Série 2

Olivier Cloux

236079

Problème 2.1

1. Pour les codes A,B,D, l'alphabet du code est $\{0, 1\}$ (la source sera "traduite" dans dans cet alphabet), alors que pour l'alphabet du code C est $\{0, 1, 2\}$. Pour tous les codes, l'alphabet de la source est $\{a, b, c, d, e, f\}$ (chaque caractère de la source vient de cet alphabet). Finalement, chaque code a un dictionnaire spécifique :

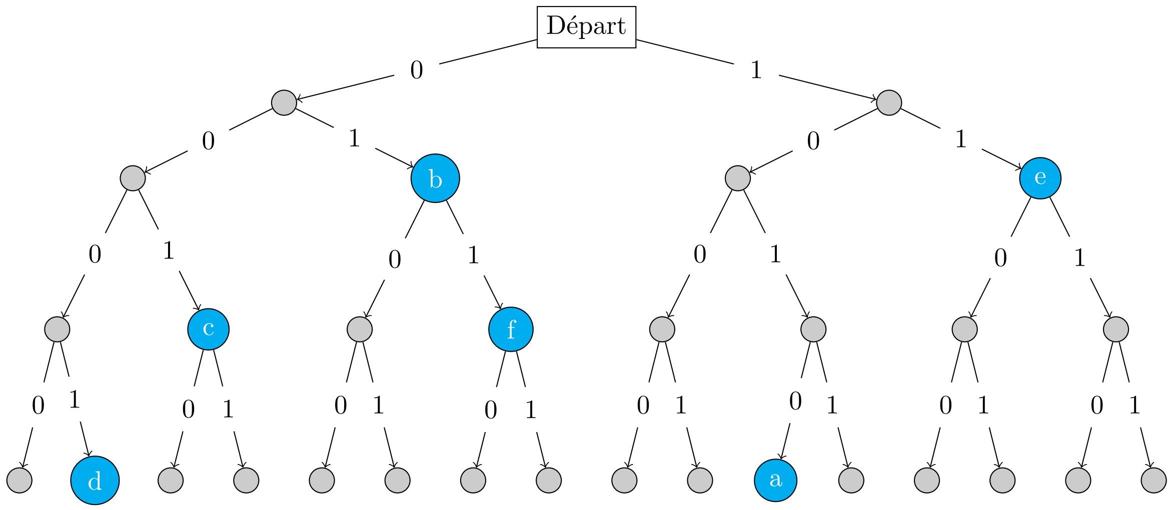
Code A $\{1010, 01, 001, 0001, 11, 011\}$

