L'inégalité de Grothendieck en bref...

1ère partie. Des normes matricielles à la présentation détaillée de l'inégalité

par Jean-Baptiste Hiriart-Urruty [1] et Patrice Lassère [2]

Résumé. Nous présentons d'une manière accessible, graduée et commentée la dite inégalité de Grothendieck (1953) ; les points d'appui sont les normes matricielles qui permettent une explicitation plus claire et perceptible de cette inégalité. En Annexe, nous en donnons une preuve, ainsi que des références pour accéder à d'autres manières de démontrer. Dans une deuxième partie (article à part), nous montrerons l'utilité et l'usage de l'inégalité de Grothendieck dans un exemple en Optimisation combinatoire moderne.

Introduction

Le nom de Grothendieck [3] est habituellement associé à ses contributions en *Géométrie algébrique* mais, avant la longue période de ses travaux dans ce domaine, il avait également travaillé profondément en *Analyse fonctionnelle*. La présente note concerne une inégalité, essentiellement matricielle, qui porte son nom et date de cette première période. Certains pensent même que la partie Géométrie algébrique a quelque peu occulté l'apport de Grothendieck en Analyse fonctionnelle et que le théorème-inégalité que nous évoquerons est un peu oublié ([14]).

Nous énonçons l'inégalité en question de manière élémentaire, progressive et commentée, sans toutefois en donner une démonstration complète dans le corps du texte, mais que nous proposons en Annexe et pour laquelle nous suggérons quelques références additionnelles précises.

Si nous revisitons cette inégalité, c'est parce qu'elle a une utilité dans l'Optimisation combinatoire moderne ; c'est donc l'occasion (dans la deuxième partie, un article à part) d'en fournir un exemple compréhensible par le plus grand nombre de lecteurs potentiels.

1. Au début étaient les normes sur $\mathbb{R}^n$ et sur $\mathcal{M}_{m,n}(\mathbb{R})$

1.1 Questionnez autour de vous au sujet d'exemples de normes sur $\mathbb{R}^n$... Il est probable qu'on vous répondra avec la norme euclidienne usuelle $\|.\|_2$ ou, plus généralement, pour $p \geqslant 1$, les normes $\|.\|_p$ définies comme suit :

$$\mathbb{R}^n \ni x = (x_1, x_2, ..., x_n) \mapsto \|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}. \qquad (1)$$

1. Département de mathématiques, université Paul Sabatier de Toulouse.
http ://www.math.univ-toulouse.fr/˜jbhu/

2. Département de mathématiques, université Paul Sabatier de Toulouse.
https ://perso.math.univ-toulouse.fr/lassere/

3. Grothendieck vécut la dernière partie de sa vie dans le village de Lasserre en Ariège (région Occitanie). Lasserre est un nom courant dans les contrées du sud-ouest et du sud de la France ; en occitan, "*sèrre*" ou "*sèrra*" [prononcé *sèrro*] désigne une hauteur de forme allongée, une crête dentelée de montagne, ce qui est le cas pour le village de Lasserre ou d'autres en Ariège.... Le second auteur de la présente note est donc *un Lasserre sans en avoir l'r...*

Le succès et la popularité de ces normes vient de leur utilisation en Analyse fonctionnelle, ne serait-ce que pour définir les dits espaces ℓ_p et L^p ; c'est pour cela d'ailleurs qu'on les appelle "les normes ℓ_p". Dans la note [8] par exemple, sont étudiés les comportements, en fonction de la dimension n de l'espace sous-jacent, d'éléments associés aux boules unités des normes $\|.\|_1$,$\|.\|_2$,$\|.\|_\infty$ dans $\mathbb{R}^n$, tels que le nombre de sommets, le diamètre et le volume.

On ne doit pas ici sous-estimer la difficulté qu'il peut y avoir dans le calcul numérique de $\|x\|_p$ pour p grand, en raison de l'opposition entre les puissances p et $1/p$ dans l'expression $(\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$. Essayez pour voir de calculer la norme $\|(4,7)\|_{10}$ du vecteur $(4,7)$ de $\mathbb{R}^2$!

Le cas où $p = 1$ est un peu à part puisque c'est le seul cas où la norme $\|.\|_p$ est *polyédrale* (on dit aussi *polytopale*) : $\|.\|_1$ est le maximum d'un nombre fini de formes linéaires sur $\mathbb{R}^n$, plus précisément

$$\|(x_1, x_2, ..., x_n)\|_1 = \max_{\varepsilon_i \in \{-1,1\}} (\varepsilon_1 x_1 + \varepsilon_2 x_2 + ... + \varepsilon_n x_n)\,. \tag{2}$$

Il y a exactement 2^n formes linéaires apparaissant dans le membre de droite de (2). La boule unité pour la norme $\|.\|_1$, notée $B_1(0,1)$, est ainsi définie via 2^n inégalités linéaires

$$\varepsilon_1 x_1 + \varepsilon_2 x_2 + ... + \varepsilon_n x_n \leqslant 1 \quad (\varepsilon_i \in \{-1,1\} \text{ pour tout } i = 1, ..., n).$$

La boule unité $B_1(0,1)$ est un polytope qui a exactement $2n$ sommets, lesquels sont $\pm e_i = (0, 0, ..., \pm 1, ..., 0)$, $i = 1, ..., n$.

Considérer le cas limite, $p \to +\infty$, est l'occasion d'un intéressant exercice, tout à fait habituel d'ailleurs : pour tout $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$,

$$\lim_{p \to +\infty} \|x\|_p = \max_{i=1,...,n} |x_i|\,. \tag{3}$$

Ce résultat est précisément la raison pour laquelle on utilise la notation $\|x\|_\infty$ pour $\max_{i=1,...,n} |x_i|$. De fait, $\|.\|_\infty$ est une nouvelle norme, polyédrale comme $\|.\|_1$, appelée "norme ℓ_∞" ou "norme du max". On pourrait écrire

$$\|(x_1, x_2, ..., x_n)\|_\infty = \max_{\varepsilon_i \in \{-1,1\}} (\varepsilon_1 x_1, \varepsilon_2 x_2, ..., \varepsilon_n x_n)\,. \tag{4}$$

Il y a exactement $2n$ formes linéaires apparaissant dans le membre de droite de (4). La boule unité pour la norme $\|.\|_\infty$, notée $B_\infty(0,1)$, est ainsi définie via $2n$ inégalités linéaires

$$\varepsilon_i x_i \leqslant 1 \quad (\varepsilon_i \in \{-1,1\} \text{ pour tout } i = 1, ..., n).$$

La boule unité $B_\infty(0,1)$ est un polytope, plus précisément l'hypercube n-dimensionnel $[-1,1]^n$ qui a exactement 2^n sommets, lesquels sont $(\pm 1, \pm 1, ..., \pm 1, ..., \pm 1)$, $i = 1, ..., n$.

On observe déjà la "dualité" ou "polarité" (*nombre d'inégalités linéaires la définissant*) versus (*nombre de sommets*) entre les boules unités $B_1(0,1)$ et $B_\infty(0,1)$, ce n'est pas un hasard. D'une manière générale, moyennant la relation $1/p + 1/p^* = 1$ (y compris le cas

limite où $p = 1$ et $p^* = \infty$), les normes $\|.\|_p$ et $\|.\|_{p^*}$ sont reliées par la relation de dualité suivante, sous-produit de l'inégalité de HÖLDER :

$$\|u\|_{p^*} = \max_{\|v\|_p \leqslant 1} \langle u, v \rangle \,, \tag{5}$$

où $\langle u, v \rangle$ désigne le produit scalaire usuel entre les vecteurs u et v de $\mathbb{R}^n$. Evidemment, $(p^*)^* = p$, et ce jeu de va-et-vient entre p et son conjugué p^* est d'un usage constant lorsqu'on manipule les normes ℓ_p.

Quand $1 < p < +\infty$, il en est de même de p^* ; ainsi les normes $\|.\|_p$ et $\|.\|_{p^*}$ vivent dans un même monde, celui des normes "lisses", c'est-à-dire celui des normes définissant des boules dont les frontières sont des surfaces lisses (*i.e.*, avec un hyperplan tangent en tout point) ; le cas $p = 2$ (et donc $p^* = 2$) en est l'exemple emblématique.

Les cas "extrêmes" où $p = 1$ ou $p = +\infty$ sont les deux seuls polyédraux de la famille des normes $\|.\|_p$, cette "extrémalité" les rend encore plus intéressantes que les autres. La relation de dualité (5) prend une forme simple ici :

$$\|u\|_\infty = \max_{\|v\|_1 \leqslant 1} \langle u, v \rangle \ ; \ \|u\|_1 = \max_{\|v\|_\infty \leqslant 1} \langle u, v \rangle \,. \tag{6}$$

Le comportement avec n du nombre de sommets des boules unités $B_1(0,1)$ et $B_\infty(0,1)$, dont on ne prend pas vraiment conscience quand n est petit, est un point important à noter et, à vrai dire, la source des difficultés futures : le nombre $2n$ de sommets de $B_1(0,1)$ croît "de manière polynomiale" avec n, tandis que le nombre 2^n de sommets de $B_\infty(0,1)$ croît "de manière exponentielle" avec n. C'est déjà la manifestation de ce qu'on appelle "l'explosion combinatoire" dans certains domaines d'applications (comme l'Optimisation combinatoire).

