CAMUS

Volume 7 Cahiers mathématiques de l'Université de Sherbrooke

CaMUS **7**, 1 – 20

Une première approche aux champs vectoriels combinatoires

Dominic Desjardins Côté

RÉSUMÉ Dans cet article, on définit une approche aux systèmes dynamiques combinatoires en utilisant les complexes simpliciaux et la théorie de Morse discrète selon Forman. De plus, on introduit quelques concepts pour étudier les systèmes dynamiques combinatoires.

1 Introduction

Depuis la dernière décennie, le concept de champ vectoriel combinatoire, introduit par Robin Forman en 1998, est devenu un outil très efficace pour la discrétisation des problèmes continus en mathématiques appliquées, en imagerie et la visualisation des systèmes dynamiques. Donc, on s'intéresse à discrétiser les systèmes dynamiques et, plus précisément, à faire des liens entre la théorie classique de Morse [Mat02] et la théorie de Morse discrète, selon Forman, dans le contexte des systèmes dynamiques. De plus, on veut être capable de transformer un système dynamique continu en un système dynamique discret sans modifier la dynamique du système et vice-versa. On débutera par une brève introduction aux systèmes dynamiques. Par la suite, on décrira les complexes simpliciaux [Jam84] qui est un espace topologique que nous utilisons pour appliquer le champ vectoriel combinatoire. Ensuite, on présentera la théorie de Morse discrète, selon Forman [Rob02, Rob98]. Pour finir, nous modifierons quelques définitions de la théorie de Morse discrète pour introduire le concept de système dynamique discret et quelques méthodes pour les étudier [TMT16] [BTMW16].

2 Les systèmes dynamiques

Tout d'abord, qu'est-ce qu'un système dynamique ? Ce qui nous intéresse ce sont les solutions $x\left(t, x_0\right)$ du système suivant :

$$\begin{cases} \frac{dx}{dt} = f\left(t,x\right) \\ x\left(t_0\right) = x_0 \end{cases},$$

où t est le temps et x_0 est la valeur initiale. Donc, on s'intéresse à voir l'évolution d'une valeur dans le temps. Cette valeur pourrait être la position d'un objet ou

Je remercie l'Institut des Sciences Mathématiques pour leur financement. Je remercie Tomasz Kaczynski pour m'avoir aidé dans mes travaux de recherche et pour son financement.

même des conditions météorologiques. Ainsi, pour simplifier, on définit $\varphi(t,x) : \mathbb{R} \times X \to X$ et on obtient :

$$\begin{cases} \frac{d\varphi}{dt} = f(\varphi) \\ \varphi(0,x) = x \end{cases} .$$

Définition 2.1. Soient X un espace métrique et $\varphi : \mathbb{R} \times X \to X$ une fonction continue. Alors, φ est un *flot* ou un *système dynamique à temps continu* si les propriétés suivantes sont satisfaites :

1. $\varphi(0,x) = \mathrm{Id}_x$;
2. $\varphi(s+t,x) = \varphi(s, \varphi(t,x))$, pour tout t et s.

Donc, la première condition indique qu'au temps initial, on obtient la valeur initiale x_0. Pour la deuxième, si on additionne les temps s et t, alors cela revient à la composition des fonctions aux temps s et t.

Ainsi, on se pose la question suivante : Que peut-on étudier sur un système dynamique ? D'abord, on peut s'intéresser aux futurs du système dynamique, c'est-à-dire $t \to \infty$, ou même aux passés du système quand $t \to -\infty$. De plus, on peut trouver les points d'attractions, les points répulsifs et vérifier s'il y contient des cycles.

En outre, on peut discrétiser le temps du système dynamique. Ainsi, le flot devient $\varphi : \mathbb{Z} \times X \to X$, qu'on appelle un système dynamique à temps discret. De plus, si $f : X \to X$ est donnée par $f(x) = \varphi_1(x)$, alors, on obtient : $\underbrace{f \circ f \circ \ldots \circ f(x)}_{n \text{ compositions}} = \varphi_1 \circ \varphi_1 \circ \ldots \circ \varphi_1(x) = \varphi_n(x)$.

Donc, on remarque qu'il est simple de discrétiser le temps. Si on a la fonction f qui définit une unité de temps, alors on applique la fonction t fois pour obtenir le temps t. Par contre, il est plus complexe de discrétiser l'espace X, car si on ne fait pas attention à l'espace utilisé, alors on perd de l'information quand on discrétise un système continu. Ainsi, le but de cet article est de définir un système dynamique combinatoire ayant un temps et un espace discrétisés.

3 Les complexes simpliciaux

Tout d'abord, avant de définir les champs vectoriels combinatoires, nous devons définir l'espace topologique où nous allons appliquer le champ vectoriel combinatoire. Donc, on a besoin d'un espace géométrique qui a un aspect combinatoire, mais qui va permettre de garder l'information du système dynamique continu. Ainsi, on utilise l'espace des complexes simpliciaux. Avant de les définir, on a besoin de quelques concepts.

Définition 3.1. Un ensemble de points $\{v_0, v_1, \ldots, v_n\}$ dans $\mathbb{R}^N$, avec $n \leq N$, est *géométriquement indépendant*, si tous les scalaires t_i, pour $i = 0, 1, \ldots, n$,

satisfont à :

$$\sum_{i=0}^{n} t_i = 0 \quad \text{et} \quad \sum_{i=0}^{n} t_i v_i = 0,$$

ce qui implique $t_0 = t_1 = \ldots = t_n = 0$.

Lemme 3.2. *Un ensemble $\{v_0, v_1, \ldots, v_n\}$ est géométriquement indépendant si et seulement si l'ensemble des vecteurs $\{v_1 - v_0, v_2 - v_0, \ldots, v_n - v_0\}$ est une famille linéairement indépendant dans le sens de l'algèbre linéaire.*

Démonstration. Soit un ensemble $\{v_0, v_1, \ldots, v_n\}$ géométriquement indépendant. On obtient :

$$\sum_{i=0}^{n} t_i v_i = 0 \implies \sum_{i=1}^{n} t_i v_i = -t_0 v_0 \implies \sum_{i=1}^{n} t_i (v_i - v_0) = -\sum_{i=0}^{n} t_i v_0.$$

De plus, $\sum_{i=0}^{n} t_i = 0$. Par conséquent, on a que $\sum_{i=1}^{n} t_i (v_i - v_0) = 0$ et $t_1 = t_2 = \ldots = t_n = 0$. Donc, $\{v_1 - v_0, \ldots, v_n - v_0\}$ est une famille linéairement indépendante.

Si $\{v_1 - v_0, v_2 - v_0, \ldots, v_n - v_0\}$, alors on obtient

$$\sum_{i=1}^{n} t_i (v_i - v_0) = 0 \implies \sum_{i=1}^{n} t_i v_i - \sum_{i=1}^{n} t_i v_0 = 0.$$

Posons $t_0 = -\sum_{i=1}^{n} t_i$. De plus, $\{v_1 - v_0, \ldots, v_n - v_0\}$ est une famille linéairement indépendante. Alors, $t_1 = t_2 = \ldots = t_n = 0$, ce qui entraîne que $t_0 = -\sum_{i=1}^{n} t_i = 0$. Donc, $\sum_{i=0}^{n} t_i v_i = 0$ et $\sum_{i=0}^{n} t_i = -\sum_{i=1}^{n} t_i + \sum_{i=1}^{n} t_i = 0$. Ce qui implique que $t_0 = t_1 = \ldots = t_n = 0$. Ainsi, $\{v_0, v_1, \ldots, v_n\}$ est géométriquement indépendant. □

Définition 3.3. Soit $\{v_0, v_1, \ldots, v_n\}$ un ensemble géométriquement indépendant dans $\mathbb{R}^N$. On définit un *n-simplexe* σ, un sous-espace engendré par les éléments $v_0, v_1, \ldots, v_n$, qui est l'ensemble de tous les points $x \in \mathbb{R}^N$ tel que

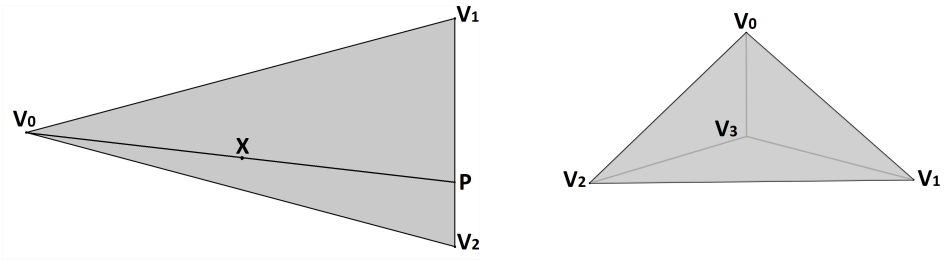
$$x = \sum_{i=0}^{n} t_i v_i,$$

où $\sum_{i=0}^{n} t_i = 1$ et $t_i \geq 0$ pour tout i.

Remarque 3.4. Les t_i de la définition précédente sont appelées les *coordonnées barycentriques* du point x de σ.

Proposition 3.5. *Les coordonnées barycentriques sont uniques.*

Démonstration. Supposons que les coordonnées barycentriques de x ne sont pas uniques. Alors, ils existent t_i et λ_i pour $i = 1, \ldots, n$ telles que $x = \sum_{i=0}^{n} t_i v_i$ avec $\sum_{i=0}^{n} t_i = 1$, $x = \sum_{i=0}^{n} \lambda_i v_i$ avec $\sum_{i=0}^{n} \lambda_i = 1$ et $t_i \neq \lambda_i$ pour au moins un i. On obtient $0 = x - x = \sum_{i=0}^{n} (t_i - \lambda_i) v_i$ et $\sum_{i=0}^{n} (t_i - \lambda_i) = 1 - 1 = 0$. De plus, les v_i sont géométriquement indépendant, alors $t_i - \lambda_i = 0 \forall i$. Donc, les coordonnées barycentrique sont uniques. □



(a) Un 2-simplexe (b) Un 3-simplexe

FIGURE 1 : Deux exemples de simplexes géométriques.

Exemple 3.6. Le 0-simplexe est engendré par v_0. Comme $t_0 = 1$, alors le seul point est $x = v_0$. Donc, le 0-simplexe est un seul point.

Exemple 3.7. Le 1-simplexe, engendré par v_0 et v_1, consiste en tous les points de la forme $x = t_0v_0+t_1v_1$. Par contre, on a $t_0+t_1 = 1$ impliquant que $t_1 = 1-t_0$. On obtient $x = t_0v_0 + (1 - t_0)\, v_1$. Ainsi, le 1-simplexe est l'ensemble de tous les points entre v_0 et v_1, qui est la ligne rejoignant les deux points x_0 et x_1.

Exemple 3.8. Le 2-simplexe σ, engendré par v_0, v_1 et v_2, est l'ensemble des points $x = \sum_{i=0}^{2} t_iv_i$. Donc, $t_0 + t_1 + t_2 = 1$ entraîne que $1 - t_0 = t_1 + t_2$. Si $t_0 \neq 1$, alors

$$\frac{1 - t_0}{1 - t_0} = \frac{t_1 + t_2}{1 - t_0}.$$

Ainsi, $x = t_0v_0 + (1 - t_0)\left(\left(\frac{t_1}{1-t_0}\right) v_1 + \left(\frac{t_2}{1-t_0}\right) v_2\right)$ représente un point entre v_0 et p, où p est le point entre v_1 et v_2. Donc, le 2-simplexe est un triangle. Dans la Figure 1(a), le point x est aux valeurs de $t_0 = \frac{1}{2}$, $t_1 = \frac{1}{6}$ et $t_2 = \frac{1}{3}$.

Exemple 3.9. En suivant une démarche similaire du 2-simplexe, on montre que le 3-simplexe est un tétrahèdre (voir Figure 1(b).)

Définition 3.10. Soit σ un n-simplexe avec les points $\{v_0, v_1, \ldots, v_n\}$:

1. Les points $v_0, v_1, \ldots, v_n$ sont appelées les *sommets* ;
2. Le nombre n est la *dimension* d'un n-simplexe σ. On note $\sigma^{(p)}$ un simplexe de dimension p. De plus, on écrit seulement la dimension du simplexe lorsque nécessaire ;
3. Tous les sous-ensembles de $\{v_0, v_1, \ldots, v_n\}$ sont appelés les *faces*. En particulier, les faces différentes de σ sont les *faces propres*. De plus, on note $\tau < \sigma$ si τ est une face de σ ;
4. *L'adhérence* de σ est l'union de toutes les faces de σ, c'est-à-dire

$$\text{Cl } \sigma := \bigcup_{\tau<\sigma} \tau;$$

5. La *frontière* de σ est l'union de toutes les faces propres de σ. Elle est notée Bd σ ;

6. *L'intérieur* de σ est défini par Int $\sigma = \sigma -$ Bd σ.

Définition 3.11. Un *complexe simplicial* K dans $\mathbb{R}^N$ est un ensemble de simplexes dans $\mathbb{R}^N$ tel que :

1. Tous les faces d'un simplexe dans K est dans K ;

2. L'intersection de deux simplexes de K est une face pour chaque simplexe.

Exemple 3.12. Dans les Figures 2(a) et 2(b) sont des complexes simpliciaux. Par contre, les Figures 2(c) et 2(d) ne sont pas des complexes simpliciaux. Pour les Figures 2(c) et 2(d), la condition 2 n'est pas respectée, car l'intersection entre les deux triangles est une face pour le petit triangle, mais il n'est pas une face du grand triangle. Pour la condition 1, on veut que tous les simplexes aient leurs faces à l'intérieur du complexe simplicial. Dans le cas des complexes simpliciaux géométriques, cette condition est respectée. Par contre, dans le cas abstrait, on veut que tous les sous-simplexes des simplexes soient dans le complexe simplicial.

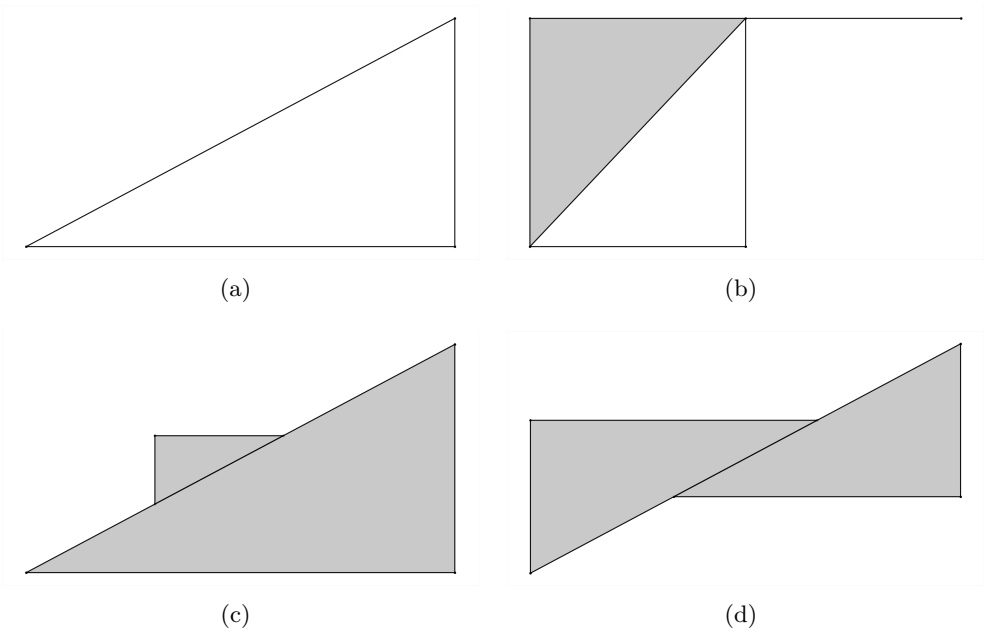
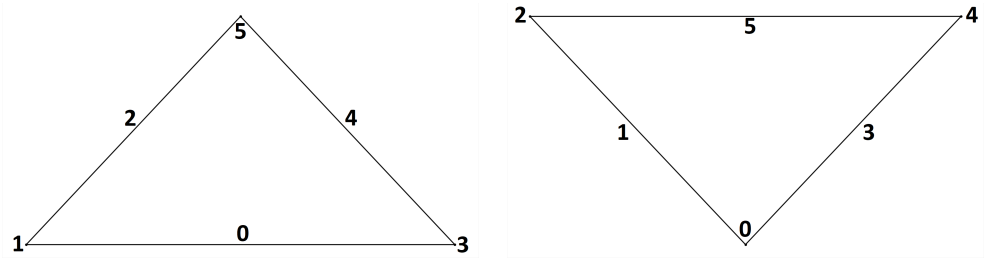


FIGURE 2 : Les Figures (a) et (b) sont des exemples de complexes simpliciaux. Les Figures (c) et (d) ne sont pas des complexes simpliciaux.

Ainsi, avec les complexes simpliciaux, on va définir un champ vectoriel combinatoire en utilisant la capacité combinatoire de cet espace. De plus, il est possible d'utiliser d'autres types d'espaces comme les complexes cubiques ou les complexes cellulaires.



(a) Contre-exemple (b) Exemple

FIGURE 3 : La Figure (a) n'est pas une fonction de Morse discrète. La Figure (b) est une fonction de Morse discrète.

4 La théorie de Morse discrète

4.1 Les champs vectoriels combinatoires selon Forman

Dans cette section, on présente la théorie de Morse avec un point de vue combinatoire. Le but de la théorie de Morse est d'étudier la topologie des variétés différentielles à l'aide des lignes de niveaux d'une fonction et de ses points critiques. Plusieurs résultats importants permettant de mieux comprendre sa topologie comme l'inégalité de Morse. Pour cet article, nous allons voir une approche discrète de la théorie de Morse qui est amené par Robin Forman. Donc, nous allons voir une première esquisse des champs vectoriels combinatoire.

Définition 4.1. Soit K un complexe simplicial. Une fonction $f : K \to \mathbb{R}$ est une *fonction de Morse discrète* si, pour tous $\alpha^{(p)} \in K$, il respecte les deux conditions suivantes :

1. $H_f(\alpha) := \text{card}\left(\left\{\beta^{(p+1)} > \alpha \mid f(\beta) \leq f(\alpha)\right\}\right) \leq 1$;
2. $T_f(\alpha) := \text{card}\left(\left\{\gamma^{(p-1)} < \alpha \mid f(\gamma) \geq f(\alpha)\right\}\right) \leq 1$.

Exemple 4.2. À la Figure 3(a), la fonction f n'est pas une fonction de Morse discrète, car le 0-simplexe σ de valeur 5 ne respecte pas la condition 1, avec $H_f(\sigma) = 2$. De plus, le 1-simplexe τ de valeur 0 ne respecte pas la condition 2, avec $T_f(\tau) = 2$. D'une autre part, la Figure 3(b) respecte les deux conditions. Alors, la fonction f est une fonction de Morse discrète.

Définition 4.3. Un simplexe $\alpha^{(p)}$ est *critique* si les deux conditions suivantes sont satisfaites :

1. $H_f(\alpha) = 0$;
2. $T_f(\alpha) = 0$.

À noter, dans les figures, que les simplexes en rouge sont critiques.

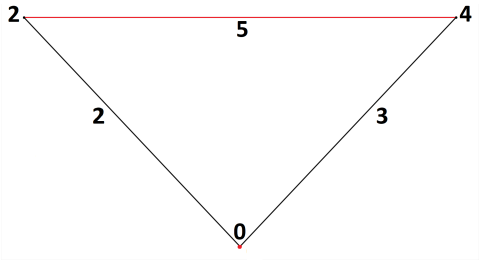


FIGURE 4 : Le 1-simplexe avec la valeur 5 et le 0-simplexe avec la valeur 0 sont critique.

Exemple 4.4. À la Figure 4, on a deux simplexes critiques. Plus précisément, il y a un 0-simplexe et un 1-simplexe. On remarque que le 0-simplexe représente un minimum local et le 1-simplexe représente un maximum local. Dans le cas général, le n-simplexe critique indique qu'il y a n directions descendantes. Donc, si le simplexe de dimension maximale dans le complexe simplicial est critique, alors c'est un maximum local. Pour le simplexe de dimension minimale, il est toujours de dimension 0 et il est un minimum local.

Lemme 4.5. *Soit K un complexe simplicial avec une fonction de Morse f. Alors, pour tous les simplexes $\alpha^{(p)}$, soit $H_f(\alpha) = 0$ ou $T_f(\alpha) = 0$.*

Démonstration. Par contradiction, supposons qu'il existe $\tau \in K$ tel que l'on ait $H_f(\tau) = 1$ et $T_f(\tau) = 1$. Donc, il existe $\sigma \in K$ tel que $\sigma > \tau$ et $f(\sigma) \leq f(\tau)$. De la même façon, il existe γ tel que $\gamma < \tau$ et $f(\gamma) \geq f(\tau)$. De plus, il existe $\tau' \in K$ tel que $\tau' \neq \tau$ et $\gamma < \tau' < \sigma$. On a que $f(\sigma) \leq f(\tau)$ et $f(\gamma) \geq f(\tau)$. Ainsi, on obtient que $f(\tau') < f(\sigma)$ et $f(\tau') > f(\gamma)$. Si $f(\tau') \geq f(\sigma)$, alors $T_f(\sigma) = 2$ ce qui est impossible, car f est une fonction de Morse discrète. De la même façon avec γ, on a que $f(\tau') > f(\gamma)$. Par conséquent, on obtient $f(\tau) \leq f(\gamma) < f(\tau') < f(\sigma) \leq f(\tau)$ d'où la contradiction. □

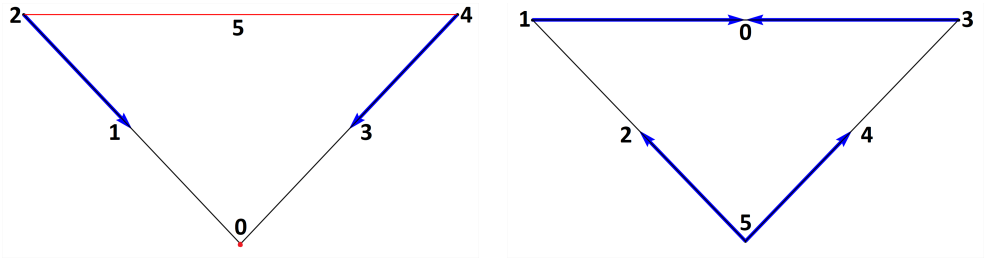
Nous pouvons maintenant définir le champ vectoriel combinatoire selon Forman. Nous utilisons la règle suivante pour dessiner les flèches du champ vectoriel : Soit $\alpha^{(p)}$ une face de $\beta^{(p+1)}$ et $f(\beta) \leq f(\alpha)$, alors on dessine une flèche de α vers β.

À l'aide du lemme précédent, on a que α satisfait à seulement un de ces trois cas :

1. α est la queue d'une seule flèche ;
2. α est la tête d'une seule flèche ;
3. α n'est pas la tête ni la queue d'une flèche.

Remarque 4.6. Dans le cas 3, on a que α est un simplexe critique.

Maintenant, nous pouvons définir les champs vectoriels combinatoires.



(a) (b)

FIGURE 5 : Dans la Figure (a) sont les flèches dessinées avec une fonction de Morse discrète. Dans la Figure (b) sont les flèches dessinées avec une fonction qui ne satisfait pas la définition d'une fonction de Morse discrète.

Définition 4.7. Un *champ vectoriel combinatoire* V sur un complexe simplicial K est un ensemble de pairs $\left\{\alpha^{(p)} < \beta^{(p+1)}\right\}$ de simplexes dans K tel que chaque simplexe est au plus qu'une seule pair dans V.

Ainsi, si on a une fonction de Morse discrète sur un complexe simplicial, alors on est assuré d'avoir un champ vectoriel combinatoire bien défini. Par contre, on s'intéresse aussi au cas inverse. Donc, si on a un champ vectoriel combinatoire V avec les flèches définies plus haut, alors il existe une fonction de Morse discrète qui produit le même champ vectoriel combinatoire que V.

Remarque 4.8. Si le champ vectoriel combinatoire est définie par une fonction de Morse discrète, alors on dit que V est un champ vectoriel combinatoire gradient.

Définition 4.9. Un *v-chemin* est une suite de simplexe :

$$\alpha_0^{(p)}, \beta_0^{(p+1)}, \alpha_1^{(p)}, \beta_1^{(p+1)}, \ldots, \alpha_r^{(p)}, \beta_r^{(p+1)}, \alpha_{r+1}^{(p)} \tag{1}$$

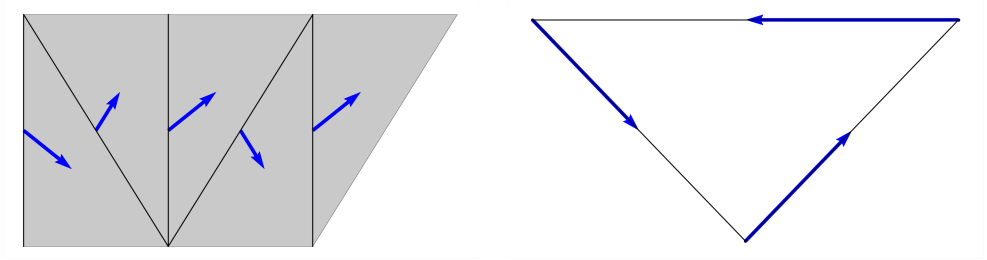
tel que pour chaque $i = 0, 1, \ldots, r$, on a $\{\alpha_i < \beta_i\} \in V$ et $\beta_i > \alpha_{i+1} \neq \alpha_i$.

Remarque 4.10. Si $r \geq 0$ et $\alpha_0 = \alpha_{r+1}$, alors on dit que le v-chemin est un *v-chemin fermé non-trivial*.

Théorème 4.11. *Soit V un champ vectoriel combinatoire gradient d'une fonction de Morse f. Alors, la suite (1) est un v-chemin si et seulement si $\alpha_i < \beta_i > \alpha_{i+1}$ pour $i = 0, 1, \ldots, r$ et*

$$f(\alpha_0) \geq f(\beta_0) > f(\alpha_1) \geq f(\beta_1) > \ldots > f(\alpha_r) \geq f(\beta_r) > f(\alpha_{r+1}).$$

Démonstration. Soit une suite de simplexe qui est un v-chemin. Par la définition de v-chemin, on a $\alpha_i < \beta_i > \alpha_{i+1}$ et $\{\alpha_i < \beta_i\} \in V$. On a aussi $\alpha_{i+1} \neq \alpha_i$, ce qui entraîne que $\{\alpha_{i+1} < \beta_i\} \notin V$. Alors, on obtient $f(\beta_i) > f(\alpha_{i+1}), \forall i$. De l'autre côté, si on a $\{\alpha_i < \beta_i\} \in V$, alors $f(\alpha_i) \geq f(\beta_i), \forall i$. Ainsi, on obtient la suite suivante : $f(\alpha_0) \geq f(\beta_0) > f(\alpha_1) \geq f(\beta_1) > \ldots > f(\alpha_r) \geq f(\beta_r) > f(\alpha_{r+1})$.



(a) (b)

FIGURE 6 : La Figure (a) est un exemple de v-chemin. Le début du v-chemin est le 1-simplexe à l'extrémité à gauche et se termine au 2-simplexe à droite. La Figure (b) est un exemple de v-chemin fermé non-trivial

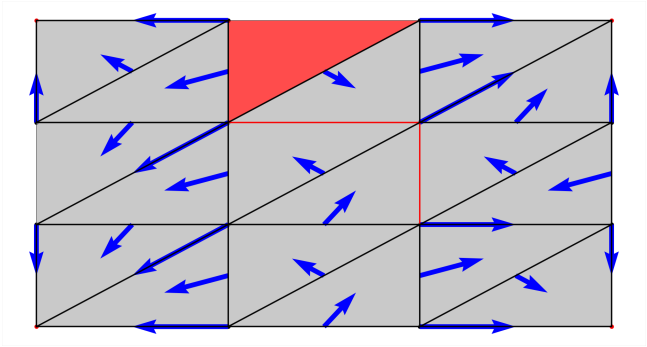


FIGURE 7 : C'est un exemple d'une fonction de Morse discrète sur un tore.

De l'autre côté, on considère une suite de simplexes telle que l'on a les inégalités $f(\alpha_0) \geq f(\beta_0) > f(\alpha_1) \geq f(\beta_1) > \ldots > f(\alpha_r) \geq f(\beta_r) > f(\alpha_{r+1})$. Si $f(\alpha_i) \geq f(\beta_i)$, alors on obtient $\{\alpha_i < \beta_i\} \in V$. De plus, si $f(\beta_i) > f(\alpha_{i+1})$, alors on a $\{\alpha_{i+1} < \beta_i\} \notin V$, ce qui entraîne que $\alpha_i \neq \alpha_{i+1}$. Donc, la suite : $\alpha_0, \beta_0, \alpha_1, \beta_1, \ldots, \alpha_r, \beta_r, \alpha_{r+1}$ est un v-chemin. □

Remarque 4.12. La fonction est décroissante le long du v-chemin.

Théorème 4.13. *Un champ vectoriel combinatoire V est un champ vectoriel gradient issu d'une fonction de Morse discrète si et seulement s'il ne possède aucun v-chemin fermé non-trivial.*

Nous allons démontrer ce résultat à la Section 4.2.

Exemple 4.14. La Figure 7 est un tore représenté dans $\mathbb{R}^2$ sur l'intervalle $[0,1] \times [0,1]$. Pour l'obtenir, il faut coller les côtés verticaux ensemble et les côtés horizontaux, c'est-à-dire, que les côtés opposés sont équivalents de cette manière : $(t,0) \sim (t,1)$ et $(0,t) \sim (1,t)$ pour $\forall\, t \in [0,1]$.

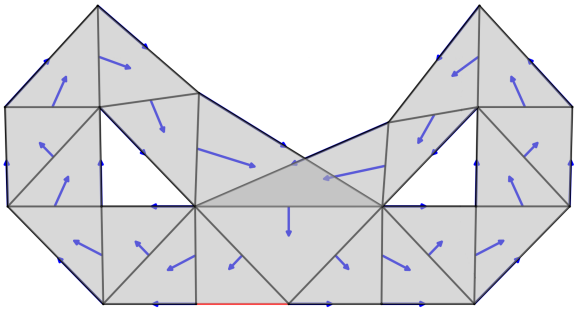


FIGURE 8 : C'est un exemple d'une fonction qui n'est pas une fonction de Morse discrète.

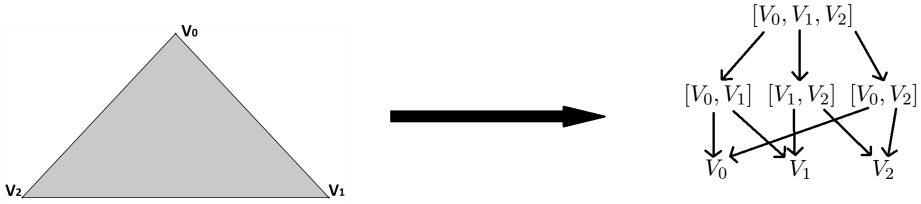


FIGURE 9 : C'est le diagramme de Hasse d'un complexe simplicial.

Exemple 4.15. À la Figure 8, on remarque qu'il existe des v-chemins fermés non-triviaux. Alors, la fonction ne genère pas un champ vectoriel produit par une fonction de Morse discrète. De plus, c'est un exemple combinatoire des attracteurs de Lorenz qui montre qu'il existe des comportements chaotiques.

4.2 Le diagramme de Hasse modifiée

Maintenant, transformons un complexe simplicial à un diagramme de Hasse et appliquons quelques modifications pour obtenir les mêmes informations qu'un champ vectoriel combinatoire.

D'abord, on construit le diagramme de Hasse à l'aide de la relation des faces de co-dimension 1 entre les simplexes, c'est-à dire qu'on trace une flèche ayant comme source le $(p-1)$-simplexe et comme destination le p-simplexe, si le p-simplexe est une face du $(p+1)$-simplexe. Donc, on obtient la Figure 9.

Ensuite, on utilise le champ vectoriel du complexe simplicial pour interchanger la source et la destination. S'il y a une flèche $\left\{\alpha^{(p)} < \beta^{(p+1)}\right\}$, alors on inverse la flèche sur le diagramme de Hasse. Donc, si on applique cette procédure pour toutes les flèches du champ vectoriel, alors on obtient le diagramme de Hasse modifié.

Ainsi, avec le diagramme de Hasse modifié, on obtient la même information que des v-chemins à l'intérieur du champ vectoriel combinatoire, car, en suivant les chemins à l'intérieur du diagramme de Hasse modifié, on obtient les che-

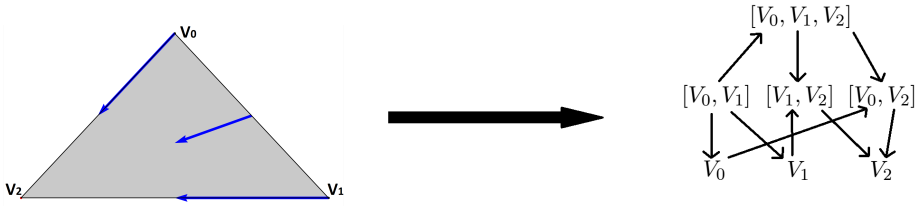


FIGURE 10 : C'est le diagramme de Hasse modifié d'une fonction de Morse discrète.

mins qui sont décroissants dans le graphe, comme les v-chemins dans le champ vectoriel combinatoire. Alors, on obtient le résultat suivant :

Théorème 4.16. *Il n'y a pas de v-chemin fermé non-trivial si et seulement s'il n'y a pas de chemin orienté non-trivial dans le diagramme de Hasse modifié.*

Donc, le Théorème 4.13 est équivalent à un théorème connu dans la théorie des graphes.

Théorème 4.17. *Soit $\mathcal{G}$ un graphe orienté. Alors, il y a une fonction à valeur réelle qui est strictement décroissante le long des chemins orientés si et seulement s'il n'existe pas de cycle dans $\mathcal{G}$.*

Par exemple, on peut utiliser un algorithme de parcours en profondeur pour trouver s'il n'y a pas de cycle et on peut conclure si la fonction du champ vectoriel est un champ vectoriel combinatoire gradient issu d'une fonction de Morse discrète ou non.

Bref, on remarque que la théorie de Morse discrète de Forman utilise le complexe simplicial (on peut aussi utiliser le complexe cellulaire) et appliquer une fonction de Morse discrète pour obtenir un champ vectoriel combinatoire. Par contre, les définitions ne pourront pas être utilisées pour définir les systèmes dynamiques combinatoires.

