le cnam

Codes et Automates finis

Cours 9 Lemme d'Arden

- Introduction

On se donne un alphabet A, c'est à dire un ensemble fini non vide. Le "mot sans lettre" ou "mot vide" est noté ε. Un langage L sur un alphabet A est un ensemble de mots qui utilisent des lettres de l'alphabet A et aussi éventuellement le mot sans lettre. L'étoile L^* réunit la suite infinie des langages L^i pour tous les entiers i : $L^* = \cup_{i\geq 0} L^i = \{\varepsilon\} \cup L \cup L^2 \cup L^3 \cup \ldots$.
On rappelle que $L^* = LL^* \cup \{\varepsilon\}$.

- Lemme d'Arden (première version) [Dean Arden, USA, 1925-2018]

On se donne deux langages K et L sur l'alphabet A : $K \subset A^*$, $L \subset A^*$. On suppose de plus que le langage K ne contient pas le mot sans lettre : $\varepsilon \notin K$. On cherche un langage $X \subset A^*$ solution de l'équation $X = K.X + L$. Le langage X satisfait donc aux deux conditions suivantes : $L \subset X$ et $\forall k \in K$, $\forall x \in X$, $k.x \in X$. Le lemme d'Arden énonce qu'il existe un et un seul langage X qui satisfait ces conditions : $X = K^* L$; le langage $X = K^* L$ est l'unique solution de l'équation $X = K.X + L$.

- Preuve du théorème

Montrons d'abord que si un langage X est solution de l'équation $X = K.X + L$, alors le langage $K^* L$ est inclus dans X : $K^* L \subset X$.
En effet, considérons $Y \subset X$. Alors $K.Y \subset X$ car $K.Y \subset K.X \subset X$. De même, on a la chaîne d'inclusions $K.(KY) \subset K.X \subset X$, donc $K^2 Y \subset X$.
Montrons par récurrence pour tout entier $n \geq 0$, $K^n Y \subset X$. Tout d'abord, la propriété est vraie à l'ordre $n = 0, 1, 2$. Ensuite, si elle est vraie à l'ordre n, alors $K^{n+1} Y = K.(K^n Y) \subset K.X \subset X$ et $K^{n+1} Y \subset X$, ce qui montre que la propriété est vraie à l'ordre suivant et achève la preuve par récurrence.
On a donc $\cup_{n\geq 0} K^n Y = K^* Y \subset X$. Ceci est vrai en particulier pour le langage $Y = L \subset X$ et on a l'inclusion $K^* L \subset X$.
Montrons maintenant que si X est solution de l'équation $X = KX + L$, alors $X = K^* L$. On sait déjà que $K^* L \subset X$. Montrons que si il existe au moins un mot dans la différence $X \setminus (K^* L)$, alors on aboutit à une contradiction.
Nous introduisons $w \in X \setminus (K^* L)$ de longueur minimale. On a donc $w \in X$ et $w \notin K^* L$. Comme $w \notin L \subset K^* L$, on en déduit que $w \in KX$ puisque $X = KX \cup L$. Donc il existe $k \in K$ et $x \in X$ de sorte que $w = kx$. Comme k ne peut pas être le mot sans lettre puique $\varepsilon \notin K$, ceci montre que la longueur $|w|$ du mot w est strictement positive, c'est à dire que w ne peut pas être égal au mot sans lettre. Le mot $x \in X$ est de longueur strictement inférieure à $|w|$ puisque $|w| = |x| + |k|$ et $|k| \geq 1$. Donc x ne peut appartenir à l'ensemble $X \setminus (K^* L)$ puisque nous avons choisi w de

longueur minimale. Donc $x \in K^*L$ et $w = kx \in K^*L$. Cette appartenance est une négation de l'hypothèse $w \notin K^*L$ et la contradiction est établie. Donc $X \setminus (K^*L) = \emptyset$ et $X = K^*L$.
Montrons maintenant que $X = K^*L$ satisfait bien à l'équation $X = K.X + L$. On part de la relation $K^* = K.K^* + \varepsilon$ et on effectue le produit de concaténation à droite avec le langage L. On obtient : $K^*L = (K.K^* + \varepsilon).L = (K.K^*).L + (\varepsilon.L)$ compte tenu de la distributivité de la concaténation relativement à l'addition. Or $(K.K^*).L = K.(K^*L)$ car la concaténation des langages est associative. De plus, ε est élément neutre pour la concaténation, donc $\varepsilon.L = L$. On en déduit $K^*L = K.(K^*L) + L$ ce qui établit que le langage $X = K^*L$ est effectivement solution de l'équation $X = K.X + L$. □

