Théorie des codes

Nicolas Daviaud

17 décembre 2010

1 Définitions

Définition 1.1. Un *alphabet* est un ensemble. Ses éléments sont appelés les lettres de cet alphabet. Dans la suite, on supposera de plus qu'aucune des lettres d'un alphabet ne peut être obtenue en concaténant d'autres lettres de cet alphabet.

Exemple 1.1. Avec ces hypothèses, $\{a\}$ et $\{a, b, c\}$ sont des alphabets, tandis que $\{a, b, ab\}$ et $\{a, aa\}$ n'en sont pas.

Définition 1.2. Soit A un alphabet. On note A^* l'ensemble des mots sur A c'est-à-dire l'ensembles des suites finies d'éléments de A.
On appelle mot vide, noté ϵ la suite vide.

Définition 1.3. Un *monoïde* est un ensemble non vide muni d'une loi de composition interne associative qui admet un élément neutre. Par convention, on notera la loi multiplicativement dans ce qui suit.

Exemple 1.2. En particulier, tout groupe est un monoïde. Ce n'est cependant pas le cas en général et la proposition 1.7 donne un exemple de monoïde dont les éléments ne sont pas inversibles.

Définition 1.4. Une partie P d'un monoïde M est appelé une *base* de M si tout élément de M admet une unique décomposition comme produit d'éléments de P. Une base est donc en particulier une partie génératrice de M. Elle est unique à l'ordre près.
Un monoïde est dit *libre* s'il admet une base.

Définition 1.5. Soit M un monoïde. Un *sous-monoïde* de M est une partie de M contenant l'élément neutre de M et stable par la loi de composition interne.

Exemple 1.3. En particulier, pour tout monoïde M, $\{1_M\}$ est un sous-monoïde de M, appelé monoïde trivial.

Définition 1.6. Soit P une partie d'un monoïde M. On appelle *monoïde engendré par* P le plus petit sous-monoïde de M contenant P.

Proposition 1.7. *A^* muni de la concaténation est un monoïde libre de base A et qui admet pour élément neutre le mot vide.*

Démonstration. A^* est par construction le monoïde engendré par A. La liberté vient de l'hyptohèse supplémentaire faite sur l'alphabet dans la définition 1.1. □

Proposition 1.8. *Soit X une partie de A^*. Le monoïde engendré par X est X^*, c'est-à-dire l'ensemble des produits d'éléments de X.*

Démonstration. X^* est par construction un sous-monoïde de A^*. De plus, tout sous-monoïde contenant X contient X^*. □

Définition 1.9. Soit X une partie de A^*. On dit que X est un *code* si X^* est libre de base X.

Proposition 1.10. *Dire que X est un code revient à dire que tout mot sur X se décompose de manière unique sur X :*

$$x_1 \cdots x_n = x'_1 \cdots x'_m \Rightarrow \begin{cases} m = n \\ \forall i \leq n, x'_i = x_i \end{cases}$$

Exemple 1.4. Soit $A = \{a, b\}$ un alphabet.

- Les ensembles suivants sont des codes :
  - A d'après la proposition 1.7
  - $\{a, ba\}$ car dans le sous-monoïde engendré, tout b est le début d'un lexème ba et les autres lexèmes sont des a.
- Les ensembles suivants ne sont pas des codes :
  - $\{a, aa\}$ trivialement
  - $\{a, ba, ab\}$ car aba admet deux décompositions.

Définition 1.11. Un code est dit *maximal* s'il n'est inclu strictement dans aucun code.

Exemple 1.5. Voici quelques exemples de codes maximaux :

$X_1 = \{aa, ab, bb, ba\}$ est un code maximal fini (en effet, si un surcode contient un autre mot w alors ww admet deux décompositions car étant de longueur paire).

$X_2 = ba^*$ est un code maximal infini.

$X_3 = \{a, ba\}$ est un code mais n'est pas maximal (cela reste un code si on lui ajoute bb).

Dans la suite, on continuera à utiliser ces exemples.

