Mathématiques Expérimentales et Calcul Formel

Notes du modal de Mathématiques Expérimentales à l'École polytechnique[1]

Version 2017-2018

Bruno Salvy

1. Fondées initialement sur un cours de Calcul Formel par Alin Bostan et Bruno Salvy dans la fin des années 2000. Une première version d'une partie des chapitres avait été rédigée par les brillants élèves de l'année 2008–2009 : Samuel Baumard, Roland Casalis, Sary Drappeau, Pierre Lairez, Bruno Le Floch, Henri Guenancia, Nicolas Mascot, Arnaud de Mesmay, Michaël Monereau, Aurel Page, Guillaume Scerri, Olivier Taïbi. Ensuite, ce cours s'est adapté au format du Modal et aux élèves de l'École polytechnique, avec des sections plus ou moins détaillées selon les années.

Table des matières

COURS 0

Présentation du cours

Résumé

Les mathématiques expérimentales exploitent les capacités de calcul exact ou à grande précision des systèmes de calcul formel pour explorer des problèmes mathématiques. La démarche expérimentale consiste à calculer des approximations à grande précision (par exemple en nombre de termes de développements en série, ou en nombre de décimales), puis à repérer plus ou moins automatiquement des motifs, et ainsi former des conjectures, et enfin à prouver ces conjectures. Pour cela, le calcul formel doit être utilisé avec doigté, pour éviter les questions d'indécidabilité qui rodent aux alentours.

1. Mathématiques expérimentales

Les expériences ont toujours été un ingrédient important de la découverte en mathématique. Euler, Gauss, Ramanujan et de nombreux autres calculaient énormément d'exemples à de grandes précisions pour raffiner leur intuition des phénomènes qu'ils étudiaient. Le mathématicien hongrois George Pólya disait quant à lui *"Finished mathematics consists of proofs, but mathematics in the making consists of guesses"*. Aujourd'hui, les systèmes de calcul formel allègent énormément la charge de calcul, et permettent de ce fait d'atteindre des exemples qui ne soient pas de simples jouets. Cette possibilité change la donne. Ainsi, les records récents de calculs de décimales de pi exploitent une formule trouvée expérimentalement en 1995 ; une vieille conjecture de combinatoire sur l'énumération de certaines marches dans le quart de plan a été résolue récemment par des expériences de grande taille, menant d'abord à une conjecture sur l'algébricité d'une série, puis à une preuve informatisée.

Ce cours développe l'utilisation d'approximations à grande précision comme structure de donnée intermédiaire pour représenter des objets exacts. Nous présentons les principaux outils permettant de passer d'équations, sommes ou intégrales à des bonnes approximations (des centaines de coefficients exacts ou de décimales selon les cas), puis à l'inverse, à partir d'approximations, de reconstruire des équations qu'elles résolvent de manière approchée ou des "formules exactes" qu'elles approchent. Les approximations à grande précision sont souvent suffisantes pour que les expressions reconstruites soient elles-mêmes non plus des approximations, mais des formules exactes. Nous présentons pour conclure des algorithmes permettant de prouver des identités entre nombres algébriques, séries hypergéométriques ou plus généralement différentiellement finies, ou de répondre à des questions sur des solutions de systèmes polynomiaux. Ceci fournit la base des outils de preuve que propose le calcul formel.

Le cours commence par présenter l'approche expérimentale sur les séries, où les calculs sont approchés, mais exacts, avant de passer aux valeurs numériques, où les preuves de convergence sont plus délicates, et simplifiées si l'on a étudié d'abord le cas des séries. La dernière partie du cours est consacrée aux outils de preuve. Le plan d'ensemble est donc :

1. Outils d'expérimentation sur les séries
   Séance 1. Approximations : Calculs de séries par itération de Newton symbolique
   Séance 2. Conjectures : Approximants de Padé et de Padé-Hermite
2. Outils d'expérimentation sur les constantes
   Séance 3. Accélération de convergence et fractions continues
   Séance 4. Vecteurs courts dans les réseaux euclidiens et applications
3. Calcul exact et preuve
   Séance 5. Résultants
   Séance 6. Séries différentiellement finies
   Séance 7. Identités hypergéométriques : l'algorithme de Zeilberger
   Séance 8. Bases de Gröbner

On peut aussi lire les quatre premiers chapitres avec un autre plan en tête, qui suit mieux les trois étapes de la démarche expérimentale en mathématiques :

1. Calcul d'approximations à grande précision
   Séance 3. Accélération de convergence et fractions continues
   Séance 1. Calculs de séries par itération de Newton symbolique
2. Aides à la conjecture
   Séance 2. Approximants de Padé et de Padé-Hermite
   Séance 4. Vecteurs courts dans les réseaux euclidiens et applications
3. Calcul exact et preuve

Au delà d'une compréhension des outils théoriques, le cours insistera beaucoup sur leur pratique et c'est pourquoi une bonne partie du cours se déroulera sur machine face à des expériences. Chaque chapitre comprend donc un cours, suivi d'un sujet de TP à effectuer sur ordinateur à l'aide d'un système de calcul formel du choix de l'élève ; le système Maple sera utilisé par ceux qui n'ont pas d'autre préférence.

En plus des cours et des TPs, chaque élève, éventuellement en binôme, devra monter une expérience sur un sujet mathématique de son choix, qui sera ensuite présentée devant l'ensemble des élèves.

2. Calcul formel

Pour conclure cette introduction, voici un petit panorama de ce qu'est le calcul formel et de ce qu'il permet de calculer ou non.

2.1. Les limites de la machine. D'une certaine manière, le calcul formel est fondé sur une contrainte d'origine logique.

THÉORÈME 1 (Richardson-Matiyasevich). *Dans la classe des expressions formées à partir d'une variable X et de la constante 1 par les opérations d'anneau $+, -, \times$ et la composition avec les fonctions* sin *et la valeur absolue $|\cdot|$, le test d'équivalence à 0 est indécidable.*

Autrement dit, il n'existe pas d'algorithme permettant pour toute expression de cette classe de déterminer en temps fini si elle vaut 0 ou non. Plus généralement tout test d'égalité peut bien entendu se ramener à tester l'égalité à zéro dès que la soustraction existe. Cette limitation de nature théorique explique la difficulté et parfois la frustration que rencontrent les utilisateurs débutants des systèmes de calcul formel face à des fonctions de « simplification », qui ne peuvent être qu'heuristiques [1].

Pour effectuer un calcul, il est pourtant souvent crucial de déterminer si des expressions représentent 0 ou non, en particulier pour évaluer une fonction qui possède des singularités (comme la division). L'approche du calculateur formel expérimenté consiste à se ramener autant que faire se peut à des opérations d'un domaine dans lequel le test à zéro est décidable. Le calcul formel repose ainsi de manière naturelle sur des constructions algébriques qui préservent la décidabilité du test à 0. En particulier, les opérations courantes sur les vecteurs, matrices, polynômes, fractions rationnelles, ne nécessitent pas d'autre test à 0 que celui des coefficients.

2.2. Structures et constructions de base. Les objets les plus fondamentaux sont assez faciles à représenter en machine de manière exacte. Nous considérons tour à tour les plus importants d'entre eux, en commençant par les plus basiques. Ils s'assemblent ensuite à l'aide de tableaux ou de listes pour en former de plus complexes.

Entiers machine. Les entiers fournis par les processeurs sont des entiers modulo une puissance de 2 (le nombre de bits d'un mot machine, typiquement 32 ou 64). Ils sont appelés des *entiers machine*. Les opérations de base (addition, soustraction, multiplication et division) sur ces entiers sont fournies par la plupart des langages de programmation.

Entiers. Pour manipuler des entiers dont la taille dépasse celle d'un mot machine, il est commode de les considérer comme écrits dans une base B assez grande :

$$N = a_0 + a_1 B + \cdots + a_k B^k.$$

L'écriture est unique si l'on impose $0 \leq a_i < B$. (Le signe est stocké séparément.) Ces nombres peuvent être stockés dans des tableaux d'entiers machine. Les objets obtenus sont des entiers de taille arbitraire. L'addition et le produit peuvent alors être réduits à des opérations sur des entiers inférieurs à B^2, qui peut être choisi par exemple pour tenir dans un mot machine.

Entiers modulaires. Les calculs avec des polynômes, des fractions rationnelles ou des matrices à coefficients entiers souffrent souvent d'une maladie propre au calcul formel : la croissance des expressions intermédiaires. Les entiers produits comme coefficients des expressions intervenant lors du calcul sont de taille disproportionnée par rapport à ceux qui figurent dans l'entrée et dans la sortie.

1. Pour cette raison, un petit guide des fonctions de simplification en Maple est fourni sur la page web du cours.

Exemple 1. Voici le déroulement typique du calcul du plus grand diviseur commun (pgcd) de deux polynômes à coefficients entiers par l'algorithme d'Euclide :

$$
\begin{aligned}
P_0 &= 7x^5 - 22x^4 + 55x^3 + 94x^2 - 87x + 56,\\
P_1 &= 62x^4 - 97x^3 + 73x^2 + 4x + 83,\\
P_2 &= \mathrm{rem}(P_0, P_1) = \frac{113293}{3844}x^3 + \frac{409605}{3844}x^2 - \frac{183855}{1922}x + \frac{272119}{3844},\\
P_3 &= \mathrm{rem}(P_1, P_2) = \frac{18423282923092}{12835303849}x^2 - \frac{15239170790368}{12835303849}x + \frac{10966361258256}{12835303849},\\
P_4 &= \mathrm{rem}(P_2, P_3) = -\frac{216132274653792395448637}{44148979404824831944178}x - \frac{631179956389122192280133}{88297958809649663888356},\\
P_5 &= \mathrm{rem}(P_3, P_4) = \frac{20556791167692068695002336923491296504125}{3639427682941980248860941972667354081}.
\end{aligned}
$$

Chaque étape calcule le reste (noté rem pour *remainder*) de la division euclidienne des deux polynômes précédents. Les coefficients de ces polynômes intermédiaires font intervenir des entiers qui croissent de manière exponentielle, alors que le résultat recherché est 1.

Les entiers modulaires remédient à ce problème de deux manières. D'une part, pour un calcul de décision, de dimension, ou de degré, l'exécution de l'algorithme sur la réduction de l'entrée modulo un nombre premier donne un algorithme *probabiliste* répondant à la question. Cette technique peut aussi servir de base à un algorithme *déterministe* lorsque les nombres premiers pour lesquels la réponse est fausse peuvent être maîtrisés. C'est le cas du pgcd : en évitant les premiers qui divisent les coefficients de tête des deux polynômes, le degré du pgcd modulaire est le même que le degré du pgcd exact.

D'autre part, les entiers modulaires sont utilisés dans les algorithmes reposant sur le théorème des restes chinois. Ce théorème indique qu'un entier inférieur au produit de nombres premiers $p_1 \cdots p_k$ peut être reconstruit à partir de ses réductions modulo $p_1, \ldots, p_k$. Lorsqu'une borne sur la taille du résultat est disponible, il suffit d'effectuer le calcul modulo suffisamment de nombres premiers (choisis assez grands pour que leur nombre soit faible et assez petits pour que les opérations tiennent dans un mot machine), pour ensuite reconstruire le résultat, court-circuitant de la sorte toute croissance intermédiaire.

Vecteurs et matrices. Une fois donnée une représentation exacte pour des coefficients, il est facile de construire des vecteurs ou matrices comme des tableaux, ou plus souvent comme des tableaux de pointeurs sur les coefficients. Les opérations de produit par un scalaire, de produit de matrices ou de produit d'une matrice par un vecteur se réduisent aux opérations d'addition et de multiplication sur les coefficients. Il en va de même de la recherche de noyau ou d'inverse de matrices.

Polynômes. Les polynômes peuvent être stockés de plusieurs manières, et la meilleure représentation dépend des opérations que l'on souhaite effectuer. Pour un polynôme en une variable, les choix principaux sont :

— la représentation dense : comme pour les entiers, le polynôme est représenté comme un tableau de (pointeurs sur les) coefficients ;

— la représentation creuse : le polynôme est représenté comme une liste de paires (coefficient, exposant) généralement triée par les exposants.

Par exemple, le système Maple utilise la seconde représentation par défaut, sans trier les exposants. En outre, il ne développe pas les produits automatiquement, et il faut le lui demander explicitement par la commande `expand`. Une autre commande utile sur les polynômes est `collect` qui regroupe les coefficients et permet d'y appliquer une fonction.

Récursivement, on construit bien sûr les polynômes multivariés.

Fractions rationnelles. Les rationnels peuvent être stockés comme des paires où numérateur et dénominateur sont des entiers de taille arbitraire. Les opérations d'addition et de multiplication se réduisent aux opérations analogues sur les entiers et le test d'égalité à zéro se réduit au test d'égalité à 0 sur le numérateur. De même, les fractions rationnelles sont représentées par des paires de polynômes. Les opérations d'addition, produit, division se réduisent aux additions et multiplications sur les coefficients.

Ces constructions sont possibles dès que les coefficients sont disponibles. Il est donc possible par exemple de manipuler des polynômes dont les coefficients sont des rationnels, des entiers modulaires, ou des matrices.

En Maple, les rationnels sont simplifiés automatiquement ; les fractions rationnelles le sont par la commande `normal`.

Séries tronquées. Les séries tronquées

$$\sum_{k=0}^{N} a_k X^k + O(X^{N+1})$$

se représentent pratiquement comme des polynômes. La différence principale apparaît lors du produit : les coefficients des termes d'exposant au moins $N+1$ n'ont pas besoin d'être calculés, ni stockés. Cette structure de données joue un rôle très important non seulement pour des calculs d'approximations, mais aussi comme une représentation *exacte*. En voici trois exemples importants qui seront abordés dans le Cours sur les approximants de Padé et de Padé-Hermite :

1. Une fraction rationnelle dont les numérateurs et dénominateurs ont degré borné par d peut être reconstruite à partir d'un développement en série à l'ordre $2d+1$.
2. Un polynôme en deux variables peut être reconstruit à partir du développement en série d'une solution.
3. Il est possible de reconstruire une équation différentielle linéaire à coefficients polynomiaux à partir du développement en série d'une solution et de bornes sur l'ordre et le degré des coefficients. De façon analogue, il est possible de reconstruire une récurrence linéaire à coefficients polynomiaux à partir des premières valeurs d'une de ses solutions.

2.3. Équations comme structures de données. Une fois construits les objets de base que sont les polynômes, les séries ou les matrices, il est possible d'aborder des objets mathématiques construits *implicitement*. Ainsi, il est bien connu qu'il n'est pas possible de représenter toutes les solutions de polynômes de haut degré

par radicaux, mais de nombreuses opérations sur ces solutions sont aisées en prenant le polynôme lui-même comme structure de données. Ce point de vue permet d'étendre le domaine d'application du calcul formel pourvu que des algorithmes soient disponibles pour effectuer les opérations souhaitées (typiquement addition, multiplication, multiplication par un scalaire, test d'égalité) par manipulation des équations elles-mêmes.

Nombres algébriques. C'est ainsi que l'on nomme les solutions de polynômes univariés à coefficients entiers. Les opérations d'addition et de multiplication peuvent être effectuées à l'aide de résultants (Cours 5). La division s'obtient par l'algorithme d'Euclide sur les polynômes (Cours 2), et le test à zéro se déduit du pgcd. Par exemple, il est possible de prouver assez facilement une identité comme

$$\frac{\sin\frac{2\pi}{7}}{\sin^2\frac{3\pi}{7}} - \frac{\sin\frac{\pi}{7}}{\sin^2\frac{2\pi}{7}} + \frac{\sin\frac{3\pi}{7}}{\sin^2\frac{\pi}{7}} = 2\sqrt{7} \tag{1}$$

une fois que l'on reconnaît qu'il s'agit d'une égalité entre nombres algébriques.

Systèmes polynomiaux. De nombreuses questions naturelles sur un système de polynômes, comme l'existence de solutions, la dimension de l'espace des solutions (qui indique s'il s'agit d'une surface, d'une courbe, ou de points isolés), le degré, ou le calcul d'une paramétrisation de l'ensemble des solutions trouvent une réponse algorithmique en utilisant comme structure de données des bases de Gröbner, qui seront abordées dans le Cours 8.

Il est également possible d'éliminer une ou des variables entre des polynômes. Cette opération peut s'interpréter géométriquement comme une projection. Dans le cas le plus simple, elle permet de calculer un polynôme s'annulant sur les abscisses des intersections de deux courbes. Une autre application est l'implicitisation, qui permet par exemple de calculer une équation pour une courbe donnée sous forme paramétrée.

Équations différentielles linéaires. Cette structure de données permet de représenter de nombreuses fonctions usuelles transcendantes (exponentielle, fonctions trigonométriques et trigonométriques hyperboliques, leurs réciproques) ainsi que de nombreuses fonctions spéciales de la physique mathématique (fonctions de Bessel, de Struve, d'Anger, ..., fonctions hypergéométriques et hypergéométriques généralisées), ainsi bien sûr que de multiples fonctions auxquelles n'est pas attaché un nom classique. Les opérations d'addition et de produit sont effectuées par des variantes noncommutatives du résultant qui se ramènent à de l'algèbre linéaire élémentaire (Cours 6). Le test à zéro se ramène à tester l'égalité d'un nombre fini de conditions initiales. En d'autres termes, des structures de données finies permettent de manipuler ces objets infinis et d'en tester l'égalité ou la nullité.

Ainsi, des identités élémentaires comme $\sin^2 x + \cos^2 x = 1$ sont non seulement facilement prouvables algorithmiquement, mais elles sont également calculables, c'est-à-dire que le membre droit se calcule à partir du membre gauche. Les relations étroites entre équations différentielles linéaires et récurrences linéaires — les séries solutions des unes ont pour coefficients les solutions des autres — amènent aux mêmes réponses algorithmiques à des questions sur des suites. Par exemple,

l'identité de Cassini sur les nombres de Fibonacci

$$F_{n+2}F_n - F_{n+1}^2 = (-1)^{n+1}, \qquad n \geq 0$$

est exactement du même niveau de difficulté que $\sin^2 x + \cos^2 x = 1$.

En conclusion, les exemples ci-dessus illustrent bien la manière dont le calcul formel parvient à effectuer de nombreux calculs utiles dans les applications malgré l'indécidabilité révélée par le théorème de Richardson-Matiyasevich.

2.4. Taille et vitesse. En pratique, la calculabilité n'indique que la faisabilité. Il faut disposer d'algorithmes efficaces et d'une bonne implantation pour pouvoir effectuer des calculs de grande taille. Les progrès tant théoriques que pratiques du calcul formel ces dernières années amènent aujourd'hui à des temps de calcul très impressionnants sur des calculs simples, et plus longs sur des questions pour lesquelles notre compréhension est encore insuffisante.

Voici par exemple ce qui peut être calculé en *une minute* avec le système Maple, en notant $\mathbb{K}$ le corps $\mathbb{Z}/p\mathbb{Z}$ à p éléments, $p = 67108879$ étant un nombre premier de 26 bits (dont le carré tient sur un mot machine) :

1. Entiers :
   — produit de deux entiers avec 1 000 000 000 de chiffres ;
   — factorielle de 40 000 000 (environ 280 000 000 de chiffres) ;
   — factorisation d'un entier de 72 chiffres (366 bits).
2. Polynômes dans $\mathbb{K}[x]$:
   — produit de deux polynômes de degré 2 000 000 ;
   — pgcd de deux polynômes de degré 100 000 ;
   — factorisation d'un polynôme de degré 2 000.
3. Polynômes dans $\mathbb{K}[x, y]$:
   — produit de deux polynômes de degré total 1 100 ;
   — factorisation d'un polynôme de degré 600 en deux variables.
4. Matrices :
   — déterminant d'une matrice $5\,500 \times 5\,500$ à coefficients dans $\mathbb{K}$;
   — polynôme caractéristique d'une matrice $1\,700 \times 1\,700$ à coefficients dans $\mathbb{K}$;
   — déterminant d'une matrice 650×650 à coefficients des entiers 32 bits.

TP 0

Coefficients de $(\sqrt{x^2-1})^{(n)}$

L'objectif de ce TP est double : d'une part il s'agit d'effectuer des premiers pas en Maple en alternant manipulations simples et recherches dans l'aide en ligne ; d'autre part, il montre comment utiliser le calcul formel pour conjecturer puis prouver une formule. L'exemple traité est celui de la dérivée $n$ième de $\sqrt{x^2-1}$.

1. Calculer les dix premières dérivées de $\sqrt{x^2-1}$ et observer qu'elles sont de la forme

$$\frac{d^n}{dx^n}\sqrt{x^2-1} = \frac{P_n(x)}{(x^2-1)^{\alpha_n}}, \tag{E}$$

   où P_n est un polynôme. Conjecturer les valeurs du degré de P_n et de α_n.

Dans la suite, on se concentrera sur le cas où n est un entier pair, le cas des valeurs impaires se traite de manière similaire.

2. Calculer le polynôme P_{100}.
3. Conjecturer une récurrence pour les valeurs des coefficients de P_{100} (à l'aide de la fonction `seriestorec` du package `gfun`). L'algorithme utilisé sera présenté plus tard dans le cours.
4. Factoriser les coefficients de cette récurrence, et en déduire une récurrence plausible pour les coefficients de P_n pour n pair arbitraire.
5. Résoudre cette récurrence.
6. Il reste à déterminer les conditions initiales $P_n(0)$. Pour cela, à nouveau, calculer les premières valeurs, conjecturer une récurrence et la résoudre.
7. Combiner ces conditions initiales avec les valeurs trouvées plus tôt pour donner une formule plausible pour l'équation (E) lorsque n est pair.
8. Utiliser le système pour prouver cette formule par récurrence.

Première partie

Outils d'expérimentation sur les séries

COURS 1

Calculs de séries par itération de Newton formelle

Résumé

Ce cours définit soigneusement les séries formelles, et montre comment une version formelle de l'itération de Newton permet de calculer facilement de nombreuses séries tronquées.

1. Séries formelles

Dans cette section et la suivante les séries ont leurs coefficients dans un anneau qui n'est pas nécessairement commutatif, sauf lorsque c'est explicitement indiqué. La motivation pour cette généralité est qu'une matrice de séries peut alors être vue comme une série à coefficients des matrices, ce qui permet de proposer des algorithmes travaillant directement sur des systèmes. À partir de la section 3, l'utilisation de la formule de Taylor contraint à se restreindre au cas commutatif.

Définition. Si $\mathbb{A}$ est un anneau, on note $\mathbb{A}[[X]]$ l'ensemble des séries formelles sur $\mathbb{A}$. Ses éléments sont les suites $(f_i)_{i\in\mathbb{N}}$ de $\mathbb{A}$, notées

$$F(X) = \sum_{i\geq 0} f_i X^i.$$

Le coefficient f_i est appelé le i^{e} coefficient de $F(X)$, le coefficient f_0 est appelé terme constant de $F(X)$, et parfois noté $F(0)$.

L'indice du premier coefficient non-nul est appelé la *valuation* de F et noté $\operatorname{val} F$. Par convention, $\operatorname{val} 0 = \infty$.

Les opérations de $\mathbb{A}[[X]]$ sont l'addition des suites et une multiplication (appelée parfois produit de Cauchy) qui généralise la multiplication des polynômes :

$$\sum_{i\geq 0} f_i X^i \times \sum_{i\geq 0} g_i X^i = \sum_{i\geq 0} h_i X^i, \qquad \text{avec} \quad h_i = \sum_{j+k=i} f_j g_k,$$

la dernière somme étant finie.

Exemple 1. La série formelle

$$1 = 1 \cdot X^0 + 0 \cdot X + 0 \cdot X^2 + \cdots,$$

où 1 est l'unité de $\mathbb{A}$, est élément neutre pour la multiplication de $\mathbb{A}[[X]]$. La formule donnant les coefficients du produit se réduit alors à un terme.

Exemple 2. Si $F = 1 + X$ et $G = 1 - X + X^2 - X^3 + \cdots$, alors le produit vaut $H = FG = 1$: le coefficient de X^n dans H vaut $1 - 1 = 0$ pour $n > 0$ et 1 sinon. La série G est donc l'inverse de F pour la multiplication, que l'on peut noter $(1 + X)^{-1}$.

Métrique. Lorsque les coefficients sont des nombres complexes, la question de la convergence des séries entières représentées par ces séries formelles ne se posera pas pour les algorithmes considérés dans ce chapitre. En revanche, quel que soit l'anneau de coefficients $\mathbb{A}$, il est possible de définir une distance entre deux séries F et G par $d(F,G) = 2^{-\operatorname{val}(F-G)}$.

EXERCICE 1. Vérifier que la fonction d définie ci-dessus est bien une distance.

Muni de cette distance, l'ensemble des séries formelles forme un espace métrique *complet* (les suites de Cauchy convergent). En effet, dire qu'une suite (S_n) de séries est de Cauchy signifie qu'étant donné $k \in \mathbb{N}$, il existe un entier K tel que pour tous $m, n \geq K$, on ait $d(S_m, S_n) < 2^{-k}$, autrement dit après l'indice K, les k premiers termes des S_n sont fixés, ce sont ceux de la série limite.

Composition. Si F et G sont deux séries formelles, avec terme constant $G(0) = 0$, on définit la composition comme

$$F(G(X)) = f_0 + f_1 G(X) + f_2 G(X)^2 + \cdots$$

Les points de suspension signifient que l'on considère la limite des sommes H_n obtenues en arrêtant la somme au terme $f_n G(X)^n$. Cette limite existe puisque pour $m \geq n$, la distance obéit à $d(H_m, H_n) \leq 2^{-n}$, ce qui fait de $(H_n)_n$ une suite de Cauchy.