5 Les systèmes dynamiques combinatoires

Dans la théorie de Morse discrète de Forman, on décrit seulement le champ vectoriel gradient, car dans la théorie de Morse continue, il y a un lien important entre le champ vectoriel gradient et la fonction de Morse. De plus, le v-chemin considère seulement les trajectoires de dimension p et $p+1$. Ainsi, dans notre cas, on veut étudier les systèmes dynamiques combinatoires. Alors, il faut généraliser la définition de champ vectoriel combinatoire et ajouter des concepts propres aux systèmes dynamiques, mais nous utiliserons l'intuition que nous avons vue dans les définitions précédentes. De plus, à l'aide de la décomposition de Morse, nous pourrons comprendre le dynamisme du système.

Définition 5.1. Soit K un complexe simplicial. Alors, $\mathcal{V} : K \to K$ est un *champ vectoriel combinatoire* si les trois conditions suivantes sont satisfaites :

1. Pour tout simplexe $\sigma \in \text{Dom } \mathcal{V}$, nous avons $\mathcal{V}(\sigma) = \sigma$ ou σ est une face de co-dimension 1 de $\mathcal{V}(\sigma)$;
2. $\text{Dom } \mathcal{V} \cup \text{Im } \mathcal{V} = K$;
3. $\text{Dom } \mathcal{V} \cap \text{Im } \mathcal{V} = \text{Fix } \mathcal{V}$, où $\text{Fix } \mathcal{V}$ est l'ensemble des points fixes du champ vectoriel combinatoire, c'est-à-dire les points critiques dans le cas du champ vectoriel combinatoire de Forman.

Autrement dit, la première condition signifie qu'on permet aux points fixes de ne pas bouger. De plus, on veut que les flèches du champ vectoriel pointent vers un simplexe de dimension supérieure à un. Pour la deuxième condition, on ne veut pas de simplexe qui ne soit ni un point fixe, ni la source d'une flèche et ni la destination d'une flèche. Pour la dernière, on veut seulement que les points fixes soient à l'intérieur de l'image du domaine. De plus, on ne veut pas qu'un simplexe soit la source et la destination de deux flèches.

Définition 5.2. Un *multiflot combinatoire* associé à un champ vectoriel combinatoire $\mathcal{V}$ est défini par une multifonction $\Pi_{\mathcal{V}} : K \to K$

$$\Pi_{\mathcal{V}}(\sigma) := \begin{cases} \text{Cl } \sigma & \text{si } \sigma \in \text{Fix } \mathcal{V} \\ \text{Cl Bd } \sigma \setminus \{\mathcal{V}^{-1}(\sigma)\} & \text{si } \sigma \in \text{Im } \mathcal{V} \setminus \text{Fix } \mathcal{V} \\ \{\mathcal{V}(\sigma)\} & \text{si } \sigma \in \text{Dom } \mathcal{V} \setminus \text{Fix } \mathcal{V} \end{cases}.$$

Exemple 5.3. À la Figure 11, soit $\sigma_1 = [V_2, V_3, V_5] \in \text{Fix } \mathcal{V}$, alors

$$\Pi_{\mathcal{V}}(\sigma_1) = \{[V_2], [V_3], [V_5], [V_2, V_3], [V_3, V_5], [V_2, V_5], [V_2, V_3, V_5]\}.$$

Soit $\sigma_2 = [V_1, V_2, V_3] \in \text{Im } \mathcal{V}$, alors

$$\Pi_{\mathcal{V}}(\sigma_2) = \{[V_1, V_3], [V_1, V_2], [V_1], [V_2], [V_3]\}.$$

Soit $\sigma_3 = [V_4, V_6] \in \text{Dom } \mathcal{V}$, alors

$$\Pi_{\mathcal{V}}(\sigma_3) = \{[V_4, V_5, V_6]\}.$$

Maintenant, nous pouvons définir un chemin à l'aide du multiflot combinatoire. On remarque que c'est une généralisation du v-chemin. Il est maintenant possible de faire un chemin contenant des simplexes de plusieurs dimensions, tandis que le v-chemin pouvait seulement contenir les simplexes de dimensions p et $p+1$ pour $p \geq 0$. De plus, le v-chemin ne pouvait pas posséder de point critique sauf pour le dernier simplexe de la suite du v-chemin.

Définition 5.4. Une *solution* d'un multiflot combinatoire $\Pi_{\mathcal{V}}$, aussi appelée orbite, d'un champ vectoriel combinatoire $\mathcal{V}$ est une fonction $\varrho : I \to K$ tel que I est un intervalle dans $\mathbb{Z}$ et $\varrho_{i+1} \in \Pi_{\mathcal{V}}(\varrho_i)$ pour $i, i+1 \in I$. De plus, c'est une *solution complète* si $I = \mathbb{Z}$.

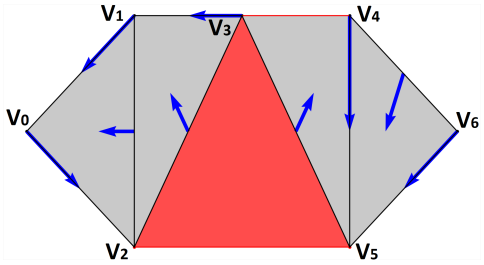


FIGURE 11 : C'est un exemple d'un champ vectoriel combinatoire.

Exemple 5.5. Dans la Figure 11, on a une solution : $[V_2, V_3, V_5] \to [V_3, V_5] \to [V_3, V_4, V_5] \to [V_3, V_4] \to [V_3, V_4] \to [V_4] \to [V_4, V_5] \to [V_5]$ où l'intervalle est $I = \{0, 1, 2, 3, 4, 5, 6, 7\}$.

Maintenant, on aimerait étudier ces systèmes dynamiques combinatoires plus en profondeur. Plus précisément, on voudrait connaître les zones du système dynamique tel que les solutions convergent vers une zone ou même les solution complète qu'ils demeurent entièrement à l'intérieur d'une zone. C'est une manière de mieux comprendre le comportement des systèmes dynamiques.

Définition 5.6. Soit $\mathcal{V}$ un champ vectoriel combinatoire sur un complexe simplicial K. Alors, un ensemble $S \subset K$ est un *invariant* associé à un multiflot $\Pi_{\mathcal{V}}$, si pour chaque simplexe $\sigma \in S$, il existe une solution complète $\varrho : \mathbb{Z} \to K$ passant par σ telle qu'elle est complètement contenu dans S.

Exemple 5.7. À la Figure 11, $S = \{[V_5], [V_2, V_5], [V_2, V_3, V_5], [V_3, V_5], [V_3, V_4, V_5], [V_3, V_4]\}$ est un invariant, car il y a une solution complète pour chaque simplexe. $\varrho_1 : \ldots \to [V_2, V_3, V_5] \to [V_2, V_3, V_5] \to [V_3, V_5] \to [V_3, V_4, V_5] \to [V_3, V_4] \to [V_3, V_4] \to \ldots$ et $\varrho_2 : \ldots \to [V_2, V_5] \to [V_2, V_5] \to [V_5] \to [V_5] \to [V_5] \to \ldots$. Les deux solutions complètes restent complètement dans S et ils atteignent tous les simplexes de S. Ainsi, S est un invariant.

Définition 5.8. Soient $\mathcal{V}$ un champ vectoriel combinatoire sur un complexe simplicial K et $S \subset K$ un ensemble invariant pour un multiflot $\Pi_{\mathcal{V}}$. Définissons l'ensemble des sorties de S par :

$$\text{Ex } S := \text{Cl } S \setminus S.$$

Alors, S est un *invariant isolé* si les deux conditions suivantes sont satisfaites :

1. L'ensemble des sorties Ex S est fermé dans K ;
2. Il n'existe pas de solution $\varrho : [-1,1] \cap \mathbb{Z} \to K$ de $\Pi_{\mathcal{V}}$ tel que $\varrho(-1) \in S$, $\varrho(1) \in S$ et $\varrho(0) \in \text{Ex } S$.

La deuxième condition signifie qu'on ne veut pas avoir de trajectoires tangentes à l'intérieur de l'invariant isolé, c'est-à-dire qu'on ne veut pas qu'un point,

en un temps t, soit dans Ex S et que les points aux temps $t+1$ et $t-1$ soient dans l'invariant isolé. Cette définition est une adaptation, dans un point de vue combinatoire, du cas continu.

Exemple 5.9. À la Figure 11, soit $S_1 = \{[V_2, V_3V_5], [V_2, V_5], [V_2], [V_5]\}$. Alors, Ex $S_1 = \{[V_2, V_3], [V_3, V_5], [V_3]\}$ et Cl Ex $S_1 = \{[V_2, V_3], [V_3, V_5], [V_3], [V_5], [V_2]\}$ et on obtient Cl Ex $S_1 \neq$ Ex S_1. Donc, S_1 n'est pas un invariant isolé.

Soit $S_2 = \{[V_4], [V_5], [V_3, V_4], [V_4, V_5]\}$. Alors, on a que Ex $S_2 = \{[V_3]\}$ et également que Cl Ex $S_2 = \{[V_3]\} =$ Ex S_2. Donc, la première condition est satisfaite. De plus, il n'existe pas de solution complète ϱ telle que $\varrho(-1) \in S_2$, $\varrho(1) \in S_2$ et $\varrho(0) \in$ Ex S_2. Ainsi, S_2 est un invariant isolé.

Cette approche de trouver des invariants isolés n'est pas toujours évidente, car il est parfois difficile de démontrer qu'il n'existe pas de solution complète contenue complètement dans le complexe simplicial pour chaque simplexe de l'invariant isolé. Alors, on va définir de nouveaux simplexes permettant d'avoir une meilleure approche combinatoire.

Définition 5.10. Définissons les simplexes suivants :

$$\sigma^+ := \begin{cases} \mathcal{V}(\sigma) & \text{si } \sigma \in \text{Dom } \mathcal{V} \\ \sigma & \text{sinon} \end{cases} \quad \text{et} \quad \sigma^- := \begin{cases} \sigma & \text{si } \sigma \in \text{Dom } \mathcal{V} \\ \mathcal{V}^{-1}(\sigma) & \text{sinon} \end{cases}.$$

Le simplexe σ^- est la source et le simplexe σ^+ est la destination de la flèche associée à σ. Donc, si σ n'est pas la source ou la destination, alors, selon le cas, σ^+ ou σ^- est égale à σ.

Lemme 5.11. *Soit $\mathcal{V}$ un champ vectoriel combinatoire sur un complexe simplicial K. Alors, pour un simplexe arbitraire $\sigma \in K$, on a que $\sigma^- \subset \sigma \subset \sigma^+$ et au moins une des inclusions est une égalité.*

Démonstration. Si $\sigma \in$ Dom $\mathcal{V}$, alors $\sigma^+ = \mathcal{V}(\sigma)$. Mais, par la Définition 5.1 de champ vectoriel combinatoire $\mathcal{V}$, on a que $\mathcal{V}(\sigma) = \sigma$ ou σ est une face de co-dimension 1 de $\mathcal{V}(\sigma)$. Cela implique que $\sigma \subseteq \mathcal{V}(\sigma)$. De plus, on a $\sigma^- = \sigma$. Ainsi, on obtient $\sigma = \sigma^- = \sigma \subseteq \sigma^+ = \mathcal{V}(\sigma)$.

Si $\sigma \notin$ Dom $\mathcal{V}$, alors $\sigma^+ = \sigma$ et $\sigma^- = \mathcal{V}^{-1}(\sigma)$. Mais par la définition de champ vectoriel combinatoire $\mathcal{V}$, $\mathcal{V}^{-1}(\sigma) = \sigma$ ou $\mathcal{V}^{-1}(\sigma)$ est une face de co-dimension 1 de σ, ce qui implique que $\mathcal{V}^{-1}(\sigma) \subseteq \sigma$. Ainsi, $\mathcal{V}^{-1}(\sigma) = \sigma^- \subseteq \sigma = \sigma^+ = \sigma$. Donc, on obtient au moins une égalité dans les deux cas. □

Dans le cas des points fixes, on obtient $\sigma^- = \sigma = \sigma^+$.

Lemme 5.12. *Soit $\mathcal{V}$ un champ vectoriel combinatoire sur un complexe simplicial K et S un invariant isolé d'un multiflot combinatoire $\Pi_{\mathcal{V}}$. Alors, pour tous les simplexes $\sigma \in K$, on a que $\sigma^+ \in S$ si et seulement si $\sigma^- \in S$.*

En d'autres mots, on veut que σ^+ et σ^- soient à l'intérieur ou à l'extérieur de l'invariant isolé.

Démonstration. [TMT16, Lemme 3.6, p.12] Tout d'abord, si $\sigma \in \text{Fix } \mathcal{V}$, alors $\sigma = \sigma^+ = \sigma^-$. Donc, le résultat est trivial à l'aide du Lemme 5.11.

Supposons que $\sigma \notin \text{Fix } \mathcal{V}$. Alors, on a $\sigma^+ \neq \sigma^-$, et par la Définition 5.10 de σ^+ et σ^-, on obtient que $\mathcal{V}(\sigma^-) = \sigma^+$.

Soit $\sigma^- \in S$. On a que S est un invariant isolé. Alors, il existe une solution complète $\varrho : \mathbb{Z} \to S$ telle que $\sigma^- = \varrho(\sigma)$. Par le multiflot combinatoire $\Pi_{\mathcal{V}}$, on obtient $\varrho(1) \in \Pi_{\mathcal{V}}(\varrho(0)) = \Pi_{\mathcal{V}}(\sigma^-)$. Si $\sigma \in \text{Dom } \mathcal{V}$, alors $\sigma^- = \sigma$. Donc $\Pi_{\mathcal{V}}(\sigma^-) = \{\mathcal{V}(\sigma^-)\} = \sigma^+$. Sinon, $\sigma \in \text{Im } \mathcal{V}$. Alors, $\sigma^- = \mathcal{V}^{-1}(\sigma) \in \text{Dom } \mathcal{V}$. Donc, $\Pi_{\mathcal{V}}(\sigma^-) = \Pi_{\mathcal{V}}(\mathcal{V}^{-1}(\sigma)) = \sigma = \sigma^+$. Ainsi, $\sigma^+ = \varrho(1) \in S$.

De l'autre côté, soit $\sigma^+ \in S$. De la même manière, S est un invariant isolé, alors il existe une solution $\varrho : \mathbb{Z} \to S$ telle que $\sigma^+ = \varrho(0)$. Soit $\tau = \varrho(-1)$. On obtient $\sigma^+ = \varrho(0) \in \Pi_{\mathcal{V}}(\varrho(-1)) = \Pi_{\mathcal{V}}(\tau)$. On a deux cas à vérifier.

Pour le premier cas, si $\tau \in \text{Dom } \mathcal{V} \setminus \text{Fix } \mathcal{V}$, alors, $\Pi_{\mathcal{V}}(\tau) = \{\mathcal{V}(\tau)\}$ et $\sigma^+ = \mathcal{V}(\tau)$. Donc $\sigma^- = \tau \in S$.

Pour le deuxième cas, supposons que $\tau \notin \text{Dom } \mathcal{V} \setminus \text{Fix } \mathcal{V}$. D'abord, on a que si $\sigma \notin \text{Fix } \mathcal{V}$, alors $\sigma^+ \notin \text{Fix } \mathcal{V}$. Donc, $\sigma^+ < \tau$ est une face de co-dimension 1. De plus, si on a que $\sigma^- = \mathcal{V}^{-1}(\sigma^+)$, alors $\sigma^- < \sigma^+$ est une face de co-dimension 1 de même que pour τ. Ainsi, $\sigma^- \in \Pi_{\mathcal{V}}(\tau)$, car les deux images possibles du multiflot combinatoire contiennent tous les simplexes et toutes ses faces.

On construit $\phi : [-1, 1] \cap \mathbb{Z} \to K$, en posant $\phi(-1) := \tau$, $\phi(0) := \sigma^-$ et $\phi(1) := \sigma^+ \in S$, alors ϕ est une solution de $\Pi_{\mathcal{V}}$. Comme S est un invariant isolée et $\sigma^- \in \text{Cl } \sigma^+$ par la Définition 5.8 , ainsi $\sigma^- \in S$. □

À l'aide de ce lemme, nous pouvons alléger la définition d'invariant isolé (plus précisément la deuxième condition) pour faciliter la recherche d'invariant isolé. Nous allons admettre la prochaine proposition sans démonstration.

Proposition 5.13. *Un invariant S est un invariant isolé s'il satisfait les deux conditions suivantes :*

1. *L'ensemble des sorties* Ex S *est fermé dans le complexe simplicial K ;*

2. *Pour tous simplexes $\sigma \in K$, on a que $\sigma^- \in S$ si et seulement si $\sigma^+ \in S$.*

Exemple 5.14. À la Figure 12, soient $S_1 = \{[V_4], [V_1, V_4], [V_4, V_6], [V_1, V_6], [V_1, V_4, V_6], [V_1, V_2, V_6]\}$ et $S_2 = \{[V_0], [V_0, V_1], [V_0, V_2], [V_0, V_3], [V_0, V_1, V_2], [V_0, V_2, V_3]\}$ et vérifions, à l'aide de la Proposition 5.13, si S_1 et S_2 sont des invariants isolés.

Pour S_1, on a Ex $S_1 = \{[V_1], [V_2], [V_6], [V_1, V_2], [V_2, V_6]\} = \text{Cl Ex } S$. Alors, la condition 1 est satisfaite. Calculons σ^+ et σ^- de $[V_6, V_4]$. On a $\sigma^+ = [V_6, V_4] \in S$ et $\sigma^- = [V_6] \notin S$. Alors la deuxième condition n'est pas satisfaite. Donc, S_1 n'est pas un invariant isolé.

Pour S_2, on a Ex $S_2 = \{[V_1], [V_2], [V_3], [V_1, V_2], [V_2, V_3]\} = \text{Cl Ex } S_2$. Alors, la condition 1 est satisfaite. Calculons les σ^+ et σ^- de S_2.

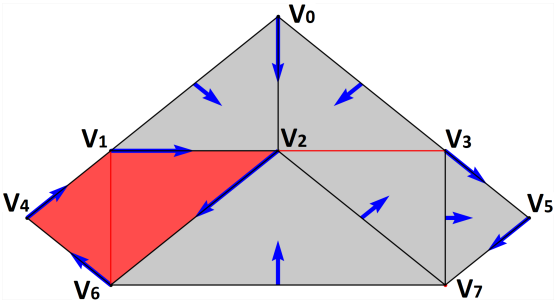


FIGURE 12 : C'est un exemple pour vérifier qu'un ensemble est un invariant isolé en utilisant la Proposition 5.13.

$$
\begin{array}{lll}
\sigma_1 = [V_0]\,, & \sigma_1^+ = [V_0, V_2] \in S_2 & \text{et } \sigma_1^- = [V_0] \in S_2 \\
\sigma_2 = [V_0, V_1]\,, & \sigma_2^+ = [V_0, V_1, V_2] \in S_2 & \text{et } \sigma_2^- = [V_0, V_1] \in S_2 \\
\sigma_3 = [V_0, V_2]\,, & \sigma_3^+ = [V_0, V_2] \in S_2 & \text{et } \sigma_3^- = [V_0] \in S_2 \\
\sigma_4 = [V_0, V_3]\,, & \sigma_4^+ = [V_0, V_2, V_3] \in S_2 & \text{et } \sigma_4^- = [V_0, V_3] \in S_2 \\
\sigma_5 = [V_0, V_1, V_2]\,, & \sigma_5^+ = [V_0, V_1, V_2] \in S_2 & \text{et } \sigma_5^- = [V_0, V_1] \in S_2 \\
\sigma_6 = [V_0, V_2, V_3]\,, & \sigma_6^+ = [V_0, V_2, V_3] \in S_2 & \text{et } \sigma_6^- = [V_0, V_3] \in S_2
\end{array}
$$

Donc, S_2 est un invariant isolé.

Ainsi, cette proposition nous permet d'avoir une meilleure méthode pour calculer les invariants isolés qui reflètent mieux la philosophie de la combinatoire, car de prouver qu'il n'existe pas de solution complète peut être très complexe et fastidieux quand l'invariant qu'on vérifie a une très grande quantité de simplexes. Donc, pour chaque simplexe, on doit seulement vérifier σ^+ et σ^-, qui consiste en deux calculs simples et on a seulement besoin de savoir si σ est dans le domaine ou dans l'image du champ vectoriel combinatoire.

Maintenant qu'on a une méthode simple pour calculer les invariants isolés, on aimerait comprendre l'interaction entre les invariants isolés dans le système dynamique. D'abord, on a besoin de comprendre le comportement des solutions complètes dans le futur et dans le passé, c'est-à-dire quand le temps tend à l'infini ou à moins l'infini.

Définition 5.15. Soit $\varrho : \mathbb{Z} \to K$ une solution complète d'un multiflot combinatoire $\Pi_{\mathcal{V}}$. On définit les deux ensembles limites suivants :

1. L'ensemble *α-limite de la solution complète* ϱ est :
$$\alpha(\varrho) := \bigcap_{n \in \mathbb{Z}} \{\varrho(k) \,|k \leq n\}\,.$$
2. L'ensemble *ω-limite de la solution complète* ϱ est :
$$\omega(\varrho) := \bigcap_{n \in \mathbb{Z}} \{\varrho(k) \,|k \geq n\}\,.$$

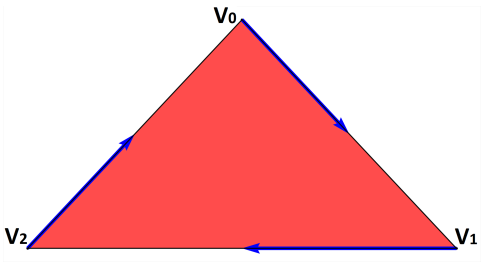


FIGURE 13 : C'est un exemple de calcul des ensembles α-limite et ω-limite.

Cette notion d'ensemble limite vient de la dynamique classique. Par contre, pour le contexte de la dynamique combinatoire, tenant compte du fait que l'intersection est prise sur une famille décroissante d'un nombre fini des ensembles, les ensembles α-limite et ω-limite sont atteints dans un temps entier fini. Par exemple, ces ensembles peuvent être des simplexes critiques répulsifs pour l'ensemble α-limite, des simplexes critiques attractifs pour l'ensemble ω-limite ou des trajectoires cycliques.

Exemple 5.16. À la Figure 13, soit ϱ la solution complète suivante : $\varrho(\mathbb{Z}) = \ldots \to [v_0, v_1, v_2] \to [v_0, v_1, v_2] \to [v_0, v_1, v_2] \to [v_0] \to [v_0, v_1] \to [v_1] \to [v_1, v_2] \to [v_2] \to [v_2, v_0] \to [v_0] \to [v_0, v_1] \to [v_1] \to \ldots$. Donc, on obtient les ensembles limites de cette solution complète :

$$\alpha(\varrho) = \{[v_0, v_1, v_2]\} \text{ et}$$

$$\omega(\varrho) = \{[v_0], [v_1], [v_2], [v_0, v_1], [v_1, v_2], [v_0, v_2]\}.$$

Ainsi, on a tous les ingrédients nécessaires pour définir la décomposition de Morse. Cette décomposition nous permettra de diviser le complexe simplicial en invariant isolé et de comprendre à l'intérieur de quel invariant isolé proviennent les solutions complètes ou l'avenir des solutions complètes.

Définition 5.17. Soit S un ensemble contenant des ensembles invariants isolés de $\Pi_{\mathcal{V}} : K \to K$. On dit qu'une famille $\mathcal{M} := \{\mathcal{M}_i \mid i \in \mathbb{I}\}$, indexé par un ensemble partiellement ordonné $\mathbb{I}$, est une *décomposition de Morse de* S si les conditions suivantes sont satisfaites :

1. Les éléments de M sont mutuellement disjoints et des sous-ensembles invariants isolés de S ;

2. Pour toutes les solutions complètes ϱ contenues dans K, il existe $r, r' \in \mathbb{I}$, $r \leq r'$, tel que $\alpha(\varrho) \subseteq \mathcal{M}_{r'}$ et $\omega(\varrho) \subseteq \mathcal{M}_r$;

3. Si pour une solution complète quelconque ϱ contenu dans K et $r \in \mathbb{I}$, on obtient que $\alpha(\varrho) \cup \omega(\varrho) \subseteq \mathcal{M}_r$, alors Im $\varrho \subseteq \mathcal{M}_r$.

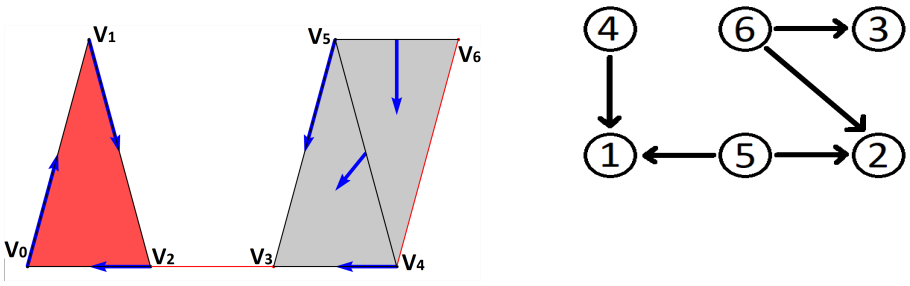


FIGURE 14 : C'est un champ vectoriel combinatoire avec son graphe de Morse.

La première condition signifie qu'on veut que les invariants isolés soient disjoints pour mieux comprendre la source et la destination des solutions complètes. La deuxième condition décrit l'interaction des ensembles limites d'une solution complète entre les invariants isolés. La troisième condition décrit les solutions complètes qui sont complètement contenues dans un invariant isolé. De plus, on veut que si une solution complète sort d'un invariant isolé, alors la solution complète doit se rendre à un autre invariant isolé et elle doit absolument converger dans cet invariant isolé.

Avec ces données, on peut maintenant construire un graphe de Morse permettant de visualiser le dynamisme du système. On a que les noeuds du graphe sont les invariants isolés. Les arêtes du graphe sont orientées et elles sont construites à l'aide de la deuxième condition de la Définition 5.17, de cette manière : s'il existe une solution complète qui satisfait la condition, alors on dessine une arête orientée ayant comme source $\mathcal{M}_{r'}$ et comme destination $\mathcal{M}_r$.

Exemple 5.18. À la Figure 14, on obtient la décomposition de Morse avec les invariants isolés suivants : $\mathcal{M}_4 = \{[V_0, V_1, V_2]\}$, $\mathcal{M}_1 = \{[V_0], [V_1], [V_2], [V_0, V_1], [V_1, V_2], [V_0, V_2]\}$, $\mathcal{M}_5 = \{[V_2, V_3]\}$, $\mathcal{M}_2 = \{[V_3]\}$, $\mathcal{M}_3 = \{[V_6]\}$ et $\mathcal{M}_6 = \{[V_4, V_6]\}$. De plus, on a les trajectoires suivantes entre les invariants isolés : $\mathcal{M}_4 \to \mathcal{M}_1$, $\mathcal{M}_5 \to \mathcal{M}_1$, $\mathcal{M}_5 \to \mathcal{M}_2$, $\mathcal{M}_6 \to \mathcal{M}_2$ et $\mathcal{M}_6 \to \mathcal{M}_3$. Donc, on a que cette décomposition respecte les conditions de la décomposition de Morse. De plus, remarquons qu'il y a trois ensembles attractifs ($\mathcal{M}_1$, $\mathcal{M}_2$ et $\mathcal{M}_3$). D'autre part, les ensembles ($\mathcal{M}_4$, $\mathcal{M}_5$ et $\mathcal{M}_6$) sont répulsifs quand la solution complète sort de l'invariant isolé.

En résumé, cette décomposition de Morse nous permet de comprendre le dynamisme du multiflot combinatoire associé aux champs vectoriels combinatoires. Plus précisément, on peut prédire les avenir possibles des solutions en connaissant l'existence des zones attractives, des zones répulsives et les liens entre eux.

Pour finir, nous avons vu une généralisation de la définition de champ vectoriel combinatoire et des v-chemins dans la théorie de Morse discrète selon Forman. À l'aide de cette intuition, cela nous a permis de définir le concept de systèmes dynamiques combinatoires et le multiflot combinatoire. Ensuite, nous

avons défini les solutions pour pouvoir comprendre les trajectoires à l'intérieur du système dynamique. Par la suite, on a introduit les invariants isolés qui ont permis de décrire la décomposition de Morse pour mieux comprendre la dynamique d'un système.

6 Conclusion

On a vu une approche aux systèmes dynamiques combinatoires en utilisant un espace construit par des complexes simpliciaux et en s'inspirant de la théorie de Morse discrète de Forman. De plus, on a défini les orbites d'un système dynamique discret et ses invariants. Par la suite, on a défini la décomposition de Morse pour décrire le dynamisme du système. Il y a encore quelques problèmes que nous n'avons pas considérés. Tout d'abord, il y a d'autres types de champ vectoriel. Il y a les champs multi-vecteurs, c'est-à dire qu'il pourrait avoir plusieurs vecteurs partant de la même source et destination. De plus, on a seulement considéré des fonctions ayant une image dans $\mathbb{R}$, alors qu'on pourrait considérer les fonctions $f : K \to \mathbb{R}^n$. En outre, il reste d'autres concepts à définir dans les systèmes dynamiques pour les champs vectoriels. Plus précisement, pour la décomposition de Morse, on peut calculer les indices de Conley sur les invariants isolés pour davantage comprendre le système dynamique. De plus, on peut construire un flot F semi-continu fortement supérieur ayant le même dynamisme que le système dynamique discret défini par le multiflot combinatoire. Ainsi, les systèmes dynamiques combinatoires sont utiles pour expliquer plusieurs phénomènes combinatoires avec la puissance de l'informatique augmentant jour après jour.

Références

[BTMW16] Batko Bogdan, Kaczynski Tomasz, Mrozek Marian et Thomas Wanner : Linking Combinatorial and Classical Dynamics : Conley Index and Morse Decompositions. *preprint arXiv :1710.05802 [math.DS]*, page 43, 2016.

[Jam84] R. Munkres James : *Elements of Algebraic Topology.* Addison-Weslay, Cambridge, 1984.

[Mat02] Yukio Matsumoto : *An Introduction to Morse Theory.* American Mathematical Society, États-Unis, 2002.

[Rob98] Forman Robin : Morse Theory for Cell Complexes. *Advances in Mathematics*, 134(AI971650):90–145, 1998.

[Rob02] Forman Robin : A User's Guide to Discrete Morse Theory. *Séminaire Lotharingien de Combinatoire*, 48(B48c):35, 2002.

[TMT16] Kaczynski Tomasz, Mrozek Marian et Wanner Thomas : Towards a Formal Tie Between Combinatorial and Classical Vector Field Dynamics. *Journal of Computational Dynamics*, 3(1):17–50, 2016.

Dominic Desjardins Côté
Département de mathématiques, Université de Sherbrooke
Courriel: dominic.desjardins.cote@usherbrooke.ca

CaMUS **7**, 21 – 32

Fonctions booléennes d'addition binaire

Julien Corriveau-Trudel

RÉSUMÉ L'objectif de cet article est de familiariser le lecteur avec l'algèbre booléenne et les fonctions booléennes utilisées en informatique, aussi appelées « portes logiques ». L'exemple au coeur de l'article est l'additionneur binaire. Seul un niveau collégial en mathématiques est requis pour comprendre cet article.

1 Introduction

En 1937, Claude Shannon montre dans sa thèse de maîtrise [Sha37] qu'il y a des applications électroniques de l'algèbre booléenne, donnée par George Boole en 1854 dans The Laws of Thought [Boo54], notamment dans les circuits logiques et numériques. Aujourd'hui, on n'a qu'à suivre un cours d'introduction à la programmation pour découvrir que les ordinateurs conventionnels fonctionnent à l'aide de 1 et de 0 (de « on » et de « off », de circuit passant et non passant). Il est possible de décrire les opérations à la base du calcul informatique à l'aide de fonctions booléennes, soit des fonctions qui prennent des 1 et des 0 en entrées et qui renvoient 1 ou 0. En particulier, l'addition, lorsqu'appliquée aux chiffres d'un nombre en représentation binaire, s'écrit comme une composition de fonctions booléennes. C'est cette addition qui est décrite et vulgarisée dans le présent article.

Dans les quelques chapitres qui suivent, quelques concepts de base de l'algèbre booléenne, la notation binaire des nombres ainsi que les fonctions booléennes qui régissent l'addition binaire sont expliqués.

2 Algèbre booléenne

2.1 Espace booléen et fonctions booléennes

Soit E l'ensemble contenant les valeurs 0 et 1,

$$E = \{0, 1\}.$$

Cet ensemble est la base de l'algèbre booléenne, et toutes les fonctions booléennes agissent sur cet ensemble.

Définition 2.1. Une *fonction booléenne* est une fonction de E^k vers E, c'est-à-dire qu'elle peut prendre k éléments de E en entrée et renvoyer un 1 ou un 0. Une telle fonction est appelée fonction booléenne de *degré* k.

Définition 2.2. Définissons trois fonctions dites booléennes sur cet ensemble.

$$\begin{aligned}
\text{Et}, \wedge : \quad E \times E &\to E \\
(x,y) &\mapsto \begin{cases} 1, \text{ si } x = y = 1 \\ 0, \text{ sinon} \end{cases}
\end{aligned}$$

$$\begin{aligned}
\text{Ou}, \vee : \quad E \times E &\to E \\
(x,y) &\mapsto \begin{cases} 0, \text{ si } x = y = 0 \\ 1, \text{ sinon} \end{cases}
\end{aligned}$$

$$\begin{aligned}
\text{Non}, \neg : \quad E &\to E \\
x &\mapsto \begin{cases} 1, \text{ si } x = 0 \\ 0, \text{ si } x = 1 \end{cases}
\end{aligned}$$

L'opération Et est l'équivalent de l'opérateur $\wedge$ dans la logique. Si l'on remplace 1 par Vrai et 0 par Faux, $\wedge$ a exactement le même comportement. Les deux termes entrant, soit (x, y), sont les termes à gauche et à droite de l'opérateur. Il en va de même de Ou et l'opérateur $\vee$. L'opération Non est l'équivalent de l'opérateur $\neg$, dont le seul argument est placé à droite. Le résultat est souvent appelé le *complément* [1]. Les fonctions $\wedge$ et $\vee$ sont de degré 2, alors que $\neg$ est de degré 1.

Considérant qu'il n'y a pas beaucoup de valeurs possibles pour un élément de E, générons les tables de valeurs de ces trois opérations :

$\wedge$	0	1
0	0	0
1	0	1

$\vee$	0	1
0	0	1
1	1	1

x	$\neg$ x
0	1
1	0

TABLE 1: Tables de valeurs des opérations $\wedge$, $\vee$ et $\neg$.

On peut définir d'autres fonctions booléennes à partir de ces fonctions de base.

Définition 2.3. Définissons les fonctions Ssi et XOu, qui représentent respectivement le « Si et seulement si » et le « Ou exclusif » de la logique.