Code B $\{0, 11, 101, 100, 001, 010\}$

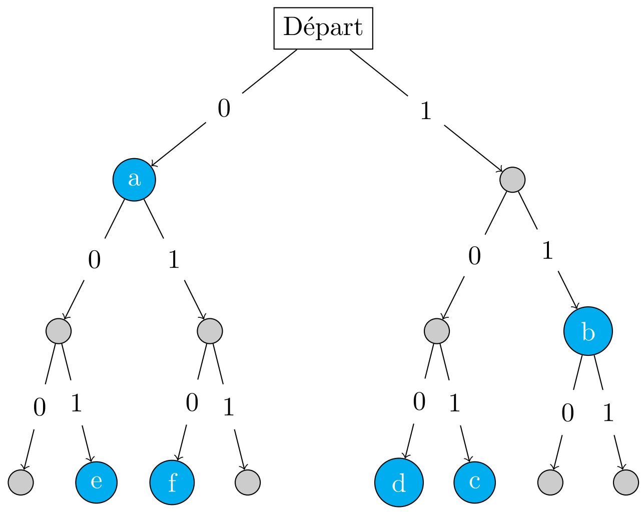
Code C $\{1, 12, 122, 10, 102, 100\}$

Code D $\{10, 011, 111, 00, 0101, 1100\}$

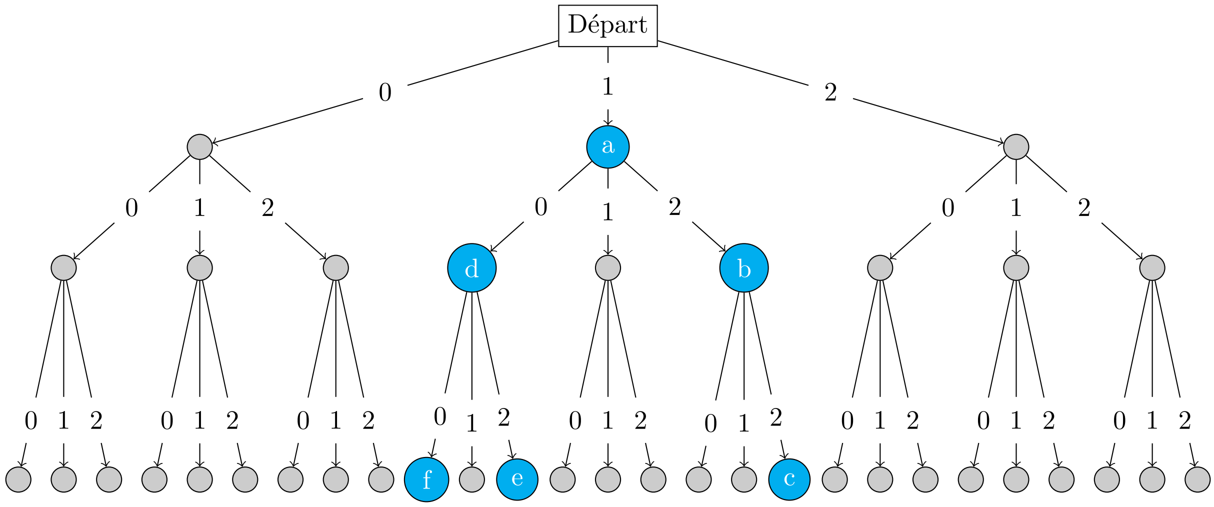
2. • Code A :



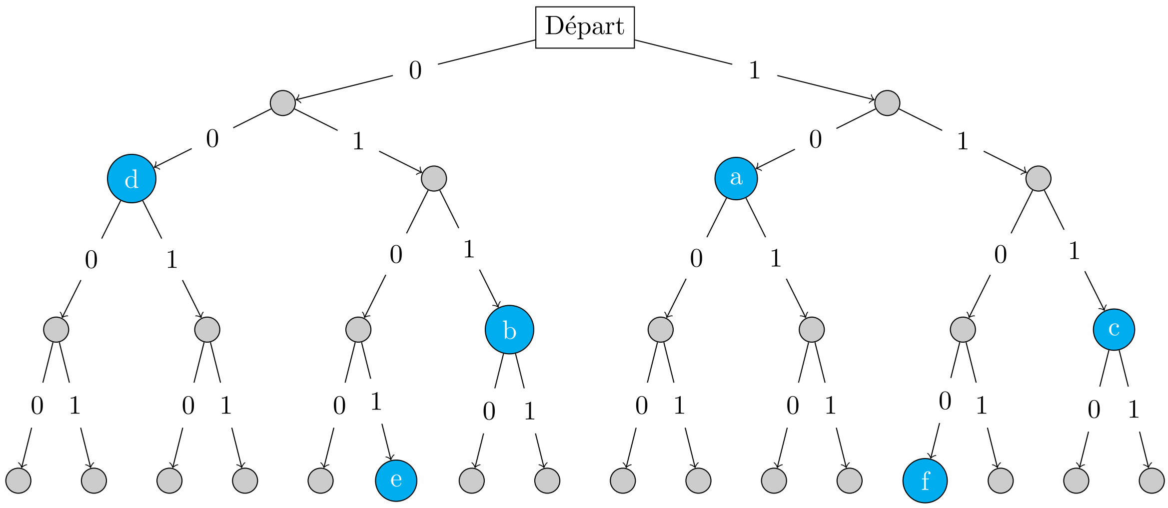
• Code B



- Code C:



- Code D:



3. (a) Code A : $2^{-4}+2^{-2}+2^{-3}+2^{-4}+2^{-2}+2^{-3}=\frac{7}{8}\leq 1$

   $\rightarrow$ Le code A respecte l'inégalité de Kraft.