En faisant croître p de 1 à $+\infty$, on a une chaîne d'inégalités

$$\|.\|_\infty \leqslant ... \leqslant \|.\|_{p_2} \leqslant \|.\|_{p_1} \leqslant ... \leqslant \|.\|_1 \,, \text{ dès lors que } p_2 \geqslant p_1, \tag{7}$$

permettant de passer de $\|.\|_\infty$ à $\|.\|_1$.

Ce n'est certes pas la seule façon de faire, on peut par exemple considérer les normes "intermédiaires" que voici : avec k entier compris entre 1 et n,

$$\|(x_1, x_2, ..., x_n)\|_{(k)} = \max \left\{ |x_{i_1}| + |x_{i_2}| + ... + |x_{i_2}| : 1 \leqslant i_1 < ... < i_k \leqslant n \right\}. \tag{8}$$

Ces normes $\|.\|_{(k)}$ ont l'avantage - ou l'inconvénient suivant les cas - d'être toutes polyédrales, et elles permettent comme au-dessus de passer de $\|.\|_\infty = \|.\|_{(1)}$ à $\|.\|_1 = \|.\|_{(n)}$ par la chaîne d'inégalités suivante :

$$\|.\|_\infty = \|.\|_{(1)} \leqslant ... \leqslant \|.\|_{(k)} \leqslant \|.\|_{(\ell)} \leqslant ... \leqslant \|.\|_{(n)} = \|.\|_1 \,, \text{ dès lors que } \ell \geqslant k. \tag{9}$$

Certaines applications, comme celles traitant de la dite parcimonie des données, font usage plutôt de ces normes dont la combinatoire (nombre de sommets des boules unités, nombre d'inégalités linéaires les définissant) est bien étudiée à présent ([9]).

1.2 Quand on passe de $\mathbb{R}^n$ à $\mathcal{M}_{m,n}(\mathbb{R})$

Sur $\mathcal{M}_{m,n}(\mathbb{R})$, espace des matrices à m lignes et n colonnes, on peut définir autant de normes qu'on veut, puisqu'au fond $\mathcal{M}_{m,n}(\mathbb{R})$ peut être vu comme $\mathbb{R}^{n\times m}$; une première qui vient à l'esprit est

$$A = [a_{ij}] \longmapsto \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|\,. \tag{10}$$

Mais il y a une classe de normes sur $\mathcal{M}_{m,n}(\mathbb{R})$ plus intéressante, ce sont celles construites à partir de normes (bien séparées) sur $\mathbb{R}^n$ et sur $\mathbb{R}^m$. En effet, A peut être vue comme une application linéaire de $\mathbb{R}^n$ sur $\mathbb{R}^m$, $x \in \mathbb{R}^n \mapsto Ax \in \mathbb{R}^m$ en notations les plus légères possibles. Alors, une fois qu'on a choisi une norme $\|.\|_d$ sur l'espace de départ $\mathbb{R}^n$ et une norme $\|.\|_a$ sur l'espace d'arrivée $\mathbb{R}^m$, on définit

$$A \in \mathcal{M}_{m,n}(\mathbb{R}) \mapsto |||A|||_{d,a} = \sup_{x\neq 0} \frac{\|Ax\|_a}{\|x\|_d}. \tag{11.1}$$

Il se trouve que $|||.|||_{d,a}$ est une norme sur $\mathcal{M}_{m,n}(\mathbb{R})$, qualifiée de "*déduite des* (ou *associée aux*, ou *subordonnée aux*) normes $\|.\|_d$ et $\|.\|_a$. Nous utilisons la notation $|||.|||_{d,a}$, avec des triples barres, pour bien distinguer ce qui se passe sur $\mathcal{M}_{m,n}(\mathbb{R})$ de ce qui se passe sur les espaces $\mathbb{R}^n$ et $\mathbb{R}^m$, pour lesquels nous réservons la notation avec les doubles barres. Il est clair que $|||A|||_{d,a}$ mesure la dilatation maximale de longueur d'un vecteur $x \in \mathbb{R}^n$ transformée dans $\mathbb{R}^m$ par l'application A ; évidemment ce facteur de dilatation dépend des outils de mesure $\|.\|_d$ et $\|.\|_a$ choisis sur $\mathbb{R}^n$ et $\mathbb{R}^m$. Certains auteurs vont même jusqu'à insister sur ces distinguos en adoptant la notation $|||A|||_{d\to a}$. L'homogénéité des normes, à savoir $\|\lambda x\| = |\lambda| \times \|x\|$ pour un réel λ, conduit facilement aux formes alternatives suivantes de $|||A|||_{d,a}$:

$$|||A|||_{d,a} = \sup_{\|x\|_d=1} \|Ax\|_a \;; \tag{11.2}$$

$$|||A|||_{d,a} = \sup_{\|x\|_d\leqslant 1} \|Ax\|_a \,. \tag{11.3}$$

Avec les choix des normes ℓ_p, nous avons donc une famille de normes $|||.|||_{p,q}$ sur $\mathcal{M}_{m,n}(\mathbb{R})$ définie de la manière suivante :

$$|||A|||_{p,q} = \sup_{x\neq 0} \frac{\|Ax\|_q}{\|x\|_p} = \sup_{\|x\|_p=1} \|Ax\|_q \;= \sup_{\|x\|_p\leqslant 1} \|Ax\|_q \,. \tag{12}$$

Au vu de la chaîne d'inégalités (7) sur les normes $\|.\|_r$, il est clair que $|||.|||_{p,q} \leqslant |||.|||_{\infty,1}$ pour tout p et q. C'est déjà un point à noter pour cette norme $|||.|||_{\infty,1}$, qui va jouer un rôle plus important que les autres dans la suite.

Suivant les domaines d'application, les normes $|||A|||_{p,q}$ à calculer ne sont pas forcément les mêmes. Par exemple (*cf.* [10]) : on a besoin d'accéder à $|||.|||_{2,4}$ pour faire la distinction entre les états intriqués et séparables dans la théorie de l'informatique quantique ; la norme $|||.|||_{2,q}$ avec q pair et supérieur ou égal à 4 est directement liée au développement en petits

ensembles dans la théorie des graphes et à la dite conjecture des jeux uniques en théorie de complexité informatique.

Il arrive que dans la littérature (anglo-saxonne notamment) on appelle "p-norme matricielle de HÖLDER" la norme que nous avons définie comme $|||A|||_{p,p}$.

On a déjà vu que le calcul de $\|x\|_p$ n'est pas toujours facile, alors oublions d'imaginer que le calcul effectif de $|||A|||_{p,q}$ est facile, sauf pour quelques cas particuliers de p et q. Voici brièvement quelques faits établis à propos de ces calculs.

Fait 1. *Le calcul de $|||A|||_{p,q}$ est un problème de maximisation d'une fonction convexe sur un ensemble convexe* (et donc a priori bien plus difficile que celui de la minimisation d'une fonction convexe sur un ensemble convexe).

En effet, les propriétés de base d'une norme $\|.\|$ font que

$$\|A(\alpha x + (1-\alpha)y)\| \leqslant \alpha \|Ax\| + (1-\alpha)\|Ay\|$$

pour x, y quelconques et $\alpha \in [0,1]$. Par ailleurs, la boule unité $B_p(0,1) = \left\{x : \|x\|_p \leqslant 1\right\}$ est un convexe compact. Donc, au vu de la troisième expression dans (12), le calcul de $|||A|||_{p,q}$ est celui du maximum de la fonction convexe $x \mapsto \|Ax\|_q$ sur le convexe compact $B_p(0,1)$.

Fait 2. *Pour déterminer le maximum de $\|Ax\|_q$ sur $B_p(0,1)$, il suffit de calculer le maximum de $\|A\overline{x}\|_q$ sur les points extrémaux $\overline{x}$ de $B_p(0,1)$.*

Rappelons qu'un *point extrémal* d'un convexe C est un point de C qui ne peut s'écrire comme une combinaison convexe (ou barycentrique à coefficients strictement positifs) de 2 points distincts de C ; dans le cas où C est un polytope, c'est simplement un *sommet*. D'après un célèbre théorème de H. MINKOWSKI, lorsque C est un convexe compact, tout point de C est une combinaison convexe de points extrémaux de C ([7, page 42]). Considérons donc $x \in B_p(0,1)$ écrit comme combinaison convexe de points extrémaux $\overline{x}_i$ de $B_p(0,1)$, bref $x = \sum_{i=1}^k \alpha_i \overline{x}_i$, avec $\alpha_i \geqslant 0$ et $\sum_{i=1}^k \alpha_i = 1$. Grâce à la convexité de la fonction $x \mapsto \|Ax\|_q$, on a

$$\begin{array}{rcl} \|Ax\|_q & \leqslant & \sum_{i=1}^k \alpha_i \|A\overline{x}_i\|_q \\ & \leqslant & \max\limits_{\overline{x} \text{ point extrémal}} \|A\overline{x}\|_q . \end{array}$$

Ainsi $\max_{\|x\|_p \leqslant 1} \|Ax\|_q = \max_{\overline{x} \text{ point extrémal}} \|A\overline{x}\|_q$.

