Université Paris 8 J. Lavauzelle

Master 1 Mathématiques et applications — année 2023–24

Théorie de l'Information – Solutions feuille de TD 3

02/10/2023

Retrouvez le sujet du TD et d'autres exercices à l'adresse :

`https://lvzl.fr/teaching/2023-24/ti.html`

(⋆) exercice fondamental (⋆⋆) pour s'entraîner (⋆⋆⋆) pour aller plus loin 💻 sur machine

Exercice 1. (⋆) Codes préfixes et uniquement décodables.

Parmi les codes suivants, lesquels sont préfixes ? Lesquels sont uniquement décodables ?

1. $\{0, 10, 11\}$
2. $\{0, 01, 11\}$
3. $\{0, 01, 10\}$
4. $\{0, 101, 111, 100\}$
5. $\{00, 10, 100, 11, 110\}$

Solutions de l'Exercice 1.

1. Le code est préfixe, donc uniquement décodable.
2. Le code n'est pas préfixe : 0 est un préfixe de 01. En revanche, il est uniquement décodable. Cela peut être vu en « renversant » les mots de codes. On obtient le code $\{0, 10, 11\}$, qui lui est préfixe donc uniquement décodable. Or, l'opération de « renversement » ne modifie pas le caractère uniquement décodable d'un code. Plus formellement, si $C^+(x) = C(x_1) \dots C(x_n)$ et si l'on note $R(C(x_i))$ le renversé de $C(x_i)$, alors on observe que R est une bijection sur $\{0,1\}^+$, et que $R(C^+(x)) = R(C(x_n)) \dots R(C(x_1))$. Ensuite, le résultat provient du fait que C^+ est injective si et seulement si $R \circ C^+$ est injective.
3. Le code n'est pas préfixe car 0 est un préfixe de 01. Il n'est pas uniquement décodable, car 010 a deux antécédents par C^+ : les mots x_1x_3 et x_2x_1.
4. Le code est préfixe donc uniquement décodable.
5. Le code n'est pas préfixe : 10 est un préfixe de 100. Le code est uniquement décodable mais c'est plus difficile à démontrer. Soit $\boldsymbol{x} \in \{0,1\}^+$. On présente une procédure permettant de déterminer de manière unique les premiers mots qui constituent $\boldsymbol{x}$:
   - (A) si $\boldsymbol{x}$ commence par un 0, alors nécessairement $\boldsymbol{x}$ commence par 00, car 00 est le seul mot du code qui commence par un 0 ;
   - (B) si $\boldsymbol{x}$ commence par un 1, alors :
     - (B1) si ce 1 est suivi d'un nombre impair de 0, on doit décoder ce préfixe de $\boldsymbol{x}$ comme 10 suivi d'une suite de 00 ;
     - (B2) si ce 1 est suivi d'un nombre pair (non nul) de 0, on doit décoder ce préfixe de $\boldsymbol{x}$ comme 100 suivi d'une suite de 00 ;
     - (B3) si ce 1 est suivi d'un autre 1 :
       - (B3a) si $\boldsymbol{x}$ commence par un nombre impair $2m+1$ de 1, alors on peut éliminer m fois le préfixe 11, puis revenir au cas (B)
       - (B3b) si $\boldsymbol{x}$ commence par un nombre pair $2m$ de 1, alors :
         - (B3b1) si $\boldsymbol{x}$ commence par $2m$ fois 1 et un nombre impair $2p+1$ de 0, on ne peut décoder ce préfixe que comme $m-1$ fois 11, puis 110, puis p fois 00 ;
         - (B3b2) si $\boldsymbol{x}$ commence par $2m$ fois 1 et un nombre pair $2p$ de 0, on ne peut décoder ce préfixe que comme m fois 11, puis p fois 00.

Exercice 2. $(\star)$ Arbre binaire d'un code.

On considère le code suivant sur l'alphabet $\mathcal{X} = \{a, b, c, d, e, f, g\}$

a	b	c	d	e	f	g
00	01	100	101	1100	1101	1111

Question 1.– Ce code est-il préfixe ? Est-il uniquement décodable ? Vérifie-t'il l'inégalité de Kraft ?

Question 2.– Décoder le texte encodé suivant, tout en essayant de bien comprendre comment vous arrivez à « séparer » les lettres :

$$010011110011111100.$$

Question 3.– Construire l'arbre binaire correspondant à ce code.

Question 4.– Est-il possible de rendre ce code plus efficace, quelle que soit la distribution de la source X sur $\mathcal{X}$? Si oui, que deviendrait le texte encodé de la question 2 ?

Solutions de l'Exercice 2.

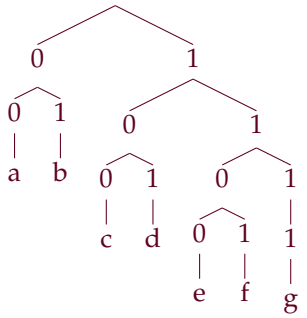
Solution Q1.

