

5. Lemme de l'étoile

Enseignant: Arnaud Casteigts

*Assistants: E. Arm & M. Marseloo
Moniteurs: A. Hugener & O. Oternaud*

Note : La première partie de la séance portant sur la transformation $AFD \rightarrow ER$ a été ajoutée aux notes du cours précédent.

Le **lemme de l'étoile** (pumping lemma, en anglais), est un outil central en langages formels. Il désigne une propriété que *tous* les langages réguliers *doivent* satisfaire. Par conséquent, pour montrer qu'un langage n'est pas régulier, il suffit de montrer qu'il ne satisfait pas cette propriété.

5.1 Lemme de l'étoile

Le lemme de l'étoile établit que tout langage régulier doit satisfaire une certaine propriété. Quelle est cette propriété ?

Réfléchissons en termes d'automates. Si un langage L est régulier, alors il existe un AFD $A(L)$ qui reconnaît L . Si on regarde le chemin pris par cet automate pour reconnaître un mot $w \in L$, il y a deux possibilités :

- Soit $A(L)$ accepte w sans jamais passer deux fois par le même état,
- Soit $A(L)$ accepte w en passant au moins deux fois par le même état.

Si la longueur du mot w est grande (par exemple plus grande que le nombre k d'états de l'automate), alors nécessairement, on est dans le deuxième cas et il y aura au moins un état répété, disons l'état q . On peut alors découper le mot w en trois facteurs $x \cdot y \cdot z$ tels que :

- La lecture de x va de l'état initial à l'état q ,
- La lecture de y commence et termine sur l'état q ,
- La lecture de z va de l'état q à un état final.

Il est possible, dans des cas particuliers, que x ou z soient des mots vides, par exemple si q est un état initial ou final (ou les deux). Par contre, $|y|$ est strictement positif car nous avons vu que l'exécution passe *plusieurs* fois par l'état q . Supposons que $x \cdot y \cdot z$ est accepté (avec y qui commence et termine sur le même état). Que se passerait-il si au lieu de lire $x \cdot y \cdot z$, on lisait le mot $x \cdot y \cdot y \cdot z$? Réponse : $x \cdot y \cdot y \cdot z$ serait nécessairement accepté aussi,

car il va suivre le même chemin (hormis la répétition). Il en va de même pour $x \cdot y \cdot y \cdot y \cdot z$, et en fait, tous les mots de la forme $x \cdot y^i \cdot z$ (et même $x \cdot y^0 \cdot z = xz$, d'ailleurs).

Récapitulons : si un mot de L est suffisamment long, alors il existe un facteur dans ce mot qui peut être répété un nombre arbitraire de fois sans sortir de L . Plus formellement :

Lemme 5.1 (Lemme de l'étoile). *Si un langage L est régulier, alors il existe une longueur k telle que tout mot $w \in L$ avec $|w| \geq k$ peut s'écrire sous la forme $x \cdot y \cdot z$ avec :*

1. $|y| > 0$,
2. $x \cdot y^i \cdot z \in L$ pour tout $i \geq 0$.
3. $|x \cdot y| \leq k$

Nous avons déjà expliqué les deux premières propriétés. La troisième est souvent utile également (bien que pas systématiquement utilisée). Elle s'explique par le fait qu'on peut forcer le choix de x et y de sorte à ce qu'aucune autre répétition ne se produise avant la lecture du dernier symbole de y , donc en ne visitant que des états différents jusqu'à ce stade, le nombre possible de ces états étant au plus le nombre total d'états de l'automate k .

5.2 Utilisation

Ainsi, tout langage régulier doit satisfaire le lemme de l'étoile. On peut donc l'utiliser pour montrer qu'un langage L n'est pas régulier. La démarche à suivre est classique : par l'absurde, on suppose d'abord que L est régulier, ce qui implique que le lemme de l'étoile est satisfait, puis on arrive à une contradiction : l'automate reconnaissant L acceptera des mots qui ne sont pas dans L . Le langage L n'est donc pas régulier.