Inverse. Si F est de la forme $F = 1 + GX$ avec $G \in \mathbb{A}[[X]]$, alors F est inversible et son inverse est donné par

$$1 - GX + G^2X^2 - \cdots,$$

composition de $(1+X)^{-1}$ par GX. La preuve découle de cette composition : $(1 + GX)(1 - GX + \dots) = H(GX)$ où $H = (1+X)(1+X)^{-1} = 1$.

Si $a = F(0)$ est inversible, la série F se récrit $F = a(1 + GX)$ avec $G = a^{-1}(F - a)/X$, donc F est inversible, d'inverse $(1 + GX)^{-1}a^{-1} = a^{-1} - Ga^{-1}X + G^2a^{-1}X^2 + \cdots$.

Ces ingrédients mènent à la structure d'anneau de l'ensemble de séries formelles.

PROPOSITION 1. *L'ensemble $\mathbb{A}[[X]]$ des séries formelles à coefficients dans $\mathbb{A}$ est un anneau, qui est commutatif si $\mathbb{A}$ l'est. Ses éléments inversibles sont les séries de terme constant inversible.*

DÉMONSTRATION. L'addition est commutative comme celle de $\mathbb{A}$. L'associativité et la distributivité du produit sont obtenues comme pour les polynômes. L'unité pour le produit est la série 1. La première partie de la proposition est donc prouvée.

Si $F = \sum f_i X^i \in \mathbb{A}[[X]]$ est inversible, et $G = \sum g_i X^i$ est son inverse, alors l'extraction du coefficient de X^0 dans l'identité $FG = 1$ donne $f_0 g_0 = 1$ ce qui prouve qu'une série inversible a un terme constant inversible. La réciproque a été présentée ci-dessus. □

Séries à plusieurs variables. Comme $\mathbb{A}[[X]]$ est un anneau, il peut être utilisé comme anneau de base pour définir des séries formelles en une autre variable Y, ce qui définit de la même manière l'anneau des séries formelles à deux variables $\mathbb{A}[[X]][[Y]]$, noté aussi $\mathbb{A}[[X,Y]]$.

Dérivation. La dérivée d'une série est définie formellement coefficient par coefficient via l'identité

$$\left(\sum_{i\geq 0} f_i X^i\right)' = \sum_{i\geq 0} (i+1) f_{i+1} X^i.$$

Les relations habituelles $(F+G)' = F' + G'$ et $(FG)' = F'G + FG'$ sont prouvées comme pour les polynômes. Dans le cas de séries en plusieurs variables, on utilise la notation des dérivées partielles $\partial F/\partial X$ pour désigner la dérivée par rapport à la variable X.

Troncatures. Algorithmiquement, on ne manipule pas de série à précision « infinie », mais seulement des troncatures, c'est-à-dire un certain nombre des premiers termes.

Étant donnée la série

$$S = \sum_{i\geq 0} a_i X^i,$$

on notera $S \bmod X^N$ le polynôme

$$S \bmod X^N := \sum_{0\leq i<N} a_i X^i.$$

On notera $S = f + O(X^N)$ si les séries ou polynômes S et f coïncident jusqu'au terme de degré $N-1$, et même plus généralement $S = O(T^k)$ si $S = O(X^{k\,\mathrm{val}(T)})$.

Formule de Taylor. Dans le cas où l'anneau de coefficients $\mathbb{A}$ est commutatif, le lien entre la composition et la dérivation est donné par la formule de Taylor. Si F, G, H sont trois séries formelles, avec $G(0) = H(0) = 0$, et les entiers $2, 3, \ldots, k-1$ sont inversibles dans $\mathbb{A}$, alors

$$F(G+H) = F(G) + F'(G)H + F''(G)\frac{H^2}{2!} + \cdots + O(H^k).$$

La formule est classique pour les polynômes, et les coefficients de X^N dans les deux membres de la formule sont les mêmes que ceux de l'identité entre polynômes obtenue en considérant les troncatures de F, G et H modulo X^{N+1}, ce qui permet de conclure.

Pour cette identité, la commutativité de $\mathbb{A}$ est cruciale comme le montre l'exemple de $F = X^2$: on a alors $(G+H)^2 = G^2+GH+HG+H^2$ et la partie linéaire $GH+HG$ est en général différente de $2GH = F'(G)H$.

Intégration. Il s'agit de l'opération inverse de la dérivation. On suppose que les entiers sont inversibles dans $\mathbb{A}$ et on définit alors

$$\int \sum_{i\geq 0} f_i X^i = \sum_{i\geq 0} f_i \frac{X^{i+1}}{i+1}.$$

Exemple 3. La série $\log(1+X)$ vaut

$$\log(1+X) = \int (1+X)^{-1} = X - \frac{1}{2}X^2 + \frac{1}{3}X^3 + \cdots$$

Inverse multiplicatif de séries formelles

Entrée : un entier $N > 0$, $F \bmod X^N$ une série tronquée ;

Sortie : $F^{-1} \bmod X^N$.

Si $N = 1$, alors renvoyer f_0^{-1}, où $f_0 = F(0)$.
Sinon,

1. Calculer récursivement l'inverse G de $F \bmod X^{\lceil N/2 \rceil}$;
2. Renvoyer $G + (1 - GF)G \bmod X^N$.

FIGURE 1. Inverse de série par itération de Newton.

2. La méthode de Newton pour le calcul d'inverses

LEMME 1. *Soit $F \in \mathbb{A}[[X]]$ une série formelle de terme constant inversible et G une série telle que $G - F^{-1} = O(X^n)$ $(n \geq 1)$, alors la série*

$$\mathcal{N}(G) = G + (1 - GF)G \tag{1}$$

vérifie $\mathcal{N}(G) - F^{-1} = O(X^{2n})$.

DÉMONSTRATION. Par hypothèse, on peut définir $H \in \mathbb{A}[[X]]$ par $1 - GF = HX^n$. Il suffit alors de récrire $F = G^{-1}(1 - HX^n)$ et d'inverser :

$$F^{-1} = (1 + HX^n + O(X^{2n}))G = G + HGX^n + O(X^{2n})G = \mathcal{N}(G) + O(X^{2n}).$$

□

2.1. Algorithme. L'algorithme d'inversion de la figure 1 se déduit du lemme. L'avantage de cette approche par rapport à l'utilisation de la formule par composition de la section précédente est qu'elle utilise beaucoup moins de multiplications et s'avère ainsi beaucoup plus rapide en pratique (en théorie aussi, mais nous ne faisons pas d'analyse de complexité dans ce cours).

2.2. Division de séries. Le quotient H/F où H et F sont des séries et F est inversible, peut être obtenu comme le produit $H \times F^{-1}$.

2.3. Application à la division euclidienne. L'algorithme de la figure 1 est l'élément clé de la division euclidienne rapide, en changeant les polynômes à diviser en des séries "à l'infini". On obtient ainsi le quotient, et le reste par multiplication et soustraction.

2.4. Application au logarithme. Si $F \in \mathbb{A}[[X]]$ est telle que $F(0) = 0$, on définit la série $\log(1+F)$ par composition avec $\log(1+X)$. Pour calculer cette série lorsque $\mathbb{A}$ est commutatif, il suffit d'utiliser l'identité

$$\log(1 + F) = \int \frac{F'}{1 + F}.$$

EXERCICE 2. Prouver cette identité à partir des définitions de log, $\int$ et de la dérivation données plus haut.

Le calcul demande une division de séries, une dérivation et une intégration, mais ces deux dernières opérations sont de complexité linéaire.

2.5. Inverse de matrices. L'anneau $\mathbb{A}$ n'étant pas supposé commutatif, il est possible de traiter directement des matrices de séries comme des séries de matrices : le lemme 1 et l'algorithme de la figure 1 s'appliquent.

3. Résolution d'équation par itération de Newton formelle

THÉORÈME 2 (Itération de Newton sur les séries). *Soit $\Phi \in \mathbb{A}[[X,Y]]$ une série à deux variables en X et Y avec $\mathbb{A}$ commutatif, telle que $\Phi(0,0) = 0$ et $\frac{\partial\Phi}{\partial Y}(0,0)$ est inversible dans $\mathbb{A}$. Il existe alors une unique série $S \in \mathbb{A}[[X]]$, telle que $\Phi(X,S) = 0$ et $S(0) = 0$. Si F est une série telle que $S - F = O(X^n)$ $(n \geq 1)$, alors*

$$\mathcal{N}(F) = F - \frac{\Phi(X,F)}{\frac{\partial\Phi}{\partial Y}(X,F)}$$

vérifie $S - \mathcal{N}(F) = O(X^{2n})$.

Ce résultat est un résultat local en $(0,0)$: si ce point est solution, il s'étend en une courbe solution. Par translation, d'autres situations où la solution n'est pas en $(0,0)$ s'y ramènent. La première partie est une version du théorème des fonctions implicites pour les séries formelles.

Comme pour l'inverse, on peut se passer de ce résultat, et calculer la solution par une méthode de coefficients indéterminés. Là encore, l'utilisation de l'itération de Newton permet d'économiser de nombreuses multiplications et de calculer bien plus rapidement (ou, ce qui est équivalent, d'atteindre des ordres plus grands dans un temps raisonnable).

DÉMONSTRATION. D'abord il faut observer que l'itération est bien définie : la composition de Φ et de $\frac{\partial\Phi}{\partial Y}$ avec F sont possibles parce que $F(0) = S(0) = 0$. Pour la même raison, $\frac{\partial\Phi}{\partial Y}(X,F)$ est inversible puisque son terme constant $\frac{\partial\Phi}{\partial Y}(0,0)$ est inversible. De ce fait, on déduit aussi $\operatorname{val}(\mathcal{N}(F) - F) = \operatorname{val}(\Phi(X,F))$.

Ensuite, la preuve se déroule en deux temps : l'itération est d'abord utilisée pour montrer l'existence d'une série solution S, puis on montre l'unicité en même temps que la vitesse de convergence vers S.

La valuation de $\Phi(X,F)$ est doublée par l'itération : la formule de Taylor donne

$$\begin{aligned}\Phi(X,\mathcal{N}(F)) &= \Phi(X,F) + \frac{\partial\Phi}{\partial Y}(X,F)(\mathcal{N}(F) - F) + O((\mathcal{N}(F) - F)^2) \\ &= O(\Phi(X,F)^2).\end{aligned} \tag{2}$$

La suite définie par $S_0 = 0$ et $S_{k+1} = \mathcal{N}(S)$ vérifie donc $\operatorname{val}(S_{k+2} - S_{k+1}) = \operatorname{val}(\Phi(X,S_{k+1})) \geq 2\operatorname{val}(S_{k+1} - S_k)$. Cette suite est donc de Cauchy et sa limite S existe et vérifie $\Phi(X,S) = \lim(\Phi(X,S_k)) = 0$.

La vitesse de convergence vers S vient encore de la formule de Taylor. En effet, l'égalité

$$0 = \Phi(X,S) = \Phi(X,F) + \frac{\partial\Phi}{\partial Y}(X,F)(S - F) + O((S - F)^2) \tag{3}$$

donne $\operatorname{val}(S - F) = \operatorname{val}(\Phi(X,F))$. On en déduit $S - \mathcal{N}(F) = O(\Phi(X,\mathcal{N}(F))) = O(X^{2n})$ au vu des hypothèses et de (2). Cette égalité donne aussi l'unicité de la solution : si F est une autre solution avec $F(0) = 0$, elle doit vérifier $S - F = O((S-F)^2)$, ce qui entraîne $\operatorname{val}(S - F) = \infty$ et donc $F = S$. □

De même que pour l'inverse, on déduit de ce résultat un algorithme récursif pour calculer les N premiers coefficients de la série solution.

3.1. Applications. Voici quelques exemples d'applications :
— l'inverse de la section précédente avec $\Phi(X,Y) = f - 1/(f(0)^{-1}+Y)$ (lorsque $\mathbb{A}$ est commutatif) ;
— $\exp(f)$ lorsque $f(0) = 0$;
— f^α lorsque $f(0) = 1$.
La puissance se déduit de l'exponentielle et du logarithme par la formule $f^\alpha = \exp(\alpha \log f)$.

Pour l'exponentielle, définie par

$$\exp(F) = 1 + F + F^2/2! + F^3/3! + \cdots$$

lorsque $2, 3, \ldots$ sont inversibles dans $\mathbb{A}$, l'idée est d'appliquer une méthode de Newton avec

$$\Phi(Y) = F - \log Y, \quad Y(0) = 1.$$

EXERCICE 3. Montrer que les conditions du Théorème 2 sont vérifiées.

Il s'ensuit une convergence quadratique vers $\exp(F)$ pour l'itération

$$E_{k+1} = \mathcal{N}(E_k) = E_k + E_k(F - \log E_k).$$

EXERCICE 4. Donner une itération de Newton qui calcule directement la racine carrée, sans passer par l'exponentielle et le logarithme.

3.2. Systèmes. Le théorème 2 s'étend à des systèmes d'équations.

THÉORÈME 3. *Soient $\Phi = (\Phi_1, \ldots, \Phi_k)$ des séries en $k+1$ indéterminées X et $Y = (Y_1, \ldots, Y_k)$, telles que $\Phi(0) = 0$ et la matrice jacobienne $(\frac{\partial \Phi}{\partial Y})$ est inversible dans $\mathbb{A}$ en 0. Alors, le système*

$$\Phi_1(X, Y_1, \ldots, Y_k) = \cdots = \Phi_k(X, Y_1, \ldots, Y_k) = 0$$

admet une solution $S = (S_1, \ldots, S_k) \in \mathbb{A}[[X]]^k$ telle que $S(0) = 0$. Si $F = (F_1, \ldots, F_k)$ est tel que $S - F = O(X^n)$ $(n \geq 1)$, alors

$$\mathcal{N}(F) = \begin{pmatrix} F_1 \\ \vdots \\ F_k \end{pmatrix} - \left(\frac{\partial \Phi}{\partial Y}(X, F_1, \ldots, F_k)\right)^{-1} \cdot \begin{pmatrix} \Phi_1(X, F_1, \ldots, F_k) \\ \vdots \\ \Phi_k(X, F_1, \ldots, F_k) \end{pmatrix}$$

vérifie $S - \mathcal{N}(F) = O(X^{2n})$.

La preuve est la même que celle du théorème 2, la formule de Taylor pour les fonctions de plusieurs variables s'exprimant alors à l'aide de la matrice jacobienne.

Pour calculer cet inverse, on fait bien entendu appel à l'itération de Newton, et on réutilise les itérés précédents.

TP 1

Calcul de séries par itération de Newton

Ce TP ne porte pas sur une expérience, mais est une collection de petites questions sur lesquelles utiliser l'itération de Newton pour la résolution en série. Une motivation importante est fournie par les séries génératrices d'énumération en combinatoire : il s'agit de séries formelles de la forme

$$\sum_{n=0}^{\infty} a_n z^n \qquad \text{ou} \qquad \sum_{n=0}^{\infty} a_n \frac{z^n}{n!},$$

où a_n désigne le nombre d'objets de taille n dans une certaine famille. Dans le premier cas, la série génératrice est dite *ordinaire* ; elle est *exponentielle* dans le second.

1. Introduction : arbres binaires

Un arbre binaire à n sommets est soit vide ($n = 0$), soit composé d'une racine et de deux sous-arbres binaires de taille k et $n - k - 1$. Les nombres C_n d'arbres binaires de taille n s'appellent les nombres de Catalan.

1. Écrire une procédure prenant une taille n en argument et renvoyant C_n à partir d'une récurrence non-linéaire simple. (En Maple, on aura intérêt à utiliser l'`option remember`.)
2. À partir de $C_0, \ldots, C_{10}$, deviner un polynôme dont la série génératrice $\sum C_n z^n$ est solution (à l'aide de la fonction `listtoalgeq` du package `gfun`).
3. Résoudre ce polynôme et vérifier que sa solution donne bien les C_n jusqu'à $n = 100$. (Voir la documentation de `solve`).
4. Deviner une récurrence linéaire satisfaite par les C_n (à nouveau avec une fonction de `gfun`), et la résoudre.

2. Arbres d'arité 5 et équations algébriques

Ces arbres sont définis comme les arbres binaires, chaque sommet interne ayant cinq sous-arbres et non plus 2.

5. Calculer le nombre B_n d'arbres d'arité 5 à n sommets, pour $n = 0, \ldots, 20$.
6. Deviner un polynôme dont la série génératrice $\sum B_n z^n$ est solution (il vaut mieux télécharger une version récente de `gfun`).
7. À l'aide de ce polynôme et d'une itération de Newton, calculer les nombres B_n, $n = 0, \ldots, 63$.
8. (À la fin, s'il reste du temps). Utiliser les 50 premiers de ces nombres pour deviner une récurrence sur les B_n, la résoudre, et confirmer la solution en comparant la valeur pour $n = 60$.

3. Des arbres bicolorés et un système

Il est possible de considérer des règles plus ou moins complexes de construction. Par exemple, des arbres dont les sommets peuvent être bleus ou verts, ont une arité finie mais illimitée, avec la contrainte qu'une racine bleue a au moins un fils vert, alors qu'une racine verte a au moins deux fils bleus. Ces arbres ont des séries génératrices d'énumération qui sont solution de

$$B(z) = z + \frac{zV(z)}{(1-V(z))(1-B(z))}, \qquad V(z) = z + \frac{zB(z)^2}{(1-V(z))(1-B(z))}.$$

9. Construire la matrice jacobienne du système ;
10. l'utiliser pour former une itération de Newton ;
11. calculer le nombre d'arbres bicolorés dont la racine est bleue pour les tailles $n = 0, \ldots, 31$.

COURS 2

Reconstruction rationnelle, approximants de Padé et de Padé-Hermite

Résumé

L'algorithme d'Euclide étendu permet le calcul de la reconstruction rationnelle. En particulier, il permet de calculer des approximants de Padé. Une généralisation de ces derniers, les approximants de Padé-Hermite, est brièvement présentée.

1. L'algorithme d'Euclide étendu

L'algorithme d'Euclide, qui permet le calcul du pgcd de deux polynômes, peut être enrichi légèrement pour calculer deux suites très utiles également. La version étendue se présente ainsi

Entrée : A et B dans $\mathbb{K}[X]$ où $\mathbb{K}$ est un corps.

Sortie : Un pgcd G de A et B, et les cofacteurs U et V de l'identité de Bézout $UA + VB = G$.

1. $R_0 := A$; $U_0 := 1$; $V_0 := 0$;
2. $R_1 := B$; $U_1 := 0$; $V_1 := 1$;
3. $i := 1$;
4. Tant que R_i est non nul, faire :
   $R_{i-1} = qR_i + R_{i+1}$ # *division euclidienne qui calcule* (q, R_{i+1})
   $U_{i+1} := U_{i-1} - qU_i$; $V_{i+1} := V_{i-1} - qV_i$;
   $i := i + 1$
5. Renvoyer $R_{i-1}, U_{i-1}, V_{i-1}$.

PROPOSITION 1. *L'algorithme d'Euclide étendu est correct, c'est-à-dire qu'il renvoie bien ce qu'il annonce.*

La preuve de cette proposition repose sur une partie du lemme suivant qui regroupe des propriétés de l'algorithme ; les autres parties serviront dans la suite.

LEMME 1. *Pour tout $i \geq 0$, les polynômes calculés par l'algorithme d'Euclide étendu vérifient*

1. $\deg R_{i+1} < \deg R_i$;
2. $R_i = U_i A + V_i B$;
3. $\deg V_i = \deg A - \deg R_{i-1}$ *pour* $i > 0$;
4. U_i *et* V_i *sont premiers entre eux.*

Démonstration. Le premier point est une conséquence de la définition de la division euclidienne.

Le second point se prouve par récurrence : la propriété est clairement vérifiée pour $i = 0$ et $i = 1$ et il suffit d'injecter la définition de U_{i+1} et V_{i+1} dans la définition de R_{i+1} pour voir que si elle est vérifiée pour $i-1$ et i, alors elle l'est pour $i+1$.

Le troisième point est également prouvé par récurrence. Il est vérifié en $i = 1$ ($R_0 = A$ et $V_1 = 1$) et $i = 2$ ($R_1 = B$ et V_2 est le quotient de A par B). Ensuite, le membre droit de l'identité $V_{i+1} := V_{i-1} - qV_i$ avec $\deg q = \deg R_{i-1} - \deg R_i$ comporte deux termes, le premier de degré $\deg A - \deg R_{i-2}$ et le second de degré $\deg q + \deg A - \deg R_{i-1} = \deg A - \deg R_i$ qui est plus grand par décroissance des degrés des R_i ; c'est donc lui le degré de V_{i+1}, ce qui conclut.

Le quatrième point s'obtient en prenant le déterminant de chaque côté de l'identité

$$\begin{pmatrix} U_{i+1} & V_{i+1} \\ U_i & V_i \end{pmatrix} = \begin{pmatrix} -q & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} U_i & V_i \\ U_{i-1} & V_{i-1} \end{pmatrix},$$

ce qui prouve que

$$U_{i+1}V_i - U_iV_{i+1} = -(U_iV_{i-1} - U_{i-1}V_i) = (-1)^{i-1}(U_1V_0 - U_0V_1) = (-1)^i,$$

ce qui entraine que U_i et V_i sont premiers entre eux. □

Démonstration du théorème. La décroissance des degrés des R_i prouve la terminaison de l'algorithme. L'observation que la division euclidienne entraine $\operatorname{pgcd}(R_{i-1}, R_i) = \operatorname{pgcd}(R_i, R_{i+1})$ montre par récurrence le résultat est le pgcd. La propriété souhaitée pour les autres polynômes renvoyés est le point 2 du lemme. □

Inversion modulaire. Une application des coefficients de Bézout U et V renvoyés par l'algorithme d'Euclide étendu est l'*inversion modulaire* : si $UA+VB = 1$, alors B est premier avec A et $VB = 1 - UA$ montre que son inverse modulo A est V. Par exemple,

$$(a - bX)(a + bX) + b^2(1 + X^2) = a^2 + b^2$$

montre que l'inverse de $a + bX$ modulo $1 + X^2$ est $(a - bX)/(a^2 + b^2)$. On retrouve la formule habituelle pour l'inversion des nombres complexes (faire $X = i = \sqrt{-1}$ dans la formule).

2. Reconstruction rationnelle

2.1. Le problème.

Définition 1. *Soit $\mathbb{K}$ un corps, $A \in \mathbb{K}[X]$ un polynôme de degré $n > 0$ et $B \in \mathbb{K}[X]$ de degré $< n$. Pour un $k \in \{1, \ldots, n\}$ fixé, la* reconstruction rationnelle de B modulo A *est la recherche d'un couple de polynômes $(R, V) \in \mathbb{K}[X]^2$ vérifiant :*

$$\text{(RR)} \qquad \operatorname{pgcd}(V, A) = 1, \quad \deg(R) < k, \quad \deg(V) \leq n - k \quad \textit{et} \quad \frac{R}{V} \equiv B \bmod A.$$

Deux cas particuliers sont particulièrement importants :

1. Un *approximant de Padé* de type (m, ℓ) d'une série $S \in K[[X]]$ est une fraction rationnelle R/V telle que

$$\frac{R}{V} = S + O(X^{m+\ell+1})$$

où les polynômes R et V ont degré inférieur ou égal à m et ℓ et $V(0) \neq 0$; c'est la reconstruction rationnelle de $S \mod X^n$ modulo $A = X^n$ avec $n = m + \ell + 1$ et $k = m + 1$.

2. L'*interpolation de Cauchy* consiste, étant donnés k et $((u_1, v_1), \ldots, (u_n, v_n))$, à trouver une fraction rationnelle R/V avec les contraintes de degré comme ci-dessus, telle que
$$\frac{R(u_i)}{V(u_i)} = v_i, \quad i \in \{1, \ldots, n\};$$
c'est la reconstruction rationnelle du polynôme d'interpolation des (u_i, v_i) modulo $A = \prod_i (X - u_i)$.

L'unicité de la fraction solution au problème (RR) est facile à établir : si (R_1, V_1) et (R_2, V_2) sont deux solutions, alors $R_1/V_1 \equiv R_2/V_2 \mod A$, donc A divise $R_1V_2 - V_1R_2$. Ce dernier polynôme ayant un degré strictement inférieur à celui de A, il doit être identiquement nul. Donc les fractions R_1/V_1 et R_2/V_2 coïncident.

En revanche, il est possible que (RR) n'admette aucune solution. C'est le cas pour $n = 3, k = 2$, $A = X^3$, et $B = X^2 + 1$. En effet, si $V(X) = aX + b$, nécessairement $b \neq 0$ puisque V doit être premier avec A, alors $R \equiv (aX + b)(X^2 + 1) = bX^2 + aX + b \mod X^3$, ce qui est incompatible avec $\deg(R) \leq 1$.

Pour contourner temporairement ce problème d'existence, il est utile d'introduire un problème plus simple :

$$\text{(RRS)} \qquad \deg(R) < k, \quad \deg(V) \leq n - k \quad \text{et} \quad R \equiv VB \mod A,$$

où, à la différence de (RR), on a mis de côté la contrainte sur le pgcd de A et de V.

Ce problème (RRS) admet *toujours* une solution non triviale $(R, V) \neq (0, 0)$: en effet, il se traduit en termes d'algèbre linéaire en un système linéaire homogène ayant $k+(n-k+1) = n+1$ inconnues (les coefficients de R et de V) et n équations. Par ailleurs, la preuve d'unicité pour R/V continue de s'appliquer.

2.2. Calcul. Il se trouve que l'algorithme d'Euclide permet de résoudre ces deux problèmes en évitant le recours à des méthodes générales mais potentiellement coûteuses d'algèbre linéaire.