- Oui, ce code est préfixe : il n'existe aucune paire de mots distincts (u,v) telle que u est un préfixe de v.
- On peut en déduire qu'il est uniquement décodable, d'après un résultat du cours.
- Toujours d'après un théorème du cours, la séquence des longueurs du code doit donc satisfaire l'inégalité de Kraft. On peut également vérifier :

$$\sum_{i=1}^{7} 2^{-n_i} = 2 \times \frac{1}{4} + 2 \times \frac{1}{8} + 3 \times \frac{1}{16} = \frac{15}{16} < 1\,.$$

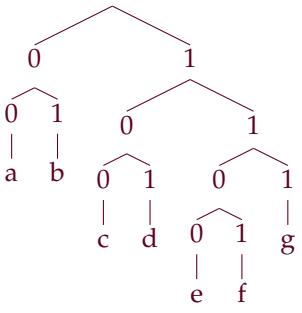
Solution Q2. On obtient le mot « bagage ».

Solution Q3. On obtient l'arbre binaire suivant.



Solution Q4.

On observe qu'il est possible d'obtenir un arbre strictement plus petit, en remontant d'un niveau la feuille portant l'étiquette g. On obtient alors l'arbre ci-contre :



Avec ce code « raccourci », le mot « bagage » s'encoderait en

$$0100111001111100$$

soit un gain de 2 bits.

Exercice 3. ($\star\star$) Code unaire.

Considérons le code suivant sur l'ensemble discret $\mathbb{N}$. À tout entier $n \in \mathbb{N}$, on associe la séquence de bits :

$$U(n) = \underbrace{0 \cdots\cdots 0}_{\longleftarrow\ n\ \longrightarrow} 1 \in \{0,1\}^{n+1}\,.$$

Question 1.– Démontrer que le code U est préfixe.

Question 2.– Sous quelle condition sur la source X, à valeurs dans $\mathbb{N}$, la longueur moyenne du code est-elle finie ?

Question 3.– Déterminer une source d'entropie finie pour laquelle le code U est optimal.

Solutions de l'Exercice 3.

Solution Q1. Pour tous $n < m$, le mot $U(n)$ est plus court que le mot $U(m)$, mais ce n'est pas un préfixe de $U(m)$: en effet, le $(n+1)$-ème bit de $U(n)$ est 1 tandis que le $(n+1)$-ème bit de $U(m)$ est 0..

Solution Q2. Soit $(p_n)_{n\in\mathbb{N}}$ la distribution de la source X. La longueur moyenne de U est égale à

$$\overline{\ell}(U) = \sum_{n=0}^{+\infty} p_n\,\ell(U(n)) = \sum_{n=0}^{+\infty} p_n(n+1) = \sum_{n=0}^{+\infty} np_n + \underbrace{\sum_{n=0}^{+\infty} p_n}_{=1}\,.$$

Donc $\overline{\ell}(U)$ est finie si et seulement si la série de terme général np_n converge, autrement dit si l'espérance de X est finie.

On peut donner des exemples (en n'oubliant pas qu'il faut également satisfaire la contrainte $\sum_n p_n = 1$).

- Les distributions telles que p_n est de la forme $p_n = C \cdot \frac{1}{n^\alpha}$, où $1 < \alpha < 2$ et $C > 0$, donnent une longueur moyenne infinie. Par exemple, c'est le cas de $p_n = \frac{6}{\pi^2} \cdot \frac{1}{n^2}$.
- Au contraire, celles où $p_n = C' \cdot \frac{1}{n^\beta}$ où $\beta > 2$ et $C' > 0$ donnent une longueur moyenne finie. C'est également le cas des distributions géométriques pour lesquelles $p_n = (1-\gamma)\gamma^n$ avec $0 < \gamma < 1$.

Solution Q3. Pour atteindre l'optimalité, on peut chercher X telle que $\overline{\ell}(U) = H(X)$. L'entropie s'écrit $H(X) = \sum_{n\in\mathbb{N}} p_n \log_2 \frac{1}{p_n}$ tandis que la longueur moyenne est de la forme $\overline{\ell}(X) = \sum_{n\in\mathbb{N}} p_n(n+1)$. Il suffit donc de choisir p_n telle que $n+1 = \log_2 \frac{1}{p_n}$, autrement dit :

$$p_n = 2^{-(n+1)}\,.$$

On vérifie également que $\sum_{n\in\mathbb{N}} p_n = 1$ pour que $(p_n)_{n\in\mathbb{N}}$ soit bien la distribution d'une variable aléatoire. D'ailleurs, on remarque qu'une telle variable suit la loi géométrique de paramètre $\gamma = \frac{1}{2}$.

Exercice 4. ($\star\star$) Code gamma.

