**EISTI** **Année scolaire 2008-2009**

**DÉPARTEMENT "INFORMATIQUE"**

---

# THÉORIE DE L'INFORMATION

**Devoir surveillé du 26 mai 2009.** *CORRIGE.*

*une solution qui vous démolit vaut mieux que n'importe quelle incertitude*
B. Vian

---

## Exercice 1. Dans le brouillard

Le soir, le Hérisson allait chez son ami l'Ourson pour admirer ensemble les étoiles en buvant le thé. Ce soir là, sur le chemin, le Hérisson s'est perdu dans le brouillard. Il s'est retrouvé soudain dans un monde merveilleux, où des créatures qu'il rencontrait lui paraissaient fantastiques et mystérieuses. Voici les différents personnages qu'il peut rencontrer au cours de son voyage et les probabilités de leur apparition :

| Cheval | Eléphant | Chien | Hibou | Papillon | Luciole | Poisson |
|---|---|---|---|---|---|---|
| 0.3 | 0.05 | 0.15 | 0.25 | 0.05 | 0.1 | 0.1 |

1. **1 Pt** Lors d'une rencontre quelle est l'incertitude moyenne que le hérisson a sur l'identité du personnage ?
2. **1 Pt** Quelle est la quantité d'information associée à la rencontre avec :

   a) un mammifère b) un insecte

3. Le hérisson souhaite deviner qui se cache derrière la silhouette indéfinie qui est devant lui.Le personnage accepte que le hérisson lui pose des questions auxquelles on peut répondre par "oui" ou par "non".

   (a) **1 Pt** Quelle est l'entropie associée à une question dont la probabilité de "oui" est $p$ ? Quelle est la valeur maximale de cette entropie ? Pour quelle valeur de $p$ elle est atteinte ?

   (b) **1 Pt** Quelle est la meilleure première question a poser ? Justifiez votre choix.

   (c) **1 Pt** Proposez une stratégie de questions qui permet au hérisson de deviner le personnage. Chaque nouvelle question doit apporter le plus possible d'information. Justifiez vos choix de questions. Représentez la stratégie sous forme d'arbre.

   (d) **1 Pt** Calculez le nombre moyen de questions à poser pour votre stratégie.

(e) **1 Pt** Quelle est la borne inférieure pour le nombre moyen de questions binaires à poser pour deviner le personnage ? Justifiez votre réponse.

---

**Solution de l'exercice 1**

1. Lors d'une rencontre quelle est l'incertitude moyenne que le hérisson a sur l'identité du personnage ? L'incertitude moyenne est l'entropie :

   $$H(X) = -0.3\log_2(0.3) - 2{\cdot}0.05\log_2(0.05) - 0.15\log_2(0.15) - 0.25\log_2(0.25) - 2{\cdot}0.1\log_2(0.1) \simeq 2.52$$

2. Quelle est la quantité d'information associée à la rencontre avec :

   a) un mammifère b) un insecte

   La probabilité de rencontrer un mammifère est

   $$P[mammifre] = P[Cheval] + P[Chien] + P[Elphant] = 0.5$$

   Alors l'information associée est

   $$h(Mammifre) = -\log_2(P[mammifre]) = 1$$

   La probabilité de rencontrer un insecte est

   $$P[mammifre] = P[Luciole] + P[papillon] = 0.15$$

   Alors l'information associée est

   $$h(insecte) = -\log_2(P[insecte]) = 2.73$$

3. Le hérisson souhaite deviner qui se cache derrière la silhouette indéfinie qui est devant lui.Le personnage accepte que le hérisson lui pose des questions auxquelles on peut répondre par "oui" ou par "non".

