m)| = l_m$ si et seulement si

$$\frac{1}{d^{l_1}} + \frac{1}{d^{l_2}} + \ldots + \frac{1}{d^{l_m}} = \sum_{i=1}^{m} d^{-l_i} \leq 1$$

2. *Inégalité de McMillan:* En revanche, dès lors que

$$\frac{1}{d^{l_1}} + \frac{1}{d^{l_2}} + \ldots + \frac{1}{d^{l_m}} = \sum_{i=1}^{m} d^{-l_i} > 1,$$

il n'y a pas de code *uniquement déchiffrable* $\varphi : \mathcal{V} \longrightarrow \Sigma^*$ faisant apparaître des mots de longueurs respectives $l_1, l_2, \ldots, l_m$.

En conséquence, certains codes-source binaires peuvent être immédiatement écartés parce que les mots de code proposés sont trop courts pour que l'on puisse avoir affaire à un code uniquement déchiffrable. Ainsi est-on certain, par exemple, de ne pas pouvoir coder de manière uniquement déchiffrable les 26 caractères de notre alphabet latin en utilisant seulement des configurations binaires de longueur 1, 2, 3 ou 4. En effet, en appelant $l_1, l_2, \ldots, l_{26}$ les longueurs de ces configurations binaires, on aura

$$\sum_{i=1}^{26} \frac{1}{2^{l_i}} \geq \sum_{i=1}^{26} \frac{1}{2^4} = 26 \cdot \frac{1}{16} > 1$$

Remarquons aussi que dans les exemples élémentaires à 4 valeurs ($\mathcal{V} = \{a, b, c, d\}$) proposés au début de cette section, on avait affaire à deux codes préfixes pour lesquels l'inégalité de Kraft est une égalité.
Nous disposons déjà de tous les ingrédients permettant d'énoncer le

1er Théorème de Shannon (*"Source-coding Theorem"*) :

Considérons une source discrète sans mémoire $\mathcal{W} = (X_1, X_2, \ldots, X_k, \ldots)$ utilisant l'alphabet source $\mathcal{V} = \{v_1, v_2, \ldots, v_m\}$; supposons en outre que l'Entropie de $\mathcal{W}$ vaut $H(\mathcal{W}) = h$, et que l'on dispose de $d = 2$ symboles pour coder $\mathcal{W}$ (mots de code *binaires*). Alors

1. La *longueur moyenne* $l(\varphi)$ des mots apparaissant dans un code uniquement déchiffrable $\varphi : \mathcal{V} \longrightarrow \mathcal{S}^*$ vérifie nécessairement
$$l(\varphi) \geq h$$
2. Il existe un code uniquement déchiffrable $\varphi : \mathcal{V} \longrightarrow \mathcal{S}^*$ vérifiant
$$l(\varphi) \leq h + 1$$

En se restreignant toujours au cas particulier où $\mathcal{S} = \{0, 1\}$, un code (binaire) optimal φ a donc une longueur moyenne $l(\varphi)$ vérifiant
$$H(\mathcal{W}) \leq l(\varphi) \leq H(\mathcal{W}) + 1$$

En fait, dans la situation présente (absence de mémoire), en passant à des extensions d'ordre supérieur (concaténations), on obtient
$$H(\mathcal{W}^{(n)}) \leq l(\varphi^{(n)}) \leq H(\mathcal{W}^{(n)}) + 1,$$
et il reste à diviser par n pour voir que
"le nombre moyen de bits utilisés par caractère source peut être rendu arbitrairement proche de l'entropie $H(\mathcal{W})$" !!
Dans tout ce qui précède, on entend par code (binaire) *optimal* un code-source $\varphi : \mathcal{V} \longrightarrow \{0, 1\}^*$ uniquement déchiffrable et de longueur moyenne minimale parmi tous les codes-source binaires uniquement déchiffrables ; on parle alors aussi de codes *compacts*.

Remarquons aussi que ce premier théorème classique de Shannon s'énonce plus généralement dans un contexte où les symboles de codage utilisés ne sont pas nécessairement binaires : si l'ensemble des symboles de codages utilisés est $\mathcal{S} = \{\sigma_1, \sigma_2, \ldots, \sigma_d\}$, on pourra affirmer que

1. La longueur moyenne $l(\varphi)$ des mots apparaissant dans un code uniquement déchiffrable $\varphi : \mathcal{V} \longrightarrow \mathcal{S}^*$ satisfait nécessairement
$$l(\varphi) \geq \frac{h}{\log d}$$
2. Il existe un code uniquement déchiffrable $\varphi : \mathcal{V} \longrightarrow \mathcal{S}^*$ satisfaisant
$$l(\varphi) \leq \frac{h}{\log d} + 1$$

En d'autres termes, il conviendra, dans une telle situation, de calculer l'entropie de la source en utilisant des logarithmes *en base* d ; moyennant cette adaptation, on pourra encore affirmer qu'un code optimal $\varphi : \mathcal{V} \longrightarrow \mathcal{S}^*$ doit avoir une longueur moyenne $l(\varphi)$ vérifiant

$$H_d(\mathcal{W}) \leq l(\varphi) \leq 1 + H_d(\mathcal{W})$$

(l'écriture $H_d(\mathcal{W})$ sert à préciser que l'on a utilisé des logarithmes en base d dans le calcul de l'entropie de la source).

2.3 Construction d'un code-source optimal : l'Algorithme de Huffman

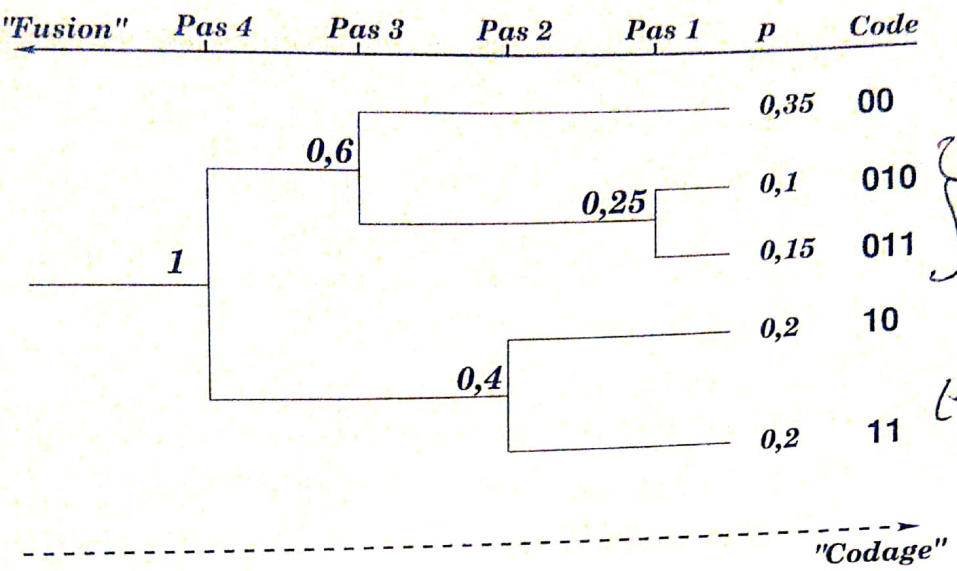
Au vu de l'énoncé du 1er Théorème de Shannon, il est naturel de se poser la question suivante :
"*Comment pourrait-on construire un code-source de longueur moyenne minimale en prenant en compte les différentes fréquences d'apparition de valeurs à la source ?*"
La bonne nouvelle de cette fin de chapitre, c'est que l'*Algorithme de Huffman* permet de construire un code-source *compact* qui a le bon goût d'être *instantané* (i.e. *préfixe*) !
Naturellement, cet algorithme vise à attribuer des mots de code "légers en bits" à ceux des caractères-source apparaissant fréquemment, les caractères-source plus rares pouvant se voir attribuer des mots de code plus lourds. En l'occurence, Huffman parvient à un résultat optimal en utilisant une construction arborescente dont le principe est simple : à chaque étape, il s'agit de regrouper les deux caractères-source les plus rares en additionnant leurs fréquences d'apparition ; ce faisant, on passe à un alphabet-source plus restreint, et l'algorithme touche à sa fin lorsque cet alphabet-source ne contient plus qu'un seul caractère. Dans cette situation finale, on se trouve en fait à la racine d'un arbre dont les feuilles sont justement étiquetées par les caractères de l'alphabet-source d'origine. Reste alors à parcourir cet arbre depuis la racine jusque vers chacune des feuilles pour parvenir à attribuer de manière optimale un mot de code binaire à chacun des caractères-source, en suivant une convention : dans l'exemple ci-dessous, les "montées" correspondent à un '0' et les descentes à un '1'.

$\rightsquigarrow$ Algorithme de Huffman : exemple

dans le cas de 5 caractères $\alpha, \beta, \gamma, \delta, \varepsilon$ apparaissant avec les fréquences respectives $0,35$, $0,1$, $0,15$, $0,2$, $0,2$, quelle pourrait être la *longueur moyenne* des mots (binaires) apparaissant dans un code-source uniquement déchiffrable optimal ??
Le calcul de l'entropie de notre source donne ici

$$\begin{aligned} H(\mathcal{W}) &= -(0,35 \cdot \log 0,35 + 0,1 \cdot \log 0,1 + 0,15 \cdot \log 0,15 + 2 \times 0,2 \cdot \log 0,2) \\ &\approx 2,2 \end{aligned}$$

Il n'y a donc aucun espoir de dépenser moins de $2,2$ bits par valeur-source (en moyenne), mais on peut espérer s'approcher de cette borne inférieure !
Une application de l'algorithme de Huffman dans ce contexte donne alors l'arbre ci-après :



et donc le code-source $\varphi : \mathcal{V} \longrightarrow \{0,1\}^*$ tel que

$$\varphi(\alpha) = 00,\ \varphi(\beta) = 010,\ \varphi(\gamma) = 011,\ \varphi(\delta) = 10,\ \varphi(\varepsilon) = 11$$

Ce code-source a le bon goût d'être instantané, tout en fournissant des mots de code ayant une ***longueur moyenne minimale*** au regard de tous les codes-sources binaires uniquement déchiffrables que l'on pourrait envisager !
Cette longueur moyenne vaut ici

$$\ell(\varphi) = 2 \times 0,35 + 3 \times 0,1 + 3 \times 0,15 + 2 \times 0,2 + 2 \times 0,2 = 2,25\, bits\,,$$

on est donc déjà "tout près" de la borne inférieure infranchissable $H(\mathcal{W}) \approx 2,2$, et il n'y a aucun espoir de dépenser en moyenne moins de *2,25 bits* par valeur-source ... à moins d'avoir recours à un codage de la source par couples de valeurs, ou encore par triplets de valeurs, etc.
Dans le cas présent, les couples de valeurs-source sont au nombre de 25 (de $\alpha\alpha$ à $\varepsilon\varepsilon$), et chacun des couples apparaît avec une probabilité produit en raison de l'indépendance des termes successifs de $\mathcal{W}$
($\alpha\alpha$ apparaît avec prob. $0,35^2 = 0,1225$, $\alpha\beta$ avec prob. $0,35 \cdot 0,1 = 0,035$, ...).
L'algorithme de Huffman appliqué à un "dédoublement" de cette source nous conduirait donc à traiter un arbre binaire comportant 25 feuilles au lieu de 5 seulement ... mais on se trouverait in fine en présence d'un code binaire où, en moyenne, légèrement moins de $4,5$ bits sont dépensés par *couple de valeurs-source*, ce qui nous rapprocherait de la borne infranchissable ($H(\mathcal{W}) \approx 2,2$) et améliorerait un tout petit peu la situation !

2.4 Quelques exercices sur le 1er Thm de Shannon et l'Alg. de Huffman

2.4.1 Un code de Huffman

Une variable aléatoire W à valeurs dans $\mathcal{A} = \{a, b, c, d, e\}$ se comporte de la façon suivante :

$$\mathbb{P}\{W = a\} = \frac{1}{3},\ \mathbb{P}\{W = b\} = \frac{1}{4},\ \mathbb{P}\{W = c\} = \frac{1}{6},\ \mathbb{P}\{W = d\} = \frac{1}{6},\ \mathbb{P}\{W = e\} = \frac{1}{12}$$

1. Evaluer $H(W)$.
2. Construire un code de Huffman pour les valeurs possibles de W, puis comparer la longueur moyenne des mots de ce code binaire avec $H(W)$.
3. Décoder le message 00101100001 en utilisant ce code binaire.