Dans cet exercice, on souhaite coder efficacement une source à valeur dans l'ensemble des entiers naturels, sans avoir d'information préalable sur les entiers émis par la source.

Usuellement, pour coder un entier $n \in \mathbb{N}$ sous forme de chaîne de bits, on décompose l'entier en base 2 :

$$n = \sum_{i=0}^{k} n_i 2^i\,,$$

et on retourne la séquence $B(n) = (n_0, \ldots, n_k) \in \{0,1\}^{k+1}$ de longueur $k+1$, où $k = \lfloor \log_2(n) \rfloor$ (excepté pour $n = 0$, pour lequel on a $k = 0$).

Question 1.– Le code $B : \mathbb{N} \to \{0,1\}^+$ est-il préfixe ? Est-il uniquement décodable ? Pourquoi ?

Le code Gamma, introduit par Elias, propose une solution au problème précédent. Avant d'encoder l'entier n sous la forme de sa décomposition en base 2, on précise à l'aide d'un code unaire la longueur de n. Ainsi, le code $\Gamma : \mathbb{N} \to \{0,1\}^+$ est défini par :

$$\Gamma(n) = \underbrace{0\ \cdots\cdots\cdots\ 0}_{\leftarrow\ 1+\lfloor \log_2(n) \rfloor\ \rightarrow} 1\, B(n) \quad \text{pour } n \geq 1,$$

et $\Gamma(0) = 10$.

Question 2.– Le code Γ est-il préfixe ?

Question 3.– Quelle est la longueur de $\Gamma(n)$ pour $n \in \mathbb{N}$? Donner une condition sur la distribution (p_n) de la source pour que la longueur moyenne du code Γ soit finie.

Question 4.– 💻 Calculer numériquement des valeurs approchées de la longueur moyenne du code Γ lorsque :

— X suit une loi uniforme sur $\{0, 1, \dots, N\}$ (et $p_X(n)$ est nulle pour $n \geq N$),
— X suit une loi géométrique,
— X suit une loi de Poisson.

Question 5.– $(\star\star\star)$ Le code Gamma peut être vu comme le codage unaire de la taille en bits de n, concaténé avec la représentation binaire de n. Peut-on itérer ce procédé pour obtenir un code encore plus court ? Quelle est la longueur du codage de l'entier n obtenu ?

Solutions de l'Exercice 4.

Solution Q1. Le code B n'est pas préfixe, car par exemple, $B(1) = 1$ est un préfixe de $B(2) = 11$. Il n'est pas non plus uniquement décodable, car

— par exemple, on peut voir que la chaîne de caractère 110 est l'image par B^+ des messages $(1,1,0)$, $(1,3)$, $(3,0)$ et 6,
— on observe aussi que la séquence de longueurs de B ne satisfait pas l'inégalité de Kraft : la quantité

$$\begin{aligned}\sum_{n=0}^{2^k-1} 2^{-\ell(B(n))} &= 2^{-\ell(B(0))} + \sum_{i=0}^{k-1}\sum_{j=0}^{2^i-1} 2^{-\ell(B(2^i+j))} \\ &= 1 + \sum_{i=0}^{k-1}\sum_{j=0}^{2^i-1} 2^{-(i+1)} \\ &= 1 + \sum_{i=0}^{k-1} \frac{1}{2} = \frac{k}{2} + 1\end{aligned}$$

diverge lorsque $k \to +\infty$.

Solution Q2. Soit n et m deux entiers, et supposons que $\Gamma(n)$ soit un préfixe de $\Gamma(m)$. Alors, $\Gamma(n)$ et $\Gamma(m)$ débutent par le même nombre de zéros, donc n et m sont encadrés par les mêmes puissances de 2. Autrement dit : $2^k \leq n, m \leq 2^{k+1} - 1$ où $k = \lfloor \log_2 n \rfloor$.

Puis, nécéssairement $B(n)$ est un préfixe de $B(m)$. Mais d'après ce qui précède, $B(n)$ et $B(m)$ doivent avoir la même longueur. Ceci induit que $B(n) = B(m)$, puis que $n = m$.

Solution Q3. Le mot $\Gamma(0)$ a longueur 2, et les mots $\Gamma(n)$ ont longueur $3 + 2\lfloor \log_2(n) \rfloor$ pour $n \geq 1$.

On a :

$$\overline{\ell}(\Gamma) = \sum_{n \in \mathbb{N}} p_n \, \ell(\Gamma(n)) = 2p_0 + \sum_{n \geq 1} p_n (3 + 2\lfloor \log_2 n \rfloor) = 3 - p_0 + 2 \sum_{n \geq 1} p_n \lfloor \log_2 n \rfloor \, .$$

De manière générale, la longueur moyenne de Γ est donc finie si et seulement si la série de terme général $p_n \lfloor \log_2 n \rfloor$ converge.

Solution Q4. À venir

Solution Q5. À venir