Devoir surveillé #6

Lundi 2026-06-01 ; durée : quatre heures

Compression entropique

Préliminaires

Ce sujet s'intéresse à la compression et à la décompression de mots sur un alphabet fini $\mathcal{S}$ de cardinal $|\mathcal{S}| \geq 2$. Plus précisément, on se donne une fonction q de $\mathcal{S}$ dans $\mathbb{N}$ et on s'intéresse aux mots (a priori *non vides*) de $\mathcal{S}^*$ tels que chaque lettre $\sigma \in \mathcal{S}$ a exactement $q(\sigma)$ occurrences dans ces mots. On note $\mathcal{S}^q$ l'ensemble de ces mots. Remarque : les mots de $\mathcal{S}^q$ ont pour longueur $N = \sum_{\sigma\in\mathcal{S}} q(\sigma)$.
Une lettre de $\mathcal{S}$ peut être représentée avec $\lceil \log_2 |\mathcal{S}| \rceil$ bits, où $\lceil x \rceil$ désigne la partie entière supérieure de x, c'est-à-dire l'entier tel que $\lceil x \rceil - 1 < x \leq \lceil x \rceil$. Si l'on concatène les bits représentant chacune des lettres d'un mot de $\mathcal{S}^q$, on peut donc représenter ce mot de façon non ambiguë à l'aide d'une séquence de bits de taille :

$$\sum_{\sigma\in\mathcal{S}} q(\sigma) \times \lceil \log_2 |\mathcal{S}| \rceil = \lceil \log_2 |\mathcal{S}| \rceil \times N.$$

Il s'agit d'une représentation non compressée. La théorie de l'information affirme qu'il faut au moins $\sum_{\sigma\in\mathcal{S}} q(\sigma) \times \log_2(N/q(\sigma))$ bits pour représenter tous les mots de $\mathcal{S}^q$ de façon non ambiguë. Ce sujet explore quelques façons de compresser les mots pour s'approcher de cette borne théorique.

La partie I s'intéresse à la fonction q, c'est-à-dire au comptage des lettres d'un mot. La partie II étudie les arbres binaires dont les feuilles sont étiquetées par des lettres de $\mathcal{S}$. La partie III utilise ces arbres pour (dé)compresser des mots de $\mathcal{S}^*$. La partie IV s'intéresse aux arbres qui donnent les meilleurs taux de compression pour les mots de $\mathcal{S}^q$. Certains de ces arbres ont une représentation compacte ; c'est l'objet de la partie V. D'autres arbres sont encore plus compacts, mais au prix d'un moindre taux de compression, comme le montre la partie VI. Finalement, la partie VII attaque le problème d'une façon complètement différente afin de s'approcher un peu plus du taux de compression théorique optimal. Les différentes parties sont indépendantes : il n'est pas nécessaire d'avoir répondu aux questions d'une partie pour répondre aux questions d'une autre partie ; cependant les notions abordées dans une partie sont souvent utiles aux parties suivantes.

Vous avez deux possibilités pour traiter ce sujet (chacune notée sur 20) :

— soit vous traitez **uniquement** les parties I à IV
— soit vous traitez l'intégralité du sujet

Dans les deux cas, vous devez indiquer clairement votre choix au début de votre copie.

OCaml

L'intégralité de ce sujet est à traiter en langage OCaml. On rappelle ici quelques fonctions de la bibliothèque standard :

— `Array.length t` renvoie la longueur du tableau `t`.
— `Array.make n v` crée un tableau de `n` cases qui sont toutes initialisées avec `v`.
— `Array.make_matrix n m v` crée un tableau à deux dimensions de `n` lignes et `m` colonnes dont toutes les cases sont initialisées avec `v`.
— `Array.of_list l` renvoie un tableau initialisé avec les valeurs de la liste `l`, tandis que `Array.to_list` réalise l'opération inverse.
— `Array.copy a` renvoie une copie du tableau `a`.
— La case numéro `i` du tableau `t` peut être accédée avec `t.(i)`. Les cases sont numérotées à partir de zéro.

— `List.rev l` renvoie le miroir de la liste `l` en argument, en temps linéaire en sa longueur.
— `List.map f x` renvoie la liste `[f x0; f x1 ; ...]` obtenue en appliquant la fonction `f` à chacun des éléments de la liste `x = [x0; x1; ...]`
— `failwith s` avec `s` une chaîne de caractères lève une exception `Failure s`.