2.4.2 Un entretien d'embauche

Proche de la retraite, M. Nelson travaille depuis de nombreuses années comme présentateur météo sur la chaîne *Bandrika TV*. Au fil de ses nombreuses années de carrière, les prévisions qu'il avait à communiquer ne se sont pas toujours réalisées, et leur adéquation est résumée à travers le tableau de probabilités croisées ci-dessous, où P désigne une variable de *Prévision* tandis que R désigne la variable de *Météo Réelle* correspondante :

$P \setminus R$	'Beau Temps'	'Mauvais Temps'
'Beau Prédit'	5/8	1/16
'Mauvais Prédit'	3/16	1/8

1. Alphonse, jeune étudiant débordant d'ambition, obtient un rendez-vous auprès de Mme Whynot, directrice de l'information à *Bandrika TV*. Il lui tient les propos suivants :
   "Embauchez-moi ! Je saurai chaque soir prédire du beau temps pour le lendemain avec mon sourire le plus exquis, et ce faisant je me tromperai moins souvent que le pauvre M. Nelson."
   Alphonse a-t-il raison de tenir de tels propos ?
2. Mme Whynot rassemble ses esprits et, à la lumière des cours de Théorie de l'Information qu'elle a pu suivre dans sa jeunesse, décide de ne pas engager Alphonse pour remplacer M. Nelson. Pourquoi ??
3. Après avoir éconduit Alphonse, Mme Whynot ordonne d'archiver les valeurs de P puis les valeurs de R correspondant aux 1'000 dernières journées de travail de M. Nelson. Quel est le nombre de bits requis pour un tel archivage ?
4. Un stagiaire intelligent décide d'appliquer l'algorithme de Huffman dans le but de stocker *conjointement* les valeurs de P et de R apparues au cours des 1'000 dernières journées de travail de M. Nelson.
   Quel est alors le nombre de bits requis ?

2.4.3 Codes de Huffman

On considère une source produisant aléatoirement des valeurs au sein de l'alphabet $\mathcal{A} = \{a, b, c, d, e\}$, et les différents codes binaires ci-dessous sont proposés pour les 5 valeurs de la source :

1. 110 , 1110 , 0 , 100 , 1111
2. 111 , 100 , 0 , 101 , 110
3. 10 , 110 , 01 , 111 , 00
4. 10 , 0 , 110 , 111 , 101

A-t-on affaire à des *codes préfixes* ?
Peut-on considérer chacun de ces quatre codes comme des *codes de Huffman* pour la source en question ?
Tenter un décodage du message '11000101000111' en utilisant successivement chacun des codes proposés. Que se passe-t-il dans le dernier cas de figure ?

2.4.4 Applications ludiques du 1er Thm de Shannon

1. Quel est le nombre moyen minimum de questions à poser à Alice pour savoir où celle-ci a placé le roi sur un échiquier ? (L'échiquier comporte naturellement 64 cases, et chaque question posée appelle une réponse binaire : OUI/NON).
   Quelle stratégie précise appliqueriez-vous pour être certain de poser, en moyenne, un nombre minimum de questions ?

2. Quel est le nombre moyen minimum de questions à poser à Bob pour identifier précisément les 4 cartes qui lui ont été distribuées (à partir d'un jeu de 32 cartes) ?

3. Combien de questions seront nécessaires, en moyenne, pour parvenir à déterminer l'ordre dans lequel un croupier a mélangé un paquet de 52 cartes ?

4. Appliquer l'algorithme de Huffman dans le but de définir une suite de questions appelant des réponses binaires *(OUI/NON)* et visant à déterminer la valeur apparue dans un lancer de dé conventionnel.

5. Charlie a lancé deux dés à six faces équilibrés et enregistré la somme des deux valeurs apparues. Il vous est demandé d'identifier cette valeur en posant à Charlie un minimum de questions (appelant une réponse binaire). Quelle stratégie adopteriez-vous ?
   (***Indication*** : ici encore, l'algorithme de Huffman pourra s'avérer utile).

6. 81 pièces d'argent (identiques en apparence) sont placées sur une table, ainsi qu'une balance à plateaux. Il se trouve que l'une de ces pièces est plus lourde que chacune des autres, qui quant à elles sont toutes de même poids. On vous demande d'identifier la pièce singulière en utilisant un nombre minimum de pesées, la balance ne pouvant fournir que l'une des trois réponses : "plateau droit plus lourd", "plateaux équilibrés" ou "plateau gauche plus lourd".
   Quelle stratégie adopteriez-vous ??

2.4.5 Identification de codes de Huffman

On suppose que les mots (binaires) d'un code de Huffman ont pour longueurs respectives $l_1 = 3$, $l_2 = 2$, $l_3 = 4$, $l_4 = 1$, $l_5 = 4$. Utiliser cette seule information pour identifier le code en question.
Quelle méthode pourrait-on mettre en oeuvre pour procéder à de telles identifications dans d'autres cas de figure (où seules les longueurs des mots binaires du code de Huffman sont connues) ?

2.4.6 Algorithme de Huffman et dédoublement de la source

On considère une source discrète et sans mémoire $\mathcal{W}$ susceptible de produire les valeurs α, β, γ avec prob. $0,5$, $0,3$ et $0,2$ respectivement.

1. Calculer l'entropie $H(\mathcal{W})$ de cette source.
2. Appliquer l'algorithme de Huffman dans le but de construire un code-source compact pour $\mathcal{W}$, et évaluer la longueur moyenne des mots de ce code-source. Conclusion ?
3. Appliquer l'algorithme de Huffman à un dédoublement $\mathcal{W}^{(2)}$ de la source $\mathcal{W}$, les 9 valeurs-sources $\alpha\alpha$, $\alpha\beta$, ..., $\gamma\gamma$ de $\mathcal{W}^{(2)}$ apparaissant avec prob. $0,25$, $0,15$, ..., $0,04$.
4. Evaluer la longueur moyenne des mots du code-source obtenu à la question précédente. Conclusion ?
5. Reprendre les expériences évoquées aux questions précédentes dans le cas où les prob. d'apparition de α, β, γ s'élèvent à $0,45$, $0,33$ et $0,22$.

2.4.7 Codes-source binaires et ternaires

Une source discrète et sans mémoire $\mathcal{W}$ fait apparaître les valeurs A, B, C, D, E avec les prob. respectives $0,3$, $0,3$, $0,2$, $0,1$, $0,1$.

1. Construire un code-source binaire optimal pour $\mathcal{W}$ en appliquant l'algorithme de Huffman. Comparer $H(\mathcal{W})$ et la longueur moyenne des mots de ce code-source.
2. Comment adapter l'algorithme de Huffman afin de passer à un codage ternaire des valeurs-source A, B, C, D, E, en utilisant les symboles $-1/0/1$?

2.4.8 Codages d'une source binaire

On considère une source discrète et sans mémoire produisant la valeur 0 avec prob. $0,9$ et la valeur 1 avec prob. $0,1$.

1. On décide de coder le bloc 000000 avec le seul symbole `'0'` et tout autre bloc $uvwxyz$ avec les sept symboles `'1uvwxyz'`. Quel est alors le nombre moyen de bits dépensés par valeur-source ? Comment se compare ce nombre moyen avec l'entropie de la source ?
2. Utiliser l'algorithme de Huffman appliqué à des extensions d'ordre k de cette source ($k = 2, 3, \ldots$), puis calculer à nouveau le nombre moyen de bits dépensés par valeur-source.

3 Les canaux bruités et leurs capacités

3.1 De la notion d'entropie à celle d'information

Disposant maintenant, avec la fonction d'entropie H, de la seule façon raisonnable de mesurer une incertitude, nous pouvons nous poser la question suivante : *"Etant donné deux variables aléatoires X et Y (éventuellement interdépendantes), comment mesurer la quantité d'information au sujet de X fournie par la connaissance de Y ?"*

Pour reprendre nos expériences standard des paragraphes précédents, on pourrait s'imaginer par exemple que X désigne le nombre tiré au sort par un lancer de dé à 20 faces, tandis que Y est une variable de Bernoulli indiquant seulement si ce nombre est premier ou non ($Y = 1$ si le résultat du lancer est un nombre premier, $Y = 0$ sinon). Bien entendu, la connaissance de Y nous renseigne (partiellement) quant à X ; comment quantifier convenablement l'information fournie par Y au sujet de X ?

La réponse naturelle, dans ce contexte, consiste à mesurer cette information en calculant la "quantité d'incertitude sur X" ayant disparu au moment où l'on prend connaissance de la valeur prise par Y. En termes mathématiques, on définit donc $I(X;Y)$, information portée par Y sur X, à travers la formule

$$I(X;Y) = H(X) - H(X|Y)\,,$$

le terme $H(X|Y)$ désignant l'entropie de X *conditionnellement à* Y ; cette entropie *conditionnelle* pourra être calculée comme suit :

si le couple (X,Y) est à valeurs dans $E \times F$, où $E = \{v_1, v_2, \ldots, v_m\}$, $F = \{w_1, w_2, \ldots, w_n\}$, et si

$$\mathbb{P}\{X = v_i, Y = w_j\} = \pi_{i,j} \qquad (1 \leq i \leq m, 1 \leq j \leq n),$$

on définit, pour chaque $1 \leq j \leq n$:

$$H(X|Y = w_j) = -\sum_{i=1}^{m} \mathbb{P}\{X = v_i|Y = w_j\} \cdot \log \mathbb{P}\{X = v_i|Y = w_j\}\,,$$

et l'on a ensuite, en effectuant une moyenne sur les valeurs prises par Y:

$$H(X|Y) = \sum_{j=1}^{n} \mathbb{P}\{Y = w_j\} \cdot H(X|Y = w_j)$$

Ces considérations constituent en quelque sorte un "cheminement naturel" permettant de calculer l'entropie conditionnelle $H(X|Y)$ puis l'information $I(X;Y)$. Fort heureusement, les propriétés du logarithme nous donnent ensuite l'occasion de simplifier les calculs de $H(X|Y)$ et $I(X;Y)$. Il s'avère en effet que

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

où $H(X,Y)$ désigne l'entropie de la variable discrète $Z = (X,Y)$, à valeurs dans $E \times F$!

On aura donc

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - [H(X,Y) - H(Y)] \\ &= H(X) + H(Y) - H(X,Y), \end{aligned}$$

en sorte que

$$I(Y;X) = I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Il y a égalité entre l'information portée par X sur Y et celle portée par Y sur X, ce qui n'était pas si évident au départ.
Avant de passer aux canaux bruités puis aux calculs de capacités, prenons le temps de récapituler certaines des propriétés les plus significatives de H et I.

Quelques propriétés importantes de H et I :

- **(P1):** on a $H(X) \geq 0$, l'égalité $H(X) = 0$ n'ayant lieu que si l'on est certain de la valeur prise par X ($p_1 = 1$).
- **(P2):** pour $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$ variant dans Σ_k, on a $H(\boldsymbol{p}) \leq \log k$, l'égalité ayant lieu seulement pour $(p_1, p_2, \ldots, p_k) = (\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$.
- **(P3):** étant donnés $p, q \in]0;+\infty[$, $(p_1, p_2, \ldots, p_k) \in]0;+\infty[^k$, $(q_1, q_2, \ldots, q_l) \in]0;+\infty[^l$ tels que

$$p = \sum_{i=1}^{k} p_i, q = \sum_{j=1}^{l} q_j, p + q = \sum_{i=1}^{k} p_i + \sum_{j=1}^{l} q_j = 1,$$

on a

$$\begin{aligned} H(p_1, p_2, \ldots, p_k, q_1, q_2, \ldots, q_l) &= H(p,q) + pH(\tfrac{p_1}{p}, \tfrac{p_2}{p}, \ldots, \tfrac{p_k}{p}) \\ &\quad + qH(\tfrac{q_1}{q}, \tfrac{q_2}{q}, \ldots, \tfrac{q_l}{q}) \end{aligned}$$

- **(P4):** pour m et n entiers strictement positifs:

$$H(\frac{1}{mn}, \frac{1}{mn}, \ldots, \frac{1}{mn}) = H(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}) + H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$$

- **(P5):** si X est une v.a. à valeurs dans $E = \{v_1, v_2, \ldots, v_m\}$ et Y une v.a. à valeurs dans $F = \{w_1, w_2, \ldots, w_n\}$, en posant $Z = (X, Y)$ et

$$\mathbb{P}\{X = v_k, Y = w_l\} = \pi_{k,l} \qquad (1 \leq k \leq m, 1 \leq l \leq n),$$

on a

$$H(Z) = H(X,Y) = -\sum_{k,l} \pi_{k,l} \log \pi_{k,l} \leq H(X) + H(Y),$$

l'égalité ayant lieu si et seulement si X et Y sont indépendantes !