   (b) Code B : $2^{-1}+2^{-2}+2^{-3}+2^{-3}+2^{-3}+2^{-3}=\frac{5}{4}=1.25>1$

   $\rightarrow$ Le code B ne respecte pas l'inégalité de Kraft.

   (c) Code C :$3^{-1}+3^{-2}+3^{-3}+3^{-2}+3^{-3}+3^{-3}=\frac{2}{3}=0,\overline{6}\leq 1$

   $\rightarrow$ Le code C respecte l'inégalité de Kraft.

(d) Code D : $2^{-2} + 2^{-3} + 2^{-3} + 2^{-2} + 2^{-4} + 2^{-4} = \frac{7}{8} \leq 1$

$\rightarrow$ Le code D respecte l'inégalité de Kraft.

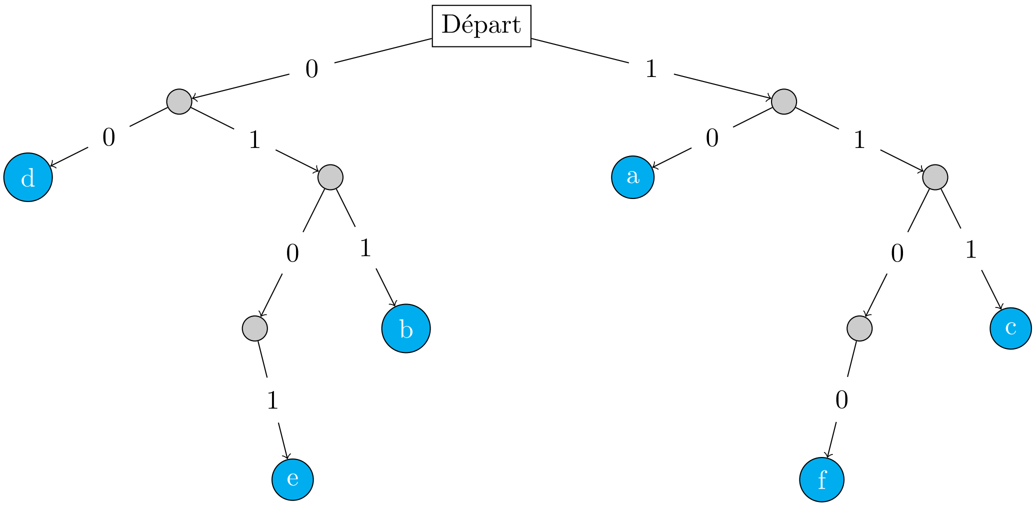
Rappelons ici quelques implications générales, relatives à tous les codes que nous étudions ici:
sans préfixe $\iff$ instantané $\rightarrow$ à décodage unique
code à décodage unique $\rightarrow$ inégalité de Kraft respectée
Évidemment, si leurs contraposées sont exactes, leurs réciproques elles sont de manière générale fausses.

4. La première partie du théorème de Kraft, grossièrement, dit que si le code est à décodage unique, alors il satisfait l'inégalité de Kraft. Par contraposée, le fait que l'inégalité ne soit pas respectée implique que le code n'est pas à décodage unique. Ainsi, le code B n'est pas à décodage unique.

5. Un code est à décodage unique si on ne peut jamais "mal interpréter" une suite valide. De plus, nous savons qu'un code à décodage instantané est à décodage unique également, raison pour laquelle nous pouvons affirmer que le code D est à décodage unique (preuve dans la question suivante).
   Pour le code B, la séquence $ad = 0100 = fa$, donc il n'est pas à décodage unique.
   Pour le A, la séquence $bab = 01101001 = fbc$, donc le code n'est pas à décodage unique.
   En revanche, le code C est lui à décodage unique car les "1" sont uniquement au début de tous les mots ; pour déchiffrer un code il n'y a donc qu'à repérer les séquences délimitées par les "1".