THÉORÈME 4. *Si (R_i, U_i, V_i) est la suite calculée par l'algorithme d'Euclide étendu sur l'entrée A, B et j le plus petit indice tel que $\deg(R_j) < k$, alors*

1. *(R_j, V_j) est une solution non-nulle de (RRS) ;*
2. *Si $\operatorname{pgcd}(A, V_j) = 1$, alors (R_j, V_j) est aussi solution de (RR).*
3. *Si (RR) admet une solution (R, V) alors $\operatorname{pgcd}(A, V_j) = 1$ et $R_j/V_j = R/V$.*

EXEMPLE 1. Avec $A = X^5$ et $B = 1 + X + 2X^2 + 3X^3 + 5X^4$ (le début de la série génératrice des nombres de Fibonacci), on obtient successivement

$$\begin{aligned}
(U_0, V_0) &= (1, 0), & R_0 &= A \\
(U_1, V_1) &= (0, 1), & R_1 &= B \\
(U_2, V_2) &= (1, -\frac{1}{25}(5X - 3)), & R_2 &= -\frac{1}{25}(X^3 - X^2 + 2X - 3) \\
(U_3, V_3) &= (125X + 200, -25(X^2 + X - 1)), & R_3 &= 25 \\
(U_4, V_4) &= (\frac{B}{25}, -\frac{A}{25}), & R_4 &= 0.
\end{aligned}$$

La suite des R_i/V_i pour $i > 0$ donne les approximants de Padé

$$B \equiv \frac{X^3 - X^2 + 2X - 3}{5X - 3} \equiv \frac{1}{1 - X - X^2} \bmod X^5.$$

En particulier, le dernier approximant "retrouve" la série génératrice complète des nombres de Fibonacci.

DÉMONSTRATION. Le premier point découle de l'identité $R_j = U_j A + V_j B$, du fait que le degré de V_j est égal à $n - \deg R_{j-1}$, et que par définition de j, $\deg R_{j-1} \geq k$.

Pour le point 2, l'invertibilité de V_j donne le résultat en divisant la même identité par V_j modulo A.

Enfin, si (R, V) est une solution de (RR), alors $R/V \equiv B \bmod A$ montre qu'il existe U tel que $R = UA + VB$. Il se trouve alors que $U/V = U_j/V_j$. En effet, si cette égalité n'a pas lieu, alors la matrice du système

$$\begin{pmatrix} U_j & V_j \\ U & V \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} R_j \\ V_j \end{pmatrix}$$

est inversible, ce qui permet d'utiliser les formules de Cramer pour obtenir

$$A = \frac{R_j V - R V_j}{U_j V - U V_j}.$$

Alors le degré de A est borné par celui du numérateur du membre droit, lui même borné strictement par $k + n - k$, ce qui est une contradiction.

Donc $UV_j = U_j V$ et par conséquent V_j divise $U_j V$. D'après le lemme sur les propriétés de l'algorithme d'Euclide étendu, V_j est premier avec U_j, donc V_j divise V, et est donc premier avec A si V l'est. Ensuite, il suffit d'écrire

$$\frac{R}{V} = B + \frac{U}{V} A = B + \frac{U_j}{V_j} A = \frac{R_j}{V_j}. \qquad \square$$

2.3. L'algorithme de Berlekamp-Massey. Si on connaît les $2d$ premiers éléments $a_0, a_1, \ldots, a_{2d-1}$ d'une suite qui vérifie une récurrence linéaire à coefficients constants inconnue

$$f_0 a_{i+d} + \cdots + f_d a_i = 0, \qquad i \geq 0$$

et supposée d'ordre d minimal, alors on peut reconstruire cette récurrence à l'aide d'un approximant de Padé de type $(d-1, d)$. Un tel approximant R/V de $a_0 + a_1 X + \cdots + a_{2d-1} X^{2d-1}$ vérifie

$$R = V(a_0 + a_1 X + \cdots + a_{2d-1} X^{2d-1}) + O(X^{2d}).$$

Si $V = v_0 + v_1 X + \cdots + v_\ell X^\ell$ avec $\ell \leq d$, alors l'extraction du coefficient de X^i dans l'identité précédente pour $\deg R < d \leq i \leq 2d - 1$ donne

$$v_0 a_i + v_1 a_{i-1} + \cdots + v_\ell a_{i-\ell} = 0,$$

c'est-à-dire que les coefficients du dénominateur de l'approximant de Padé sont ceux de la récurrence.

3. Approximants de Padé-Hermite

Il s'agit d'une généralisation des approximants de Padé.

DÉFINITION 2. *Soit $F = (f_1, \dots, f_n) \in \mathbb{K}[[X]]^n$, et soit $D = (d_1, \dots, d_n) \in \mathbb{N}^n$. Un vecteur non nul $P = (p_1, \dots, p_n) \in \mathbb{K}[X]^n$ est appelé* approximant de Padé-Hermite de type D de F *si :*

$$p_1 f_1 + \cdots + p_n f_n = O(X^{(d_1+1)+\cdots+(d_n+1)-1})$$

et $\deg(P_i) \leq d_i$ *pour tout* $1 \leq i \leq n$.

Par exemple, si $R/V \in \mathbb{K}(X)$ est un approximant de Padé de type $(k, n-k)$ de $B \in \mathbb{K}[[X]]$, alors (R, V) est un approximant de Padé-Hermite pour $(-1, B)$, de type $(k, n-k)$.

De même que les approximants de Padé permettent de reconstruire des fractions rationnelles, si S est une série, le calcul d'approximant de Padé-Hermite de $(S, S', S'', \dots, S^{(d)})$ permet de reconstruire une équation différentielle linéaire que satisferait S. De même un approximant de Padé-Hermite de $(1, S, S^2, \dots, S^d)$ reconstruit un polynôme que cette série annule. Ces opérations sont à la base des calculs de "devinette" sur les séries utilisés dans les TDs.

PROPOSITION 2. *Tout vecteur de séries formelles $F = (f_1, \dots, f_n) \in \mathbb{K}[[X]]^n$ admet un approximant de Padé-Hermite de type $D = (d_1, \dots, d_n) \in \mathbb{N}^n$ donné.*

DÉMONSTRATION. On procède par coefficients indéterminés : en écrivant $P_i = \sum_{j=0}^{d_i} p_{i,j} X^j$, on obtient un système linéaire homogène à $\sigma = \sum_i (d_i + 1) - 1$ équations en les $\sigma + 1$ inconnues $p_{i,j}$. Puisqu'il a moins d'équations que d'inconnues, ce système admet forcément une solution non triviale. □

Comme pour les approximants de Padé, il est possible de calculer les approximants de Padé-Hermite plus efficacement que par l'algèbre linéaire généraliste, mais nous ne détaillerons pas ces algorithmes ici.

TP 2

Approximants de Padé et de Padé-Hermite

1. Approximation

L'objectif de cet exercice est l'exploration des capacités d'approximation numérique des approximants de Padé bien au-delà du disque de convergence d'une série formelle. On prendra comme exemple la série de Taylor $S = \tan x + O(x^{32})$.

1. Calculer un approximant de Padé de type (15,16) de S (voir ?pade). On appelle ensuite F la fraction obtenue.
2. Tracer sur un même dessin les graphes de F, de S et de la fonction tangente, pour $-20 \leq x \leq 20$.
3. Comparer la qualité des approximations de $\tan$ par F et par S en $x = 1$, c'est-à-dire à l'intérieur du disque de convergence de S.
4. Estimer la vitesse de convergence de la suite d'approximants de Padé diagonaux (c'est-à-dire de type (d, d)) en $x = 1$.
5. Effectuer les mêmes opérations en $x = 10$, c'est-à-dire *en dehors* du disque de convergence de la série.

Les approximants de Padé sont souvent employés pour localiser les singularités et les zéros de fonctions sur lesquelles on dispose de peu d'information.

6. Calculer numériquement les racines du dénominateur de F, et comparer les plus petites en valeur absolue aux valeurs attendues.
7. Faire de même pour le numérateur.
8. Construire une animation permettant de visualiser la convergence des approximants de Padé diagonaux vers la tangente.

2. Singularités d'une intégrale

Une version simplifiée d'un calcul de susceptibilité magnétique d'un modèle d'Ising a conduit des physiciens à s'interroger sur la position des singularités de la fonction

$$\phi_n(w) = \frac{1}{\pi} \int_{-1}^{1} F_n(w,t) \frac{dt}{\sqrt{1-t^2}} \quad \text{où} \quad F_n(w,t) = \frac{1}{1 - x(w,t)^{n-1} x(w, T_{n-1}(t))},$$

$T_{n-1}(t)$ est un polynôme de Tchebychev de 1ère espèce (c'est un polynôme de degré $n-1$ défini par $\cos((n-1)t) = T_{n-1}(\cos(t))$) et

$$x(w,t) = \frac{2w}{1 - 2wt + \sqrt{(1-2wt)^2 - 4w^2}}.$$

Des arguments généraux permettent d'assurer que $\phi_n(w)$ satisfait une équation différentielle linéaire à coefficients polynomiaux, c'est-à-dire de la forme

$$a_k(w)\phi_n^{(k)}(w) + \cdots + a_0(w)\phi_n(w) = 0, \tag{E}$$

où les a_i sont des polynômes. La théorie de ces équations affirme alors que les singularités de $\phi_n(w)$ ne peuvent se trouver que parmi les racines du terme de tête $a_k(w)$ de (E).

Le but de cet exercice est de calculer une telle équation dans le cas d'intérêt le plus simple, c'est-à-dire lorsque $n = 3$. La taille des calculs est déjà telle qu'il faudra souvent aider Maple dans les étapes intermédiaires. L'approche consiste à calculer un développement en série de $\phi_n(w)$, puis reconstruire (E) par un approximant de Padé-Hermite de $(\phi_n, \phi_n', \ldots, \phi_n^{(k)})$.

2.1. Premiers essais d'approximation. Tout d'abord, il est possible de se faire une idée approximative de la position des singularités par un simple calcul d'approximant de Padé.

9. Calculer un développement en série de $F_3(w,t)$ à l'ordre 15 par rapport à w ;
10. intégrer terme à terme pour obtenir le développement de $\phi_3(w)$ (on ne demande pas de justifier l'interversion des signes somme et intégrale) ;
11. calculer un approximant de Padé de cette série et évaluer numériquement la position des pôles de l'approximant.

La suite de l'exercice consiste à trouver un polynôme à coefficients entiers dont ces pôles approchent les racines qui correspondent vraiment aux singularités de ϕ_3.

2.2. Développement en série à grande précision. Pour obtenir l'équation (E) que nous cherchons par un calcul d'approximant de Padé-Hermite, il est nécessaire de disposer d'une bonne centaine de termes du développement en série de $\phi_n(w)$ (cette valeur est trouvée en tâtonnant). Les méthodes directes utilisées dans la section précédente ne peuvent pas aller à de très grands ordres. Il faut aider Maple à développer F_3 par rapport à w, puis à intégrer terme à terme.

Pour développer F_n, l'idée est d'utiliser son caractère algébrique, qui entraîne l'existence d'une récurrence linéaire sur ses coefficients (les preuves et algorithmes seront présentés dans un cours ultérieur).

12. Calculer un polynôme $P(w,t,y)$ tel que $P(w,t,F_3(w,t)) = 0$ (`?algfuntoalgeq`) ;
13. en déduire une équation différentielle (`?algeqtodiffeq`), en précisant que la solution qui nous intéresse est celle qui satisfait $y(0) = 1, y'(0) = 0$;
14. en déduire une récurrence sur les coefficients (`?diffeqtorec`), puis une procédure pour dérouler cette récurrence (`?rectoproc`, avec l'option `evalfun` à positionner à `expand`) ;
15. calculer enfin les 200 premiers coefficients du développement
$$F_3(w,t) = 1 + w^3 + (4t^2 + 4t - 2)w^4 + \ldots.$$
16. Pour aider Maple à intégrer terme à terme cette série, calculer symboliquement l'intégrale
$$I_k = \frac{1}{\pi}\int_{-1}^{1} \frac{t^k}{\sqrt{1-t^2}}\, dt;$$

transformer la série ci-dessus en un polynôme en t à coefficients des polynômes en w (`?collect`), puis intégrer terme à terme et retransformer en série en w. On doit trouver les premiers termes

$$\phi_3(w) = 1 + w^3 + 11w^5 + 7w^6 + \dots.$$

2.3. Approximant de Padé-Hermite.

17. Calculer l'approximant de Padé-Hermite souhaité (`?seriestodiffeq`) ;
18. En déduire une conjecture sur les positions des singularités de ϕ_3 ;
19. Certaines des singularités trouvées à la question précédente ne sont pas vraiment des singularités des solutions ; on les appelle des *singularités apparentes*. Pour s'en débarrasser, calculer une équation différentielle d'ordre plus élevé (à l'aide de `gfun:-Parameters`), et calculer le pgcd des coefficients de tête. Ceci mène à une meilleure conjecture ;
20. Pour conforter cette conjecture, calculer les pôles d'un approximant de Padé (40,40) de la série obtenue en question (16) et les afficher dans le plan complexe, avec les singularités de la question ci-dessus.

Deuxième partie

Outils d'expérimentation sur les constantes

COURS 3

Accélération de convergence

Résumé

L'accélération de convergence est une technique d'analyse numérique qui s'avère utile en calcul formel, en conjonction avec la précision arbitraire et les outils de conjecture à base de l'algorithme LLL présenté au cours suivant. Une partie des convergences spectaculaires de ces méthodes s'explique par leur relation avec les approximants de Padé.

Le principe de l'accélération de convergence est assez simple : on connaît une suite réelle $(S_n)_{n\in\mathbb{N}}$ qui converge vers une valeur S_∞ et on cherche une nouvelle suite T_n qui tende aussi vers S_∞, mais (beaucoup) plus rapidement. Autrement dit, $T_n - S_\infty = o(S_n - S_\infty)$ lorsque $n \to \infty$. Il est possible de trouver de telles suites T_n si l'on dispose d'hypothèses supplémentaires sur la régularité avec laquelle S_n tend vers sa limite.

1. Exemple : Archimède, Huygens et le calcul de π

1.1. La suite à accélérer. Pour encadrer π, le point de départ de la méthode d'Archimède consiste à considérer deux polygones réguliers, l'un inscrit et l'autre circonscrit à un cercle de rayon 1. Une petite figure permet de se convaincre que pour tout n, ces polygones ont pour périmètres $2n\sin\pi/n$ et $2n\tan\pi/n$, d'où découle l'inégalité suivante

$$n\sin\frac{\pi}{n} < \pi < n\tan\frac{\pi}{n}.$$

Ensuite, Archimède, avec bien moins d'outils que ce dont nous disposons, se rend compte qu'il est possible de calculer simultanément les éléments des deux suites

$$s_k := \sin(\alpha/2^k), \qquad t_k := \tan(\alpha/2^k)$$

à l'aide des relations

$$\frac{1}{\tan\frac{x}{2}} = \frac{1}{\tan x} + \frac{1}{\sin x} \quad \text{et} \quad \sin\Big(\frac{x}{2}\Big)^2 = \frac{1}{1+\frac{1}{\tan\left(\frac{x}{2}\right)^2}}.$$

En partant de $\alpha = \pi/3$, Archimède parvient ainsi à pousser le calcul jusqu'à $k = 5$, ce qui correspond à des polygones à 96 côtés ! Il obtient ainsi l'encadrement $3.141 < \pi < 3.142$.

1.2. Principe de l'accélération. Le développement de Taylor de tan donne directement le développement asymptotique

$$n\tan\frac{\pi}{n} = \pi + \frac{\pi^3}{3n^2} + O\left(\frac{1}{n^4}\right), \qquad n \to \infty.$$

Lorsque l'angle est divisé par 2, n est doublé, ce qui mène à

$$2n\tan\frac{\pi}{2n} = \pi + \frac{1}{4}\frac{\pi^3}{3n^2} + O\left(\frac{1}{n^4}\right), \qquad n \to \infty.$$

L'idée de la méthode consiste à éliminer le terme en $1/n^2$ par une combinaison linéaire. Si $t_k^{(0)} := 3\cdot 2^k \tan(\pi/(3\cdot 2^k))$, une suite à convergence plus rapide est ainsi fournie par

$$t_k^{(1)} := \frac{4t_{k+1}^{(0)} - t_k^{(0)}}{3} = \pi + O\left(\frac{1}{16^k}\right), \qquad k \to \infty.$$

Cette idée, jointe à une manipulation analogue sur le sinus, est due à Huygens qui s'en est servi en 1654 pour calculer 35 décimales de π. De nombreuses années plus tard, en 1936, Kommerell se rend compte qu'il est possible de réitérer cette transformation, en considérant cette fois

$$s_k^{(2)} := \frac{16s_{k+1}^{(1)} - s_k^{(1)}}{15} = \pi + O\left(\frac{1}{64^k}\right), \qquad k \to \infty,$$

et ainsi de suite. Toujours avec les mêmes 5 valeurs initiales, cette idée produit l'approximation

$$3.141592653589793$$

correcte à 15 décimales !

2. Méthodes d'extrapolation linéaires

Le calcul de Huygens est un cas particulier des méthodes d'extrapolation linéaires. Le terme *linéaire* qualifie ici l'application d'accélération envoyant une suite sur une suite accélérée, et on parle d'extrapolation pour signifier que l'on cherche à déterminer une valeur en dehors du domaine des valeurs de départ.

2.1. Méthode d'Euler. Il s'agit sans doute de la plus vieille des méthodes d'accélération. Elle s'applique à des suites dont le comportement asymptotique est de la forme suivante (ou est supposé l'être) :

$$S_n = S_\infty + r^n\varphi(n) \quad \text{avec} \quad \frac{\varphi(n+1)}{\varphi(n)} \to 1 \text{ et } r \neq 1 \text{ connu.}$$

(La convergence exige en outre $|r| \leq 1$.)

L'accélération repose sur le principe simple suivant.

Proposition 1. *Dans une telle situation, la suite $T_n := (S_{n+1} - \rho S_n)/(1-\rho)$ vérifie $T_n - S_\infty = o(S_n - S_\infty)$ dès lors que $\rho \neq 1$ et $|\rho - r| < |\rho - 1|$ (et en particulier si $\rho = r$).*

Démonstration. Il suffit de développer :

$$T_n = \frac{S_{n+1} - \rho S_n}{1-\rho} = S_\infty + (S_n - S_\infty)\frac{r\frac{\varphi(n+1)}{\varphi(n)} - \rho}{1-\rho}.$$

La fraction de droite tend vers $(r-\rho)/(1-\rho)$, et donc T_n se rapproche de S_∞ plus vite que S_n. □

Exemple 1. La série géométrique peut être accélérée par cette méthode. Si $S_n = 1 + \alpha + \alpha^2 + \cdots + \alpha^n$, alors le choix $r = \alpha$ donne $T_n = S_\infty$.

La méthode d'Euler consiste à réitérer ce procédé pour obtenir une suite de suites accélérées en posant

$$T_n^{(0)} := S_n, \quad T_n^{(k)} := \frac{T_{n+1}^{(k-1)} - r_k T_n^{(k-1)}}{1 - r_k}, \quad (k \geq 1). \tag{1}$$

EXEMPLE 2. Le calcul de π par Kommerell est obtenu en prenant $r_k = 1/4^k$.

De manière plus surprenante, cette méthode fonctionne également très bien sur des suites qui convergent vraiment lentement, ou même qui divergent complètement, comme

$$\sum_{i=0}^{n} \frac{(-1)^i}{i+1} \qquad \text{et} \qquad \sum_{i=0}^{n} (-1)^i \frac{2^{i+1}}{i+1},$$

pour lesquelles elle fournit (avec $r = -1$ et $r = -2$) des bonnes approximations de $\ln 2$ et $\ln 3$.

2.2. Méthode de Richardson (1910). *Cette méthode n'a pas été traitée à l'oral. Elle permet une introduction plus douce à la méthode de Shanks plus loin.*

Il s'agit d'une généralisation de la méthode précédente au cas où la suite se comporte comme

$$S_n = S_\infty + c_1 r_1^n + \cdots + c_k r_k^n, \tag{2}$$

les r_i étant connus, distincts, différents de 1, et de modules décroissants (mais pas nécessairement strictement décroissants). Le principe est ici d'éliminer les r_i non pas les uns après les autres comme dans la méthode d'Euler, mais tous simultanément. La suite accélérée $T_n^{(k)}$ est fournie par les formules de Cramer : le système linéaire est donné par (2) évalué en $n, n+1, \ldots, n+k$ ayant pour inconnues S_∞ et les c_i. Il s'ensuit une expression en terme de déterminants :

$$T_n^{(k)} = \frac{\begin{vmatrix} S_n & 1 & \ldots & 1 \\ \vdots & \vdots & & \vdots \\ S_{n+k} & r_1^k & \ldots & r_k^k \end{vmatrix}}{\begin{vmatrix} 1 & 1 & \ldots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & r_1^k & \ldots & r_k^k \end{vmatrix}}.$$

2.3. Convergence plus lente. Dans le cas très fréquent où la convergence de la suite est de la forme

$$S_n = S_\infty + \frac{\alpha}{n} + \frac{\beta}{n^2} + \cdots,$$

la méthode d'Euler ne s'applique pas directement puisqu'il faudrait prendre $r = 1$. L'idée est alors de poser d'abord $\widetilde{S_n} = S_{2^n}$, ce qui permet alors d'utiliser la méthode d'Euler sur la suite $\widetilde{S_n}$ avec $r_i = 1/2^i$. L'équation devient alors

$$T_n^{(0)} = S_{2^n}, \quad T_n^{(k+1)} := \frac{2^{k+1} T_{n+1}^{(k)} - T_n^{(k)}}{2^{k+1} - 1}.$$

EXERCICE 5. Employer cette méthode pour calculer une dizaine de décimales de la constante γ d'Euler, définie comme limite de la suite

$$S_n = \sum_{k=1}^{n} \frac{1}{k} - \log n.$$

2.4. Méthode de Romberg (1955). C'est la méthode précédente, appliquée au calcul d'une intégrale par la méthode des trapèzes. La suite S_n est alors

$$S_n = \frac{h}{2}(f(a)+f(b)) + h\sum_{k=1}^{n-1} f(a+kh), \quad h = \frac{b-a}{n}.$$

Il faut cependant faire attention que si la fonction ne se comporte pas aimablement sur l'intervalle $[a,b]$, la convergence peut ne pas être assez bonne pour appliquer la méthode.

3. Méthodes non-linéaires

3.1. Méthode Δ^2 d'Aitken (1926). Cette méthode s'applique aux suites dont le comportement est de la forme

$$S_n = S_\infty + r^n\varphi(n), \qquad \text{avec} \quad \frac{\varphi(n+1)}{\varphi(n)} \longrightarrow 1,$$

mais cette fois-ci, r n'est pas supposé connu. La méthode peut être vue comme un calcul simultané de r et de l'accélération d'Euler.

Pour approcher r, on considère la suite

$$\Delta S_n = S_{n+1} - S_n = r^n(r\varphi(n+1) - \varphi(n)).$$

La notation Δ est classique pour cet opérateur aux différences. On observe alors que $\Delta S_{n+1}/\Delta S_n \longrightarrow r$. Cette propriété est très utile, même en relation avec des méthodes linéaires : si l'on parvient à identifier r en contemplant les premières valeurs de cette suite, alors il vaut mieux exploiter cette valeur.

La transformée d'Aitken consiste donc naturellement à appliquer la Proposition 1, mais en remplaçant r par sa valeur approchée ci-dessus, ce qui donne

$$T_n = \frac{S_{n+1} - \frac{\Delta S_{n+1}}{\Delta S_n} S_n}{1 - \frac{\Delta S_{n+1}}{\Delta S_n}} = S_n - \frac{(\Delta S_n)^2}{\Delta^2 S_n}.$$

Ici, Δ^2 désigne la composition de l'opérateur Δ avec lui-même :

$$\Delta^2 u_n = \Delta(u_{n+1} - u_n) = u_{n+2} - 2u_{n+1} + u_n.$$

Exemple 3. En partant des 11 premiers termes de la formule de Leibniz pour π :

$$\pi = 4 - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \cdots,$$

et en appliquant cette méthode tant qu'il reste assez de termes, on obtient 9 décimales de π, c'est-à-dire qu'on atteint une précision pour laquelle il aurait fallu plus de 10^{10} termes de la série de départ.

3.2. Méthode de Shanks (1949). La méthode de Shanks est à la méthode d'Aitken ce que la méthode de Richardson est à la méthode d'Euler : il s'agit d'éliminer plusieurs termes perturbateurs, mais cette fois-ci les r_i sont inconnus. Alors que la méthode d'Aitken peut se récrire

$$T_n = \frac{\begin{vmatrix} S_n & S_{n+1} \\ \Delta S_n & \Delta S_{n+1} \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ \Delta S_n & \Delta S_{n+1} \end{vmatrix}},$$

le cas général devient

$$T_{n,k} = \frac{\begin{vmatrix} S_n & \dots & S_{n+k} \\ \Delta S_n & \dots & \Delta S_{n+k} \\ \vdots & & \vdots \\ \Delta S_{n+k-1} & \dots & \Delta S_{n+2k-1} \end{vmatrix}}{\begin{vmatrix} 1 & \dots & 1 \\ \Delta S_n & \dots & \Delta S_{n+k} \\ \vdots & & \vdots \\ \Delta S_{n+k-1} & \dots & \Delta S_{n+2k-1} \end{vmatrix}}. \tag{3}$$

La vitesse de convergence augmente avec k. Cette méthode était semble-t-il déjà connue de Jacobi.