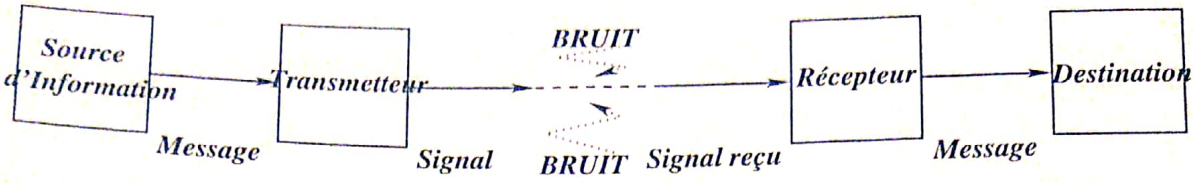
- **(P6):** dans le même contexte, on pourra tout aussi bien remarquer que

$$H(X|Y) \leq H(X),$$

avec égalité seulement dans le cas où X et Y sont indépendantes.
(Intuitivement : l'incertitude sur X diminue dès lors que l'on a pris connaissance de la valeur prise par Y, sauf si X et Y sont indépendantes, auquel cas cette incertitude est inchangée).

3.2 Quelques modèles de canaux bruités

Revenons au "2ème maillon" du schéma de communication de Shannon, où une suite de symboles (e.g. des bits) est transmise sur un canal bruité.



Comment convient-il de modéliser le bruit présent sur le canal??

Définition:

Restreignons-nous aux canaux bruités *discrets* et *sans mémoire*, pour lesquels :

1. Le message à transmettre est une suite de symboles pris dans un ensemble *fini* $\mathcal{S}$.
2. Chacun des symboles de cette suite est transmis fidèlement ou encore altéré, et cela de manière *i.i.d.* pour toute la suite de symboles à transmettre.

Plus précisément, si $\mathcal{S} = \{\sigma_1, \ldots, \sigma_d\}$ est l'ensemble de symboles original et $\mathcal{S}' = \{\sigma'_1, \ldots, \sigma'_{d'}\}$ un nouvel ensemble de symboles, le traitement du message est modélisé au moyen d'une matrice $P = (p_{i,j})_{1 \leq i \leq d, 1 \leq j \leq d'}$ de taille $d \times d'$, en sorte que

$$\mathbb{P}\left\{Symbole\ reçu = \sigma'_j | Symbole\ émis = \sigma_i\right\} = p_{i,j},$$

pour chaque terme de la suite finie de symboles à transmettre.

Remarque et définition: d'après la définition même de la matrice $P = (p_{i,j})_{1 \leq i \leq d, 1 \leq j \leq d'}$, on a

$$\forall (i,j) \in \{1, \ldots d\} \times \{1, \ldots d'\}, \quad p_{i,j} \geq 0$$

et

$$\forall i \in \{1, \ldots d\}, \quad \sum_{j=1}^{d'} p_{i,j} = 1$$

(Les coeff. de P sont positifs ou nuls, et la somme des coeff. de chaque ligne vaut 1).

Lorsque $d' = d$ (matrice carrée), on dit que P est une *matrice stochastique* (cf étude des *chaînes de Markov*).

Quelques exemples célèbres de canaux bruités:

1. *Canal binaire symétrique:*
   ici $d = d' = 2$ et $\Sigma = \Sigma' = \{0; 1\}$, chaque symbole binaire à transmettre est soit transmis fidèlement (avec probabilité p), soit transformé en l'autre symbole (avec probabilité $q = 1 - p$), en sorte que

$$P = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$$

2. *Canal binaire asymétrique:*
on introduit une asymétrie dans le modèle précédent en changeant la matrice P en

$$\tilde{P} = \begin{pmatrix} p & q \\ q' & p' \end{pmatrix},$$

où $p' \neq p$ (et $q' = 1 - p'$).

3. *Canal d'effacement binaire:*
ici $d = 2$ et $\Sigma = \{0; 1\}$, mais $d' = 3$ et $\Sigma' = \{0; 1; \iota\}$. Chaque symbole binaire à transmettre est soit transmis fidèlement (avec probabilité $1 - \varepsilon$), soit transformé en un symbole illisible ι (avec probabilité ε), en sorte que

$$P = \begin{pmatrix} 1-\varepsilon & 0 & \varepsilon \\ 0 & 1-\varepsilon & \varepsilon \end{pmatrix}$$

Comme précédemment, on pourra considérer des extensions d'ordre n où l'on passe de Σ à Σ^n, de Σ' à Σ'^n et de P à $P^{(n)}$, matrice de taille $d^n \times d'^n$.
Par exemple, dans l'extension d'ordre 3 du dernier canal ci-dessus, 000 est transformé en 000 avec prob. $(1-\varepsilon)^3$, en 001 avec prob. $\varepsilon(1-\varepsilon)^2$, en 101 avec prob. $\varepsilon^2(1-\varepsilon)$...

N.B.: dans le 1er modèle (binaire symétrique), la transmission est satisfaisante si le paramètre p est proche de 1 ou encore proche de 0...
Par contre si $p = \frac{1}{2}$ il n'y a plus rien à faire, le message est entièrement perdu!

⇝Deux questions fondamentales :

1. Comment mesurer la *qualité de transmission* sur un canal bruité donné ?
2. Comment *remédier au bruit* pour retrouver le signal original ??

⇝Quelques suggestions relatives à la deuxième question :

- Découper la suite finie de $N = k \cdot n$ symboles à transmettre en k "tranches" ayant chacune n symboles.
- Transformer (allonger) chaque tranche de n symboles en une nouvelle tranche à n' symboles ($n' > n$), dans le but d'introduire une certaine fiabilité dans les transmissions.

Exemples tout simples :
considérons un canal binaire symétrique et fixons $n = 3$.

1°) Première possibilité: on choisit de répéter chaque tranche à transmettre, en sorte que

$$000 \mapsto 000000, \quad 001 \mapsto 001001, \ldots 011 \mapsto 011011, \quad 111 \mapsto 111111$$

Les tranches de 3 bits sont transformées en tranches de 6 bits, à l'arrivée les erreurs de transmission éventuelles sont *parfois* détectées.

2°) Variante: on répète deux fois ou plus chaque tranche à transmettre, e.g.

$$000 \mapsto 000000000, \quad 001 \mapsto 001001001, \ldots \quad 111 \mapsto 111111111$$

Les tranches de 3 bits sont transformées en tranches de $3(m+1)$ bits, où m désigne le nombre de répétitions.

3°) Autre possibilité: on choisit de rajouter *un seul bit* à la fin de chaque tranche à transmettre, ce dernier bit valant la somme (binaire) des 3 bits de la tranche:

$$000 \mapsto 0000, \quad 001 \mapsto 0011, \ldots 011 \mapsto 0110, \quad 111 \mapsto 1111$$

Les tranches de 3 bits sont transformées en tranches de 4 bits, à l'arrivée les erreurs de transmission éventuelles sont *parfois* détectées.

Avantage de la dernière méthode: on est beaucoup plus "économe en bits"!
Avantage de la méthode à m répétitions: plus de fiabilité ...
D'où les questions fondamentales ci-dessous :
1°) *Comment concevoir un code correcteur d'erreurs qui permette d'atteindre "un certain niveau de fiabilité" tout en ne prenant "pas trop de place" ?*
2°) *Quel est l'allongement minimum des messages qui est rendu nécessaire par l'exigence d'une fiabilité élevée ?*
3°) *Comment concevoir de tels codes correcteurs à fiabilité élevée ?*
Le 2nd Théorème de Shannon fournit une réponse précise à la 2ème de ces questions en mettant en avant la notion de *Capacité* d'un canal bruité.

Définition:
on considère encore un canal de transmission bruitée discret et sans mémoire donné par une matrice $P = (p_{i,j})_{1\le i\le d, 1\le j\le d'}$, les ensembles de symboles entrants et sortants étant $\mathcal{S} = \{\sigma_1, \ldots, \sigma_d\}$ et $\mathcal{S}' = \{\sigma'_1, \ldots, \sigma'_{d'}\}$.

- Si X est une v.a. à valeurs dans $\mathcal{S}$, désignons par $\Phi(X)$ la v.a. à valeurs dans $\mathcal{S}'$ obtenue en appliquant à X la transformation *stochastique* donnée à travers P, en sorte que

$$\mathbb{P}\left\{\Phi(X) = \sigma'_j | X = \sigma_i\right\} = p_{i,j} \quad (1 \le i \le d, 1 \le j \le d')$$

- La capacité C du canal de transmission considéré est alors définie par

$$C = \sup_X I(X, \Phi(X)) = \max_X I(X, \Phi(X))$$

Quelques remarques:
Souvenons-nous de ce que

$$I(X, \Phi(X)) = H(X) - H(X|\Phi(X)) = H(X) + H(\Phi(X)) - H(X, \Phi(X)) \ge 0$$

La capacité C du canal considéré est donc un nombre réel ≥ 0, et $I(X, \Phi(X))$ dépend de X seulement à travers sa loi, donnée par un vecteur de probabilité $(p_1, p_2, \ldots, p_d)$ t.q.

$$\mathbb{P}\{X = \sigma_i\} = p_i, \quad i = 1, 2, \ldots, d$$

L'identité

$$\sup_X I(X, \Phi(X)) = \max_X I(X, \Phi(X))$$

signifie qu'il y a un (ou plusieurs) vecteur(s) $(p_1, p_2, \ldots, p_d)$ pour le(s)quel(s) le sup est atteint.

Exemples de calculs de capacités :

1°) **Capacité d'un canal binaire symétrique :**

dans le cas d'un canal binaire symétrique où

$$P = \begin{pmatrix} (1-\varepsilon) & \varepsilon \\ \varepsilon & (1-\varepsilon) \end{pmatrix}$$

on vérifie que

$$C = C(\varepsilon) = \max_X I(X, \Phi(X)) = 1 + \varepsilon \log \varepsilon + (1-\varepsilon) \log(1-\varepsilon)$$

Remarquons qu'une telle expression de la capacité C en fonction de ε est tout à fait convaincante intuitivement : dans le cas où $\varepsilon = \frac{1}{2}$ (brouillage complet !), on obtient $C = 0$, tandis que pour les cas extrêmes ($\varepsilon = 0$ ou $\varepsilon = 1$), C vaut 1, ce qui correspond à la possibilité de transmettre à la perfection (sans perte d'information). Dans tous les autres cas, C est située dans l'intervalle $]0; 1[$.

Esquisse de preuve : en supposant que la source émette un $'1'$ avec prob. p et un $'0'$ avec prob. $q = (1-p)$, on obtiendra une variable de sortie $Y = \Phi(X)$ prenant la valeur $'1'$ avec prob. $p(1-\varepsilon) + q\varepsilon$ et la valeur $'0'$ avec prob. $p\varepsilon + q(1-\varepsilon)$.
Il en résulte que

$$I(X,Y) = \varepsilon \log \varepsilon + (1-\varepsilon) \log(1-\varepsilon) - [p(1-\varepsilon) + q\varepsilon] \log [p(1-\varepsilon) + q\varepsilon] - [p\varepsilon + q(1-\varepsilon)] \log [p\varepsilon + q(1-\varepsilon)]$$

et une maximisation en p (à ε fixé) fournit alors le résultat annoncé.

2°) **Capacité d'un canal d'effacement binaire :**

dans le cas d'un canal d'effacement binaire où

$$P = \begin{pmatrix} 1-\varepsilon & 0 & \varepsilon \\ 0 & 1-\varepsilon & \varepsilon \end{pmatrix}$$

on vérifie que

$$C = C(\varepsilon) = \max_X I(X, \Phi(X)) = 1 - \varepsilon$$

Remarquons qu'ici encore, nous obtenons une expression de la capacité C en fonction de ε qui est tout à fait conforme à l'intuition : en l'absence d'effacement ($\varepsilon = 0$) cette capacité est pleine ($C = 1$), à mesure que le paramètre ε grandit (de plus en plus d'effacement) la capacité C diminue, jusqu'à atteindre la valeur 0 pour $\varepsilon = 1$ (effacement complet).

3°) **Capacité d'un canal de "brouillage ternaire" :**

dans le cas d'un canal ternaire pour lequel

$$P = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

on vérifie que

$$C = \max_X I(X, \Phi(X)) = 0$$

3.3 Codage sur un canal bruité : le second Théorème de Shannon

Dans cette dernière section, par commodité, on se restreindra pour l'essentiel aux canaux binaires symétriques.
Comme nous l'avons déjà vu, pour transmettre une suite de bits de manière fiable sur un canal bruité, on pourra être amené à découper la suite donnée en tranches de M bits, puis à transformer chacune de ces tranches de M bits une nouvelle tranche (plus longue) de N bits avant d'effectuer la transmission.
Par exemple, si la suite originale est découpée par tranches de 3 bits, et si chacune de ces tranches est transformée en une tranche de 5 bits, il va s'agir de choisir 8 tranches de 5 bits parmi les 32 tranches de 5 bits à disposition.