   (a) Quelle est l'entropie associée à une question dont la probabilité de "oui" est $p$ ? Quelle est la valeur maximale de cette entropie ? Pour quelle valeur de $p$ elle est atteinte ?

   $$H_2(p) = -p\log_2(p) - (1-p)\log_2(1-p), \quad \max_{p\in[0,1]} H_2(p) = H_2(0.5) = 1$$

   (b) Quelle est la meilleure première question a poser ? Justifiez votre choix. La meilleure question à poser au début est celle qui apporte en moyenne le plus d'information, donc celle dont l'entropie associée est la plus élevée possible. Dans le cas idéal la meilleure question est celle qui divise l'ensemble de choix possibles en deux groupes équiprobables.

   Dans cette exemple, la question "Est-ce un mammifère ? " a la probabilité de "Oui" égale justement à 0.5, donc c'est une première question optimale.

   Il est tout de même possible dans ce cas particulier de choisir comme première question "Es tu Cheval, Luciole, Poisson ? ". En effet, on peut facilement remarquer que la probabilité de ce sous-ensemble de personnages est aussi égale à 0.5.

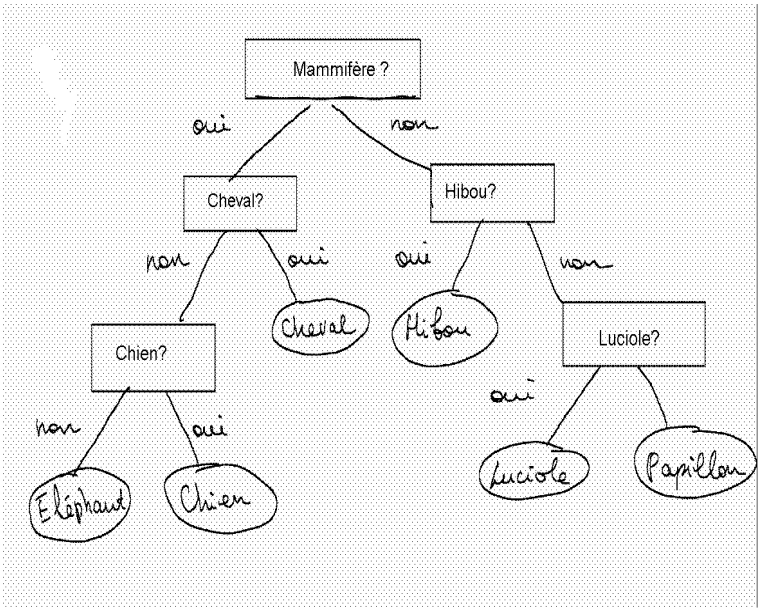


FIGURE 1 – Stratégie de questions

(c) Proposez une stratégie de questions qui permet au hérisson de deviner le personnage. Chaque nouvelle question doit apporter le plus possible d'information. Justifiez vos choix de questions. Représentez la stratégie sous forme d'arbre. **Justification.** A chaque étape de construction on choisit la question qui divise l'ensemble de réponses possibles en deux parties équiprobables. Si l'équiprobabilité exacte n'est pas possible, on choisit la question dont le probabilité de "Oui" est aussi proche que possible de 0.5.

Par exemple, si la réponse à la première question "es tu un mammifère ? " est positive l'ensemble de candidats restants est

| personnage, m | Cheval | Eléphant | Chien |
|---|---|---|---|
| P(m) | 0.3 | 0.05 | 0.15 |
| $P(m \mid mammifre)$ | $\frac{0.3}{0.5}$ | $\frac{0.05}{0.5}$ | $\frac{0.15}{0.5}$ |

Il n'est pas possible de diviser l'ensemble en deux parties complémentaires et équiprobables. La question qui se rapproche le mieux d'équiprobabilité de "oui" et de "non" est "Es tu Cheval ? ". La probabilité de "oui" est alors de $3/5$. Si la réponse à la première question "es tu un mammifère ? " est négative l'ensemble de candidats restants est

| personnage, m | Hibou | Papillon | Luciole | Poisson |
|---|---|---|---|---|
| P(m) | 0.25 | 0.05 | 0.1 | 0.1 |
| $P(m \parallel non\ mammifre)$ | $\frac{0.25}{0.5}$ | $\frac{0.05}{0.5}$ | $\frac{0.1}{0.5}$ | $\frac{0.1}{0.5}$ |

Il possible ici de diviser l'ensemble en deux parties complémentaires et équiprobables. La question à poser est donc est "Es tu Hibou ? ". La probabilité de "oui" est alors de $0.5$.