[1] À partir de cet endroit, les opérateurs $\wedge, \vee$ et $\neg$ seront utilisés dans le document.

$$\begin{aligned} \text{Ssi, } \Leftrightarrow\text{:} \quad E \times E &\to E \\ (x,y) &\mapsto \begin{cases} 1, \text{ si } x = y \\ 0, \textit{sinon} \end{cases} \end{aligned}$$

$$\begin{aligned} \text{XOu, } \oplus : \quad E \times E &\to E \\ (x,y) &\mapsto \begin{cases} 1, \text{ si } x \neq y \\ 0, \textit{sinon} \end{cases} \end{aligned}$$

Voyons ensuite les tables de valeurs de ces opérateurs.

$\Leftrightarrow$	0	1
0	1	0
1	0	1

$\oplus$	0	1
0	0	1
1	1	0

TABLE 3: Tables de valeurs des opérations $\Leftrightarrow$ et $\oplus$.

Remarque 2.4. Il en va de soi que l'opérateur $\oplus$ est égal à la composition de $\neg$ et $\Leftrightarrow$, tel que $x \oplus y = \neg(x \Leftrightarrow y)$, mais il sera plus agréable à l'oeil d'éliminer les $\neg$. De plus, on peut décomposer la fonction Ssi comme une composition de fonction :

$$\text{Ssi}(x,y) = \text{Ou}\Big(\text{Et}(x,y), \text{Et}\Big(\text{Non}(x), \text{Non}(y)\Big)\Big) = (x \wedge y) \vee (\neg x \wedge \neg y) = x \Leftrightarrow y$$

On comprend que la composition des fonctions ci-dessus est l'équivalent de la phrase suivante : $x = 1$ et $y = 1$, ou $x = 0$ et $y = 0$, ce qui revient à la définition de l'opération Ssi. De plus, il est évident que l'écriture avec $\neg$, $\wedge$ et $\vee$ est beaucoup plus lisible.

Maintenant que les opérateurs $\wedge$, $\vee$, $\neg$, $\Leftrightarrow$ et $\oplus$ sont définis, établissons quelques règles pour travailler avec ces opérateurs.

2.2 Simplification des fonctions booléennes

Regardons certains cas particuliers de simplification de fonctions.

Fonction	Idempotence	Tautologie/contradiction
$\vee$	$x \vee x = x$	$x \vee \neg x = 1$
$\wedge$	$x \wedge x = x$	$x \wedge \neg x = 0$
$\Leftrightarrow$	-	$x \Leftrightarrow x = 1$
$\oplus$	-	$x \oplus x = 0$

TABLE 5: Certaines identités des fonctions booléennes $\vee$, $\wedge$, $\Leftrightarrow$ et $\oplus$.

Ces 6 cas sont évidents et découlent des Tables de valeurs 1 et 3 de la Section 2. Voyons certaines propriétés des fonctions booléennes de base.

Nom	Identités
Loi de De Morgan	$\neg(x \vee y) = \neg x \wedge \neg y$
	$\neg(x \wedge y) = \neg x \vee \neg y$
Commutativité	$x \wedge y = y \wedge x$
	$x \vee y = y \vee x$
Associativité	$(x \wedge y) \wedge z = x \wedge (y \wedge z)$
	$(x \vee y) \vee z = x \vee (y \vee z)$
Distributivité	$(x \wedge y) \vee z = (x \vee z) \wedge (y \vee z)$
	$(x \vee y) \wedge z = (x \wedge z) \vee (y \wedge z)$

TABLE 6: Certaines propriétés des fonctions booléennes $\wedge$, $\vee$ et $\neg$.

Ces identités et propriétés servent à rendre une équation, une fonction booléenne composée ou toute expression logique, moins complexe.

Exemple 2.5. On les applique sur la fonction suivante pour la simplifier :

$$\text{Ssi}(\text{XOu}(1,0),0) = (1 \oplus 0) \Leftrightarrow 1 = (1) \Leftrightarrow 1 = 1.$$

Note 2.6. Le respect de la priorité des parenthèses est de mise.

Avant de pouvoir appliquer ces fonctions sur l'addition binaire, il faut expliquer ce que sont les nombres binaires.

3 Nombres binaires

Avant de décrire les nombres binaires, il serait intéressant de présenter une remarque sur les nombres utilisés couramment : les *nombres décimaux*. Les chiffres comme on les connaît sont 0, 1, 2, 3, 4, 5, 6, 7, 8 et 9. De plus, on sait que lorsqu'on additionne au dernier chiffre, soit 9, un 1, on obtient une *dizaine*. C'est le système commun de nombre, dit *en base 10*. Chaque position d'un nombre représente des puissances de la base, soit 10. L'unité est 10^0, la dizaine est 10^1, etc.

Or, qu'arriverait-il si la base était 2 ? La dizaine s'obtenait lorsqu'on additionne 1 à 1 : $1 + 1 = 10$. De même, à chaque position du nombre, lorsqu'on additionne un 1 avec un 1 , on doit ajouter 1 dans la position supérieure et conserver 0 dans la courante position, de sorte que $10 + 10 = 100$ et $100 + 100 = 1000$. On appelle ce système de nombres : les nombres *binaires*.

Aussi, comme le système en base 10, chaque position d'un nombre en base 2 représente une puissance de 2. Pour calculer la quantité représentée par un nombre en base 2, on prend le nombre à une certaine position et on le multiplie par l'ordre de grandeur de sa position. Par exemple, le nombre 1000101 est égal en nombre décimal à $(1 \times 2^6 + 1 \times 2^2 + 1 \times 2^0)_{10} = (64 + 4 + 1)_{10} = (69)_{10}$.

Notation 3.1. Les nombres décimaux qui peuvent porter à confusion seront notés $(X)_{10}$, à moins qu'il ne soit mentionné autrement.

De plus, une addition binaire peut se faire de la même façon qu'une addition décimale. On additionne les unités, on note la retenue qu'on additionne à la position suivante. Par exemple :

$$\begin{array}{rr} & 1 \\ + & 0 \\ \hline & 1 \end{array} \qquad \begin{array}{rr} & {\scriptstyle 1} \\ & 1 \\ + & 1 \\ \hline 1 & 0 \end{array}$$

TABLE 7: Exemples d'addition unitaire en système binaire.

$$\begin{array}{rcccccc} & {\scriptstyle 1} & {\scriptstyle 1} & {\scriptstyle 1} & & & \\ & & 1 & 0 & 1 & 0 & 1 \\ + & & 1 & 1 & 1 & 1 & 0 \\ \hline & 1 & 1 & 0 & 0 & 1 & 1 \end{array}$$

TABLE 8: Exemple d'addition à la main en système binaire.

À la Table 8, on a additionné

$$10101 = (16 + 4 + 1)_{10} = (21)_{10}$$

et

$$11110 = (16 + 8 + 4 + 2)_{10} = (30)_{10}.$$

En additionnant les nombres décimaux, on obtient 51. Le résultat binaire traduit en nombre décimal est

$$110011 = (32 + 16 + 2 + 1)_{10} = (51)_{10}.$$

Ainsi, une addition procède par le même principe en base 2 qu'en base 10, et on obtient le même résultat, à une traduction près. Toutefois, l'utilisation de la base binaire assure que les chiffres utilisés soient les éléments de $E = \{0,1\}$. Ainsi, il est possible d'appliquer des fonctions booléennes sur ces chiffres. Ceci mène au prochain chapitre, dans lequel est développé un additionneur de nombre binaire défini avec des fonctions booléennes.

4 Addition binaire

4.1 Définitions

Avant de décrire l'addition binaire en fonctions booléennes, on va définir l'addition binaire.

Définition 4.1. Une *addition binaire* est une addition en base 2 de deux nombres, soit x et y, chacun formé de $n + 1$ chiffres issus de $E = \{0, 1\}$. La

longueur $n+1$ est sans perte de généralité, car on accepterait qu'un nombre commence avec des 0. Les retenues de l'addition sont notées r_k. La $k^{\text{ème}}$ retenue est la retenue de l'addition à la position $k-1$. Il y a des retenues de la position 1 à la position $n+2$. La somme de l'addition est de longueur maximale $n+2$, allant de la position 0 à la position $n+1$.

Remarque 4.2. On peut illustrer les termes des positions de l'addition binaire à 2 nombres, chacun de position au plus n, comme dans la Table 9 suivante. Ces mêmes termes seront réutilisés plus tard.

	r_{n+1}	r_n	$\cdots$	r_2	r_1	
		x_n	$\ldots$	x_2	x_1	x_0
$+$		y_n	$\cdots$	y_2	y_1	y_0
	s_{n+1}	s_n	$\ldots$	s_2	s_1	s_0

TABLE 9: Les termes des positions de l'addition binaire à 2 nombres de longueurs n.

4.2 Position de l'unité

Le premier objectif est de trouver une fonction booléenne qui imite l'addition à la position des unités. Toutefois, comme on peut le voir dans l'exemple de la Table 7, il est possible qu'une addition unitaire renvoie 2 chiffres. Or, une fonction booléenne telle que définie dans cet article ne peut renvoyer qu'une valeur. Il faudra donc deux fonctions : une qui renvoie la valeur unitaire, et l'autre qui renvoie la retenue de l'addition. Évaluons les résultats possibles d'addition unitaire afin de construire ces fonctions booléennes.

s_0	0	1
0	0	1
1	1	0

TABLE 10: Valeurs de l'*unité* (s_0) suite à l'addition binaire unitaire.

r_1	0	1
0	0	0
1	0	1

TABLE 11: Valeurs de la *retenue* (r_1) suite à l'addition binaire unitaire.

Les fonctions booléennes liées à ces résultats sont facilement identifiables.

Proposition 4.3. *Supposons une addition unitaire binaire $x+y=s$. Soit $x_0 \in E$ la valeur à la position des unités du premier nombre de l'addition et*

$y_0 \in E$ la valeur à la position des unités du deuxième nombre de l'addition. La valeur à la position des unités de la somme est donnée par s_0, définie :

$$\begin{array}{rccc} s_0 : & E \times E & \to & E \\ & (x_0, y_0) & \mapsto & x_0 \oplus y_0 \end{array}$$

Proposition 4.4. *Supposons une addition unitaire binaire $x + y = s$. Soit $x_0 \in E$ la valeur à la position des unités du premier nombre de l'addition et $y_0 \in E$ la valeur à la position des unités du deuxième nombre de l'addition. La valeur de la retenue à la position des dizaines de l'addition est donnée par r_1, définie :*

$$\begin{array}{rccc} r_1 : & E \times E & \to & E \\ & (x_0, y_0) & \mapsto & x_0 \wedge y_0 \end{array}$$

En résumé, la retenue qui sera utilisée à la position des dizaines dans l'addition est donnée par la fonction r_1 et la valeur de l'unité de la somme est donnée par s_0.

4.3 Position supérieure

On va maintenant chercher à déterminer deux fonctions : une qui renvoie s_1 et l'autre qui renvoie r_2. On va ensuite les généraliser.

La différence entre la somme de la position des dizaines et la somme à la position des unités, c'est que dans l'addition de la dizaine, on doit additionner la retenue. On cherche donc des fonctions booléennes de *degré 3*.

Prenons tous les cas possibles de valeurs des dizaines (x_1 et y_1) et de la retenue (r_1) et regardons les valeurs des dizaines de la somme (s_1) qui en résulte.

r_1	x_1	y_1	s_1
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

TABLE 12: Tables de valeurs de la position des dizaines (r_1, x_1, y_1 et s_1) dans l'addition binaire.

Comme on peut le constater dans la Table 12, la valeur à la position des

dizaines de la somme est 1 lorsque :

$$\begin{aligned} y_1 = 1 \text{ et } r_1 = x_1 = 0, \\ \text{ou } x_1 = 1 \text{ et } r_1 = y_1 = 0, \\ \text{ou } r_1 = 1 \text{ et } x_1 = y_1 = 0, \\ \text{ou } r_1 = x_1 = y_1 = 1 \end{aligned}$$

On se rappelle qu'on peut transformer les « et » et les « ou » en forme logique. Ainsi, comme on vient d'exprimer la fonction logique sous forme de mot, il serait possible de la réécrire sous forme de fonction booléenne, comme ceci :

$$s_1 = (\neg r_1 \wedge \neg x_1 \wedge y_1) \vee (\neg r_1 \wedge x_1 \wedge \neg y_1) \vee (r_1 \wedge \neg x_1 \wedge \neg y_1) \vee (r_1 \wedge x_1 \wedge y_1) \quad (1)$$

Toutefois, on peut réfléchir à comment pré-simplifier cette fonction, en utilisant un fait qu'on connaît déjà : la somme unitaire binaire à 2 termes est déjà définie :

$$s_0 = x_0 \oplus y_0$$

Comme on ne cherche que les unités (c'est-à-dire pas la retenue), si on prenait l'opération $\oplus$ (Ou exclusif), qu'on applique sur y_1 et x_1, puis qu'on réapplique sur ce résultat avec la retenue r_1, cela fonctionnerait-il ? Autrement dit, on connaît l'addition à 2 termes. On cherche à prendre le résultat de cette addition et on refait une addition à 2 termes. Voici la formule présentée :

$$s_1 = (x_1 \oplus y_1) \oplus r_1 \quad (2)$$

Vérifions si cette formule renvoie les mêmes valeurs que l'addition à laquelle on s'attend, c'est-à-dire que les Équations 1 et 2 sont équivalentes.

r_1	x_1	y_1	$x_1 \oplus y_1$	$(x_1 \oplus y_1) \oplus r_1$	s_1
0	0	0	0	0	0
0	0	1	1	1	1
0	1	0	1	1	1
0	1	1	0	0	0
1	0	0	0	1	1
1	0	1	1	0	0
1	1	0	1	0	0
1	1	1	0	1	1

TABLE 13: Tables de valeurs de la position des dizaines (r_1, x_1, y_1, s_1 et $(x_1 \oplus y_1) \oplus r_1$) dans l'addition binaire.

On voit qu'on obtient exactement les mêmes valeurs avec les deux fonctions booléennes. Ces fonctions sont donc équivalentes.

Afin de généraliser le résultat, il est possible de pointer que l'addition à 3 termes se fasse sous les mêmes règles à la position des dizaines ou à la position

des centaines. Les seules positions qui ne respectent pas ces règles sont la position des unités, car elle n'a pas de retenue à prendre en compte, et la position $n+1$, car il ne reste que la retenue. Ceci permet la proposition suivante, qui généralise la somme sans retenue à chaque position de l'addition binaire.

Proposition 4.5. *Supposons une addition binaire $x+y=s$. Soit $i \in 1, 2, \ldots, n$ et soient :*

$x_i \in E$ *la valeur à la $i^{ème}$ position du 1^{er} nombre de l'addition (x),*
$y_i \in E$ *la valeur à la $i^{ème}$ position du $2^{ème}$ nombre de l'addition (y) et*
$r_i \in E$ *la valeur à la $i^{ème}$ position des retenues de l'addition binaire.*

La valeur à la $i^{ème}$ position de la somme de l'addition est donnée par s_i, définie :

$$\begin{array}{rccc} s_{i,\ i\in\{1,2,\ldots,n\}}: & E\times E\times E & \to & E \\ & (x_i, y_i, r_i) & \mapsto & (x_i \oplus y_i) \oplus r_i \end{array}$$

Approchons d'un autre angle la retenue. On a précédemment comparé les tables de valeurs de deux fonctions afin d'en déduire l'équivalence. Plus loin, on utilisera plutôt le principe de simplification pour déduire l'équivalence.

Ainsi, considérons la retenue à la position 2 (qui sera ensuite généralisée). De même qu'à la recherche de la fonction de la somme, on va commencer par regarder les valeurs possibles de r_2. Prenons tous les cas possibles de valeurs des dizaines (x_1 et y_1) et de la retenue à la position des dizaines (r_1) et regardons le résultat attendu de la retenue (r_2).

r_1	x_1	y_1	r_2
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

TABLE 14: Tables de valeurs de la position des dizaines et de la retenue des centaines (r_1, x_1, y_1 et r_2) dans l'addition binaire.

Il y a une retenue quand au moins 2 des trois valeurs sont 1. C'est normal, puisque pour changer de position, la somme doit être d'au moins 2. Voyons tous les cas possibles explicitement.

La retenue de la position supérieure (r_2) est 1 lorsque

$$\begin{aligned} y_1 = x_1 = 1 \text{ et } r_1 = 0, \\ \text{ou } x_1 = r_1 = 1 \text{ et } y_1 = 0, \\ \text{ou } r_1 = y_1 = 1 \text{ et } x_1 = 0, \\ \text{ou } r_1 = x_1 = y_1 = 1. \end{aligned}$$

Changeons ces cas en fonction booléenne.

$$r_2 = (r_1 \wedge \neg x_1 \wedge y_1) \vee (\neg r_1 \wedge x_1 \wedge y_1) \vee (r_1 \wedge x_1 \wedge \neg y_1) \vee (r_1 \wedge x_1 \wedge y_1)$$

Maintenant, simplifions cette équation en détail, afin de voir un processus de simplification, en utilisant les propriétés énumérées à la Section 2.2.

$$\begin{aligned} r_2 &= (r_1 \wedge \neg x_1 \wedge y_1) \vee (\neg r_1 \wedge x_1 \wedge y_1) \vee (r_1 \wedge x_1 \wedge \neg y_1) \vee (r_1 \wedge x_1 \wedge y_1) \\ &= \Big[\Big((r_1 \wedge \neg x_1) \vee (\neg r_1 \wedge x_1)\Big) \wedge y_1\Big] \vee (r_1 \wedge x_1 \wedge y_1) \vee (r_1 \wedge x_1 \wedge \neg y_1) \\ &= \Big[\Big((r_1 \wedge \neg x_1) \vee (\neg r_1 \wedge x_1) \vee (r_1 \wedge x_1)\Big) \wedge y_1\Big] \vee (r_1 \wedge x_1 \wedge \neg y_1) \\ &= \Big[\Big((r_1 \wedge \neg x_1) \vee ((\neg r_1 \vee r_1) \wedge x_1)\Big) \wedge y_1\Big] \vee (r_1 \wedge x_1 \wedge \neg y_1) \\ &= \Big[\big((r_1 \wedge \neg x_1) \vee x_1\big) \wedge y_1\Big] \vee (r_1 \wedge x_1 \wedge \neg y_1) \\ &= \Big[\big((r_1 \vee x_1) \wedge (\neg x_1 \vee x_1)\big) \wedge y_1\Big] \vee (r_1 \wedge x_1 \wedge \neg y_1) \\ &= \Big[(r_1 \vee x_1) \wedge y_1\Big] \vee (r_1 \wedge x_1 \wedge \neg y_1) \\ &= \Big[(r_1 \wedge y_1) \vee (x_1 \wedge y_1)\Big] \vee (r_1 \wedge x_1 \wedge \neg y_1) \\ &= (r_1 \wedge y_1) \vee \Big[x_1 \wedge \Big((y_1 \vee (r_1 \wedge \neg y_1)\Big)\Big] \\ &= (r_1 \wedge y_1) \vee \Big[x_1 \wedge \Big((y_1 \vee r_1) \wedge (y_1 \vee \neg y_1)\Big)\Big] \\ &= (r_1 \wedge y_1) \vee \Big[x_1 \wedge (y_1 \vee r_1)\Big] \\ &= (r_1 \wedge y_1) \vee (x_1 \wedge y_1) \vee (x_1 \wedge r_1) \end{aligned}$$

La comparaison de la table des valeurs de la formule initiale et de celle qui est simplifiée est inutile, car la simplification ne change pas les résultats. L'équation a passé d'une formule plutôt rébarbative à une forme dont on peut tirer une certaine logique. Intuitivement, le nouveau résultat propose que dès qu'il y a

une paire dont les termes valent tous les deux 1, alors il y a une retenue. Le résultat a du sens. D'ailleurs, on peut comprendre qu'il n'est pas nécessaire non plus de connaître la valeur du 3e terme, car il n'affectera pas la retenue si deux termes valent déjà 1.

Il est important de noter que la règle de retenue est respectée pour toutes positions sauf la position 1. Ceci inclut la position r_{n+1}. On peut donc généraliser la fonction de retenue trouvée précédemment aux positions i, i allant de 1 à $n+1$.

Proposition 4.6. *Supposons une addition binaire $x+y=s$. Soit $i \in 1, 2, \ldots, n$ et soit :*

$x_i \in E$ *la valeur à la $i^{ème}$ positions du 1^{er} nombre de l'addition (x),*
$y_i \in E$ *la valeur à la $i^{ème}$ position du $2^{ème}$ nombre de l'addition (y) et*
$r_i \in E$ *la valeur à la $i^{ème}$ position des retenues de l'addition binaire.*

La valeur à la $(i+1)^{ème}$ retenue de l'addition est donnée par r_{i+1}, défini :

$$\begin{array}{rccl} r_{i+1,\ i\in\{1,2,\ldots,n-1\}}: & E\times E\times E & \to & E \\ & (x_i, y_i, r_i) & \mapsto & (r_i \wedge y_i) \vee (x_i \wedge y_i) \vee (x_i \wedge r_i) \end{array}$$

Comme à la position $n+1$ il n'y a pas de valeur de x et y, s_{n+1} devient r_{n+1} dans l'addition.

5 Conclusion

Il a été vu quelques théories de l'algèbre booléenne, notamment les fonctions booléennes de base ET, OU, NON, SSI et XOU, et leur simplification. Aussi, on a vu le principe de nombres binaires, les bases de nombre et l'addition booléenne sur les nombres binaires. Les formules pour chaque position de la retenue de l'addition binaire ainsi que la somme ont été construites avec les fonctions booléennes de base.

Pour résumé toutes les fonctions construites, on peut les rassembler les fonctions de retenues et de sommes à chaque position dans le schéma de l'addition binaire à n position suivant :

$$\begin{array}{cccccccc} & r_{n+1} & r_n & \cdots & r_2 & r_1 & \\ & & x_n & \ldots & x_2 & x_1 & x_0 \\ + & & y_n & \ldots & y_2 & y_1 & y_0 \\ \hline & s_{n+1} & s_n & \ldots & s_2 & s_1 & s_0 \end{array}$$

où x_i est la valeur à la $i^{\text{ème}}$ position du nombre x écrit sous forme binaire ($i \in \{0, 1, 2, \ldots, n\}$), y_i est la valeur à la $i^{\text{ème}}$ position du nombre y écrit sous forme binaire ($i \in \{0, 1, 2, \ldots, n\}$),

$$s_i = \begin{cases} x_i \oplus y_i, & \text{si } i = 0 \\ (x_i \oplus y_i) \oplus r_i, & \text{si } i \in \{1, 2, \ldots, n\} \text{ et} \\ r_i, & \text{si } i = n+1 \end{cases}$$

$$r_i = \begin{cases} x_i \wedge y_i, & \text{si } i = 1 \\ (r_i \wedge y_i) \vee (x_i \wedge y_i) \vee (x_i \wedge r_i), & \text{si } i \in \{2, \ldots, n+1\}. \end{cases}$$

Références

[Boo54] George Boole : *An Investigation of the Laws of Thought : On which are Founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberly, 1854.

[Sha37] Claude Elwood Shannon : A symbolic analysis of relay and switching circuits. Master's thesis, Massachusetts Institute of Technology, 1937.

Julien Corriveau-Trudel
Département de Mathématiques, Université de Sherbrooke
Courriel: Julien.Corriveau-Trudel@USherbrooke.ca

CaMUS **7**, 33 – 44

Estimation de la copule gaussienne dans le cadre Bayésien

Marwa Hamdi

Résumé Parmi les différentes familles de copules, on étudie en particulier dans cet article la copule gaussienne. Après une introduction générale, on montrera comment estimer efficacement les paramètres de la copule gaussienne à l'aide de l'approche bayésienne, en se basant sur l'algorithme de Metropolis Hastings.

1 Introduction

La dépendance est un concept fondamental en statistique. Souvent, des mesures comme le coefficient de corrélation, le tau de Kendall et le rho de Spearman sont utilisées pour évaluer la force de la dépendance. Aussi, les modèles de régression, comme la régression linéaire, permettent de relier une variable réponse et plusieurs variables explicatives. Cependant, les mesures de dépendance et les modèles de régression sont restreints.

Les copules sont des fonctions qui modélisent la structure de dépendance entre deux ou plusieurs variables aléatoires. Elles ont été développées à partir d'un problème de probabilité, énoncé par Maurice Fréchet [Fré51] dans le cadre des espaces métriques aléatoires et comme solution introduite par Abde Sklar en 1959 [Skl59] dans la théorie des lois multidimensionnelles. L'idée principale des copules a été introduite par ce dernier. Elle consiste à séparer la modélisation des lois marginales et de la structure de dépendance d'un vecteur aléatoire. En raison de sa flexibilité à modéliser des relations complexes entre les variables, les copules ont plusieurs applications dans multiples domaines comme : l'analyse de la survie, l'actuariat, la finance, le marketing, etc.

Comme l'idée des copules est récente, la plupart des recherches se sont concentrées sur le développement, les propriétés de ses fonctions et leur performance pour résoudre des problèmes. Cependant, moins d'attention a été portée sur comment estimer efficacement les paramètres des copules.

L'approche bayésienne à l'avantage de permettre l'estimation simultanée des lois marginales et de la copule des modèles multidimensionnelles. Le but de ce

En tout premier lieu, je tiens à remercier le Professeur Bernard Colin pour ses précieux conseils dans la structuration et la supervision de ce travail. Je tiens également à remercier mon directeur de maîtrise, le Professeur Taoufik Bouezmarni, pour ses différents commentaires et suggestions visant à améliorer la qualité de ce travail.

projet est de présenter brièvement les copules, en particulier la copule gaussienne, et d'appliquer l'approche bayésienne (méthode MCMC) pour estimer ses paramètres de la copule gaussienne.

2 Définition et propriétés de la Copule

2.1 Copules : définitions et propriétés

Définition 2.1. Une *copule bidimensionnelle*, notée C, est une fonction de répartition définie sur $I = [0,1]^2$ dont les lois marginales sont égales à la loi uniforme sur $[0,1]$, c'est-à-dire que pour tout $u, v \in [0,1]^2$, on a :

$$C(u,v) = P(U \leq u, V \leq v),$$

où U et V sont deux variables aléatoires uniformes sur $[0,1]$.

Autrement dit, une copule est une fonction C définie sur $[0,1]^2$ vérifiant les caractéristiques suivantes :

1. $\forall u, v \in [0,1]$, $C(u,0) = 0 = C(0,v)$;
2. $\forall u, v \in [0,1]$, $C(u,1) = u$ et $C(1,v) = v$;
3. $\forall (u_1,u_2), (v_1,v_2) \in [0,1]^2$ tels que $u_1 \leq v_1$ et $u_2 \leq v_2$ on a :
$$C(v_1,v_2) - C(v_1,u_2) - C(u_1,v_2) + C(u_1,u_2) \geq 0.$$

Dans la littérature, plusieurs familles de copules ont été introduites : les copules elliptiques (par exemple, la copule gaussienne et de Student) et les copules archimédiennes (par exemple, la copule de Clayton, Gumbel et Frank). L'exemple suivant présente la copule comonotone, notée M, qui est la borne supérieure de Fréchet-Hoeffding, c'est-à-dire que pour toute copule C on a

$$C(u,v) \leq M(u,v).$$

Exemple 2.2.
Soit la fonction $M(u,v) = \min(u,v)$, pour $u,v \in [0,1]$. On vérifie que M définit une copule. On a,

1. $\forall u, v \in [0,1]$, $\min(u,0) = 0 = \min(0,v)$;
2. $\forall u, v \in [0,1]$, $C(u,1) = u$ et $C(1,v) = v$;
3. $\forall (u_1,u_2), (v_1,v_2) \in [0,1]^2$ tels que $u_1 \leq v_1$ et $u_2 \leq v_2$ on a :
$$\min(v_1,v_2) - \min(v_1,u_2) - \min(u_1,v_2) + \min(u_1,u_2) \geq 0.$$

En effet, $\min(v_1,v_2) - \min(v_1,u_2) - \min(u_1,v_2) + \min(u_1,u_2)$ est égal à

- $v_1 - v_1 - u_1 + u_1 \,(\geq 0)$ si $u_1 \leq v_1 \leq u_2 \leq v_2$;
- $v_2 - u_2 - v_2 + u_2 \,(\geq 0)$ si $u_2 \leq v_2 \leq u_1 \leq v_1$;
- $v_1 - u_2 - v_1 + v_1 \,(\geq 0)$ si $u_1 \leq u_2 \leq v_1 \leq v_2$;
- $v_2 - u_2 - u_1 + v_2 \,(\geq 0)$ si $u_2 \leq u_1 \leq v_2 \leq v_1$.

2.2 Théorème de Sklar (1959)

L'idée du théorème est de décomposer la fonction de répartition conjointe en deux composantes : la copule et ses lois marginales. Ceci permet de construire les fonctions de répartitions multidimensionnelles en choisissant séparément les lois marginales et la structure de dépendance.

Théorème 2.3 (Sklar(1958) [Skl59]). *Soit $\boldsymbol{X} = (X_1, \ldots, X_d)$ un vecteur aléatoire et F une fonction de répartition conjointe sur celui-ci, dont les lois marginales sont $F_1, \ldots, F_d$. Alors, il existe une fonction de copule C, telle que :*

$$\forall (x_1, \ldots, x_d) \in \mathbb{R}^d,\ F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

Remarque 2.4. Si les lois marginales, $F_1, F_2, \ldots, F_d$ sont continues, alors C est unique. Sinon elle est unique seulement sur $\prod_{i=1}^{d} Im(F_i)$, où $Im(F_i)$ est l'image de la fonction F_i.

Réciproquement, si C est une fonction de copule et $F_1, \ldots, F_d$ sont des fonctions de répartition, alors :

$$\forall u_i \in [0,1],\ C(u_i) = F(F_1^{-1}(x_1), \ldots, F_d^{-1}(x_d)),$$

où F_i^{-1} est la fonction réciproque de F_i, appelée aussi la fonction quantile.

L'exemple suivant montre comment on peut construire une copule à partir d'une fonction de répartition bidimensionnelle. Ici, la copule construite est la copule de Clayton de paramètre $\phi = 1$.

Exemple 2.5. On considère la distribution bidimensionnelle

$$F(x,y) = (1 + e^{-x} + e^{-y})^{-1},$$

où $x, y \in [0,\infty)$. Les fonctions marginales sont égales à

$$F_1(x) = (1 + e^{-x})^{-1} \quad \text{et} \quad F_2(y) = (1 + e^{-y})^{-1}.$$

Alors, d'après le théorème de Sklar, on a que

$$\begin{aligned} C(u,v) &= F(F_1^{-1}(u), F_2^{-1}(v)) \\ &= \left(1 + e^{\ln\left(\frac{1}{u}-1\right)} + e^{\ln\left(\frac{1}{v}-1\right)}\right) \\ &= \frac{uv}{u + v - uv}, \end{aligned}$$

qui définit bien une copule.

2.3 Densité d'une copule

On peut modéliser la structure de dépendance d'un vecteur aléatoire en utilisant la densité de la copule C qui est définie par :

$$c(u_1,\ldots,u_d) = \frac{\partial^d}{\partial u_1 \ldots \partial u_d} C(u_1,\ldots,u_d).$$

On peut exprimer la densité conjointe f du vecteur aléatoire $\boldsymbol{X}$ en fonction de sa densité de copule, de ses fonctions de répartitions et de ses densités marginales. En effet, d'après le théorème de Sklar et la définition de la densité de copule, f peut s'écrire de la façon suivante :

$$\begin{aligned} f(x_1,...,x_d) &= \frac{\partial^d}{\partial x_1 \ldots \partial x_d} F(x_1,\ldots,x_d) \\ &= c(F_1(x_1),\ldots,F_d(x_d)) \prod_{i=1}^{d} f_i(x_i). \end{aligned}$$

En utilisant la densité conjointe f, on peut réexprimer la densité de copule :

$$c(u_1,\ldots,u_d) = \frac{f(F_1^{-1}(u_1),\ldots,F_d^{-1}(u_d))}{\prod_{i=1}^{d} f_i(F_i^{-1}(u_i))}.$$

2.4 Mesure de dépendance

Plusieurs coefficients ont été développés pour mesurer la dépendance entre les variables aléatoires. Parmi ces mesures, on trouve le coefficient de corrélation de Pearson entre deux variables aléatoires. La proposition suivante établit que le coefficient de corrélation de Pearson peut être déduit à partir de la fonction de copule et de ses distributions marginales. Notons que les deux autres coefficients cités dans l'introduction, le tau de Kendall et le rho de Spearman, peuvent être calculés en utilisant seulement la fonction de copule.

Proposition 2.6.
Soit (X_1,X_2) *un vecteur aléatoire de copule* C *et de fonctions de répartition marginales* F_1 *et* F_2*. Le coefficient de corrélation entre deux variables aléatoires* X_1 *et* X_2 *est donné par :*

$$\rho(X_1,X_2) = \frac{1}{\sqrt{var(X_1)var(X_2)}} \int_0^1 \int_0^1 (C(u_1,u_2) - u_1u_2)\, dF_1^{-1}(u_1)\, dF_2^{-1}(u_2).$$

En effet,

$$\begin{aligned} \rho(X_1,X_2) &= \frac{cov(X_1,X_2)}{\sqrt{var(X_1)var(X_2)}} \\ &= \frac{1}{\sqrt{var(X_1)var(X_2)}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (F(x_1,x_2) - F_1(x_1)F_2(x_2))\mathrm{d}x_1\mathrm{d}x_2 \\ &= \frac{1}{\sqrt{var(X_1)var(X_2)}} \int_0^1 \int_0^1 (C(u_1,u_2) - u_1u_2)\mathrm{d}F_1^{-1}(u_1)\mathrm{d}F_2^{-1}(u_2). \end{aligned}$$

On trouve la dernière équation en appliquant le théorème de Sklar et le changement de variables $u_1 = F_1(x_1)$ et $u_2 = F_2(x_2)$.

2.5 Modèles de copules

Parmi les modèles de copules les plus importants, on trouve les copules archimédiennes et les copules elliptiques. Dans cette section, nous allons présenter brièvement les copules les plus utilisées.

2.5.1 Copules archimédiennes

La classe des copules archimédiennes permet de construire de nombreuses familles de copules qui sont capables de modéliser plusieurs structures de dépendance. Les intéressantes propriétés et la facilité de construire les copules archimédiennes les rendent très attractives en pratique. Pour plus de détails, voir [Nel06].

Définition 2.7.
Soit la fonction $\varphi : [0,1] \to \mathbb{R}^+$, dite la fonction génératrice, telle que :

$$\varphi(1) = 0,\ \varphi'(u) < 0 \text{ et } \varphi''(u) > 0.$$

La *copule archimédienne avec le générateur* φ est alors définie par :

$$C(u_1,\ldots,u_d) = \begin{cases} \varphi^{-1}\left(\sum_{i=1}^{d} \varphi(u_i)\right) & \text{si } \sum_{i=1}^{d} \varphi^{-1}(u_i) \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

Exemple 2.8. Dans la littérature, on trouve plusieurs copules archimédiennes. Dans cet exemple, on en cite quelques-unes.