Définition 1.12. On appelle préfixe (respectivement suffixe) d'un mot w tout mot u tel qu'il existe $v \in A^*$ tel que $w = uv$ (resp. $w = vu$). Si u est préfixe de w, on note

$$u \leq w.$$

Un préfixe ou suffixe est dit propre s'il est non vide et différent du mot entier. Si u est préfixe propre de w, on note

$$u < w.$$

Définition 1.13. Un sous-ensemble P de A^* est dit *ensemble préfixe* si aucun mot de P n'est préfixe propre d'un mot de P, ie pour tous mots u et v dans P,

$$u \leq v \Rightarrow u = v.$$

Exemple 1.6. $\{a, ba\}$ est un ensemble préfixe alors que $\{a, ab\}$ n'en est pas un.

Définition 1.14. On appelle mot sans bord un mot dont aucun préfixe propre n'est un suffixe propre.

Exemple 1.7. *abb* est un mot sans bord tandis que *abab* n'en est pas un.

Proposition 1.15. *Si l'alphabet A comporte au moins deux lettres, alors tout mot $u \in A^*$ peut être complété en un mot sans bord uv avec $v \in A^*$.*

Démonstration. Notons a la première lettre de u et b une lettre de A différente de a. Le mot $w = uab^{|u|}$ est sans bordure. En effet, soit t un préfixe non vide de w. Sa première lettre est donc un a. Une condition nécessaire pour qu'il soit un suffixe de w est que $|t| > |u|$ au vu du nombre de b à la fin de w. Mais dans ce cas, t s'écrit comme suffixe $sab^{|u|}$ et comme préfixe $uab^{|s|}$ avec $s \in A^*$. On obtient ainsi $|s| = |u|$ et finalement $t = w$. Un préfixe non vide de w qui en est également un suffixe est donc nécessairement w entier. w est donc sans bord. □

Ce résultat nous sera très utile pour démontrer la proposition 3.2.

2 Mesure d'un code

Dans cette partie, on introduit une notion très proche de la théorie de la mesure qui permettra de caractériser les codes grâce à la proposition 2.4. On y définit et caractérise également la notion de maximalité pour les codes.

Définition 2.1. Une *distribution de Bernoulli* sur A^* est un morphisme de monoïde

$$\pi : A^* \to \left(\mathbb{R}_+, \times\right)$$

qui vérifie la relation

$$\sum_{a \in A} \pi(a) = 1.$$

Elle est *strictement positive* si $\forall a \in A, \pi(a) > 0$

Soit π une distribution de Bernoulli. On étend π à $\mathscr{P}(A^*)$ en posant

$$\forall L \subset A^*, \pi(L) = \sum_{l \in L} \pi(l).$$

Proposition 2.2. *π est une mesure positive sur $(A^*, \mathscr{P}(A^*))$.*

Démonstration. π vérifie les conditions de la définition d'une mesure positive :

(i) $\forall L \subset A^*, \pi(L) \geq 0$

(ii) $\pi(\varnothing) = 0$

(iii) Pour toute famille $(L_i)_{i \in I}$ de parties disjointes de A^* ,

$$\pi\left(\bigcup_{i \in I} L_i\right) = \sum_{i \in I} \pi(L_i).$$

□

Remarque. Si les parties ne sont pas disjointes, on a seulement

$$\pi\left(\bigcup_{i \in I} L_i\right) \leq \sum_{i \in I} \pi(L_i).$$

Exemple 2.1. Si A est fini, on appelle distribution uniforme la distribution de Bernoulli définie par

$$\forall a \in A, \pi(a) = \frac{1}{card(A)}$$

Proposition 2.3. *Soit $X \subset A^+ = A^* \setminus \{\epsilon\}$ et π une distribution de Bernoulli sur A^**

1. *Si X est un code, alors*

$$\forall n, \pi(X^n) = \pi(X)^n;$$

$$\pi\left(X^*\right) = \sum_{n\geq 0} \pi(X)^n;$$

et en particulier,

$$\pi\left(X^*\right) < \infty \Longleftrightarrow \pi(X) < 1.$$

2. *Réciproquement, si π est strictement positive, si $\pi(X)$ est finie et si*

$$\forall n, \pi\left(X^n\right) = \pi(X)^n$$

alors X est un code.

Démonstration. Notons S_n le produit cartésien de n fois X.