Le résultat exprimé en Fait 2 n'apporte rien de nouveau dans le calcul de $|||A|||_{p,q}$ lorsque $1 < p < +\infty$ puisque dans ce cas tous les points frontières de $B_p(0,1)$, à savoir ceux vérifiant $\|x\|_p = 1$, sont exactement les points extrémaux de $B_p(0,1)$. Mais la donne change lorsque $B_p(0,1)$ est un polytope, pour $p = 1$ ou $p = +\infty$. Voici quelques exemples de calculs explicites de $|||A|||_{p,q}$:

- Lorsque $p = 1$, la boule unité $B_1(0,1)$ étant un polytope à $2n$ sommets $\pm e_i = (0, 0, ..., \pm 1, ..., 0)$, $i = 1, ..., n$,

$$\left\{ \begin{array}{c} |||A|||_{1,q} = \max_{j=1,...,n} \|A_{*j}\|_q , \\ \text{où } A_{*j} \text{ est la } j\text{-ème colonne de } A ; \end{array} \right. \tag{13}$$

en particulier

$$|||A|||_{1,1} = \max_{j=1,...,n} \left(\sum_{i=1}^{m} |a_{ij}| \right) \text{ (sommes sur les colonnes)} ; \tag{14}$$

$$|||A|||_{1,\infty} = \max_{\substack{i=1,...,m \\ j=1,...,n}} (|a_{ij}|) . \tag{15}$$

- Lorsque $p = +\infty$, la boule unité $B_\infty(0,1) = [-1,1]^n$ étant un polytope qui a 2^n sommets qui sont $(\pm 1, \pm 1, ..., \pm 1, ..., \pm 1)$, $i = 1, ..., n$,

$$|||A|||_{\infty,q} = \max_{\overline{x}=(\pm 1,...,\pm 1)} \|A\overline{x}\|_q ;$$

cela n'aide pas beaucoup car il y a 2^n sommets à explorer. Voici un cas particulier qui se simplifie tout de même :

$$|||A|||_{\infty,\infty} = \max_{i=1,...,m} \left(\left| \sum_{j=1}^{n} \pm a_{ij} \right| \right)$$

$$= \max_{i=1,...,m} \left(\sum_{j=1}^{n} |a_{ij}| \right) \text{ (sommes sur les lignes).} \tag{16}$$

Fait 3. *Le calcul de $|||A|||_{p,q}$ peut être vu comme un problème de maximisation bilinéaire sur le produit cartésien de 2 boules ℓ_r.*

Constatant que $\|Ax\|_q = \sup_{\|y\|_{q^*} \leqslant 1} \langle Ax, y \rangle$ (en utilisant (5)), nous reformulons $|||A|||_{p,q}$ comme suit :

$$|||A|||_{p,q} = \sup_{\|x\|_p \leqslant 1, \|y\|_{q^*} \leqslant 1} \langle Ax, y \rangle . \tag{17}$$

Il s'agit ici de maximiser la forme bilinéaire $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mapsto \langle Ax, y \rangle$ sur le produit cartésien des 2 boules $B_p(0,1)$ et $B_{q^*}(0,1)$. Cette expression (17) est une formulation alternative fort intéressante de $|||A|||_{p,q}$ mais, pour autant, cela ne rend pas son calcul effectif plus facile.

Une conséquence directe de (17) néanmoins. Remarquant que $\langle Ax, y \rangle = \langle x, A^T y \rangle$, nous pouvons réécrire (17) en

$$\sup_{\|x\|_p \leqslant 1, \|y\|_{q^*} \leqslant 1} \langle Ax, y \rangle = \sup_{\|y\|_{q^*} \leqslant 1, \|x\|_p \leqslant 1,} \left\langle A^T y, x \right\rangle , \tag{17^T}$$

et nous y reconnaissons dans le second membre $\left|||A^T|||\right|_{q^*,p^*}$. Autrement dit, nous avons démontré que

$$|||A|||_{p,q} = \left|||A^T|||\right|_{q^*,p^*} . \tag{18}$$

Résultat de va-et-vient curieux et intéressant. C'est exactement ce qui s'est passé entre (14) et (16), dans lesquels $p = q = 1$, et donc $p_* = q_* = +\infty$.

Fait 4. *Le calcul de $|||A|||_{p,q}$ peut être vu comme un problème de maximisation quadratique sur le produit cartésien de 2 boules ℓ_r.*

Considérons la matrice symétrique à $m+n$ lignes et $m+n$ colonnes $\mathcal{A}$ structurée en blocs comme suit

$$\mathcal{A} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}.$$

Elle définit une forme quadratique sur $\mathbb{R}^m \times \mathbb{R}^n$ de la manière suivante :

$$(y,x) \in \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathcal{A}(y,x) = \langle Ax, y\rangle + \left\langle A^T y, x\right\rangle.$$

Ainsi, au vu de (17) et (17^T), nous avons

$$2\,|||A|||_{p,q} = \max_{(y,x)\in B_{q^*}(0,1)\times B_p(0,1)} \mathcal{A}(y,x). \tag{19}$$

La formulation ci-dessus est alléchante tant il est vrai qu'en Optimisation on aime bien des fonctions à maximiser (ou minimiser) qui sont quadratiques. Le hic ici est que la contrainte $(y,x) \in B_{q^*}(0,1) \times B_p(0,1)$ met en jeu une boule de $\mathbb{R}^m \times \mathbb{R}^n$ qui n'est pas euclidienne...., donc la difficulté de notre calcul initial de $|||A|||_{p,q}$ n'est que déplacée.

Exemple-Exercice 1. Pour la matrice $\mathcal{A}$ construite à partir de A comme plus haut, montrer que

$$|||\mathcal{A}|||_{p,p} = \max\left(|||A|||_{p,p}\,, |||A|||_{p^*,p^*}\right). \tag{20}$$

Parlant de maximisation de fonctions quadratiques sur des boules, voici un aparté qui règle le cas particulier du calcul de $|||A|||_{2,2}$. En effet, la formulation variationnelle (c'est-à-dire en termes de maximisation ou de minimisation de fonctions) de la plus grande valeur propre $\lambda_{\max}(M)$ d'une matrice symétrique M est comme suit :

$$\lambda_{\max}(M) = \sup_{x\neq 0} \frac{\langle Mx, x\rangle}{\|x\|_2} = \sup_{\|x\|_2=1} \langle Mx, x\rangle \;\; = \sup_{\|x\|_2\leqslant 1} \langle Mx, x\rangle\,. \tag{20}$$

Partant de $A \in \mathcal{M}_{m,n}(\mathbb{R})$, appliquons ce qui est au-dessus à la matrice symétrique A^TA (ou AA^T d'ailleurs) :

$$\begin{aligned} \lambda_{\max}(A^TA) &= \sup_{\|x\|_2\leqslant 1} \left\langle A^TAx, x\right\rangle \\ &= \sup_{\|x\|_2\leqslant 1} \langle Ax, Ax\rangle \\ &= \sup_{\|x\|_2\leqslant 1} \left(\|Ax\|_2\right)^2 = \left(\sup_{\|x\|_2\leqslant 1} \|Ax\|_2\right)^2. \end{aligned}$$

On y reconnaît le carré de $|||A|||_{2,2}$. Comme A^TA est (symétrique) semidéfinie positive, soit toutes ses valeurs propres sont positives ou nulles, on déduit de ce qui est au-dessus :

$$|||A|||_{2,2} = \sqrt{\lambda_{\max}(A^TA)}. \tag{21}$$

Les valeurs propres de A^TA sont toutes réelles et positives (car A^TA est symétrique et semidéfinie positive) ; leurs racines carrées s'appellent les *valeurs singulières* de A. Ici, avec $|||A|||_{2,2}$, nous avons donc la plus grande valeur singulière de A.

Dans un contexte à peine plus général,

$$|||A|||_{2,q} = \sqrt{\sup_{\|x\|_{q^*}\leqslant 1} \langle A^T A x, x\rangle}.$$

Fait 5. *Simplification dans le cas particulier où $m = n$ et $A \in \mathcal{M}_{n,n}(\mathbb{R})$ est symétrique semidéfinie positive.*

Dans ce cas, nous disons que

$$|||A|||_{p,p^*} = \sup_{\|x\|_p\leqslant 1} \langle Ax, x\rangle . \tag{22}$$

Un exemple est celui où $p = +\infty$; cela donne

$$|||A|||_{\infty,1} = \sup_{\|x\|_\infty\leqslant 1} \langle Ax, x\rangle . \tag{22$_\infty$}$$

La formulation est encore une maximisation de fonction quadratique sur une boule ℓ_p, mais la difficulté du calcul reste entière, par exemple avec $p = +\infty$ où $B_\infty(0,1) = [-1,1]^n$ avec ses 2^n sommets qui font exploser la combinatoire du problème.