- Lien avec l'Analyse mathématique

Si on remplace formellement les langages par des nombres réels, notés usuellement avec des lettres minuscules, qu'on se donne $k \neq 1$ et $\ell \in \mathbb{R}$, l'équation $x = kx + \ell$ a une solution unique $x = \frac{1}{1-k}\,\ell$. Nous allons voir que la fraction $\frac{1}{1-k}$ est reliée formellement à l'expression de l'opérateur étoile appliqué au langage K, c'est à dire au langage K^*.

- Suite géométrique

On se donne un nombre réel q: $q \in \mathbb{R}$. Une suite géométrique "de raison q" est une suite de la forme 1, q, q^2, ... , q^n, ... À chaque étape, on multiplie le terme courant de la suite par la raison q.
Si $q = 1$, la suite est constante égale à 1. Si $q = -1$, c'est une suite alternée :
1, -1, 1, -1, *etc*. Si $|q| > 1$, alors $|q^n|$ tend vers l'infini si l'entier n tend vers l'infini ; penser par exemple au cas $q = 2$ où l'on a 1, 2, 4, 8, ... Si $|q| < 1$, alors q^n tend vers zéro si n tend vers l'infini. Par exemple pour $q = \frac{1}{2}$, on obtient la suite 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, ...

- Série associée à une suite géométrique

On se donne une suite géométrique q^n pour $q \in \mathbb{R}$ fixé et $n \in \mathbb{N}$. On pose $S_0 = 1$, $S_1 = 1 + q^1$, $S_2 = 1 + q + q^2 = S_1 + q^2$, ... , $S_n = 1 + q + q^2 + \cdots + q^n = S_{n-1} + q^n$. On observe que la somme S_n comporte $(n+1)$ termes. Si $q = 1$, $S_n = n + 1$, ce qui est d'un intérêt limité. Mais si $q \neq 1$, on a toujours, pour tout entier $n \in \mathbb{N}$, $S_n = \frac{1-q^{n+1}}{1-q}$.
En effet, $S_n = 1 + (q + q^2 + \cdots + q^n)$ et en multipliant cette somme par la raison q, il vient $qS_n = (q + q^2 + \cdots + q^n) + q^{n+1}$. Quand on fait la différence entre ces deux expressions, les termes entre parenthèses s'éliminent et on trouve $S_n - qS_n = 1 - q^{n+1}$, c'est à dire
$(1-q)S_n = 1 - q^{n+1}$, ce qui montre la propriété. □
Maintenant, si $|q| < 1$ et $n \longrightarrow \infty$, on a $S_n \longrightarrow \frac{1}{1-q}$ puisque q^{n+1} tend vers zéro si n tend vers l'infini. On peut écrire cette propriété sous la forme $1 + q + q^2 + \cdots + q^n \longrightarrow \frac{1}{1-q}$ pour $|q| < 1$. On dit que la série géométrique $1 + q + q^2 + \cdots + q^n$ converge vers $\frac{1}{1-q}$.

- Expression de la solution d'une équation du premier degré

On reprend l'expression $x = \frac{1}{1-k}\,\ell$ de la solution x de l'équation $x = kx + \ell$. Si $|k| < 1$, on peut remplacer $\frac{1}{1-k}$ par la série $1 + k + k^2 + \cdots + k^n + \ldots$. Alors $x = (1 + k + k^2 + \cdots + k^n + \ldots)\,\ell$.
L'analogie avec les langages et le lemme d'Arden est alors explicite, puisque
$K^* = \varepsilon + K + K^2 + \cdots + K^n + \ldots$, à la façon dont on somme une série géométrique. Mais la somme a été remplacée par la réunion ensembliste et le produit ordinaire des nombres par le

produit de concaténation ! De plus, la condition $k \neq 1$ a pour analogue l'hypothèse $\varepsilon \notin K$.