Définitions :

- Pour chaque entier $N \geq 1$, on note $\mathcal{U}_N$ l'ensemble des suites finies de N bits: $\mathcal{U}_N = (\mathbb{Z}/2\mathbb{Z})^N$.
- Un **code-bloc** de longueur N est un sous-ensemble $\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$ de $\mathcal{U}_N$. Le **taux** du code-bloc $\mathcal{C}$ vaut alors $\tau(\mathcal{C}) = \frac{\log|\mathcal{C}|}{N} = \frac{\log K}{N}$ (où les logarithmes sont encore et toujours pris en base 2).
- On appelle alors **règle de décodage** pour le code-bloc $\mathcal{C}$ toute partition $\mathcal{R} = \{R_1, R_2, \ldots, R_K\}$ de $\mathcal{U}_N$ en sous-ensembles disjoints $R_1, R_2, \ldots, R_K$ tels que

$$\boldsymbol{c}_1 \in R_1, \quad \boldsymbol{c}_2 \in R_2, \quad \ldots, \quad \boldsymbol{c}_K \in R_K.$$

Une première idée intuitive gouvernant ces choix de codes-blocs et de règles de décodage sera de choisir des mots de code $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K$ bien "éloignés les uns des autres" (au sens de la distance de Hamming), et de faire en sorte que chacun de ces mots de code $\boldsymbol{c}_i$ soit situé "bien au chaud à l'intérieur de la région R_i" (et non à proximité de la frontière de cette région avec une autre région R_j).

Exemples de règles de décodage :

- **Règle de l'observateur idéal:** le bloc reçu $\boldsymbol{d} = (d_1, \ldots, d_N)$ est décodé comme $\boldsymbol{c}_i$ si et ssi

$$\mathbb{P}\{\boldsymbol{c}_i \text{ envoyé} \,|\, \boldsymbol{d} \text{ reçu}\} = \max_{1 \leq j \leq K} \mathbb{P}\{\boldsymbol{c}_j \text{ envoyé} \,|\, \boldsymbol{d} \text{ reçu}\}$$

  Autrement dit: $\boldsymbol{d} \in R_i$ si et ssi

$$\forall j \neq i, \quad \mathbb{P}\{\boldsymbol{c}_j \text{ envoyé} \,|\, \boldsymbol{d} \text{ reçu}\} \leq \mathbb{P}\{\boldsymbol{c}_i \text{ envoyé} \,|\, \boldsymbol{d} \text{ reçu}\}$$

- **Règle du maximum de vraisemblance:** $\boldsymbol{d} = (d_1, \ldots, d_N)$ est décodé comme $\boldsymbol{c}_i$ si et ssi la probabilité conditionnelle $\mathbb{P}\{\boldsymbol{d} \text{ reçu} \,|\, \boldsymbol{c}_j \text{ envoyé}\}$ est maximale pour $j = i$.
- **Règle de la distance minimum:** $\boldsymbol{d} = (d_1, \ldots, d_N)$ est décodé comme $\boldsymbol{c}_i$ si et ssi la *distance de Hamming* $\Delta_H(\boldsymbol{d}, \boldsymbol{c}_j)$ est minimale pour $j = i$.

Illustration sur un canal bin. sym. avec $K = 8$, $N = 5$:
Supposons que

$$\mathcal{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_8\} = \{00000, 00101, 01010, 01111, 10000, 10101, 11010, 11111\}$$

et que le bloc $\boldsymbol{d} = 10010$ est reçu après transmission.
L'application des règles du maximum de vraisemblance et de la distance minimum donnent alors le même résultat: selon ces deux règles, $\boldsymbol{d} = 10010$ est décodé en $\boldsymbol{c}_7 = 11010$, ou encore en $\boldsymbol{c}_5 = 10000$.
De manière générale, dans le cas d'un canal binaire symétrique, les règles du maximum de vraisemblance et de la distance minimum fournissent le même décodage.
En revanche, sans une connaissance des probabilités d'apparition de chaque bloc $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_8$ à l'entrée du canal bruité, on n'est pas en mesure de fournir un décodage selon la règle de l'observateur idéal... Dans le cas d'*apparitions équiprobables* pour $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_8$, les règles de l'observateur idéal et du maximum de vraisemblance coïncident.

Définitions (suite):

- Etant donné un code-bloc $\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K\}$ de longueur N et une règle de décodage $\mathcal{R}$ pour $\mathcal{C}$, la **probabilité moyenne d'erreur** est donnée par

$$e(\mathcal{C}, \mathcal{R}) = \frac{1}{K} \sum_{j=1}^{K} \mathbb{P}\{\text{erreur de décodage} \,|\, \boldsymbol{c}_j \text{ envoyé}\}$$

  Cette définition n'est pas des plus satisfaisantes lorsque les blocs $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K$ n'apparaissent pas avec les mêmes fréquences.

- On utilisera donc plutôt la **probabilité maximale d'erreur**, donnée par

$$e_{max}(\mathcal{C}, \mathcal{R}) = \max_{1 \le j \le K} \mathbb{P}\{\text{erreur de décodage} \,|\, \boldsymbol{c}_j \text{ envoyé}\}$$

Nous disposons maintenant de tous les ingrédients permettant d'énoncer le 2nd Théorème de Shannon dans le contexte des canaux binaires symétriques.

Théorème (2nd Thm de Shannon, ou *"Noisy Coding Thm"*):
Considérons un canal binaire symétrique de capacité $C > 0$, et fixons $0 < \rho < C$. Il existe alors une suite $\{(\mathcal{C}_N, \mathcal{R}_N)\,; N \geq N_0\}$ de codes-bloc et de règles de décodage telle que:

1. Chaque code-bloc $\mathcal{C}_N$ est de longueur N et a un taux $\tau(\mathcal{C}_N)$ satisfaisant $\tau(\mathcal{C}_N) \leq \rho$.
2. Les probabilités maximales d'erreur sont telles que

$$\lim_{N \to +\infty} e_{max}(\mathcal{C}_N, \mathcal{R}_N) = 0.$$

Autrement dit:
si l'on accepte de transmettre à un taux situé sous la capacité du canal, on pourra atteindre une fiabilité arbitrairement élevée!

Utilisation du 2nd Thm de Shannon sur un exemple:
Supposons que l'on cherche à transmettre de manière fiable une longue suite de bits sur un canal de capacité $C = 0,77$.
Il convient alors de procéder comme suit:

1°) Découper la suite originale par tranches de M bits, où M est "raisonnablement grand".

2°) Attribuer à chaque tranche de M bits un bloc de N bits (mot de code), où $N = \lceil \frac{4}{3} M \rceil$, et spécifier une règle de décodage $\mathcal{R}_N$ pour le code correspondant, en sorte que $e_{max}(\mathcal{C}_N, \mathcal{R}_N)$ soit "petit".

3°) Envoyer des mots de code de longueur N sur le canal de transmission bruité.

Dès la parution de l'article fondateur de Shannon en 1948, chacun aura pu remarquer que la preuve de ce second Théorème requérait nettement plus de travail que celle du 1er Théorème. En outre, la preuve originale fournie par Shannon pour ce 2nd Thm était manifestement *non-constructive*, et cette fois il ne s'est trouvé personne (pas même Huffman) pour proposer dans les années suivantes un algorithme permettant de réaliser les prédictions du Thm en matière de codes correcteurs ... En fait, il aura fallu près d'un demi-siècle d'efforts de recherche pour que soient découverts les premiers codes correcteurs ayant un taux proche de la borne de Shannon (turbo-codes, présentés pour la 1ère fois par Claude Berrou et Alain Glavieux lors de l'*International Conference on Communications* de Genève, été 1993).

Complément au 2nd Thm de Shannon:
Réciproquement, si $\{(\mathcal{C}_N, \mathcal{R}_N)\,; N \geq N_0\}$ est une suite de codes-bloc ayant des taux $\tau(\mathcal{C}_N)$ satisfaisant $\tau(\mathcal{C}_N) \geq \rho'$ pour un ρ' fixé dans l'intervalle $]C; +\infty[$, on a nécessairement

$$e_{max}(\mathcal{C}_N, \mathcal{R}_N) \underset{N \to +\infty}{\longrightarrow} 1.$$

Autrement dit:
avec des taux situés au-dessus de la capacité du canal, il n'y a plus de transmission fiable qui puisse être envisagée !

3.4 Un exemple de code correcteur

Au sein des codes correcteurs, les codes *linéaires*, dont les mots de code forment un sous-espace vectoriel de $(\mathbb{Z}/2\mathbb{Z})^N$, ont rapidement joué un rôle important. Voici donc un exemple de code linéaire (code de Hamming), traité sous forme d'exercice : on considère le code-bloc de longueur 7 défini comme étant le sous-ensemble $\mathcal{C}$ de $(\mathbb{Z}/2\mathbb{Z})^7$ constitué de tous les vecteurs binaires $\boldsymbol{u} = (u_1, u_2, \ldots, u_7)$ satisfaisant la condition

$$H \times {}^t\boldsymbol{u} = H \times \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

où H est la matrice à coefficients dans $\mathbb{Z}/2\mathbb{Z}$ donnée par

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

1. Trouver quelques mots de code $\boldsymbol{u}$ figurant dans $\mathcal{C}$.

2. Trouver le rang de la matrice binaire H. En déduire qu'il y a en tout 32 mots de code dans $\mathcal{C}$. Quel est le taux de transmission correspondant?

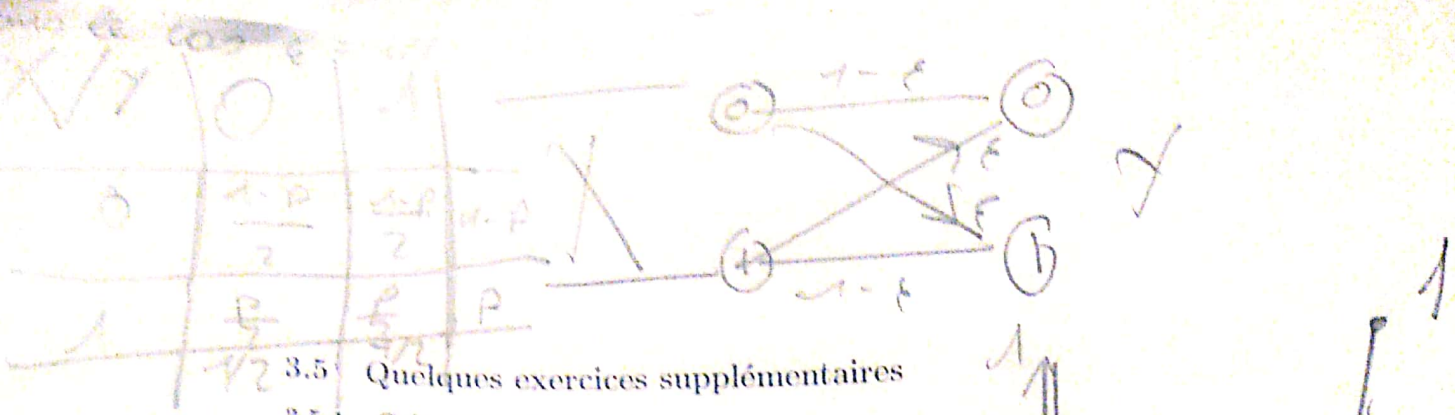
3. Montrer que la distance de Hamming séparant deux mots de codes distincts vaut au moins 3 :

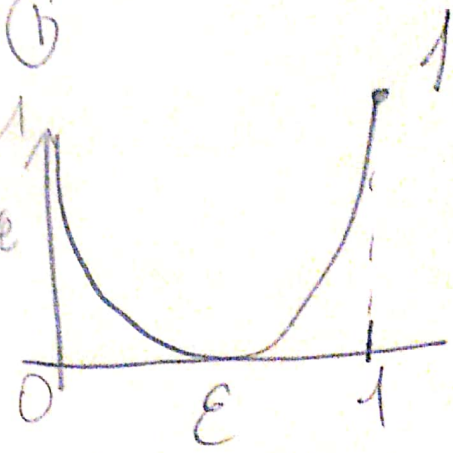
$$\forall (\boldsymbol{u}, \boldsymbol{u}') \in \mathcal{C}^2, \quad \boldsymbol{u} \neq \boldsymbol{u}' \Longrightarrow \Delta_H(\boldsymbol{u}, \boldsymbol{u}') \geq 3$$

4. Montrer que si une erreur au plus est commise lors de la transmission d'un mot de code sur un canal binaire bruité, le mot reçu est plus proche du mot de code envoyé que de tout autre mot de code.

5. On suppose que le mot de code $\boldsymbol{u} \in \mathcal{C}$ est envoyé sur ce canal binaire, et qu'une erreur exactement est commise lors de la transmission, en sorte que l'on reçoit le mot $\boldsymbol{r} = \boldsymbol{u} + \boldsymbol{e}_i$, le terme d'erreur $\boldsymbol{e}_i$ comportant un '1' à l'emplacement i et des '0' aux autres emplacements.
Montrer que l'on parvient à retrouver l'emplacement de cette erreur en effectuant le produit $H \times {}^t\boldsymbol{r}$, en sorte que $\mathcal{C} \subset (\mathbb{Z}/2\mathbb{Z})^7$ fournit un code-bloc permettant de corriger les erreurs isolées.