1. Soit $n \in \mathbb{N}$. Supposons que X est un code.

$$\pi(X)^n = \sum_{\bar{x}\in S_n} \left(\pi(x_1)\cdots\pi(x_n)\right) = \sum_{\bar{x}\in S_n} \pi(x_1\cdots x_n)$$

Or par définition d'un code, X^n est en bijection à S_n. Ceci fournit le résultat

$$\pi(X)^n = \sum_{u\in X^n} \pi(u) = \pi\left(X^n\right).$$

De plus les $X^n, n \in \mathbb{N}$ forment une partition finie de X^* car X est un code. Avec ce qui précède, cela prouve

$$\pi(X)^* = \sum_{n\in\mathbb{N}} \pi(X^n) = \sum_{n\in\mathbb{N}} \pi(X)^n.$$

Le cas particulier en est une conséquence triviale (série gémométrique).

2. Supposons que π est positive, que $\pi(X)$ est finie, que

$$\forall n, \pi\left(X^n\right) = \pi(X)^n$$

mais que X n'est pas un code. Montrons que π n'est alors pas strictement positive. Soit $u \in X^*$ admettant deux décompositions sur X

$$u = x_1\cdots x_n = y_1\ldots y_m.$$

Donc uu admet (entre autres) deux décompositions de même longueur $k = n + m$ à savoir

$$uu = x_1\cdots x_n y_1\ldots y_m = y_1\ldots y_m x_1\cdots x_n.$$

L'application

$$\begin{array}{rcl} S_k & \to & X^k \\ (x_1,\ldots,x_k) & \mapsto & x_1\cdots x_k \end{array}$$

est donc surjective mais non injective. π étant à valeurs positives, on a donc

$$\pi\left(X^k\right) = \pi(X)^k = \sum_{\bar{x}\in S_k} \pi(x_1\cdots x_k) \geq \pi\left(X^k\right) + \pi(uu),$$

ce qui implique $\pi(uu) = 0$ et on obtient ainsi que π est non strictement positive.

□

Proposition 2.4. *Si X est un code sur A, alors toute distribution de Bernoulli π sur A^* vérifie $\pi(X) \leq 1$.*

Démonstration. Supposons dans un premier temps que la longueur des éléments de X est majorée par k (c'est le cas par exemple si X est fini). On a alors

$$X^n \subset \bigcup_{i=1}^{nk} A^i.$$

Donc par croissance de la mesure π et l'union étant disjointe,

$$\pi\left(X^n\right) \leq \sum_{i=1}^{nk} \pi\left(A^i\right).$$

Or A est trivialement un code sur A^* donc d'après la première partie du théorème précedent,

$$\begin{aligned} \pi\left(X^n\right) &\leq \sum_{i=1}^{nk} \pi(A)^i \\ &= \sum_{i=1}^{nk} 1. \end{aligned}$$

Le théorème précédent fournit donc

$$\pi(X)^n \leq nk.$$

Par comparaison de suite, on obtient ainsi le résultat voulu

$$\pi(X) \leq 1.$$

Revenons au cas général et considérons les codes

$$X_n = \{x \in X | \, |x| \leq n\}$$

qui vérifient l'hypothèse utilisée ci dessus. On a ainsi

$$\lim_{n\to\infty} \uparrow X_n = X \text{ et } \pi(X_n) \leq 1.$$

π étant une mesure positive, il s'agit des hypothèses exactes du théorème de convergence monotone, qui fournit enfin

$$\pi(X) \leq 1.$$

□

Proposition 2.5. *Soit X un code sur A. S'il existe une distribution de Bernoulli strictement positive sur A^* induisant une mesure de probabilité sur X (i.e. $\pi(X) = 1$) alors X est un code maximal.*

Démonstration. Supposons que X n'est pas maximal. Par définition, il existe $y \in A^+$ tel que $Y = X \cup \{y\}$ soit un code. D'après la proposition précédente, on a donc $\pi(Y) \leq 1$. D'autre part, $y \notin X$, donc

$$\pi(Y) = \pi(X) + \pi(y) = 1 + \pi(y).$$

Par conséquent, $\pi(y) = 0$, ce qui contredit le fait que π est strictement positive. □