6. Un code est à décodage instantané si et seulement s'il est sans préfixe ; c'est à dire qu'aucun mot du dictionnaire n'est le début d'un autre (par exemple le code "0" est préfixe de "0101"). Nous voyons donc bien que les codes A,B,C sont avec préfixes (donc pas à décodage instantané), alors que le code D est à décodage instantané (car sans préfixe).

7. Arbre de décodage du code D



8. Pour les codes A et C il est possible de trouver un code instantané, car ils respectent l'inégalité de kraft. Pour cela, il suffit de regarder l'arbre complet du code, et "déplacer" un caractère sur un même niveau (afin de ne pas altérer sa longueur), de manière à n'avoir aucun préfixe. Ainsi les codes A et C deviennent (par exemple):

Symbole de Source	A	C
a	0100	0
b	10	12
c	001	112
d	0001	21
e	11	102
f	011	100

De cette manière, plus aucun caractère n'est le préfixe d'un autre, et les longueurs des mots sont respectées.

En revanche, on ne peut pas trouver de code à décodage instantané pour B, l'inégalité de Kraft n'étant pas respectée pour celui-ci ; une manière simple de le comprendre visuellement (en regardant l'arbre complet de B) et de penser à a. Son encodage est soit "1" soit "0". Pour respecter la condition des préfixes, b doit alors être dans l'autre moitié de l'arbre. Il reste alors c, d, e, f à placer, mais en ne les plaçant ni sous a ni sous b, il ne reste alors que 2 places. Il est donc impossible de trouver un tel code.

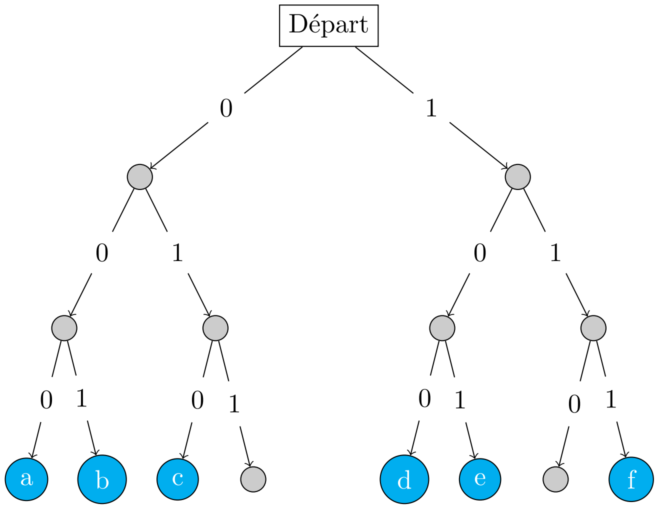
Problème 2.2

1. Nous connaissons maintenant par cœur la méthode de calcul de l'entropie, que nous pouvons donc appliquer : $6 \cdot \frac{1}{6} \log_2(6) \simeq 2.584 = H(S)$.
   Pour la longueur moyenne, nous multiplions la longueur d'un caractère encodé par sa probabilité d'apparition (et ce pour chaque caractère), ce qui nous donne :

$$\text{Code A : } \frac{1}{6} \cdot (4+2+3+4+2+3) = \frac{1}{6} \cdot 18 = \boxed{3}$$
$$\text{Code B : } \frac{1}{6} \cdot (1+2+3+3+3+3) = \frac{1}{6} \cdot 15 = \frac{15}{6} = \boxed{\frac{5}{2}}$$
$$\text{Code C : } \frac{1}{6} \cdot (1+2+3+2+3+3) = \frac{1}{6} \cdot 14 = \frac{14}{6} = \boxed{\frac{7}{3}}$$
$$\text{Code D: } \frac{1}{6} \cdot (2+3+3+2+4+4) = \frac{1}{6} \cdot 18 = \boxed{3}$$

   Pour les codes à décodage unique (C,D), la première loi de l'entropie est $L(\Gamma) \geq \frac{H(S)}{\log_2(D)}$.
   Pour le code D, cela se traduit par $L(\Gamma) \geq H(S)$ (car le code est binaire), et cette inégalité est respectée ($3 \geq 2.584$).
   C en revanche est ternaire, donc l'inégalité devient :
   $2.\overline{3} = \frac{7}{3} \geq \frac{2.584}{\log_2(3)} = 1.63$. Cette inégalité est correcte également.
   Ainsi, la première loi de l'entropie est bien respectée pour les deux codes. Les codes A et B ne sont pas considérés, car ils ne sont pas à décodage unique.