I. Comptage d'occurrences

Un texte non compressé est un mot de $\mathcal{S}^*$. Dans la suite, à chaque fois qu'il s'agira de définir une fonction en OCaml, les lettres de $\mathcal{S}$ seront représentées par des entiers `int` et un mot de $\mathcal{S}^*$ sera représenté par un tableau d'entiers `int array` ou une liste d'entiers `int list` en fonction des questions. La première étape consiste à calculer le nombre d'occurrences $q(\sigma)$ de chaque lettre σ dans un mot.

1. Supposons $\mathcal{S} = [0, 255]$. Définissez une fonction OCaml `occurrences : int list -> int array` qui reçoit en argument un mot représenté par une liste d'entiers et renvoie le nombre d'occurrences $q(\sigma)$ de chaque lettre $\sigma \in \mathcal{S}$ sous forme d'un tableau $[\![q(0); q(1); \ldots; q(255)]\!]$.

Il peut être intéressant de ne considérer que le sous-ensemble des lettres qui ont au moins une occurrence. Dans la question suivante, on ne représentera pas q par $[\![q(0); q(1); \ldots; q(|\mathcal{S}|-1)]\!]$ mais par un tableau de paires $[\![(\sigma_1, q(\sigma_1)); (\sigma_2, q(\sigma_2)); \ldots]\!]$ avec $\{\sigma_1, \sigma_2, \ldots\} = \{\sigma \in \mathcal{S} \mid q(\sigma) \neq 0\}$ et $\sigma_1 < \sigma_2 < \ldots$

2. Définissez une fonction OCaml `nonzero_occurrences : int array -> (int * int) array` qui passe de la représentation de q de la question 1 (tableau $[\![q(0); q(1); \ldots; q(|\mathcal{S}|-1)]\!]$) à la représentation sous forme d'un tableau de paires. Cette fonction devra avoir une complexité temporelle en $O(|\mathcal{S}|)$.

II. Arbres binaires

Soit $\mathcal{T}_\mathcal{S}$ l'ensemble des arbres binaires dont les feuilles sont en bijection avec un ensemble fini $\mathcal{S}$. Autrement dit, un arbre de $\mathcal{T}_\mathcal{S}$ a exactement $|\mathcal{S}|$ feuilles ; chacune de ses feuilles est étiquetée par un élément de $\mathcal{S}$; toutes les feuilles sont étiquetées par des éléments différents de $\mathcal{S}$. Ces arbres peuvent être construits comme suit :

— Une feuille étiquetée par x est notée $F(x)$. C'est un arbre de $\mathcal{T}_{\{x\}}$.
— Si g et d sont des arbres appartenant respectivement à $\mathcal{T}_{\mathcal{S}_1}$ et $\mathcal{T}_{\mathcal{S}_2}$, alors l'arbre dont le sous-arbre gauche est g et le sous-arbre droit est d, noté $N(g, d)$, est un arbre de $\mathcal{T}_{\mathcal{S}_1 \cup \mathcal{S}_2}$ à condition que $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$.

Remarque : un tel arbre binaire possède exactement $|\mathcal{S}| - 1$ nœuds internes. Par ailleurs, on étiquettera implicitement les arêtes par 0 ou 1 suivant qu'elles mènent à un sous-arbre gauche (0) ou droit (1).

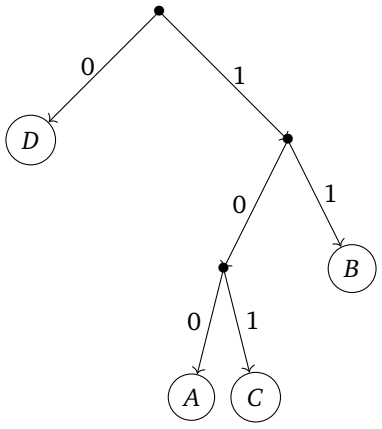

FIGURE 1 – Exemple d'arbre de $\mathcal{T}_{\{A,B,C,D\}}$

3. Montrez que l'ensemble $\mathcal{T}_{\{A,B,C,D\}}$ contient 120 arbres.

Le type OCaml utilisé pour représenter les arbres est le suivant :

```
type tree = F of int | N of tree * tree
```