(d) Calculez le nombre moyen de questions à poser pour votre stratégie. D'après l'arbre on a

| Personnage | Cheval | Eléphant | Chien | Hibou | Papillon | Luciole | Poisson |
|---|---|---|---|---|---|---|---|
| Probabilité | 0.3 | 0.05 | 0.15 | 0.25 | 0.05 | 0.1 | 0.1 |
| Nombre de questions | 2 | 3 | 3 | 2 | 4 | 4 | 3 |

Le nombre moyende questions est alors égal à

$$\overline{N} = 2 \cdot 0.3 + 3 \cdot +3 \cdot 0.15 + 2 \cdot 0.25 + 4 \cdot 0.05 + 4 \cdot 0.1 + 3 \cdot 0.1 = 2.6$$

(e) Quelle est la borne inférieure pour le nombre moyen de questions binaires à poser pour deviner le personnage ? Justifiez votre réponse.

La borne inférieure est l'entropie. En effet, un questionnaire binaire peut être assimilé à un code binaire et le nombre de questions coïncide alors avec le longueur de mots de code. D'après le théorème du cours on sait que la borne inférieure de la longueur moyenne de mots de code binaire est l'entropie.

---

**Exercice 2. Compression par codage de Huffman**

L'objectif de cet exercice est d'appliquer l'algorithme de codage de Huffman à la compression de l'image suivante :

L'image est de taille $10 \times 10$ pixels, en niveaux de gris.

On suppose que chaque pixel est codé par défaut sur 8 bits.

Ainsi la taille mémoire de référence pour cette image est de 800 bits.

Elle contient trois couleurs : Noir (N=256), Gris (G=128), Blanc (B=0). Si on représente chaque pixel

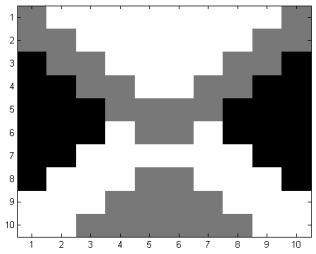

FIGURE 2 – Image à compresser

par le caractère qui représente sa couleur on obtient la matrice

$$\begin{bmatrix} G & B & B & B & B & B & B & B & B & G \\ G & G & B & B & B & B & B & B & G & G \\ N & G & G & B & B & B & B & G & G & N \\ N & N & G & G & B & B & G & G & N & N \\ N & N & N & G & G & G & G & N & N & N \\ N & N & N & B & G & G & B & N & N & N \\ N & N & B & B & B & B & B & B & N & N \\ N & B & B & B & G & G & B & B & B & N \\ B & B & B & G & G & G & G & B & B & B \\ B & B & G & G & G & G & G & G & B & B \end{bmatrix}$$

1. **Codage "pixel par pixel".**
   (a) **1 Pt** Etablir la table des fréquences des trois couleurs. Calculer l'entropie associée
   (b) **2 Pt** Construire l'arbre de code de Huffman pour cette table de fréquences. En déduire la table de code.
   (c) **1 Pt** Calculer la taille en mémoire de l'image, codée en utilisant le code de Huffman. En déduire le taux de compression.
2. **Codage par couple de pixels.** Vous allez maintenant construire un code de Huffman qui associe un mot binaire à chaque couple de pixels. Pour cela, l'image sera lue ligne par ligne, de gauche à droite et de haut en bas. Par exemple, la première ligne correspond à 5 couples :

   $$(GB), (BB), (BB), (BB), (BG)$$

   Le nouvel alphabet est ainsi constitué de tous les couples que l'on peut former avec trois couleurs : $\{N, B, G\}$.
   (a) **1.5 Pt** Former la table des fréquences pour l'alphabet des couples. En déduire l'entropie associée.
   (b) **1 Pt** Construire le code de Huffman pour cette table des fréquences.
   (c) **1.5 Pt** Calculer la taille en mémoire de l'image codée avec ce nouveau code. En déduire le taux de compression.