3.3. Cas de convergence. Le résultat suivant permet de mieux comprendre l'origine de ces formules.

PROPOSITION 2. *Si la suite S_n est définie comme $S_n = S_\infty + c_1 r_1^n + \cdots + c_k r_k^n$, où S_∞ et les $c_i \neq 0$ sont des constantes fixées, les r_i sont distincts et différents de 1, alors pour tout n, l'accélérée de Shanks $T_{n,k}$ vaut S_∞.*

DÉMONSTRATION. La suite $u_n := \Delta S_n$ est solution d'une récurrence linéaire à coefficients constants d'ordre k, dont le polynôme caractéristique a les r_i pour racines. Cette récurrence peut s'écrire

$$a_0 u_{n+k} + \cdots + a_k u_n = 0.$$

L'équation

$$\begin{pmatrix} 1 & \dots & x^k \\ u_n & \dots & u_{n+k} \\ \vdots & & \vdots \\ u_{n+k-1} & \dots & u_{n+2k-1} \end{pmatrix} \begin{pmatrix} a_k \\ \vdots \\ a_0 \end{pmatrix} = 0$$

est satisfaite pour $x = r_1, \dots, r_k$. Pour chacune de ces valeurs, le déterminant de la matrice est donc nul. Comme c'est un polynôme de degré k, il s'agit donc d'un multiple du polynôme caractéristique de la récurrence, et donc les coefficients $a_0, \dots, a_k$ de la récurrence sont tous donnés par un même multiple des mineurs obtenus en développant ce déterminant par rapport à sa première ligne. Ce multiple est non-nul, parce que la suite ne vérifie pas de récurrence d'ordre inférieur à k, puisque les c_i sont non-nuls et les r_i distincts.

Ensuite, il suffit d'observer que par linéarité

$$a_0 S_{n+k} + \cdots + a_k S_n = (a_0 + \cdots + a_k) S_\infty,$$

ce qui explique les '1' sur la première ligne du déterminant du dénominateur. Cette dernière somme est non-nulle, puisque 1 n'est pas racine du polynôme caractéristique. □

3.4. Lien avec les approximants de Padé. Toute suite convergente (t_n) peut se voir comme l'évaluation en $x = 1$ de la série entière

$$t_0 + (t_1 - t_0)x + (t_2 - t_1)x^2 + \cdots .$$

L'introduction de cette variable supplémentaire permet d'appréhender plus facilement la convergence de la méthode de Shanks.

THÉORÈME 5. *Si $S_n = u_0 + u_1 x + \cdots + u_n x^n$, alors la valeur $T_{n,k}$ calculée par la méthode de Shanks est un approximant de Padé de type $(n+k, k)$ de la série formelle dont les troncatures sont les S_n : c'est une fraction rationnelle, dont le numérateur a degré au plus $n+k$, dont le dénominateur a degré au plus k, et*

$$T_{n,k} - S_{n+2k} = O(x^{n+2k+1}).$$

DÉMONSTRATION. La différence $\Delta S_n = S_{n+1} - S_n$ vaut $u_{n+1}x^{n+1}$, de sorte qu'après avoir simplifié la formule (3) par les puissances de x identiques, la formule devient

$$T_{n,k} = \frac{\begin{vmatrix} S_n x^k & \dots & S_{n+k} \\ u_{n+1} & \dots & u_{n+k+1} \\ \vdots & & \vdots \\ u_{n+k} & \dots & u_{n+2k} \end{vmatrix}}{\begin{vmatrix} x^k & \dots & 1 \\ u_{n+1} & \dots & u_{n+k+1} \\ \vdots & & \vdots \\ u_{n+k} & \dots & u_{n+2k} \end{vmatrix}}.$$

Le déterminant N_{n+k} du numérateur est alors clairement un polynôme de degré au plus $n+k$ et de même le déterminant D_k du dénominateur est un polynôme de degré au plus k. En multipliant la deuxième ligne de N_{n+k} par x^{n+k+1}, la troisième par x^{n+k+2} et ainsi de suite jusqu'à la dernière par x^{n+2k} et en les additionnant à la première, on obtient que le déterminant se récrit

$$N_{n+k} = \begin{vmatrix} S_{n+k}x^k & \dots & S_{n+2k} \\ u_{n+1} & \dots & u_{n+k+1} \\ \vdots & & \vdots \\ u_{n+k} & \dots & u_{n+2k} \end{vmatrix},$$

c'est-à-dire que les indices sur la première ligne sont augmentés de k. En multipliant la première ligne du dénominateur par S_{n+2k} et en soustrayant, on obtient finalement

$$S_{n+2k}D_k - N_{n+k} = \begin{vmatrix} (S_{n+2k} - S_{n+k})x^k & \dots & 0 \\ u_{n+1} & \dots & u_{n+k+1} \\ \vdots & & \vdots \\ u_{n+k} & \dots & u_{n+2k} \end{vmatrix}.$$

Les polynômes sur la première ligne sont tous divisibles par x^{n+2k+1}, ce qui donne le résultat. □

3.5. Algorithme ε de Wynn (1956). L'algorithme de Wynn réalise la transformation de Shanks par une récurrence qui évite le calcul de déterminants ou d'approximants de Padé.

$$\varepsilon_{-1}^{(n)} = 0, \quad \varepsilon_0^{(n)} = S_n, \quad \varepsilon_{k+1}^{(n)} = \varepsilon_{k-1}^{(n+1)} + \frac{1}{\varepsilon_k^{(n+1)} - \varepsilon_k^{(n)}}.$$

Nous admettons alors le résultat suivant.

PROPOSITION 3. *$\varepsilon_{2k}^{(n)}$ est le $T_{n,k}$ de la méthode de Shanks.*

4. Convergence numérique des approximants de Padé

Le succès des méthodes d'accélération de convergence peut en partie s'expliquer par la convergence des approximants de Padé.

4.1. Définitions. Il est possible de présenter ces résultats avec très peu d'analyse complexe. Il faut juste la définition d'une *fonction analytique* dans un ouvert Ω de $\mathbb{C}$ comme une fonction dont la série de Taylor en tout point de Ω a un rayon de convergence non-nul ; et d'une *fonction méromorphe* comme le quotient de deux fonctions analytiques. Un point r où une fonction $f(z)$ méromorphe n'est pas analytique est appelé un *pôle*. Pour un tel point, il existe un entier $k > 0$ appelé *ordre du pôle* tel que $(z-r)^k f(z)$ est analytique au voisinage de r. Le pôle est *simple* si son ordre est 1.

4.2. Le théorème de Montessus de Ballore. Il s'agit du premier résultat classique sur ces questions. Pour tout $r > 0$, on note $D(0,r)$ le disque ouvert de centre 0 et de rayon r.

THÉORÈME 6 (Montessus de Ballore (1902)). *Soit $f(z)$ une fonction méromorphe dans un disque $D(0,\rho)$ de rayon ρ supérieur à $R > 0$, avec exactement m pôles simples distincts $z_1, \ldots, z_m$ dans le disque, et*

$$0 < |z_1| \leq \cdots \leq |z_m| < R.$$

Soit $\{P_n(z)/Q_n(z)\}_n$ la suite d'approximants de Padé de type (n,m) de f, alors

$$\lim_{n\to\infty} \frac{P_n(z)}{Q_n(z)} = f(z)$$

uniformément dans tout sous-ensemble compact de $D(0,R) \setminus \bigcup_{i=1}^m \{z_i\}$.

La preuve originelle de ce théorème est élémentaire en ce sens qu'elle n'utilise pas d'analyse complexe, mais elle est technique et un peu longue. La preuve ci-dessous, due à Karlsson et Wallin en 1977, est bien plus courte, mais nécessite un peu d'analyse complexe.

DÉMONSTRATION. Un rôle important est joué par le polynôme $Q(z) = (z - z_1)\cdots(z - z_m)$. Par définition de l'approximant de Padé,

$$Q(z)Q_n(z)f(z) - Q(z)P_n(z) = \sum_{i>m+n} e_i z^i.$$

La première partie de la preuve consiste à montrer que les racines de Q_n convergent vers les z_i.

On choisit de normaliser P_n et Q_n de sorte que

$$\max_{|z|=\rho} |Q_n(z)| = 1 \geq \max_{|z|=R} |Q_n(z)|,$$

où la seconde inégalité provient du principe du maximum. Comme le degré de QP_n est inférieur à $m+n+1$, la borne de Cauchy donne alors

$$|e_i| \leq R^{-i} \max_{|z|=R} |Q(z)f(z)|, \qquad i > m+n.$$

En sommant terme à terme, on en déduit la majoration

$$|Q(z)Q_n(z)f(z) - Q(z)P_n(z)| \leq \max_{|z|=R} |Q(z)f(z)| \left(\frac{|z|}{R}\right)^{n+m}, \qquad |z| < R, \tag{4}$$

où le premier facteur du membre droit ne dépend ni de z ni de (m,n).

Comme $(Q_n)_n$ est une suite de polynômes de degré au plus m et bornés uniformément dans $D(0,\rho)$, on peut en extraire une sous-suite (Q_{n_j}) convergeant uniformément dans tout compact de ce disque (théorème de Montel), la limite étant alors un polynôme q non nul et de degré au plus m. L'inégalité ci-dessus entraîne alors que les P_n correspondants convergent uniformément sur les compacts de $D(0,R)$ vers une fonction analytique p, et $Qqf - Qp = 0$. Pour $z = z_i$, $Q(z_i) = 0$ mais $Q(z_i)f(z_i) \neq 0$ et donc $q(z_i) = 0$, ce qui fait de q un multiple de Q, tel que $\max_{|z|=\rho} q(z) = 1$. Par le même raisonnement, de toute sous-suite de $(Q_n)_n$, on peut extraire une sous-suite convergeant uniformément vers q, et donc la suite $(Q_n)_n$ elle-même converge uniformément vers q.

Si K est un compact de $D(0,R) \setminus \{z_1,\dots,z_m\}$, il est inclus dans un $D(0,r)$ avec $r < R$, et l'inégalité (4) montre que

$$\lim_{n\to\infty} |Q(z)Q_n(z)f(z) - Q(z)P_n(z)|^{1/n} \leq \frac{r}{R},$$

et comme en plus $QQ_n \to Qq$ uniformément dans K, où $Qq > \epsilon \neq 0$, on obtient par division que

$$\lim_{n\to\infty} \left| f(z) - \frac{P_n(z)}{Q_n(z)} \right|^{1/n} \leq \frac{r}{R} < 1,$$

uniformément dans K, ce qui donne non seulement la convergence souhaitée, mais aussi une vitesse géométrique. □

Remarque 1. Le même raisonnement s'applique au cas où les pôles ne sont pas distincts, en prenant pour degré des dénominateurs la somme des multiplicités des pôles. Lorsque le degré des dénominateurs, toujours fixé, est plus grand que cette somme, ce raisonnement s'étend pour montrer que pour une sous-suite des approximants de Padé, les autres pôles ont également une limite, et la convergence a lieu dans tout compact évitant aussi ces pôles limites.

4.3. Les approximants de Padé diagonaux. Pour de nombreuses raisons et en particulier leur lien avec la théorie des fractions continues, il est fréquent que l'on utilise des approximants de Padé diagonaux (de type (n,n)) ou "paradiagonaux" (de type $(n,n+j)$ pour un $j \in \mathbb{Z}$ fixé). Le théorème de Montessus de Ballore ne donne alors aucune indication.

Un objectif naturel est le suivant.

Conjecture 1 (Baker, Gammel, Wills (1961)). *Soit f une fonction méromorphe dans le disque $D(0,1)$ et analytique en 0. Alors il existe une sous-suite infinie de la suite d'approximants de Padé diagonaux de f qui converge uniformément dans tout compact de $D(0,1)$ privé des pôles de f.*

Malheureusement, cette conjecture est fausse. Un premier contre-exemple a été trouvé en 2001 par Lubinsky, et un contre-exemple plus simple, dû à Buslaev en 2002 est

$$f(z) = \frac{-27 + 6z^2 + 3(9+j)z^3 + \sqrt{81(3-(3+j)z^3)^2 + 4z^6}}{2z(9+9z+(9+j)z^2)},$$

où $j = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$ et la branche de la racine carrée est celle qui rend $f(0)$ nul. Cette fonction est analytique dans un disque qui contient $D(0,1)$, mais il existe un

$r \in (0,1)$ tel que tous ses approximants de Padé diagonaux d'indice supérieur à 1 ont un pôle dans $D(0,r)$.

Il faut donc ajouter des hypothèses sur la fonction f pour obtenir une version correcte de la conjecture. Plusieurs tels résultats existent. Nous n'en citons qu'un, mais dont la preuve nécessite plus d'analyse complexe que souhaité pour ces notes.

THÉORÈME 7 (Nuttall (1970)). *Soit f une fonction analytique en 0 et méromorphe dans l'ensemble du plan complexe. Alors la suite de ses approximants de Padé diagonaux (P_n/Q_n) converge* en mesure *dans les sous-ensembles compacts du plan, c'est-à-dire que pour tous $r > 0, \epsilon > 0$,*

$$\text{mes}\left\{ z \mid |z| \leq r \ \textit{et} \ \left| f(z) - \frac{P_n(z)}{Q_n(z)} \right| \geq \epsilon \right\} \to 0, \quad n \to \infty.$$

Ceci explique en particulier la convergence des approximants de Padé de la tangente étudiée dans le chapitre précédent (sans toutefois rien dire sur la *vitesse* de convergence, question encore plus délicate).

5. Fractions continues

Certains résultats de convergence sur les approximants de Padé diagonaux peuvent être obtenus grâce au lien entre ces approximants et les fractions continues.

5.1. Définitions et notation.

DÉFINITION 1. *Une* fraction continue *est un triplet de suites $((a_m),(b_m),(f_m))$ avec $f_m = S_m(0)$, où la suite (S_m) est définie par $S_0(w) = b_0 + w$, $S_m(w) = S_{m-1}(a_m/(b_m + w))$. On note*

$$f_m = b_0 + \mathbb{K}_{i=1}^{m}\left(\frac{a_i}{b_i}\right) = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{\ddots \cfrac{}{b_2 + \cfrac{}{1 + \frac{a_m}{b_m}}}}}.$$

Les (a_m) et (b_m) sont appelés les éléments *de la fraction continue ; (f_m) est sa suite de* convergents.

5.2. Récurrence. Les convergents vérifient une récurrence simple qui est la base de la plupart des preuves par récurrence du reste de ce cours.

THÉORÈME 8. *Le convergent f_m s'écrit A_m/B_m, où les suites (A_m) et (B_m) vérifient la même récurrence*

$$\begin{pmatrix} A_m \\ B_m \end{pmatrix} = a_m \begin{pmatrix} A_{m-2} \\ B_{m-2} \end{pmatrix} + b_m \begin{pmatrix} A_{m-1} \\ B_{m-1} \end{pmatrix}, \qquad m \geq 1$$

avec $A_{-1} = 1$, $A_0 = b_0$, $B_{-1} = 0$, $B_0 = 1$.

DÉMONSTRATION. La preuve est par récurrence. Pour $m = 0$ on a $f_0 = b_0 = A_0/B_0$, pour $m = 1$ on a $f_1 = b_0 + a_1/b_1$, ce qui correspond bien à ce que donne la récurrence, à savoir $A_1 = a_1 + b_1 b_0$, $B_1 = b_1$. Ensuite, on observe que f_{m+1} s'obtient

en considérant f_m comme une fonction de b_m et en l'évaluant en $b_m + a_{m+1}/b_{m+1}$. La récurrence donne alors

$$\begin{aligned}\frac{A_{m+1}}{B_{m+1}} &= \frac{a_m A_{m-2} + \left(b_m + \frac{a_{m+1}}{b_{m+1}}\right) A_{m-1}}{a_m B_{m-2} + \left(b_m + \frac{a_{m+1}}{b_{m+1}}\right) B_{m-1}} \\ &= \frac{A_m + \frac{a_{m+1}}{b_{m+1}} A_{m-1}}{B_m + \frac{a_{m+1}}{b_{m+1}} B_{m-1}}.\end{aligned}$$

La première égalité est obtenue en remplaçant b_m dans f_m donné par l'hypothèse de récurrence, la seconde en réutilisant encore cette hypothèse pour faire apparaître A_m et B_m. □

Exercice 6. Montrer que le même raisonnement donne aussi

$$S_m(w) = \frac{A_m + wA_{m-1}}{B_m + wB_{m-1}}.$$

5.3. C-fractions.

Définition 2. *Une fraction continue de la forme*

$$b_0 + \mathbb{K}_{m=1}^{\infty} \left(\frac{a_m z^{\alpha_m}}{1} \right)$$

avec $a_m \neq 0$ *et* $\alpha_m \in \mathbb{N}^\star$ *est appelée* C-fraction.

Si

$$f^{[0]} = b_0 + c^{[1]} z^{\alpha_1} + \cdots$$

est une série entière, alors on peut la récrire $f^{[0]} = b_0 + c^{[1]} z^{\alpha_1} g_1$, où g_1 est une série entière telle que $g_1(0) = 1$. Donc $1/g_1 - 1$ est également une série entière $f^{[1]}$, telle que $f^{[1]}(0) = 0$. On a ainsi obtenu une série entière telle que

$$f^{[0]} = b_0 + \frac{c^{[1]} z^{\alpha_1}}{1 + f^{[1]}}, \quad f^{[1]} = \frac{c^{[1]} z^{\alpha_1}}{f^{[0]} - b_0} - 1 =: c^{[2]} z^{\alpha_2} + \cdots .$$

Ensuite, si $f^{[n-1]}$ est non-nul, on peut recommencer et plus généralement définir

$$f^{[n]} = \frac{c^{[n]} z^{\alpha_n}}{f^{[n-1]}} - 1 = c^{[n+1]} z^{\alpha_{n+1}} + \cdots$$

dès lors que $f^{[n-1]}$ est non-nul.

Ainsi, à toute série $f^{[0]}$ correspond une fraction continue. Cette construction s'arrête si l'un des $f^{[n]}$ est nul, et alors la série était le développement en série d'une fraction rationnelle qui vaut exactement le convergent de la fraction obtenue. Si elle ne s'arrête pas, on obtient un développement en fraction continue infinie, pour lequel on contrôle parfaitement la convergence.

5.3.1. *Quelques exemples.* Des fractions continues classiques sont

— la tangente

$$z \tan z = \cfrac{z^2}{1 - \cfrac{z^2/(1 \cdot 3)}{1 - \cfrac{z^2/(3 \cdot 5)}{\ddots}}},$$

qui converge pour $z \in \mathbb{C} \setminus \{(2k+1)\pi/2 \mid k \in \mathbb{Z}\}$;

— l'exponentielle

$$\exp(z) = 1 + \cfrac{z}{1 - \cfrac{z/(2\cdot 1)}{1 + \cfrac{z/(2\cdot 3)}{1 - \cfrac{z/(2\cdot 3)}{1 + \cfrac{z/(2\cdot 5)}{\ddots}}}}},$$

qui converge pour tout $z \in \mathbb{C}$;

— le logarithme

$$\log(1+z) = \cfrac{z}{1 + \cfrac{z\frac{1\cdot 2}{2\cdot 3}}{1 + \cfrac{z\frac{1\cdot 2}{3\cdot 4}}{1 + \cfrac{z\frac{2\cdot 3}{4\cdot 5}}{1 + \cfrac{z\frac{2\cdot 3}{5\cdot 6}}{\ddots}}}}}$$

qui converge pour $z \in \mathbb{C} \setminus \mathbb{R}^-$.

5.3.2. *Lien avec les approximants de Padé.*

PROPOSITION 4. *Avec la construction ci-dessus, le convergent*

$$\frac{A_m}{B_m} = b_0 + \mathbb{K}_{i=1}^m \left(\frac{c^{[m]} z^{\alpha_m}}{1} \right)$$

est une fraction rationnelle telle que

$$f^{[0]} - \frac{A_m}{B_m} = O(z^{\alpha_1 + \cdots + \alpha_{m+1}}).$$

Lorsque tous les exposants α_i valent 1, alors A_{2m} et B_{2m} ont degré au plus m.

DÉMONSTRATION. Comme (par récurrence), $B_m(0) \neq 0$, il suffit de prouver la propriété pour $B_m f^{[0]} - A_m$. Par récurrence encore, on a d'abord $A_0 = b_0 B_0$ et donc $B_0 f^{[0]} - A_0 = O(z^{\alpha_1})$. Ensuite, on utilise le fait que

$$f^{[0]} = S_m(f^{[m]}) = \frac{A_m + f^{[m]} A_{m-1}}{B_m + f^{[m]} B_{m-1}}$$

que l'on inverse pour obtenir

$$f^{[m]} = -\frac{A_m - f^{[0]} B_m}{A_{m-1} - f^{[0]} B_{m-1}} = O(z^{\alpha_{m+1}}),$$

et donc $A_m - f^{[0]} B_m = O((A_{m-1} - f^{[0]} B_{m-1}) z^{\alpha_{m+1}})$, et le résultat en découle par récurrence, de même que les degrés des A_i et B_i. □

À ce stade, on a établi un lien entre les fractions continues, les approximants de Padé, et les accélérées de Shanks.

6. Convergence numérique

Il reste à comprendre comment on détermine si une fraction continue converge ou non. Il existe des conditions nécessaires et suffisantes (mais peu commodes d'emploi) et quelques conditions suffisantes simples décrites ici.

6.1. Expression en série numérique.

PROPOSITION 5. *Le n^e convergent de $\mathbb{K}_{m=1}^{\infty}\left(a_m/b_m\right)$ vaut*

$$\frac{A_n}{B_n} = \frac{a_1}{B_1} - \frac{a_1 a_2}{B_1 B_2} + \cdots + (-1)^{n+1}\frac{a_1 \cdots a_m}{B_{n-1}B_n}.$$

DÉMONSTRATION. Par récurrence. Pour $n = 1$ on retrouve $A_1/B_1 = a_1/b_1$. Ensuite,

$$\frac{A_{n+1}}{B_{n+1}} - \frac{A_n}{B_n} = \frac{A_{n+1}B_n - A_n B_{n+1}}{B_n B_{n+1}}$$

a le dénominateur souhaité. Pour le numérateur, on reconnaît le déterminant de

$$\begin{pmatrix} A_{n+1} & B_{n+1} \\ A_n & B_n \end{pmatrix} = \begin{pmatrix} b_{n+1} & a_{n+1} \\ 1 & 0 \end{pmatrix}\begin{pmatrix} A_n & B_n \\ A_{n-1} & B_{n-1} \end{pmatrix}.$$

La conclusion s'obtient en prenant les déterminants dans cette identité et en observant que $A_1B_0 - A_0B_1 = 1$. □

6.2. Le cas des coefficients positifs.

COROLLAIRE 1. *Lorsque tous les coefficients a_m et b_m sont positifs, alors les convergents $f_m = A_m/B_m$ vérifient*

$$f_2 < f_4 < \cdots < f_{2n} < f_{2n+1} < f_{2n-1} < \cdots < f_1.$$

Donc les deux suites (f_{2m}) et (f_{2m+1}) convergent. Lorsqu'elles ont la même limite, elle est dans l'intervalle (f_{2m}, f_{2m+1}) pour tout m.

DÉMONSTRATION. D'abord, la positivité des éléments entraîne la positivité des numérateurs et dénominateurs des convergents, comme le montre la récurrence qu'ils satisfont. Ensuite, en utilisant la somme de deux termes consécutifs dans la série de la proposition ci-dessus, on obtient

$$\begin{aligned} f_{m+1} - f_{m-1} &= \frac{(-1)^{m+1}a_1 \ldots a_m}{B_{m-1}B_m} + \frac{(-1)^m a_1 \ldots a_{m+1}}{B_m B_{m+1}} \\ &= \frac{(-1)^{m+1}a_1 \ldots a_m}{B_m}\left(\frac{1}{B_{m-1}} - \frac{a_{m+1}}{B_{m+1}}\right), \end{aligned}$$

mais $B_{m+1} = a_{m+1}B_{m-1} + b_{m+1}B_m > a_{m+1}B_{m-1}$ par positivité, ce qui montre que le terme entre parenthèses est positif et donc que le signe de $f_{m+1} - f_{m-1}$ est celui de $(-1)^{m+1}$, ce qui conclut la première partie de la preuve. Ensuite (f_{2m+1}) est une suite croissante bornée et donc converge, et de même pour f_{2m}. Le dernier énoncé est clair. □

6.3. Le cas des coefficients complexes. Nous donnons sans preuve l'énoncé ci-dessous qui a le mérite d'être simple d'emploi.

THÉORÈME 9 (Worpitzky (1865)). *Si les nombres complexes a_k vérifient $|a_k| \leq 1/4$, alors la fraction continue $\mathbb{K}_{m=1}^{\infty}\left(a_k/1\right)$ converge.*

EXEMPLE 4. Pour tout $z \in \mathbb{C}$, dès que k est assez grand, $|z^2/((2k{+}1)(2k{+}3))| \leq 1/4$, donc la partie d'indice supérieur à un tel k de la fraction continue de $z \tan z$ converge. L'ensemble de la fraction continue converge donc (éventuellement vers l'infini).