6. Trouver un minorant intéressant de la **probabilité maximale d'erreur** dans le contexte d'une transmission de tels blocs de longueur 7 sur un Canal Binaire Symétrique de paramètre d'erreur $\varepsilon = 0,1$.





3.5 Quelques exercices supplémentaires

3.5.1 Calculs de Capacités :

Calculer les Capacités des canaux bruités suivants :

1. Le *Canal Binaire Symétrique*, donné par la matrice

$$P = \begin{pmatrix} (1-\varepsilon) & \varepsilon \\ \varepsilon & (1-\varepsilon) \end{pmatrix}$$

2. Le *Canal Binaire Asymétrique*, donné par la matrice

$$P = \begin{pmatrix} (1-\varepsilon) & \varepsilon \\ \varepsilon' & (1-\varepsilon') \end{pmatrix},$$

où $\varepsilon' \neq \varepsilon$.

3. Le Canal Binaire Symétrique à Effacement, donné par la matrice

$$P = \begin{pmatrix} (1-\alpha-\beta) & \alpha & \beta \\ \alpha & (1-\alpha-\beta) & \beta \end{pmatrix}$$

4. Le Canal (sans mémoire) donné par la matrice

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

5. Le Canal Ternaire donné par la matrice

$$P = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

6. Le Canal Ternaire donné par la matrice

$$P = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

7. Le Canal Ternaire donné par la matrice

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

8. Le Canal Quaternaire donné par la matrice

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{pmatrix}$$

9. Le Canal Quaternaire donné par la matrice

$$P = \begin{pmatrix} (1-\varepsilon) & \varepsilon & 0 & 0 \\ \varepsilon & (1-\varepsilon) & 0 & 0 \\ 0 & 0 & (1-\varepsilon) & \varepsilon \\ 0 & 0 & \varepsilon & (1-\varepsilon) \end{pmatrix}$$

3.5.2 Règles de Décodage :

Soit $\mathcal{C} = \{c_1, c_2, \ldots, c_K\} \subset \mathcal{S}_N$ un code-bloc de longueur N.

1. Si les mots de code $c_1, c_2, \ldots, c_K$ apparaissent avec prob. $\frac{1}{K}$, vérifier que les règles de l'observateur idéal et du maximum de vraisemblance sont identiques.

2. Si ces mots de code sont transmis sur un canal binaire symétrique, vérifier que les règles du maximum de vraisemblance et de la distance minimum sont identiques.

3.5.3 Décodage sur un Canal Binaire Symétrique :

On considère un Canal Binaire Symétrique de paramètre d'erreur $q = \varepsilon$.

1. Dans le cas d'un code-bloc de longueur 3 donné par $c_1 = 000$, $c_2 = 111$, prouver que pour chaque mot transmis la prob. d'une erreur de décodage vaut $3\varepsilon^2 - 2\varepsilon^3$ (où la règle du max. de vraisemblance est appliquée).

2. Dans le cas d'un code-bloc de longueur 5 dont les mots de code sont tous les mots binaires de longueur 5 comportant exactement deux "1", quelle est la probabilité de décoder "10001" alors que l'on a envoyé "11000" (en employant toujours le max. de vraisemblance)?

3.5.4 Règles de Décodage (suite) :

On considère le code-bloc $\mathcal{C} = \{c_1, c_2, c_3, c_4\} \subset \mathcal{S}_4$ donné par

$$c_1 = 1000,\ c_2 = 0110,\ c_3 = 0001,\ c_4 = 1111,$$

les prob. d'apparition de ces mots de code étant

$$\mathbb{P}(c_1) = \mathbb{P}(c_2) = \frac{1}{3},\ \mathbb{P}(c_3) = \mathbb{P}(c_4) = \frac{1}{6}$$

Ces mots de code sont transmis sur un canal binaire symétrique de paramètre d'erreur $q = 1 - p = 0,1$.
Le mot 1001 est reçu; comment sera-t-il décodé si l'on emploie

1. la règle de l'observateur idéal?

2. la règle du maximum de vraisemblance?

3.5.5 Utilisation du 2nd Thm de Shannon :

1. Un Canal Binaire Symétrique de paramètre d'erreur $q = 0,05$ parvient à transmettre 800 bits/seconde au Récepteur. Combien de bits/secode peut-il transmettre *de manière fiable*?

2. Un Canal Binaire Symétrique peut transmettre (physiquement) jusqu'à 800 bits/seconde au Récepteur; on sait par ailleurs que ce même Canal peut transmettre 500 bits/seconde de manière fiable.
Que peut-on en conclure?

3.5.6 Utilisation du 2nd Thm de Shannon (suite) :

Une source discrète et sans mémoire d'entropie 15 bits/mot-source est connectée à un Canal Binaire Symétrique de paramètre d'erreur $q = 0,1$ qui peut transmettre (physiquement) jusqu'à 10^3 bits/seconde au Récepteur.
Si la transmission doit être effectuée de manière fiable, combien de mots-source peuvent être émis par seconde?

3.5.7 Capacité de deux canaux montés en série :

On considère un premier canal transformant les symboles d'entrée $A/B/C$ en symboles de sortie D/E suivant la matrice $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ p & (1-p) \end{pmatrix}$, puis un second canal transformant les symboles d'entrée D/E en symboles de sortie $A/B/C$ suivant la matrice $P' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & q & (1-q) \end{pmatrix}$.

3.5.8 Capacité de canaux binaires symétriques montés en série :

On considère n canaux binaires symétriques (CBS) fonctionnant de manière indépendante avec un même paramètre d'erreur par symbole ε.
Ces canaux sont montés en série, au sens où un symbole X_0 entrant dans le premier CBS est transformé en X_1 à sa sortie, X_1 entre dans le deuxième canal et est transformé en X_2 avant d'entrer dans le second canal, ..., X_{n-1} sort de l'avant-dernier canal et est transformé en X_n à l'issue de son passage par le dernier canal.
Quelle est alors la capacité du canal correspondant à un tel montage en série ?
(Rappel : il s'agit dévaluer $\max_{X_0} I(X_0; X_n)$).
Que se passe-t-il pour $n \to +\infty$?

3.5.9 Capacité de deux canaux montés en parallèle :

On considère deux canaux bruités donnés respectivement par les matrices $P = (p_{i,j})$ et $Q = (q_{k,l})$ t.q.

$$\forall i \in [\![1; d_1]\!], \forall j \in [\![1; d'_1]\!], \quad \mathbb{P}\{X' = u'_j | X = u_i\} = p_{i,j}$$

et

$$\forall k \in [\![1; d_2]\!], \forall l \in [\![1; d'_2]\!], \quad \mathbb{P}\{Y' = v'_l | y = v_k\} = q_{k,l}$$

On suppose que le premier canal a pour capacité C_1 et que le second a pour capacité C_2.
Quelle est alors la capacité du canal correspondant à un montage en parallèle de ces deux canaux ?

Semi 2 — Modélisation Maths

X : $v_1 : p^1$, $v_2 : p^2$, …, $v_m : p^m$

$H(X) = H(p^1, \dots, p^m) = -\sum_{i=1}^{m} p^i \log_2(p^i)$

Exercice 1.2.1

Si $m = 2$, $p^2 = 1 - p^1$

$H(p, 1-p) = f(p)$

$f(p) = -x \log_2(x) - (x-1)\log_2(x-1)$

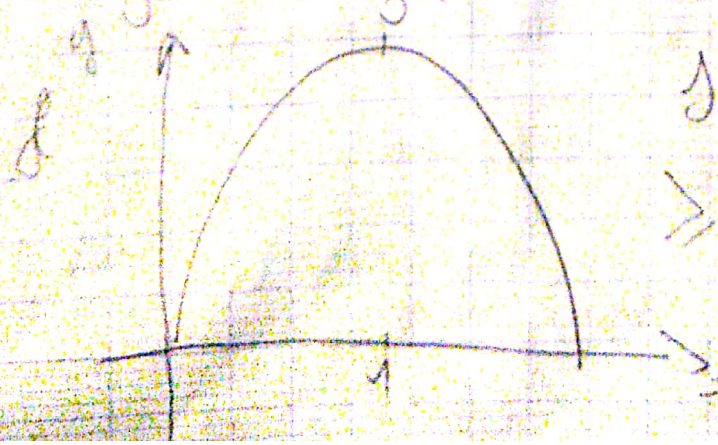
$f\left(\frac{1}{2}\right) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$

$= \frac{1}{2}\log(2) + \frac{1}{2}\log(2)$

$= \log(2)$

$= 1$

$f(1-x) = H(1-x, x) = (H \log(1-x))$



$f(\lambda x_1 + (1-\lambda)x_2)$

$\geq f(x_1) + (1-\lambda) f(x_2)$

$\geq \frac{1}{2}(H(a_1, 1-a_1) + H$

$\geq H(x_1 - x_2)$

Cas où $n = 3$

$$H(p_1, p_2, p_3) = H(x, y, 1-x-y) = g(x, y)$$

$$\max_{p=(p_1,p_2,\dots,p_n)} H(p_1, p_2, \dots, p_n) = H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) = \log(n)$$

Exercice 1.3.2

Mathématiquement

Discrets : $X = (X_1, X_2, \dots, X_{1000})$

Image : $Y = (Y_1, Y_2, \dots, Y_N)$ où

$N = 600 \times 600 = 3 \times 10^5$

1er Calcul :

$X \subset \cup(\llbracket 1 ; k \rrbracket)$ où $k = 10000^{1000}$

$Y \subset \cup(\llbracket 1, L \rrbracket)$ où $L = 16^{(3\times10^5)}$

$$H(X) = \log_2(k) = 1000 \log(10^4) = 4000 \log(10) = 4000 \log(2\times5) = 4000 \times [\log(2) + \log(5)]$$

$$= 4000 \times (1 + \log(5))$$

Sem 2 — Mod Maths — TD 2

$\simeq 13\,000$

$H(Y) = \log_2(L) = \log(16^{3\times10^5})$

$= 3\times10^5 \log(16)$

$= 12\times10^5 \log(2)$

$= 12\times10^5$

$H(X) \ll H(Y)$

2ème calcul

Qd 2 variables sont indépendantes, et ont 0 de corrélation mais ça marche que c'est quantitatif et en plus la réciproque est fausse

Mais avec l'entropie on peut dire:

$X_1, X_2, \dots X_{1000}$ sont indépendantes

$H(X_1, X_2, \dots, X_{1000}) = H(X_1) + \dots H(X_{1000})$

$= \sum_{i=1}^{1000} H(X_i)$

$= 1000 \times H(X_1)$

$= 1000 \times \log_2(16)$

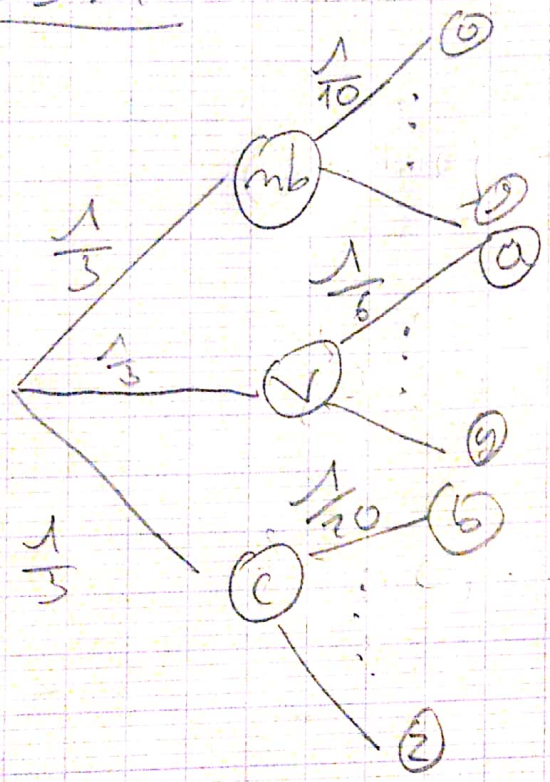
de même:

$H(Y) = NH(Y_1)$

$= 3\cdot10^5 \log(16)$

L'entropie se comporte additivement

1.3.1 h



$$H(X) = H\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) + \frac{1}{3} H\left(\frac{1}{10}, \frac{1}{10}, \dots, \frac{1}{10}\right) \quad (10\times)$$

$$+ \frac{1}{3} H\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6} \dots \frac{1}{6}\right) \quad (6\times)$$

$$+ \frac{1}{3} H\left(\frac{1}{20}, \frac{1}{20}, \dots, \frac{1}{20}\right) \quad (20\times)$$

$$= \log(3) + \frac{1}{3}(1+\log(5)) + \frac{1}{3}(1+\log 3) + \frac{1}{3}(1+\log(5))$$