**Solution de l'exercice 2**

1. **Codage "pixel par pixel".**

   (a) Etablir la table des fréquences des trois couleurs. Calculer l'entropie associée

| | B | G | N |
|---|---|---|---|
| NbOccur | 44 | 32 | 24 |
| Fréquence | 0.44 | 0.32 | 0.24 |

   Calcul de l'entropie :

$$H = -0.44 \log_2(0.44) - 0.32 \log_2(0.32) - 0.24 \log_2(0.24) = 1.54$$

   (b) Construire l'arbre de code de Huffman pour cette table de fréquences. En déduire la table de code.

| Couleur | NbOccur | Couleur | NbOccur |
|---|---|---|---|
| B | 44 | GN | 56 |
| G | 32 | B | 44 |
| N | 24 | | |

   Le code obtenu :

| | B | G | N |
|---|---|---|---|
| Fréquence | 0.44 | 0.32 | 0.24 |
| Mot Code | 1 | 01 | 00 |
| Longueur | 1 | 2 | 2 |

   La longueur moyenne de mots de code est

$$1 * 0.44 + 2 * 0.32 + 2 * 0.24 = 1.56$$

   (c) Calculer la taille en mémoire de l'image, codée en utilisant le code de Huffman. En déduire le taux de compression.

$$l(U) = 1 * 44 + 2 * 32 + 2 * 24 = 156$$

   Compression :

$$156/800 = 0.195$$

2. **Codage par couple de pixels.**

   (a) Former la table des fréquences pour l'alphabet des couples. En déduire l'entropie associée.

| Caract | a | b | c | d | e | f | g | h | i |
|---|---|---|---|---|---|---|---|---|---|
| | BB | BG | BN | GG | GB | GN | NN | NB | NG |
| NbOccur | 17 | 3 | 2 | 11 | 3 | 2 | 8 | 2 | 2 |
| Fréquence | 0.34 | 0.06 | 0.04 | 0.22 | 0.06 | 0.04 | 0.16 | 0.04 | 0.04 |

L'entropie associée est

$$H = 2.66$$

(b) Construire le code de Huffman pour cette table des fréquences.

| Caract | a | d | g | b | e | c | i | f | h |
|---|---|---|---|---|---|---|---|---|---|
| NbOccur | 17 | 11 | 8 | 3 | 3 | 2 | 2 | 2 | 2 |
| Caract | a | d | g | fh | b | e | c | i | |
| NbOccur | 17 | 11 | 8 | 4 | 3 | 3 | 2 | 3 | |
| Caract | a | d | g | fh | ci | b | e | | |
| NbOccur | 7 | 11 | 8 | 4 | 4 | 3 | 3 | | |
| Caract | a | d | g | be | fh | ci | | | |
| NbOccur | 7 | 11 | 8 | 6 | 4 | 4 | | | |
| Caract | a | d | g | fhci | be | | | | |
| NbOccur | 7 | 11 | 8 | 8 | 6 | | | | |
| Caract | a | fhcibe | d | g | | | | | |
| NbOccur | 17 | 14 | 11 | 8 | | | | | |
| Caract | dg | a | fhcibe | | | | | | |
| NbOccur | 19 | 17 | 14 | | | | | | |
| Caract | fhcibea | dg | | | | | | | |
| NbOccur | 31 | 19 | | | | | | | |

Le code qui en résulte

| Caract | a | b | c | d | e | f | g | h | i |
|---|---|---|---|---|---|---|---|---|---|
| | BB | BG | BN | GG | GB | GN | NN | NB | NG |
| Fréquence | 0.34 | 0.06 | 0.04 | 0.22 | 0.06 | 0.04 | 0.16 | 0.04 | 0.04 |
| Mot | 10 | 1111 | 11010 | 00 | 1110 | 11000 | 01 | 11001 | 11011 |
| Longueur | 2 | 4 | 5 | 2 | 4 | 5 | 2 | 5 | 5 |

Longueur moyenne : $\overline{L} = 2.72$.