Copule d'indépendance

La copule d'indépendance, notée $\Pi(u,v)$, fait partie des copules archimédiennes. Soit la fonction génératrice $\varphi(t) = -\log(t)$. Son inverse est donnée par la fonction $\varphi^{-1}(t) = \exp(-t)$. En se basant sur la définition, on trouve :

$$\begin{aligned} C(u,v) &= \exp\left(-\left[-\log(t) + (-\log(v))\right]\right) \\ &= uv \\ &:= \Pi(u,v). \end{aligned}$$

Copule de Clayton

Pour cette famille, le générateur est défini pour un paramètre réel $\phi > -1, \phi \neq 0$ et $u \in]0,1]$ par :

$$\varphi^{Cl}(u) = \phi^{-1}(u^{-\phi} - 1).$$

Ceci permet de générer la copule de Clayton dans le cas bidimensionnel de la façon suivante :

$$C_\phi^{Cl}(u_1,u_2) = (u_1^{-\phi} + u_2^{-\phi} - 1)^{-1/\phi}.$$

Dans l'Exemple 2.5, on a construit cette copule, avec $\phi = 1$, à partir d'une fonction de répartition bidimensionnelle. Comme cette fonction de copule est différentiable, elle possède une fonction de densité donnée par la forme suivante :

$$c_\phi^{Cl}(u_1,u_2) = (\phi+1)(u_1u_2)^{-\phi-1}(u_1^{-\phi} + u_2^{-\phi} - 1)^{\frac{-1}{\phi}-2}.$$

Copule de Frank

Pour $\phi \neq 0$ et pour $u \in]0,1]$, le générateur de la copule de Frank est donné par :

$$\varphi^F(u) = -\ln\left(\frac{e^{-\phi u}-1}{e^{-\phi}-1}\right).$$

La copule de Frank est donnée par :

$$C_\phi^F(u_1,u_2) = -\frac{1}{\phi}\ln\left(1 + \frac{(e^{-\phi}u_1-1)(e^{-\phi}u_2-1)}{e^{-\phi}-1}\right),$$

et sa densité de copule est égale à :

$$c_\phi^F(u_1,u_2) = \frac{\phi(1-e^{-\phi})e^{-\phi(u_1+u_2)}}{[(1-e^{-\phi}) - (e^{-\phi u_1}-1)(e^{-\phi u_2}-1)]^2}.$$

Copule de Gumbel

La dernière copule archimédienne que nous présentons est la copule de Gumbel. Sa fonction génératrice est donnée, pour $u \in]0,1]$, par :

$$\varphi^G(u) = (-\ln(u))^\phi \quad \text{avec} \quad \phi > 0.$$

Donc, la copule de Gumbel s'écrit sous la forme :

$$C_\phi^G(u_1,u_2) = e^{-[(-\ln(u_1))^\phi + (-\ln(u_2))^\phi]^{1/\phi}}.$$

Remarque 2.9. Notons que, C_ϕ^{Cl}, C_ϕ^F et C_ϕ^G convergent vers la copule comonotone introduite dans l'Exemple 2.2, si ϕ tendent vers l'infini. Aussi, C_ϕ^{Cl} et C_ϕ^F $\left(\text{resp. } C_\phi^G\right)$ tendent vers la copule d'indépendance si ϕ converge vers 0 (resp. 1.)

2.6 Copules elliptiques

Dans cette classe de copules se trouve la copule gaussienne et la copule de Student. Nous allons nous limiter à la présentation de la copule gaussienne qui nous intéresse dans cet article. Cependant, la méthodologie que nous présenterons dans la Section 3 peut être adaptée pour estimer les paramètres des autres copules.

Copule gaussienne

La copule gaussienne est la plus connue des familles des copules elliptiques et surtout la plus utilisée en pratique (voir [Smi11]). La copule gaussienne d'un vecteur $\boldsymbol{X}$ de dimension d s'écrit sous la forme suivante :

$$C(u_1,\ldots,u_d;\Gamma) \quad = \quad \boldsymbol{\Phi}_d\left(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_d);\Gamma\right), \tag{1}$$

où $\boldsymbol{\Phi}_d$ (respectivement Φ) est la distribution normale multidimensionnelle de moyenne $0_d = (0,\ldots,0)^\top$ et de matrice de corrélation Γ (respectivement la distribution normale unidimensionnelle de moyenne 0 et de variance 1). Sa densité de copule est donnée par :

$$\begin{aligned} c(\boldsymbol{u};\boldsymbol{\Phi}) &= \frac{\partial}{\partial u_1 \ldots \partial u_d} C(\boldsymbol{u};\boldsymbol{\Phi}) \\ &= |\Gamma|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{x}^\top(\Gamma^{-1}-I)\boldsymbol{x}\right\}, \end{aligned}$$

où $\boldsymbol{u} = (u_1,\ldots,u_d)$ et $\boldsymbol{x} = (\Phi^{-1}(u_1,1),\ldots,\Phi^{-1}(u_d,1))^\top$.

3 Estimation des copules

3.1 Inférence bayésienne

L'approche bayésienne est cohérente pour la résolution des problèmes d'inférence statistique. Elle permet la modélisation et l'analyse complète des incertitudes. [PCK06] et [Smi11] ont suggérés l'approche bayésienne pour estimer les paramètres de la copule gaussienne en se basant sur l'algorithme de simulation des méthodes de Monte-Carlo par chaine de Markov (MCMC). Un des objectifs de l'estimation des paramètres des copules est de construire une inférence sur les mesures de dépendances. Dans le cas de la copule gaussienne, l'inférence est construite sur le coefficient de corrélation de Pearson.

Estimation

Il existe plusieurs méthodes pour estimer les paramètres d'une copule. Une de ces méthodes, l'approche bayésienne, est basée sur la méthode MCMC.

Fonction de vraisemblance

Soit $\{\boldsymbol{y}_i = (y_{i1},\ldots,y_{id})\}_{i=1}^n$, n réalisations indépendantes et identiquement distribuées d'un vecteur aléatoire continue $\boldsymbol{Y} = (Y_1,\ldots,Y_d)$. On note par $f_j(.;\boldsymbol{\theta}_j)$, $j = 1,\ \ldots,\ d$, la densité marginale de Y_j et par $f(.;\boldsymbol{\theta},\boldsymbol{\phi})$ la fonction de densité conjointe de $\boldsymbol{Y}$, où $\boldsymbol{\phi}$ désigne le vecteur des paramètres de la copule de $\boldsymbol{Y}$ et $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d)$. On suppose que la copule de $\boldsymbol{Y}$ est gaussienne définie par (1).

Dans ce cas, la fonction de vraisemblance, avec $\boldsymbol{\phi} = \Gamma$, est de la forme suivante (voir [Smi11]) :

$$\begin{aligned} L &= \prod_{i=1}^{n} f(\boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= \prod_{i=1}^{n} \left(c(\boldsymbol{u}_i; \Gamma) \prod_{j=1}^{d} f_j(y_{ij}; \boldsymbol{\theta}_j) \right) \\ &= |\Gamma|^{-n/2} \prod_{i=1}^{n} \left[\exp\left\{ -\frac{1}{2} \boldsymbol{x}_i^\top (\Gamma^{-1} - I) \boldsymbol{x}_i \right\} \prod_{j=1}^{d} f_j(y_{ij}; \boldsymbol{\theta}_j) \right], \end{aligned}$$

pour $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})$ et $\boldsymbol{x}_i = (\Phi^{-1}(u_{i1}), \ldots, \Phi^{-1}(u_{id}))^\top$ avec $u_{ij} = F_j(y_{ij}; \boldsymbol{\theta}_j)$, où F_j est la fonction de distribution de Y_j.

3.2 Méthode de simulation de Monte-Carlo

L'estimation bayésienne peut être rendue difficile si ce n'est pas impossible, pour deux raisons :

1. Le calcul explicite de $\pi(\theta|x)$ peut être impossible.

2. Lorsque l'espace des paramètres et l'espace des décisions sont de grandes dimensions. Même si $\pi(\theta|x)$ est connue, les calculs nécessitent un temps considérable.

Pour résoudre ce problème, une solution est donnée par les méthodes de simulation.

Méthode de MCMC

Le principe de la méthode MCMC est de générer une chaîne de Markov qui suit une loi asymptotiquement distribuée selon la loi a posteriori de θ, puis approximer $\mathbb{E}(g(\theta|x))$, où g est une fonction réelle. Pour construire une telle chaîne, il existe deux algorithmes : "Algorithme de Gibbs" et "Algorithme de Metropolis-Hastings (MH)". Dans notre étude, on s'intéresse à l'algorithme de MH comme suggéré par [Smi11].

Algorithme de Metropolis-Hastings (MH)

Une chaîne de Markov peut être construite par l'algorithme (MH). Cet algorithme nécessite une connaissance partielle de la fonction de densité. Étant donné la densité $\pi(\theta|x)$, on choisit une densité conditionnelle instrumentale $q(.|x)$. Pour plus de détails sur le choix de la loi de proposition symétrique q, voir [Ra11]. L'algorithme est donné comme suit :

1. Étant donné θ^p, simuler $\tilde{\theta}$ à partir de $q(\tilde{\theta}|\theta^p)$.

2. Prendre :

$$\theta^{p+1} = \begin{cases} \tilde{\theta} & \text{avec probabilité} \quad \alpha \\ \boldsymbol{\theta}^p & \text{avec probabilité} \quad 1-\alpha \end{cases}$$

où α est la probabilité d'acceptation donnée par :

$$\alpha = \min\left\{1; \frac{\pi(\tilde{\theta}|x)q(\theta^p|\tilde{\theta})}{\pi(\theta^p|x)q(\tilde{\theta}|\theta^p)}\right\}.$$

Metropolis-Hastings pour la copule gaussienne

On utilise la statistique bayésienne pour estimer les paramètres de la copule gaussienne avec la méthode de MCMC. La méthode MCMC consiste en deux étapes :

1. D'abord, on génère à partir de la densité $f(\theta_j|\{\boldsymbol{\theta}/\theta_j\},\Gamma,\boldsymbol{y})$, où $\{\boldsymbol{\theta}/\theta_j\}$ désigne tous les éléments de $\boldsymbol{\theta}$ sauf θ_j;

2. ensuite on génère à partir de $f(\Gamma|\boldsymbol{\theta},\boldsymbol{y})$.

Pour la première étape, la loi a posteriori $f(\theta_j|\{\boldsymbol{\theta}/\theta_j\},\Gamma,y)$ est donnée par :

$$\begin{aligned} f(\theta_j|\{\boldsymbol{\theta}/\theta_j\},\Gamma,\boldsymbol{y}) &= \frac{f(\theta_j,\{\boldsymbol{\theta}/\theta_j\},\Gamma,\boldsymbol{y})}{f(\{\boldsymbol{\theta}/\theta_j\},\Gamma,\boldsymbol{y})} \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta},\Gamma)f(\boldsymbol{\theta},\Gamma) \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta},\Gamma)f(\Gamma|\boldsymbol{\theta}/\theta_j)\pi(\theta_j) \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta},\Gamma)\pi(\theta_j) \\ &\propto |\Gamma|^{-n/2}\prod_{i=1}^{n}\left[\exp\left\{-\frac{1}{2}\boldsymbol{x}_i^\top(\Gamma^{-1}-I)\boldsymbol{x}_i\right\}\prod_{j=1}^{d}f_j(y_{ij};\boldsymbol{\theta}_j)\right]\pi(\theta_j), \end{aligned} \tag{2}$$

où $\pi(\theta_j)$ est la loi marginale a priori. Notons que $\boldsymbol{x}_i = (\Phi^{-1}(u_{i1}),\ldots,\Phi^{-1}(u_{id}))^\top$, où $\boldsymbol{u}_i = (u_{i1},\ldots,u_{id})$ avec $u_{ij} = F_j(y_{ij};\boldsymbol{\theta}_j)$, donc $\boldsymbol{x}_i$ dépend de θ_j. La vraisemblance dans (2) n'est pas une forme standard et donc est difficile à utiliser en pratique.
Pour construire l'algorithme de Metropolis-Hastings, [PCK06] ont suggéré la loi Student multivariée $T_\nu(\hat{\theta}_j,V)$, où $V = -H^{-1}$,

$$H = \frac{\partial^2 \log(f(\theta_j|\{\boldsymbol{\theta}/\theta_j\},\Gamma,\boldsymbol{y}))}{\partial\theta_j(\partial\theta_j)^\top}\Big|_{\theta_j=\hat{\theta}_j}.$$

La probabilité d'acceptation est donnée par :

$$\alpha = \min\left\{1, \frac{f(\hat{\theta}_j|\boldsymbol{\theta}/\theta_j\Gamma,\boldsymbol{y})T_\nu(\theta_j^p|\hat{\theta}_j,V)}{f(\theta_j^p|\boldsymbol{\theta}/\theta_j,\Gamma,\boldsymbol{y})T_\nu(\hat{\theta}_j|\theta_j^p,V)}\right\}.$$

Dans la deuxième étape où la matrice de corrélation Γ est générée conditionnellement par rapport à θ et $\boldsymbol{y}$, plusieurs suggestions ont été proposées ([PCK06], [DP09], [Smi11]). Chaque méthode dépend de la paramétrisation de la matrice de corrélation et la fonction a priori. Une de ces suggestions sur la paramétrisation de la matrice de corrélation est celle de [DP09] donnée par :

$$\lambda_{t,s} = Corr(Y_t,Y_s|Y_{t-1},...,Y_{s+1}), \quad s < t.$$

Pour assurer une paramétrisation unique de la matrice de corrélation, on choisit

$$\Lambda = \{\lambda_{t,s} : t = 2,...m; s < t\}.$$

On utilise la méthode du MCMC comme suggérée par [Smi11], pour générer Γ à l'aide de la paramétrisation de [DP09].

- **Étape** 1 : Générer à partir de $f(\theta_j|\{\boldsymbol{\theta}/\theta_j\}, \Gamma, y)$, tel que $j = 1, \ldots, m$.
- **Étape** 2 : Générer à partir de $f(\tilde{\lambda}_{t,s},\gamma_{t,s}|\Theta, \{\tilde{\Lambda}/\tilde{\lambda}_{t,s}\}, \{\gamma/\gamma_{t,s}\}, y)$, tel que $t = 2, \ldots, m$, $s < t$.
- **Étape** 3 : Calculer Λ à partir de $(\tilde{\Lambda},\gamma)$, où $\gamma = \{\gamma_{t,s} : t = 2, \ldots, m;\ s < t\}$, $\tilde{\lambda}_{t,s} = 0 \Leftrightarrow \gamma_{t,s} = 0$, telle que $\tilde{\lambda}_{t,s}$ et $\tilde{\Lambda}$ sont des variables latentes de $\lambda_{t,s}$ et Λ respectivement.

Pour la deuxième étape, la loi de $(\tilde{\lambda}_{t,s},\gamma_{t,s})$ est $q(\tilde{\lambda}_{t,s},\gamma_{t,s}) = q_1(\gamma_{t,s})q_2(\tilde{\lambda}_{t,s})$ qui a comme probabilité d'acceptation p

$$p = \min\{1,\alpha\tfrac{\Pi(\lambda^{\ell+1})}{\Pi(\lambda^{\ell})}\kappa\}.$$

Si la densité q_2 est symétrique sur $(-1,1)$ avec une distribution Q_2, alors :

$$\kappa = \tfrac{Q_2(1-\tilde{\lambda}^{\ell})-Q_2(-1-\tilde{\lambda}^{\ell})}{Q_2(1-\tilde{\lambda}^{\ell+1})-Q_2(-1-\tilde{\lambda}^{\ell+1})}$$

$$\alpha = \begin{cases} 1 & \text{si } (\tilde{\lambda}^{\ell},\gamma^{\ell} = 0) \to (\tilde{\lambda}^{\ell+1},\gamma^{\ell+1} = 0) \\ \frac{L(\tilde{\lambda}^{\ell+1},\gamma^{\ell+1}=1)\delta_1}{L(0,\gamma^{\ell}=0)\delta_0} & \text{si } (\tilde{\lambda}^{\ell},\gamma^{\ell} = 0) \to (\tilde{\lambda}^{\ell+1},\gamma^{\ell+1} = 1) \\ \frac{L(0,\gamma^{\ell+1}=0)\delta_0}{L(\tilde{\lambda}^{\ell},\gamma^{\ell}=1)\delta_1} & \text{si } (\tilde{\lambda}^{\ell},\gamma^{\ell} = 1) \to (\tilde{\lambda}^{\ell+1},\gamma^{\ell+1} = 0) \\ \frac{L(\tilde{\lambda}^{\ell+1},\gamma^{\ell+1}=1)}{L(\lambda^{\ell},\gamma^{\ell}=0)} & \text{si } (\tilde{\lambda}^{\ell},\gamma^{\ell} = 1) \to (\tilde{\lambda}^{\ell+1},\gamma^{\ell+1} = 1) \end{cases},$$

où $\delta_0 = \Pi(\gamma_{t,s} = 0|\{\gamma/\gamma_{t,s}\})$ et $\delta_1 = \Pi(\gamma_{t,s} = 1|\{\gamma/\gamma_{t,s}\})$ sont les lois a priori de $\gamma_{t,s}$.

Une fois que Λ est généré, on peut calculer Γ (voir l'équation (2) dans l'article de [DP09] et [Joe06]).

Approcher la moyenne a posteriori est le résultat attendu. En utilisant la méthode du MCMC dans le cas de la copule gaussienne, l'itération de la méthode est :

$$\{(\Gamma^{[1]},\Theta^{[1]}),...,(\Gamma^{[L]},\Theta^{[L]})\}$$

avec $(\Gamma^{[L]},\Theta^{[L]}) \sim f(\Gamma,\Theta|y)$

$$\frac{1}{L}\sum_{\ell=1}^{L}\theta_k^{[\ell]} \overset{L\to\infty}{\longrightarrow} \mathbb{E}(\theta_k|y)$$

et

$$\frac{1}{L}\sum_{\ell=1}^{L}\Gamma^{[\ell]} \overset{L\to\infty}{\longrightarrow} \mathbb{E}(\Gamma|y)$$

4 Conclusion

Dans cet article nous avons travaillé sur l'estimation bayésienne de la copule gaussienne. Le domaine des copules est récent, donc l'estimateur bayésien n'est pas assez développé dû à la difficulté causé par le grand nombre de paramètres à estimer. Cela n'empêche pas d'appliquer ces approches sur d'autres copules.

Références

[DP09] Michael J. Daniels et Mohsen Pourahmadi : Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis*, 100:2352–2363, 2009.

[Fré51] Maurice Fréchet : Sur les tableaux dont les marges et des bornes sont données. *Ann.Univ. Lyon, Science*, 1951.

[Joe06] Harry Joe : Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97:2177–2189, 2006.

[Nel06] Roger B. Nelsen : *An Introduction to Copulas*. Springer, 2006.

[PCK06] Michael Pitt, David Chan et Robert Kohn : Efficient bayesian inference for gaussian copula regression models. *Biometrika*, 93:537–554, 2006.

[Ra11] Christian P. Robert et George Casella (auth.) : *Méthodes de Monte-Carlo avec R*. Springer Paris, 2011.

[Skl59] Abe Sklar : Fonctions de répartition à n dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 1959.

[Smi11] Michael Stanley Smith : Bayesian approaches to copula modelling. *Methodology (stat.ME)*, 2011.

Marwa Hamdi
Département de mathématiques, Université de Sherbrooke
Courriel: Marwa.Hamdi@USherbrooke.ca

CaMUS **7**, 45 – 66

Polynômes à déviation minimale sur l'union de deux intervalles

Gabriel Dupuis

RÉSUMÉ Cet article étudie la proposition qui dresse une équivalence entre une classe de polynômes à déviation minimale sur une union d'intervalles et l'équation de Pell définie à partir de cette même union d'intervalles. Après avoir énoncé cette proposition, nous tenterons de comprendre puis de montrer cet énoncé à l'aide de problèmes résolus et résultats importants de la théorie des polynômes extrêmes. Nous verrons, entre autres, les polynômes de Tchebycheff, de Akhiezer et de Tchebycheff généralisés ainsi que le théorème général de Tchebycheff. Une fois la proposition comprise et montrée, nous allons voir une construction de polynômes à déviation minimale sur l'union de deux intervalles basée sur cette même proposition.

1 Introduction

Historiquement, P.L. Tchebycheff fut le premier à poser et résoudre des problèmes dans lesquels, étant donné un sous-ensemble fermé de la droite réelle et une fonction $f(x)$ continue sur ce sous-ensemble, le but est de trouver une fonction $g(x)$, parmi une famille de fonctions, qui minimise la norme supremum de la différence entre $f(x)$ et $g(x)$. Ces problèmes ont émergé de questionnements d'ingénierie où le but était de minimiser la friction dans les jointures du parallélogramme de Watt qui est la pièce de la machine à vapeur transformant le mouvement de va-et-vient en mouvement de rotation. Les recherches de Tchebycheff ont mené à une modification des liens du parallélogramme en question avec les autres pièces de la machine à vapeur, modifications encore utilisées à ce jour [Bog05]. C'est ensuite ses étudiants, E.I. Zolotarev puis N.I. Akhiezer qui ont pris la relève en posant et résolvant d'autres problèmes du même type. Aujourd'hui, les résultats obtenus en résolvant ces problèmes sont utilisés dans la théorie de l'approximation afin de répondre à des questions relatives au génie électrique.

Avant d'énoncer certains problèmes, la notion de déviation doit être introduite. Cette notion est une façon de mesurer à quel point deux fonctions sont près ou éloignées l'une de l'autre.

Ce travail a reçu le soutien de Mitacs dans le cadre de Bourse de formation à la recherche Mitacs. Aussi, j'aimerais remercier Mme Vasilisa Shramchenko, Professeure à l'Université de Sherbrooke, pour la supervision de mon stage de recherche et pour son aide tout au long de la rédaction de cet article.

Définition 1.1. La *déviation* d'une fonction continue $g(x)$ par rapport à une fonction continue $f(x)$ sur un sous-ensemble Θ, fini ou infini, fermé de la droite réelle est définie par la quantité :

$$\sup_{x\in\Theta} |f(x) - g(x)|.$$

La déviation s'avère être la norme suprémum de la fonction $f - g$ se trouvant dans l'espace des fonctions continues définies sur Θ.

Parmi les problèmes proposés par P.L. Tchebycheff, E.I. Zolotarev et N.I. Akhiezer, nous retrouvons les cinq problèmes de la théorie des polynômes extrêmes suivants.

Problème 1. Étant donné un intervalle fermé $[a,b]$ de la droite réelle et deux fonctions à valeurs réelles $f(x)$, $s(x)$ continues sur $[a,b]$ telles que $s(x) \neq 0$, $\forall x \in [a,b]$. Considérons l'expression :

$$Q(x) = s(x)\frac{q_0x^n + q_1x^{n-1} + \ldots + q_n}{p_0x^m + p_1x^{m-1} + \ldots + p_m} \tag{1}$$

où m et n sont donnés. Trouver les paramètres réels p_0, p_1, … , p_m ; ainsi que q_0, q_1, … , q_n qui sont tels que la déviation de $Q(x)$ par rapport à $f(x)$ soit minimale.

Problème 2. Trouver le polynôme monique de degré n dont la déviation par rapport à zéro sur l'intervalle $[-1, 1]$ est minimale parmi l'ensemble des polynômes moniques de degré n.

Problème 3. Étant donné un paramètre σ réel fixé, trouver le polynôme de degré n de la forme :

$$x^n - n\sigma x^{n-1} + q_0x^{n-2} + \ldots + q_{n-2},$$

dont la déviation par rapport à zéro sur $[-1,1]$ est minimale parmi l'ensemble des polynômes de cette forme.

Problème 4. Trouver le polynôme monique de degré n dont la déviation par rapport à zéro sur les intervalles symétriques $[-1, -a]\cup[a, 1]$ est minimale parmi l'ensemble des polynômes moniques de degré n.

Problème 5. Étant donné $a,b \in \mathbb{R}$ tel que $-1 < a < b < 1$, trouver le polynôme monique de degré n dont la déviation par rapport à zéro sur les intervalles $[-1,a] \cup [b,1]$ est minimale parmi l'ensemble des polynômes moniques de degré n.

Suivant principalement les Problèmes 4 et 5, il vient naturellement un problème au cadre plus général encore.

Problème 6. Étant donné c_1, c_2, c_3, $c_4 \in \mathbb{R}$ tels que $c_4 < c_3 < c_2 < c_1$, trouver le polynôme monique de degré n dont la déviation par rapport à zéro sur les intervalles $[c_4,c_3]\cup[c_2,c_1]$ est minimale parmi l'ensemble des polynômes moniques de degré n.

Le but de cet article est de construire une formule explicite pour une famille de solutions du Problème 6. Pour définir cette famille et construire la formule explicite des polynômes qui la compose, nous allons étudier un lien étonnant entre une équation de Pell et certaines solutions du Problème 6. La proposition qui suit énonce ce lien.

Pour certaines constantes c_1, c_2, c_3, $c_4 \in \mathbb{R}$ qui satisfassent les inégalités $c_4 < c_3 < c_2 < c_1$, il existe $\hat{p}_n$ et $\hat{q}_{n-2}$, des polynômes à coefficients réels de degré n et $n-2$ respectivement, tel que l'équation de Pell :

$$\hat{p}_n^2(x) - \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x) = 1, \tag{2}$$

où

$$\hat{\mathscr{P}}_4(x) = \prod_{j=1}^{4}(x - c_j),$$

est vérifiée. Nous notons $\hat{P}_n$ la solution du Problème 6 lorsque l'on considère les intervalles $[c_4,c_3] \cup [c_2,c_1]$ et L_n sa déviation sur cette union d'intervalles.

Proposition 1.2. *Soit c_1, c_2, c_3, $c_4 \in \mathbb{R}$ tels que $c_4 < c_3 < c_2 < c_1$ et soit $\hat{p}_n$, un polynôme à coefficients réels de degré n. Il existe $\hat{q}_{n-2}$, un polynôme à coefficients réels de degré $n-2$, tel que $\hat{p}_n$ et $\hat{q}_{n-2}$ vérifie l'équation de Pell (2) si et seulement si $\hat{P}_n$, la solution du Problème 6 pour les intervalles $[c_4,c_3]\cup[c_2,c_1]$, vérifie les deux conditions suivantes :*

1. $\hat{p}_n(x) = \hat{P}_n(x)/\pm L_n$;
2. *L'ensemble $[c_4,c_3] \cup [c_2,c_1]$ est le sous-ensemble maximal de $\mathbb{R}$ pour lequel $\hat{P}_n(x)$ est le polynôme monique de degré n déviant le moins de zéro.*

Dans la première section, il sera question d'un résultat important de la théorie des polynômes extrêmes, soit le théorème général de Tchebycheff (aussi appelé théorème d'alternance). Ce théorème découle directement de la recherche d'une solution au Problème 1 et nous verrons comment appliquer ce résultat aux Problèmes 2 et 3. Le travail effectué dans cette section servira de base à plusieurs raisonnements logiques des sections suivantes en plus de nous permettre de déduire une propriété importante du polynôme $\hat{P}_n$ de la Proposition 1.2.

Dans la seconde section nous discuterons de la maximalité des intervalles $[c_4,c_3] \cup [c_2,c_1]$ dont il est question au point 2. de l'énoncé de la Proposition 1.2. Cette discussion se fera à travers la recherche de solutions aux Problèmes 4 et 5. D'un côté, la recherche d'une solution au Problème 4 nous permettra de comprendre ce concept de maximalité des intervalles de façon générale. De l'autre, la recherche d'une solution au Problème 5 permettra de distinguer une famille

de solutions particulière à ce problème, distinction qui sera étroitement liée à ce même concept de maximalité des intervalles.

Dans la quatrième section nous ferons la preuve de la Proposition 1.2 et poserons une définition de la famille de solutions vérifiant les conditions 1. et 2. de celle-ci. Il s'avèrera que cette famille de solutions sera étroitement liée à la famille trouvée au Problème 5. Ensuite, nous tenterons de construire une formule explicite pour cette famille de polynômes tout en dégageant les limites de cette même construction. Le fil logique de cette section sera guidé par des arguments utilisés dans les sections antérieures.

La dernière section conclura en donnant quelques résultats étroitement liés à la Proposition 1.2, résultats qui répondront à un questionnement qui n'aura été que brièvement abordé dans les Sections 3 et 4.

2 Théorème général de Tchebycheff

2.1 Théorème général de Tchebycheff et son corollaire

Le Problème 1 étant une généralisation des Problèmes 2, 3 et bien plus encore, il est difficile, voir impossible, d'en déterminer une solution générale. Toutefois, la recherche d'une solution à ce problème a donné lieu à la démonstration d'un théorème d'importance dans la théorie des polynômes extrêmes, soit le théorème général de Tchebycheff. Avant d'énoncer ce théorème, nous devons poser deux définitions donnant un nom aux points où la déviation d'une fonction par rapport à une autre atteint son maximum.

Définition 2.1. Soient deux fonctions continues $g(x)$ et $f(x)$ sur un sous-ensemble Θ fini ou infini fermé de la droite réelle. Un *point de déviation* est un point $x_0 \in \Theta$ tel que :

$$|f(x_0) - g(x_0)| = \sup_{x \in \Theta} |f(x) - g(x)|.$$

Définition 2.2. Soit deux fonctions continues $g(x)$ et $f(x)$ définies sur un sous-ensemble Θ fini ou infini fermé de la droite réelle. Un *ensemble de points d'alternance* est une suite finie croissante maximale de points de déviation sur laquelle le signe de $f(x) - g(x)$ alterne.

Par exemple, considérons $\Theta = [c_4,c_3] \cup [c_2,c_1]$, $f(x) = 0$ et $g(x)$ le polynôme illustré à la Figure 1. Les points de déviation sont c_4, c_3, c_2, c_1, y_1, y_2 et y_3. Il y a deux ensembles de points d'alternance, étant donnée la condition de maximalité de la définition, soit $\{c_4, y_1, y_2, y_3, c_3, c_1\}$ et $\{c_4, y_1, y_2, y_3, c_2, c_1\}$. Nous pouvons maintenant énoncer le théorème général de Tchebycheff.

Théorème 2.3 (Théorème général de Tchebycheff)**.** *Soit un intervalle* $[a,b]$ *réel fermé et une fonction* $f(x)$ *continue sur cet intervalle. Parmi les fonctions de la forme* (1)*, il existe une fonction* $P(x)$ *déviant le moins de la fonction* $f(x)$ *dans* $[a,b]$*. Les polynômes du numérateur et du dénominateur de* $P(x)$ *sont*

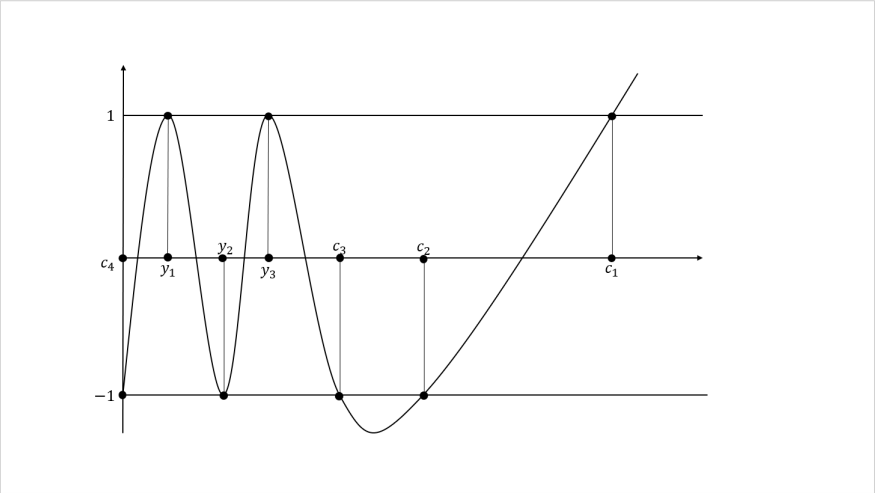


FIGURE 1 : Graphe de $\hat{p}_5(x)$ avec $(m_0,m_1) = (5,4)$.

uniquement déterminés si on suppose que la fraction est réduite. La fonction $P(x)$ est complètement caractérisée par la propriété suivante : si la fonction $P(x)$ s'exprime sous la forme :

$$P(x) = s(x)\frac{b_0x^{n-\nu} + \ldots + b_{n-\nu}}{a_0x^{m-\mu} + \ldots + a_{m-\mu}} = s(x)\frac{B(x)}{A(x)},$$

où $0 \leq \mu \leq m$, $0 \leq \nu \leq n$, $a_0 \neq 0$ et la fraction $\frac{B(x)}{A(x)}$ est irréductible, alors le nombre de points d'alternance dans $[a,b]$ est supérieur ou égal à $m+n+2-d$, où $d = \min\{\mu,\nu\}$.

Remarque 2.4. L'unicité et la caractérisation de $P(x)$ permettent d'obtenir l'équivalence suivante : étant donné $Q(x)$ de la forme (1), alors $Q(x) = P(x)$ si et seulement si pour le nombre de points d'alternance de $Q(x)$ par rapport à $f(x)$ sur $[a,b]$ est supérieur ou égal à $m+n+2-d$.
La preuve de cette équivalence se trouve à la fin de la preuve (du théorème général de Tchebycheff) à la section 34 du livre [Akh92].

Ce théorème s'adapte au cas particulier où la formule (1) représente un polynôme de degré n, adaptation qui servira tout au long de l'article étant donné que tous les problèmes que nous traitons sont polynomiaux.

Corollaire 2.5. *Soit $[a,b]$ un intervalle réel fermé et $f(x)$ une fonction continue sur cet intervalle. Si $s(x) = 1$ et $m = 0$ dans l'expression (1), alors le Problème 1 devient celui de chercher le polynôme de degré n dont la déviation sur $[a,b]$ par rapport à une fonction $f(x)$ est minimale parmi l'ensemble des polynômes de degré n. Dans ce contexte, le théorème général de Tchebycheff devient :*

Parmi l'ensemble des polynômes de degré n, il existe un unique polynôme $P(x)$ dont la déviation sur $[a,b]$ par rapport à $f(x)$ est minimale. Ce polynôme

se caractérise par le fait que le nombre de points d'alternance dans $[a,b]$ *est supérieur ou égal à* $n+2$.

Remarque 2.6. L'unicité et la caractérisation du polynôme $P(x)$ nous permettent d'obtenir l'équivalence suivante : étant donné un polynôme $Q(x)$ de degré n, alors $Q(x) = P(x)$ si et seulement si le nombre de points d'alternance de $Q(x)$ par rapport à $f(x)$ sur $[a,b]$ est supérieur ou égal à $n+2$.

