

Un peu de théorie de l'information

On considère un alphabet fini Σ de taille $K \geq 2$. On rappelle qu'un mot $u \in \Sigma^*$ est préfixe d'un mot $w \in \Sigma^*$ s'il existe $u' \in \Sigma^*$ tel que $uu' = w$. On dit qu'un langage $\Gamma \subseteq \Sigma^*$ est sans préfixe si aucun mot de Γ n'est préfixe d'un autre mot de Γ .

1. Donnez un ensemble sans préfixe sur $\Sigma = \{a, b, c\}$ contenant au moins 6 mots. Donnez un ensemble sans préfixe infini.
2. Soit $\Gamma \subseteq \Sigma^*$ un ensemble de mots sans préfixe ne contenant pas le mot vide. Montrez que Γ possède la propriété de décomposabilité unique : si on a $w = u_1 \dots u_n = v_1 \dots v_m$ avec tous les u_i, v_j dans Γ , les deux familles sont égales.
3. Étant donné un ensemble fini de mots, donnez un algorithme naïf pour tester si Γ est sans préfixe.

Un arbre K -aire est un arbre dont chaque noeud a au plus K fils. Une feuille est un noeud de l'arbre sans fils. On appelle hauteur d'un noeud sa distance à la racine, celle-ci étant de hauteur 0.

4. Soit T un arbre K -aire fini. Montrez que :

$$\sum_{\sigma \text{ feuille de } T} K^{-h(\sigma)} \leq 1$$

Dans quel cas a-t-on égalité ?

5. Montrer donc que si Γ est fini sans préfixe, alors :

$$\sum_{w \in \Gamma} K^{-|w|} \leq 1$$

On appelle cette inégalité l'inégalité de Kraft. Est-ce que l'inégalité de Kraft reste vraie pour les ensembles infinis (sans préfixe ?)

6. Donnez un algorithme en temps linéaire pour tester si un ensemble de mots finis est sans préfixe.

On va donner une preuve algorithmique de la réciproque ; si (l_i) est une famille d'entiers telle que $\sum_{i=1}^n K^{-l_i} \leq 1$, alors il existe une famille de mots sans préfixe dont les tailles sont les l_i .

7. Donner un algorithme prenant en entrée une telle suite l_i et renvoyant une telle suite u_i . On demande de plus que l'algorithme soit *en ligne* c'est-à-dire que l'algorithme lit les l_i un par un et produit u_i sans attendre les l_j . On pourra utiliser un invariant faisant intervenir à l'étape $t + 1$ l'écriture en base K de $1 - \sum_{i=1}^t K^{-l_i}$.
8. Montrer que l'inégalité de Kraft reste vraie pour tout ensemble de mots Γ vérifiant la propriété de décomposabilité unique de la question 2. On pourra considérer la série entière $\sum_n |\Gamma \cap \Sigma^n| x^n$ et ses puissances pour de grands entiers.

Solution: Un peu de théorie de l'information

1. $\Gamma = \{a^n b \mid b \in \mathbb{N}\}$.
2. On procède par récurrence sur la longueur de w . On regarde u_1 et v_1 . On sait qu'aucun n'est préfixe stricts de l'autre mais qu'ils sont préfixes, et sont donc égaux.
3. On teste pour toute paire de mots de Γ s'ils sont préfixe l'un de l'autre. La comparaison se fait en $\mathcal{O}(|u| + |v|)$ pour une complexité totale en $\mathcal{O}(n^2)$ pour n la taille mémoire de l'entrée Γ .
4. Par récurrence sur la hauteur h de l'arbre, on montre en même temps l'inégalité et le fait qu'on a égalité si et seulement si l'arbre est localement complet. Pour $h = 0$ on a le résultat. Pour $h + 1$ on observe que :

$$f(T) = \frac{1}{K} \sum_i f(T'_i)$$

la somme étant prise sur les sous-arbres non-vides de T . Par récurrence on a bien la propriété, puisque l'égalité oblige les T'_i à être localement complets.

5. Il suffit pour cela d'identifier un chemin K -aire dans un arbre à un code sur K symboles. On a égalité quand Γ est maximal, et on montre aisément que c'est le cas si et seulement si son arbre T_Γ est localement parfait.
6. Il suffit de placer les mots un par un dans un arbre K -aire comme précédemment. L'algorithme est donc clairement en temps linéaire.
7. Voir https://diplome.di.ens.fr/informatique-ens/annales/2022_InfoU-exercices.pdf

Nombres de Mots

Pour L un langage rationnel, on note l_n le nombre de mots de L de longueur n .