TP 3

Calcul numérique par accélération de convergence

1. Échauffement

1. Écrire une procédure prenant en entrée une liste L et un réel r, appliquant une étape d'accélération d'Euler à L avec cette valeur de r, et renvoyant la liste ainsi obtenue ;
2. Écrire une seconde procédure prenant à nouveau L, mais une liste de valeurs $[r_1, r_2, \dots]$, effectuant les accélérations d'Euler successivement avec ces valeurs dans l'ordre tant que c'est possible, et renvoyant un tableau des valeurs obtenues à chaque étape ;
3. Appliquer ces procédures à l'accélération du calcul d'Archimède, c'est-à-dire en partant des valeurs $3\,2^k \sin(\pi/(3\,2^k))$ et $3\,2^k \tan(\pi/(3\,2^k))$ pour $k = 1, \dots, 5$.
4. Utiliser cette méthode avec $r = -1$ pour calculer des décimales de $\ln 2$ donné comme limite de la suite
$$S_n = \sum_{i=0}^{n} \frac{(-1)^i}{i+1}.$$
On demande de calculer les 4 premières décimales de $\ln 2$ en partant seulement des 16 premières valeurs $S_0, \dots, S_{15}$.
5. Appliquer les mêmes opérations (avec un r adapté) sur la suite pourtant divergente
$$\sum_{i=0}^{n} (-1)^i \frac{2^{i+1}}{i+1}.$$
Comparer à $\ln 3$.
6. Adapter ces techniques pour calculer une dizaine de décimales de la constante γ d'Euler, définie comme limite de la suite
$$S_n = \sum_{k=1}^{n} \frac{1}{k} - \log n.$$
7. Appliquer répétitivement la méthode Δ^2 d'Aitken aux 11 premiers termes de la formule de Leibniz pour π :
$$\pi = 4 - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \cdots.$$

2. Évaluation conjecturale d'une somme

La somme

$$S := \lim_{N\to\infty} S_N \qquad \text{où} \qquad S_N := \sum_{k=1}^{N} \frac{(-1)^{k+1}}{k^2} \sum_{i=1}^{k} \frac{(-1)^{i+1}}{i}$$

est issue d'une famille de sommes similaires dont l'étude remonte à Euler. L'objet de cet exercice est d'employer d'abord des méthodes d'accélération de convergence, puis des outils de conjecture pour obtenir une formule particulièrement simple pour S.

8. Estimer l'ordre de grandeur du nombre de termes à utiliser pour calculer 30 décimales de la somme *sans* accélération de convergence. (Utiliser le caractère alterné de la somme).
9. Calculer les 1100 premiers termes de la suite S_n et déterminer une méthode d'accélération de convergence adaptée. Estimer empiriquement le nombre de décimales de S ainsi obtenu.
10. En écrivant une procédure qui n'effectue pas trop d'opérations pour calculer S_n, calculer S_{2^m} pour m jusqu'à 15.
11. Appliquer alors une accélération de convergence comme en question 9.
12. À l'aide de la valeur ainsi obtenue, il est également possible de conjecturer *une forme close* pour S. Pour cela, utiliser la fonction `identify`, dont le principe sera vu dans un cours suivant. Pour aider `identify`, il faut exploiter (par l'option `BasisSumConst`) le fait que la somme

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^2} \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i}$$

admet une expression simple, que Maple peut calculer, et dont une combinaison linéaire avec S a une forme simple.

Finalement, ce calcul a permis d'obtenir suffisamment de décimales pour arriver à la conjecture suivante [1]

$$S = \frac{1}{4}\pi^2 \ln(2) - \frac{5}{8}\zeta(3).$$

1. Des preuves de cette identité et d'autres similaires existent (R. Sitaramachandrarao. A formula of S. Ramanujan, Journal of Number Theory 25 (1987), no. 1, 1–19. P. Flajolet and B. Salvy, Euler sums and contour integral representations, Experimental Mathematics 7 (1998), no. 1, 15–35.), mais aucune n'est vraiment simple.

COURS 4

Du flottant à la forme close : LLL

1. Introduction

L'algorithme LLL, dont le nom provient des initiales de ses inventeurs (A. Lenstra, H. Lenstra, et L. Lovász) est un algorithme de réduction des réseaux, dont une des premières applications a été la factorisation de polynômes en une variable à coefficients dans $\mathbb{Q}$. Il s'avère qu'il donne en temps polynomial une réponse approchée à de nombreux problèmes difficiles.

En mathématiques expérimentales, il est précieux pour trouver des dépendances linéaires entre des constantes réelles approchées.

2. Réseaux euclidiens

DÉFINITION 1. *Un* réseau $\mathcal{L}$ *de dimension* d *de* $\mathbb{R}^n$ *est l'ensemble des combinaisons linéaires à coefficients entiers de vecteurs* $(b_1, \dots, b_d)$ *linéairement indépendants sur* $\mathbb{R}$.

Dans tout ce chapitre on utilisera la convention suivante : les vecteurs d'une base seront notés en lettres minuscules (comme $b_1, \dots, b_d$) et la lettre majuscule correspondante (ici B) sera utilisée pour noter la matrice dont les b_i sont les colonnes.

PROPOSITION 1. *Soeint* $(e_1, \dots, e_k)$ *et* $(f_1, \dots, f_\ell)$ *deux bases d'un réseau* $\mathcal{L}$ *de* $\mathbb{R}^n$. *Alors* $k = \ell$ *et il existe une matrice* $U \in \mathrm{GL}_k(\mathbb{Z})$ *telle que* $E = FU$.

DÉMONSTRATION. L'espace vectoriel sur $\mathbb{R}$ engendré par l'ensemble des éléments de $\mathcal{L}$ est

$$\mathrm{Vect}(\mathcal{L}) = \mathrm{Vect}(e_1, \dots, e_k) = \mathrm{Vect}(f_1, \dots, f_\ell).$$

L'indépendance linéaire des e_i d'une part et des f_i de l'autre entraîne alors $k = \ell$.

Ensuite, comme e_i appartient au réseau engendré par les f_j, il existe $u_{1i}, \dots, u_{ki}$ des entiers tels que

$$e_i = u_{1i} f_1 + \cdots + u_{ki} f_k,$$

ou matriciellement $E = FU$ et il reste à prouver que U est inversible. Le même raisonnement montre l'existence d'une matrice V à coefficients entiers telle que $F = EV$. Alors $F = FUV$ et comme F est de rang k, on peut simplifier par F dans cette identité, ce qui donne $UV = \mathrm{Id}$. □

DÉFINITION 2. *Soit* $\mathcal{L}$ *un réseau de base* B. *On appelle volume du réseau la quantité* $\mathrm{vol}\, L = \sqrt{\det({}^tBB)}$.

Cette définition ne dépend pas du choix de la base : si E est une autre base, alors il existe $U \in \mathrm{GL}_d(\mathbb{Z})$ inversible donc de déterminant ± 1 telle que $E = BU$ et

${}^tEE = {}^tU({}^tBB)U$. Parenthésé ainsi, il s'agit d'un produit de trois matrices carrées, et on a bien le résultat souhaité sur le produit des déterminants.

Quand le rang du réseau est n, on a bien sûr $\mathrm{vol}(\mathcal{L}) = |\det(B)|$.

2.1. Problèmes difficiles. Voici trois problèmes classiques liés aux réseaux.

- CVP : Étant donnés une base $\mathcal{B}$ d'un réseau de $\mathbb{Z}^n$ et un vecteur $Y \in \mathbb{Z}^n$, trouver un vecteur $X = (x_1, \dots, x_d) \in \mathbb{Z}^d$ tel que $\|Y - \sum_{i=1}^d x_i b_i\|$ soit minimal.
- SVP : Étant donnée une base $\mathcal{B}$ d'un réseau de $\mathbb{Z}^n$, trouver un vecteur non nul $X = (x_1, \dots, x_d) \in \mathbb{Z}^d$ tel que $\|\sum_{i=1}^d x_i b_i\|$ soit minimal.
- SubsetSum : Étant donnés des entiers $a_1, \dots, a_n$ et y, trouver, s'ils existent, $x_1, \dots, x_n$ dans $\{0, 1\}$ tels que $y = x_1 a_1 + \cdots + x_n a_n$.

Tous ces problèmes sont NP-difficiles, c'est-à-dire qu'il sont au moins aussi durs que tout problème de la classe NP. Le dernier appartient en outre à cette classe, ce qui le rend potentiellement plus facile, mais la grande question P=NP est équivalente à savoir s'il existe ou non un algorithme de complexité polynomiale pour le résoudre.

2.2. Résultat principal. Le résultat spectaculaire est le suivant.

THÉORÈME 10 (Lenstra-Lenstra-Lovász 1982). *On peut trouver un vecteur v non nul dans $\mathcal{L}$ tel que*

$$\|v\| \leq 2^{\frac{d-1}{2}} \min_{e \in \mathcal{L} \setminus \{0\}} \|e\|$$

en temps polynomial en n et en $\log \max \|b_j\|$.

Avant de prouver ce résultat, nous passons en revue quelques applications de l'algorithme.

3. Applications

3.1. Somme de sous-ensembles (SubsetSum). On considère d'abord le réseau de base

$$B = \begin{pmatrix} a_1 & a_2 & \dots & a_n & y \\ 2 & 0 & & 0 & 1 \\ 0 & 2 & & \vdots & \vdots \\ \vdots & & \ddots & & \\ 0 & 0 & \dots & 2 & 1 \end{pmatrix}.$$

Si le problème admet une solution alors ce réseau contient $B \cdot {}^t(x_1, \dots, x_n, -1) = {}^t(0, 2x_1 - 1, \dots, 2x_n - 1)$ de norme $\sqrt{n}$. Les autres vecteurs du réseau ont une norme qui satisfait

$$\|BX\|^2 = \underbrace{\Big(y - \sum a_i x_i\Big)^2}_{>0} + \sum \underbrace{(2x_i - 1)^2}_{\geq 1} \geq n + 1.$$

La différence n'est pas très grande, mais si maintenant on multiplie d'abord les a_i et y par une constante $C > 0$ alors la première estimation ne change pas, alors que la seconde devient $n + C^2$. En choisissant $C > 2^{(n-1)/2}\sqrt{n}$, l'algorithme renvoie alors un vecteur de norme inférieure à

$$2^{\frac{n-1}{2}}\sqrt{n} < C < \sqrt{n + C^2}$$

qui ne peut donc qu'être une solution.

En pratique, on utilisera un C plus petit et on regardera si les coordonnées du vecteur trouvé sont dans $\{-1, 0, 1\}$, exceptée éventuellement la première.

3.2. Relations linéaires. En entrée on a des réels $\alpha_1, \ldots, \alpha_n$ et $K > 0$ tels qu'il existe des entiers $p_i \in [-K, K]$ avec

$$p_1\alpha_1 + \cdots + p_n\alpha_n = 0$$

et il s'agit de calculer de tels p_i.

Le réseau à considérer a une matrice $(n+1) \times n$ dont la première ligne est

$$([C\alpha_1], [C\alpha_2], \ldots, [C\alpha_n]),$$

le reste étant un bloc identité. La notation $[x]$ signifie l'entier le plus proche de x, et la constante C reste à ajuster.

Une combinaison linéaire par $(p_1, \ldots, p_n)$ donne un vecteur v tel que

$$\|v\|^2 = p_1^2 + \cdots + p_n^2 + (\sum p_i[C\alpha_i])^2.$$

Comme pour tout i on a $C\alpha_i - 1/2 \leq [C\alpha_i] \leq C\alpha_i + 1/2$ en sommant on obtient

$$\left|\sum p_i[C\alpha_i]\right| \leq \frac{1}{2}\sum |p_i| \leq n\frac{K}{2}.$$

Le réseau contient donc un vecteur de norme au carré inférieure à $K^2(n + n^2/4)$.

L'algorithme LLL renvoie donc un vecteur de norme au plus $2^{n/2}K\sqrt{n + n^2/4}$ et il s'agit de choisir C assez grand pour que ce vecteur soit associé à une solution. Soit donc $(q_1, \ldots, q_n)$ un n-uplet d'entiers tel que $q_1\alpha_1 + \cdots + q_n\alpha_n$ ne soit pas nul et soit de valeur absolue M minimale parmi toutes les valeurs non nulles possibles, et, quitte à prendre l'opposé du vecteur, on peut supposer que la valeur de la somme est M et non $-M$. Le vecteur w correspondant dans le réseau a alors norme au carré égale à

$$\|w\|^2 = q_1^2 + \cdots + q_n^2 + (\sum q_i[C\alpha_i])^2.$$

Le même raisonnement que pour les p_i montre

$$\sum q_i[C\alpha_i] \geq C\sum q_i\alpha_i - \frac{1}{2}\sum |q_i| \geq CM - nm/2,$$

où $m = \max |q_i|$, où la dernière expression est positive pour C assez grand. Alors, $\|w\|^2 \geq (CM - nm/2)^2 + m^2$, quantité qui est minimisée pour $m = 2CMn/(n^2+4)$, ce qui entraîne

$$\|w\|^2 \geq \frac{4C^2M^2}{n^2+4}.$$

Dès que C est tel que cette valeur soit supérieure à $2^nK^2(n + n^2/4)$, alors le vecteur renvoyé par l'algorithme ne peut donc être qu'un vecteur solution.

Il faut noter qu'en pratique on ne connaît pas M, ni même K, mais qu'on cherche des relations "anormalement" petites.

3.3. Approximation simultanée. Le problème prend en entrée des réels $\alpha_1, \ldots, \alpha_n$ et $\epsilon > 0$ et il s'agit de trouver des entiers $p_1, \ldots, p_n$ et q tels que

$$|p_i - q\alpha_i| \leq \frac{3}{2}\epsilon, \quad 1 \leq q \leq 2^{\frac{n(n+1)}{4}}\epsilon^{-n}.$$

La matrice à considérer est de taille $(n+1) \times (n+1)$. Elle comporte un bloc $C\mathrm{Id}$ au dessus d'une ligne à 0, et est bordée à droite par un vecteur colonne $([C\alpha_1], \ldots, [C\alpha_n], 1)$. Par le même type de raisonnement le vecteur renvoyé

$(r_1, \dots, r_n, -q)$ premet de reconstruire les p_i par $r_i = Cp_i - q[C\alpha_i]$. Le choix $C\lceil 2^{\frac{n(n+1)}{4}}\epsilon^{-n-1}\rceil$ fournit le résultat.

4. L'algorithme LLL

4.1. Orthogonalisation de Gram-Schmidt. C'est l'opération classique qui transforme un d-uplet $(f_1, \dots, f_d)$ de vecteurs linéairement indépendants en une base orthogonale $(f_1^*, \dots, f_d^*)$ du même espace par projections et soustractions successives :

$$f_j = f_j^* + \sum_{i=1}^{j-1} \mu_{i,j} f_i^*, \quad \text{où} \quad \mu_{i,j} = \frac{(f_j, f_i^*)}{(f_i^*, f_i^*)}.$$

La matrice M triangulaire supérieure avec des 1 sur la diagonale et $M_{i,j} = \mu_{i,j}$ si $i < j$ fournit le changement de base $F = F^*M$.

L'idée de l'algorithme LLL est de trouver une base de L presque orthogonale, puis de choisir le plus petit vecteur de cette base. La base F^* ne convient pas, puisqu'en général les $\mu_{i,j}$ ne sont pas des entiers.

Proposition 2. *Soit $(f_j)_{j\le k}$ une base du réseau $\mathcal{L} \subset \mathbb{R}^n$. Alors $\min_{i\le k} \|f_i^*\| \le \|f\|$ pour tout $f \in \mathcal{L} \setminus \{0\}$.*

Démonstration. On écrit $f = \sum_{j=1}^k \lambda_j f_j$ avec $\lambda_j \in \mathbb{Z}$. Soit $l \le k$ le plus grand indice tel que $\lambda_l \neq 0$. Alors $f = \lambda_l f_l^* + v$ avec $v \in \mathrm{Vect}(f_1^*, \dots, f_{l-1}^*)$ qui est donc orthogonal à f_l^*. Donc

$$\|f\|^2 = \lambda_l^2 \|f_l^*\|^2 + \|v\|^2 \ge \|f_l^*\|^2 \ge \min \|f_i^*\|^2.$$

□

Proposition 3 (Inégalité de Hadamard). *Si $\mathcal{L} \subset \mathbb{R}^n$ a pour base F*

$$\mathrm{vol}(\mathcal{L}) = \sqrt{\det({}^tF \cdot F)} = \prod_{j=1}^k \|f_j^*\| \le \prod_{j=1}^k \|f_j\|.$$

Démonstration. On a $\det({}^tF.F) = \det({}^tM({}^tF^*F^*)M) = \det(M)^2 \det({}^tF^*F^*)$. Or $\det(M) = 1$, et ${}^tF^*F^* = \mathrm{diag}(\|f_j\|^2)$, car la base (f_j^*) est orthogonale. D'où l'égalité dans la proposition. Pour l'inégalité il suffit de voir que $\|f_j^*\| \le \|f_j\|$, puisque f_j^* est un projeté orthogonal de f_j. □

4.2. Bases propres.

Définition 3. *Une base (f_j) d'un réseau est dite propre si $|\mu_{i,j}| \le \frac{1}{2}$ pour tout $i < j$.*

Proposition 4 (Réduction faible). *Pour toute base $(f_j)_{j\le k}$ d'un réseau $\mathcal{L}$, il existe une base propre $(g_j)_{j\le k}$ de $\mathcal{L}$ telle que $g_j^* = f_j^*$ pour tout $j \le k$.*

Démonstration. Voir l'algorithme ReductionFaible.

La preuve de la correction se fait par récurrence ; elle est laissée au lecteur. Noter qu'on maintient l'invariant $G = F^*M$. □

On déduit un algorithme BasePropre(F) qui calcule d'abord la matrice F^* puis effectue une réduction faible.

Algorithme 1 Algorithme ReductionFaible

pour $i = 1$ à k **faire**
  $g_j := f_j$;
fin pour
pour $i = 2$ à k **faire**
  pour $j = (i-1)$ à 1 **faire**
    $g_i := g_i - [\mu_{i,j}]g_j$; $\mu_{i,j} := \mu_{i,j} - [\mu_{i,j}]$
  fin pour
fin pour

5. Base réduite

Définition 4. *Une base réduite de $\mathcal{L}$ est une base propre $(f_i)_{i\leq k}$ telle que pour chaque i, on ait $\|f_i^*\|^2 \leq 2\|f_{i+1}^*\|^2$.*

Proposition 5. *Si (f_i) est réduite, alors pour tout $f \in \mathcal{L}\setminus\{0\}$, on a $\|f_1\| \leq 2^{\frac{k-1}{2}}\|f\|$.*

Démonstration. La proposition 2 implique que

$$\|f\| \geq \min_{j\leq k}\|f_j^*\| \geq \min_{j\leq k} 2^{-\frac{j-1}{2}}\|f_1^*\| \geq 2^{-\frac{k-1}{2}}\|f_1\|,$$

car $f_1^* = f_1$. □

Pour répondre à notre problème, il suffit donc de donner un algorithme de réduction de la base et renvoyer le premier vecteur de la base.

Voici enfin l'algorithme LLL.

Algorithme 2 Algorithme LLL(F)

F :=BasePropre(F) ;
tant que G n'est pas réduite **faire**
  échanger f_i et f_{i+1} pour un i où $\|f_i^*\|^2 > 2\|f_{i+1}^*\|^2$;
  F :=BasePropre(F) ;
fin tant que

Proposition 6. *LLL termine, et calcule bien une base réduite de $\mathcal{L}$ en temps polynomial.*

Démonstration. La correction de l'algorithme est évidente. Le point crucial est la terminaison. On introduit le variant $\mathcal{V}(F) = \prod_{j=1}^{k} V_j(F)$, où $V_j(F) = \mathrm{vol}(f_1,\ldots,f_j) = \prod_{i=1}^{j}\|f_i^*\|$. Le carré de $\mathcal{V}(F)$ est un entier, qui n'est pas modifié lors d'une étape de réduction faible, car il ne dépend que des f_j^*, qui ne sont pas modifiés. Il suffit donc de montrer qu'il décroît lors des échanges.

Lors d'une étape d'échange, seul V_i est modifié. En notant (g_j) la famille obtenue à partir de (f_j) par l'échange, on obtient

$$g_i^* = g_i - \sum_{j=1}^{i-1}\frac{(g_i,g_j^*)}{(g_j^*,g_j^*)}g_j^* = f_{i+1} - \sum_{j=1}^{i-1}\frac{(f_{i+1},f_j^*)}{(f_j^*,f_j^*)}f_j^* = f_{i+1}^* + \mu_{i,i+1}f_i^*.$$

Ainsi, $\|g_i^*\|^2 = \|f_{i+1}^*\|^2 + \mu_{i,i+1}^2\|f_i^*\|^2 \leq ((1/2) + (1/2)^2)\|f_i^*\|^2 = \frac{3}{4}\|f_i^*\|^2$. Finalement, on obtient $\frac{V_i(g)}{V_i(f)} = \frac{\|g_i^*\|}{\|f_i^*\|} \leq \frac{\sqrt{3}}{2}$. D'où

$$\mathcal{V}(g)^2 \leq \frac{3}{4}\mathcal{V}(f)^2.$$

Ceci conclut quand à la terminaison. Il ne manque pas grand chose pour obtenir la complexité : simplement de remarquer qu'initialement,

$$\mathcal{V}(F) = \|f_1^*\|^k \dots \|f_k^*\|^1 \leq \|f_1\|^k \cdots \|f_k\|^1 \leq M^{\frac{k(k+1)}{2}},$$

où $M = \max \|f_j\|$. Ainsi, le nombre d'itérations est borné par

$$\log_{4/3}(M^{\frac{k(k+1)}{2}}) = O(\log(M)n^2).$$

Comme l'algorithme BasePropre est lui-même polynomial en $\log(M)$ et n, on obtient finalement bien que l'algorithme LLL est polynomial. □

TP 4

Les décimales de π en base 16

La fascination pour les décimales de π a mené à de nombreuses expérimentations. En 1995, Simon Plouffe fait ainsi la découverte d'une formule étonnante :

$$\pi = \sum_{i=0}^{\infty} \frac{1}{16^i} \left(\frac{4}{8i+1} - \frac{2}{8i+4} - \frac{1}{8i+5} - \frac{1}{8i+6} \right). \tag{E}$$

Cette formule permet de calculer le $n$ième chiffre en base 16 de π (et même quelques chiffres à partir du $n$ième) sans calculer les chiffres précédents, avec peu de mémoire. Il est donc utile de comprendre comment découvrir des formules de ce type. Dans l'article écrit par Simon Plouffe avec David Bailey et Peter Borwein où (E) est présentée et prouvée, sa découverte est décrite comme le résultat de « divination inspirée et recherche intensive ». Le but du TP est de montrer comment une telle recherche peut être menée, et indiquer comment cette identité et des identités similaires peuvent aussi être prouvées automatiquement. S'il reste du temps, l'utilisation de cette formule pour le calcul de décimales lointaines pourra aussi être abordée.

1. Découvertes automatiques

Le principe est simple, il s'agit de calculer numériquement un certain nombre de constantes et d'utiliser LLL pour trouver une relation linéaire entre elles.

1. Calculer avec cent décimales de précision les constantes
$$\Sigma_j := \sum_{i=0}^{\infty} \frac{16^{-i}}{8i+j}, \qquad j = 1, \ldots, 8.$$
2. Utiliser l'algorithme LLL pour « découvrir » la relation (E). (`?IntegerRelations,LinearDependency`)
3. Écrire une procédure prenant en argument une constante, une liste d'expressions, une précision, et renvoyant l'identité que suggère LLL utilisé sur cette constante et ces expressions évaluées à cette précision.
4. Utiliser cette procédure pour conjecturer des identités pour $\ln 2$, $\ln 3$, $\ln 5$, $\arctan 2$, $\arctan 3$, $\sqrt{2}\arctan(1/\sqrt{2})$, $\sqrt{2}\ln(1+\sqrt{2})$.

D'autres identités peuvent être conjecturées avec des jeux de constantes différents.

5. Obtenir des conjectures pour les expressions de π^2, $\ln 7$, $\ln^2 2$ en fonction des séries
$$\sum_{i=1}^{\infty} \frac{2^{-ji}}{i^m}, \qquad j = 1, \ldots, 6, \quad m = 1, \ldots, 5.$$

2. Preuves

La partie difficile du travail se situe dans la découverte : trouver la bonne classe de constantes dans laquelle l'identité a des chances d'exister. La phase de preuve est plus facile et plusieurs preuves existent. La méthode utilisée ci-dessous suggère aussi les constantes qui ont des chances d'être obtenues.

6. Calculer une forme close pour les sommes
$$S_j(z) := \sum_{i=0}^{\infty} \frac{z^{8i+j}}{8i+j}, \qquad j = 1, \ldots, 8.$$
Pour aider Maple dans cette sommation, il pourra être utile de calculer d'abord les sommes des dérivées, puis d'intégrer.
7. En déduire des expressions symboliques pour les Σ_j, puis une preuve de (E). Les autres sommes trouvées en question (5) se prouvent de la même manière.

3. Calcul rapide de décimales

Le calcul se déroule de la même manière pour tous ces nombres se décomposant en combinaison linéaire à coefficients entiers de séries du type
$$S = \sum_{i=0}^{\infty} \frac{p(i) \cdot b^{-i}}{q(i)},$$
où p et q sont des polynômes à coefficients entiers. Les chiffres en base b de S à partir du Nième sont donnés par la partie fractionnaire de $b^N S$, que nous noterons $b^N S \bmod 1$. La somme se décompose pour donner
$$b^N S \bmod 1 = \sum_{i=0}^{N} \frac{p(i) \cdot b^{N-i} \bmod q(i)}{q(i)} + \sum_{i>N} \frac{p(i)}{b^{i-N} q(i)} \bmod 1.$$
La seconde somme converge géométriquement et peu de termes sont nécessaires pour obtenir des décimales. La première s'évalue rapidement en calculant chacun des sommants par exponentiation binaire.

8. Écrire une procédure prenant deux entiers m et k et calculant $16^m \bmod k$ par exponentiation binaire ;
9. Écrire deux procédures prenant deux entiers N et p, un polynôme q et sa variable k et renvoyant les premières décimales de
$$\sum_{i=0}^{N} \frac{p \cdot 16^{N-i} \bmod q(i)}{q(i)} \qquad \text{et de} \qquad \sum_{i>N} \frac{p}{16^{i-N} q(i)}.$$
10. Écrire enfin une procédure prenant en entrée des entiers $N, b_1, \ldots, b_8$ et renvoyant les premières décimales en base 16 à partir de la Nième de
$$\sum_{i \geq 0} \frac{1}{16^i} \sum_{j=1}^{8} \frac{b_j}{8i+j}.$$
Tester cette procédure sur π.
11. Estimer la complexité de ce calcul en fonction de N.