Étant donnés un arbre t de $\mathcal{T}_\mathcal{S}$ et un élément σ de $\mathcal{S}$, on note $\ell_t(\sigma)$ la profondeur de la feuille étiquetée par σ, c'est à dire le nombre d'arêtes entre la racine de t et cette feuille. Par exemple, avec l'arbre de la Figure 1, on a $\ell_t(A) = 3$.
Comme précédemment, soit q une fonction de $\mathcal{S}$ dans $\mathbb{N}$. Pour tout arbre de $t \in \mathcal{T}_{\mathcal{S}'}$ avec $\mathcal{S}' \subseteq \mathcal{S}$, on note $c_q(t)$ la somme pondérée suivante :

$$c_q(t) = \sum_{\sigma \in \mathcal{S}'} q(\sigma)\ell_t(\sigma).$$

4. Supposons $\mathcal{S} = [0, n-1]$. Définissez une fonction OCaml `cq : tree -> int array -> int` qui reçoit deux arguments, un arbre $t \in \mathcal{T}_\mathcal{S}$ et un tableau $[\![q(0); q(1); ...; q(n-1)]\!]$ représentant q, et qui renvoie la valeur de $c_q(t)$. On cherchera à écrire une fonction efficace. Donnez et justifiez sa complexité temporelle.

Étant donné un arbre t de $\mathcal{T}_\mathcal{S}$, on représente une lettre $\sigma \in \mathcal{S}$ par la séquence des bits obtenus en parcourant t de la racine vers la feuille $F(\sigma)$. Par exemple, dans l'arbre de $\mathcal{T}_{\{A,B,C,D\}}$ de la Figure 1, D est représenté par 0 tandis que A est représenté par 100.

5. Définissez une fonction OCaml `get_path : int -> tree -> int list` qui reçoit en argument un entier $\sigma \in \mathcal{S}$ et un arbre $t \in \mathcal{T}_\mathcal{S}$ et qui renvoie la liste de 0 et 1 représentant σ dans t. Donnez et justifiez la complexité temporelle de cette fonction.

Considérons les séquences de bits définies de la façon suivante. Elles commencent par $k \geq 0$ bits valant 1, suivis d'un bit à 0, suivis de k bits arbitraires, ce que l'on notera $1^k0(0|1)^k$. Les dix premières séquences par ordre lexicographique sont donc 0, 100, 101, 11000, 11001, 11010, 11011, 1110000, 1110001, 1110010. L'arbre dont les branches constituent ces séquences et dont les feuilles sont étiquetées de gauche à droite par des entiers croissants commence comme illustré sur la Figure 2.

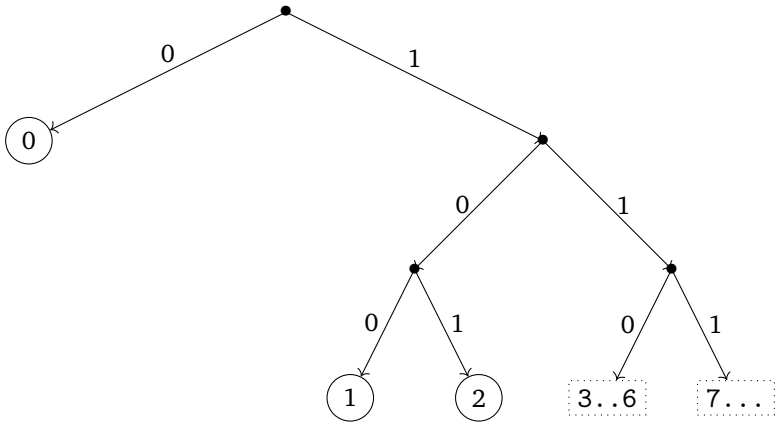

FIGURE 2 – Arbre des séquences $1^k0(0|1)^k$

Remarque : l'étiquette de chaque feuille correspond à l'indice de la séquence quand elles sont triées par ordre lexicographique. Les séquences croissant indéfiniment, l'arbre est a priori infini. C'est pourquoi la question suivante se limite aux séquences de bits $1^k0(0|1)^k$ pour lesquelles k n'excède pas une certaine borne ℓ, afin d'obtenir un arbre de profondeur bornée $2\ell + 1$.

6. Définissez une fonction OCaml `integers : int -> tree` qui prend un entier $\ell \geq 0$ en argument et renvoie l'arbre dont les branches sont les séquences de la forme $1^k0(0|1)^k$ avec $k \leq \ell$, ainsi que la séquence $1^{\ell+1}$. Les feuilles seront étiquetées de gauche à droite par les entiers 0, 1, 2, etc.

III. Codes préfixes

Étant donnés un alphabet $\mathcal{S}$ et un arbre t de $\mathcal{T}_\mathcal{S}$, un mot de $\mathcal{S}^*$ peut être représenté par la concaténation des séquences de bits représentant chacune de ses lettres dans t, de la gauche vers la droite. Supposons que l'arbre t est l'exemple de la Figure 1. Le mot « ADBDCD » est alors représenté par la séquence de bits

100|0|11|0|101|0. Remarque : les barres verticales ne servent qu'à rendre l'exemple plus lisible ; elles ne font pas partie de la séquence de bits et ne sont pas nécessaires pour retrouver le mot initial.