2. Les longueurs de mots d'un code de Shannon-Fano sont obtenues par $\lceil \log_2 \left(\frac{1}{p_i} \right) \rceil$. Ainsi, la longueur d'un mot de code de Shannon, avec une densité de probabilité de $\frac{1}{6}$ est : $\lceil \log_2 (\frac{1}{\frac{1}{6}}) \rceil = \lceil \log_2(6) \rceil \simeq \lceil 2.58 \rceil = 3$ (pour tous les mots).



Symbole de source	Code de Shannon-Fano
a	000
b	001
c	010
c	011
d	100
e	101
f	111

La longueur moyenne de ce code est $L(\Gamma_{SF}) = \frac{1}{6} \cdot (3 \cdot 6) = 3$, la même que celle de D.

Vu que nous travaillons avec un code binaire, la seconde inégalité de l'entropie devient $H(S) + 1 \geq L(\Gamma_{SF}) \geq H(S)$

Nous connaissons $H(S) \simeq 2.584$, et $L(\Gamma_{SF}) = 3$

$\rightarrow 1 + 2.584 = 3.584 \geq 3 \geq 2.584$. La seconde inégalité de l'entropie est donc respectée.

3. Le code D_1 est instantané car aucun mot n'est un préfixe d'un autre.

   Sa longueur moyenne est de $\frac{1}{6}(4 \cdot 3 + 2 \cdot 2) = \frac{16}{6} = 2,\overline{6}$. Cette longueur est plus courte que celle du code de Shannon-Fano (bien que toujours plus grande que $H(S)$). En effet, si le code de Shannon-Fano est garanti efficace, il n'est pas garanti être le meilleur ; ce code D_1 en est la preuve.

4. La longueur des mots de code de Shannon se calcule comme avant ($\lceil \log_2 \left(\frac{1}{p_i}\right) \rceil$)

Symbole de Source	a	b	c	d	e	f
Probabilité	0.3	0.2	0.01	0.08	0.01	0.4
longueur approximative	1.73	2.32	6.64	3.64	6.64	1.32
longueur arrondie	2	3	7	4	7	2

Ce qui nous donne une longueur moyenne de $L(\Gamma_{SF}) = \boxed{2.46}$

La longueur moyenne de D (appliquée à S') est de

$2 \cdot 0.3 + 3 \cdot 0.2 + 3 \cdot 0.01 + 2 \cdot 0.08 + 4 \cdot 0.01 + 4 \cdot 0.4 = \boxed{3.03}$

La longueur moyenne du code D_1 est de

$2 \cdot 0.3 + 3 \cdot 0.2 + 3 \cdot 0.01 + 2 \cdot 0.08 + 3 \cdot 0.01 + 3 \cdot 0.4 = \boxed{2.62}$

L'entropie de S' est de

$H(S') = -0.3 \log_2(0.3) - 0.2 \log_2(0.2) - 0.01 \log_2(0.01) - 0.08 \log_2(0.08) - 0.01 \log_2(0.01) - 0.4 \log_2(0.4) \simeq \boxed{1.938}$

La seconde inégalité de l'entropie est $H(S') \leq L(\Gamma_{SF}) \leq H(S') + 1$ (car nous sommes en présence d'un code binaire). Nous devons donc vérifier l'inégalité suivante :

$1.938 \leq 2.46 \leq 2.938 = 1.938 + 1$. Cette inégalité est diablement juste, nous pouvons donc affirmer que la seconde inégalité de l'entropie est respectée pour ce code.

Pensez au pauvre étudiant que je suis ! Je fais un effort en faisant de magnifiques (quoi que chronophages) arbres, alors pourquoi ne pas me récompenser ? :D