Remarque 2.7. Nous pouvons retrouver l'énoncé du Problème 1, la preuve du théorème général de Tchebycheff et l'énoncé du Corollaire 2.5 dans les sections 31 à 35 du livre [Akh92].

2.2 Application du théorème général de Tchebycheff aux Problèmes 2 et 3

Voyons ce que le théorème de Tchebycheff nous permet de déduire à propos des solutions aux Problèmes 2 et 3 qui consistent tous deux à déterminer un polynôme dont la déviation par rapport à zéro est minimale sur $[-1,1]$ parmi une certaine famille de polynômes. Dans le Problème 2, nous nous intéressons à la famille de polynômes de la forme :

$$x^n + q_0 x^{n-1} + \ldots + q_{n-1}, \quad q_0, \ldots, q_{n-1} \in \mathbb{R}$$

et nous cherchons les paramètres $q_0, \ldots, q_{n-1}$ minimisant la norme supremum de ce polynôme sur $[-1,1]$, c'est-à-dire :

$$\inf_{q_0,\ldots,q_{n-1}} \left(\sup_{[-1,1]} |x^n + q_0 x^{n-1} + \ldots + q_{n-1}| \right).$$

En posant $f(x) = x^n$ et $[a,b] = [-1,1]$, et en considérant les polynômes de degré $n-1$ dans l'énoncé du Corollaire 2.5, alors il existe un unique polynôme $P(x) = b_0 x^{n-1} + \ldots + b_{n-1}$ solution au Problème 1. De plus, $P(x)$ se caractérise par le fait qu'il y a au moins $(n-1)+2 = n+1$ points d'alternance dans $[-1,1]$. Par définition, $P(x)$ vérifie :

$$\sup_{x\in[-1,1]} |f(x) - P(x)| = \inf_{q_0,\ldots,q_{n-1}} \left(\sup_{x\in[-1,1]} |x^n - q_0 x^{n-1} - \ldots - q_{n-1}| \right).$$

Ainsi, $f(x) - P(x)$ est le polynôme monique de degré n déviant le moins de zéro sur $[-1,1]$ et se caractérise par le fait qu'il y a au moins $n+1$ points d'alternance dans cet intervalle. La Remarque 2.6 nous permet de reformuler en disant que $f(x) - P(x)$ est l'unique polynôme monique de degré n tel que le nombre de points d'alternance sur $[a,b]$ est supérieur ou égal à $n+1$.
Plus particulièrement, un point d'alternance dans $(-1,1)$ est un extrémum local, ce qui entraine qu'il y a au plus $n-1$ points d'alternance dans $(-1,1)$, comme $f(x) - P(x)$ est un polynôme de degré n. Par conséquent, il est évident que la solution $f(x) - P(x)$ du Problème 2 se caractérise par le fait qu'il y a exactement $n+1$ points d'alternance dans $[-1,1]$ dont $n-1$ points sont dans $(-1,1)$ et les deux autres sont en ± 1.

Définition 2.8. Les *polynômes de Tchebycheff*, notés $T_n(x)$ ($n = 0, 1, 2, \ldots$), sont définis comme suit :

$$T_n(x) = \cos(n\varphi), \quad x = \cos(\varphi) \quad \text{et} \quad n = 0, 1, 2, \ldots \tag{3}$$

$T_n(x)$ vérifie la relation de récurrence suivante :

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad T_0 = 1 \quad \text{et} \quad T_1 = x. \tag{4}$$

On appelle coefficient de déviation minimale, noté L_n, la quantité :

$$L_n = \begin{cases} 1 & n = 0 \\ 2^{1-n} & n = 0, 1, 2, \ldots \end{cases}.$$

On remarque alors que $L_nT_n(x)$ est monique de degré n et possède exactement $n + 1$ points d'alternance dans $[-1,1]$, points que nous notons :

$$x_k = \cos\left(\frac{(n-k)\pi}{n}\right), \quad k = 0, 1, \ldots, n.$$

La caractérisation de la solution du Problème 2 trouvée ci-haut nous permet de conclure que $L_nT_n(x)$ est la solution recherchée.

Nous pouvons trouver une caractérisation, similaire à celle du Problème 2, pour la solution du Problème 3 en posant $f(x) = x^n - n\sigma x^{n-1}$, $[a,b] = [-1,1]$ et en considérant les polynômes de degré $n - 2$ dans l'énoncé du Corollaire 2.5. La solution $f(x) - P(x)$ du Problème 3 se caractérisera alors plutôt par le fait qu'elle possède exactement n points d'alternance dans $[-1,1]$ dont $n - 2$ points sont dans $(-1,1)$ et les deux autres sont en ± 1. La solution du Problème 3 est appelée polynôme de Zolotarev, notée $z_n(x)$.

Remarque 2.9. Pour davantage de détails à propos des solutions aux Problèmes 2 et 3, voir la section 5 de l'article [VD18]. La démarche de la construction des polynômes de Tchebycheff à partir du théorème général de Tchebycheff se trouve à la section 36 du livre [Akh92]. À noter que la formule de la solution du Problème 3 utilise les fonctions elliptiques.

Les démarches effectuées ici nous montrent qu'étant donné le contexte du Problème 1, si nous trouvons une expression pour laquelle nous avons le bon nombre de points d'alternance, alors cette solution est la solution recherchée. Nous utiliserons cette idée régulièrement dans ce qui suit.

2.3 Généralisation du théorème général de Tchebycheff

La famille de polynômes parmi laquelle nous cherchons une solution au Problème 2, c'est-à-dire les polynômes de la forme :

$$x^n + q_0x^{n-1} + \ldots + q_{n-1}, \quad q_0, \ldots, q_{n-1} \in \mathbb{R},$$

est la même famille que celle parmi laquelle nous cherchons une solution aux Problèmes 4, 5 et 6. La seule différence entre ces problèmes est le sous-ensemble

réel d'intérêt, c'est-à-dire respectivement $[-1,1]$, $[-1,-a]\cup[a,1]$, $[-1,a]\cup[b,1]$ et $[c_4,c_3]\cup[c_2,c_1]$. Il s'avère que le Corollaire 2.5 s'applique également aux problèmes posés sur l'union de deux intervalles et donc que la discussion ci-dessus concernant la solution du Problème 2 est également valide pour la solution respective des Problèmes 4, 5 et 6. Par conséquent, la solution du Problème 4 (resp. 5 et 6) existe et est l'unique polynôme monique de degré n admettant au moins $n+1$ points d'alternance sur $[-1,-a]\cup[a,1]$ (resp. $[-1,a]\cup[b,1]$ et $[c_4,c_3]\cup[c_2,c_1]$).

Ces dernières lignes nous donnent ainsi une propriété fondamentale de la solution du Problème 6 et, par conséquent, du polynôme $\hat{P}_n$ de la Proposition 1.2.

3 Maximalité des intervalles vue à travers les Problèmes 4 et 5

Nous avons déterminé une propriété intéressante du polynôme $\hat{P}_n$ de la Proposition 1.2, soit qu'il est l'unique polynôme monique de degré n admettant au moins $n+1$ points d'alternance sur $[c_4,c_3]\cup[c_2,c_1]$. Toutefois, $\hat{P}_n$ doit vérifier la condition 2. de la Proposition 1.2, c'est-à-dire celle concernant la maximalité des intervalles $[c_4,c_3]\cup[c_2,c_1]$. Cette condition peut potentiellement apporter des contraintes à propos de la disposition des points d'alternance dans $[c_4,c_3]\cup[c_2,c_1]$. Nous tenterons donc, dans la présente section, de mieux comprendre la signification de cette condition de maximalité à travers la recherche de solutions aux Problèmes 4 et 5.

3.1 Problème 4

La solution du Problème 4 est appelée polynôme de Akhiezer et se définit comme suit.

Définition 3.1. Le *polynôme de Akhiezer* de degré $n \in \mathbb{N}$, noté $A_n(x,a)$, est le polynôme de la forme :

$$A_n(x,a) = x^n + b_0 x^{n-1} + \ldots + b_{n-1}, \quad b_0, \ldots, b_{n-1} \in \mathbb{R}, \tag{5}$$

qui dévie le moins par rapport à zéro sur les intervalles égaux :

$$[-1,-a]\cup[a,1] \quad \text{avec} \quad a \in (0,1).$$

La déviation de $A_n(x,a)$ par rapport à zéro est noté $L_n(a)$.

Tel que discuté à la Section 2.3, $A_n(x,a)$ existe et est unique pour toutes valeurs de a. De plus, $A_n(x,a)$ est l'unique polynôme monique de degré n comptant au moins $n+1$ points d'alternance dans $[-1,-a]\cup[a,1]$. Aussi, de l'unicité et de la symétrie des intervalles, nous pouvons facilement déduire que pour n pair (resp. impair), $A_n(x,a)$ est paire (resp. impaire).

Étudions le cas particulier $n = 2m-1$, $m \in \mathbb{N}_{>0}$. Dans ce cas, $A_{2m-1}(x,a)$ est l'unique polynôme monique de degré $2m-1$ comptant au moins $2m$ points

d'alternance dans $[-1,-a]\cup[a,1]$. Nous avons vue que le polynôme de Tchebycheff $T_{2m-1}(x)$ possède les $2m$ points d'alternance sur $[-1,1]$:

$$x_0 < x_1 < \ldots < x_{m-1} < 0 < x_m < \ldots < x_{2m-1},$$

où

$$x_k = \cos\left(\frac{((2m-1)-k)\pi}{(2m-1)}\right), \quad k = 0,\ 1,\ \ldots,\ 2m-1.$$

Donc, si $a \leq x_m (= -x_{m-1})$ alors $A_{2m-1}(x,a) = L_{2m-1}T_{2m-1}(x)$, car les $2m$ points d'alternance de $L_{2m-1}T_{2m-1}(x)$ se retrouvent dans $[-1,-a]\cup[a,1]$. Toutefois, la formule de $A_{2m-1}(x,a)$ est beaucoup plus difficile à trouver pour $a > x_m$, comme $L_{2m-1}T_{2m-1}(x)$ n'a plus tous ses $2m$ points d'alternance dans $[-1,-a]\cup[a,1]$. Ainsi, la valeur $a = x_m$ est la « frontière » à partir de laquelle $L_{2m-1}T_{2m-1}$ n'est plus à déviation minimale sur $[-1,-a]\cup[a,1]$.

Ainsi, la discussion concernant $A_{2m-1}(x,a)$ nous fait rendre compte qu'un polynôme qui est la solution d'un problème tels les Problèmes 4, 5 et 6 peut potentiellement l'être pour une multitude d'unions de deux intervalles. De plus, cette même discussion nous montre qu'il y a des bornes, pour les unions de deux intervalles, à partir desquelles ce même polynôme n'est plus la solution. En effet, nous avons vu que $L_{2m-1}T_{2m-1}(x)$ est le polynôme monique de degré $2m-1$ dont la déviation sur l'intervalle $[-1,1]$ et les unions d'intervalles $[-1,-a]\cup[a,1]$, $a \in (0,x_m]$, est minimale parmi l'ensemble des polynômes monique de degré $2m-1$. Nous pouvons donc imaginer que lorsqu'il est question de maximalité de $[c_4,c_3]\cup[c_2,c_1]$ dans la Proposition 1.2 ceci signifie que $\hat{P}_n$ est à déviation minimale sur une multitude d'unions de deux intervalles et que l'union particulier d'intervalles $[c_4,c_3]\cup[c_2,c_1]$ en question est celle à partir de laquelle $\hat{P}_n$ n'est plus à la déviation minimale.

Remarque 3.2. Les formules explicites de $A_n(x,a)$ pour tous les n se trouvent à la section 52 du livre [Akh70]. Pour les comprendre, il faut connaître les fonctions elliptiques.

3.2 Problème 5

Étant donnée l'union d'intervalles :

$$[-1,a]\cup[b,1] \quad \text{avec} \quad -1 < a < b < 1,$$

la solution du Problème 5 est un polynôme de degré n de la forme :

$$M_n(x) = x^n + b_0x^{n-1} + \ldots + b_{n-1}, \quad b_0,\ \ldots,\ b_{n-1} \in \mathbb{R}, \tag{6}$$

tel que

$$\sup_{x\in[-1,a]\cup[b,1]} |M_n(x)| = \inf_{q_0,\ \ldots,\ q_{n-1}\in\mathbb{R}} \left(\sup_{x\in[-1,a]\cup[b,1]} \left| x^n + q_0x^{n-1} + \ldots + q_{n-1} \right| \right).$$

Tel que discuté à la Section 2.3, étant donnée l'union d'intervalles donnée par $[-1,a]\cup[b,1]$, il existe un unique polynôme $M_n(x)$ tel que décrit ci-haut. De

plus, ce polynôme est le seul polynôme monique de degré n comptant au moins $n+1$ points d'alternance dans $[-1,a]\cup[b,1]$. Toutefois, il n'est pas garanti que $[-1,a]\cup[b,1]$ soit le sous-ensemble maximal de $\mathbb{R}$ sur lequel $M_n(x)$ est minimal. Examinons un cas particulier de solutions pour lesquelles la maximalité des intervalles est atteinte.

Définition 3.3. Un polynôme de la forme :

$$\mathscr{T}_n(x) = x^n + b_0x^{n-1} + \ldots + b_{n-1}, \quad b_0, \ldots, b_{n-1} \in \mathbb{R}, \tag{7}$$

est appelé *polynôme de Tchebycheff généralisé* (abrégé par T-polynôme) sur $[-1,a]\cup[b,1]$ s'il a exactement $n+2$ points de déviation dans $[-1,a]\cup[b,1]$, points notés :

$$x_1 < x_2 < \ldots < x_{n+2}.$$

Nous notons :

$$L_n = \sup_{x\in[-1,a]\cup[b,1]} |\mathscr{T}_n(x)|,$$

et disons que $\mathscr{T}_n(x)$ est un polynôme normalisé sur $[-1,a]\cup[b,1]$ si :

$$\mathscr{T}_n(x) = \mathscr{T}_n(x)/L_n.$$

Nous verrons en quoi la condition selon laquelle $\mathscr{T}_n(x)$ admet exactement $n+2$ points de déviation dans $[-1,a]\cup[b,1]$ assure que ce polynôme est solution du Problème 5. Nous verrons également le lien entre cette propriété et la maximalité des intervalles.

Proposition 3.4. *$\mathscr{T}_n(x) = x^n+\ldots+b_0x^{n-1}+\ldots+b_{n-1}$ est un T-polynôme sur $[-1,a]\cup[b,1]$ si et seulement si $\mathscr{T}_n(x)$ possède exactement $n-2$ points intérieurs $x_j \in (-1,a)\cup(b,1)$*

$$x_1 < \ldots < x_i < a < b < x_{i+1} < \ldots < x_{n-2}$$

et les quatre points frontières ±1, a et b comme points de déviation où les suites de points d'alternance sont

$$-1, x_1, \ldots, x_i, b, x_{i+1}, \ldots, x_{n-2}, 1$$

ou

$$-1, x_1, \ldots, x_i, a, x_{i+1}, \ldots, x_{n-2}, 1.$$

Démonstration. ($\Longleftarrow$) Direct par la définition de T-polynôme.
($\Longrightarrow$) Supposons que $\mathscr{T}_n(x)$ est un polynôme de Tchebycheff généralisé, alors $\mathscr{T}_n(x)$ admet exactement $n+2$ points de déviation sur $[-1,a]\cup[b,1]$. Ceci signifie qu'il y a soit $n-2$ points intérieurs et les quatre points frontières ou $n-1$ points intérieurs et trois des quatre points frontières qui sont des points de déviation. Or, ce dernier cas est impossible. En effet, si $n-1$ points intérieurs sont de déviation, alors $\mathscr{T}_n(x)$ compte $n-1$ extrémums locaux dans $(-1,a)\cup(b,1)$. De

plus, comme a et ou b est de déviation, il doit y avoir au moins un extrémum local dans $[a,b]$, ce qui signifie au moins n extrémums locaux pour un polynôme de degré n.

L'alternance des suites de points est facile à vérifier à l'aide d'arguments faisant intervenir le nombre d'extrémums locaux d'un polynôme de degré n.

□

Remarque 3.5. $\mathscr{T}_n(x)$ admet exactement un extrémum local dans (a,b) étant donné que a et b sont des points de déviation sans être tous deux des points d'alternance et que $n-2$ extrémums locaux se retrouvent dans $(-1,a)\cup(b,1)$.

La proposition précédente nous donne le nombre de points d'alternance, soit $n+1$, et la forme des ensembles de points d'alternance des polynômes de Tchebycheff généralisés sur $[-1,a]\cup[b,1]$. Ainsi, nous pouvons en conclure qu'un polynôme de Tchebycheff généralisé sur $[-1,a]\cup[b,1]$ est le polynôme monique de degré n dont la déviation est minimale sur $[-1,a]\cup[b,1]$ parmi l'ensemble des polynômes moniques de degré n. Le corollaire suivant formalise cette dernière conclusion.

Corollaire 3.6. *Soit $\mathscr{T}_n(x) = x^n + \ldots + b_0x^{n-1} + \ldots + b_{n-1}$ un T-polynôme sur $[-1,a]\cup[b,1]$ avec les points de déviation :*

$$-1 < x_1 < \ldots < x_i < a < b < x_{i+1} < \ldots < x_{n-2} < 1$$

et les points d'alternance :

$$-1,\, x_1,\, \ldots,\, x_i,\, b,\, x_{i+1},\, \ldots,\, x_{n-2},\, 1$$

resp.

$$-1,\, x_1,\, \ldots,\, x_i,\, a,\, x_{i+1},\, \ldots,\, x_{n-2},\, 1.$$

Alors $\mathscr{T}_n(x)$ est le polynôme monique de degré n dont la déviation par rapport à zéro sur $[-1,\lambda]\cup[b,1]$ où $x_i \leq \lambda \leq a$ (resp. $[-1,a]\cup[\hat{\lambda},1]$ où $b \leq \hat{\lambda} \leq x_{i+1}$) est minimale.

Démonstration. Découle directement de la discussion à la Section 2.3, comme il y a au moins $n+1$ points d'alternance dans tous les cas proposés.

□

Remarque 3.7. La réciproque n'est pas vraie, c'est-à-dire qu'un polynôme monique de degré n dont la déviation par rapport à zéro dans $[-1,a]\cup[b,1]$ est minimale n'est pas nécessairement un T-polynôme sur $[-1,a]\cup[b,1]$. Par exemple, un T-polynôme sur $[-1,a]\cup[b,1]$ n'est pas un T-polynôme sur $[-1,\lambda]\cup[b,1]$ pour $x_i \leq \lambda < a$ comme ce polynôme ne compte que $n+1$ points de déviation (et d'alternance) dans $[-1,\lambda]\cup[b,1]$ et non $n+2$.

Ce corollaire nous montre donc que $[-1,a]\cup[b,1]$ est le sous-ensemble maximal de $\mathbb{R}$ sur lequel le polynôme de Tchebycheff généralisé sur $[-1,a]\cup[b,1]$ est solution du Problème 6. Nous pouvons donc voir la condition définissant

les polynômes de Tchebycheff généralisés, soit avoir exactement $n+2$ points de déviation dans $[-1,a]\cup[b,1]$, comme étant une condition qui garantit la maximalité des intervalles $[-1,a]\cup[b,1]$. Également, dans le cas où cette condition est vérifiée, alors nous avons directement la forme des ensembles de points d'alternance. Cette condition reviendra potentiellement pour définir le polynôme $\hat{P}_n$ de la Proposition 1.2 de par son lien avec la maximalité des intervalles.

Remarque 3.8. L'ensemble des polynômes de Tchebycheff généralisés peut donc être perçu comme une famille de solutions du Problème 5. Également, lorsque nous étudions cette famille de polynômes en particulier, cela équivaut (intuitivement, car un T-polynôme est un T-polynôme que sur une unique union d'intervalles) à poser un critère sur les intervalles $[-1,a]\cup[b,1]$ à l'étude, critère qui se manifeste sous forme de formules pour a et b telles qu'énoncées à la section 5.3 de l'article [VD18].

4 Équation de Pell et Problème 6

Nous avons discuté du théorème général de Tchebycheff qui nous donnait une propriété intéressante que possède le polynôme $\hat{P}_n$ de la Proposition 1.2. Ensuite, nous avons discuté de la maximalité des intervalles dont il est question dans cette proposition et avons ainsi vu qu'elle est potentiellement en lien avec le nombre de points de déviation et la forme de l'ensemble de points d'alternance dans $[c_4,c_3]\cup[c_2,c_1]$. Reste à prouver la Proposition 1.2 afin d'avoir en main assez de propriétés concernant $\hat{P}_n$ pour en déduire une formule explicite.

4.1 Preuve de la Proposition 1.2 et propriété de $\hat{P}_n$

Étant donné l'union d'intervalles :

$$[c_4,c_3]\cup[c_2,c_1] \quad \text{avec} \quad c_4<c_3<c_2<c_1,$$

la solution du Problème 6 est un polynôme de degré n de la forme :

$$\hat{P}_n(x)=x^n+b_0x^{n-1}+\ldots+b_{n-1}, \quad b_0,\ \ldots,\ b_{n-1}\in\mathbb{R}, \tag{8}$$

tel que

$$\sup_{x\in[c_4,c_3]\cup[c_2,c_1]}|\hat{P}_n(x)|=\inf_{q_0,\,\ldots,\,q_{n-1}\in\mathbb{R}}\left(\sup_{x\in[c_4,c_3]\cup[c_2,c_1]}|x^n+q_0x^{n-1}+\ldots+q_{n-1}|\right).$$

Notons L_n sa déviation sur $[c_4,c_3]\cup[c_2,c_1]$.

Tel que discuté à la Section 2.3, étant donnée l'union particulier d'intervalles $[c_4,c_3]\cup[c_2,c_1]$ il existe un unique polynôme $\hat{P}_n$ tel que décrit ci-haut. De plus, ce polynôme est le seul polynôme monique de degré n comptant au moins $n+1$ points d'alternance dans $[c_4,c_3]\cup[c_2,c_1]$. Toutefois, il n'est pas garanti que ce polynôme vérifie les conditions 1. et 2. de la Proposition 1.2. Ainsi, montrons la Proposition 1.2 pour ensuite déterminer la famille de solutions vérifiant les points 1. et 2. de celle-ci.

Démonstration. Soit c_1, c_2, c_3, $c_4 \in \mathbb{R}$ tels que $c_4 < c_3 < c_2 < c_1$.
($\Longrightarrow$) Supposons que nous avons $\hat{p}_n(x)$ et $\hat{q}_{n-2}(x)$ vérifiant l'équation de Pell (2) et vérifions 1. et 2.

L'équation de Pell nous permet de déduire deux choses. D'abord,

$$\hat{p}_n^2(x) - \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x) = 1 \iff (\hat{p}_n(x) - 1)(\hat{p}_n(x) + 1) = \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x).$$

Notons aussi que, $|\hat{p}_n(x)| = 1$, pour $x \in \{c_4, c_3, c_2, c_1\}$, comme $\{c_4, c_3, c_2, c_1\}$ sont les racines de $\hat{\mathscr{P}}_4(x)$. Ensuite,

$$\hat{\mathscr{P}}_4(x) < 0, \quad \forall x \in (c_4,c_3) \cup (c_2,c_1)$$

étant donnée que pour $x \in (c_4,c_3)$ nous avons $(x - c_4) > 0$ et $(x - c_i) < 0$, $i = 1$, 2, 3, et pour $x \in (c_2,c_1)$ nous avons $(x - c_i) > 0$, $i = 2$, 3, 4, et $(x - c_1) < 0$. De ces deux derniers constats, nous déduisons que :

$$(\hat{p}_n(x) - 1)(\hat{p}_n(x) + 1) = \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x) \leq 0, \quad \forall x \in [c_4,c_3] \cup [c_2,c_1]$$

$$\begin{aligned}
&\Longrightarrow \hat{p}_n^2(x) - 1 = (\hat{p}_n(x) - 1)(\hat{p}_n(x) + 1) \leq 0, \quad \forall x \in [c_4,c_3] \cup [c_2,c_1] \\
&\Longrightarrow \hat{p}_n^2(x) \leq 1, \quad \forall x \in [c_4,c_3] \cup [c_2,c_1] \\
&\Longrightarrow |\hat{p}_n(x)| \leq 1, \quad \forall x \in [c_4,c_3] \cup [c_2,c_1] \\
&\qquad \text{et } |\hat{p}_n(x)| \text{ atteint 1 sur ces intervalles au moins en } x \in \{c_4, c_3, c_2, c_1\} \\
&\Longrightarrow | \pm L_n\hat{p}_n(x)| \leq L_n, \quad \forall x \in [c_4,c_3] \cup [c_2,c_1].
\end{aligned}$$

Ainsi, $\pm L_n\hat{p}_n(x)$ est un polynôme dont la déviation dans $[c_4,c_3] \cup [c_2,c_1]$ est égale à la déviation de $\hat{P}_n(x)$ dans ces mêmes intervalles. Ce dernier constat jumelé à l'unicité de $\hat{P}_n(x)$ permet d'affirmer l'égalité en 1., c'est-à-dire

$$\hat{p}_n(x) = \hat{P}_n(x)/ \pm L_n.$$

Pour montrer 2., remarquons d'abord que :

$$\hat{\mathscr{P}}_4(x) > 0, \quad \forall x \in ([c_4,c_3] \cup [c_2,c_1])^c$$

$$\begin{aligned}
&\Longrightarrow (\hat{p}_n(x) - 1)(\hat{p}_n(x) + 1) = \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x) \geq 0, \quad \forall x \in ([c_4,c_3] \cup [c_2,c_1])^c \\
&\Longrightarrow \hat{p}_n^2(x) - 1 = (\hat{p}_n(x) - 1)(\hat{p}_n(x) + 1) \geq 0, \quad \forall x \in ([c_4,c_3] \cup [c_2,c_1])^c \\
&\Longrightarrow \hat{p}_n^2(x) \geq 1, \quad \forall x \in ([c_4,c_3] \cup [c_2,c_1])^c \\
&\Longrightarrow |\hat{p}_n(x)| \geq 1, \quad \forall x \in ([c_4,c_3] \cup [c_2,c_1])^c
\end{aligned}$$

l'égalité ne tenant que pour les zéros de $\hat{q}_{n-2}$ qui sont dans $([c_4,c_3] \cup [c_2,c_1])^c$, donc que pour un nombre fini de points. Ainsi, $\forall x \in ([c_4,c_3] \cup [c_2,c_1])^c$, le point 1. et la dernière implication permettent d'affirmer que $|\hat{P}_n(x)| \geq L_n$, donc que $[c_4,c_3] \cup [c_2,c_1]$ est le sous-ensemble maximal de $\mathbb{R}$ sur lequel $\hat{P}_n$ est à déviation minimale.

($\Longleftarrow$) Supposons maintenant que $\hat{P}_n$ vérifie les points 1. et 2. et montrons que $\hat{p}_n(x)$ vérifie l'équation de Pell.

Le point 2. permet d'affirmer que $\hat{P}_n(x) = \pm L_n$ pour $x \in \{c_4,\ c_3,\ c_2,\ c_1\}$, car autrement ceci signifierait que $|\hat{P}_n| \leq |L_n|$ sur un intervalle plus large, ce qui constitue une contradiction.

Ensuite, le point 1. et le fait que $\hat{P}_n(x) = \pm L_n$ pour $x \in \{c_4,\ c_3,\ c_2,\ c_1\}$ permettent d'affirmer que :

$$\hat{p}_n(x) = \pm 1,\ \text{pour } x \in \{c_4, c_3, c_2, c_1\}$$

$$\begin{aligned}
&\Longrightarrow \hat{p}_n^2(x) = 1,\ \text{pour } x \in \{c_4, c_3, c_2, c_1\}\\
&\Longrightarrow \hat{p}_n^2(x) - 1 = 0,\ \text{pour } x \in \{c_4, c_3, c_2, c_1\}\\
&\Longrightarrow \exists \hat{q}_{n-2}(x) \text{ t.q. } \hat{p}_n^2(x) - 1 = \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x)\\
&\Longrightarrow \exists \hat{q}_{n-2}(x) \text{ t.q. } \hat{p}_n^2(x) - \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x) = 1.
\end{aligned}$$

Ainsi, $\hat{p}_n(x)$ vérifie l'équation de Pell (2).

□

Comment pouvons-nous distinguer la famille de solutions du Problème 6 vérifiant les conditions 1. et 2. de la Proposition 1.2? La proposition qui suit nous montre que cette famille de solutions est fortement similaire à la famille des polynômes de Tchebycheff généralisés.

Proposition 4.1. *Soit $\hat{P}_n$ polynôme monique de degré n à déviation minimale sur $[c_4,c_3] \cup [c_2,c_1] \subset \mathbb{R}$. Alors $\hat{P}_n$ vérifie les points 1. et 2. de la Proposition 1.2 si et seulement s'il y a exactement $n+2$ points de déviation dans $[c_4,c_3] \cup [c_2,c_1]$. Ces $n+2$ points sont $n-2$ points intérieurs $y_i \in (c_4,c_3) \cup (c_2,c_1)$, $i = 1, \ldots, n-2$ et les quatre points frontières c_4, c_3, c_2 et c_1 :*

$$c_4 < y_1 < \ldots < y_i < c_3 < c_2 < y_{i+1} < \ldots < y_{n-2} < c_1.$$

Les ensembles de points d'alternance sont alors soit

$$c_4,\ y_1,\ \ldots,\ y_i,\ c_2,\ y_{i+1},\ \ldots,\ y_{n-2},\ c_1$$

ou

$$c_4,\ y_1,\ \ldots,\ y_i,\ c_3,\ y_{i+1},\ \ldots,\ y_{n-2},\ c_1.$$

Démonstration. ($\Longleftarrow$) En posant :

$$\hat{q}_{n-2}(x) = (x - y_1)\cdots(x - y_i)(x - y_{i+1})\cdots(x - y_{n-2})$$

nous avons que l'Équation de Pell (2) est vérifiée pour $\hat{p}_n = \hat{P}_n / \pm L_n$. Ainsi, $\hat{P}_n$ vérifie les points 1. et 2.

($\Longrightarrow$) Soit $\hat{P}_n$ le polynôme monique de degré n à déviation minimale sur l'union d'intervalle $[c_4,c_3] \cup [c_2,c_1]$ et supposons qu'il vérifie les conditions 1. et 2. de la

Proposition 1.2. Alors $\hat{p}_n = \hat{P}_n / \pm L_n$ est tel qu'il existe $\hat{q}_{n-2}$, un polynôme de degré $n-2$, pour lequel l'Équation (2) est vérifiée d'après la Proposition 1.2.

D'abord, $\hat{P}_n$ est solution du Problème 6 donc il y au moins $n+1$ points d'alternance dans $[c_4,c_3] \cup [c_2,c_1]$. Ensuite, l'Équation (2) se réécrit :

$$\hat{p}_n^2(x) - \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x) = 1 \iff (\hat{p}_n(x)-1)(\hat{p}_n(x)+1) = \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x),$$

alors $|\hat{p}_n(x)| = 1$ en au plus $n+2$ points, soit en c_1, c_2, c_3, c_4 et en des racines de $\hat{q}_{n-2}$. Par conséquent, du point 1. nous avons que $\hat{P}_n$ compte au plus $n+2$ points de déviation dans $[c_4,c_3] \cup [c_2,c_1]$, soit c_1, c_2, c_3, c_4 et les racines de $\hat{q}_{n-2}$ se retrouvant dans $[c_4,c_3] \cup [c_2,c_1]$. Du point 2. nous avons que $|\hat{P}_n(x)| > |L_n|$ en $x \to c_3^+$ et $x \to c_2^-$. Nous distinguons alors trois cas possibles en ce qui concerne le nombre de points de déviation et d'alternance dans $[c_4,c_3] \cup [c_2,c_1]$.

Cas 1 : il y a $n+2$ points de déviation et $n+2$ points d'alternance dans $[c_4,c_3] \cup [c_2,c_1]$. Les points de déviation de $\hat{P}_n$ dans $[c_4,c_3] \cup [c_2,c_1]$ sont les $n-2$ racines de $\hat{q}_{n-2}$ et les quatre points frontières.

Cas 2 : il y a $n+1$ points de déviation et $n+1$ points d'alternance dans $[c_4,c_3] \cup [c_2,c_1]$. Les points de déviation de $\hat{P}_n$ dans $[c_4,c_3] \cup [c_2,c_1]$ sont $n-3$ racines de $\hat{q}_{n-2}$ et les quatre points frontières.

Or pour les cas 1 et 2), c_3 et c_2 étant des points d'alternance et $|\hat{P}_n(x)| > |L_n|$ en $x \to c_3^+$ et $x \to c_2^-$, implique que $|\hat{P}_n(x)| = |L_n|$ en au moins deux points dans $[c_3,c_2]$. Par conséquent, $\hat{q}_{n-2}$ doit posséder au moins deux racines dans $[c_3,c_2]$, ce qui est impossible comme ses $n-2$ racines (resp. n - 3 racines) sont dans $[c_4,c_3] \cup [c_2,c_1]$.

Cas 3 : il y a $n+2$ points de déviation et $n+1$ points d'alternance dans $[c_4,c_3] \cup [c_2,c_1]$. Dans ce cas, seul le cas où $n-2$ points intérieurs et trois des quatre points frontières sont des points d'alternance ne mène pas à une contradiction, d'où l'énoncé.

□

Ainsi, étant donnée une union de deux intervalles $[c_4,c_3] \cup [c_2,c_1]$ et un degré n, il existe un unique polynôme monique $\hat{P}_n$ dont la déviation par rapport à zéro sur $[c_4,c_3] \cup [c_2,c_1]$ est minimale parmi l'ensemble des polynômes moniques de degré n. Toutefois, $\hat{P}_n$ ne vérifie pas nécessairement les conditions 1. et 2. de la Proposition 1.2 pour cette union de deux intervalles. La dernière proposition permet de dire que $\hat{P}_n$ vérifie la Proposition 1.2 si et seulement si $\hat{P}_n$ admet exactement $n+2$ points de déviation dans $[c_4,c_3] \cup [c_2,c_1]$. Nous pouvons donc définir les solutions du Problème 6 vérifiant les conditions 1. et 2. de la Proposition 1.2 à l'aide de la proposition ci-haute, définition qui serait substantiellement la même que celle des polynômes de Tchebycheff généralisés.

Remarque 4.2. À la manière des polynômes de Tchebycheff généralisés, nous pouvons donc imaginer qu'étudier les polynômes à déviation minimale vérifiant les conditions 1. et 2. de la Proposition 1.2 équivaut à poser un critère sur les intervalles $[c_4,c_3] \cup [c_2,c_1]$ à l'étude, critère prenant la forme de formules sur les bornes des intervalles. Nous discuterons de ceci à la Section 5.