7. Soit t un arbre de $\mathcal{T}_\mathcal{S}$. Étant donnée une séquence de bits, montrez qu'il existe au plus un mot de $\mathcal{S}^*$ qui est représenté par cette séquence. On justifiera *soigneusement*.
8. Supposons $\mathcal{S} \subseteq \mathbb{N}$. Définissez une fonction OCaml `decomp1 : int list -> tree -> int list` qui reçoit en argument une liste non vide d'entiers 0 ou 1 et un arbre `t` de $\mathcal{T}_\mathcal{S}$ et qui renvoie la liste des éléments de $\mathcal{S}$ représentée par cette liste de bits. On sera attentif au cas où la liste en argument ne représente aucun mot (non vide) de $\mathcal{S}^*$. On a aussi toujours que $|\mathcal{S}| \geq 2$, et donc que l'argument `t` est de la forme `N (t1, t2)` pour certains arbres `t1`, `t2`. Donnez et justifiez la complexité temporelle de cette fonction.

Cette fonction est dite de *décompression*. Supposons que $\mathcal{S}$ est $\{A, B, C, D\}$ et que la fonction q donne le nombre d'occurrences de chaque lettre dans le mot « ADBDCD », par exemple $q(D) = 3$. Dans ce cas, $c_q(t)$ vaut 11 pour l'arbre exemple de la Figure 1. Il s'agit de la longueur de la séquence 10001101010 représentant le mot « ADBDCD ». Plus généralement, $c_q(t)$ est la longueur de n'importe quelle séquence représentant un mot de $\mathcal{S}^q$ compressé avec l'arbre t. Il s'avère qu'il n'existe aucun arbre $t' \in \mathcal{T}_\mathcal{S}$ tel que $c_q(t') < 11$; t est donc optimal.
Étant donnée une fonction q, l'objectif de la partie suivante sera de construire un des arbres de $\mathcal{T}_\mathcal{S}$ offrant la meilleure compression des mots de $\mathcal{S}^q$, c'est-à-dire un arbre qui minimise c_q.

IV. Arbres optimaux

Soit q une fonction dont le domaine inclut $\mathcal{S}$. Un arbre de $\mathcal{T}_\mathcal{S}$ est dit optimal pour $(q, \mathcal{S})$ s'il minimise c_q parmi tous les arbres de $\mathcal{T}_\mathcal{S}$. Autrement dit, un arbre optimal $t \in \mathcal{T}_\mathcal{S}$ vérifie $c_q(t) = \min_{t' \in \mathcal{T}_\mathcal{S}} c_q(t')$. De tels arbres optimaux existent et, dès lors que $|\mathcal{S}| \geq 2$, ils ne sont pas uniques.

9. Soit $t = N(g, d) \in \mathcal{T}_\mathcal{S}$ un arbre optimal pour $(q, \mathcal{S})$. Soit $\mathcal{S}_g$ l'ensemble des lettres qui étiquettent les feuilles du sous-arbre g. Montrez que g est un arbre de $\mathcal{T}_{\mathcal{S}_g}$ optimal pour $(q, \mathcal{S}_g)$.

Malheureusement, cette propriété ne permet pas d'en déduire un algorithme de type « diviser pour régner » pour trouver l'arbre optimal. En effet, il faudrait que l'algorithme puisse efficacement deviner comment partitionner $\mathcal{S}$ entre les feuilles du sous-arbre gauche et celles du sous-arbre droit.