La démonstration de (22) est aisée et agréable à suivre, faisons-là rapidement. Nous partons de la formulation de $|||A|||_{p,p^*}$ telle que dérivée de (17) :

$$|||A|||_{p,p^*} = \sup_{\|x\|_p\leqslant 1,\|y\|_p\leqslant 1} \langle Ax, y\rangle . \tag{23}$$

Une première inégalité, valable pour toute matrice A, qu'elle soit semidéfinie positive ou pas, vient de (23) : en choisissant de faire $y = x$ dans le membre de droite, on obtient

$$|||A|||_{p,p^*} \geqslant \sup_{\|x\|_p\leqslant 1} \langle Ax, x\rangle . \tag{24a}$$

Faisons à présent le changement de variables linéaire suivant : $u = \frac{x+y}{2}, v = \frac{x-y}{2}$. Comme A est symétrique, $\langle Au, u\rangle - \langle Av, v\rangle = \langle Ax, y\rangle$. Par ailleurs, lorsque $\|x\|_p \leqslant 1$ et $\|y\|_p \leqslant 1$, il en est de même de u et v : $\|u\|_p \leqslant 1$ et $\|v\|_p \leqslant 1$. En conséquence,

$$\begin{aligned}\sup_{\|x\|_p\leqslant 1,\|y\|_p\leqslant 1} \langle Ax, y\rangle &\leqslant \sup_{\|u\|_p\leqslant 1,\|v\|_p\leqslant 1} [\langle Au, u\rangle - \langle Av, v\rangle] \\ &\leqslant \sup_{\|u\|_p\leqslant 1} \langle Au, u\rangle - \inf_{\|v\|_p\leqslant 1} \langle Av, v\rangle .\end{aligned}$$

La semidéfinie positivité de A fait que $\langle Av, v\rangle \geqslant 0$ pour tout v. Il vient donc de l'inégalité au-dessus

$$\sup_{\|x\|_p\leqslant 1,\|y\|_p\leqslant 1} \langle Ax, y\rangle \leqslant \sup_{\|u\|_p\leqslant 1} \langle Au, u\rangle . \tag{24b}$$

En combinant (24a) et (24b), on obtient bien la formulation (22) de $|||A|||_{p,p^*}$.

Exemple-Exercice 2 (tiré de [2]). Considérons une matrice $H = [h_{ij}] \in \mathcal{M}_{n,n}(\mathbb{R})$ de HADAMARD (ou de SYLVESTER & HADAMARD), c'est-à-dire dont tous les coefficients sont ± 1 et dont les lignes sont toutes orthogonales entre elles ; en résumé, une matrice carrée $H = [h_{ij}]$ vérifiant :

$$h_{ij} = \pm 1 \text{ pour tout } i, j \,;\; H \times H^T = nI_n. \tag{25}$$

Ce type de matrices intervient parfois comme exemple ou contre-exemple dans les questions qui nous préoccupent. D'une manière claire, $|||H|||_{1,1} = |||H|||_{\infty,\infty} = n$, mais on peut aller plus loin. En effet, les inégalités fines entre les normes $\|.\|_{p_1}$ et $\|.\|_{p_2}$ dans $\mathbb{R}^n$ lorsque $p_2 \geqslant p_1$ (des améliorations de (7)) conduisent facilement au résultat suivant :

$$|||H|||_{p,p} = n^{1/p} \text{ lorsque } 1 \leqslant p \leqslant 2.$$

Comme $|||H|||_{p,p} = \left|||H^T|||\right|_{p^*,p^*}$ (voir (18)) et que H^T est aussi de HADAMARD dès que H l'est, la relation au-dessus est généralisée en :

$$|||H|||_{p,p} = \max\left(n^{1/p}, n^{1/p^*}\right) \text{ lorsque } p \geqslant 1. \tag{26}$$

Aparté. *Pour les fans de la mesure de la complexité des calculs, voici ce qu'on peut dire du calcul de* $|||A|||_{p,q}$.

Comme signalé dans le Fait 1, la calcul de $|||A|||_{p,q}$ est un problème d'optimisation difficile, car une convexité (celle de la fonction-objectif) est "à rebours" ; de ce fait les méthodes numériques d'optimisation habituelles sont susceptibles de conduire à un point de maximisation locale de $\|Ax\|_q$ sur $B_p(0,1)$, alors que c'est la maximisation globale (ou absolue) qui est en jeu ici.

Du point de vue complexité algorithmique ("ou complexité computationnelle" dans le jargon spécialisé), le calcul de $|||A|||_{p,q}$ avec $p \neq q$ est jugé "NP-difficile" ; même chose pour le calcul de $|||A|||_{p,p}$ mis à part les cas particuliers (signalés plus haut) où $p = 1, 2$ ou $+\infty$ (*cf.* [6], [10], [11], [15]).

2. En route vers l'inégalité de GROTHENDIECK

Nous avons vu au paragraphe précédent le caractère "extrémal" et polytopal des normes $\|.\|_1$ et $\|.\|_\infty$. Ces particularités induisent des difficultés additionnelles dans le calcul de la norme triple $|||.|||_{\infty,1}$ (voir [15] par exemple), laquelle norme triple est d'ailleurs la plus grande des normes $|||.|||_{p,q}$ considérées. Mais c'est précisément cette norme triple $|||.|||_{\infty,1}$ qui est intéressante du point de vue mathématique et dans certaines applications. Elle est justement l'un des ingrédients de l'inégalité de GROTHENDIECK, laquelle est explorée plus bas. Rappelons tout d'abord ce que nous avons vu au sujet de $|||A|||_{\infty,1}$:

$$|||A|||_{\infty,1} = \begin{cases} \sup_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_\infty} = \sup_{\|x\|_\infty=1} \|Ax\|_1 \;\; = \sup_{\|x\|_\infty\leqslant 1} \|Ax\|_1 \;; \\ \sup_{\|x\|_\infty\leqslant 1,\ \|y\|_\infty\leqslant 1} \langle Ax, y\rangle \;; \\ \sup_{\|x\|_\infty=1,\ \|y\|_\infty=1} \langle Ax, y\rangle \;; \\ \sup_{\substack{x_i=\pm 1,\\ y_j=\pm 1}} \left(\sum_{i=1}^m \sum_{j=1}^n a_{ij} x_i y_j\right). \end{cases} \tag{27}$$

L'autre ingrédient dans l'inégalité de GROTHENDIECK est défini comme suit. Considérons m vecteurs unitaires (euclidiens) $u^{(i)}$ et n vecteurs unitaires (euclidiens) $v^{(j)}$ de $\mathbb{R}^d$ (ou, plus généralement, m vecteurs unitaires $u^{(i)}$ et n vecteurs unitaires $v^{(j)}$ d'un espace de HILBERT $(H,\ \langle .,.\rangle)$) et formons le scalaire $\sum_{i=1}^m \sum_{j=1}^n a_{ij} \left\langle u^{(i)}, v^{(j)} \right\rangle$. Ceci n'est autre que la généralisation de l'expression entre parenthèses dans la $4^{ème}$ ligne de (27) ; dans le cas de cette dernière ligne, les $u^{(i)} = x_i = \pm 1$ et $v^{(j)} = y_j = \pm 1$ sont des "vecteurs" unitaires de $\mathbb{H} = \mathbb{R}^1$. La quantité considérée dans l'inégalité de GROTHENDIECK est définie comme

$$|||A|||_G = \sup_{\substack{\left\|u^{(i)}\right\|=1 \text{ pour tout } i=1,..,m \\ \left\|v^{(j)}\right\|=1 \text{ pour tout } j=1,..,n \\ u^{(i)} \text{ et } v^{(j)} \text{ dans } H \\ (H,\langle .,.\rangle) \text{ espace de Hilbert}}} \left(\sum_{i=1}^m \sum_{j=1}^n a_{ij} \left\langle u^{(i)}, v^{(j)} \right\rangle \right). \tag{28}$$

Le supremum au-dessus est bien une quantité finie puisque, grâce à l'inégalité de CAUCHY & SCHWARZ, $\sum_{i=1}^m \sum_{j=1}^n a_{ij} \left\langle u^{(i)}, v^{(j)} \right\rangle \leqslant \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|$.

Il se trouve que $|||.|||_G$ est une norme sur $\mathcal{M}_{m,n}(\mathbb{R})$, ce qui, d'ailleurs, est facile à vérifier directement. D'après ce que nous avons expliqué plus haut (choix de $\mathbb{H} = \mathbb{R}^1$), $|||A|||_G \geqslant |||A|||_{\infty,1}$. Plus intéressant et fondamental est que $|||.|||_G$ est majorée par un multiple de $|||.|||_{\infty,1}$ et **la constante multiplicative K_G peut être rendue indépendante de m, n et de l'espace de HILBERT H dans lequels vivent les m vecteurs unitaires $u^{(i)}$ et les n vecteurs unitaires $v^{(j)}$. C'est l'essence même de l'inégalité de GROTHENDIECK.** Ecrivons cela de façon détaillée.

Théorème IG (A. GROTHENDIECK, 1953).

Soit $A \in \mathcal{M}_{m,n}(\mathbb{R})$ vérifiant $|||A|||_{\infty,1} \leqslant 1$. Alors, pour tout espace de HILBERT *$(H,\langle .,.\rangle)$, et tout jeu de m vecteurs unitaires $u^{(i)}$ et n vecteurs unitaires $v^{(j)}$ de H, on a*

$$\left| \sum_{i=1}^m \sum_{j=1}^n a_{ij} \left\langle u^{(i)}, v^{(j)} \right\rangle \right| \leqslant K_{G.} \tag{29}$$

Cette constante (universelle) K_G ne dépend pas de m et n, ni de l'espace H (de sa dimension notamment) dans lequel vivent les vecteurs unitaires $u^{(i)}$ et $v^{(j)}$.

Dit autrement,

$$\begin{cases} |||A|||_G \leqslant K_G \ |||A|||_{\infty,1} \ \textit{pour tout } A \in \mathcal{M}_{m,n}(\mathbb{R}), \\ \qquad K_G \ \textit{constante indépendante} \text{ de } m, n. \end{cases} \tag{30}$$

Voici quelques commentaires sur l'hypothèse et la conclusion de ce Théorème IG.