4.2 Exemples explicites de solutions au Problème 6

Dans cette Section, le but est de construire une formule explicite pour les polynômes $\hat{P}_n$ vérifiant les conditions 1. et 2. de la Proposition 1.2. Ainsi, soit l'union d'intervalles $[c_4,c_3]\cup[c_2,c_1]$ tel que la solution du Problème 6, notée $\hat{P}_n$, vérifie les conditions 1. et 2. de la Proposition 1.2. Nous chercherons à déterminer la formule de $\hat{p}_n = \hat{P}_n/\pm L_n$ où L_n est la déviation de $\hat{P}_n$ sur $[c_4,c_3]\cup[c_2,c_1]$.

D'après la Proposition 4.1, nous pouvons considérer les points de déviation de $\hat{P}_n$:

$$c_4 < y_1 < \ldots < y_i < c_3 < c_2 < y_{i+1} < \ldots < y_{n-2} < c_1$$

et la suite de points d'alternance

$$c_4,\ y_1,\ \ldots,\ y_i,\ c_3,\ y_{i+1},\ \ldots,\ y_{n-2},\ c_1.$$

Notons ensuite :

- m_0+1, le nombre de points d'alternance sur $[c_4,c_1]$;
- m_1+1, le nombre de points d'alternance sur $[c_4,c_3]$;
- $\tau_1 = m_0 - m_1 - 1$, le nombre de points d'alternance sur (c_2,c_1) ;
- $\tau_2 = m_1 - 1$, le nombre de points d'alternance sur (c_4,c_3).

Notons que $m_0 = n$ et $m_1 < m_0$. D'après la formule (3.6) de l'article [VD18], il s'avère que m_0 et m_1 vérifient l'égalité suivante :

$$n\int_{c_1}^{\infty}\frac{1}{\sqrt{\hat{\mathscr{P}}_4(x)}}dx = m_1\int_{c_3}^{c_2}\frac{1}{\sqrt{\hat{\mathscr{P}}_4(x)}}dx$$

entrainant alors que $(m_0,m_1) = (n,m_1)$ correspond aux « winding numbers » d'une trajectoire de billard elliptique n-périodique dans une ellipse définie à partir de c_3 et de c_2 ou c_1 et de caustique défini par c_3, c_2 et c_1. Ceci nous permet de déduire que m_1 est pair d'après des arguments avancés dans ce même article.

4.2.1 Construction de $\hat{p}_3(x)$

Soit $[c_4,c_3]\cup[c_2,c_1]$ une union d'intervalles pour laquelle le polynôme $\hat{p}_3$ vérifie l'Équation (2). Alors $0 < m_1 < m_0 = 3$ et le fait que m_1 est pair donne que $m_1 = 2$ d'où :

- $m_0 + 1 = 4$;
- $m_1 + 1 = 3$;
- $\tau_1 = m_0 - m_1 - 1 = 0$;
- $\tau_2 = m_1 - 1 = 1$.

D'après les valeurs de τ_1 et τ_2, nous pouvons poser la suite des points de déviation :

$$c_4 < y_1 < c_3 < c_2 < c_1,$$

avec la suite de points d'alternance :

$$c_4,\, y_1,\, c_3,\, c_1.$$

Partant de ces propriétés, nous tentons de déterminer la formule explicite de $\hat{p}_3$. Sans perte de généralité (car $\hat{p}_3 = \hat{P}_3/\pm L_3$), supposons que $\hat{p}_3(c_4) = -1$. Alors, à partir de la suite de points d'alternance, nous trouvons les valeurs : $\hat{p}_3(c_4) = \hat{p}_3(c_3) = \hat{p}_3(c_2) = -1$ et $\hat{p}_3(y_1) = \hat{p}_3(c_1) = 1$. La Figure 2 illustre $\hat{p}_3$ construit à partir de ces égalités.

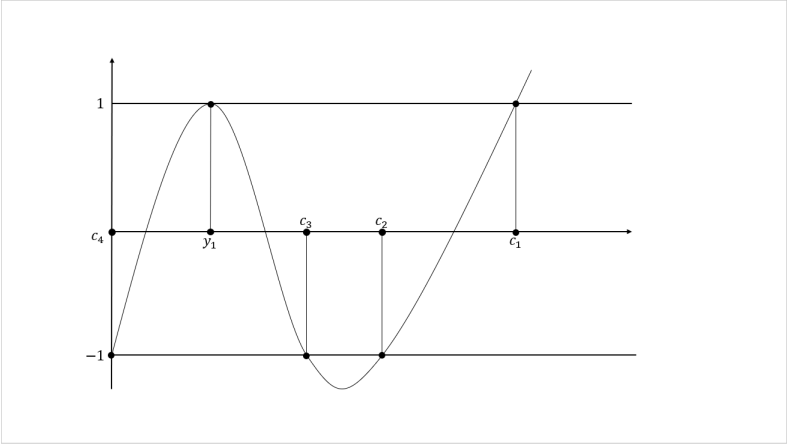


FIGURE 2 : Graphe de $\hat{p}_3(x)$.

Il reste à déterminer la formule de $\hat{p}_3$. Comme c'est un polynôme de degré 3 qui doit vérifier les égalités précédentes, $\hat{p}_3$ prend la forme :

$$\hat{p}_3(x) = kr(x) - 1,$$

où $r(x) = (x-c_4)(x-c_3)(x-c_2)$ et k est tel que $kr(y_1) = 2$ (ce qui est équivalent à $k = \frac{2}{r(y_1)}$). Ainsi,

$$\hat{p}_3(x) = 2\frac{r(x)}{r(y_1)} - 1.$$

Comme y_1 est un extrémum local, pour le déterminer il suffit de déterminer la racine de $r'(x)$ qui se trouve dans $[c_4,c_3]$ (la racine restante se trouvant dans (c_3,c_2)). Donc il suffit de résoudre une équation de degré 2. Pour obtenir l'expression de $\hat{P}_3$ il suffit de diviser $\hat{p}_3$ par son coefficient directeur.

Finalement, remarquons que c_1 doit être la première valeur x suivant c_2 telle que $r(x) = r(y_1)$, donc que c_1 doit être une certaine fonction de c_4, c_3 et c_2. Donc dans le cas $n = 3$, nous pouvons définir la famille d'intervalles de la forme $[c_4,c_3] \cup [c_2,c_1]$ pour laquelle le polynôme $\hat{p}_3$ vérifie l'équation de Pell comme étant l'ensemble des intervalles de la forme $[c_4,c_3] \cup [c_2,c_1]$ où $c_4 < c_3 < c_2 \in \mathbb{R}$ et $c_1 \in \mathbb{R}$ est la première valeur $x > c_2$ tel que $\hat{p}_3(x) = 1$.

4.2.2 Construction de $\hat{p}_4(x)$

Soit $[c_4,c_3] \cup [c_2,c_1]$ une union d'intervalles pour lesquels le polynôme $\hat{p}_4$ vérifie l'équation (2). Alors $0 < m_1 < m_0 = 4$ et m_1 pair donne que $m_1 = 2$ d'où :

- $m_0 + 1 = 5$;
- $m_1 + 1 = 3$;
- $\tau_1 = m_0 - m_1 - 1 = 1$;
- $\tau_2 = m_1 - 1 = 1$.

D'après les valeurs de τ_1 et τ_2, nous pouvons poser la suite des points de déviation :

$$c_4 < y_1 < c_3 < c_2 < y_2 < c_1,$$

avec la suite de points d'alternance :

$$c_4,\ y_1,\ c_3,\ y_2,\ c_1.$$

Partant de ces propriétés, nous tentons de déterminer la formule explicite de $\hat{p}_4$. Sans perte de généralité (car $\hat{p}_4 = \hat{P}_4/\pm L_4$), supposons que $\hat{p}_4(c_4) = -1$. Alors, à partir de la suite de points d'alternance, nous trouvons les valeurs : $\hat{p}_4(c_4) = \hat{p}_4(c_3) = \hat{p}_4(c_2) = \hat{p}_4(c_1) = -1$ et $\hat{p}_4(y_1) = \hat{p}_4(y_2) = 1$. La Figure 3 illustre $\hat{p}_4$ construit à partir de ces égalités.

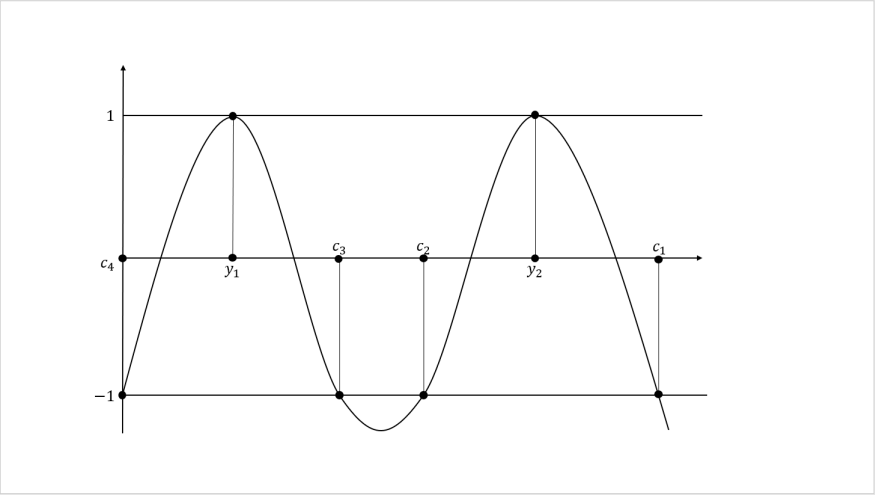


FIGURE 3 : Graphe de $\hat{p}_4(x)$.

Il reste à déterminer la formule de $\hat{p}_4$. Comme c'est un polynôme de degré 4 qui doit vérifier les égalités précédentes, $\hat{p}_4$ prend la forme :

$$\hat{p}_4(x) = kr(x) - 1,$$

où $r(x) = (x - c_4)(x - c_3)(x - c_2)(x - c_1)$ et k est tel que $kr(y_1) = 2$ (ce qui revient à $k = \frac{2}{r(y_1)}$). Ainsi,

$$\hat{p}_4(x) = 2\frac{r(x)}{r(y_1)} - 1.$$

Comme y_1 et y_2 sont des extrémums locaux, pour les déterminer il faut déterminer les racines de $r'(x)$ qui se trouvent dans $[c_4,c_3]$ et $[c_2,c_1]$ respectivement (la racine restante se trouvant dans (c_3,c_2)). Donc il faut ici résoudre une équation de degré 3. Pour obtenir l'expression de $\hat{P}_4$ il suffit de diviser $\hat{p}_4$ par son coefficient directeur.

4.2.3 Construction de $\hat{p}_5(x)$

Soit $[c_4,c_3] \cup [c_2,c_1]$ une union d'intervalles pour laquelle le polynôme $\hat{p}_5$ vérifie l'Équation (2). Alors $0 < m_1 < m_0 = 5$ et m_1 pair donne que $m_1 = 2$ ou $m_1 = 4$. Pour $m_1 = 2$:

- $m_0 + 1 = 6$;
- $m_1 + 1 = 3$;
- $\tau_1 = m_0 - m_1 - 1 = 2$;
- $\tau_2 = m_1 - 1 = 1$.

D'après les valeurs de τ_1 et τ_2, nous pouvons poser la suite de points de déviation :

$$c_4 < y_1 < c_3 < c_2 < y_2 < y_3 < c_1,$$

avec la suite de points d'alternance :

$$c_4,\ y_1,\ c_3,\ y_2,\ y_3,\ c_1.$$

Partant de ces propriétés, nous tentons de déterminer la formule explicite de $\hat{p}_5$. Sans perte de généralité (car $\hat{p}_5 = \hat{P}_5 / \pm L_5$), supposons que $\hat{p}_5(c_4) = -1$. Alors, à partir de la suite de points d'alternance, nous trouvons les valeurs : $\hat{p}_5(c_4) = \hat{p}_5(c_3) = \hat{p}_5(c_2) = \hat{p}_5(y_3) = -1$ et $\hat{p}_5(y_1) = \hat{p}_5(y_2) = \hat{p}_5(c_1) = 1$. La Figure 4 illustre $\hat{p}_4$ construit à partir de ces égalités.

Il reste à déterminer la formule de $\hat{p}_5$ avec $(m_0,m_1) = (5,2)$. Comme $\hat{p}_5$ est un polynôme de degré 5 qui doit vérifier les égalités précédentes, $\hat{p}_5$ prend la forme :

$$\hat{p}_5(x) = kr(x) - 1,$$

où $r(x) = (x - c_4)^h(x - c_3)^i(x - c_2)^j(x - y_3)^l$, avec $\{h,\ i,\ j,\ l\} = \{1,\ 1,\ 1,\ 2\}$ et k est tel que $kr(y_1) = 2$ (ce qui revient à $k = \frac{2}{r(y_1)}$). Ainsi,

$$\hat{p}_5(x) = 2\frac{r(x)}{r(y_1)} - 1.$$

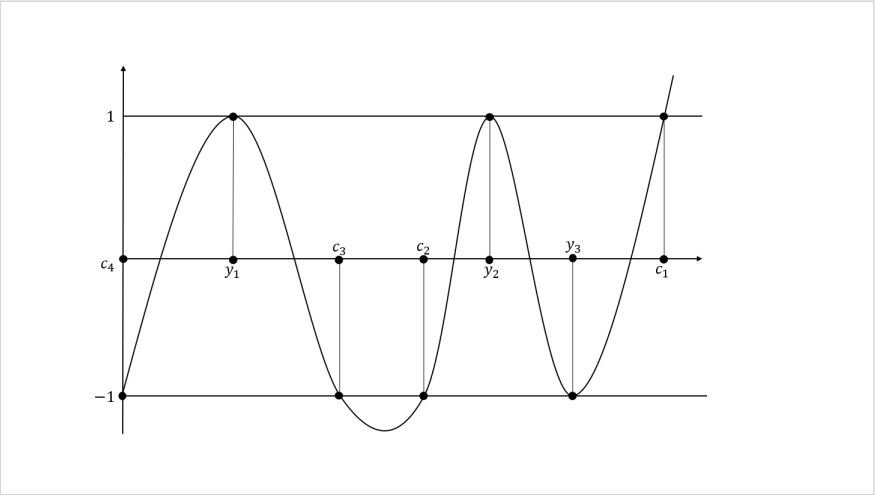


FIGURE 4 : Graphe de $\hat{p}_5(x)$ avec $(m_0,m_1) = (5,2)$.

De l'équation de Pell, $(\hat{p}_5(x)-1)(\hat{p}_5(x)+1) = \hat{\mathscr{P}}_4(x)\hat{q}^2_{n-2}(x)$, il est évident que $h = i = j = 1$ et $l = 2$ dans l'expression de $r(x)$. Comme y_1, y_2 et y_3 sont des extrémums locaux, pour les déterminer il faut déterminer la racine de $r'(x)$ qui se trouve dans $[c_4,c_3]$ et les deux se trouvant dans $[c_2,c_1]$ (la racine restante se trouvant dans (c_3,c_2)). Le degré de $r'(x)$ est égal à 4 et nous voyons ainsi qu'il devient de plus en plus difficile de trouver les racines que nous devons trouver pour notre construction. En réalité, le degré de $r'(x)$ augmente de 1 à chaque fois que nous augmentons de 1 la valeur de n. Également, le fait que y_3, c'est-à-dire un point dans $(c_4,c_3) \cup (c_2,c_1)$, serve à définir $r(x)$ augmentera potentiellement la difficulté dans la recherche des valeurs de y_1, y_2 et y_3. Pour $m_1 = 4$:

- $m_0 + 1 = 6$;
- $m_1 + 1 = 5$;
- $\tau_1 = m_0 - m_1 - 1 = 0$;
- $\tau_2 = m_1 - 1 = 3$.

D'après les valeurs de τ_1 et τ_2, nous pouvons poser la suite de points de déviation :

$$c_4 < y_1 < y_2 < y_3 < c_3 < c_2 < c_1,$$

avec la suite de points d'alternance :

$$c_4,\, y_1,\, y_2,\, y_3,\, c_3,\, c_1.$$

Partant de ces propriétés, nous tentons de déterminer la formule explicite de $\hat{p}_5$. Sans perte de généralité (car $\hat{p}_5 = \hat{P}_5 / \pm L_5$), supposons que $\hat{p}_5(c_4) = -1$. Alors, à partir de la suite de points d'alternance, nous trouvons les valeurs :

$\hat{p}_5(c_4) = \hat{p}_5(c_3) = \hat{p}_5(c_2) = \hat{p}_5(y_2) = -1$ et $\hat{p}_5(y_1) = \hat{p}_5(y_3) = \hat{p}_5(c_1) = 1$. La Figure 1 illustre $\hat{p}_4$ construit à partir de ces égalités.

Il reste à déterminer la formule de $\hat{p}_5$ avec $(m_0,m_1) = (5,4)$. Comme $\hat{p}_5$ est un polynôme de degré 5 qui doit vérifier les égalités précédentes, $\hat{p}_5$ prend la forme :

$$\hat{p}_5(x) = kr(x) - 1,$$

où $r(x) = (x - c_4)(x - c_3)(x - c_2)(x - y_2)^2$ et k est tel que $kr(y_1) = 2$ (ce qui revient à $k = \frac{2}{r(y_1)}$). Ainsi,

$$\hat{p}_5(x) = 2\frac{r(x)}{r(y_1)} - 1.$$

La même problématique pour déterminer y_1, y_2 et y_3 que pour $(m_0, m_1) = (5, 2)$ se pose ici.

5 Conclusion

Ainsi, nous avons une construction pour les polynômes $\hat{p}_n$ avec $n = 3$, $n = 4$ et presque pour $n \geq 5$ (comme nous n'arrivons pas à déterminer y_1, y_2 et y_3 pour ce dernier cas) vérifiant l'équation de Pell :

$$\hat{p}_n^2(x) - \hat{\mathscr{P}}_4(x)\hat{q}_{n-2}^2(x) = 1,$$

où

$$\hat{\mathscr{P}}_4(x) = \prod_{j=1}^{4}(x - c_j).$$

Toutefois, cette construction suppose le fait que l'union d'intervalles $[c_4,c_3] \cup [c_2,c_1]$ posée est connue et est tel que $\hat{p}_n$ vérifie l'Équation (2), or il est difficile de déterminer une telle union d'intervalles. Pour $\hat{p}_3(x)$ il est possible de fixer c_4, c_3, c_2 puis poser c_1 comme étant la première valeur $x > c_2$ telle que $r(x) = r(y_1)$ et nous avons alors que la construction trouvée fonctionne. Par contre, une telle astuce ne fonctionne pas pour $\hat{p}_4$.

Le questionnement à savoir si étant donnée l'union d'intervalles $[c_4,c_3]\cup[c_2,c_1]$ il existe un polynôme à déviation minimale vérifiant les conditions 1. et 2. de la Proposition 1.2 est abordée dans l'article de V. Dragovic et M. Radnovic [VD18]. L'approche utilisée pour déterminer les formules de c_1 en fonction de c_2, c_3 et c_4 pour les différents degrés n est celle du billard elliptique et des courbes elliptiques.

Une fois les formules de c_1 déterminées, une transformation affine liant les polynômes à déviation minimale de degré n sur $[c_4,c_3] \cup [c_2,c_1]$ aux polynômes de Tchebycheff généralisés sur $[-1,a] \cup [b,1]$ est donnée dans ce même article.

Références

[Akh70] N.I. Akhiezer : *Elements of the Theory of Elliptic Functions*. American Mathematical Society, Providence, Rhode Island, 1970.

[Akh92] N.I. Akhiezer : *Theory of Approximation*. Dover publications, INC., New York, New York, 1992.

[Bog05] Andrei Bogatyrev : *Extremal Polynomials and Riemann Surfaces*. Springer, Moscou, Russie, 2005.

[VD18] M. Radnovic V. Dragovic : Caustics of poncelet polygons and classical extremal polynomials. *arXiv :1812.02907v1 [math.DS]*, 2018.

Gabriel Dupuis
Département de mathématiques, Université de Sherbrooke
Courriel: Gabriel.Dupuis2@USherbrooke.ca

CaMUS **7**, 67 – 97

Le treillis des structures partiellement exactes

Souheila Hassoun et Élodie Lapointe

RÉSUMÉ Dans le but d'avancer l'étude du treillis des structures partiellement exactes sur une catégorie additive, on prouve que ce dernier est isomorphe au treillis des sous-bimodules d'un certain bimodule sur l'algèbre d'Auslander de la catégorie.

1 Introduction

Cet article résume le travail des deux auteures durant le stage de recherche effectué par la deuxième sous la supervision de la première pendant l'hiver 2020.

Dans un cadre plus général, les recherches effectuées par le groupe de travail du professeur Thomas Brüstle sont dans le domaine de la théorie des représentations des algèbres.
La théorie des représentations, étudiée dans le livre [ASS06], est une branche des mathématiques qui étudie les structures algébriques abstraites en représentant leurs éléments comme des transformations linéaires d'espaces vectoriels. Cette théorie étudie également les modules sur ces structures algébriques abstraites.
La théorie des représentations est un outil puissant. En effet, elle permet de réduire des problèmes d'algèbre abstraite à des problèmes d'algèbre linéaire, un domaine qui est bien compris.
Dans notre projet, on étudie *le treillis des structures partiellement exactes* introduit dans [BBGH20]. Ce treillis est isomorphe au treillis des sous-bifoncteurs du bifoncteur additif d'extension sur une catégorie additive. Notre but est de simplifier l'étude de ces treillis en proposant un troisième treillis. Celui-ci est le treillis de bimodules sur une algèbre qui est isomorphe aux deux autres. Ce dernier nous offre une nouvelle façon de voir le treillis des structures partiellement exactes d'une catégorie additive et de nouvelles manières pour étudier leurs propriétés.

Cet article est divisé en 6 sections.

Les auteures remercient le professeur Thomas Brüstle, le directeur de recherche de la première auteure au doctorat, pour sa présence et son aide. Les auteures aimeraient aussi remercier Rose-Line Baillargeon et Samuel Lalumière Lavoie pour leur participation dans le projet. Les auteures sont supportées par : Bishop's University et l'Université de Sherbrooke. La première auteure est supportée par la bourse "thésards étoiles" de l'ISM.

Dans la première section, celle dans laquelle on se trouve, on organise notre projet et on énonce les grandes lignes de notre article.

Dans la deuxième section, on aborde la théorie des modules à gauche et à droite sur une algèbre et, plus généralement, sur un anneau. Cela permet de définir les bimodules. On mentionne également plusieurs notions et propriétés qui nous permettront de mieux comprendre ce qu'est un bimodule sur une algèbre.

Dans la troisième section, on définit une catégorie exacte ; une paire formée par une catégorie additive et une structure exacte de Quillen. Cette dernière nous permet de définir ce qu'est une structure partiellement exacte ; une classe de suites exactes courtes formées par des objets de la catégorie additive, satisfaisant quelques axiomes de Quillen.

Dans la quatrième section on discute des sous-bifoncteurs du bifoncteur d'extension Ext en général et des sous-bifoncteurs fermés parmi eux.

Dans la cinquième section, on définit ce qu'est une structure de treillis sur un ensemble partiellement ordonné. On présente aussi le treillis des structures partiellement exactes d'une catégorie additive et le treillis des sous-bifoncteurs de Ext. Par la suite, on construit un ensemble de bimodules admettant une structure de treillis isomorphe aux deux dernières structures mentionnées. On construit explicitement un isomorphisme de treillis. On prend bien soin d'expliquer notre preuve étape par étape. On obtient donc comme résultat trois treillis isomorphes.

Enfin, dans la dernière section, on donne un exemple concret du treillis étudié dans les sections précédentes. On refère au livre [ASS06] pour une meilleure compréhension des notions impliquées dans l'exemple.

Tous les concepts mathématiques décrits dans cette introduction sont définis et décris en détail dans la suite de notre article.

2 Les bimodules

Nous commençons par définir ce qu'est une structure de bimodule sur un ensemble et nous donnons des exemples d'une telle structure algébrique. Deux exemples connus de modules sont donnés par les groupes abéliens et les espaces vectoriels. Chaque groupe abélien est un $\mathbb{Z}-$module, c'est-à-dire un module sur l'anneau des entiers et chaque espace vectoriel est un module sur le corps des scalaires.
Notons que les notions introduites dans cette section sont issues de [Ass97].

2.1 Définitions

Avant de se plonger dans la définition d'un bimodule, on commence par définir quelques concepts importants. Tout d'abord, on rappelle très brièvement ce qu'est un anneau et une algèbre. Par la suite, on définit un module à droite, un module à gauche. On termine par définir les bimodules.

2.1.1 Les anneaux

Contrairement aux groupes qui ne possèdent qu'une opération, les anneaux représentent les ensembles qui possèdent deux opérations : l'addition et la multiplication. Bien que légèrement plus complexe, cette représentation est beaucoup plus générale. En effet, la majorité des ensembles possèdent plus d'une opération. Pensez notamment aux entiers ou aux nombres réels.

Définition 2.1. Un ensemble A admet une structure d'*anneau* s'il est muni de 2 opérations nommées l'addition et la multiplication $(A,+,\cdot)$. Ces opérations doivent respecter les conditions suivantes :

1. $(A,+)$ est un groupe abélien,
2. $(A,\cdot)$ est un monoïde[1],
3. La multiplication est distributive à gauche et à droite sur l'addition : $a(b+c) = ab + ac$ et $(b+c)a = ba + ca$.

2.1.2 Les modules

Lorsqu'on étudie les espaces vectoriels, on choisit les scalaires parmi les éléments d'un corps et ces scalaires agissent sur les éléments de l'ensemble de base d'un module. On a voulu généraliser cette construction, mais en ne limitant pas l'ensemble des scalaires à un corps. Le concept de module a fait son apparition lorsqu'on a voulu que les scalaires appartiennent plutôt à un anneau (voir 2.1). Plus précisément, un module M est un groupe abélien additif muni d'une multiplication à gauche ou à droite par les éléments d'un anneau. Plus formellement, on définit ce concept de la manière suivante.

Définition 2.2. Soit A un anneau, un *A-module* (à gauche) est composé d'un groupe abélien M, dont l'opération est l'addition, ainsi que d'une multiplication à gauche par les éléments de A (opération externe). Il existe donc une application $A \times M \to M$ définit par $(a,x) \mapsto ax$ ($a \in A$ et $x \in M$) telle que les conditions suivantes sont respectées pour tous a, $b \in A$ et pour tous x, $y \in M$:

1. $(ab)x = a(bx)$;
2. $1_A \cdot x = x$;
3. $(a+b)x = ax + bx$;
4. $a(x+y) = ax + ay$.

Note 2.3. La notation utilisée pour spécifier si M est un module à gauche est la suivante : ${}_AM$. Il est également à noter que la définition pour un A-module à droite est similaire à celle énoncée ci-haut. La seule différence est que les éléments de l'anneau A sont placés à droite et non à gauche des éléments de M. On note ce module M_A.

[1] Un monoïde est un ensemble E muni d'une opération $\star$ telle que $\star$ est associative et E possède un élément neutre

Exemple 2.4. Soient A et B deux anneaux. De plus, considérons qu'il existe un morphisme d'anneaux $\varphi : A \to B$.
Dans cet exemple, on suppose que tous les modules sont des modules à droite.

1. Posons $\star$ comme la loi externe de A sur M et $*$ celle de B sur M.
2. On sait que pour avoir M_A, on a besoin d'une opération externe bien définie entre les éléments de A et de M. On a vu que, en général, pour avoir un module, il doit exister une application telle que $M \times A \to M$ définie par $(x, a) \mapsto x \star a$ ($a \in A$ et $x \in M$). Dans le cas de cet exemple, on connait seulement l'existence d'un module M_B et d'un morphisme φ. On doit donc utiliser ces informations pour redéfinir l'application $(x, a) \mapsto x \star a$. On prendra $x \star a = x * \varphi(a)$. On obtiendra alors une application bien définie puisque le module M_B existe. Cela implique que $x * \varphi(a)$ existe et est bien contenu dans M.

Remarque 2.5.

1. Dans l'exemple précédent, on dit que le B-module de M induit la structure de A-module par *changement des scalaires*.
2. Il existe deux cas particuliers à l'exemple précédent.
   (a) Si $A \subseteq B$ et φ est l'inclusion.
   (b) Si $B = A/I$ avec $I \trianglelefteq A$ un idéal bilatère de A et $\varphi : A \to A/I$ la projection canonique, alors on définira $x \star a = x * \varphi(a) = x * (a + I)$.

2.1.3 Les sous-modules

Définition 2.6. Soit N un ensemble qui contient au moins un élément. Alors N est un *sous-module* si $\forall a, b \in A$ et $\forall x, y \in N$, on a $ax + by \in N$.

Définition 2.7. Soient A un anneau et M un A-module. On dénote par $\mathcal{S}(M)$ l'ensemble des sous-modules de M ordonné par l'inclusion ordinaire des ensembles.

Exemple 2.8. Choisissons, pour cet exemple, une famille de sous-modules de M suivante $(M_\lambda)_{\lambda \in \Lambda}$. Il est intéressant de noter que dans cette situation, l'intersection de tous les sous-modules $\bigcap_{\lambda \in \Lambda} M_\lambda$ est également un sous-module de M. De plus, ce sous-module est le plus grand de l'ensemble M qui soit contenu dans chaque M_λ.
Essayons de trouver un autre sous-module à partir de cette famille. Prenons, la somme $\sum_{\lambda \in \Lambda} M_\lambda$ définit comme l'ensemble des sommes qui s'écrivent de la manière suivante, $\sum_{\lambda \in \Lambda} x_\lambda$ avec $x_\lambda \in M_\lambda$ $\forall \lambda \in \Lambda$. De plus, dans cette définition, $(x_\lambda)_{\lambda \in \Lambda}$ doit être une famille d'éléments de M à support fini. Si ces conditions sont respectées, alors on peut facilement vérifier que $\sum_{\lambda \in \Lambda} M_\lambda$ est aussi un sous-module de M. De plus, celui-ci est le plus petit sous-module de M qui contient tous les M_λ.

2.1.4 Les algèbres

Définition 2.9. Soit K un corps. Une *K-algèbre* est un ensemble qui est à la fois un K-module et un anneau A. De plus, ces deux structures doivent être compatibles, c'est-à-dire $\forall a,b \in A$ *et* $\alpha \in K$ la condition suivante est respectée :

$$(ab)\alpha = a(b\alpha) = (a\alpha)b.$$

Définition 2.10. Soit A une algèbre. On appelle *algèbre opposée*, notée A^{op}, une algèbre qui possède la même structure de K-module que A. Cependant, la multiplication $*$ de A^{op} est définie par $a * b = ba$, $\forall a, b \in A$.

Remarque 2.11.

1. L'algèbre A est commutative $\Leftrightarrow A = A^{op}$.
2. De la même manière qu'on a défini un A-module sur un anneau en général 2.2, on peut considérer les A-modules sur une algèbre.

Définition 2.12. [ASS06, A.2(2.10)] Soit A une K-algèbre de représentation finie, c'est-à-dire telle qu'il n'existe, à isomorphisme près, qu'un nombre fini de A-module indécomposables, et soient $X_1, \ldots, X_n$ l'ensemble des modules indécomposables[2] sur A. On définit *l'algèbre d'Auslander de* A la K-algèbre de dimension finie donnée par :

$$B = End(\overset{n}{\underset{j=1}{\oplus}} X_j).$$

2.1.5 Les bimodules

Maintenant que le concept de module a été introduit, il est plus simple de comprendre les bimodules. En effet, un bimodule comme le suggère son nom est constitué de deux structures de modules sur le même ensemble sous-jacent, une structure de module à droite et une structure de module à gauche. Cependant, ces deux structures doivent satisfaire une propriété qui permet d'assurer que ces structures de modules à gauche et à droite sont compatibles.

Définition 2.13. Soit A et B deux anneaux et M un groupe abélien. Un *(A-B) bimodule* (aussi noté ${}_AM_B$) est un ensemble M qui est à la fois un A-module à gauche, ${}_AM$, et un B-module à droite, M_B, et ces deux structures sont compatibles, c'est-à-dire que la condition suivante est satisfaite :

$$a \star (x * b) = (a \star x) * b,$$

avec $a \in A$, $b \in B$ et $x \in M$ où $\star$ est la loi externe de A sur M et $*$ celle de B sur M.

Exemple 2.14. Soit A un anneau et $I \trianglelefteq A$ alors I a une structure ${}_AI_A$.

[2] c'est-à-dire si $X_j = S \oplus S'$ alors $S = 0$ ou $S' = 0$

Remarque 2.15. Si l'anneau A est commutatif, la distinction entre un module à droite M_A et un module à gauche ${}_AM$ n'est qu'une question d'écriture. Dans ce cas, un tel module M peut être considéré comme un bimodule ${}_AM_A$ dont les deux lois externes sont identiques à la loi du A-module donné.

Démonstration. Soit A un anneau commutatif. Supposons que les deux opérations externes sont définies de la manière suivante avec (a, $b \in A$ et $x \in M$) :

$$\star : A \times M \to M; \qquad (a, x) \mapsto a \star x$$

$$* : M \times A \to M; \qquad (x, b) \mapsto x * b = b \star x.$$

On a alors que :

$$a \star (x * b) = a \star (b \star x) = (ab) \star x = (ba) \star x = b \star (a \star x) = (a \star x) * b.$$

On voit donc que lorsque A est commutatif, si M est un A-module, M peut aussi être considéré comme un bimodule ${}_AM_A$.

□

Définition 2.16. Soient M et N deux (A-B) bimodules avec $\diamond$ comme la loi externe de A sur M et $*$ celle de B sur M. Si une application $f : M \to N$ respecte les conditions suivantes :

1. $f(x + y) = f(x) + f(y)$, $\forall x, y \in M$;
2. $f(a \diamond x * b) = a \diamond f(x) * b$, $\forall x \in M$, $a \in A$ et $b \in B$;

alors on dit que f est un *morphisme* de (A-B) bimodules (ou application A-B linéaire)

On peut aussi réécrire ces deux conditions en la condition suivante :

$$f(a \diamond x * b + a' \diamond y * b') = a \diamond f(x) * b + a' \diamond f(x) * b'.$$

Remarque 2.17. Il est a noter qu'un morphisme de modules ${}_AM$ (ou application A-linéaire) respecte la condition 1 de 2.16 ainsi que la condition $f(a \diamond x) = a \diamond f(x)$, $\forall x \in M$, $a \in A$.

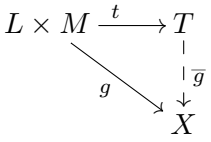
Définition 2.18. Soit A une K-algèbre et L_A ainsi que ${}_AM$ deux modules. Une application $g : L \times M \to X$, avec $L \times M$ et X des K-modules, est dite A-*bilinéaire* si les conditions suivantes sont respectées.