5.2.1 Exemple : $L = \{a^n b^n \mid n \in \mathbb{N}\}$ sur l'alphabet $\Sigma = \{a, b\}$

Prenons l'exemple du langage $L = \{a^n b^n \mid n \in \mathbb{N}\}$ sur l'alphabet $\Sigma = \{a, b\}$, autrement dit le langage $L = \{\varepsilon, ab, aabb, aaabbb, aaaabbbb, \dots\}$, qui est infini.

Si L est régulier, alors le lemme de l'étoile nous dit qu'il existe une longueur k au delà de laquelle tout mot du langage peut être décomposé sous la forme $x \cdot y \cdot z$, avec $|y| > 0$ et $x \cdot y^i \cdot z \in L$ pour tout i . Prenons un mot $w = aa \dots aabb \dots bb$ de longueur supérieure à k tel que $w \in L$. Quelle que soit la décomposition choisie, il y a trois possibilités pour le facteur y :

1. y ne contient que des **a**,
2. y ne contient que des **b**,
3. y commence par des **a** et termine par des **b**.

Examinons chaque cas. Si y n'a que des **a**, alors $x \cdot y^2 \cdot z$ aura plus de **a** que de **b**, ce qui contredit la propriété 2 du lemme de l'étoile. Si y n'a que des **b**, alors $x \cdot y^2 \cdot z$ aura plus de **b** que de **a** (idem). Enfin, si y commence par des **a** et termine par des **b**, alors $x \cdot y^2 \cdot z$ alternera des **a**, puis des **b**, puis des **a**, puis des **b**, il n'est donc pas non plus dans L (là encore, contradiction). Dans chacun des cas, on arrive à la contradiction que L contient un mot... qui n'est pas dans L (glups). La conclusion est qu'aucune décomposition $x \cdot y \cdot z$ permet de satisfaire le lemme de l'étoile, donc L n'est pas régulier.

5.2.2 $L' = \{w \mid w \text{ a autant de a que de b}\}$ sur l'alphabet $\Sigma = \{a, b\}$

En exercices, vous montrez que L' n'est pas régulier, en utilisant le lemme de l'étoile et en utilisant cette fois la troisième propriété ($|xy| \leq k$). En attendant, on peut démontrer cela autrement, en faisant un lien utile avec le langage L de la section 5.2.1.

Définissons d'abord un autre L'' correspondant à l'expression régulière a^*b^* . Ce langage est bien régulier, puisqu'on l'a défini à partir d'une expression régulière.

Que vaut le langage $L' \cap L''$? Autrement dit tous les mots qui ont autant de **a** que de **b** (mots de L') et qui ont d'abord des a puis des b (mots de L''). Réponse : les mots de cette intersection sont exactement les mots de $L = \{a^n b^n : n \in \mathbb{N}\}$, on a donc $L' \cap L'' = L$.

On sait que l'intersection de deux langages réguliers est toujours un langage régulier. Sachant que L'' est régulier, on peut donc dire que si L' était régulier, alors $L' \cap L'' = L$ le serait aussi. Mais on sait déjà que L n'est pas régulier, donc L' ne peut pas être régulier.

5.3 Conclusion

Certain langages ne sont pas réguliers et ne peuvent donc pas être reconnus par des automates finis. Par exemple, le langage $\{a^n b^n \mid n \in \mathbb{N}\}$ n'est pas régulier. Le langage de tous les palindromes est un autre exemple. Plus tard dans le cours, nous étudierons des modèles de machines plus puissantes, qui permettent de reconnaître ces langages. Puis nous verrons (à leur tour) les limites de ces machines. Ultimement, nous montrerons que même les ordinateurs d'aujourd'hui (et ceux de demain...) ne peuvent pas reconnaître certains langages.