Troisième partie

Preuves automatiques

COURS 5

Résultants

Résumé

Dans le cadre de l'utilisation des polynômes comme structures de données pour représenter leurs solutions, les résultants fornissent des outils efficaces pour les opérations principales.

Dans ce cours, $\mathbb{A}$ désigne un anneau commutatif. Deux polynômes A et B de $\mathbb{A}[X]$ interviendront souvent ; ils seront notés

$$A(X) = a_m X^m + \cdots + a_0, \qquad B(X) = b_n X^n + \cdots + b_0,$$

avec $a_m \neq 0$ et $b_n \neq 0$.

1. Définition

1.1. Matrice de Sylvester.

Définition 1. *Avec les notations ci-dessus, la* matrice de Sylvester *de A et B est la matrice*

$$\mathrm{Syl}(A,B) = \begin{pmatrix} a_m & \cdots\cdots\cdots\cdots & a_0 & & 0 \\ & \ddots & & \ddots & \\ 0 & & a_m & \cdots\cdots\cdots\cdots & a_0 \\ b_n & \cdots\cdots\cdots & b_0 & & 0 \\ & \ddots & & \ddots & \\ & & \ddots & & \ddots \\ 0 & & & b_n & \cdots\cdots\cdots & b_0 \end{pmatrix}$$

où les n premières lignes sont occupées par les coefficients de A et les m suivantes par ceux de B. La matrice $\mathrm{Syl}(A,B)$ est alors carrée et de taille $m+n$.

Cette matrice est la transposée de la matrice dans les bases des monômes de l'application linéaire de $\mathbb{A}_{n-1}[X] \times \mathbb{A}_{m-1}[X]$ dans $\mathbb{A}_{m+n-1}[X]$ définie par $(U,V) \mapsto AU + BV$.

1.2. Résultant.

Définition 2. *Le* résultant *de A et B est le déterminant de la matrice de Sylvester de A et B ; c'est un élément de $\mathbb{A}$ que l'on note* $\mathrm{Res}(A,B)$ *ou* $\mathrm{Res}_X(A,B)$ *lorsqu'il est utile de préciser la variable.*

Les coefficients de la matrice sont dans un anneau $\mathbb{A}$ qui n'est pas nécessairement un corps, mais les définitions et propriétés habituelles du déterminant persistent dans ce cadre (c'est une forme multilinéaire alternée, le déterminant de la matrice identité vaut 1, on peut le développer par rapport à une ligne ou une

colonne, il ne change pas par addition de combinaisons linéaires de lignes ou de colonnes, et s'exprime comme somme sur les permutations).

EXEMPLE 1. On a immédiatement $\operatorname{Res}(aX+b, cX+d) = ad - bc$.

EXEMPLE 2. Le discriminant d'un polynôme A est lié au résultant de A et A'. Par exemple, si $A = ax^2 + bx + c$, alors

$$\operatorname{Res}(A, A') = \begin{vmatrix} a & b & c \\ 2a & b & 0 \\ 0 & 2a & b \end{vmatrix} = -a(b^2 - 4ac).$$

Plus généralement, si $A = aX^m + \cdots$, alors on définit son discriminant par

$$\operatorname{Res}(A, A') = (-1)^{m(m-1)/2} a \operatorname{disc}(A).$$

PROPOSITION 1. *Si $\mathbb{A}$ est un corps, alors* $\operatorname{Res}(A, B) = 0$ *si et seulement si* $\deg \operatorname{pgcd}(A, B) > 0$.

DÉMONSTRATION. Si $G = \operatorname{pgcd}(A, B)$ a degré strictement positif, alors une division fournit deux polynômes non-nuls $\tilde{A}$ avec $\deg \tilde{A} < \deg A$ et $\tilde{B}$ avec $\deg \tilde{B} < \deg B$ tels que $A = G\tilde{A}$ et $B = G\tilde{B}$. Alors $(-\tilde{B}, \tilde{A})$ est un élément non-nul du noyau de la transposée de $\operatorname{Syl}(A, B)$, dont le déterminant doit être nul.

À l'inverse, si le résultant est nul, alors ce noyau est non vide et contient $(U, V) \neq (0, 0)$ tels que $UA + VB = 0$. Nécessairement, on a alors $V \neq 0$. Par l'absurde, si $\operatorname{pgcd}(A, B) = 1$ alors il existe une identité de Bézout

$$uA + vB = 1.$$

Multiplier par V et remplacer VB par $-UA$ mène à

$$(uV - vU)A = V.$$

Dans cette identité, le membre droit est non nul et de degré moindre que le membre gauche ($\deg V < \deg A$), ce qui est une contradiction. □

2. Propriétés principales

PROPOSITION 2. *Il existe deux polynômes U et V de $\mathbb{A}[X]$ tels que* $\operatorname{Res}(A, B) = UA + VB$.

DÉMONSTRATION. En notant $C_1, \ldots, C_{m+n}$ les colonnes de la matrice de Sylvester, ajouter à la dernière colonne la somme des $X^{m+n-i}C_i$ pour $i = 1, \ldots, m+n-1$ y fait apparaître des multiples de A et B. Le développement du déterminant par rapport à cette dernière colonne donne le résultat. □

La plupart des propriétés du résultant se déduisent de la formule de Poisson.

THÉORÈME 11 (Formule de Poisson). *Si*

$$A = a_m \prod_{i=1}^{m} (X - x_i) \qquad \text{et} \qquad B = b_n \prod_{j=1}^{n} (X - y_j),$$

alors

$$\operatorname{Res}(A, B) = a_m^n b_n^m \prod_{i,j} (x_i - y_j) = a_m^n \prod_{i=1}^{m} B(x_i) = (-1)^{mn} b_n^m \prod_{j=1}^{n} A(y_j).$$

Démonstration. Les deux dernières égalités sont des conséquences immédiates de la première.

Le coefficient $a_m^n b_n^m$ est obtenu par la multilinéarité du déterminant. On peut donc se contenter de regarder ensuite des polynômes unitaires. Le principe de la preuve est de prouver la formule pour A et B dans $\mathcal{A}[X]$, où

$$\mathcal{A} = \mathbb{Z}[x_1, \dots, x_m, y_1, \dots, y_n].$$

Cette identité entre polynômes se spécialise ensuite en y appliquant un morphisme d'anneaux de $\mathcal{A}$ vers $\mathbb{A}$, défini par les images des x_i et des y_j.

En raisonnant d'abord dans $\mathbb{Q}(x_1, \dots, x_m, y_1, \dots, y_n)[X]$, la proposition 1 montre que le résultant s'annule si on l'évalue en $x_i = y_j$. Il s'ensuit que $\prod(x_i - y_j)$ est un polynôme en les x_i et y_j qui divise $\operatorname{Res}(A, B)$. Ce polynôme a degré n par rapport à chacun des x_i et m par rapport à chacun des y_j. Par ailleurs, ces nombres sont des bornes supérieures sur les degrés de $\operatorname{Res}(A, B)$ par rapport à ces variables : le degré des n premières lignes de $\operatorname{Syl}(A, B)$ par rapport à un x_i vaut 1 et il est 0 pour les m lignes suivantes, et de même pour les y_j. Donc le résultant est égal à ce produit à une constante près. Pour trouver la constante, il suffit de considérer le cas particulier $A = X^m$. La matrice de Sylvester est alors triangulaire et son déterminant vaut $b_0^m = B(0)^m$, ce qui montre que la constante est 1. □

3. Calcul

Proposition 3. *Si $\mathbb{A}$ est un corps, et $A = BQ + R$ la division euclidienne de A par B (où $\deg R < \deg B$), alors*

$$\operatorname{Res}(A, B) = (-1)^{mn} b_n^{m - \deg R} \operatorname{Res}(B, R).$$

Démonstration. Il s'agit d'une application directe de la formule de Poisson,

$$\operatorname{Res}(A, B) = (-1)^{mn} b_n^m \prod_{B(\alpha)=0} A(\alpha) = (-1)^{mn} b_n^m \prod_{B(\alpha)=0} R(\alpha) = (-1)^{mn} b_n^{m - \deg R} \operatorname{Res}(B, R).$$

□

Grâce à cette proposition, on peut écrire un algorithme itératif semblable à l'algorithme d'Euclide, qui permet de calculer en $O(d^2)$ opérations le résultant de deux polynômes de degré inférieur ou égal à d, ce qui est plus rapide qu'un calcul de déterminant.

4. Applications

4.1. Calcul de projections. Algébriquement, la « résolution » d'un système polynomial se ramène souvent à une question d'*élimination*. Lorsqu'elle est possible, l'élimination successive des variables amène le système d'entrée sous une forme triangulaire. De proche en proche, la résolution d'un tel système se réduit alors à la manipulation de polynômes à une variable, pour lesquels compter, isoler, …, les solutions est bien plus facile.

Géométriquement, l'élimination correspond à une projection. Le schéma général est représenté en Figure 1 et correspond à la proposition suivante.

Proposition 4. *Soient $A = a_m Y^m + \cdots$ et $B = b_n Y^n + \cdots$ où les coefficients a_i et b_j sont dans $\mathbb{K}[X]$ où $\mathbb{K}$ est un corps algébriquement clos. Alors les racines du polynôme $\operatorname{Res}_Y(A, B) \in \mathbb{K}[X]$ sont d'une part les abscisses des solutions du système $A = B = 0$, d'autre part les racines des coefficients de tête a_m et b_n.*

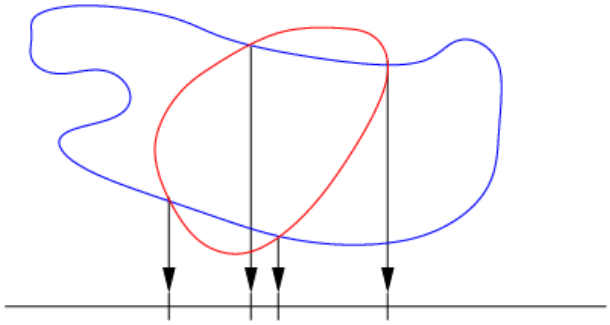


FIGURE 1. Le résultant calcule des projections.

DÉMONSTRATION. Les racines des coefficients de tête a_m et b_n sont racines du résultant d'après la formule de Poisson.

D'après la proposition 2, il existe U et V dans $\mathbb{K}[X,Y]$ tels que

$$\mathrm{Res}_Y(A,B) = UA + VB.$$

En évaluant cette identité en (x,y) tels que $A(x,y) = B(x,y) = 0$, on voit qu'un tel x est racine de $\mathrm{Res}_Y(A,B)$. À l'inverse, si x annule le résultant, alors d'après la proposition 1, les polynômes $A(x,Y)$ et $B(x,Y)$ ont un pgcd de degré au moins 1. Comme le corps est algébriquement clos, il existe alors $y \in \mathbb{K}$ racine de ce pgcd tel que $A(x,y) = B(x,y) = 0$. □

4.2. Implicitation de courbes unicursales. Supposons donnée une courbe plane paramétrée par des fractions rationnelles, $\{x = p_1(t)/q_1(t), y = p_2(t)/q_2(t)\}$. On peut calculer une équation implicite polynomiale dont les points de la courbe seront solutions grâce au résultant

$$\mathrm{Res}_t(p_1(t) - xq_1(t), p_2(t) - yq_2(t)).$$

Par exemple, le cercle unité est paramétré rationnellement par

$$x = \frac{1-t^2}{1+t^2}, \quad y = \frac{2t}{1+t^2}$$

et on retrouve bien l'équation du cercle :

$$\mathrm{Res}_t((1+t^2)X - (1-t^2), (1+t^2)Y - 2t) = 4(X^2 + Y^2 - 1).$$

4.3. Calcul avec des nombres algébriques. L'idée que les polynômes sont de bonnes structures de données pour représenter leur racines amène à chercher des algorithmes pour effectuer les opérations de base sur ces racines, comme la somme ou le produit. Le résultant répond à cette attente.

PROPOSITION 5. *Soient $A = \prod_i (X - \alpha_i)$, $B = \prod_j (X - \beta_j)$ et C des polynômes unitaires de $\mathbb{K}[X]$, avec C et A premiers entre eux. Alors*

$$\begin{aligned}
\mathrm{Res}_X(A(X), B(T - X)) &= \prod_{i,j} (T - (\alpha_i + \beta_j)), \\
\mathrm{Res}_X(A(X), B(T + X)) &= \prod_{i,j} (T - (\beta_j - \alpha_i)), \\
\mathrm{Res}_X(A(X), X^{\deg B} B(T/X)) &= \prod_{i,j} (T - \alpha_i \beta_j), \\
\mathrm{Res}_X(A(X), C(X)T - B(X)) &= (-1)^m C(0) \prod_i \left(T - \frac{B(\alpha_i)}{C(\alpha_i)} \right).
\end{aligned}$$

DÉMONSTRATION. C'est une application directe de la formule de Poisson. □

TP 5

Utilisation de résultants

1. Manipulation de nombres algébriques

1. Conjecturer puis prouver une expression algébrique de
$$\frac{\sin\frac{2\pi}{7}}{\sin^2\frac{3\pi}{7}} - \frac{\sin\frac{\pi}{7}}{\sin^2\frac{2\pi}{7}} + \frac{\sin\frac{3\pi}{7}}{\sin^2\frac{\pi}{7}}.$$
2. Prouver l'identité suivante due à Ramanujan :
$$\sqrt[3]{\cos a} + \sqrt[3]{\cos 2a} + \sqrt[3]{\cos 4a} = \sqrt[3]{\frac{5 - 3\sqrt[3]{7}}{2}}, \quad \text{où } a = 2\pi/7.$$
3. Calculer
$$\sum_{P(\alpha)=0} F(\alpha), \qquad \text{où} \quad F(\alpha) = \frac{\alpha^{10}}{\alpha^2+1} \quad \text{et} \quad P(\alpha) = \alpha^4 + p\alpha + q.$$
Deux approches seront employées : d'abord via le calcul du polynôme
$$\prod_{P(\alpha)=0} (T - F(\alpha));$$
ensuite l'utilisation de la série génératrice des sommes de Newton
$$\frac{XP'(X)}{P(X)} = \sum_{\substack{P(\alpha)=0 \\ i\geq 0}} \alpha^i X^{-i}.$$

2. L'épicycloïde de Huygens

L'enveloppe d'un ensemble de rayons lumineux s'appelle une *caustique*. Les caustiques sont souvent très nettement visibles. Ainsi, des rayons de lumière parallèles se réfléchissant sur les bords circulaires d'une tasse à café dessinent à la surface une épicycloïde de Huygens encore appelée *néphroïde* à cause de sa forme. Les points de l'enveloppe d'une famille de courbes d'équation $f(x, y, t) = 0$ satisfont
$$f(x, y, t) = \frac{\partial}{\partial t} f(x, y, t) = 0.$$

4. En prenant la tasse à café comme un cercle centré en 0 et de rayon 1, écrire les coordonnées du point $P = (\cos\theta, \sin\theta)$ en fonction de $t = \tan\theta/2$.
5. Les rayons lumineux sont pris horizontaux. Écrire l'équation de la droite reflétée en P.
6. En déduire une équation pour la caustique ; isoler le facteur pertinent.
7. Représenter sur une figure la tasse et la caustique.

3. Polynômes de Tchebychev entiers

Soit $\mathbb{Z}_k[X]$ l'ensemble des polynômes à coefficients entiers et de degré au plus k. Il existe (au moins) un polynôme $P_k \in \mathbb{Z}_k[X]$ tel que

$$\mu_k := \max_{x\in[0,1]} |P_k(x)|$$

soit minimal. De tels polynômes sont appelés polynômes de Tchebychev entiers dans l'intervalle $[0,1]$. Il est possible de calculer un certain nombre de facteurs de tels polynômes grâce à des observations simples sur les résultants. Cet exercice décrit une partie de ce calcul pour $k = 2m = 34$. Par symétrie, il n'est pas difficile de voir que $P_k(X) = Q_m(T)$ où $T = X(1-X)$, et $Q_m(T) \in \mathbb{Z}_m[T]$, ce qui réduit de moitié le degré des polynômes à chercher. L'intervalle d'étude change un peu puisque

$$\mu_k = \max_{x\in[0,1]} |P_k(x)| = \max_{t\in[0,\frac{1}{4}]} |Q_m(t)|.$$

Si Q_m se factorise en $Q_m = AB$ avec A et B dans $\mathbb{Z}[T]$ et que c_m est une borne connue sur $|Q_m|$, alors pour tout $t \in [0, 1/4]$,

$$|A(t)| \cdot |B(t)| \leq c_m.$$

Si B est un facteur de Q_m, cette inégalité entraîne alors une inégalité sur le facteur A qui peut permettre de prouver sa nullité en des points bien choisis.

8. On admet que $\mu_{34} < 0.33 \times 10^{-12}$. Montrer que T est un facteur de Q_{17}.
9. Montrer que $4T - 1$ et $5T - 1$ sont aussi des facteurs de Q_{17}.
10. Montrer enfin que $29T^2 - 11T + 1$ est un facteur de Q_{17}. (C'est ici qu'il faut utiliser les résultants de manière non triviale.)

Markov a donné une inégalité sur les dérivées d'un polynôme : si P est un polynôme de degré n à coefficients réels, alors

$$\max_{a\leq x\leq b} |Q^{(r)}(x)| \leq \frac{2^r}{(b-a)^r} \frac{n^2(n^2-1^2)\cdots(n^2-(r-1)^2)}{(2r-1)!!} \max_{a\leq x\leq b} |Q(x)|,$$

où $(2i+1)!! = 1 \cdot 3 \cdot 5 \cdots (2i+1)$.

11. Montrer que $T^4 | Q_{17}$.
12. Continuer à utiliser cette inégalité pour augmenter les multiplicités des facteurs $4T-1$, $5T-1$, et T de Q_{17} qui peuvent être trouvés par cette méthode.

COURS 6

Identités de fonctions spéciales et séries différentiellement finies

Résumé

Les équations différentielles linéaires et les récurrences linéaires fournissent des structures de données permettant de calculer avec des fonctions ou des suites, et en particulier de prouver des identités sur ces objets.

1. Définitions

1.1. Rappels et complément sur les séries formelles. Si $\mathbb{K}$ désigne un corps, l'anneau des séries formelles à coefficients dans $\mathbb{K}$ est noté $\mathbb{K}[[X]]$. Ses principales propriétés ont été présentées au Cours 1. Son corps des fractions, noté $\mathbb{K}((X))$ est égal à $\mathbb{K}[[X]][1/X]$. Ses éléments sont appelés des séries de Laurent formelles. C'est une algèbre non seulement sur $\mathbb{K}$, mais aussi sur le corps des fractions rationnelles $\mathbb{K}(X)$.

Dans tout ce cours on suppose $\mathbb{K}$ de caractéristique nulle. On peut donc penser sans rien perdre aux idées à $\mathbb{K} = \mathbb{Q}$.

1.2. Séries différentiellement finies.

Définition 1. *Une série formelle $A(X)$ à coefficients dans un corps $\mathbb{K}$ est dite* différentiellement finie *(ou D-finie) lorsque ses dérivées successives A, A', …, engendrent un espace vectoriel de dimension finie sur le corps $\mathbb{K}(X)$ des fractions rationnelles.*

De manière équivalente, cette série est solution d'une équation différentielle linéaire à coefficients dans $\mathbb{K}(X)$: si c'est le cas alors l'équation différentielle permet de récrire toute dérivée d'ordre supérieur à celui de l'équation en termes des dérivées d'ordre moindre (en nombre borné par l'ordre), à l'inverse, si l'espace est de dimension finie, alors pour m suffisamment grand, $A, A', \ldots, A^{(m)}$ sont liées et une relation de liaison entre ces dérivées est une équation différentielle linéaire.

En pratique, l'équation différentielle est utilisée pour les calculs, et la caractérisation de la définition pour les preuves d'existence.

1.3. Suites P-récursives.

Définition 2. *Une suite $(a_n)_{n\geq 0}$ d'éléments d'un corps $\mathbb{K}$ est appelée suite* polynomialement récursive *(ou P-récursive) si elle satisfait une récurrence de la forme*

$$p_d(n)a_{n+d} + p_{d-1}(n)a_{n+d-1} + \cdots + p_0(n)a_n = 0, \qquad n \geq 0, \tag{1}$$

où les p_i sont des polynômes de $\mathbb{K}[X]$.

2. Équivalence entre séries D-finies et suites P-récursives

THÉORÈME 12. *Une série formelle est D-finie si et seulement si la suite de ses coefficients est P-récursive.*

DÉMONSTRATION. Soit $A(X) = a_0 + a_1 X + \cdots$ une série D-finie et

$$q_0(X)A^{(m)}(X) + \cdots + q_m(X)A(X) = 0 \tag{2}$$

une équation différentielle qui l'annule. En notant $[X^n]f(X)$ le coefficient de X^n dans la série $f(X)$ avec la convention que ce coefficient est nul pour $n < 0$, on a pour $n \geq 0$

$$[X^n]f'(X) = (n+1)[X^{n+1}]f(X), \qquad [X^n]X^k f(X) = [X^{n-k}]f(X). \tag{3}$$

Par conséquent, l'extraction du coefficient de X^n de (2) fournit une récurrence linéaire sur les a_n valide pour tout $n \geq 0$ avec la convention $a_k = 0$ pour $k < 0$. Pour obtenir une récurrence de la forme (1) il faut décaler les indices de $n_0 := \max_{0\leq i\leq m}(\deg q_i + i - m)$ s'il est strictement positif. Les équations obtenues alors pour les indices moindres fournissent des contraintes linéaires sur les premiers coefficients a_n pour qu'ils correspondent aux coefficients d'une série solution de (2).

À l'inverse, soit (a_n) une suite vérifiant la récurrence (1). Les identités analogues à (3) sont maintenant

$$\sum_{n\geq 0} n^k a_n X^n = \left(X\frac{d}{dX}\right)^k A(X), \quad \sum_{n\geq 0} a_{n+k}X^n = (A(X) - a_0 - \cdots - a_{k-1}X^{k-1})/X^k,$$

où A est la série génératrice des coefficients a_n et la notation $(Xd/dX)^k$ signifie que l'opérateur Xd/dX est appliqué k fois. En multipliant (1) par X^n et en sommant pour n allant de 0 à ∞, puis en multipliant par une puissance de X on obtient donc une équation différentielle linéaire de la forme

$$q_0(X)A^{(d)}(X) + \cdots + q_d(X)A(X) = p(X),$$

où le membre droit provient des conditions initiales. Il est alors possible, quitte à augmenter l'ordre de l'équation de 1, de faire disparaître ce membre droit, par une dérivation et une combinaison linéaire. □

Ce calcul permet aussi d'observer le résultat suivant.

LEMME 1. *Si $A(X)$ est une série D-finie solution de (2) et $q_0(0) \neq 0$, alors le coefficient de tête de la récurrence (1) satisfaite par ses coefficients est le polynôme $q_0(0)(n+1)\cdots(n+m)$.*

DÉMONSTRATION. D'après (3), un terme $cX^i A^{(j)}(X)$ intervient dans la récurrence sur les coefficients sous la forme $c(n-i+1)\cdots(n-i+j)a_{n-i+j}$. L'indice maximal est donc atteint pour $j-i$ maximal et donc pour $j = m$ si $i = 0$. □

EXEMPLE 1. L'équation $y' - x^k y = 0$ ($k \in \mathbb{N}$) donne la récurrence $(n+1)a_{n+1} - a_{n-k} = 0$ valide pour tout $n \geq 0$ avec la convention que les a_n d'indice négatif sont nuls. On en déduit que a_0 est libre, puis les contraintes $a_1 = \cdots = a_k = 0$, et les coefficients suivants sont fournis par la récurrence décalée $(n+k+1)a_{n+k+1} - a_n = 0$, valide pour $n \geq 0$. On reconnaît ainsi les coefficients de $a_0 \exp(x^{k+1}/(k+1))$.

3. Test d'égalité

LEMME 2. *Si (u_n) et (v_n) sont deux suites solutions de (1), $u_0 = v_0, \ldots, u_{d-1} = v_{d-1}$ et $0 \notin p_d(\mathbb{N})$, alors ces suites sont égales.*

DÉMONSTRATION. Par récurrence, puisque $0 \notin p_d(\mathbb{N})$ permet d'inverser le coefficient de tête de la récurrence et donc d'exprimer chaque terme à partir de l'indice d en fonction des précédents. □

COROLLAIRE 1. *Si $f(X)$ et $g(X)$ sont deux séries formelles solutions de (2), $f(0) = g(0), \ldots, f^{(m-1)}(0) = g^{(m-1)}(0)$ et $q_0(0) \neq 0$, alors ces séries sont égales.*

DÉMONSTRATION. D'après le Lemme 1, le coefficient de tête de la récurrence sur les coefficients des solutions séries formelles de (2) ne s'annule pas sur $\mathbb{N}$. Les contraintes linéaires sur les conditions initiales jusqu'à l'indice n_0 introduit dans la preuve du Théorème 12 définissent les coefficients d'indice m à $m + n_0$ à partir des précédents et le Lemme 2 s'applique pour les valeurs suivantes. □

4. Somme et Produit

THÉORÈME 13. *L'ensemble des séries D-finies à coefficients dans un corps $\mathbb{K}$ est une algèbre sur $\mathbb{K}$. L'ensemble des suites P-récursives d'éléments de $\mathbb{K}$ est aussi une algèbre sur $\mathbb{K}$.*

DÉMONSTRATION. Les preuves pour les suites et les séries sont similaires. Les preuves pour les sommes sont plus faciles que pour les produits, mais dans le même esprit. Nous ne donnons donc que la preuve pour le produit $h = fg$ de deux séries D-finies f et g. Par la formule de Leibniz, toutes les dérivées de h s'écrivent comme combinaisons linéaires de produits entre une dérivée $f^{(i)}$ de f et une dérivée $g^{(j)}$ de g. Les dérivées de f et de g étant engendrées par un nombre fini d'entre elles, il en va de même pour les produits $f^{(i)}g^{(j)}$, ce qui prouve la D-finitude de h. □

En outre, cette preuve permet de borner l'ordre des équations : l'ordre de l'équation satisfaite par une somme est borné par la somme des ordres des équations satisfaites par les sommants, et l'ordre de l'équation satisfaite par un produit est borné par le produit des ordres.