1. $g(x_1\alpha_1 + x_2\alpha_2, y) = g(x_1, y)\alpha_1 + g(x_2, y)\alpha_2$;
2. $g(x, y_1\beta_1 + y_2\beta_2) = g(x, y_1)\beta_1 + g(x, y_2)\beta_2$;
3. $g(xa, y) = g(x, ay)$,

$\forall x, x_1, x_2 \in L$, $y, y_1, y_2 \in M$, $\alpha_1, \alpha_2, \beta_1, \beta_2 \in K$ et $a \in A$.

Définition 2.19. Un *produit tensoriel* de L_A et ${}_AM$ est défini par la donnée d'une paire (T,t), où T est un K-module et t est une application A-bilinéaire de $L \times M$ vers T, telle que, pour toute paire (X,g), avec X un K-module et g une application A-bilinéaire de $L \times M$ vers X, il existe, à isomorphisme près, une unique application K-linéaire $\overline{g} : T \to X$ telle que $\overline{g}t = g$.

Le diagramme ci-dessous illustre les applications de la Définition 2.19.



On notera $T = L \otimes_A M$ le produit tensoriel de L et M.

Proposition 2.20. *Soient A et B deux K-algèbres. Tout bimodule ${}_AM_B$ peut être vu comme un $A \otimes B^{op}$-module $M_{A\otimes B^{op}}$.*

Démonstration. On sait que le produit tensoriel de deux algèbres est une algèbre satisfaisant les propriétés suivantes pour tous $a \otimes b$ avec $a \in A$ et $b \in B$:

i. $(\lambda a) \otimes b = a \otimes (\lambda b) = \lambda(a \otimes b)$;

ii. $(a + a') \otimes b = a \otimes b + a' \otimes b$.

À partir de ces propriétés, on peut montrer que tout (A-B) bimodule ${}_AM_B$ donne lieu à un $A \otimes B^{op}$-module à partir du même groupe abélien M et de la multiplication définie par $m \cdot (a \otimes b) := amb$. Inversement, un $A \otimes B^{op}$-module N permet de créer un (A-B) bimodule ${}_AN_B$ muni de la multiplication externe suivante $anb := n \cdot (a \otimes b)$. Donc, un bimodule ${}_AM_B$ peut être vu comme un $A \otimes B^{op}$-module $M_{A\otimes B^{op}}$.

Remarque 2.21. On dénote $End(M_A)$ l'ensemble des applications A-linéaires de M_A vers lui-même. Posons $B = End(M_A)$, cet ensemble forme une algèbre, pour la composition usuelle d'applications et l'addition d'applications linéaires. Il est intéressant de noter que tout A-module M_A a également une structure naturelle de (B-B) bimodule. Voir [Ass97] pour plus de détails.

□

3 Les catégories partiellement exactes

Dans cette section, on rappelle les notions de base de la théorie des catégories. Le but est de définir les catégories exactes et partiellement exactes.

3.1 La théorie des catégories

Définition 3.1. Une *catégorie* $\mathcal{C}$ est définie par la donnée de :

1. Une classe d'objets $\mathcal{C}_0$ aussi notée $Ob(\mathcal{C})$.

2. Une classe de flèches (de morphismes) , définie par la donnée pour chaque paire d'objets (X,Y) d'un ensemble $Hom_{\mathcal{C}}(X,Y)$.
   Notant que si $(X,Y) \neq (X',Y')$, alors $Hom_{\mathcal{C}}(X,Y) \cap Hom_{\mathcal{C}}(X',Y') = \emptyset$.

3. Une application bien définie pour chaque triplet d'objets (X,Y,Z) de $\mathcal{C}$, appelée *la composition* des morphismes :

$$\circ : Hom_{\mathcal{C}}(X,Y) \times Hom_{\mathcal{C}}(Y,Z) \to Hom_{\mathcal{C}}(X,Z)$$

$$(f,g) \longmapsto g \circ f.$$

   Cette composition doit également respecter les deux conditions suivantes :

   (a) L'associativité de la composition : si on a $f \in Hom_{\mathcal{C}}(U,V)$, $g \in Hom_{\mathcal{C}}(V,W)$, $h \in Hom_{\mathcal{C}}(W,X)$, alors $h \circ (g \circ f) = (h \circ g) \circ f$.

   (b) Pour chaque objet X de $\mathcal{C}$, il existe un morphisme appelé *l'identité sur* X noté :

$$1_X \in Hom_{\mathcal{C}}(X,X).$$

   Et ce morphisme joue le rôle de l'unité à gauche et à droite, c'est-à-dire que si $f \in Hom_{\mathcal{C}}(X,Y)$ et $g \in Hom_{\mathcal{C}}(W,X)$, alors $f \circ 1_X = f$ et $1_X \circ g = g$.

Exemple 3.2. Les bimodules (voir 2.13) sur une algèbre A (voir 2.9) forment une catégorie dont les objets sont les bimodules et les flèches sont les morphismes de bimodules (voir 2.16).

Remarque 3.3. La correspondance introduite en 2.20 définit alors des équivalences inverses entre les catégories de $(A\text{-}B)$ bimodules et $A \otimes B^{op}$-modules.

Proposition 3.4. *Soit M et N deux A-modules. L'ensemble des applications A-linéaires de M dans N se note $Hom_{Mod(A)}(M,N)$. On définit sur cet ensemble les opérations suivantes :*

i. La somme de $f,g : M \to N$ est définie par $(f+g)(x) = f(x) + g(x)$ avec $x \in M$;

ii. Le produit de $f : M \to N$ par $\alpha \in K$ est défini par $(f\alpha)(x) = \alpha f(x)$ avec $x \in M$.

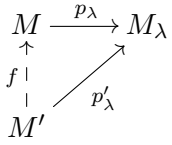
Comme ces opérations satisfont les axiomes de la Définition 2.2, elles définissent une structure de K-module sur l'ensemble $Hom_{ModA}(M,N)$.

3.2 Les catégories additives

Tout d'abord, on va définir quelques concepts comme le produit et la somme directe pour une catégorie. Ensuite, on va énoncer ce qu'est une catégorie linéaire. Ces notions sont essentielles pour comprendre la définition d'une catégorie additive.

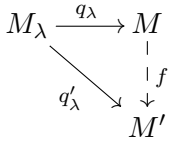
3.2.1 Définitions et concepts importants

Définition 3.5 ([Ass97])**.** Soit $(M_\lambda)_{\lambda\in\Lambda}$ une famille d'objets d'une catégorie $\mathcal{C}$. Un *produit* de $(M_\lambda)_{\lambda\in\Lambda}$ est défini lorsqu'on a une paire $(M,(p_\lambda)_{\lambda\in\Lambda})$ donnée. De plus, si on a une autre paire $(M',(p'_\lambda)_{\lambda\in\Lambda})$ donnée alors il existera un unique morphisme $f:\ M'\to M$. En outre, ce morphisme est tel que $p_\lambda\circ f = p'_\lambda,\ \forall\ \lambda\in\Lambda$



Dans un tel cas, on dit que M est un objet *universel*. On peut également dire que celui-ci est universellement *attirant*. On note généralement le produit par $\prod\limits_{\lambda\in\Lambda} M_\lambda$.

Définition 3.6 ([Ass97])**.** Soit $(M_\lambda)_{\lambda\in\Lambda}$ une famille d'objets d'une catégorie $\mathcal{C}$. Une *somme directe*, ou coproduit, de $(M_\lambda)_{\lambda\in\Lambda}$ est définie lorsqu'on a une paire $(M,(q_\lambda)_{\lambda\in\Lambda})$ donnée. De plus, si on a une autre paire $(M',(q'_\lambda)_{\lambda\in\Lambda})$ donnée alors il existe un unique morphisme $f:\ M\to M'$. En outre, ce morphisme est tel que $f\circ q_\lambda = q'_\lambda,\ \forall\ \lambda\in\Lambda$



Dans un tel cas, on dit aussi que M est un objet *universel*. Cependant, on dit plutôt que celui-ci est universellement *repoussant*. On note généralement l'opération de la somme directe par $M = \bigoplus\limits_{\lambda\in\Lambda} M_\lambda$ ou encore $\coprod\limits_{\lambda\in\Lambda} M_\lambda$.

Définition 3.7. Soit K un anneau commutatif. Une catégorie $\mathcal{C}$ est **additive** si :

1. Pour toute paire (X,Y), avec X, Y des objets de $\mathcal{C}$, l'ensemble $Hom_C(X,Y)$ est un groupe abélien.

2. La composition de $\mathcal{C}$ respecte les conditions suivantes :

   (a) $g\circ(f_1\alpha_1 + f_2\alpha_2) = (g\circ f_1)\alpha_1 + (g\circ f_2)\alpha_2$

   (b) $(g_1\beta_1 + g_2\beta_2)\circ f = (g_1\circ f)\beta_1 + (g_2\circ f)\beta_2$

avec f, f_1, $f_2 : X \to Y$, g, g_1, $g_2 : Y \to Z$, et α_1, α_2, β_1, $\beta_2 \in K$

3. Toute famille finie d'objets de $\mathcal{C}$ admet un produit ($\prod$) et une somme directe ($\bigoplus$) dans $\mathcal{C}$.

Voici quelques exemples comparant les notions de produit et de co-produit, qui ne coïncident pas toujours nécessairement :

Exemple 3.8. On considère la catégorie des groupes $\mathcal{G}r$ où le produit est donné par le produit cartésien des groupes tandis que la somme directe (co-produit) est le produit libre de groupes.

Exemple 3.9. On considère la catégorie des $\mathbb{R}-$espaces vectoriels où la somme d'une famille infinie de copies de $\mathbb{R}$ est isomorphe à l'espace vectoriel $\mathbb{R}[x]$, qui est un espace vectoriel de dimension infinie, mais où chaque élément lui-même n,a qu'un nombre fini de coefficients non nuls, tandis que leur produit est isomorphe aux suites de nombre réels $\mathbb{R}^{\mathbb{N}}$.

Définition 3.10. Soit K un anneau commutatif. Une catégorie $\mathcal{C}$ est *additive* si :

1. Pour toute paire (X,Y), avec X, Y des objets de $\mathcal{C}$, l'ensemble $Hom_C(X,Y)$ est un groupe abélien.

2. La composition de $\mathcal{C}$ respecte les conditions suivantes :

   (a) $g \circ (f_1\alpha_1 + f_2\alpha_2) = (g \circ f_1)\alpha_1 + (g \circ f_2)\alpha_2$;

   (b) $(g_1\beta_1 + g_2\beta_2) \circ f = (g_1 \circ f)\beta_1 + (g_2 \circ f)\beta_2$;

   avec f, f_1, $f_2 : X \to Y$, g, g_1, $g_2 : Y \to Z$, et α_1, α_2, β_1, $\beta_2 \in K$.

3. Toute famille finie d'objets de $\mathcal{C}$ admet un produit ($\prod$) et une somme directe ($\bigoplus$) dans $\mathcal{C}$.

On a terminé la définition des concepts importants à la compréhension d'une catégorie K-linéaire. Il est donc possible d'énoncer la signification de cette notion.

Définition 3.11. Une catégorie **K-linéaire** est une petite catégorie additive dont le groupe abélien $Hom_C(X,Y)$ est un K-module pour toute paire d'objets. Une catégorie additive est donc une catégorie $\mathbb{Z}-$linéaire.

Suivant 2.12 on introduit :

Définition 3.12. Soit $\mathcal{A}$ une catégorie additive *Krull-Schmidt* de représentation finie, c'est-à-dire que chaque objet se décompose en somme directe finie d'objets indécomposables d'une manière unique à permutation près. De plus, soient X_1,...,X_n les représentants des classes d'isomorphismes d'objets indécomposables de $\mathcal{A}$ et soit $X = \bigoplus_{j=1}^{n} X_j$.
On définit **l'algèbre d'Auslander** de $\mathcal{A}$ par : $B = \mathrm{End}\,(X)$.

3.3 Les suites exactes

Définition 3.13. Soit une suite de A-modules et d'applications A-linéaires,

$$... \longrightarrow M_{i+1} \xrightarrow{f_{i+1}} M_i \xrightarrow{f_i} M_{i-1} \xrightarrow{f_{i-1}} ...$$

celle-ci est dite *exacte en* M_i si $Im(f_{i+1}) = Ker(f_i)$. Elle est dite *exacte* si elle l'est en tout M_i.

Définition 3.14. Une *suite exacte courte* ou s.e.c pour simplifier les notations, est une suite exacte dans le sens de la Définition 3.13, c'est-à-dire que f est un monomorphisme, g est un épimorphisme et $Im(f) = Ker(g)$:

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{g} P \longrightarrow 0.$$

Note 3.15.

1. Tout morphisme surjectif de modules $g : N \to P$ induit une suite exacte courte suivante, avec j l'inclusion :

$$0 \longrightarrow Ker\ g \xhookrightarrow{j} N \xrightarrow{g} P \longrightarrow 0.$$

2. Tout morphisme injectif de modules $f : M \to N$ induit une suite exacte courte suivante, avec p la projection :

$$0 \longrightarrow M \xrightarrow{f} N \xrightarrow{p} Coker\ f \longrightarrow 0.$$

Définition 3.16. Deux suites exactes courtes sont dites *équivalentes* s'il existe un morphisme $\beta : N \to N'$ qui rend le diagramme suivant commutatif :

$$\begin{array}{ccccccccc} 0 & \longrightarrow & M & \xrightarrow{f} & N & \xrightarrow{g} & P & \longrightarrow & 0 \\ & & \| & & \downarrow \beta & & \| & & \\ 0 & \longrightarrow & M & \xrightarrow{f'} & N' & \xrightarrow{g'} & P & \longrightarrow & 0 \end{array}$$

Définition 3.17. Une suite exacte courte est dite *scindée* s'il existe un isomorphisme $h : N \to M \oplus P$ qui rend le diagramme suivant commutatif :

$$\begin{array}{ccccccccc} 0 & \longrightarrow & M & \xrightarrow{f} & N & \xrightarrow{g} & P & \longrightarrow & 0 \\ & & \| & & \downarrow h & & \| & & \\ 0 & \longrightarrow & M & \xrightarrow{q_1} & M \oplus P & \xrightarrow{p_2} & P & \longrightarrow & 0 \end{array}$$

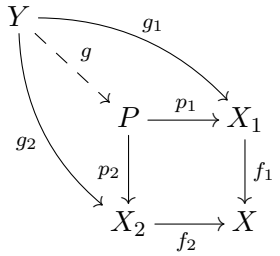
En utilisant [Büh10, Corollaire 3.2], il suffit qu'il existe n'importe quel morphisme h qui fait commuter le diagramme.

3.3.1 Produits fibrés et sommes amalgamées

Nous définirons ici le produit fibré et la somme amalgamée sur $\mathcal{C}$, lorsque $\mathcal{C}$ désigne une catégorie additive.

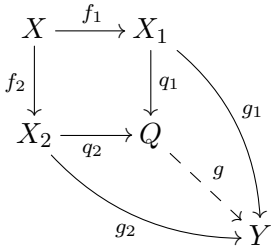
Définition 3.18. Soit $f_1 : X_1 \to X$, $f_2 : X_2 \to X$ deux morphismes de $\mathcal{C}$. Un *produit fibré* (ou *pull back* en anglais) de f_1 et f_2 est la donnée d'un objet P et de deux morphismes $p_1 : P \to X_1$, $p_2 : P \to X_2$ tels que :

1. $f_1 p_1 = f_2 p_2$,
2. pour tout objet Y et toute paire de morphismes $g_1 : Y \to X_1$, $g_2 : Y \to X_2$ tels que $f_1 g_1 = f_2 g_2$, il existe un unique morphisme $g : Y \to P$ tel que $p_1 g = g_1$ et $p_2 g = g_2$.



Définition 3.19. Soit $f_1 : X \to X_1$, $f_2 : X \to X_2$ deux morphismes de $\mathcal{C}$. Une *somme amalgamée* (ou *push out* en anglais) de f_1 et f_2 est la donnée d'un objet Q et de deux morphismes $q_1 : X_1 \to Q$, $q_2 : X_2 \to Q$ tels que :

1. $q_1 f_1 = q_2 f_2$,
2. pour tout objet Y et toute paire de morphismes $g_1 : X_1 \to Y$, $g_2 : X_2 \to Y$ tels que $g_1 f_1 = g_2 f_2$, il existe un unique morphisme $g : Q \to Y$ tel que $g q_1 = g_1$ et $g q_2 = g_2$.



Exemple 3.20. Le produit fibré constitue un sous-ensemble du produit cartésien et dans des catégories ensemblistes, telles celles des espaces topologiques ou des espaces vectoriels, le produit fibré constitue lui-même un objet de la catégorie. Voir [Ass97] pour plus d'exemples.

3.4 Les structures exactes de Quillen

3.4.1 Définition et propriétés de base

Ici, nous rappellerons suivant [Büh10], la définition d'une catégorie exacte au sens de Quillen [Qui73] et nous donnerons quelques exemples.

Définition 3.21. Soit $\mathcal{A}$ une catégorie additive. Une paire noyau-conoyau (i, d) dans $\mathcal{A}$ est une paire de morphismes composables tels que i est le noyau de d et d est le conoyau de i. Si une classe $\mathcal{E}$ de paires de noyau-conoyau sur $\mathcal{A}$ est fixée, un *monomorphisme admissible* est un morphisme i pour lequel il existe un morphisme d tel que $(i, d) \in \mathcal{E}$. On écrira :

$$A \overset{i}{\rightarrowtail} B \ .$$

On définit un *épimorphisme admissible* de façon duale à la Définition 3.21. De plus, on écrira :

$$B \overset{d}{\twoheadrightarrow} C \ .$$

Définition 3.22. Une *structure exacte* $\mathcal{E}$ sur $\mathcal{A}$ est une classe de paires de noyau-conoyau (i, d) dans $\mathcal{A}$, fermée à isomorphismes près et satisfaisant aux axiomes suivants :

(E0) Pour tout objet A dans $\mathcal{A}$, l'identité 1_A est un monomorphisme admissible.

(E0)^{op} Pour tout objet A dans $\mathcal{A}$, l'identité 1_A est un épimorphisme admissible.

(E1) La classe des monomorphismes admissibles est fermée sur la composition.

(E1)^{op} La classe des épimorphismes admissibles est fermée sur la composition.

(E2) La somme amalgamée d'un monomorphisme admissible $i : A \rightarrowtail B$ et d'un morphisme quelconque $t : A \longrightarrow C$ existe et s_C est un monomorphisme admissible :

$$\begin{array}{ccc} A & \overset{i}{\rightarrowtail} & B \\ {\scriptstyle t}\downarrow & \text{SA} & \downarrow{\scriptstyle s_B} \\ C & \underset{s_C}{\rightarrowtail} & S \end{array}$$

(E2)^{op} Le produit fibré d'un épimorphisme admissible $h : A \twoheadrightarrow C$ et d'un morphisme quelconque $t : B \longrightarrow C$ existe et p_B est un épimorphisme admissible :

$$\begin{array}{ccc} P & \overset{p_B}{\twoheadrightarrow} & B \\ {\scriptstyle P_A}\downarrow & \text{PF} & \downarrow{\scriptstyle t} \\ A & \underset{h}{\twoheadrightarrow} & C. \end{array}$$

On dénote par $Ex(A)$ l'ensemble (jamais vide) formé par toutes les structures exactes $\mathcal{E}$ sur $\mathcal{A}$.

Définition 3.23. Une *catégorie exacte* est une paire $(\mathcal{A}, \mathcal{E})$ constituée d'une catégorie additive $\mathcal{A}$ et d'une structure exacte $\mathcal{E}$ sur $\mathcal{A}$, c'est-à-dire une classe de suites exactes courtes satisfaisant les axiomes $(E0)$, $(E0)^{op}$, $(E1)$, $(E1)^{op}$,$(E2)$ et $(E2)^{op}$ de la Définition 3.22.

Remarque 3.24. Il est intéressant de savoir que $\mathcal{E}$ est une structure exacte sur $\mathcal{A}$ si et seulement si $\mathcal{E}^{op}$ est une structure exacte sur $\mathcal{A}^{op}$, alors pour toute catégorie exacte $(\mathcal{A}, \mathcal{E})$ son opposée $(\mathcal{A}^{op}, \mathcal{E}^{op})$ est aussi exacte.

Définition 3.25. [BBGH20, 4.10] Soit $\mathcal{A}$ une catégorie additive. Une *structure partiellement exacte* $\mathcal{W}$ sur $\mathcal{A}$ est une classe de paires noyau-conoyau (i, d) dans $\mathcal{A}$, fermée sous les isomorphismes, fermée pour la somme directe et satisfaisant les axiomes $(E0)$, $(E0)^{op}$,$(E2)$ et $(E2)^{op}$ de la Définition 3.22. On dénote par $Wex(A)$ l'ensemble formé par toutes les structures partiellement exactes $\mathcal{W}$ sur $\mathcal{A}$ qui sont inclueses dans $\mathcal{E}_{max}$ du théorème 3.26.

3.4.2 La structure exacte minimale

Il est bien connu que toute catégorie additive admet une structure exacte *minimale* pour l'inclusion des classes.

Définition 3.26. La *structure exacte minimale* est la classe de toutes les suites exactes courtes scindées (voir 3.17), on la note $\mathcal{E}_{min}$. Celle-ci forme une structure exacte sur toute catégorie additive $\mathcal{A}$.

Proposition 3.27. *[Büh10, exemple 13.1] Pour toute catégorie additive $\mathcal{A}$ les suites isomorphes à une suite de la forme :*

$$A \overset{f}{\rightarrowtail} A \oplus B \overset{g}{\twoheadrightarrow} B$$

forment une structure exacte $\mathcal{E}_{min}$, appelée la structure exacte scindée.

En effet, toute structure exacte sur $\mathcal{A}$ contient les suites exactes courtes scindées [Büh10, Lemma 2.7], ce qui fait que $\mathcal{E}_{min}$ est la borne inférieure du treillis $Ex(\mathcal{A})$ formé par toutes les structures exactes sur $\mathcal{A}$.

3.4.3 La structure exacte maximale

Théorème 3.28. *[Rum11, Corollaire 2] Toute catégorie additive admet une unique structure exacte maximale $\mathcal{E}_{max}$.*

4 Les sous-bifoncteurs fermés de Ext^1

4.1 Définitions et exemples

Définition 4.1. Soient $\mathcal{C}, \mathcal{D}$ deux catégories.

1. Un *foncteur covariant* $F : \mathcal{C} \to \mathcal{D}$ est défini par la donnée pour chaque objet $X \in \mathcal{C}$ d'un objet $FX \in \mathcal{D}$ et pour chaque morphisme $(f : X \to Y) \in \mathcal{C}$ d'un morphisme $(F(f) : FX \to FY) \in \mathcal{D}$ de sorte que

   (a) Si $g \circ f$ est bien définie dans $\mathcal{C}$, alors $F(g) \circ F(f)$ est bien définie dans $\mathcal{D}$ et $F(g \circ f) = F(g) \circ F(f)$.

   (b) $\forall$ objet $X \in \mathcal{C}$, $F(1_X) = 1_{FX}$.

   De façon plus visuelle :

$$\begin{array}{ccc} X & \longmapsto & FX \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle F(f)} \\ Y & \longmapsto & FY \end{array}$$

2. Un *foncteur contravariant* $F : \mathcal{C} \to \mathcal{D}$ est défini par la donnée pour chaque objet $X \in \mathcal{C}$ d'un objet $FX \in \mathcal{D}$ et pour chaque morphisme $(f : X \to Y) \in \mathcal{C}$ d'un morphisme $(F(f) : FY \to FX) \in \mathcal{D}$ de sorte que

(a) Si $g \circ f$ est bien définie dans $\mathcal{C}$, alors $F(f) \circ F(g)$ est bien définie dans $\mathcal{D}$ et $F(g \circ f) = F(f) \circ F(g)$.

(b) $\forall$ objet $X \in \mathcal{C}$, $F(1_X) = 1_{FX}$.

De façon plus visuelle :

$$\begin{array}{ccc} X & \longmapsto & FX \\ {\scriptstyle f}\downarrow & & \uparrow{\scriptstyle \boldsymbol{F(f)}} \\ Y & \longmapsto & FY \end{array}$$

Définition 4.2. Un *bifoncteur* est un foncteur à deux variables. Il est contravariant dans la première variable et covariant dans la seconde.

Exemple 4.3. Le foncteur d'abélianisation $A : \mathbb{G}r \to \mathcal{A}b$ qui à chaque groupe G associe son groupe abélianisé $G^{ab} = G/G'$, où son groupe dérivé

$$G' = \{aba^{-1}b^{-1} | a,b \in G\}.$$

Exemple 4.4. Le foncteur covariant L : **Ens** $\to Vect_K$ qui envoie chaque ensemble vers son espace vectoriel (ou module) libre.

Vu que les sous-bifoncteurs de Ext^1 sont considérés en profondeur dans la sous-section prochaine, voici une exemple accessible illustrant la notion d'un sous-foncteur en général.

Exemple 4.5. On considère le foncteur identité sur la catégorie des groupes abéliens $I : \mathcal{A}b \to \mathcal{A}b$. Le foncteur envoyant un groupe abélien vers son sous-groupe de torsion forme un sous-foncteur de I.

Définition 4.6. Soient $\mathcal{C}$ une catégorie et X un objet de $\mathcal{C}$. Le bifoncteur $Hom_{\mathcal{C}}$**(-,-)** associe à chaque paire d'objets de $\mathcal{C}$ l'ensemble des flèches entre eux. Plus précisément, on définit le foncteur covariant[3] :

$$Hom_C(X,-) : \mathcal{C} \longrightarrow Ens$$

$$M \longmapsto Hom_C(X,M).$$

De plus, pour chaque morphisme $(f : M \to N)$, on a une application

$$Hom_C(X,f) : Hom_C(X,M) \to Hom_C(X,N)$$

$$g \mapsto g \circ f.$$

Dualement, on définit le foncteur contravariant :

$$Hom_C(-,X) : \mathcal{C} \longrightarrow Ens$$

$$M \longmapsto Hom_C(M,X).$$

[3] Attention, on note $Hom_C(X,M)$ comme étant l'ensemble des morphismes qui vont de l'ensemble X à l'ensemble M voir la Proposition 3.4

De plus, pour chaque morphisme $(f : M \to N)$, on a une application

$$Hom_C(f,X) : Hom_C(N,X) \to Hom_C(M,X)$$

$$g \mapsto g \circ f.$$

Note 4.7. Il est à noter que le foncteur covariant $Hom_A(M,-)$ préserve les produits. D'autre part, le foncteur contravariant $Hom_A(-,N)$ transforme les sommes en produits.

Exemple 4.8. Un cas particulier du foncteur $Hom_{\mathcal{C}}(-,-)$ est celui où l'on prend comme catégorie $\mathcal{C} = ModA$ avec A une K-algèbre. Pour des A-modules X et Y, l'ensemble $Hom_{\mathcal{C}}(X,Y)$ est un k-module, alors on a l'application suivante ;

$$Hom_{ModA}(X,-) : ModA \longrightarrow Modk$$

$$M \longmapsto Hom_{ModA}(X,M).$$

Il en va de même pour le foncteur $Hom_{ModA}(-,X)$ qui est le dual de celui présenté.

4.2 Les sous-bifoncteurs de $\mathrm{Ext}^1_{\mathcal{E}_{max}}$

Soit $\mathcal{A}$ une catégorie additive au sens de la Définition 3.10.

Définition 4.9. [BBGH20, Definition 3.1] On considère le bifoncteur additif suivant :

$$\mathrm{Ext}^1_{\mathcal{E}_{max}} : \mathcal{A} \times \mathcal{A}^{op} \longrightarrow Ab$$

$$(C,A) \mapsto \{\overline{(i,d)} | (i,d) \in \mathcal{E}_{max}\},$$

où $\overline{(i,d)}$ est la classe d'équivalence (voir 3.16) de la suite exacte courte (i,d). Ce bifoncteur est bien défini puisque $\mathcal{E}_{max}$ existe pour toute catégorie additive $\mathcal{A}$ (voir 3.28).

Et d'une manière plus générale :

Définition 4.10. [BBGH20, Definition 4.1] Soit $\mathcal{W}$ une structure partiellement exacte

$$\mathrm{Ext}^1_{\mathcal{W}} : \mathcal{A} \times \mathcal{A} \to Ab$$

$$\mathbb{W}(C,A) = \left\{ \overline{(i,d)} \quad | \quad A \xrightarrow{i} B \xrightarrow{d} C \quad \in \mathcal{W} \right\},$$

on note par $\overline{(i,d)}$ la classe d'équivalence de la suite exacte courte (i,d).

Proposition 4.11. *[BBGH20, Lemme 3.2] Soit $\mathcal{V}$ et $\mathcal{W}$ des structures exactes sur $\mathcal{A}$ avec $\mathcal{V} \subseteq \mathcal{W}$. En assumant que la construction de la Définition 4.9 donne un foncteur additif qui est un groupe abélien. Alors il s'avère que le foncteur $\mathbb{V} = \mathrm{Ext}^1_{\mathcal{V}}(-,-) : \mathcal{A}^{op} \times \mathcal{A} \to Ab$ est en fait un* ***sous-bifoncteur*** *additif de $\mathbb{W} = \mathrm{Ext}^1_{\mathcal{W}}(-,-)$.*

Définition 4.12. [BH61, DRS+99, BBGH20, Définition 3.5] Un sous-bifoncteur $F \in \mathrm{Sub}(\mathrm{Ext}^1_{\mathcal{A}})$ est dit *fermé* si, pour n'importe quelle suite exacte courte

$$E:\ A \xrightarrow{\ i\ } B \xrightarrow{\ d\ } C$$

avec $\overline{(i,d)} \in F(C,A)$ et pour n'importe quel objet $X \in \mathcal{A}$, les suites qui suivent sont exactes dans la catégorie des groupes abéliens :

$$F(X,A) \to F(X,B) \to F(X,C)$$

et

$$F(C,X) \to F(B,X) \to F(A,X).$$

5 Rappels sur les treillis

Dans cette section, on fournit les prérequis combinatoitres en rappelant quelques notions et définitions nécessaires.

5.1 Rappels

On rappelle ici ce qu'est une structure de treillis sur un ensemble. Ce concept est indispensable pour construire des isomorphismes de treillis dans la suite de cette section.

Définition 5.1.

1. La *plus petite borne supérieure*, aussi appelée *suprémum*, d'une partie F d'un ensemble E est un élément $1 \in E$ tel que :
   (a) 1 est une borne supérieure de F ;
   (b) pour toute autre borne supérieure $m \in E$ de F, on a que $1 \leq m$.
2. La *plus grande borne inférieure*, aussi appelée *infimum*, d'une partie F d'un ensemble E est un élément $0 \in E$ tel que :
   (a) 0 est une borne inférieure de F ;
   (b) pour toute autre borne inférieure $m \in E$ de F, on a que $0 \geq m$.

Définition 5.2. Soit M un ensemble ordonné. Cet ensemble est un *treillis* si pour toute paire de deux éléments de M, il existe une borne supérieure (suprémum) et une borne inférieure (infimum)(voir note 5.1 ci-dessous). Autrement dit s'il existe deux opérations binaires $\vee$ et $\wedge$ vérifiant les axiomes suivants :

1. $\vee$ et $\wedge$ sont deux opérateurs désignant deux opérations binaires de la forme $M \times M \to M$.
2. La loi $\vee$ est associative et commutative[4]

[4] L'axiome d'associativité de la loi $\vee$ implique qu'il existe un élément identité I_a.

3. La loi $\wedge$ est associative et commutative.

4. Les opérations doivent respecter l'équation suivante [5]

$$m \vee (m \wedge n) = m = m \wedge (m \vee n), \forall m, n \in M.$$

Remarque 5.3. On nomme généralement l'équation précedente "loi d'absorption".

Exemple 5.4. L'ensemble des parties d'un ensemble muni de l'inclusion forme un treillis où la borne supérieure est l'union et la borne inférieure l'intersection.

Exemple 5.5. L'ensemble des entiers naturels muni de la relation de divisibilité forme un treillis, où la borne supérieure est le PPCM et la borne inférieure est le PGCD.

Définition 5.6. Un treillis est dit *borné* s'il possède un maximum et un minimum.

Exemple 5.7. L'ensemble des entiers naturels muni de la relation d'ordre $\leq$ n'est pas borné, mais le même ensemble muni de la relation d'ordre de divisibilité est un treillis borné dont le minimum est 1 et le maximum 0.

Définition 5.8. Un treillis est dit *complet* si toute partie possède une borne supérieure, ou encore si toute partie de E possède une borne inférieure.

Définition 5.9. Un treillis borné distributif (la loi $\vee$ est distributive par rapport à la loi $\wedge$,) et complémenté (chacun de ses éléments m possède un complément n vérifiant $m \wedge n = 0$ et $m \vee n = 1$) est dit *booléen*.

Définition 5.10. Un treillis est dit *modulaire* si *la loi de modularité* suivante est satisfaite :

$$s \leq b \implies s \vee (m \wedge n) = (s \vee m) \wedge b.$$

Proposition 5.11 ([Ass97], [AL09])**.** *Lorsqu'on considére un idéal bilatère I d'un anneau A, il existe une bijection croissante, dont l'inverse est aussi croissante et formant donc un isomorphisme de treillis :*

$$A \to A/I;$$

$$J \mapsto J/I$$

entre les idéaux J de A contenant I d'une part $I \subset J \subset A$, et les idéaux de A/I d'autre part $J/I \subset A/I$.

Théorème 5.12. *[Ass97, Lemme 1.4] $\mathcal{S}(M)$ est un treillis complet borné modulaire.*

[5] Comme conséquence de ces axiomes on aura que $m \vee m = m$ et $m \wedge m = m$, $\forall m \in M$.

Démonstration. L'ensemble $\mathcal{S}(M)$ défini en 2.7 ordonné par l'inclusion des ensembles forme un poset. Où les sous-modules $\bigcap_{\lambda\in\Lambda} M_\lambda$ et $\sum_{\lambda\in\Lambda} M_\lambda$ représentent respectivement l'infimum et le suprémum de chaque famille de sous-modules formant la structure de treillis complet sur $\mathcal{S}(M)$.
De plus, celui-ci est borné admettant 0 comme son plus petit élément et M comme son plus grand.
Enfin $\mathcal{S}(M)$ est modulaire vu qu'il satisfait la propriété modulaire suivante ; si $M_1, M_2, M_3 \in \mathcal{S}(M)$ et que $M_1 \subseteq M_2$, alors :

$$M_2 \cap (M_1 + M_3) = M_1 + (M_2 \cap M_3).$$

□

Définition 5.13. [DP02, 2.16] Soit L et K des treillis, alors une application $f : L \to K$ est un *morphisme de treillis* si $\forall a,b \in L$ on a :

$$f(a \vee b) = f(a) \vee f(b) \qquad \text{et} \qquad f(a \wedge b) = f(a) \wedge f(b).$$

Définition 5.14. [DP02, 2.16, 2.17] Soit L et K des treillis, une application $f : L \to K$ est un *isomorphisme de treillis* si f est un morphisme bijectif. De plus, si f est un isomorphisme de treillis bijectif alors son inverse est aussi un isomorphisme de treillis bijectif.