- Lemme d'Arden, deuxième version

On se donne deux langages $K \subset A^*$ et $L \subset A^*$ avec $\varepsilon \notin K$. On cherche $X \subset A^*$ tel que $X = X.K + L$. On a donc $L \subset X$ et $\forall x \in X$, $\forall k \in K$, $x.k \in X$. Il existe un et un seul langage X qui satisfait à cette équation, le langage $X = L.K^*$.
La preuve est exactement calquée sur la preuve de la première version. C'est un bon exercice !

- Lemme d'Arden, cas général

On se donne deux langages $K \subset A^*$ et $L \subset A^*$ sur l'alphabet A, avec $\varepsilon \in K$. Le langage $X \subset A^*$ est solution de l'équation $X = K.X + L$, si et seulement si il existe $Y \subset A^*$ tel que $X = K^*.(L+Y)$.
On a aussi un résultat analogue pour l'équation $X = X.K + L$; on peut écrire toute solution sous la forme $X = (L+Y).K^*$, où $Y \subset A^*$ est un langage sur l'alphabet A.
On a perdu l'unicité dans la "division par zéro" due au fait que $\varepsilon \in K$.

- Résolution d'équations linéaires

Une fois que l'on sait résoudre une équation de la forme $X = K.X + L$ (avec $\varepsilon \notin K$), on peut envisager la résolution d'un système linéaire d'équations entre langages.

- Exemple 1. Système de deux équations

Sur l'alphabet $A = a + b$, on se propose de résoudre le système de deux équations
$X_1 = AX_2$ et $X_2 = aX_2 + bX_1 + \varepsilon$.
On reporte la première équation dans la seconde : $X_2 = aX_2 + b(AX_2) + \varepsilon$. On factorise cette expression à l'aide de la distributivité à droite de la concaténation : $X_2 = (a + bA)X_2 + \varepsilon$. On a maintenant une équation d'Arden de la forme $X = K.X + L$, avec $K = a + bA = a + ba + bb$ et $L = \varepsilon$. On constate que K ne contient pas le mot sans lettre et on en déduit
$X_2 = K^*\varepsilon = K^* = (a + ba + bb)^*$. On peut expliciter les mots du langage X_2 de plus petite longueur : $X_2 = \varepsilon + a + ba + bb + aba + abb + a^2 + a^3 + a^4 + baba + bbba + \ldots$ Puis $X_1 = AX_2$ s'obtient en concaténant à gauche tous les mots de X_2 par la lettre a, puis par la lettre b avant de réunir les deux ensembles obtenus :
$X_1 = \{a, a^2, aba, abb, a^2ba, a^2bb, a^3, \ldots\} \cup \{b, ba, bba, bbb, baba, babb, ba^2, \ldots\}$.

- Exemple 2. Système de trois équations

On va maintenant résoudre le système suivant de trois équations $X_1 = bX_2 + aX_3$, $X_2 = bX_3 + aX_1 + \varepsilon$, $X_3 = AX_3$ sur l'alphabet $A = \{a, b\}$.
La troisième équation peut s'écrire $X_3 = AX_3 + \emptyset$. Elle peut être résolue à l'aide du lemme d'Arden puisque l'ensemble A ne contient pas le mot vide. Donc $X_3 = A^*\emptyset$ et le langage X_3 est vide : $X_3 = \emptyset$.
On reporte cette valeur $X_3 = \emptyset$ dans les deux autres équations. On obtient : $X_1 = bX_2$,
$X_2 = aX_1 + \varepsilon$. On reporte alors dans la seconde équation l'expression de X_1 proposée à la première : $X_2 = a(bX_2) + \varepsilon = (ab)X_2 + \varepsilon$. On déduit du lemme d'Arden la relation $X_2 = (ab)^*$ puis $X_1 = bX_2 = b(ab)^*$. La solution du système proposé est donc explicitée : $X_1 = b(ab)^*$, $X_2 = (ab)^*$ et $X_3 = \emptyset$.