10. Supposons qu'il existe une lettre $\sigma_0 \in \mathcal{S}$ telle que $q(\sigma_0) > \sum_{\sigma \in \mathcal{S} \setminus \{\sigma_0\}} q(\sigma)$. Montrez que, pour tout arbre de $\mathcal{T}_\mathcal{S}$ optimal pour $(q, \mathcal{S})$, la feuille $F(\sigma_0)$ est à profondeur 1, c'est-à-dire directement attachée à la racine.
11. Soient σ_1 et σ_2 deux éléments différents de $\mathcal{S}$ tels que, pour tout $\sigma \in \mathcal{S} \setminus \{\sigma_1, \sigma_2\}$, on a $q(\sigma) \geq q(\sigma_1)$ et $q(\sigma) \geq q(\sigma_2)$. Soit $\mathcal{S}' = \mathcal{S} \setminus \{\sigma_1, \sigma_2\} \cup \{\sigma_3\}$ avec σ_3 un tout nouvel élément. On définit q' de telle sorte que $q'(\sigma_3) = q(\sigma_1) + q(\sigma_2)$ et $\forall \sigma \in S \setminus \{\sigma_1, \sigma_2\}$, $q'(\sigma) = q(\sigma)$.
    i. Montrez qu'il existe un arbre t de $\mathcal{T}_\mathcal{S}$ optimal pour $(q, \mathcal{S})$ ayant un nœud $N(F(\sigma_1), F(\sigma_2))$.
    ii. Soit un arbre t' de $\mathcal{T}'_\mathcal{S}$ optimal pour $(q', \mathcal{S}')$. Montrez que l'arbre $t \in \mathcal{T}_\mathcal{S}$ obtenu en remplaçant dans t' la feuille $F(\sigma_3)$ par $N(F(\sigma_1), F(\sigma_2))$ est optimal pour $(q, \mathcal{S})$.

On définit la fonction $\bar{q}$ en étendant la fonction q aux arbres de la façon suivante. Pour une feuille, $\bar{q}(F(\sigma))$ vaut $q(\sigma)$. Pour un nœud interne, $\bar{q}(N(g, d))$ vaut récursivement $\bar{q}(g) + \bar{q}(d)$. Remarque : en général, $c_q(t) \neq \bar{q}(t)$.

La question précédente montre qu'il est possible de construire un arbre optimal pour $(q, \mathcal{S})$ à l'aide de l'algorithme suivant. On initialise un ensemble d'arbres E avec toutes les feuilles $F(\sigma)$ avec $\sigma \in \mathcal{S}$. À chaque étape, on retire de E deux arbres t_1 et t_2 qui minimisent $\bar{q}$; puis on ajoute à E l'arbre $N(t_1, t_2)$. On répète cette procédure jusqu'à ce qu'il ne reste qu'un seul arbre dans E. Il s'agit alors d'un arbre optimal pour $(q, \mathcal{S})$.

12. Implantez l'algorithme décrit ci-dessus en définissant les deux fonctions OCaml suivantes : `insert` et `optimal`.
    i. La fonction `insert : (int * tree) -> (int * tree) list -> (int * tree) list` insère dans une liste son premier argument. La liste en argument est supposée triée par rapport à la première composante de chacune de ses paires. La liste renvoyée doit l'être aussi.
    ii. La fonction `optimal : (int * int) list -> tree` renvoie un arbre optimal. La liste passée en argument est de taille $|\mathcal{S}|$; elle contient toutes les paires $(\sigma, q(\sigma))$ pour $\sigma \in \mathcal{S}$; ces paires sont triées par valeur croissante de $q(\sigma)$.

iii. Donnez et justifiez la complexité temporelle de la fonction `optimal`.

Les deux questions suivantes montrent que, quand la liste initiale des $(\sigma, q(\sigma))$ est triée par valeur croissante de $q(\sigma)$, de simples listes et/ou tableaux suffisent pour écrire une fonction `optimal` dont la complexité temporelle est en $O(|\mathcal{S}|)$.

13. Montrez que, lors de l'exécution de l'algorithme décrit ci-dessus, les arbres $t = N(t_1, t_2)$ sont ajoutés à l'ensemble E par valeur croissante de $\bar{q}(t)$.

14. Expliquez comment écrire la fonction `optimal` pour que sa complexité temporelle soit linéaire. Le code OCaml n'est pas demandé.

Cette partie a permis de calculer un arbre qui minimise la longueur c_q de la séquence de bits représentant un mot de $\mathcal{S}^q$. Mais pour décompresser ce mot, il faut disposer de l'arbre et donc l'avoir stocké quelque part. Parmi tous les arbres optimaux, il est donc important d'en choisir un qui nécessitera le moins de bits pour être représenté ; c'est l'objet de la partie suivante.