≈ 5

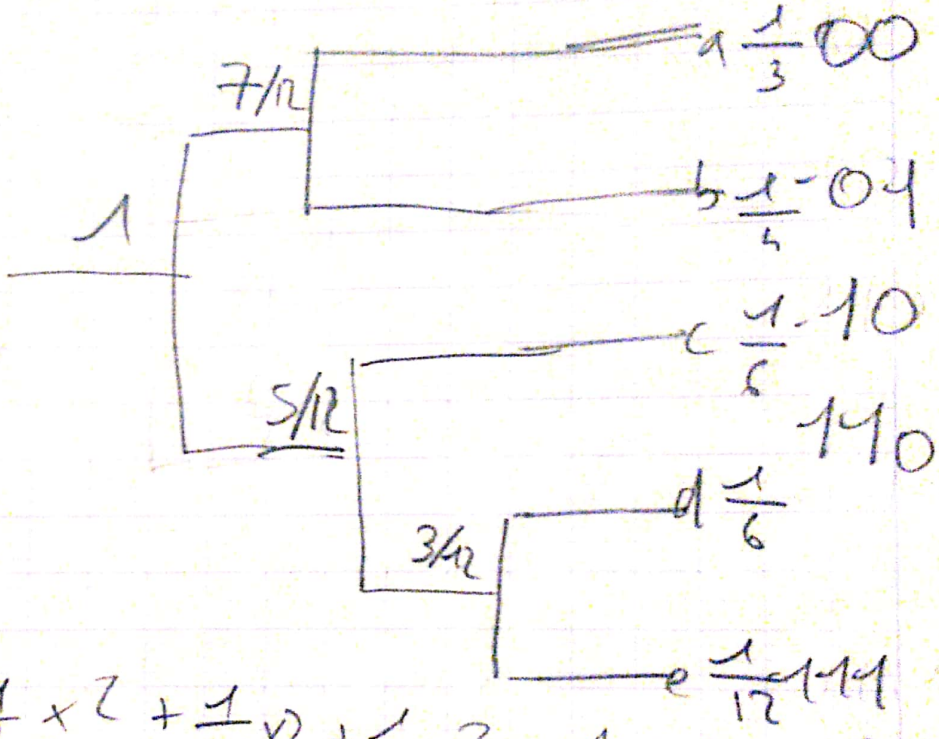
Dem 2 Med Mat

$= -\left(\frac{1}{3}\log 3 + \frac{2}{4} + \frac{2}{6}\log 6 + \frac{1}{12}\log 12\right)$

$\vdots$

$= 2,06$

e)



$\ell(\ell) = \frac{1}{3} \times 2 + \frac{1}{4} \times 2 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{12} \times 3 = 2,5$

3) a c d a b

X/Y	a	b	c	
α	1/8	1/8	1/4	1/2
β	0	3/8	1/8	1/2
	1/8	1/2	3/8	

La somme des probas vaut 1.

X et Y pas indépendant $X \not\perp Y$ donc $\hat{I}(X,Y) > 0$

$$H(X) = H\left(\frac{1}{2}, \frac{1}{2}\right) = -\left(\frac{1}{2}\log\left(\frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{1}{2}\right)\right)$$

$$= -\frac{1}{2} \times (-1) - \frac{1}{2} \times (-1)$$

$$= 1$$

$$H(Y) = H\left(\frac{1}{8}, \frac{1}{2}, \frac{3}{8}\right) = -\left(\frac{1}{8}\log\left(\frac{1}{8}\right) + \frac{1}{2}\log\left(\frac{1}{2}\right) + \frac{3}{8}\log\left(\frac{3}{8}\right)\right)$$

$$= \frac{1}{8} \times 3 + \frac{1}{2} \times 1 + \frac{3}{8}(\log 8 - \log 3)$$

$$= 1{,}41$$

$$P(X = \alpha \mid Y = c) = \frac{P((X = \alpha) \cap (Y = c))}{P(Y = c)}$$

$$= \frac{(1/4)}{(3/8)}$$

$$H(X|Y) = \sum_{\text{valeurs de } Y} H(X|Y=v) \times P(Y=v)$$

$$= H(1,0) \times P(Y=a) + H\left(\frac{1}{4}, \frac{3}{4}\right) \times P(Y=b) + H\left(\frac{2}{3}, \frac{1}{3}\right) \times P(Y=c)$$

$$= 0 \times \frac{1}{8} + \left(\frac{1}{4}\log 4 + \frac{3}{4}(\log 4 - \log 3)\right) \times \frac{1}{2} + \left(\frac{2}{3}(\log 3 - \log 2) + \frac{1}{3}\log 3\right) \times \frac{3}{8}$$

$$= 0{,}75$$

$$I(X,Y) = H(X) - H(X|Y)$$
$$= 1 - 0{,}75$$
$$= 0{,}25$$

$H(Y) = H(X,Y) - H(Y)$

en effet

$$H(X,Y) = \frac{3}{8}\log 8 + \frac{1}{4}\log 4 + \frac{3}{8}(\log 8 - \log 3)$$
$$= \frac{22}{8} - \frac{3}{8} \times \log 3$$
$$= 2{,}16$$

On a bien

$$H(X|Y) = H(X,Y) - H(Y)$$
$$0{,}75 = 2{,}16 - 1{,}41$$

donc $I(X,Y) = H(X) - \left(H(X,Y) - H(Y)\right)$
$= H(X) + H(Y) - H(X,Y)$

1.3.13

X est à valeur dans $V = [\![0;9]\!] \cup \{a, b, \dots, z\}$

$|X(\Omega)| = |V| = 36$, mais X n'est pas uniforme (vecteur de probabilité de taille 36 mais pas équiprobable)

En effet :

$P(\text{Un chiffre}) = \frac{1}{3} \times \frac{1}{10} = \frac{1}{30}$

$P(\text{voyelle}) = \frac{1}{3} \times \frac{1}{6} = \frac{1}{18}$

$P(\text{consonne}) = \frac{1}{3} \times \frac{1}{20} = \frac{1}{60}$

c'est n'est pas la loi uniforme

$H(X) < \log(36) = H(U)$

$H(X) = -\sum_{\text{valeurs}} \text{proba} \log(\text{proba})$

$= -\left(10 \times \frac{1}{30} \log\left(\frac{1}{30}\right) + 6 \times \frac{1}{18} \log\left(\frac{1}{18}\right) + 20 \times \frac{1}{60} \log\left(\frac{1}{60}\right) \right)$

$= \frac{1}{3}\log(36) + \frac{1}{3}\log(18) + \frac{1}{3}\log(60)$

$= \frac{1}{3}(1 + \log(3) + \log(5)$

$+ \frac{1}{3}(1 + 2\log(3))$

$+ \frac{1}{3}(2 + \log 3 + \log 5)$

$= \frac{4}{3}(1 + \log) + \frac{2}{3}\log 5$

$\approx 5 < \log(36)$

1, 3, 6

1)

$H(X,Y) = H(0, \frac{3}{4}, \frac{1}{8}, \frac{1}{8})$

$= -\left(\frac{3}{4}\log\left(\frac{3}{4}\right) + 2\left(\frac{1}{8}\log\left(\frac{1}{8}\right)\right)\right)$

$= -\left(\frac{3}{4}(\log 3 - \log 4)\right.$

$\left. + \frac{2}{8}(\log 1 - \log 8)\right)$

$= 1,06$

2) $H(X) = H(\frac{3}{4}, \frac{1}{4})$

$= -\left(\frac{3}{4}\log\left(\frac{3}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right)\right)$

$= 2 - \frac{3}{4}\log(3)$

$= 0,81$

Mat Mod

Sem 3 $H(X) = H\left(\frac{1}{8}, \frac{7}{8}\right) = \frac{1}{8}\log 8 + \frac{7}{8}(\log 8 - \log 7)$

$= 3 - \frac{7}{8}\log 7 \approx 0,54$

b)

$H(X \mid Y=a) = H(0,1) = 0$

$H(X \mid Y=\beta) = H\left(\frac{6}{7}, \frac{1}{7}\right)$

$= \frac{6}{7}(\log 6 - \log 7) + \frac{1}{7}\log 7$

$\leq 0,59$

Loi Maud

$H(X|Y) = \sum_{\text{valeurs}} H(X \mid Y=y) \times P(Y=y)$

2.4.1

1) $H(W) = H\left(\frac{1}{3}, \frac{1}{4}, \frac{1}{6}, \frac{1}{6}, \frac{1}{12}\right)$

$= \left(\frac{1}{3}\log 3 + \frac{1}{4}\log 4 + \frac{2}{6}\log 6 + \frac{1}{12}\log 12\right)$

Lun 4 — Mod Maths — TD 1

L'entropie de la source

$$H(W) = \lim_{N\to\infty} \frac{1}{N} H(X_1, X_2, \dots, X_N)$$

$$\overset{id}{=} \lim_{N} \frac{1}{N} \left(H(X_1) + \dots + H(X_N) \right)$$

$$= \lim_{N} \frac{1}{N} \left(N \times H(X_1) = H(X_1) \right)$$

Exercice 2.4.4

1) $X \sim U(\dots 64 \text{ valeurs} \dots)$

$H(X) = \log 64 = 6$

D'après le THM 1 de Shannon, il n'existe pas de stratégie permettant de dépenser en moyenne moins de 6 questions, il doit exister une stratégie permettant de dépenser en moyenne (en dichotomie).

2) $X \sim U(N \text{ valeurs})$ avec $N = \binom{32}{4}$

$$H(X) = \log \binom{32}{4} = \log \frac{32!}{28!\,4!}$$

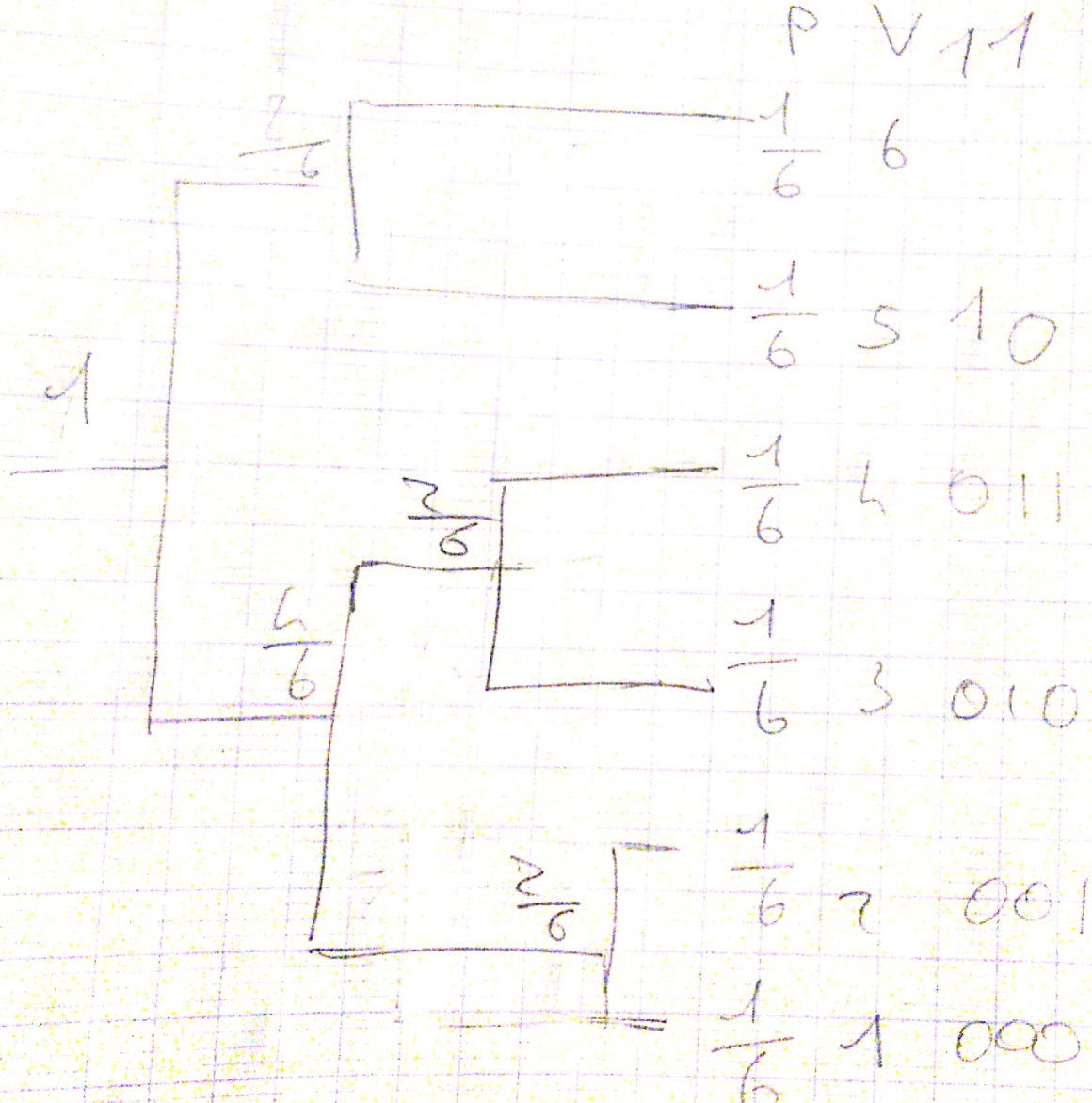
$$= \log 35\,960$$

$$= 15,13$$

3) X ~ U(... 52! valeurs ...)