Exemple 2.2. Si l'on considère les codes de l'exemple 1.5 et la distribution uniforme sur $A = \{a, b\}$, on a bien :

- $X_1 = \{aa, ab, bb, ba\}$ est maximal car

$$\pi(X_1) = 4 \cdot \left(\frac{1}{2} \cdot \frac{1}{2}\right) = 1$$

- $X_2 = ba^*$ est maximal car

$$\pi(X_2) = \sum_{n=0}^{\infty} 2 \cdot \frac{1}{2^n} = 1$$

- $X_3 = \{a, ba\}$ n'est pas maximal et

$$\pi(X_3) = \frac{3}{4}$$

3 Ensembles complets

Il ne reste plus qu'à définir une notion de densité pour les codes pour aboutir au théorème souhaité.

Définition 3.1. 1. Un sous-ensemble P d'un monoïde M est dit *dense* dans M si pour tout élément m de M

$$\begin{array}{ll} & \exists u, v \in M : umv \in P \\ \text{i.e.} & MmM \cap P \neq \varnothing \end{array}$$

Un sous-ensemble d'un monoïde est dit mince s'il n'est pas dense.

2. Un sous-ensemble P de M est dit *complet* (dans M) si le monoïde engendré par P est dense dans M.

Remarque. Un code peut être mince et complet, ces notions ne sont pas contradictoires, l'exemple le plus simple étant le cas d'un alphabet fini.

Exemple 3.1. Si l'on considère toujours les codes de l'exemple 1.5,
- $X_1 = \{aa, ab, bb, ba\}$ est mince ($m = abb$) mais complet car $X_1^* = \{uu | u \in A^*\}$
- $X_2 = ba^*$ est mince ($m = bb$) mais complet (cf proposition ci-dessous)
- $X_3 = \{a, ba\}$ est mince ($m = bb$) et non complet ($m = bb$)

Proposition 3.2. *Un code maximal est complet.*

Afin de montrer cette proposition, nous aurons besoin du lemme suivant qui permet de compléter un code quelconque en un code complet.

Lemme 3.3. *Soit X un code sur A^*. Soit y un mot sans bord tel que $A^*yA^* \cap X^* = \varnothing$. Si on pose*

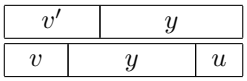
$$\begin{array}{ll} & U = A^* \setminus (X^* \cup A^*yA^*) \\ et & Y = X \cup y(Uy)^*, \end{array}$$

alors Y est un code complet.

Démonstration. Posons

$$V = A^* \setminus A^*yA^*.$$

On a donc par hypothèse $X^* \subset V$ et $U = V \setminus X^*$.
- Montrons que $Z = Vy$ est un ensemble préfixe.
  Soient $vy, v'y \in Z$. Supposons que $vy < v'y$. Supposons que l'on ait



  Un suffixe propre de y en est donc également un préfixe propre, ce qui contredit le fait que y soit un mot sans bord.
  vy est donc un préfixe propre de v'. Mais dans ce cas, v' est dans $A^*yA^* \cap$

$V = \varnothing$ ce qui est absurde.
On a ainsi

$$\forall z, z' \in Z, z \leq z' \Rightarrow z = z'.$$

Z est donc un ensemble préfixe.

– Montrons à présent que Y est un code.
Supposons que cela ne soit pas le cas et qu'un mot de Y^* s'écrive

$$y_1 \cdots y_n = y'_1 \cdots y'_m$$

dans Y. Quitte à le choisir de longueur minimale, on peut supposer $y_1 \neq y'_1$. X étant un code, un des lexèmes est dans $Y \setminus X$. Supposons par symétrie qu'il soit dans $y_1, \ldots, y_n$ et soit p l'indice minimum tel que $y_p \in y(Uy)^*$. Sachant que $A^* y_p A^* = A^* y A^*$, on a donc $A^* y_p A^* \cap X^* = \varnothing$. Donc $y'_1 \cdots y'_m$ n'est pas non plus dans $X*$ et par conséquent, l'un de ses lexèmes est dans $y(Uy)^*$. Soit q l'indice minimal tel que $y'_q \in y(Uy)^*$. Dans ce cas,