• Exemple 3. Système de cinq équations

On se donne cette fois un système de cinq équations : $X_1 = aX_4 + bX_2$, $X_2 = aX_3 + bX_4$, $X_3 = bX_5 + aX_4 + \varepsilon$, $X_4 = AX_4$, $X_5 = aX_3 + bX_4$ sur l'alphabet $A = a + b$. Malgré le nombre d'équations, la résolution peut se mener sans difficulté majeure.

De façon analogue à l'exemple précédent, on remarque que la 4e équation $X_4 = AX_4$ a pour solution $X_4 = \emptyset$. On reporte cette valeur dans les autres équations et on obtient : $X_1 = bX_2$, $X_2 = aX_3$, $X_3 = bX_5 + \varepsilon$, $X_5 = aX_3$.

On substitue la valeur de X_5 dans l'équation pour X_3 : $X_3 = b(aX_3) + \varepsilon = (ba)X_3 + \varepsilon$. Comme ba est un langage formé du seul mot "ba" et ne contient pas le mot sans lettre, on déduit du lemme d'Arden une expression pour X_3 : $X_3 = (ba)^* \varepsilon = (ba)^*$.

On reporte cette expression dans la seconde équation : $X_2 = a(ba)^*$, puis dans la première : $X_1 = b(a(ba)^*) = ba(ba)^* = (ba)^+$. On en déduit la solution du système : $X_1 = (ba)^+$, $X_2 = a(ba)^*$, $X_3 = (ba)^*$, $X_4 = \emptyset$, $X_5 = a(ba)^*$.

On note à l'issue de ces trois exemples que l'utilisation du lemme d'Arden doit être faite le plus tard possible dans le processus de résolution du système linéaire.

Exercices

• Résolution d'un système d'équations

On se donne l'alphabet $A = \{a, b\}$. On note ε le mot sans lettre. On cherche les langages X_1 et X_2 de sorte que $X_1 = bX_1 + aX_2$ et $X_2 = bX_1 + aX_2 + \varepsilon$.

a) Utiliser la première équation pour exprimer X_1 en fonction de X_2.

b) Reporter l'expression obtenue pour en déduire X_2 en fonction des mots du langage et des opérateurs rationnels (réunion, concaténation, étoile).

c) En déduire une expression du langage X_1.

d) Remarquer que $X_2 = X_1 + \varepsilon$.

e) Reporter cette expression dans la première équation et en déduire une nouvelle expression du langage X_1. [$A^* a$]

f) Vérifier que les deux expressions trouvées aux questions c) et e) définissent bien le même langage.

• Un autre système d'équations

On se donne l'alphabet $A = \{a, b\}$ et ε désigne le mot sans lettre. Pour un langage quelconque L sur l'alphabet A, on note $\widetilde{L}$ le langage obtenu en enlevant le mot sans lettre de la clôture L^* : $\widetilde{L} = L^* \backslash \{\varepsilon\}$.

On considère quatre langages X_1, X_2, X_3 et X_4 sur l'alphabet A. On suppose qu'ils satisfont au système d'équations suivant :

$X_1 = aX_1 + bX_2 + \varepsilon$, $X_2 = aX_1 + bX_3 + \varepsilon$, $X_3 = bX_2 + aX_4$, $X_4 = AX_4$.

On se propose de résoudre ce système d'équations.

a) Montrer que l'on a $b\left((b^2)^*\right) = \left((b^2)^*\right) b$.

b) Décrire en une phrase les mots des langages $L = \varepsilon + b\left((b^2)^*\right)$ et $\widetilde{L}$.

c) Que vaut le langage X_4 ? [$\emptyset$]

d) Exprimer X_2 uniquement en fonction de X_1, X_2, des mots de l'alphabet A et du mot sans lettre.

e) En déduire X_2 en fonction de X_1, des mots de l'alphabet A et du mot sans lettre.

f) Quelle est la valeur du langage X_1 ? $[(La)^* L]$

g) Terminer la résolution du système d'équations

$X_1 = aX_1 + bX_2 + \varepsilon$, $X_2 = aX_1 + bX_3 + \varepsilon$, $X_3 = bX_2 + aX_4$, $X_4 = AX_4$

en précisant la valeur des langages X_2 et X_3.