$\log(52!) = 225.581$

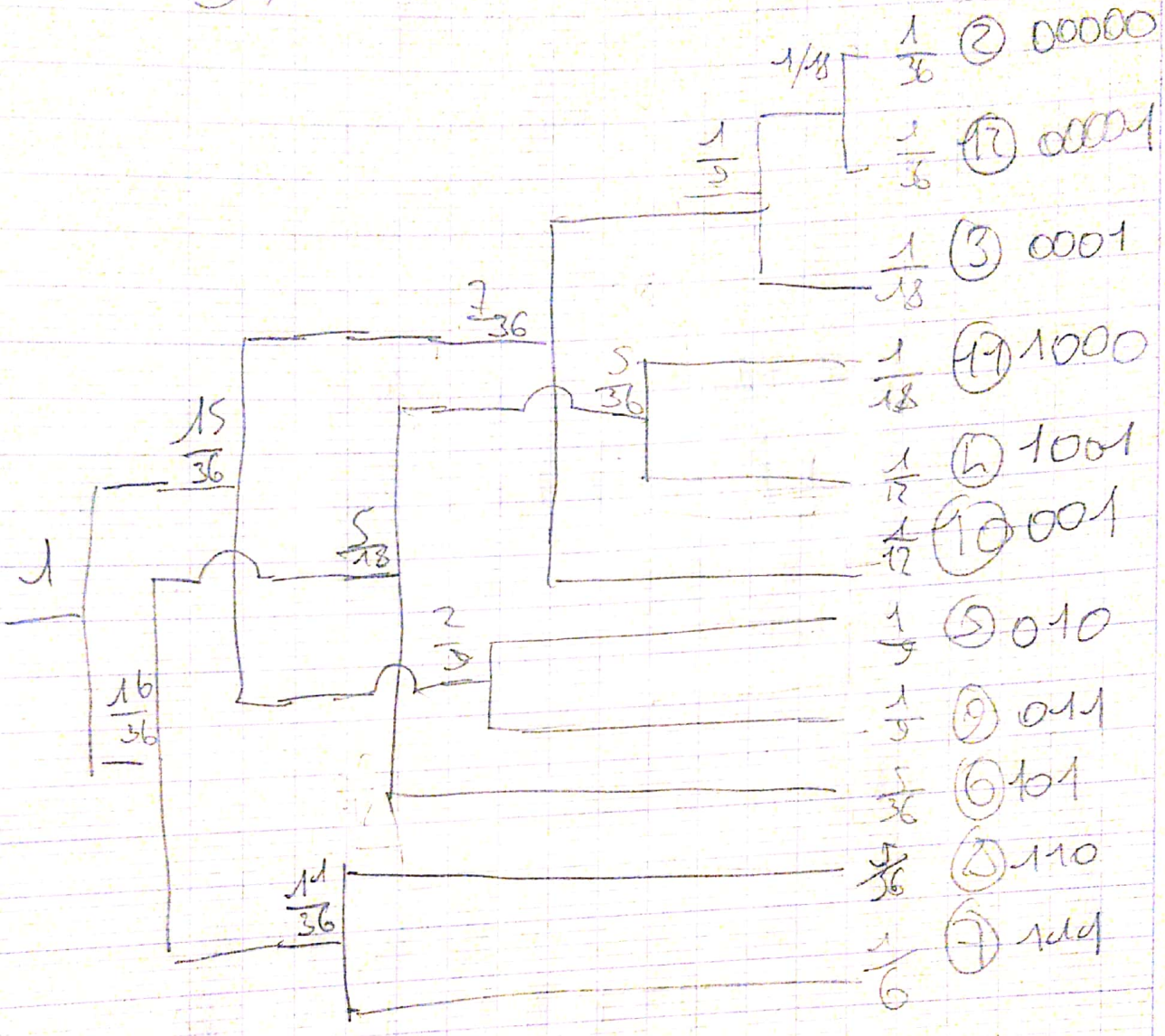
4)



$E(Y) = \frac{5}{6} \times (4 \times 3 + 2 \times 2) = \frac{16 \times 2,7}{6}$

Sm1 Mod Maths TD 2

5)



$H(X) = 3,27$

$\ell(\varphi) = 3,30$

Exercice 2.4.2

p\R	b	m	
b	5/8	1/16	11/16
m	3/16	1/8	5/16
	13/16	3/16	

$P\{\text{Alphonse se plante}\}$

$= P\{R = m\}$

$= \frac{3}{16}$

$P\{\text{Nelson se plante}\}$

$= P\{(P=b) \cap (R=m) \cup (P=m) \cap (R=b)\}$

$= \frac{1}{16} + \frac{3}{16}$

$= \frac{1}{4}$

Alphonse a raison

2) Dans le contexte de M. Nelson :

$I(R,P) = H(R) - H(R|P)$

$= H(R) - (H(R,P) - H(P))$

$= H(R) + H(P) - H(R,P)$

$= H\left(\frac{13}{16}, \frac{3}{16}\right) + H\left(\frac{11}{16}, \frac{5}{16}\right)$

$= 0,7 + 0,9 - 1,5$

$= 0,1 > 0$

Alors que Alphonse propose une variable qui ne donne pas d'information, son détenteur triste

Sem 3n Mod maths TD 3

3) Stockage séparé de $P_1, P_2, \dots P_{1000}$
$R_1, \dots R_{1000}$

Le nombre de bits minimum d'après le THM 1 de Shannon:

$H(P_1, P_2 \dots P_{1000}) = H(P_1) \times 1000 = 1000$ bits

$+ H(R_1, R_2 \dots R_{1000}) = H(R_1) \times 1000 = 1000$ bits

$\Rightarrow$ 1600 bits

4) $1000 \times H(P, R) = 1000 \times 1,5$
$= 15000$ bits

TD 1

Sem 5 Modélisation

Exercice 2.4.5

$V_4 \mapsto 0$

$V_2 \mapsto 10$

$V_1 \mapsto 110$

$V_3 \mapsto 1110$

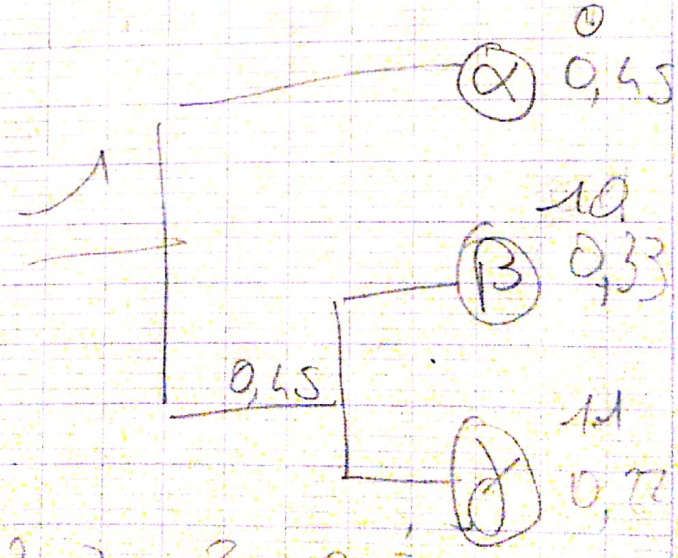
$V_5 \mapsto 1111$

Exercice 2.4.6

$H(w) = H(0,45 ; 0,33 ; 0,22)$

$= -(0,45 \log(0,45) + 0,33 \log(0,33) + 0,22 \log(0,22))$

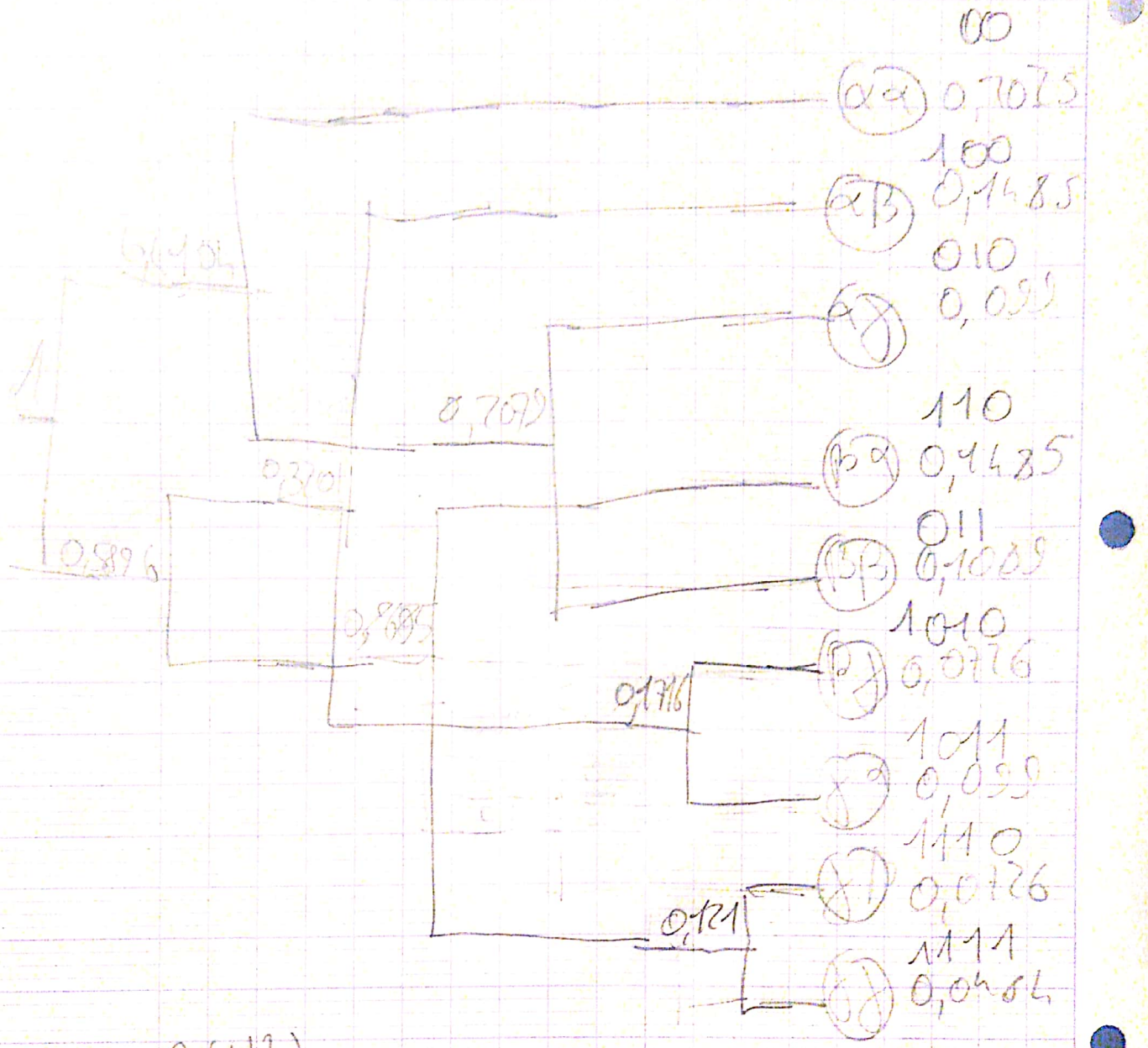
$= 1,53$



$P(0) = 0,45 \times 1 + 0,33 \times 2 + 2 \times 0,22$

$= 1,55 > H(w)$

Ici on peut décoder les lettres



$$\ell(\Psi) = 2 \times 0,7075 + 3 \times 0,1485 + 2 \times 0,1405 + \ldots + 4 \times 0,0484$$