- L'hypothèse $|||A|||_{\infty,1} \leqslant 1$, vue sous le format

$$|||A|||_{\infty,1} = \sup_{\substack{x_i=\pm 1, \\ y_j=\pm 1}} \left(\sum_{i=1}^m \sum_{j=1}^n a_{ij} x_i y_j \right) \leqslant 1, \tag{31}$$

est piégeuse... En effet, on pourrait penser qu'avec un choix judicieux de $x_i = \pm 1$ et $y_j = \pm 1$ on pourrait faire $a_{ij}x_ix_j = |a_{ij}|$, de sorte que $\sup_{\substack{x_i=\pm 1,\\ y_j=\pm 1}} \left(\sum_{i=1}^m \sum_{j=1}^n a_{ij}x_ix_j\right)$ soit égal à $\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|$. Il n'en est rien - avouons que nous nous sommes faits avoir dans un premier temps - , on a seulement

$$\sup_{\substack{x_i=\pm 1,\\ y_j=\pm 1}} \left(\sum_{i=1}^m \sum_{j=1}^n a_{ij}x_iy_j\right) \leqslant |||A|||_1 = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|\,. \tag{32}$$

La raison en est que les choix des $x_i = \pm 1$ et $y_j = \pm 1$ pour avoir $a_{ij}x_ix_j = |a_{ij}|$ (il y a $m+n$ conditions non linéaires simultanées) peuvent être incompatibles.

Exemple-Exercice 3. Prenons un exemple avec $A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, encore une matrice de Hadamard. Avoir $a_{11}x_1y_1 = |a_{11}|$ impose $x_1y_1 = -1$, soit x_1 et y_1 de signe contraire ; avoir $a_{12}x_1y_2 = |a_{12}|$ impose $x_1y_2 = -1$, soit x_1 et y_2 de signe contraire ; en conséquence, y_1 et y_2 devraient être de même signe. Par ailleurs, avoir $a_{21}x_2y_1 = |a_{21}|$ impose $x_2y_1 = -1$, avoir $a_{22}x_2y_2 = |a_{22}|$ impose $x_2y_2 = 1$; par conséquent, y_1 et y_2 devraient être de signes contraires. D'où une contradiction. On est donc assuré que pour cet exemple, $|||A|||_{\infty,1} < |||A|||_1$; de fait, $|||A|||_1 = 4$ tandis que $|||A|||_{\infty,1} = 2$.

- $|||.|||_1$ est certes une norme sur $\mathcal{M}_{n,n}(\mathbb{R})$ mais elle ne peut être de la famille $|||.|||_{p,p}$ puisque $|||I_n|||_{p,p} = 1$ alors que $|||I_n|||_1 = n$. Remplacer dans le Théorème IG l'hypothèse $|||A|||_{\infty,1} \leqslant 1$ par l'hypothèse plus forte $|||A|||_1 \leqslant 1$ est bien tentant... , on aurait alors facilement, grâce à l'inégalité de Cauchy & Schwarz : $|||A|||_G \leqslant |||A|||_1$ pour tout $A \in \mathcal{M}_{m,n}(\mathbb{R})$. Rien à voir avec le Théorème IG si ce n'est un air de ressemblance dans l'hypothèse et la conclusion.

- Une formulation équivalente de l'hypothèse du Théorème IG (voir sa formulation (31)) est

$$\left|\sum_{i=1}^m \sum_{j=1}^n a_{ij}x_iy_j\right| \leqslant \|x\|_\infty \times \|y\|_\infty \ \textit{pour tout } x \in \mathbb{R}^m \textit{ et } y \in \mathbb{R}^n. \tag{33}$$

C'est une hypothèse plus forte qu'il n'y paraît..., même $A = I_n$ ne la satisfait pas (prendre pour x et y des vecteurs dont toutes les coordonnées valent 1).

Dans la même lignée, une formulation équivalente de la conclusion du Théorème IG est : *Pour tout espace de* Hilbert *$(H,\langle .,.\rangle)$, et tout jeu de m vecteurs $u^{(i)}$ et n vecteurs $v^{(j)}$ de H, on a*

$$\left|\sum_{i=1}^m \sum_{j=1}^n a_{ij} \left\langle u^{(i)}, v^{(j)}\right\rangle\right| \leqslant K_G \max_i \left\|u^{(i)}\right\| \times \max_i \left\|v^{(j)}\right\|. \tag{34}$$

- Le calcul effectif de $|||A|||_G$ est éminemment difficile, c'est pour cela qu'une majoration comme celle du Théorème IG (voir sa formulation (30)) est prisée..., si possible avec la meilleure estimation possible de K_G. Le grand choix possible de vecteurs unitaires $u^{(i)}$

et $v^{(j)}$ dans la formulation de $|||A|||_G$ (voir son expression dans (28)) conduit à quelques minorations de K_G. A titre d'exemple, prenons $A = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix}$, une matrice déjà considérée plus haut, au coefficient $1/2$ près (Exemple-Exercice 3). Nous avons $|||A|||_{\infty,1} = 1$ et pouvons vérifier que $|||A|||_G \geqslant \sqrt{2}$, ce qui avec l'inégalité (30) conduira à $K_G \geqslant \sqrt{2}$. Considérons en effet $H = \mathbb{R}^2$ muni du produit scalaire habituel $\langle ., . \rangle$ et les vecteurs unitaires suivants : $u^{(1)} = (1, 0)$, $u^{(2)} = (0, 1)$, $v^{(1)} = \left(\sqrt{2}/2, \sqrt{2}/2\right)$, $v^{(2)} = \left(\sqrt{2}/2, -\sqrt{2}/2\right)$. Alors, $\sum_{i=1}^{2} \sum_{j=1}^{2} a_{ij} \left\langle u^{(i)}, v^{(j)} \right\rangle = \sqrt{2}$.

Le Théorème IG est un théorème d'*existence d'une constante* K (indépendante de m et n) telle que

$$|||A|||_G \leqslant K \ |||A|||_{\infty,1} \text{ pour tout } A \in \mathcal{M}_{m,n}(\mathbb{R}).$$

La "meilleure" de ces constantes, c'est-à-dire la plus petite, est celle notée K_G et appelée *constante de* Grothendieck. Malheureusement, contrairement à la plupart des constantes universelles (ou absolues) du monde mathématique (voir [3]), elle n'est pas exprimée sous forme d'une intégrale ou d'une somme de série, ce qui l'aurait rendue approximable numériquement avec la précision voulue. En se limitant à 3 décimales après la virgule des données numériques, ce que l'on sait est :

- K_G est assurément strictement inférieur à $1,783$ (meilleure majoration numérique connue à ce jour, due à J.-L. Krivine en 1979 ([12])).

- K_G est assurément strictement supérieur à $1,676$. Nous avons facilement démontré plus haut que $K_G \geqslant \sqrt{2} \backsimeq 1,414$; des auteurs démontrent que $K_G \geqslant \pi/2 \backsimeq 1,570$; la meilleure minoration numérique connue à ce jour est $K_G > 1,676$.

En résumé, **le meilleur encadrement numérique connu à ce jour de K_G est :**

$$\mathbf{1,676 < K_G < 1,783}. \tag{35}$$

Les démonstrations du Théorème IG existent et sont variées. N'allez-pas la voir dans l'article originel de Grothendieck ([5]) consacré essentiellement aux produits tensoriels d'espaces de Banach, vous allez vous y perdre... et ne pas reconnaître l'inégalité cherchée. Fort heureusement, les mathématiciens ont depuis "digéré" le résultat et rendu sa démonstration plus accessible ; citons les articles ou notes de cours suivants : [2], [4], [12], [13] [16, Exercice 26.6] ; d'ailleurs la constante universelle K (la majorante) exhibée dans ces démonstrations n'est pas toujours la même, elle va de 288 dans [16] (une première démonstration) à $\mathrm{sh}(\pi/2) \backsimeq 2,302$ dans [5] et [16] (une deuxième démonstration) et à $\pi/(2\ln(1+\sqrt{2})) \backsimeq 1,783$ dans [12]. On a d'ailleurs cru (conjecturé) pendant des années que $K_G = \pi/(2\ln(1+\sqrt{2}))$, constante (majorante) trouvée par Krivine dans sa publication de 1979 ; on sait depuis 2013 que $K_G < \pi/(2\ln(1+\sqrt{2}))$ sans pour autant proposer une expression numérique plus précise que celle déjà exhibée.

Une variante du Théorème IG consiste à ne considérer que des espace de Hilbert $(H, \langle ., . \rangle)$ dont la dimension est au plus d ; dans ce cas, la constante universelle $K_G(d)$ mise

en évidence pour l'inégalité analogue à (29) ou (30) est, bien sûr, plus petite que K_G. Par exmple, $K_G(2) = \sqrt{2} \backsimeq 1,414$ et $K_G(3) \leqslant 1,517$.

Une dernière remarque. Tout ce que nous avons présenté l'est dans un contexte réel ; si l'on remplace les nombres réels par les nombres complexes, il y a aussi une inégalité de Grothendieck, avec une constante universelle $K_G^{\mathbb{C}}$ forcément différente de $K_G^{\mathbb{R}}$. Mais comme l'écrivait une étudiante sur sa copie : *"Oh non pas dans $\mathbb{C}$... c'est assez complexe comme ça dans $\mathbb{R}$".*

Conclusion

L'inégalité de Grothendieck, dont les versions et présentations habituelles peuvent paraître déroutantes ou mystérieuses à un lecteur lambda, est présentée ici sous l'angle des normes matricielles. Cet aspect étant davantage étudié et connu des étudiants, nous espérons que notre manière de faire leur permettra de mieux appréhender cette inégalité de Grothendieck.