6 Les structures de treillis isomorphes

6.1 Le treillis des structures exactes

Théorème 6.1. *[BHLR20, Théorème 5.3, corollaire 5.4] L'ensemble partiellement ordonné $Ex(\mathcal{A})$ est un treillis borné et complet $(Ex(\mathcal{A}), \subseteq, \wedge, \vee)$.*

Démonstration. Nous ne ferons pas la preuve au complet, pour plus de détails allez voir la référence. L'ensemble partiellement ordonné $Ex(\mathcal{A})$ muni de l'inclusion des classes comme relation d'ordre forme un treillis pour les opérations suivantes :

- *la borne inférieure* $\wedge$ est définie par $\mathcal{E} \wedge \mathcal{E}' = \mathcal{E} \cap \mathcal{E}'$,
- *la borne supérieure* $\vee$ est définie par
  $\mathcal{E} \vee \mathcal{E}' = \cap\{\mathcal{E}'' \in Ex(\mathcal{A}) \mid \mathcal{E} \subseteq \mathcal{E}'', \mathcal{E}' \subseteq \mathcal{E}''\}$.

□

Théorème 6.2. *[BBGH20, Théorème 4.3] Soit $\mathcal{A}$ une catégorie additive, l'ensemble des sous-bifoncteurs fermés de $Ext^1_{\mathcal{E}_{max}}(-,-)$ que l'on notera $Cbf(\mathcal{A})$ forme un treillis $(Cbf(\mathcal{A}), \leq, \wedge, \vee)$.*

Démonstration. Tout comme pour le théorème précédent, nous ne ferons pas cette preuve dans l'article. Par contre, on y énoncera les éléments importants. On considère l'ensemble des sous-bifoncteurs de $Ext^1_{\mathcal{E}_{max}}$ partiellement ordonné par la relation d'ordre suivante :

$$F \leq F' \Longleftrightarrow F(C,A) \subseteq_{Ab} F'(C,A),$$

où $F,\ F' \in Cbf(\mathcal{A})$ des bifoncteurs et $F(C,A)$ est un sous-groupe abélien de $F'(C,A)$. Ce poset forme un treillis pour les opérations suivantes :

- *la borne inférieure* $\wedge$ est définie par

$$F \wedge F' = F \cap F'$$

  telle que $(F \wedge F')(C,A) = F(C,A) \cap_{Ab} F'(C,A)$ pour tous objets A, C de $\mathcal{A}$.

- *la borne supérieure* $\vee$ est définie par

$$F \vee F' = \cap\{F'' \in Cbf(\mathcal{A}) \mid F \leq F'', F' \leq F''\}.$$

□

Théorème 6.3. *[BBGH20, Théorème 4.4] Les deux treillis* $(Ex(\mathcal{A}), \subseteq, \wedge, \vee)$ *et* $(Cbf(\mathcal{A}), \leq, \wedge, \vee)$ *sont isomorphes.*

Démonstration. Dans la preuve du [BBGH20, Théorème 4.4] on vérifie que l'application

$$\phi : Ex(\mathcal{A}) \longrightarrow Cbf(\mathcal{A})\,;\ \mathcal{E} \longmapsto Ext_{\mathcal{E}}(-,-)$$

forme un isomorphisme de treillis. □

6.2 Le treillis des structures partiellement exactes

Théorème 6.4. *[BBGH20, Théorème 4.13] L'ensemble des structures partiellement exactes sur* $\mathcal{A}$ *forme un treillis*

$$(Wex(\mathcal{A}), \subseteq, \wedge, \vee_W).$$

Démonstration. L'ensemble partiellement ordonné $Wex(\mathcal{A})$ muni de l'inclusion des classes comme relation d'ordre forme un treillis pour les opérations suivantes :

- *la borne inférieure* $\wedge$ est définie par $\mathcal{W} \wedge \mathcal{W}' = \mathcal{W} \cap \mathcal{W}'$,

- *la borne supérieure* $\vee_W$ est définie par
  $\mathcal{W} \vee_W \mathcal{W}' = \cap\{\mathcal{W}'' \in Wex(\mathcal{A}) \mid \mathcal{W} \subseteq \mathcal{W}'', \mathcal{W}' \subseteq \mathcal{W}''\}$.

□

Maintenant on introduit la structure de treillis suivante permettant de voir le treillis des structures partiellement exactes d'un point de vue fonctoriel.

Théorème 6.5. *[BBGH20, Théorème 4.8] Soit $\mathcal{A}$ une catégorie additive, l'ensemble des sous-bifoncteurs de $Ext^1_{\mathcal{E}_{max}}(-,-)$ que l'on notera $Bf(\mathcal{A})$ est un treillis $(Bf(\mathcal{A}), \leq, \wedge, \vee_{Bf})$.*

Démonstration. Tout comme pour les théorèmes précédents, nous ne ferons pas cette preuve dans l'article. Par contre, on y énoncera les éléments importants. On considère l'ensemble des sous-bifoncteurs de $Ext^1_{\mathcal{E}_{max}}(-,-)$ partiellement ordonné par la relation d'ordre suivante :

$$F \leq F' \iff F(C,A) \subseteq_{Ab} F'(C,A),$$

où $F,\ F' \in Cbf(\mathcal{A})$ des bifoncteurs et $F(C,A)$ est un sous-groupe abélien de $F'(C,A)$. Ce poset forme un treillis pour les opérations suivantes :

- *la borne inférieure* $\wedge$ est définie par

  $$F \wedge F' = F \cap F'$$

  telle que $(F \wedge F')(C,A) = F(C,A) \cap_{Ab} F'(C,A)$ pour tous objets A, C de $\mathcal{A}$.

- *la borne supérieure* $\vee Bf$ est définie par

  $$F \vee_{Bf} F'$$

  telle que $F \vee_{Bf} F' = \wedge\{F'' \in Bf(\mathcal{A}) \mid F \leq F'',\ F' \leq F''\}$. Autrement dit $(F \vee_{Bf} F')(C,A) = F(C,A) +_{Ab} F'(C,A)$ pour tous objets A, C de $\mathcal{A}$.

□

Le résultat suivant établit le lien entre les deux treillis 6.4 et 6.5 :

Théorème 6.6. *[BBGH20, Théorème 4.16] Il y a un isomorphismes de treillis entre les treillis $(Wex(\mathcal{A}), \subseteq, \wedge, \vee_{\mathcal{W}})$ et $(Bf(\mathcal{A}), \leq, \wedge, \vee_{Bf})$.*

Démonstration. □

6.3 Le troisième treillis

Cette section a pour but d'introduire un nouveau treillis isomorphe aux deux treillis $Wex(\mathcal{A})$ et $Bf(\mathcal{A})$.

On construit, étape par étape, un nouvel isomorphisme de treillis entre l'ensemble des sous-bifoncteurs de $Ext^1_{\mathcal{E}_{max}}$ basé sur une catégorie additive fixée $\mathcal{A}$ et un certain treillis de bimodules. Ce dernier est défini comme l'ensemble des sous-bimodules d'un module M, qu'on construit en fonction de $\mathcal{A}$. Cet isomorphisme nous permet d'avoir un isomorphisme direct entre le treillis des structures partiellement exactes sur $\mathcal{A}$ et le nouveau treillis des sous-bimodules de M.

Cela nous permet de voir les structures partiellement exactes de plusieurs façons, ce qui facilite l'étude de leurs propriétés.

6.3.1 Définition du treillis $Bim(B)$

Soient $\mathcal{A}$ une catégorie additive Krull-Schmidt, avec $X = X_1 \oplus \cdots \oplus X_n$ et $B = \mathrm{End}\,(X)$ est son algèbre d'Auslander définie en 3.12.

Définition 6.7. Pour tous les objets A et C, on considère les morphismes $\Delta_C : C \to C \oplus C$ et $\nabla_A : A \oplus A \to A$ et les deux suites exactes courtes $E : 0 \to A \xrightarrow{\mu} B \xrightarrow{\lambda} C \to 0$ et $E' : 0 \to A \xrightarrow{\mu'} B' \xrightarrow{\lambda'} C \to 0$. On définit la *somme de Baer* donnée par

$$E + E' = \nabla(E \oplus E')\Delta.$$

La somme de Baer correspond donc à prendre la somme amalgamée de la suite exacte courte $E \oplus E'$ selon ∇_A puis le produit fibré selon Δ_C.

Lemme 6.8. *L'ensemble M qui suit*

$$\bigoplus_{j,k=1}^{n} \{\overline{(i,d)} \mid X_j \overset{i}{\rightarrowtail} Y \overset{d}{\twoheadrightarrow} X_k \in \mathcal{E}_{max}, X_j, X_k \in Ind(\mathcal{A}), Y \in Obj(\mathcal{A})\}$$

$$\bigoplus_{j,k=1}^{n} \{X_j \overset{i}{\rightarrowtail} \overset{d}{\twoheadrightarrow} X_k | \overline{(i,d)} \in \mathcal{E}_{max}, X_j, X_k \in Ind(\mathcal{A})\}$$

muni de la somme de Baer forme un bimodule sur l'algèbre d'Auslander B.

Démonstration. On considère les preuves dans [Mit65, Page 165], mais pour $Ext^1_{\mathcal{E}_{max}}$. Alors, les propriétés prouvées sont valides pour des suites exactes courtes dans les groupes abéliens $E, E' \in Ext^1_{\mathcal{E}_{max}}(X, X)$ avec $X = \bigoplus_{j=1}^{n} X_j$, $\forall j$ entre 0 et n avec $X_j \in Ind(\mathcal{A})$.
Donc [Mit65, Théorème 1.5, page 165,166] implique que M est un groupe abélien pour la somme de Baer. La preuve dans cette référence montre que la classe d'équivalence (pour la relation d'équivalence définie 3.16) de la suite exacte courte scindée

$$0 \longrightarrow X \longrightarrow X \oplus X \longrightarrow X \longrightarrow 0$$

représente l'élément neutre de ce groupe. [Mit65, Lemme 1.3, Lemme 1.4 (ii),(iii)] implique que M est un B-module à gauche et à droite selon la Définition 2.2, où la multiplication extérieure est donnée par le produit fibré et la somme amalgamée. De plus, [Mit65, Lemme 1.4 (iii)*] implique que ces deux structures de B-modules à gauche et à droite sont compatibles donc que M est un $(B\text{-}B)$ bimodule selon la Définition 2.13.

□

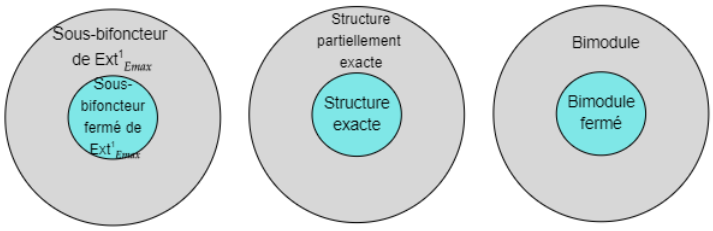
Théorème 6.9. *L'ensemble de tous les sous-bimodules de ${}_BM_B$ défini en 6.8, que l'on notera $Bim(B)$, forme un treillis borné complet modulaire*

$$(Bim(B), \leq, \wedge_{Bim(B)}, \vee_{Bim(B)}).$$

Démonstration. L'ensemble $Bim(B)$ muni de l'inclusion d'ensembles $\subseteq$ comme relation d'ordre forme un poset. De plus, ce poset forme un treillis comme 2.7 pour les opérations suivantes :

- *la borne inférieure* $\wedge_{Bim(B)}$ est définie par $N \wedge N' = N \cap N'$,
- *la borne supérieure* $\vee_{Bim(B)}$ est définie par $N \vee N' = N + N'$.

□



Le diagramme ci-dessus explicite les relations entre les ensembles définis précédemment et leur sous-ensemble respectif.

6.3.2 Construction de l'isomorphisme

Soit $\mathcal{A}$ une catégorie additive Krull-Schmidt et B l'algèbre d'Auslander définie en 3.12.

Pour cette construction, on considère l'ensemble $Bf(\mathcal{A})$ de tous les sous-bifoncteurs de $Ext^1_{\mathcal{E}_{max}}(-,-)$ défini en 4.9. On prend également l'ensemble $Bim(B)$ de tous les sous-bimodules du $(B\text{-}B)$ bimodule M construit en 6.8.
Remarquons que l'isomorphisme qu'on construit ici est donné par l'application de chaque bifoncteur à la somme directe de tous les objets indécomposables de la catégorie $\mathcal{A}$. Par ailleurs, puisque nos bifoncteurs sont biadditifs, les images des indécomposables de $\mathcal{A}$ résument toutes les informations nécessaires pour définir le bifoncteur en tout sur la catégorie $\mathcal{A}$. Notons que cette idée est similaire à celle utilisée pour montrer l'équivalence de catégories en [ASS06, A.2, exemple 2.10].

Théorème 6.10. *La correspondance suivante :*

$$\Phi : Bf(\mathcal{A}) \longrightarrow Bim(B)$$

$$F \mapsto F(X,X)$$

est un isomorphisme de treillis.

Démonstration.

1. Application bien définie :

   On considère un élément $F \in Bf(\mathcal{A})$, c'est-à-dire un sous-bifoncteur de $Ext^1_{\mathcal{E}_{max}}(-,-)$. Par l'isomorphisme 6.6, il existe une structure partiellement exacte $\mathcal{W}$ telle que $F = Ext^1_{\mathcal{W}}(-,-)$. Cet F appliqué à l'objet

$X = \overset{n}{\underset{j=1}{\oplus}} X_j$ donne l'ensemble des suites exactes courtes suivant

$$F(X,X) = Ext^1_{\mathcal{W}}(X,X) = Ext^1_{\mathcal{W}}(\overset{n}{\underset{j=1}{\oplus}} X_j, \overset{n}{\underset{k=1}{\oplus}} X_k) = \overset{n}{\underset{j,k=1}{\oplus}} Ext^1_{\mathcal{W}}(X_j, X_k)$$

$$= \overset{n}{\underset{j,k=1}{\oplus}} \{\overline{(i,d)} \mid X_j \overset{i}{\rightarrowtail} Y \overset{d}{\twoheadrightarrow} X_k \in \mathcal{W}\}.$$

Pour prouver que $F(X,X) \in Bim(B)$, montrons que $F(X,X)$ est un sous-bimodule du $(B\text{-}B)$ bimodule M défini en 6.8. C'est-à-dire montrons que c'est un sous-ensemble de M ayant une structure de bimodule par les mêmes opérations de la structure de $(B\text{-}B)$ bimodule sur M restreintes à l'ensemble $F(X,X)$:

- sous-ensemble de M :
  le fait que $\mathcal{W} \subseteq \mathcal{E}_{max}$ implique que

$$F(X,X) = \overset{n}{\underset{j,k=1}{\oplus}} \{\overline{(i,d)} \mid X_j \overset{i}{\rightarrowtail} Y \overset{d}{\twoheadrightarrow} X_k \in \mathcal{W}\} \subseteq$$

$$\overset{n}{\underset{j,k=1}{\oplus}} \{\overline{(i,d)} \mid X_j \overset{i}{\rightarrowtail} Y \overset{d}{\twoheadrightarrow} X_k \in \mathcal{E}_{max}\} = M$$

- $(B\text{-}B)$ bimodule :
  L'argument des preuves dans [Mit65, Lemme 1.3, Lemme 1.4, Théorème 1.5] est valide pour $Ext^1_{\mathcal{W}}(-,-)$ vu qu'il n'utilise pas les axiomes $(E1)$ et $(E1)^{op}$ d'une structure exacte (voir 3.22). Par le fait même, l'argument s'applique à la structure partiellement exacte $\mathcal{W}$. Cela implique que $F(X,X) = Ext^1_{\mathcal{W}}(\overset{n}{\underset{k=1}{\oplus}} X_k, \overset{n}{\underset{j=1}{\oplus}} X_j)$ forme un $(B\text{-}B)$ bimodule pour les mêmes opérations que M vu en 6.8.

En plus $F = Ext^1_{\mathcal{W}}(-,-)$ est un foncteur bien défini alors Φ envoie chaque $F \in Bf(\mathcal{A})$ à une *unique* image $F(X,X) \in Bim(B)$ et donc Φ est une application bien définie.

2. bijection :

   - injectif :
     On considère deux sous-bifoncteurs $F,\ G \in Bf(\mathcal{A})$ tel que leurs images sous Φ sont des $(B\text{-}B)$ bimodules égaux :

$$\bigoplus_{r,s=1}^{n} F(X_r,X_s)^{X(r)X(s)} \cong F(X,X)$$

$$= G(X,X)$$

$$\cong \bigoplus_{r,s=1}^{n} G(X_r,X_s)^{X(r)X(s)},$$

où $X \cong X_1^{X(1)} \oplus \cdots \oplus X_n^{X(n)}$.

On considère deux objets Y, Z de $\mathcal{A}$, qu'est une catégorie Krull-Schmidt, alors ils se décomposent en somme directe finie d'objets indécomposables de $\mathcal{A}$, et vu que $X = \overset{n}{\underset{i=1}{\oplus}} X_i$ est la somme directe de tous les objets indécomposables de $\mathcal{A}$ alors tous objets Y, Z s'écrivent comme somme directe de facteurs directs de X, ou autrement sont des sous objets de X :

$$Y \cong X_1^{y(1)} \oplus \cdots \oplus X_n^{y(n)}$$

$$Z \cong X_1^{z(1)} \oplus \cdots \oplus X_n^{z(n)}.$$

Sachant que F et G sont des bifoncteurs biadditifs, les images de (Y, Z) se décomposent :

$$F(Y,Z) \cong \bigoplus_{j,k=1}^{n} F(X_k,X_j)^{y(k)z(j)}$$

$$G(Y,Z) \cong \bigoplus_{j,k=1}^{n} G(X_k,X_j)^{y(k)z(j)}.$$

L'hypothèse $F(X,X) = G(X,X)$ et l'additivité des foncteurs F et G implique l'égalité sur tous les facteurs directs :

$$F(X_k,X_j) = G(X_k,X_j)$$

$\forall j,k$ on obtient $F(Y,Z) = G(Y,Z)$ $\forall Y$, Z dans $\mathcal{A}$ et alors $F = G$, d'où Φ est injectif.

- surjectif :

  Soit $N \in Bim(B)$ un sous-bimodule de M dont la structure de (B-B) bimodule est étudiée en 6.8; étant un groupe abélien pour la somme de Baer dont l'élément neutre est donné par la classe d'équivalence de la suite exacte courte scindée basée sur X et les deux opérations de multiplication externes par des morphismes $b \in B$ sont données par la somme amalgamée et le produit fibré. Alors par la définition d'un sous-module 2.6 les suites exactes courtes scindées appartiennent à N et toutes les suites exactes courtes de N satisfont les axiomes $(E0)$, $(E0)^{op}$,$(E2)$ et $(E2)^{op}$ et leurs sommes directes forment par 3.25 une structure partiellement exacte $\mathcal{W}$. De plus, par 6.6 il existe un bifoncteur additif $F = Ext^1_{\mathcal{W}}(-,-) \in Bf(\mathcal{A})$ tel que $N \cong \bigoplus_{j,k=1}^{n} Ext^1_{\mathcal{W}}(X_k, X_j) = F(X,X)$.

3. morphisme d'ensembles partiellement ordonnés :

   Ici, on montre que Φ préserve l'ordre, c'est-à-dire que pour F, $F' \in Bf(\mathcal{A})$ tel que F est un sous-bifoncteur de F' alors $F(X,X)$ est un sous-bimodule

de $F'(X,X)$ dans M.

Par la définition de l'ordre sur le poset $Bf(\mathcal{A})$, on a que $F(C,A)$ est un sous-groupe abélien de $F'(C,A)$ pour tous objets A, C de $\mathcal{A}$. C'est vrai en particulier pour $A = C = \overset{n}{\underset{j=1}{\oplus}} X_j = X$. On obtient donc que $F(X,X)$ est un sous-groupe abélien de $F'(X,X)$.
Il reste à montrer que $F(X,X)$ est stable pour la multiplication externe par des éléments de l'algèbre d'Auslander B. Soit $b \in B$ et $E \in F(X,X)$, on considère le diagramme suivant :

$$\begin{array}{ccccccccc} E: 0 & \longrightarrow & X & \overset{i}{\rightarrowtail} & & \overset{d}{\twoheadrightarrow} & X & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle b} & & \downarrow & & \| & & \\ bE: 0 & \longrightarrow & X & \overset{p}{\rightarrowtail} & S.A & \overset{q}{\twoheadrightarrow} & X & \longrightarrow & 0 \end{array}$$

où la somme amalgamée existe grâce à l'axiome $(E2)$ de la Définition 3.25. On a donc $(p,q) \in F(X,X)$ ce qui implique que $F(X,X)$ est un module à gauche. Duallement, grâce à $(E2)^{op}$, $F(X,X)$ est également un module à droite. D'où $F(X,X)$ est un sous-bimodule de $F'(X,X)$. On obtient ainsi que l'implication suivante est vraie.

$$F \leq F' \Longrightarrow \Phi(F) \subseteq \Phi(F').$$

4. morphisme de treillis :

   On veut montrer que Φ préserve la structure de treillis, c'est-à-dire que Φ préserve *la borne inférieure* $\wedge$ ainsi que *la borne supérieure* $\vee$:

   - Pour ce faire, on montre qu'en appliquant Φ à la borne inférieure de n'importe quels deux éléments du treillis $Bf(\mathcal{A})$, on obtient la borne inférieure de leurs images dans le treillis $Bim(B)$. Par les Théorèmes 6.2 et 6.5, la borne inférieure de $Bf(\mathcal{A})$ est donnée par $(F \wedge_{Bf} F')(C,A) = F(C,A) \cap_{Ab} F'(C,A)$ pour tous objets A, C de $\mathcal{A}$. C'est vrai en particulier pour $A = C = \overset{n}{\underset{j=1}{\oplus}} X_j = X$. Donc, on a :

$$\begin{aligned} \Phi(F \wedge_{Bf} F') &= (F \wedge_{Bf} F')(X,X) \\ &= F(X,X) \cap_{Ab} F'(X,X) \\ &= F(X,X) \cap_{Mod(B)} F'(X,X) \\ &= F(X,X) \wedge_{Bim(B)} F'(X,X) \\ &= \Phi(F) \wedge_{Bim(B)} \Phi(F'). \end{aligned}$$

     Or, l'intersection des bimodules représente la borne inférieure du treillis de $Bim(\mathcal{A})$ par 6.9.

- Pour ce faire, on montre qu'en appliquant Φ à la borne supérieure de n'importe quels deux éléments du treillis $Bf(\mathcal{A})$, on obtient la borne supérieure de leurs images dans le treillis $Bim(B)$. Par le Théorème 6.2, pour tous objets A, C de $\mathcal{A}$, la borne supérieure de $Bf(\mathcal{A})$ est donnée par $(F \vee_{Bf} F')(C,A) = F(C,A) +_{Ab} F'(C,A)$. C'est vrai en particulier pour $A = C = \bigoplus_{j=1}^{n} X_j = X$. Donc, on a :

$$\begin{aligned}\Phi(F \vee_{Bf} F') &= (F \vee_{Bf} F')(X,X)\\ &= F(X,X) +_{Ab} F'(X,X)\\ &= F(X,X) +_{Mod(B)} F'(X,X)\\ &= F(X,X) \vee_{Bim(B)} F'(X,X)\\ &= \Phi(F) \vee_{Bim(B)} \Phi(F').\end{aligned}$$

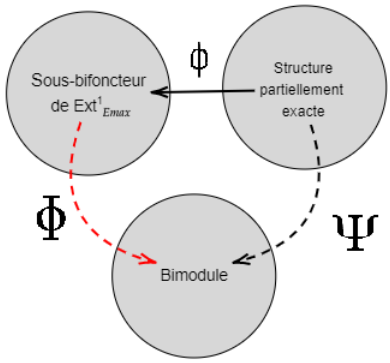
  Or, la somme des bimodules représente la borne supérieure du treillis de $Bim(\mathcal{A})$ par 6.9.

D'où l'application Φ est bien un isomorphisme de treillis par 5.14.

□

Corollaire 6.11. *Les deux treillis $Wex(\mathcal{A})$ et $Bim(B)$ sont isomorphes.*

Démonstration. Par le Théorème 6.6, il existe un isomorphisme de treillis ϕ entre le treillis $Wex(\mathcal{A})$ étudié en 6.4 et le treillis $Bf(\mathcal{A})$ étudié en 6.5. De plus, on vient de montrer, dans le Théorème 6.10, qu'il existe également un isomorphisme de treillis Φ entre le treillis $Bf(\mathcal{A})$ étudié en 6.5 et le treillis $Bim(B)$ étudié en 6.9. On considère alors l'isomorphisme $\Psi = \Phi \circ \phi$ donné par la composition. On obtient un isomorphisme de treillis entre $Wex(\mathcal{A})$ et $Bim(B)$. □



La flèche rouge Φ de ce diagramme représente l'application qu'on a construite dans cet article. La flèche pleine noire ϕ représente l'application qui existait déjà. La flèche pointillée noire représente la composition de ces deux applications.

Remarque 6.12. L'isomorphisme de treillis Φ étudié en 6.10 est un morphisme de treillis bijectif. On a montré la bijectivité en montrant que c'est injectif et surjectif. Un moyen alternatif est de considérer l'application inverse suivante

$$\Phi^{-1} : Bim(B) \longrightarrow Bf(\mathcal{A})$$

$$_BN_B \longmapsto F_N(-,-)$$

avec

$$F_N(-,-) : \mathcal{A} \times \mathcal{A}^{op} \to Ab$$

$$\{\overline{(i,d)} | A \overset{i}{\rightarrowtail} B \overset{d}{\twoheadrightarrow} C \in \mathcal{E}_{max}, \overline{(i,d)} \in N\}$$

où

$$0 \longrightarrow A \xrightarrow{i} B \xrightarrow{d} C \longrightarrow 0$$

tel que $\Phi \circ \Phi^{-1} = 1_{Bim(B)}$ et $\Phi^{-1} \circ \Phi = 1_{Bf(\mathcal{A})}$.

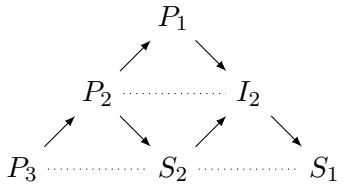
6.4 Exemple

Dans cette section, on présente un exemple concret du treillis étudié dans les sections précédentes de cet article. Par contre, pour plus de détails sur les concepts utilisés, on suggère au lecteur d'aller voir la référence [ASS06]. Cela est indispensable pour une profonde compréhension de cette section.

Soit la catégorie additive $\mathcal{A} = \operatorname{rep} Q$ des représentations du carquois suivant :

$$Q : \qquad 1 \longrightarrow 2 \longrightarrow 3.$$

Le carquois d'Auslander-Reiten de $\mathcal{A}$ est le suivant :



Il existe, à équivalence près, exactement cinq suites exactes courtes indécomposables non-scindées. Les trois premières énoncées ci-dessous sont celles d' Auslander-Reiten :

(α) $$0 \longrightarrow P_3 \xrightarrow{a} P_2 \xrightarrow{c} S_2 \longrightarrow 0$$

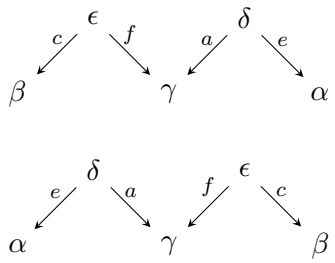
(β) $$0 \longrightarrow S_2 \xrightarrow{e} I_2 \xrightarrow{f} S_1 \longrightarrow 0$$

(γ) $$0 \longrightarrow P_2 \longrightarrow P_1 \oplus S_2 \longrightarrow I_2 \longrightarrow 0$$

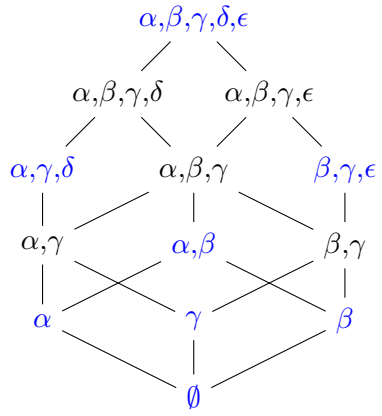
$$(\delta) \qquad 0 \longrightarrow P_3 \longrightarrow P_1 \xrightarrow{d} I_2 \longrightarrow 0$$

$$(\epsilon) \qquad 0 \longrightarrow P_2 \xrightarrow{b} P_1 \longrightarrow S_1 \longrightarrow 0$$

À isomorphisme près, un foncteur additif est uniquement déterminé par ses valeurs sur des objets indécomposables. Pour étudier les sous-bifoncteurs additifs de $\mathrm{Ext}^1_{\mathcal{A}}$, il suffit d'examiner la structure des bimodules sur l'espace vectoriel généré par ces cinq suites exactes courtes indécomposables non-scindées. Selon la notation de [M65,p.163] et aussi de [BBH, déf 3.1] la notation habituelle est de mettre le morphisme à gauche pour la somme amalgamée et à droite pour le produit fibré : $\delta e = \alpha$, $a\delta = \gamma$, $\epsilon f = \gamma$, $c\epsilon = \beta$:



On voit que $Ext^1_{\mathcal{E}_{max}}(-,-)$ admet treize sous-bifoncteurs (incluant l'élément neutre et lui-même). Le treillis $Bim(B)$, défini en 6.9, est représenté par le diagramme suivant où chaque sous-bimodule est représenté par son sous-ensemble de générateurs :



Les huit sous-bimodules en bleu parmi tous les sous-bimodules de $Bim(B)$ représentent les huit structures exactes de $Ex(\mathcal{A})$ parmi toutes les structures partiellement exactes de $Wex(\mathcal{A})$. D'une manière équivalente, ceux-ci représentent les huit sous-bifoncteurs fermés de $Cbf(\mathcal{A})$ parmi tous les sous-bifoncteurs de $Bf(\mathcal{A})$.

Références

[AL09] Ibrahim Assem et Pierre Yves Leduc : *Cours d'algèbre : groupes, anneaux, modules et corps*. Presses inter Polytechnique, 2009.

[Ass97] Ibrahim Assem : *Algèbres et modules*. Masson, Paris, 1997.

[ASS06] Ibrahim Assem, Daniel Simson et Andrzej Skowroński : *Elements of the representation theory of associative algebras. Vol. 1*, volume 65 de *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 2006. Techniques of representation theory.

[BBGH20] Rose-Line Baillargeon, Thomas Brüstle, Mikhail Gorsky et Souheila Hassoun : On the lattices of exact and weakly exact structures. *arXiv preprint arXiv :2009.10024*, 2020.

[BH61] Michael Charles Richard Butler et Geoffrey Horrocks : Classes of extensions and resolutions. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 254(1039):155–222, 1961.

[BHLR20] Thomas Brüstle, Souheila Hassoun, Denis Langford et Sunny Roy : Reduction of exact structures. *Journal of Pure and Applied Algebra*, 224(4):106212, 2020.

[Büh10] Theo Bühler : Exact categories. *Expositiones Mathematicae*, 28(1): 1–69, 2010.

[DP02] Brian A Davey et Hilary A Priestley : *Introduction to lattices and order*. Cambridge university press, 2002.

[DRS+99] Peter Dräxler, Idun Reiten, Sverre Smalø, Øyvind Solberg, B Keller *et al.* : Exact categories and vector space categories. *Transactions of the American Mathematical Society*, 351(2):647–682, 1999.

[Mit65] Barry Mitchell : *Theory of categories*. Academic Press, 1965.

[Qui73] Daniel Quillen : Higher algebraic k-theory : I. *In Higher K-theories*, pages 85–147. Springer, 1973.

[Rum11] Wolfgang Rump : On the maximal exact structure of an additive category. *Fundamenta Mathematicae*, 214:77–87, 2011.

Souheila Hassoun
Département de mathématiques, Université de Sherbrooke
Courriel: Souheila.Hassoun@USherbrooke.ca

Élodie Lapointe
Département de mathématiques, Université de Sherbrooke
Courriel: Elodie.Lapointe@USherbrooke.ca

La revue **CaMUS** souhaite remercier ses commanditaires, sans qui l'achèvement de ce projet aurait été beaucoup plus ardu. Pour ce volume, nous avons eu l'appui de trois généreux partenaires :

L'Association Générale des Étudiantes et Étudiants en Sciences

Le Regroupement des étudiants-chercheurs en sciences de l'Université de Sherbrooke

Le Fonds d'appui à l'engagement étudiant de l'Université de Sherbrooke

Également, la revue **CaMUS** aimerait remercier les arbitres ayant donné généreusement de leur temps afin de réviser en profondeur les articles publiés dans ce volume.

CaMUS est une revue mathématique publiée par le Département de mathématiques de l'Université de Sherbrooke. Le but de ces cahiers est de permettre aux étudiants et étudiantes de présenter leurs travaux effectués dans le cadre d'activités tels les stages de recherche du premier cycle, les présentations au Club Mathématique et les cours d'initiation à la recherche. Les personnes auteures sont principalement des étudiants et étudiantes au premier cycle, notamment en mathématiques et au Baccalauréat en enseignement au secondaire avec profil en mathématiques. Cette revue est publiée à l'intention de tous ceux et celles qui s'intéressent aux mathématiques, à l'Université de Sherbrooke et ailleurs. Ses objectifs sont de favoriser :

- l'apprentissage de la rédaction d'articles dès le premier cycle,
- l'amélioration de la communication écrite,
- le développement de la rigueur d'expression,
- la motivation du personnel étudiant pour la recherche en général.

Information aux auteurs : Les articles doivent constituer des présentations originales, mais on ne demande pas qu'ils contiennent des résultats originaux : pour la prépublication de ces derniers, la série de Rapports de recherche du Département de mathématiques est un meilleur médium. Les articles doivent être rédigés en LaTeX et soumis suivant les instructions données à la page Web de CaMUS :

http ://camus.math.usherbrooke.ca

Abonnement, achat d'exemplaires et autres renseignements : CaMUS est une publication électronique sans frais avec un nombre limité d'exemplaires imprimés à vendre. Pour des informations sur le prix courant, le mode de paiement ou pour nous contacter, rendez-vous à l'adresse Web indiquée ci-dessus.

CaMUS · Département de mathématiques · Faculté des sciences · Université de Sherbrooke · 2500, boulevard de l'Université · Sherbrooke (Québec), Canada J1K 2R1