1. Calculer la suite l_n dans le cas où L est l'ensemble des mots sur $\{a, b\}$ ne contenant pas deux a consécutifs.
2. Donner un automate fini déterministe reconnaissant ce langage.

La série génératrice de la suite des longueurs est la série $f_L(z) = \sum_{n=0}^{\infty} l_n z^n$.

3. Calculer f_L pour le langage précédent. On se souviendra que $\varphi = \frac{1+\sqrt{5}}{2}$ vérifie $\varphi^2 - \varphi - 1 = 0$.
4. Montrer que la série génératrice des longueurs est une fraction rationnelle dans le cas où L est rationnel.
5. Montrer que la série génératrice des mots du langage généré par la grammaire $S \rightarrow aSbS \mid \varepsilon$ n'est pas rationnelle. On notera qu'elle est algébrique.

Solution: Nombres de Mots

1. C'est la suite des nombres de Fibonacci.
2. $f_L(z) = \sum_{n=0}^{\infty} F_n z^n = \frac{1+z}{1-z-z^2}$
3. On procède aisément par induction, en remarquant que la concaténation est le produit de Cauchy, l'union est l'addition, le complémentaire est la soustraction à $|\Sigma|^n z^n$. Pour l'étoile

Génération Aléatoire de Mots

On veut démontrer le théorème suivant :

Théorème 0.1 — Loi de Zipf-Mandelbrot La probabilité d'un mot dans une langue est équivalente à $\frac{C}{(r(X)+B)^\alpha}$ pour certaines constantes α, B, C .

Ici $r(X)$ désigne le rang du mot X , rangés par fréquence décroissante.

On considère donc un alphabet fini Σ de taille $S + 1$. On suppose que Σ contient un certain symbole **EOW** qui marque la fin d'un mot. On va tirer uniformément au hasard des lettres dans Σ pour former des mots, ces mots étant séparés par **EOW**. Soit X une variable aléatoire représentant la distribution des mots dans un texte aléatoire.

1. Donner $\mathbb{P}(X = w)$ pour $w \in \Sigma^*$. En déduire la probabilité que w soit de longueur n .
2. Calculer la fonction $n \mapsto r(n)$ qui a un mot de longueur n associe ses rangs possibles.
3. En déduire le théorème.

On change de paradigme, on considère un texte généré aléatoirement dont on ne connaît pas la distribution.

4. Donner une fonction qui permette de mesurer l'information portée par un mot de probabilité p dans un texte. On vérifiera que la fonction I est décroissante, vaut 0 en 1 et vérifie $I(p_1 p_2) = I(p_1) + I(p_2)$.
5. Donnez une mesure de la longueur moyenne d'un mot pour un code optimal associé à la distribution mesurée.

Solution: Génération Aléatoire de Mots

Consider a finite alphabet Σ of size $S + 1$. We will suppose that Σ contains a symbol **EOW** that marks the end of a word. Suppose we take letters uniformly at random in Σ to make words, and that we say that words are separated by **EOW**. Let X be a random variable representing the distribution of words in a random text : Then the probability of a given word w in a text is stated by (as it needs to start with **EOW** and end with **EOW**) :

$$\mathbb{P}(X = w) = \frac{1}{S(S+1)^{|w|}}$$

and thus the probability of a word being of length n is :

$$\mathbb{P}(|X| = n) = \sum_{w \in \Sigma^*, |w|=n} \mathbb{P}(X = w) = \frac{S^{n-1}}{(S+1)^n}$$

If we represent the rank of a word, we see that since its probability only depends on its length, its rank cannot be fully determined, just as a function $r(n)$ of its length :

$$\sum_{i=1}^{n-1} S^i < r(n) \leq \sum_{i=1}^n S^i$$

that is :

$$\frac{S}{S-1} (S^{n-1} - 1) < r(n) \leq \frac{M}{M-1} (M^n - 1)$$

Taking the log :

$$n - 1 < \log_S \left(\frac{S-1}{S} r(n) + 1 \right) \leq n$$

Raising each $1/S + 1$ to the power of this :

$$\frac{1}{(S+1)^{n-1}} > \left(\frac{1}{S+1} \right)^{\log_S \left(\frac{S-1}{S} r(n) + 1 \right)} \geq \frac{1}{(S+1)^n}$$

or :

$$\mathbb{P}(|X| = n) < \frac{C}{(r(n) + B)^\alpha} \leq \mathbb{P}(|X| = n - 1)$$

where :

$$\alpha = \frac{\log(S+1)}{\log S}, B = \frac{S}{S-1} \text{ and } C = \frac{M^{\alpha-1}}{(M-1)^\alpha}$$