V. Arbres canoniques

On suppose maintenant qu'il existe un ordre alphabétique noté $<$ sur les éléments de $\mathcal{S}$. Pour $\mathcal{S} \subseteq \mathbb{N}$, il s'agira de l'ordre usuel sur $\mathbb{N}$. Soit t un arbre de $\mathcal{T}_\mathcal{S}$. On définit la relation $<_t$ entre éléments de $\mathcal{S}$ de la façon suivante : pour toute paire $(\sigma_1, \sigma_2) \in \mathcal{S}^2$,

$$\sigma_1 <_t \sigma_2 \Leftrightarrow \ell_t(\sigma_1) < \ell_t(\sigma_2) \text{ ou } (\ell_t(\sigma_1) = \ell_t(\sigma_2) \text{ et } \sigma_1 < \sigma_2).$$

L'arbre t est dit canonique si, en parcourant les feuilles de la gauche vers la droite, leurs étiquettes respectent la relation $<_t$. Autrement dit, plus une feuille est à gauche, plus elle est proche de la racine ; et les étiquettes des feuilles à une même profondeur sont triées de gauche à droite par ordre alphabétique.
L'arbre de la Figure 1 vérifie bien la deuxième partie de la propriété : les feuilles à une même profondeur sont triées de gauche à droite par ordre alphabétique. Mais il ne respecte pas la première partie : la feuille étiquetée par A est plus profonde que celle étiquetée par B alors qu'elle est plus à gauche. Cet arbre n'est donc pas canonique. L'arbre de la Figure 3 suivante est par contre canonique :

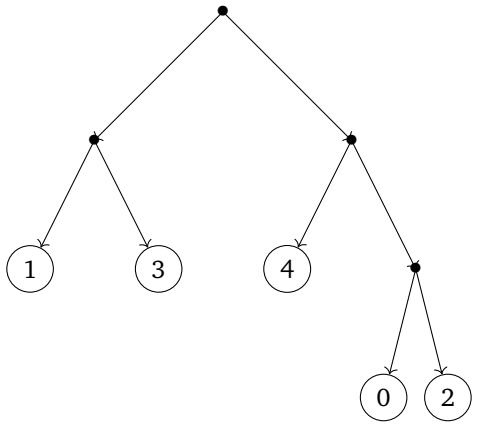

FIGURE 3 – Exemple d'arbre canonique

Un arbre canonique peut être représenté par deux tableaux d'entiers. Le premier tableau associe à chaque indice i le nombre de lettres $\sigma \in \mathcal{S}$ telles que $\ell_t(\sigma) = i$. Le deuxième tableau contient les éléments de $\mathcal{S}$ obtenus en parcourant les feuilles de l'arbre de gauche à droite. Ainsi, l'arbre de $\mathcal{T}_{[0,4]}$ de la Figure 3 est représenté par les deux tableaux $[\![0; 0; 3; 2]\!]$ et $[\![1; 3; 4; 0; 2]\!]$.

15. Définissez une fonction OCaml `canonical : int array -> int array -> tree` qui, étant donnés deux tels tableaux, renvoie l'arbre canonique qu'ils représentent, s'il existe. Donnez et justifiez la complexité temporelle de cette fonction.

VI. Arbres alphabétiques

L'arbre de la Figure 1 était optimal pour le mot « ADBDCD ». Un autre arbre optimal pour ce mot est celui de la Figure 4.

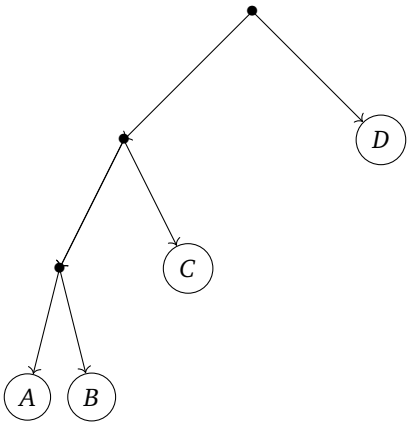

FIGURE 4 – Arbre à la fois optimal pour le mot « ADBDCD » et alphabétique

Il n'est pas canonique mais il a une propriété très intéressante : ses feuilles sont dans l'ordre alphabétique. Cela signifie que lors du stockage de l'arbre, il n'est pas nécessaire de stocker les étiquettes des feuilles, il suffit de stocker la forme de l'arbre. Un tel arbre est dit *alphabétique*.
Plus généralement, un arbre de $\mathcal{T}_\mathcal{S}$ est *alphabétique* si, en parcourant ses feuilles de gauche à droite, les étiquettes apparaissent dans l'ordre alphabétique. On note $\mathcal{A}_\mathcal{S}$ l'ensemble de ces arbres. On dit d'un arbre qu'il est *alphabétique-optimal* pour $(q, \mathcal{S})$ s'il minimise c_q parmi tous les arbres de $\mathcal{A}_\mathcal{S}$. Remarque : un arbre alphabétique-optimal pour $(q, \mathcal{S})$ n'est pas nécessairement optimal pour $(q, \mathcal{S})$.
Soit $n \in \mathbb{N}$ et $\mathcal{S} = [0, n-1]$. On se donne une fonction $q : \mathcal{S} \to \mathbb{N}$. Étant donnés deux entiers i et j tels que $0 \leq i \leq j < n$, on note $m_{i,j}$ la valeur de $c_q(t)$ pour n'importe quel arbre t de $\mathcal{A}_{[i,j]}$ alphabétique-optimal pour $(q, [i, j])$. En particulier, si t est un arbre de $\mathcal{A}_\mathcal{S}$. alphabétique-optimal pour $(q, \mathcal{S})$, il vérifie $c_q(t) = m_{0,n-1}$.