Références

1. A. Chambert-Loir et S. Fermigier, *Exercices pour l'agrégation. Analyse 3*. Masson (1996).

2. D. Dadush, *Geometric functional analysis*. Notes de cours téléchargeables à partir de son site web, Université d'Utrecht (Pays-Bas).

3. S. R. Finch, *Mathematical constants*. Cambridge University Press (2003).

4. S. Friedland, L-H Lim and J. Zhang, *An elementary and unified proof of Grothendieck's inequality*. L'Enseignement Mathématique (2) 64 (2018), $327 - 351$.

5. A. Grothendieck, *Résumé de la théorie métrique des produits tensoriels topologiques*. Bol. Soc. Math. Sao Paulo 8 (1953), $1-79$. Réimprimé dans Resenhas 2, n°4 (1996), $401 - 480$.

6. N. J. Higham, *Estimating the matrix p-norm* $|||A|||_{p,p}$. Numerische Mathematik, vol. 62 (1992), $539 - 555$.

7. J.-B. Hiriart-Urruty et C. Lemaréchal, *Fundamentals of convex analysis*. Grundlehren Text Editions, Springer (2001).

8. J.-B. Hiriart-Urruty et M. Pradel, *Les boules !* Quadrature 54 (2004), $8 - 11$.

9. M. Gaudioso and J.-B. Hiriart-Urruty, *Deforming* $\|.\|_1$ *into* $\|.\|_\infty$ *via polyhedral norms : a pedestrian approach.* SIAM Review 64, n°3 (2022), $713 - 727$.

10. L. Guth, D. Maldague and J. Urschel, *Estimating the matrix* $p \to q$ *norm.* ArXiv (nov. 2023).

11. J. M. Hendrickx and A. Olshevsky, *Matrix p-norms are NP-hard to approximate if* $p \neq 1, 2, \infty$. SIAM J. Matrix Anal. Appl. vol. 31, n°5 (2010), $2802 - 2812$.

12. J.-L. Krivine, *Constantes de Grothendieck et fonctions de type positif sur les sphères.* Advances in Math. 31 (1979), $16 - 30$.

13. J. Lindenstrauss and A. Pełczyński, *Absolutely summing operators in $\mathcal{L}^p$-spaces and their applications*. Studia Math. 29 (1968), 275 – 326.

14. Ph. Pajot, *La revanche d'un théorème oublié*. La Recherche 501 + 502 (juillet-août 2015).

15. J. Rohn, *Computing the norm $|||.|||_{\infty,1}$ is NP-hard.* Linear and Multilinear Algebra 47 (3), (2000), 195 – 204.

16. R. Vershynin, *High-dimensional probability. An introduction with applications in data science* (2024). Livre téléchargeable à partir de son site web, Université de Californie à Irvine (USA).

Addendum au moment de la correction des épreuves (par JBHU). La consultation d'anciens numéros de Quadrature me fait croiser, et me conduit à ajouter, une référence sur l'apport de Grothendieck en Analyse fonctionnelle ; la voici :

17. H. Queffélec, *Grothendieck et l'Analyse fonctionnelle*. Quadrature n°101 (2016), 11 – 18.

Remerciements

Nous remercions Djalil Chafaï, ancien étudiant de l'université Paul Sabatier de Toulouse, professeur d'université détaché à l'ENS Paris, de nous avoir signalé les références [2] et [16].

Annexe : Une démonstration du Théorème IG

Nous partons de l'hypothèse du Théorème IG, formulée de la manière suivante : Soit $A = [a_{ij}] \in \mathcal{M}_{n,m}(\mathbb{R})$ vérifiant

$$\left\{ \begin{array}{c} \left| \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} x_i y_j \right| \leqslant 1 \\ \text{dès que } |x_i| \leqslant 1 \text{ pour tout } i \text{ et } |y_j| \leqslant 1 \text{ pour tout } j. \end{array} \right. \qquad (\mathcal{H})$$

Notre objectif est de démontrer que pour tout jeu de n vecteurs unitaires X_i et m vecteurs unitaires Y_j d'un espace de Hilbert $(H, \langle ., . \rangle)$, on a

$$\left| \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} \langle X_i, Y_j \rangle \right| \leqslant \mathrm{sh}\left(\frac{\pi}{2}\right). \qquad (\mathcal{C})$$

Bref, la constante majorante K (supérieure à la constante universelle K_G de Grothendieck) exhibée par la démonstration sera $\mathrm{sh}(\pi/2)$. C'était d'ailleurs la constante majorante proposée dans l'article originel de Grothendieck ([5]). La "digestion" de la démonstration initiale par les auteurs de [13] et le cheminement proposé dans [1] sont les principaux ingrédients de ce qui va être présenté ici. Cette démonstration, au demeurant fort subtile, utilise les résultats et techniques de l'Analyse réelle classique (changements de variables dans des calculs d'intégrales multiples, fonctions circulaires réciproques, fonctions hyperboliques, intégrales usuelles comme $\int_{\mathbb{R}} e^{-t^2/2} dt = (2\pi)^{1/2}$).

Pour démontrer que $(\mathcal{H})$ implique $(\mathcal{C})$, on rajoute une étape intermédiaire : on montrera que $(\mathcal{H})$ implique $(\mathcal{I})$, puis que $(\mathcal{I})$ implique $(\mathcal{C})$, où $(\mathcal{I})$ désigne l'inégalité que voici :

$$\left|\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\arcsin\langle X_i, Y_j\rangle\right| \leqslant \frac{\pi}{2},\ \text{pour tous } X_i, Y_j \text{ de norme 1.} \qquad (\mathcal{I}\ )$$

La démonstration de $(\mathcal{I})$ se fait en plusieurs étapes : une étape préliminaire pour obtenir une superbe égalité (dite égalité de GROTHENDIECK), puis trois autres étapes montrant bien que $(\mathcal{I})$ implique $(\mathcal{C})$.

• **Etape préliminaire.** Pour $U, V \in \mathbb{R}^n$ de norme euclidienne 1 :

$$(\bigstar) \qquad \boxed{I := \int_{\mathbb{R}^n} \operatorname{sign}\langle U, X\rangle \operatorname{sign}\langle V, X\rangle \frac{e^{-\|X\|^2/2}}{(2\pi)^{n/2}} dX = \frac{2}{\pi}\arcsin\langle U, V\rangle.}$$

On reconnaît dans $\frac{e^{-\|X\|^2/2}}{(2\pi)^{n/2}}dX$ la mesure gaussienne sur $\mathbb{R}^n$, et on peut s'étonner ici de la "randomisation" d'un sujet qui ne l'est pas *a priori.*

Première constation facile : le résultat $(\star)$ est immédiat si U et V (unitaires) sont colinéaires, c'est-à-dire si $U = V$ ou $U = -V$; dans ce cas, on a bien le résultat annoncé puisque $\arcsin(1) = \pi/2$ et $\arcsin(-1) = -\pi/2$.

L'apparition de $\arcsin\langle U, V\rangle$ dans le résultat de l'évaluation de I a un côté mystérieux. Aussi, avant de donner une démonstration générale de $(\star)$, on va esquisser une preuve dans le cas bidimensionel.

Esquisse dans le cas où $n = 2$ (suivant [2, pages $3-4$]).

Considérons deux vecteurs unitaires $\overrightarrow{u}$ et $\overrightarrow{v}$ linéairement indépendants de $\mathbb{R}^2$. Lorsqu'on considère le vecteur (non nul) générique $\overrightarrow{X} \in \mathbb{R}^2$, ce qui importe pour évaluer $\operatorname{sign}\langle\overrightarrow{u}, \overrightarrow{X}\rangle$ ou $\operatorname{sign}\langle\overrightarrow{v}, \overrightarrow{X}\rangle$ n'est pas $\overrightarrow{X}$ lui-même mais le vecteur unitaire $\overrightarrow{n} = \frac{\overrightarrow{X}}{\|\overrightarrow{X}\|}$ le dirigeant : $\operatorname{sign}\langle\overrightarrow{u}, \overrightarrow{X}\rangle = \operatorname{sign}\langle\overrightarrow{u}, \overrightarrow{n}\rangle$ et $\operatorname{sign}\langle\overrightarrow{v}, \overrightarrow{X}\rangle = \operatorname{sign}\langle\overrightarrow{v}, \overrightarrow{n}\rangle$. Suivre maintenant sur la Figure 1 plus bas.