$$= 3,0901$$

On $3,0901/2 = 1,54505$

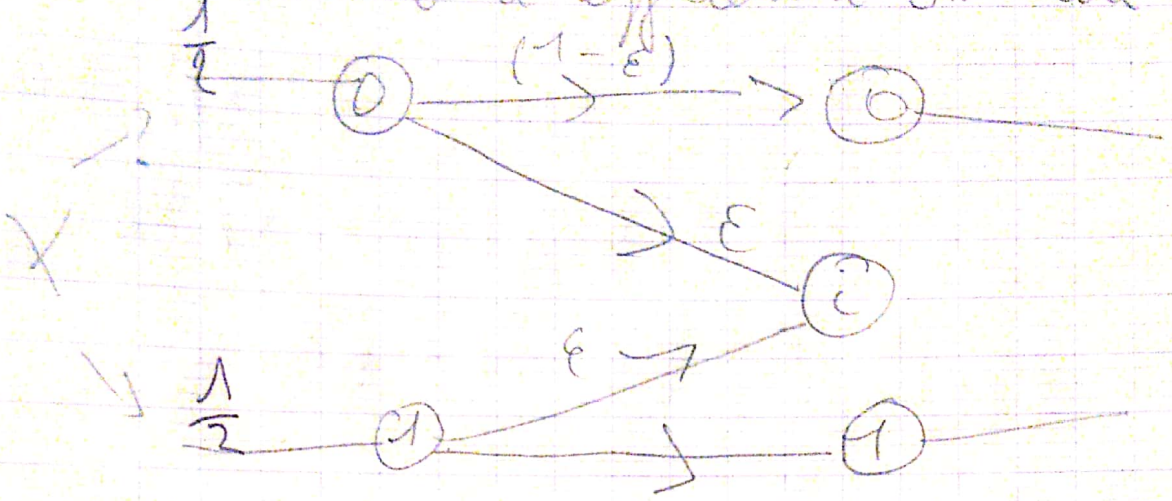
$$1 + H(W) = 2,53 > \ell(\Psi) = 1,55 > \frac{\ell(\Psi)}{2} = 1,54505$$

$$> 1,53$$

Jan 5 — Modélisation — TD 2

Exercice 3.5.1

Canal à effacement binaire



$$P = \begin{pmatrix} (1-\varepsilon) & 0 & \varepsilon \\ 0 & (1-\varepsilon) & \varepsilon \end{pmatrix}$$

$$C(\varepsilon) = \max_{X} I(X, \underline{\Phi}(X))$$

$\underline{\Phi}(X)$ variable à la sortie du canal

On fait un tableau avec

X \ $\underline{\Phi}(X)$	0	1	ε	
0	$\frac{1}{2}(1-\varepsilon)$	0	$\frac{1}{2}\varepsilon$	1/2
1	0	$\frac{1}{2}(1-\varepsilon)$	$\frac{1}{2}\varepsilon$	1/2
			ε	

⇔ X est symétrique

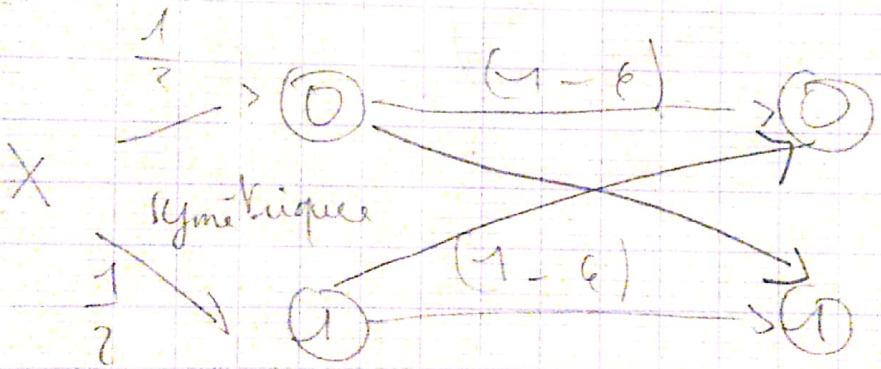
$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

$$= H\left(\frac{1}{2}\varepsilon, \frac{1}{2}\varepsilon\right) + H\left(\frac{1}{2}(1-\varepsilon), \frac{1}{2}(1-\varepsilon)\right) - H\left(\frac{1}{2}\varepsilon, \frac{1}{2}\varepsilon, \frac{1}{2}(1-\varepsilon), \frac{1}{2}(1-\varepsilon), 0, 0\right)$$

$$= 1 - (1-\varepsilon)\log\left(\frac{1}{2}(1-\varepsilon)\right) - \varepsilon\log\varepsilon + (1-\varepsilon)\log\left(\frac{1}{2}(1-\varepsilon)\right) + \varepsilon\log\frac{\varepsilon}{2}$$

$$C(\varepsilon) = \begin{matrix} \varepsilon \\ 1-\varepsilon \end{matrix} \log\varepsilon - \varepsilon\log 2$$

Canal binaire symétrique



$$P = \begin{pmatrix} (1-\varepsilon) & \varepsilon \\ \varepsilon & (1-\varepsilon) \end{pmatrix}$$

X \ Φ(X)	0	1	
0	(1-ε)/2	ε/2	1/2
1	ε/2	(1-ε)/2	1/2
	1/2	1/2	1

$$I(X; \Phi(X)) = H(X) + H(\Phi(X)) + H(X, \Phi(X)) = 1 + 1 + (1-\varepsilon)\log\left(\frac{1-\varepsilon}{2}\right) + \varepsilon\log\left(\frac{\varepsilon}{2}\right)$$

$$C(\varepsilon) = -1 - (1-\varepsilon)\log(1-\varepsilon) + \varepsilon\log\varepsilon$$

Sem 6

Math / Modélisation

TD 1

Exo 3.5.1 (suite)

Cours

Canal binaire symétrique

$C(\varepsilon) = 1 + \varepsilon \log \varepsilon + (1-\varepsilon)\log(1-\varepsilon)$

Si $\varepsilon = \frac{1}{2}$ qq soit loi de X, on a $Y \perp X$

$C = \max_X I(X,Y)$

Cas général :

X/X	0	1	
0	$q(1-\varepsilon)$	$q\varepsilon$	$q=(1-p)$
1	$p\varepsilon$	$p(1-\varepsilon)$	p
	$1+2p\varepsilon-(p+\varepsilon)$ $= P(Y=0)$	$p+\varepsilon-2p\varepsilon$ $P(Y=1)$	

$I(X;Y) = H(X) + H(Y) - H(X,Y)$

$= (-p\log p - 1+p\log(1-p)$

$- (1+2p\varepsilon-(p+\varepsilon))\log(1+2p\varepsilon-(p+\varepsilon))$

$- (p+\varepsilon-2p\varepsilon)\log(p+\varepsilon-2p\varepsilon)$

$+ q(1-\varepsilon)\log(q(1-\varepsilon))$

$+ q\varepsilon\log(q\varepsilon) + p\varepsilon\log(p\varepsilon)$

$+ p(1-\varepsilon)\log(p(1-\varepsilon))$

$$\frac{dI}{dp} = (\varepsilon - 1 - \varepsilon) \log\left(\frac{p(1-\varepsilon) + 1 - p}{p\varepsilon + 1 - p)(1-\varepsilon)}\right)$$

$$\Leftrightarrow p = \frac{1}{2}$$

Exercice 3.5.5

1)

$$C(\varepsilon) = H\left(\frac{1}{2}, \frac{1}{2}\right) + H\left(\frac{1}{2}, \frac{1}{2}\right) - H\left(\frac{1-\varepsilon}{2}, \frac{\varepsilon}{2}, \frac{1-\varepsilon}{2}, \frac{\varepsilon}{2}\right)$$

$$= 1 + (1-\varepsilon)\log(1-\varepsilon) - \varepsilon \log \varepsilon$$

$$= 1 + 0,95 \log 0,95 - 0,05 \log 0,05$$

$$\simeq 0,7136$$

On transmettre un taux de bits significatif de 71,36 %

$800 \times 0,7136 \simeq 571$ bits fiables.

2) Taux de bits significatifs :

$\frac{500}{800} = 0,625$ or $0,625 < 0,7136$

Et n'est fiable

On en déduit que $\varepsilon = 0,072$

(à tâton par dichotomie)

Sem 6 Mod Math TD 2

Exercice 3.5.6

Il faut en moyenne 15 bits par mot de la source

$C(0,1) = 0,531$

$1000 \times 0,531 = 531$

→ 531 bits/s significatifs

$531/15 \simeq 35$ mots de la source par seconde.

Sem 7 — Mod / Maths — TD 1

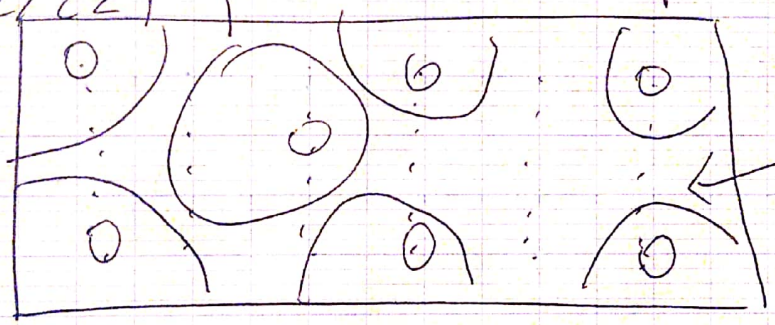
Codes correcteurs d'erreurs

Vitesse signif: débit — Code block: fraction de code. Paquet de message (ex: 100 bits) taille réelle remplacé par un truc en plus taille d'bloc plus long pour rendre la transmission résistante au bruit

Quand on utilise un canal bruité on envoie des trucs d'un certain nb de bits.

Distance minimum pour décoder:

$U_n = (\mathbb{Z}/2\mathbb{Z})^n$



↑ Espace vectoriel binaire

Combinaisons binaires à N termes

On choisit ceux qui sont les + éloignés les uns des autres (au sens de la distance Hamming (on compte le nb de site où les 2 config diffèrent

Sous espace vectoriel:
$\forall \vec{u}, \vec{v} : V$
$\forall \alpha, \beta \in \mathbb{R}$
$S(\alpha \cdot u + \beta \cdot v) \in V$

Les mots

3.4 p 31 : code linéaire de Hamming

$\mathcal{C} \subset U_n = (\mathbb{Z}/2\mathbb{Z})^N$ si u et v sont des mots de code $u+v$ est encore un mot de code.

$\mathcal{C} = \{ u = (u_1, u_2, \dots, u_7) \in U_7 / H \times {}^t u = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \}$

Sem 7 Mod/Maths TD 2

3) ~~Soit~~

remarquons que
$\Delta_{Ht}(u, u') = \Delta_H(\underbrace{u - u'}_{v \in \mathcal{C}}, \vec{0})$

$\forall v \in \mathcal{C}, v \neq 0 \rightarrow v$ comporte au moins 3 bits non nuls

car il y a dans H pas de colonne $\begin{pmatrix}0\\0\\0\end{pmatrix}$ donc pas deux fois 1

elles sont 2 à 2 distinctes donc on peut n'avoir 2 fois 1.

donc on a au moins 3x la valeur 1.

4.1)

$Ht \times \begin{pmatrix}1\\1\\1\\0\\0\\0\end{pmatrix} = \begin{pmatrix}0\\0\\1\end{pmatrix}$ C'est la colonne n°1 de Ht. Donc on enlève le 1 on a

$\begin{matrix}0\\1\\1\\1\\0\\0\end{matrix}$ → mot de code

Sem 7 — Mod Maths — Cours 9

$$H(W) = \lim_{n\to\infty} \frac{1}{n} H(H_1, \dots, H_n)$$

$$= \lim_{n\to\infty} \frac{1}{n} H(X_1 \dots X_n)$$

$$= \lim_{n\to\infty} \frac{1}{n} \left[H(X_n | X_{n-1} \dots X_1) + H(X_{n-1} | \dots \right.$$

Markov $= \lim_{n\to\infty} \frac{1}{n} \Big[H(X_n | X_{n-1})$
$+ H(X_{n-1} | X_{n-2})$
$+ H(X_{n-2} | X_{n-3})$
$\dots$
$+ H(X_2 | X_1)$
$+ H(X) = \sum_{i=1}^{N} \pi_i H(X_2 | X_1 = e_i)$

Doblin : si X_n est irréductible, apériodique

Rapidement, loi$(X_k) \to \pi$ (inv).

$$H(X_k | X_{k-1}) = \sum_{i=1}^{N} H(X_k | X_{k-1} = e_i) \times P(X_{k-1} = e_i)$$

$$\to \sum_{i=1}^{N} H(X_k | X_{k-1} = e_i) \times \pi_i$$

$$H(W) = \sum_{i=1}^{N} \pi_i H(p_{i,\cdot}) \leq H(\pi)$$

Rappel d'enchaînement

$H(X,Y) = H(X|Y) + H(Y)$

$H(X|Y) = \sum_{y\in Y} H(X|Y=y) \cdot P(Y=y)$