16. Exprimez la valeur de $m_{i,j}$ en fonction des valeurs de $m_{i',j'}$ avec $[i', j'] \subset [i, j]$. Déduisez en une fonction OCaml `alpha_optimal : int array -> tree` qui calcule un arbre de $\mathcal{A}_{[0,n-1]}$ alphabétique-optimal pour $(q, [0, n-1])$. La fonction q est donnée par le tableau de taille n passé en argument. La complexité temporelle de la fonction devra être en $O(n^3)$.

Un arbre alphabétique de $\mathcal{A}_{[0,n-1]}$ peut être représenté par la séquence de $2n-1$ bits obtenue en effectuant un parcours en profondeur préfixe, c'est-à-dire que la racine d'un sous-arbre est visité avant ses feuilles et qu'un sous-arbre gauche est parcouru avant un sous-arbre droit. Pour chaque nœud, les chiffres 0 et 1 indiquent respectivement s'il s'agit d'un nœud interne ou d'une feuille. Par exemple, l'arbre de la Figure 4 est représenté par la séquence 0001111.

17. Définissez une fonction OCaml `alpha : int array -> tree` qui reçoit en argument un tableau de 0 et 1 et renvoie l'arbre alphabétique correspondant. On sera attentif au cas où le tableau en argument ne représente aucun arbre. Cette fonction devra avoir une complexité temporelle en $O(n)$ avec n la taille du tableau.

VII. Codes arithmétiques

Soient un alphabet $\mathcal{S}$ et une fonction $q : \mathcal{S} \to \mathbb{N}$. Comme précédemment, $\mathcal{S}^q$ est le sous-ensemble des mots de $\mathcal{S}^*$ qui contiennent exactement $q(\sigma)$ occurrences de σ pour chaque $\sigma \in S$.

18. Supposons que $\mathcal{S}$ est $\{A, B, C\}$ et que q satisfait $q(A) = 15$, $q(B) = 4$, $q(C) = 1$.

i. Combien de bits faut-il pour représenter un mot de $\mathcal{S}^q$ en utilisant un arbre optimal pour $(q, \mathcal{S})$?
ii. Combien y a-t-il de mots dans $\mathcal{S}^q$? On pourra exprimer ce nombre sous forme d'un produit de nombres premiers.

iii. Montrez que 17 bits suffisent pour représenter chaque entier entre 0 et $|\mathcal{S}^q| - 1$ (et donc n'importe quel mot de $\mathcal{S}^q$). Remarque : $15 \times 17 = 2^8 - 1$.

L'exemple de la question précédente montre que, dans certains cas, par exemple quand une lettre est bien plus fréquente que les autres, l'utilisation d'un code préfixe n'est pas l'approche optimale. Il faut alors se tourner vers d'autres mécanismes de compression.
On se restreint maintenant au cas où $\mathcal{S} = [0, n-1]$. Soit $N = \sum_{\sigma \in \mathcal{S}} q(\sigma)$. La fonction $E_q : \mathbb{N} \times \mathcal{S} \to \mathbb{N}$ est définie de la façon suivante pour $q(\sigma) \neq 0$:

$$E_q(x, \sigma) = \lfloor x/q(\sigma) \rfloor \times N + (x \mod q(\sigma)) + \sum_{0 \leq k < \sigma} q(k),$$

où $\lfloor a \rfloor$ désigne la partie entière de a.
La fonction E_q peut être étendue en une fonction $C_q : \mathbb{N} \times \mathcal{S}^* \to \mathbb{N}$ qui travaille sur les mots de la façon suivante :

$$C_q(x, \sigma_0 \sigma_1 \cdots \sigma_k) = E_q(\ldots E_q(E_q(x, \sigma_0), \sigma_1), \ldots \sigma_k).$$