Images : (cinématique) $\overrightarrow{n}$ est supposé parcourir le cercle unité (de longueur 2π) à vitesse constante : ou bien (probabiliste) $\overrightarrow{n}$ est distribué uniformément sur le cercle unité (densité $1/\,2\pi$). Pour un $\overrightarrow{n}$ donné, les signes de $\langle\overrightarrow{u}, \overrightarrow{n}\rangle$ et de $\langle\overrightarrow{v}, \overrightarrow{n}\rangle$, et donc le produit $\operatorname{sign}\langle\overrightarrow{u}, \overrightarrow{n}\rangle$ $\operatorname{sign}\langle\overrightarrow{v}, \overrightarrow{n}\rangle$, sont "lisibles" sur la Figure 1 : $\langle\overrightarrow{u}, \overrightarrow{n}\rangle$ et $\langle\overrightarrow{v}, \overrightarrow{n}\rangle$ sont de même signe (1 ou -1), et donc $\operatorname{sign}\langle\overrightarrow{u}, \overrightarrow{n}\rangle$ $\operatorname{sign}\langle\overrightarrow{v}, \overrightarrow{n}\rangle = 1$, lorsque $\overrightarrow{n}$ se trouve sur les 2 morceaux de cercle opposés (en rouge) de longueur $\pi - \theta$ chacun (où θ désigne $\arccos\langle\overrightarrow{u}, \overrightarrow{v}\rangle$) ; $\langle\overrightarrow{u}, \overrightarrow{n}\rangle$ et $\langle\overrightarrow{v}, \overrightarrow{n}\rangle$ sont de signes contraires, et donc $\operatorname{sign}\langle\overrightarrow{u}, \overrightarrow{n}\rangle$ $\operatorname{sign}\langle\overrightarrow{v}, \overrightarrow{n}\rangle = -1$ lorsque $\overrightarrow{n}$ se trouve sur les 2 morceaux de cercle opposés (en bleu) de longueur θ chacun. "Sommons" le tout de manière à avoir

$$\begin{aligned}\frac{1}{2\pi}\left[2(\pi-\theta)\times 1 + 2\theta\times(-1)\right] &= 1 - \frac{2}{\pi}\theta \\ &= 1 - \frac{2}{\pi}\arccos\langle\overrightarrow{u}, \overrightarrow{v}\rangle.\end{aligned}$$

Comme $\arccos\langle\overrightarrow{u},\overrightarrow{v}\rangle = \frac{\pi}{2} - \arcsin\langle\overrightarrow{u},\overrightarrow{v}\rangle$, le résultat de la sommation au-dessus vaut bien $\arcsin\langle\overrightarrow{u},\overrightarrow{v}\rangle$.

Démonstration générale.

On commence par une série de changements de variables pour donner une forme plus sympathique à notre intégrale à calculer.

$\circ$ $\mathcal{B} = \{e_1, \dots, e_n\}$ désigne la base canonique de $\mathbb{R}^n$. Les vecteurs U et V étant de norme 1, choisissons comme changement de variable $X' = \Omega X$ où Ω est la matrice orthogonale par laquelle $\mathcal{B}$ a pour image la base orthonormée $\mathcal{B}' = \{e'_1 = U, e'_2, \dots, e'_n\}$ telle que $V \in \mathrm{Vect}\{e'_1, e'_2\}$. Pour y arriver deux rotations suffisent, Ω est bien orthogonale et donc $\frac{e^{-\|X\|^2/2}}{(2\pi)^{n/2}}dX = \frac{e^{-\|X'\|^2/2}}{(2\pi)^{n/2}}dX'$. Le vecteur $V \in \mathrm{Vect}\{e'_1, e'_2\}$ étant de norme 1 par construction, il va s'écrire $V = (\sin(\alpha), \cos(\alpha), 0, \dots, 0)$ dans $\mathcal{B}'$, où $\alpha \in]-\pi, \pi]$.

$\circ$ Quitte à remplacer V par $-V$ (ce qui laisse l'expression dans ($\bigstar$) invariante) on peut supposer que $\alpha \in [-\pi/2, \pi/2]$ et même $\alpha \in [0, \pi/2]$ (en remplaçant e'_2 par $-e'_2$, faire un dessin pour voir cela).

$\circ$ On fait ensuite le changement de variables suivant :

$$(x_1, x_2, x_3, \dots, x_n) = (r\cos(\theta), r\sin(\theta), x_3, \dots, x_n),\ r \in \mathbb{R}_+,\ \theta \in]-\pi, \pi],\ x_3, \dots, x_n \in \mathbb{R},$$

qui transforme I en

$$\begin{aligned}
I &= \int_{\mathbb{R}^n} \mathrm{sign}\langle U, X\rangle \,\mathrm{sign}\langle V, X\rangle \frac{e^{-\|X\|^2/2}}{(2\pi)^{n/2}}dX \\
&= \int_{\mathbb{R}^n} \mathrm{sign}(r\cos(\theta))\,\mathrm{sign}(r[\sin(\alpha)\cos(\theta) + \cos(\alpha)\sin(\theta)])\frac{e^{-r^2/2}e^{-x_3^2/2}\cdots e^{-x_n^2/2}}{(2\pi)^{n/2}} r\,dr\,d\theta\,dx_3\cdots dx_n \\
&= \left(\int_{-\pi}^{\pi} \mathrm{sign}(\cos(\theta))\,\mathrm{sign}(\sin(\alpha+\theta))d\theta\right)\left(\int_{\mathbb{R}_+} re^{-r^2/2}dr\right)\left(\int_{\mathbb{R}_+} e^{-t^2/2}dt\right)^{n-2}(2\pi)^{-n/2} \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{sign}(\cos(\theta))\,\mathrm{sign}(\sin(\alpha+\theta))d\theta := \frac{J}{2\pi}.
\end{aligned}$$

Enfin

$$\begin{aligned}
I &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{sign}(\cos(\theta))\,\mathrm{sign}(\sin(\alpha+\theta))d\theta \\
&= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \mathrm{sign}(\sin(\alpha+\theta))d\theta - \frac{1}{2\pi}\int_{\pi/2}^{3\pi/2} \mathrm{sign}(\sin(\alpha+\theta))d\theta \\
&= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \mathrm{sign}(\sin(\alpha+\theta))d\theta - \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \mathrm{sign}(\sin(\alpha+\phi+\pi))d\varphi, \quad \varphi = \theta - \pi, \\
&= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \mathrm{sign}(\sin(\alpha+\theta))d\theta + \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \mathrm{sign}(\sin(\alpha+\phi))d\varphi \\
&= \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \mathrm{sign}(\sin(\alpha+\theta))d\theta.
\end{aligned}$$

Comme $\alpha \in]0,\pi/2[,\ \theta \in]-\pi/2,\pi/2[$, $\sin(\alpha+\theta)>0$ équivaut à $-\alpha<\theta<\pi/2$, tandis que $\sin(\alpha+\theta)<0$ équivaut à $-\pi/2<\theta<-\alpha$; on a finalement

$$\begin{aligned} I &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{sign}(\cos(\theta))\operatorname{sign}(\sin(\alpha+\theta))d\theta = \frac{1}{\pi}\int_{-\pi/2}^{\pi/2}\operatorname{sign}(\sin(\alpha+\theta))d\theta \\ &= \frac{1}{\pi}\left(-\int_{-\pi/2}^{-\alpha} d\theta + \int_{-\alpha}^{\pi/2} d\theta\right) = \frac{2\alpha}{\pi} = \frac{2\alpha}{\pi} = \frac{2\arcsin(\sin(\alpha))}{\pi}, \quad \text{car } \alpha\in]0,\pi/2[, \\ &= \frac{2\arcsin(\langle U,V\rangle)}{\pi}, \quad \text{car } U=e_1, V=(\sin(\alpha),\cos(\alpha),0,\dots,0). \end{aligned}$$

soit ($\bigstar$), et fin de l'étape préliminaire.

• Deuxième étape : on montre que ($\mathcal{H}$) implique ($\mathcal{I}$)
Soient $A\in\mathcal{M}_{n,m}(\mathbb{R})$ vérifiant ($\mathcal{H}$), H un espace de Hilbert, et $X_1,\dots,X_n,Y_1,\dots,Y_m\in H$ de norme 1.
Avant de montrer que A vérifie ($\mathcal{I}$) il est bon d'observer qu'étant donné l'objectif, il ne coûte rien de supposer que $H=\mathrm{Vect}\{X_1,\dots,X_n,Y_1,\dots,Y_m\}$, autrement dit remplacer H par $\mathbb{R}^N$ (avec $N=n+m$) car deux espaces de Hilbert de même dimension (finie) sont isométriques.
∘ Soit $A\in\mathcal{M}_{n,m}(\mathbb{R})$ vérifiant ($\mathcal{H}$) et une famille de vecteurs $X_1,\dots,X_n,Y_1,\dots,Y_m\in\mathbb{R}^N$ de norme 1 ; alors, au vu de ($\bigstar$) :

$$\begin{aligned} \left|\sum_{i=1}^n\sum_{j=1}^m a_{i,j}\arcsin\langle X_i,Y_j\rangle\right| &= \left|\frac{\pi}{2}\sum_{i=1}^n\sum_{j=1}^m a_{i,j}\int_{\mathbb{R}^N}\operatorname{sign}\langle X_i,X\rangle\operatorname{sign}\langle Y_j,X\rangle\frac{e^{-\|X\|^2/2}}{(2\pi)^{n/2}}dX\right| \\ &\le \frac{\pi}{2}\int_{\mathbb{R}^N}\left|\sum_{i=1}^n\sum_{j=1}^m a_{i,j}\operatorname{sign}\langle X_i,X\rangle\operatorname{sign}\langle Y_j,X\rangle\right|\frac{e^{-\|X\|^2/2}}{(2\pi)^{n/2}}dX \end{aligned}$$

Mais A vérifiant ($\mathcal{H}$) on a :

$$\left|\sum_{i=1}^n\sum_{j=1}^m a_{i,j}\operatorname{sign}(\langle X_i,X\rangle)\operatorname{sign}(\langle Y_j,X\rangle)\right|\le 1, \qquad \forall\, X\in\mathbb{R}^N,$$

donc

$$\begin{aligned} \left|\sum_{i=1}^n\sum_{j=1}^m a_{i,j}\arcsin\langle X_i,Y_j\rangle\right| &\le \frac{\pi}{2}\int_{\mathbb{R}^N}\left|\sum_{i=1}^n\sum_{j=1}^m a_{i,j}\operatorname{sign}\langle X_i,X\rangle\operatorname{sign}\langle Y_j,X\rangle\right|\frac{e^{-\|X\|^2/2}}{(2\pi)^{n/2}}dX \\ &\le \frac{\pi}{2}\int_{\mathbb{R}^N}\frac{e^{-\|X\|^2/2}}{(2\pi)^{n/2}}dX = \frac{\pi}{2}\left(\int_{\mathbb{R}}\frac{e^{-t^2/2}}{(2\pi)^{1/2}}dt\right)^n = \frac{\pi}{2}. \end{aligned}$$

Ainsi l'inégalité ($\mathcal{I}$) est démontrée.