$$y_1 \cdots y_{p-1} \text{ et } y'_1 \cdots y'_{q-1} \in X* \subset V$$

par minimalité de p et q. Par définition de Z,

$$y_1 \cdots y_{p-1} y \text{ et } y'_1 \cdots y'_{q-1} y \in Z.$$

Or

$$y_1 \cdots y_n = y'_1 \cdots y'_m \Rightarrow \left\{ \begin{array}{lcl} & y_1 \cdots y_{p-1} y & < \quad y'_1 \cdots y'_{q-1} y \\ ou & & \\ & y'_1 \cdots y'_{q-1} y & < \quad y_1 \cdots y_{p-1} y \end{array} \right.$$

Z étant un ensemble préfixe, on obtient ainsi

$$y_1 \cdots y_{p-1} = y'_1 \cdots y'_{q-1}$$

en simplifiant par y. Or X est un code donc $p = q$ et si $p - 1 \neq 0$, alors en particulier $y_1 = y'_1$ ce qui contredit l'hypothèse faite plus haut. Donc $p = q = 1$.
y_1 et y'_1 s'écrivent ainsi

$$\begin{array}{rcl} y_1 & = & y u_1 y \cdots y u_k y \\ y'_1 & = & y u'_1 y \cdots y u'_l y \end{array}$$

avec $u_1, \ldots, u'_l \in U \subset V$ donc en particulier, y n'est pas un sous-mot de l'un de ces lexèmes. Comme $p = q$ on peut supposer par symétrie que $y_1 < y'_1$. La remarque précédente montre que y joue le rôle d'un symbole extérieur entre les $u_1, \ldots, u'_l$. Ainsi, on voit aisément que

$$\forall i \leq k, u_i = u'_i$$

car y_1 est préfixe de y'_1.
Posons $t = u'_{k+1} y \cdots y u'_l y$. On a alors

$$y_2 \cdots y_n = t y'_2 \cdots y'_m.$$

y est un sous-mot de t donc de $y_2 \cdots y_n$. Ainsi, $y_2 \cdots y_n$ n'est pas dans X^*. l'un de ses lexèmes est donc dans $y(Uy)^*$. Soit r l'indice minimal tel que $y_r \in y(Uy)^*$. $y_2 \cdots y_{r-1}y$ et $u'_{k+1}y$ sont donc préfixes du même mot donc l'un est préfixe de l'autre. Or ils sont tous deux dans Z qui est un ensemble préfixe. Ainsi

$$y_2 \cdots y_{r-1} = u'_{k+1}.$$

Or par minimalité de r, $y_2 \cdots y_{r-1}$ est dans X^*. On obtient ainsi

$$u'_{k+1} \in X^* \text{ et } u'_{k+1} \in U = V \setminus X^*$$

ce qui est absurde. Y est donc un code.

– Montrons à présent qu'il est complet.
Soit $w \in A^*$. w s'écrit

$$v_1 y v_2 y \cdots y v_{n-1} y v_n$$

avec $v_i \in A^* \setminus A^* y A^*$. Notons $v_{i_1} \dots v_{i_k}$ les v_i qui sont dans X^*. Alors ywy s'écrit

$$(yv_1y \cdots vv_{i_1-1}y)\, v_{i_1}\, (yv_{i_1+1}y \cdots vv_{i_2-1}y) \cdots v_{i_k}\, (yv_{i_k+1}y \cdots vv_n y)$$

Les sous mots entre parenthèses sont par hypothèse dans $y(Uy)^*$ et les autres sous-mots sont exactement les v_{i_j} qui sont par hypothèse dans X^*. ywy est donc une concaténation de mots de Y, c'est-à-dire un élément de Y^*.

□

Démonstration de la proposition 3.2. Soit X un code. Supposons-le non complet. Il existe alors v tel que $A^*vA^* \cap X^* = \varnothing$. D'après la proposition 1.15 il existe $u \in A^*$ tel que $y = vu$ soit un mot sans bord. y vérifie encore $A^*yA^* \cap X^* = \varnothing$. Le théorème précédent donne un sur-code strict de X car contenant au moins y en plus. X n'est donc pas maximal. □

Exemple 3.2. X_1 et X_2 sont maximaux donc complets. X_3 n'est pas complet donc pas maximal.