(c) Calculer la taille en mémoire de l'image codée avec ce nouveau code. En déduire le taux de compression. Taille en mémoire :

$$T = 136$$

Compression

$$\frac{136}{800} = 0.17$$

---

**Exercice 3. Code Linéaire.** Soit un code linéaire

$$C = \{0000000, 0001111, 0010011, 0011100, 0100101, 0101010, 0110110, 0111001,$$
$$1000110, 1001001, 1010101, 1011010, 1100011, 1101100, 1110000, 1111111\}$$

.

1. **3 Pt** Construire sa matrice génératrice
2. **2 Pt** Déterminer ses paramètres $(n, k, d)$.

---

**Solution de l'exercice 3**

1. Construire sa matrice génératrice.

   On peut déduire la dimension du code à partir de nombre de mots. On sait qu'un code linéaire de dimension $k$ a $2^k$ mots. Ici il y a 16 mots donc la dimension $k = 4$. la matrice génératrice est formée de $4$ mots formant une base de $C$ placés en lignes. On a donc

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

2. Déterminer ses paramètres $(n, k, d)$.

   On a $n = 7$, c'est la longueur des mots, $k = 4$, la dimension (voir la question précédente) et $d = \min_{x \in C,\ x \neq 0} dist(x, 0)$ où $dist(x, 0)$ est la distance de Hamming entre $x$ et le mot nul, autrement dit c'est le nombre de $1$ dans le mot $x$. Ainsi $d$ est le plus petit nombre de $1$ dans un mot non nul. Ici on a $d = 3$.

---

**ANNEXE. Table de logarithme de base 2**

| $p$ | $\log_2(p)$ | $plog_2(1/p)$ |
|---|---|---|
| 0.01 | -6.6438 | 0.0664 |
| 0.02 | -5.6438 | 0.1128 |
| 0.03 | -5.0588 | 0.1517 |
| 0.04 | -4.6438 | 0.1857 |
| 0.05 | -4.3219 | 0.2160 |
| 0.06 | -4.0588 | 0.2435 |
| 0.07 | -3.8365 | 0.2685 |
| 0.08 | -3.6438 | 0.2915 |
| 0.09 | -3.4739 | 0.3126 |
| 0.10 | -3.3219 | 0.3321 |
| 0.11 | -3.1844 | 0.3502 |
| 0.12 | -3.0588 | 0.3670 |
| 0.13 | -2.9434 | 0.3826 |
| 0.14 | -2.8365 | 0.3971 |
| 0.15 | -2.7369 | 0.4105 |
| 0.16 | -2.6438 | 0.4230 |
| 0.17 | -2.5563 | 0.4345 |
| 0.18 | -2.4739 | 0.4453 |
| 0.19 | -2.3959 | 0.4552 |
| 0.20 | -2.3219 | 0.4643 |
| 0.21 | -2.2515 | 0.4728 |
| 0.22 | -2.1844 | 0.4805 |
| 0.23 | -2.1202 | 0.4876 |
| 0.24 | -2.0588 | 0.494134 |
| 0.25 | -2.00 | .50 |
| 0.26 | -1.9434 | 0.5052 |
| 0.27 | -1.8889 | 0.5100 |
| 0.28 | -1.8365 | 0.5142 |
| 0.29 | -1.7858 | 0.5179 |
| 0.30 | -1.7369 | 0.5210 |
| 0.31 | -1.6896 | 0.5237 |
| 0.32 | -1.6438 | 0.5260 |
| 0.33 | -1.5994 | 0.5278 |