• **Troisième étape : on peut itérer $(\mathcal{I})$.**
Si $A \in \mathcal{M}_{n,m}(\mathbb{R})$ vérifie $(\mathcal{H})$ et donc (c'est le résultat de la seconde étape) $(\mathcal{I})$

$$\left|\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}\arcsin\langle X_i, Y_j\rangle\right| \leq \frac{\pi}{2},$$

alors elle vérifie aussi pour $X_1, \ldots, X_n, Y_1, \ldots, Y_m \in \mathbb{R}^N$ de norme 1 :

$$\left|\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}\arcsin^p\langle X_i, Y_j\rangle\right| \leq \left(\frac{\pi}{2}\right)^p, \qquad \forall\, p \in \mathbb{N}. \tag{$\mathcal{I}'$}$$

∘ Pour s'en convaincre, il suffit d'observer que si $A = [a_{ij}] \in \mathcal{M}_{n,m}(\mathbb{R})$ vérifie $(\mathcal{H})$, alors pour tous $s_1, \ldots, s_n, t_1, \ldots, t_m \in [-1, 1]$ la nouvelle matrice $B = [b_{i,j} = a_{i,j}s_it_j]$ vérifie encore $(\mathcal{H})$. En effet pour $s'_1, \ldots, s'_n, t'_1, \ldots, t'_m \in [-1, 1]$,

$$\left|\sum_{i=1}^{n}\sum_{j=1}^{m} b_{i,j}s'_it'_j\right| = \left|\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}(s_is'_i)(t_jt'_j)\right| \leq 1$$

car A vérifie $(\mathcal{H})$ et $|s_is'_i| \leq 1$ pour tout i et $|t_jt'_j| \leq 1$ pour tout j.
Donc (voir $2^{ème}$ étape) $B = [b_{i,j} = a_{i,j}s_it_j]$ vérifie $(\mathcal{I})$:

$$\left|\sum_{i=1}^{n}\sum_{j=1}^{m} b_{i,j}\arcsin\langle X_i, Y_j\rangle\right| = \left|\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}s_it_j\arcsin\langle X_i, Y_j\rangle\right| \leq \frac{\pi}{2}.$$

Autrement dit la matrice $A_1 = \left[\frac{2}{\pi}a_{i,j}\arcsin\langle X_i, Y_j\rangle\right]$ vérifie $(\mathcal{H})$ et donc $(\mathcal{I})$:

$$\left|\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}\arcsin^2\langle X_i, Y_j\rangle\right| \leq \left(\frac{\pi}{2}\right)^2.$$

En réitérant cet argument il vient que pour toute matrice $A \in \mathcal{M}_{n,m}(\mathbb{R})$ vérifiant $(\mathcal{H})$:

$$\left|\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}\arcsin^p\langle X_i, Y_j\rangle\right| \leq \left(\frac{\pi}{2}\right)^p, \qquad \forall\, p \in \mathbb{N}.$$

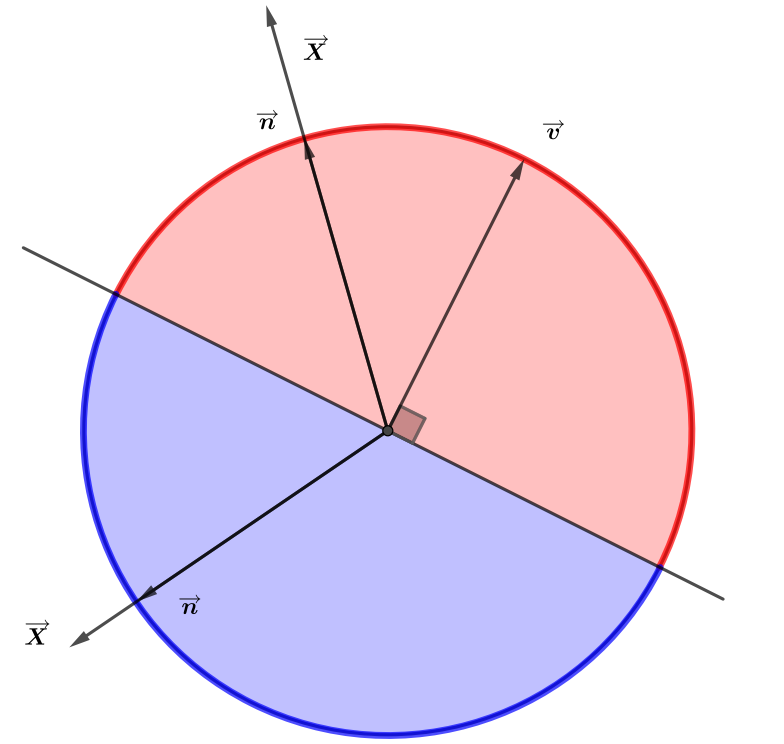
• **Quatrième étape : vers la conclusion.**
Soit $A \in \mathcal{M}_{n,m}(\mathbb{R})$ vérifiant $(\mathcal{H})$ et $X_1, \ldots, X_n, Y_1, \ldots, Y_m \in \mathbb{R}^N$ de norme 1. Ces vecteurs étant de norme 1 :

$$\langle X_i, Y_j\rangle = \sin(\arcsin\langle X_i, Y_j\rangle) = \sum_{p=0}^{+\infty} \frac{(-1)^p}{(2p+1)!}\arcsin^{2p+1}\langle X_i, Y_j\rangle,$$

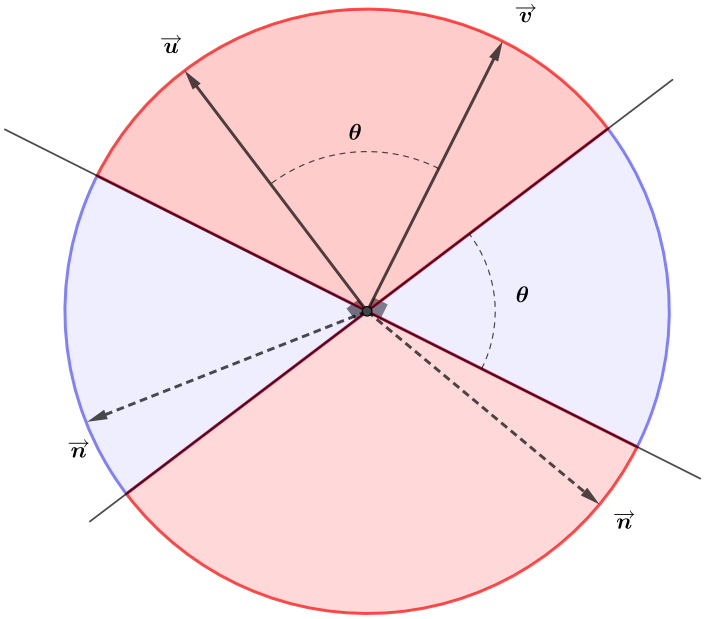
soit

$$\begin{aligned}\left|\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}\langle X_i, Y_j\rangle\right| &= \left|\sum_{p=0}^{\infty} \frac{(-1)^p}{(2p+1)!}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}\arcsin^{2p+1}\langle X_i, Y_j\rangle\right)\right| \\ &\leq \sum_{p=0}^{\infty} \frac{1}{(2p+1)!}\underbrace{\left|\sum_{i=1}^{n}\sum_{j=1}^{m} a_{i,j}\arcsin^{2p+1}\langle X_i, Y_j\rangle\right|}_{\leq\left(\frac{\pi}{2}\right)^{2p+1}\ \textit{étape 3,}} \\ &\leq \sum_{p=0}^{\infty} \frac{(\pi/2)^{2p+1}}{(2p+1)!} = \sinh(\pi/2).\end{aligned}$$

CQFD. ■



— : zone où $\langle \vec{n}, \vec{v} \rangle \geq 0$. — : zone où $\langle \vec{n}, \vec{v} \rangle \leq 0$.



— : zone où $\langle \vec{n}, \vec{u} \rangle$ et $\langle \vec{n}, \vec{v} \rangle$ sont de même signe

— : zone où $\langle \vec{n}, \vec{u} \rangle$ et $\langle \vec{n}, \vec{v} \rangle$ sont de signe contraire

FIGURE 1 – Visualisation du calcul quand $n = 2$