Cette preuve donne également un algorithme pour trouver l'équation différentielle (resp. la récurrence) cherchée : il suffit de calculer les dérivées (resp. les décalées) successives en les récrivant sur un ensemble fini de générateurs. Une fois leur nombre suffisant (c'est-à-dire au pire égal à la dimension plus 1), il existe une relation linéaire entre elles. À partir de la matrice dont les lignes contiennent les coordonnées des dérivées successives (resp. des décalés successifs) sur cet ensemble fini de générateurs, la détermination de cette relation se réduit alors à celle du noyau de la transposée.

EXEMPLE 2. Voici comment prouver (et même découvrir) l'identité

$$\arcsin(x)^2 = \sum_{k\geq 0} \frac{k!}{\left(\frac{1}{2}\right)\cdots\left(k+\frac{1}{2}\right)} \frac{x^{2k+2}}{2k+2}.$$

Le calcul consiste à partir d'une équation satisfaite par $\arcsin(x)$, en déduire une équation satisfaite par son carré, traduire cette équation en récurrence sur les coefficients, et conclure en constatant que cette récurrence est satisfaite par les coefficients de la série.

Le point de départ est la propriété $(\arcsin(x))' = 1/\sqrt{1-x^2}$, qui permet de représenter arcsin par l'équation différentielle $(1-x^2)y''-xy' = 0$ avec les conditions initiales $y(0) = 0$, $y'(0) = 1$.

Ensuite, en posant $h = y^2$, les dérivations et réductions successives donnent

$$\begin{aligned}
h' &= 2yy', \\
h'' &= 2y'^2 + 2yy'' = 2y'^2 + \frac{2x}{1-x^2}yy', \\
h''' &= 4y'y'' + \frac{2x}{1-x^2}(y'^2 + yy'') + \left(\frac{2}{1-x^2} + \frac{4x^2}{(1-x^2)^2}\right)yy', \\
&= \left(\frac{4x}{1-x^2} + \frac{2x^2}{(1-x^2)^2} + \frac{2}{1-x^2} + \frac{4x^2}{(1-x^2)^2}\right)yy' + \frac{2x}{1-x^2}y'^2.
\end{aligned}$$

Les quatre vecteurs h, h', h'', h''' sont combinaisons linéaires des trois vecteurs y^2, yy', y'^2. Ils sont donc liés et une relation de liaison s'obtient en calculant le noyau de la matrice 3×4 qui découle de ce système. Le résultat est

$$(1-x^2)h''' - 3xh'' - h' = 0.$$

La récurrence qui s'en déduit est

$$(n+1)(n+2)(n+3)a_{n+3} - (n+1)^3 a_{n+1} = 0.$$

Comme le facteur $(n+1)$ ne s'annule pas sur $\mathbb{N}$, il est possible de simplifier pour obtenir la récurrence équivalente

$$(n+2)(n+3)a_{n+3} - (n+1)^2 a_{n+1} = 0.$$

La vérification que les coefficients de la série ci-dessus vérifient cette identité est facile.

5. Séries algébriques

THÉORÈME 14. *Si la série $Y(X)$ annule un polynôme $P(X,Y)$ de degré d en Y, alors elle est solution d'une équation différentielle linéaire d'ordre au plus d.*

DÉMONSTRATION. La preuve est algorithmique. Quitte à diviser d'abord P par son pgcd avec sa dérivée P_Y par rapport à Y, il est possible de le supposer premier avec P_Y (car la caractéristique est nulle !). En dérivant $P(X,Y) = 0$ et en isolant Y', il vient

$$Y' = -\frac{P_X}{P_Y}.$$

Par *inversion modulaire* de P_Y (voir le Cours 2), cette identité se récrit via un calcul de pgcd étendu en

$$Y' = R_1(Y) \bmod P,$$

où R_1 est un polynôme en Y de degré au plus d et à coefficients dans $\mathbb{K}(X)$. Ceci signifie que Y' s'écrit comme combinaison linéaire de $1, Y, Y^2, \ldots, Y^{d-1}$ à coefficients dans $\mathbb{K}(X)$. Dériver à nouveau cette équation, puis récrire Y' et prendre le reste de la division par P mène à nouveau à une telle combinaison linéaire pour Y'' et plus généralement pour les dérivées successives de Y. Les $d+1$ vecteurs $Y, Y', \ldots, Y^{(d)}$ sont donc linéairement dépendants et la relation de liaison est l'équation cherchée. □

Les mêmes arguments que ci-dessus mènent à une autre propriété de clôture des séries D-finies.

COROLLAIRE 2. *Si f est une série D-finie et y une série algébrique sans terme constant, alors $f \circ y$ est D-finie.*

La preuve consiste à observer que les dérivées successives de $f \circ y$ s'expriment comme combinaisons linéaires des $f^{(i)}(y)y^j$ pour un nombre fini de dérivées de f (par D-finitude) et de puissances de y (par la même preuve que pour le théorème 14). Cette preuve fournit encore un algorithme.

TP 6

La moyenne arithmético-géométrique et les séries hypergéométriques

1. La moyenne arithmético-géométrique

Si a et b sont deux réels tels que $0 \leq b \leq a$, les deux suites définies par

$$a_{n+1} = \frac{a_n + b_n}{2}, \quad b_{n+1} = \sqrt{a_n b_n}, \quad a_0 = a, \quad b_0 = b$$

ont une limite commune dont l'existence se déduit de $b_n \leq b_{n+1} \leq a_{n+1} \leq a_n$ et

$$a_{n+1}^2 - b_{n+1}^2 = \left(\frac{a_n - b_n}{2}\right)^2.$$

Cette limite est notée $M(a, b)$. Cette fonction est clairement homogène : pour $\lambda > 0$, $M(\lambda a, \lambda b) = \lambda M(a, b)$ ce qui permet de concentrer l'étude sur la fonction de une variable $M(1, x)$, dont on admet qu'elle est analytique au voisinage de $x = 1$.

1. Déduire de l'homogénéité et de $M(a_1, b_1) = M(a_0, b_0)$ avec $a_0 = 1+x$ et $b_0 = 1 - x$ une équation fonctionnelle satisfaite par $M(1, \cdot)$;
2. Utiliser cette équation fonctionnelle pour calculer les 10 premiers coefficients de Taylor de la fonction $A(x) = 1/M(1, \sqrt{1-x})$ (ou $M(1, x) = 1/A(1-x^2)$) à l'origine ;
3. Utiliser ces coefficients pour conjecturer une équation différentielle linéaire satisfaite par $A(x)$;
4. En utilisant la clôture des solutions d'équations différentielles linéaires par substitution algébrique, prouver que la série solution de cette équation différentielle avec les conditions initiales $y(0) = 1$, $y'(0) = 1/4$ satisfait l'équation fonctionnelle satisfaite par A ;
5. La série hypergéométrique est définie comme

$$F(a, b; c; z) := \sum_{n \geq 0} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}, \quad \text{où} \quad (x)_n = x(x+1)\cdots(x+n-1).$$

Calculer une récurrence satisfaite par les coefficients de Taylor de $A(x)$ et en déduire l'identité (due à Gauss)

$$M(a, b) = \frac{a}{F(\frac{1}{2}, \frac{1}{2}; 1; 1 - \frac{b^2}{a^2})}.$$

2. La relation de Legendre sur les intégrales elliptiques

Les *intégrales elliptiques complètes de première et de seconde espèce* sont

$$K(k) := \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}, \qquad E(k) := \int_0^1 \frac{\sqrt{1-k^2t^2}}{\sqrt{1-t^2}}\, dt.$$

6. Calculer une récurrence linéaire satisfaite par les intégrales

$$B_n = \int_0^1 \frac{t^{2n}\, dt}{\sqrt{1-t^2}};$$

7. En calculant une récurrence sur les coefficients du développement des intégrandes par rapport à k puis sur ceux de l'intégrale terme à terme, en déduire que pour $|k|<1$,

$$K(k) = \frac{\pi}{2} F(\frac{1}{2}, \frac{1}{2}; 1; k^2), \qquad E(k) = \frac{\pi}{2} F(-\frac{1}{2}, \frac{1}{2}; 1; k^2).$$

En conclure, après Gauss, que $K(k)$ peut se calculer par une moyenne arithmético-géométrique (π étant donné).

Les questions qui suivent permettent de prouver la *relation de Legendre*, qui s'écrit

$$E(k)K(k') + E(k')K(k) - K(k)K(k') = \frac{\pi}{2}, \quad \text{avec } k' = \sqrt{1-k^2}, \qquad 0 \le k \le 1.$$

8. En considérant les récurrences satisfaites par leurs coefficients montrer que $F(x) := \frac{2}{\pi} K(\sqrt{x})$ et $G(x) := \frac{2}{\pi} E(\sqrt{x})$ sont liés par

$$G(x) = (1-x)(2xF'(x) + F(x));$$

9. Injecter cette valeur de G dans le membre gauche de la relation de Legendre pour constater que celle-ci se récrit

$$x(1-x) \begin{vmatrix} F(x) & F(1-x) \\ F'(x) & F'(1-x) \end{vmatrix} = \text{cte};$$

10. Observer que $F(x)$ et $F(1-x)$ sont toutes deux solutions d'une même équation différentielle linéaire d'ordre 2 et en déduire l'existence de cette constante ;

11. Pour obtenir le membre droit de la relation de Legendre en faisant tendre k vers 0, on admettra que

$$K(k') = -\ln(k) + 2\ln 2 + O(k^2 \ln k), \quad k \to 0.$$

12. Déduire de la relation de Legendre une relation entre $K(1/\sqrt{2})$, $E(1/\sqrt{2})$ et $\pi/2$.

3. L'algorithme de Brent-Salamin pour le calcul de π

En introduisant la suite $T_n = 2^n a_n \left(E(b_n/a_n) - K(b_n/a_n)\right)$ et en établissant que $T_{n+1} - T_n = 2^n c_n^2 K(b/a)$, on obtient l'identité suivante que l'on admettra :

$$E(k) = \left(1 - \sum_{n=1}^{\infty} 2^k c_k^2\right) K(k), \tag{4}$$

où $c_k := \sqrt{a_k^2 - b_k^2} = c_{k-1}^2/(4a_k)$.

En mettant ensemble les équations des questions (7), (12) et celle ci-dessus, on obtient finalement

$$\frac{\pi^2}{4M^2(1,1/\sqrt{2})}\left(1-2\sum_{n=1}^{\infty}2^k c_k^2\right)=\frac{\pi}{2},$$

base de l'algorithme de Brent-Salamin pour le calcul de π. La moyenne arithmético-géométrique convergeant quadratiquement (grosso modo, le nombre de décimales correctes double à chaque étape), il s'agit du meilleur algorithme connu du point de vue de la complexité. L'itération est donnée par :

$$\pi=\lim_{n\to\infty}\pi_n,\qquad \pi_n:=\frac{2a_{n+1}^2}{1-\sum_{k=0}^{n}2^k c_k^2},$$

où les a_k et b_k sont les suites de la moyenne arithmético-géométrique pour $a_0=1$ et $b_0=1/\sqrt{2}$. Pour obtenir une bonne complexité, il faut disposer d'une multiplication rapide et calculer inverse et racine carrée par itération de Newton.

13. Calculer π_n pour $n=1,\dots,7$ avec 60 décimales de précision ;
14. Pour bénéficier d'arithmétique rapide en Maple, il vaut mieux effectuer les calculs sur des entiers que sur des flottants. Il suffit pour cela de démarrer le calcul avec $a_0=10^D$, $b_0=\sqrt{10^{2D}/2}$ où D est de l'ordre du nombre de décimales cherchées et d'utiliser des entiers tout au long du calcul (via `iquo` et `isqrt`). Écrire cette itération et la tester avec $D=10^4$, $D=2\cdot 10^4$, $D=4\cdot 10^4$, et observer l'évolution du temps de calcul avec D.

COURS 7

Sommation hypergéométrique

Résumé

Les suites hypergéométriques sont très courantes dans les applications. Leur algorithmique fait partie des succès du calcul formel. Les deux problèmes principaux sont la sommation indéfinie et la sommation définie. Les algorithmes correspondants sont dus à Gosper et Zeilberger.

Dans tout ce qui suit, $\mathbb{K}$ est un corps de caractéristique nulle.

1. Sommation indéfinie

1.1. Problème de la sommation indéfinie. La sommation indéfinie est l'analogue discret du calcul de primitives.

Définition 1. *Étant donnée une suite* $(f_k) \in \mathbb{K}^{\mathbb{N}}$*, on appelle* $(F_k) \in \mathbb{K}^{\mathbb{N}}$ *une* somme indéfinie *de* (f_k) *si*

$$\forall k \in \mathbb{N}, \quad F_{k+1} - F_k = f_k.$$

Le lien entre sommation indéfinie et somme est le même qu'entre primitive et intégrale : si $m \leq p \in \mathbb{N}$,

$$\sum_{k=m}^{p} f_k = \sum_{k=m}^{p} (F_{k+1} - F_k) = F_{p+1} - F_m.$$

1.2. Sommation indéfinie hypergéométrique.

Définition 2. *Une suite* $(u_k) \in \mathbb{K}^{\mathbb{N}}$ *est dite* hypergéométrique *sur* $\mathbb{K}$ *s'il existe une fraction rationnelle* $r(k) = p(k)/q(k) \in \mathbb{K}(k)$ *telle que*

$$\forall k \in \mathbb{N}, \quad q(k)u_{k+1} = p(k)u_k.$$

Exemple 1.
— Une suite géométrique est hypergéométrique : prendre $r(k) = \alpha$ où α est la raison de la suite. En ce sens, les suites hypergéométriques sont une généralisation des suites géométriques.
— La suite factorielle est hypergéométrique :

$$\forall k \in \mathbb{N}, \quad \frac{(k+1)!}{k!} = k + 1.$$

— $k \in \mathbb{N}$ étant fixé, les coefficients binomiaux $\binom{n}{k}$ forment une suite hypergéométrique :

$$\forall n \in \mathbb{N}, \quad \frac{\binom{n+1}{k}}{\binom{n}{k}} = \frac{\frac{(n+1)!}{k!(n+1-k)!}}{\frac{n!}{k!(n-k)!}} = \frac{n+1}{n+1-k}.$$

— $n \in \mathbb{N}$ étant fixé, les coefficients binomiaux $\binom{n}{k}$ forment une suite hypergéométrique :

$$\forall k \in \mathbb{N}, \quad \frac{\binom{n}{k+1}}{\binom{n}{k}} = \frac{\frac{n!}{(k+1)!(n-k-1)!}}{\frac{n!}{k!(n-k)!}} = \frac{n-k}{k+1}.$$

— Le cas général de suites hypergéométriques sur $\mathbb{C}$ s'écrit

$$u_k = CA^k \frac{\prod_{i=1}^{p} (a_i)(a_i+1)\cdots(a_i+k)}{\prod_{j=1}^{q} (b_i)(b_i+1)\cdots(b_i+k)},$$

avec $C, A, a_1, \ldots, a_p, b_1, \ldots, b_q$ dans $\mathbb{C}$. Il est clair que cette suite est hypergéométrique, et réciproquement, toute fraction rationnelle peut s'écrire $AP(k)/Q(k)$ où P et Q sont unitaires et on obtient la formule ci-dessus en nommant $a_1, \ldots, a_p$ les racines de P et $b_1, \ldots, b_q$ celles de Q.

Problème 1. On se restreint aux suites hypergéométriques.
— **Entrée** : $r(k) = p(k)/q(k) \in \mathbb{K}(k)$, et sous-entendue $(u_k) \in \mathbb{K}^{\mathbb{N}}$ telle que $\forall k \in \mathbb{N}$, $q(k)u_{k+1} = p(k)u_k$.
— **Sortie** : une suite hypergéométrique (v_k) telle que $\forall k \in \mathbb{N}$, $v_{k+1} - v_k = u_k$, ou FAIL si une telle suite n'existe pas. Lorqu'elle existe, on dit que (v_k) est une somme hypergéométrique de (v_k).

Exemple 2. Des exemples de telles paires (f_k, F_k) sont

$$\left(\frac{1}{4^k}\binom{2k}{k}, \frac{2k+1}{4^k}\binom{2k}{k}\right), \quad \left(\frac{4^k}{\binom{2k}{k}}, \frac{4^{k+1}}{3}\frac{2k+1}{\binom{2k+2}{k+1}}\right).$$

1.3. Algorithme de Gosper.

Lemme 1. *Soit $(u_k) \in \mathbb{K}^{\mathbb{N}}$ une suite hypergéométrique. Si (u_k) admet une somme hypergéométrique (v_k), alors il existe une fraction rationnelle $t(k) \in \mathbb{K}(k)$ telle que $\forall k \in \mathbb{N}$, $v_k = t(k)u_k$.*

Démonstration. Soit $s(k) \in \mathbb{K}(k)$ telle que $v_{k+1}/v_k = s(k)$. Alors :

$$\forall k \in \mathbb{N}, \quad v_{k+1} - v_k = (s(k) - 1)v_k = u_k$$

D'où le résultat en prenant $t(k) = 1/(s(k) - 1)$. □

Poursuivant le calcul, on obtient alors l'équation que doit vérifier $t(k)$:

$$\forall k \in \mathbb{N}, \quad t(k+1)r(k)u_k - t(k)u_k = u_k$$

d'où

$$t(k+1)r(k) - t(k) = 1$$

(équation d'inconnue $t(k)$ rationnelle).

Idée 1 (Gosper, 1978). *On se débarasse des différences entières entre racines et pôles en écrivant $r(k)$ sous la forme (appelée* forme de Gosper*)*

$$r(k) = \frac{a(k)}{b(k)}\frac{c(k+1)}{c(k)}, \qquad a, b, c \in \mathbb{K}[k] \tag{1}$$

avec $\forall m \in \mathbb{N}$, $\operatorname{pgcd}(a(k), b(k+m)) = 1$, puis on cherche $t(k)$ sous la forme $t(k) = b(k-1)x(k)/c(k)$.

L'équation sur $t(k)$ devient

$$\frac{b(k)}{c(k+1)}x(k+1)\frac{a(k)}{b(k)}\frac{c(k+1)}{c(k)} - \frac{b(k-1)}{c(k)}x(k) = 1$$

soit finalement

$$a(k)x(k+1) - b(k-1)x(k) = c(k). \tag{2}$$

THÉORÈME 15 (Gosper 1978). *Avec les notations précédentes, si $x(k)$ est une fraction rationnelle solution de (2), alors c'est un polynôme.*

DÉMONSTRATION. On écrit $x(k) = p(k)/q(k)$ avec $p, q \in \mathbb{K}[k]$ premiers entre eux, et on injecte dans l'équation :

$$a(k)p(k+1)q(k) - b(k-1)q(k+1)p(k) = q(k)q(k+1)c(k).$$

On a donc $q(k)|b(k-1)q(k+1)p(k)$ d'où $q(k)|b(k-1)q(k+1)$. De même $q(k+1)|a(k)q(k)$. En itérant, pour tout $K \in \mathbb{N}^\star$ on déduit

$$q(k)|b(k-1)b(k)\cdots b(k+K-2)q(k+K), \quad q(k)|a(k-1)a(k-2)\cdots a(k-K)q(k-K).$$

En choisissant K assez grand pour que $\operatorname{pgcd}(q(k), q(k+K)) = 1$, on en déduit

$$q(k)|\operatorname{pgcd}(b(k-1)b(k)\cdots b(k+K-2), a(k-1)a(k-2)\cdots a(k-K)) = 1,$$

où la dernière égalité est conséquence des hypothèses sur a et b. □

Le principe de l'algorithme de Gosper est donné en Algorithme 1. Il dépend de deux autres algorithmes, pour le calcul de la forme de Gosper et pour la recherche de solutions polynomiales.

Algorithme 1 Algorithme de Gosper

ENTRÉES: $r(k) \in \mathbb{K}(k)$ telle que $\forall k \in \mathbb{N}$, $u_{k+1}/u_k = r(k)$.
SORTIES: $f(k)$ telle que la suite hypergéométrique définie par $\forall k \in \mathbb{N}$, $v_k = f(k)u_k$ vérifie $\forall k \in \mathbb{N}$, $v_{k+1} - v_k = u_k$, ou FAIL si (u_k) n'admet pas de somme hypergéométrique.

1: Calculer une forme de Gosper de $r(k)$.
2: Trouver une solution polynomiale $x(k) \in \mathbb{K}[k]$ de (2).
3: **si** il n'y a pas de solution **alors**
4: **renvoyer** FAIL
5: **sinon**
6: **renvoyer** $\frac{b(k-1)}{c(k)}x(k)$
7: **fin si**

Le calcul de la forme de Gosper est donné en Algorithme 2. Dans un premier temps, il calcule l'ensemble des différences entières possibles entre les racines du numérateur et du dénominateur. Dans un second temps (la boucle), ces différences sont accumulées dans le polynôme c. La correction de l'algorithme provient d'un invariant simple à observer par récurrence : à chaque itération, (1) est satisfaite, et après l'étape i, $\operatorname{pgcd}(a(k), b(k+h_i)) = 1$. Ces deux propriétés restent donc vraies à la fin de l'algorithme.

La recherche de solutions polynomiales de l'équation (2) procède aussi en deux temps. D'abord, on va chercher une borne sur le degré de x, ensuite on peut résoudre

Algorithme 2 Forme de Gosper

Entrées: $r(k) = P(k)/Q(k)$, $\operatorname{pgcd}(P, Q) = 1$
Sorties: $a, b, c \in \mathbb{K}[k]$ obéissant à (1) et tels que $\forall m \in \mathbb{N}$, $\operatorname{pgcd}(a(k), b(k+m)) = 1$.

$R(h) := \operatorname{Res}_k(P(k), Q(k+h)) \in \mathbb{K}[h]$
Calculer $0 < h_1 < h_2 < \cdots < h_N$ les racines entières positives de R.
$a(k) := P(k); b(k) := Q(k); c(k) := 1$
pour i allant de 1 à N **faire**
  $g_i := \operatorname{pgcd}(a(k), b(k+h_i))$
  $a(k) := a(k)/g_i(k)$
  $b(k) := b(k)/g_i(k-h_i)$
  $c(k) := c(k)g_i(k-1)g_i(k-2)\cdots g_i(k-h_i)$
fin pour
renvoyer a, b, c.

par coefficients indéterminés. Pour obtenir une borne sur le degré, il est commode de récrire la récurrence sous la forme

$$a(k)(x(k+1) - x(k)) + (a(k) - b(k-1))x(k) = c(k),$$

ce qui permet d'exploiter le fait que si $x(k) \sim \gamma k^D$, alors $x(k+1) - x(k) \sim \gamma D k^{D-1}$. On note alors $a(k) \sim \lambda k^d$, $a(k) - b(k-1) \sim \mu k^\delta$ et on considère les termes asymptotiquement dominants :

$$\underbrace{a(k)}_{\sim\lambda k^d}\underbrace{(x(k+1) - x(k))}_{\gamma D k^{D-1}} + \underbrace{(a(k) - b(k-1))}_{\sim\mu k^\delta}\underbrace{x(k)}_{\sim\gamma k^D} = c(k).$$

La discussion sur le degré distingue donc deux cas :

1. Cas facile : $d - 1 \neq \delta$, alors $D \leq \deg c - \max(\delta, d-1)$;
2. Cas restant : il est possible d'avoir des solutions de degré élevé, la borne devenant $D \leq \max(\deg c - \delta, -\mu/\lambda)$, où le second terme n'apparaît que s'il est entier.

Une fois le degré de x borné, on peut résoudre l'équation par coefficients indéterminés, ce qui mène à un système *linéaire* sur les coefficients.

2. Sommation définie

Dans bien des cas, il n'existe pas de somme indéfinie hypergéométrique, et il faut avoir recours à d'autres méthodes pour la sommation définie.

2.1. Problème de la sommation définie.

Définition 3. *On dit qu'une suite* $(F_{n,k}) \in \mathbb{K}^{\mathbb{N}\times\mathbb{N}}$ *est* hypergéométrique *si* $F_{n+1,k}/F_{n,k}$ *et* $F_{n,k+1}/F_{n,k}$ *sont des fractions rationnelles dans* $\mathbb{K}(n,k)$.