| $p$ | $\log_2(p)$ | $plog_2(1/p)$ |
|---|---|---|
| 0.34 | -1.5563 | 0.5291 |
| 0.35 | -1.5145 | 0.5301 |
| 0.36 | -1.4739 | 0.5306 |
| 0.37 | -1.4344 | 0.5307 |
| 0.38 | -1.3959 | 0.5304 |
| 0.39 | -1.3584 | 0.5297 |
| 0.40 | -1.3219 | 0.5287 |
| 0.41 | -1.2863 | 0.5273 |
| 0.42 | -1.2515 | 0.5256 |
| 0.43 | -1.2175 | 0.5235 |
| 0.44 | -1.1844 | 0.5211 |
| 0.45 | -1.1520 | 0.5184 |
| 0.46 | -1.1202 | 0.5153 |
| 0.47 | -1.0892 | 0.5119 |
| 0.48 | -1.0588 | 0.5082 |
| 0.49 | -1.0291 | 0.5042 |
| 0.50 | -1.00 | 0.50 |
| 0.51 | -0.9714 | 0.4954 |
| 0.52 | -0.9434 | 0.4905 |
| 0.53 | -0.9159 | 0.4854 |
| 0.54 | -0.8889 | 0.4800 |
| 0.55 | -0.8624 | 0.4743 |
| 0.56 | -0.8365 | 0.4684 |
| 0.57 | -0.8109 | 0.4622 |
| 0.58 | -0.7858 | 0.4558 |
| 0.59 | -0.7612 | 0.4491 |
| 0.60 | -0.7369 | 0.4421 |
| 0.61 | -0.7131 | 0.4350 |
| 0.62 | -0.6896 | 0.4275 |
| 0.63 | -0.6665 | 0.4199 |
| 0.64 | -0.6438 | 0.4120 |
| 0.65 | -0.6214 | 0.4039 |
| 0.66 | -0.5994 | 0.3956 |

| $p$ | $\log_2(p)$ | $plog_2(1/p)$ |
|---|---|---|
| 0.67 | -0.5777 | 0.3871 |
| 0.68 | -0.5563 | 0.3783 |
| 0.69 | -0.5353 | 0.3693 |
| 0.70 | -0.5145 | 0.3602 |
| 0.71 | -0.4941 | 0.3508 |
| 0.72 | -0.4739 | 0.3412 |
| 0.73 | -0.4540 | 0.3314 |
| 0.74 | -0.4344 | 0.3214 |
| 0.75 | -0.4150 | 0.3112 |
| 0.76 | -0.3959 | 0.3009 |
| 0.77 | -0.3770 | 0.2903 |
| 0.78 | -0.3584 | 0.2795 |
| 0.79 | -0.3400 | 0.2686 |
| 0.80 | -0.3219 | 0.2575 |
| 0.81 | -0.3040 | 0.2462 |
| 0.82 | -0.2863 | 0.2347 |
| 0.83 | -0.2688 | 0.2231 |
| 0.84 | -0.2515 | 0.2112 |
| 0.85 | -0.2344 | 0.1992 |
| 0.86 | -0.2175 | 0.1871 |
| 0.87 | -0.2009 | 0.1747 |
| 0.88 | -0.1844 | 0.1622 |
| 0.89 | -0.1681 | 0.1496 |
| 0.90 | -0.1520 | 0.1368 |
| 0.91 | -0.1360 | 0.1238 |
| 0.92 | -0.1202 | 0.1106 |
| 0.93 | -0.1046 | 0.0973 |
| 0.94 | -0.0892 | 0.0839 |
| 0.95 | -0.0740 | 0.0703 |
| 0.96 | -0.0588 | 0.0565 |
| 0.97 | -0.0439 | 0.0426 |
| 0.98 | -0.0291 | 0.0285 |
| 0.99 | -0.0145 | 0.0143 |