Proposition 3.4. *Soit X un sous-ensemble mince et complet de A^*. Soit w un mot tel que $A^*wA^* \cap X^* \neq \varnothing$. Alors*

$$\begin{aligned} &\forall z \in A^*, \exists (d,g) \in D \times G : dzg \in X^* \\ i.e.\quad &A^* = D^{-1}X^*G^{-1} = \bigcup_{(d,g)\in D\times G} d^{-1}X^*g^{-1} \end{aligned}$$

où D et G sont l'ensemble des suffixes (respectivement préfixes) de w.

Démonstration. Soit $z \in A^*$. X est complet donc X^* est dense donc le mot wzw se complète dans X^*. Il existe ainsi $u, v \in A^*$ tels que

$$uwzwv \in X^*.$$

Or w n'est pas un sous-mot d'un élément de X donc la décomposition de $uwzwv$ dans X coupe nécessairement les w en au moins deux morceaux :

$$uwzwv = ug'dzgd'v \text{ avec } ug', dzg, d'v \in X^*.$$

On obtient ainsi bien le résultat voulu car $dzg \in X^*$ et $(d, g) \in D \times G$. □

Proposition 3.5. *Soit X un sous-ensemble mince et complet de A^*. Pour toute distribution de Bernoulli π strictement positive, on a*

$$\pi(X) \geq 1.$$

Démonstration. D'après la proposition précédente et π étant une mesure positive,

$$\infty = \pi\left(A^*\right) = \pi\left(\bigcup_{(d,g)\in D\times G} d^{-1}X^*g^{-1}\right) \leq \sum_{(d,g)\in D\times G} \pi\left(d^{-1}X^*g^{-1}\right).$$

$D \times G$ est cependant fini car w est un mot (donc fini). Il existe donc un couple $(d, g) \in D \times G$ tel que $\pi\left(d^{-1}X^*g^{-1}\right) = \infty$. Or

$$d\left(d^{-1}X^*g^{-1}\right)g \subset X^*.$$

Sachant que $\pi(d)\pi(g) \neq 0$ car π est strictement positive, cela implique

$$\infty = \pi(d)\pi\left(d^{-1}X^*g^{-1}\right)\pi(g) \leq \pi\left(X^*\right).$$

On a cependant

$$\pi\left(X^*\right) \leq \sum_{n\geq 0} \pi\left(X^n\right) \leq \sum_{n\geq 0} \left(\pi(X)\right)^n,$$

et donc

$$\infty \leq \sum_{n\geq 0} \left(\pi(X)\right)^n.$$

Ceci montre finalement que $\pi(X) \geq 1$. □

Nous sommes maintenant armés pour montrer le théorème suivant :

Théorème 3.6. *Soit X un code mince. Sont équivalentes :*

- *(i)* *X est un code maximal.*
- *(ii)* *Il existe une distribution de Bernoulli définie positive π telle que $\pi(X) = 1$.*
- *(iii)* *Toute distribution de Bernoulli définie postive π vérifie $\pi(X) = 1$.*
- *(iv)* *X est complet.*

Démonstration. On procède comme suit :

$(i) \Rightarrow (iv)$	Il s'agit exactement de la proposition 3.2.
$(iv) \Rightarrow (iii)$	La proposition 2.4 fournit $\pi(X) \leq 1$ car X est un code, tandis que la proposition 3.5 fournit l'autre sens $\pi(X) \geq 1$ car X est mince et complet.
$(iii) \Rightarrow (ii)$	L'ensemble des distributions de Bernoulli strictement positives est non vide.
$(ii) \Rightarrow (i)$	La proposition 2.5 permet de conclure. □

Ce théorème permet ainsi de déterminer très facilement si un code mince est maximal : il suffit de prendre une distribution de Bernoulli strictement positive et de vérifier si la mesure du code est 1 ou pas. Ainsi, par exemple, X_3 est maximal.

Références

[1] J. Berstel and D. Perrin. *Theory of Codes.* Academic Press, 1984.