Ainsi, après avoir choisi x arbitrairement, un mot $\sigma_0 \cdots \sigma_{N-1} \in \mathcal{S}^q$ peut être compressé en un entier $y = C_q(x, \sigma_0 \cdots \sigma_{N-1})$. Il est alors possible de retrouver le mot original à partir de y en calculant progressivement σ_{N-1}, σ_{N-2}, etc, jusqu'à σ_0.
Considérons par exemple le mot « 0120000 » ($N = 7$, $q = [\![5, 1, 1]\!]$). Si l'on part de $x = 0$, sa version compressée sera l'entier 151 :

$$0 \xrightarrow{0} 0 \xrightarrow{1} 5 \xrightarrow{2} 41 \xrightarrow{0} 57 \xrightarrow{0} 79 \xrightarrow{0} 109 \xrightarrow{0} 151$$

où $x \xrightarrow{\sigma} y$ signifie $E_q(x, \sigma) = y$.

19. Définissez une fonction OCaml `decomp2 : int array -> int -> int -> int * int array` qui reçoit en argument un tableau représentant q et deux entiers y et k, et renvoie un entier x et un tableau $[\![\sigma_0, \sigma_1, \ldots, \sigma_{k-1}]\!]$ tels que $C_q(x, \sigma_0 \sigma_1 \cdots \sigma_{k-1}) = y$

Le nombre de bits de l'entier $C_q(x, \sigma_0 \ldots \sigma N - 1)$ est proche de l'optimum théorique $\sum_{\sigma \in S} q(\sigma) \times \log_2(N/q(\sigma))$ mais un problème se pose en pratique. Sauf pour de tout petits mots sur de tout petits alphabets, le calcul de $C_q(x, \sigma_0 \cdots \sigma_{N-1})$ va nécessiter de manipuler des entiers très grands. Pour éviter ce problème, une solution consiste, avant chaque appel à E_q, à se débarrasser d'un certain nombre k_i de bits de poids faible. Plus précisément, on construit les suites (x_n) et (y_n) suivantes :

$$\begin{cases} y_i = & \lfloor x_i / 2^{k_i} \rfloor \\ x_{i+1} = & E_q(y_i, \sigma_i) \end{cases}$$

Comme précédemment, la valeur de x_0 est fixée arbitrairement. Le mot $\sigma_0 \cdots \sigma_{N-1}$ peut alors être représenté par d'une part x_N et d'autre part la concaténation des séquences de bits suivantes, les bits de poids forts apparaissant en premier :

$$\underbrace{x_{N-1} \mod 2^{k_{N-1}}}_{k_{N-1}\text{bits}}; \ \ldots; \ \underbrace{x_2 \mod 2^{k_2}}_{k_2\text{bits}}; \ \underbrace{x_1 \mod 2^{k_1}}_{k_1\text{bits}}.$$

Pour chaque i en ordre décroissant, le décompresseur déduit de x_{i+1} les valeurs de y_i et σ_i (par exemple avec `decomp2`, question 19, avec $k = 1$), puis il lit k_i bits dans le mot compressé et les concatène à y_i pour obtenir x_i, et ainsi de suite jusqu'à avoir décompressé tout le mot.
Dans la suite, on supposera que N est une puissance de deux : $N = 2^K$. On supposera par ailleurs que le processeur n'est efficace que pour des entiers ne dépassant pas 2^B pour un certain B bien plus grand que K. En particulier, $x_{i+1} = E_q(y_i, \sigma_i)$ ne doit jamais atteindre cette borne 2^B lors de la compression.
Pour que le taux de compression se rapproche de l'optimum théorique, il faut que chaque k_i soit le plus petit possible (idéalement 0) et donc que chaque y_i soit le plus grand possible. Par ailleurs, les différents k_i ne font pas partie du mot compressé. Autrement dit, en ne connaissant que y_i, le décompresseur doit pouvoir deviner le nombre k_i de bits qui doivent être lus dans le mot compressé pour reconstruire x_i. Une solution correcte mais mauvaise serait, par exemple, de fixer tous les k_i à la constante K ; le décompresseur pourrait alors deviner trivialement les k_i mais le mot résultant ne serait absolument pas compressé.

20. Proposez une façon pour le compresseur de choisir k_i en fonction de x_i et σ_i. Expliquez comment le décompresseur choisit k_i en fonction de y_i et prouvez que ce choix correspond bien à celui du compresseur. Si cela a une importance, expliquez comment le compresseur doit choisir x_0.

Remarque : la contrainte imposant que N soit une puissance de deux ne pose aucune difficulté en pratique. En effet, en matière de compression entropique, la seule chose qui importe est que $q(\sigma)/N$ soit proche de la proportion de la lettre σ dans le mot à compresser.

Fin du sujet