Problème 2. Étant donnée une suite hypergéométrique $(F_{n,k}) \in \mathbb{K}^{\mathbb{N}\times\mathbb{N}}$, on cherche une récurrence linéaire à coefficients polynomiaux pour la suite (u_n) définie par

$$\forall n \in \mathbb{N}, \quad u_n = \sum_{k\in\mathbb{N}} F_{n,k}$$

Exemple 3. Des identités typiques de ce qui est calculable dans ce contexte sont :

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n, \qquad \sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}.$$

2.2. Algorithme de Zeilberger.

Idée 2. *(Zeilberger) On cherche des polynômes $t_0(n), \dots, t_p(n) \in \mathbb{K}[n]$ et une suite $G_{n,k}$ hypergéométrique tels que*

$$t_0(n)F_{n,k} + t_1(n)F_{n+1,k} + \cdots + t_p(n)F_{n+p,k} = G_{n,k+1} - G_{n,k}. \tag{3}$$

Si l'on dispose d'une telle récurrence, alors en sommant sur k les deux membres, on voit apparaître des décalées de la somme définie à gauche, et une somme qui se téléscope à droite. Moyennant de bonnes propriétés de G, ce téléscopage donne 0, ce qui permet de trouver une solution du problème de sommation. L'algorithme de Zeilberger part de l'observation que $G_{n,k}$ est une somme indéfinie en k d'une suite hypergéométrique. Posant des coefficients indéterminés pour les $t_i(n)$ (qui sont inconnus), on déroule alors l'algorithme de Gosper en cherchant les conditions sur les t_i pour qu'il existe une somme indéfinie hypergéométrique G. Plus en détail, avec $g_{n,k}$ le membre gauche de (3), on a

$$\frac{g_{n,k+1}}{g_{n,k}} = \frac{\sum_i t_i(n)F_{n+i,k+1}}{\sum_i t_i(n)F_{n+i,k}}$$

qui se récrit

$$\frac{g_{n,k+1}}{g_{n,k}} = \frac{\sum_{i=0}^{p} t_i(n)\frac{F_{n+i,k+1}}{F_{n,k+1}}}{\sum_{i=0}^{p} t_i(n)\frac{F_{n+i,k}}{F_{n,k}}} \frac{F_{n,k+1}}{F_{n,k}} \in \mathbb{K}(n,k).$$

Le membre gauche est donc bien hypergéométrique. L'idée est alors d'utiliser l'algorithme de Gosper dans le corps $\mathbb{K}(n)$, avec des t_i indéterminés. Étape par étape, le déroulement est le suivant :

1. Posons $P_n(k) = \sum_{i=0}^{p} t_i(n)F_{n+i,k}/F_{n,k}$, $Q_n(k) = F_{n,k+1}/F_{n,k}$ et $R_n(k) = g_{n,k+1}/g_{n,k}$. Alors P_n dépend linéairement des t_i, Q_n n'en dépend pas, et on a
$$R_n(k) = \frac{P_n(k+1)}{P_n(k)} Q_n(k)$$
Ainsi, en mettant Q_n sous forme de Gosper
$$Q_n(k) = \frac{A_n(k)}{B_n(k)} \frac{c_n(k+1)}{c_n(k)}$$
et en posant $C_n(k) = P_n(k)c_n(k)$, on obtient une forme de Gosper pour R_n
$$R_n(k) = \frac{A_n(k)}{B_n(k)} \frac{C_n(k+1)}{C_n(k)}$$
où A_n et B_n ne dépendent pas des t_i, et C_n en dépend linéairement.
2. L'équation qu'il faut alors résoudre est
$$A_n(k)X(k+1) - B_n(k-1)X(k) = C_n(k)$$
Dans cette équation, seul C_n dépend des t_i, et de façon linéaire. De plus, dans l'algorithme de Gosper, les bornes sur le degré du polynôme inconnu dépendent uniquement de A_n et de B_n, qui ne dépendent pas des t_i, et du

degré en k de C_n qui peut être borné indépendamment des t_i. En passant à des coefficients indéterminés pour X, on obtient alors un système linéaire, non seulement en les coefficients de X mais aussi en les t_i, qu'on peut résoudre.

3. Il suffit alors d'essayer des ordres p successivement dans l'espoir de trouver une récurrence.

Un théorème de Wilf & Zeilberger assure l'existence d'une telle récurrence (et donc la terminaison de l'algorithme) pour une sous-classe des suites hypergéométriques appelées "proprement" hypergéométriques, c'est-à-dire celles qui peuvent s'écrire sous la forme

$$P(n,k)A^k \frac{\prod_{i=1}^{p}(a_i n + b_i k + c_i)!}{\prod_{j=1}^{q}(u_j n + v_j k + w_j)!},$$

où P est un polynôme, les a_i, b_i, u_j, v_j sont des entiers, et p et q sont des entiers positifs ou nuls.

TP 7

Séries pour la fonction Zêta de Riemann aux entiers positifs

La fonction ζ de Riemann est définie pour $\Re s > 1$ par

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}. \tag{Z}$$

Cette série converge, mais sa vitesse de convergence n'est pas très élevée. Il existe des formules plus rapides, dont beaucoup ont été prouvées dans les vingt dernières années via l'algorithme de Zeilberger.

1. Une première série pour $\zeta(3)$

Andrei Andreevich Markov a prouvé en 1890 l'identité

$$\zeta(3) = \frac{5}{2} \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{\binom{2n}{n} n^3}. \tag{Z_3}$$

1. Évaluer numériquement les vingt premiers sommants, et la somme. En déduire la précision obtenue. Comparer au résultat de l'évaluation des mille premiers sommants de (Z) en $s = 3$.

Pour prouver (Z_3), l'algorithme de Zeilberger part de

$$F_{n,k} = \frac{(-1)^k k!^2 (n-k-1)!}{(n+k+1)!(k+1)}. \tag{A}$$

2. Calculer une suite $G_{n,k}$ telle que

$$F_{n+1,k} - F_{n,k} = G_{n,k+1} - G_{n,k}.$$

(On pourra utiliser l'implantation de l'algorithme de Zeilberger disponible en Maple dans le package `SumTools`).

3. (Cette question n'utilise pas de calcul formel, et pourra être admise dans un premier temps). En sommant cette identité d'abord de $k = 0$ à $n - 1$, puis de $n = 0$ à l'infini, montrer que si les sommes concernées convergent, alors

$$\sum_{n=0}^{\infty} G_{n,0} = \sum_{n \geq 0} (G_{n,n} + F_{n+1,n}). \tag{S_1}$$

4. En déduire l'équation (Z_3) en faisant attention aux formes indéterminées.
5. (À la fin, s'il reste du temps). Vérifier les convergences requises sur l'exemple.

2. Une famille de séries de plus en plus rapidement convergentes

6. Calculer le comportement asymptotique du n^{e} sommant de (Z_3).

Une généralisation de (Z_3) proposée par Amdeberhan en 1996 part de

$$F_{n,k} = \frac{(-1)^k k!^2 (sn-k-1)!}{(sn+k+1)!(k+1)}.$$

Le cas particulier $s=1$ redonne (A).

7. Appliquer la même méthode que ci-dessus pour $s=2$. En déduire une représentation sommatoire de $\zeta(3)$. Calculer le comportement asymptotique du n^{e} sommant.
8. Recommencer avec $s=3$.

3. D'autres valeurs de $\zeta(2n+3)$

De nombreuses identités plus ou moins récentes existent également sur les valeurs $\zeta(2n+3)$ [1]. Un article de 2008 de Kh. et T. Hessami Pilehrood montre comment obtenir certaines de ces identités par l'algorithme de Zeilberger.

9. Appliquer l'algorithme de Zeilberger à

$$F_{n,k} = \frac{(-1)^n (1+a)_n (1-a)_n}{\Gamma(1+a)\Gamma(1-a)} \frac{k!}{(2n+k+1)!((n+k+1)^2-a^2)},$$

où $(x)_n = x(x+1)\cdots(x+n-1)$ est le symbole de Pochhammer.

10. En sommant d'abord sur $n \geq 0$, puis sur $k \geq 0$, on obtient comme en question 3

$$\sum_{k=0}^{\infty} F_{0,k} = \sum_{n=0}^{\infty} G_{n,0}. \tag{S_2}$$

Vérifier que la formule obtenue est équivalente à

$$\sum_{k=1}^{\infty} \frac{1}{k(k^2-a^2)} = \sum_{n=0}^{\infty} \frac{(-1)^n (1+a)_n (1-a)_n (5(n+1)^2-a^2)}{(2n+2)!(2n+2)((n+1)^2-a^2)}.$$

11. Développer en série par rapport à a et obtenir par extraction des coefficients de a non seulement l'identité (Z_3), mais une identité pour $\zeta(5)$. (On pourrait bien sûr continuer et obtenir des formules pour tous les $\zeta(2n+3)$).
12. La généralisation suivante étend la convergence du cas $s=2$ de la question 7 à tous les $\zeta(2n+3)$. Mener le calcul en partant de

$$F_{n,k} = \frac{(-1)^n k!(n-1)!(2n)!(1+a)_k (1-a)_k (1+a)_n (1-a)_n (1+a)_{2n} (1-a)_{2n}}{(2n+k+1)!(3n-1)!(1+a)_{2n+k+1}(1-a)_{2n+k+1}}.$$

---

1. Les valeurs de ζ aux entiers pairs sont connues depuis Euler : elle s'écrivent

$$\zeta(2n) = (-1)^{n+1} \frac{B_{2n}}{2(2n)!} (2\pi)^{2n}, \qquad n = 1, 2, 3, \ldots$$

où les constantes $B_{2n}/(2n)!$ sont les coefficients de la série $z/(\exp(z)-1)$, et les B_n sont appelés *nombres de Bernoulli*.

COURS 8

Bases de Gröbner

Résumé

Les bases de Gröbner sont un outil très important du calcul formel. Elles permettent de nombreux calculs avec des idéaux d'anneaux de polynômes, ce qui en fait une structure de données utile pour manipuler les solutions de systèmes polynomiaux. Une des applications importantes, en relation avec la géométrie, est l'appartenance au radical d'un idéal. Quant au calcul de ces bases, il est permis par un algorithme simple de Buchberger, dont la correction et la terminaison nécessitent un peu de travail.

La division euclidienne, l'algorithme d'Euclide et l'algorithme d'Euclide étendu rendent effectifs de nombreux calculs dans $\mathbb{K}[x]$ ($\mathbb{K}$ est un corps). En particulier, ces algorithmes fournissent

— un test de divisibilité dans $\mathbb{K}[x]$;
— un test d'appartenance à l'idéal $(P) \subset \mathbb{K}[x]$, où $P \in \mathbb{K}[x]$;
— un calcul de forme normale dans $\mathbb{K}[x]/(P)$;
— un calcul d'élimination (les résultants).

Les bases de Gröbner permettent une généralisation de ces opérations à l'anneau $\mathbb{A} = \mathbb{K}[x_1, \ldots, x_n]$ des polynômes à n variables et à coefficients dans $\mathbb{K}$.

On utilisera la notation multi-exposant : si $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, on notera $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}$.

1. Définitions

1.1. Ordres monomiaux.

Définition 1. *On appelle*
— monôme *: un élément de $\mathbb{A}$ de la forme x^α où $\alpha \in \mathbb{N}^n$;*
— terme *: un élément de $\mathbb{A}$ de la forme λm où $\lambda \in \mathbb{K}$ et m est un monôme ;*
— ordre monomial *: un ordre total sur les monômes qui est compatible avec le produit (i.e. $m_1 \prec m_2 \Rightarrow mm_1 \prec mm_2$) et tel que toute suite décroissante de monômes est stationnaire.*

Une conséquence simple de cette définition est que si $\prec$ est un ordre monomial, alors 1 est le plus petit élément. En effet, si $m \prec 1$ pour un certain monôme m, alors en multipliant par m, $m^2 \prec m \prec 1$. On construit de cette manière une suite infinie strictement décroissante.

D'autre part, si $n = 1$, il n'y a qu'un ordre monomial possible, à savoir l'ordre donné par le degré.

EXEMPLE 1. L'ordre lexicographique est un ordre monomial. Il s'agit de l'ordre défini par $x^\alpha \prec x^\beta$ si et seulement si le premier coefficient non nul de $\alpha - \beta$ est négatif. En Maple, cet ordre est noté $\mathrm{plex}(x_1, \ldots, x_n)$. Par exemple, pour $\mathrm{plex}(x, y, z)$:

$$1 \prec z \prec z^2 \prec \cdots \prec y \prec yz \prec \cdots \prec y^2 \prec \cdots \prec x \prec \cdots .$$

EXEMPLE 2. L'ordre du degré lexicographique inverse est également un ordre monomial. Il s'agit de l'ordre définit par $x^\alpha \prec x^\beta$ si et seulement si $\sum \alpha_i < \sum \beta_i$ ou $\sum \alpha_i = \sum \beta_i$ et le *dernier* élément non nul de $\alpha - \beta$ est positif. En Maple, cet ordre est noté $\mathrm{tdeg}(x_1, \ldots, x_n)$. Par exemple, pour $\mathrm{tdeg}(x, y, z)$:

$$1 \prec z \prec y \prec x \prec z^2 \prec \cdots \prec x^2 \prec z^3 \prec \cdots \prec y^3 \prec \cdots \prec x^2 y \prec x^3.$$

DÉFINITION 2. *Un ordre monomial sur $\mathbb{A}$ étant fixé, soit $f \in \mathbb{A}$ un polynôme. On appelle*

- monôme de tête *de f : le plus grand monôme de f. On le note* $\mathrm{LM}(f)$ *(pour* leading monomial*).*
- terme de tête *de f : le terme correspondant au monôme de tête. On le note* $\mathrm{LT}(f)$ *(pour* leading term*).*
- coefficient de tête *de f : le coefficient correspondant au monôme de tête. On le note* $\mathrm{LC}(f)$ *et on a* $\mathrm{LT}(f) = \mathrm{LC}(f)\,\mathrm{LM}(f)$.

La compatibilité de l'ordre monomial avec le produit entraîne la relation $\mathrm{LT}(fg) = \mathrm{LT}(f)\,\mathrm{LT}(g)$ pour tous $f, g \in \mathbb{A}$.

1.2. Bases de Gröbner.

DÉFINITION 3. *Un ordre monomial sur $\mathbb{A}$ étant fixé, un sous-ensemble fini $G = \{g_1, \ldots, g_k\}$ d'un idéal $I \subset \mathbb{A}$ est une* base de Gröbner *de I si $\langle \mathrm{LT}(G) \rangle = \langle \mathrm{LT}(I) \rangle$. (Ici, $\langle A \rangle$ désigne l'idéal engendré par la partie A).*

Il n'y a pas unicité des bases de Gröbner. Par exemple, si G est une base de Gröbner d'un idéal I et si $g \in I$ alors $G \cup \{g\}$ est encore une base de Gröbner de I. Bien que ce ne soit pas évident d'après la définition, on verra plus loin que si G est une base de Gröbner d'un idéal I, alors G engendre I. On verra aussi que de telles bases existent toujours.

EXEMPLE 3. Si $n = 1$, alors $\mathbb{A} = \mathbb{K}[x]$ est un anneau principal, ce qui veut dire que ses idéaux peuvent être engendrés avec un élément. Si $I \subset \mathbb{K}[x]$ est un idéal, il existe donc $g \in I$ tel que $I = \langle g \rangle$ (g est le pgcd des éléments de I). Les termes de tête des éléments de I sont les termes de la forme λx^k avec k supérieur ou égal au degré de g. L'idéal qu'ils engendrent est donc $x^{\deg g}\mathbb{A} = \langle x^{\deg g} \rangle = \langle \mathrm{LM}(g) \rangle$ et donc $\{g\}$ est une base de Gröbner de I.

EXEMPLE 4. Si $A = (a_{ij})$ est une matrice en forme échelon dans $\mathbb{K}^{m \times n}$, alors l'idéal

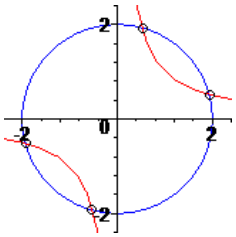
$$\left\langle \sum_{j=1}^{n} a_{ij} x_j, 1 \leq i \leq m \right\rangle \subset \mathbb{A}$$

admet l'ensemble $\{\sum_{j=1}^{n} a_{ij} x_j\}$ comme base de Gröbner pour l'ordre lexicographique. Ainsi, les bases de Gröbner généralisent à la fois le pgcd et la réduction de Gauss.

Exemple 5. Considérons le système

$$f_1 = x^2 + y^2 - 4, \quad f_2 = xy - 1$$

correspondant aux points d'intersections d'un cercle et d'une hyperbole :



On admet qu'une base de Gröbner pour plex(x, y) est donnée par le système

$$\underline{y^4} - 4y^2 + 1, \quad \underline{x} + y^3 - 4y,$$

où les monômes de tête sont soulignés. Le premier polynôme admet pour racines les ordonnées des points d'intersections. Le second permet de calculer les valeurs des abscisses correspondantes. Plus généralement, l'ordre lexicographique permet de trianguliariser les systèmes polynomiaux.

Pour l'ordre tdeg(x, y), le système suivant est une base de Gröbner :

$$\underline{x^2} + y^2 - 4, \quad \underline{xy} - 1, \quad \underline{y^3} + x - 4y,$$

et on peut vérifier qu'aucun système comportant moins de polynômes n'est une base de Gröbner de cet idéal. En particulier, on observe que le nombre de polynômes dépend de l'ordre monomial.

2. Division et forme réduite

2.1. Division.

Définition 4. *On dit qu'un polynôme $f \in \mathbb{A}$ est* réduit *par rapport à une partie $G \subset \mathbb{A}$ pour un ordre monomial donné si aucun des monômes de f n'est divisible par le monôme de tête d'un élément de G.*

On dit qu'une base de Gröbner $G = \{g_1, \ldots, g_k\}$ est réduite *si pour tout $i \in \{1, \ldots, k\}$, g_i est réduit par rapport à $G \setminus \{g_i\}$.*

Théorème 16 (Division). *Soit $G = \{g_1, \ldots, g_k\}$ un ensemble de polynômes de $\mathbb{A}$ et $F \in \mathbb{A}$. Il existe $Q = (q_1, \ldots, q_k)$ dans $\mathbb{A}^k$ et R dans $\mathbb{A}$ tels que*

$$F = q_1 g_1 + \cdots + q_k g_k + R$$

et R est réduit par rapport à G. On note $R = \overline{F}^G$ et on dit que R est le reste de la division de F par G. De plus, si G est une base de Gröbner, alors R est unique.

Démonstration. *Existence.* L'existence est donnée par l'Algorithme 2.

À chaque étape, la relation

$$F = f + r + a_1 g_1 + \cdots + a_k g_k \tag{1}$$

est maintenue ; par construction seuls des monômes réduits sont ajoutés à r ; enfin, la terminaison est assurée par la décroissance du terme de tête de f à chaque passage dans la boucle.

Algorithme 2 Algorithme de Division

Entrées: $F, G = \{g_1, \dots, g_k\}$ et l'ordre monomial correspondant
Sorties: R et $a_1, \dots, a_k$ tels que $F = a_1 g_1 + \cdots + a_k g_k + R$
1: Initialisation : $R = a_1 = \cdots = a_k = 0$; $f = F$;
2: **tant que** $f \neq 0$ **faire**
3: $S := \{i \mid \mathrm{LT}(g_i) \text{ divise } \mathrm{LT}(f)\}$;
4: **si** $S = \emptyset$ **alors**
5: $r := r + \mathrm{LT}(f)$; $f := f - \mathrm{LT}(f)$;
6: **sinon**
7: $i := \min S$; $a_i := a_i + \mathrm{LT}(f)/\mathrm{LT}(g_i)$; $f := f - g_i \mathrm{LT}(f)/\mathrm{LT}(g_i)$;
8: **fin si**
9: **fin tant que**
10: **renvoyer** $(r, a_1, \dots, a_k)$.

Unicité. Si G est une base de Gröbner d'un idéal I et F s'écrit de deux façons $F = B_1 + R_1 = B_2 + R_2$ avec R_1 et R_2 réduits par rapport à G et B_1 et B_2 dans I, alors $R_1 - R_2 = B_2 - B_1 \in I$, donc $\mathrm{LT}(R_1 - R_2) \in \langle \mathrm{LT}(I) \rangle = \langle \mathrm{LT}(G) \rangle$. Mais $\mathrm{LT}(R_1 - R_2)$ est réduit, donc il n'est divisible par le monôme de tête d'un élément de G que s'il est nul. □

Corollaire 1. *Sous les mêmes hypothèses, $F \in I$ si et seulement si $\overline{F}^G = 0$ ce qui fournit un test d'appartenance à un idéal dès que l'on possède une base de Gröbner.*

Corollaire 2. *Une base de Gröbner d'un idéal I engendre I.*

En effet il est clair que $\langle G \rangle \subset I$. Inversement, si $F \in I$, alors $F = F + 0$ est l'unique décomposition. D'après le théorème précédent, on a alors $F \in \langle G \rangle$.

2.2. Bases réduites. L'algorithme de division permet aussi de réduire les bases de Gröbner. Si G n'est pas réduite, alors on réduit chaque g_i par $G \setminus \{g_i\}$, on le supprime si le reste est nul et on le remplace par son reste sinon. Le résultat engendre le même idéal, et ne modifie pas l'ensemble des termes de tête de G.

Deux bases de Gröbner réduites pour le même ordre monomial sont identiques à des facteurs constants près. En effet, soient G, G' deux bases de Gröbner réduites. Si $g_1 \in G$ alors $\mathrm{LT}(g_1)$ est divisible par le terme de tête d'un élément de G', disons g'_1. À son tour, le terme de tête de g'_1 est divisible par le terme de tête d'un élément $g_2 \subset G$. Mais alors $\mathrm{LT}(g_2)$ divise $\mathrm{LT}(g_1)$ et comme G_1 est réduite $g_1 = g_2$. Ainsi $\mathrm{LT}(g_1) = c\,\mathrm{LT}(g'_1)$ pour un $c \in \mathbb{K}$. Posons alors $f_1 = g_1 - cg'_1 \in I$. Si $f_1 \neq 0$, son monôme de tête apparaît alors dans g_1 ou g'_1, disons g_1. Il n'est pas divisible par $\mathrm{LT}(g_1)$ puisque $\mathrm{LT}(f_1) \prec \mathrm{LT}(g_1)$, ni divisible par $\mathrm{LT}(g)$ pour $g \in G \setminus \{g_1\}$ puisque G est réduite. Ceci contredit $f_1 \in I$. Ainsi $g_1 = cg'_1$. En raisonnant de même pour tous les éléments de G, puis par symétrie, on obtient la conclusion. On obtient même l'unicité en forçant les termes de tête à être unitaires.

Cette observation fournit un test d'égalité entre idéaux : si I_1 et I_2 sont deux idéaux possédant chacune une base de Gröbner alors $I_1 = I_2$ si et seulement si les bases de Gröbner réduites à coefficient de tête unitaires sont égales.

Remarque 1. Nous n'avons toujours pas démontré l'existence de bases de Gröbner.

3. Elimination

THÉORÈME 17 (Elimination). *Soit G une base de Gröbner de $I \subset \mathbb{K}[x_1, \dots, x_n]$ pour l'ordre lexicographique. Alors $G \cap \mathbb{K}[x_q, \dots, x_n]$ est une base de Gröbner de $I \cap \mathbb{K}[x_q, \dots, x_n]$ pour l'ordre lexicographique.*

REMARQUE 2. Le théorème est valable pour d'autres ordres monomiaux appelés ordres d'élimination, qui séparent les groupes de variables $(x_1, \dots, x_{q-1})$ et $(x_q, \dots, x_n)$ lexicographiquement, mais traitent comme l'ordre du degré les variables à l'intérieur de chaque groupe. En Maple, ces ordres sont notés lexdeg.

DÉMONSTRATION. Notons $\mathbb{A}_q = \mathbb{K}[x_q, \dots, x_n]$, $G_q = G \cap \mathbb{A}_q$ et $I_q = I \cap \mathbb{A}_q$. Comme $G_q \subset I \cap \mathbb{A}_q$, $\langle G_q \rangle \subset I_q$ ($\langle G_q \rangle$ désigne ici l'idéal de $\mathbb{A}_q$ engendré par G_q).

Réciproquement, si $F \in I_q \subset I$ alors $\mathrm{LT}(F) \in \mathbb{A}_q$. En appliquant l'algorithme de division à F et G, si $g \in G$ est tel que $\mathrm{LT}(g) | \mathrm{LT}(F)$ alors $\mathrm{LT}(g) \in \mathbb{A}_q$ et par définition de l'ordre lexicographique, g lui-même appartient alors à $\mathbb{A}_q$. Donc $g \in G_q$, et l'opération de soustraction maintient f dans $\mathbb{A}_q$. Ainsi à chaque étape de l'algorithme, tous les polynômes de l'écriture (1) sont dans $\mathbb{A}_q$ et on obtient donc $F \in \langle G_q \rangle$. □

L'élimination a de nombreuses applications. En voici quelques unes.

Résultant. Soient $f, g \in \mathbb{A} := \mathbb{K}[x_1, \dots, x_n, Y]$ deux polynômes. On munit $\mathbb{A}$ de l'ordre lexicographique. Soit I l'idéal engendré par f et g et G la base de Gröbner réduite de I. Alors $G \cap \mathbb{K}[x_1, \dots, x_n]$ ne contient qu'un élément : le résultant de f et g par rapport à Y.

Implicitation. Soit un système polynomial

$$\begin{cases} x_1 = f_1(U_1, \dots, U_k) \\ \vdots \\ x_n = f_n(U_1, \dots, U_k) \end{cases}$$

et soit $I \subset \mathbb{K}[U_1, \dots, U_k, x_1, \dots, x_n]$ l'idéal engendré par ce système. L'élimination des U_i dans la base de Gröbner de I donne les équations implicites de l'ensemble algébrique défini par le système. Si les f_i sont des fractions rationnelles, $f_i = p_i/q_i$, alors on travaille avec l'idéal

$$\langle q_1 x_i - p_1, \dots, q_n x_n - p_n, 1 - t q_1 \cdots q_n \rangle \subset \mathbb{K}[t, U_1, \dots, U_k, x_1, \dots, x_n].$$

Relations de dépendance. Soient $f_1, \dots, f_m \in \mathbb{K}[x_1, \dots, x_n]$ des polynômes et $g \in \langle f_1, \dots, f_m \rangle$. L'élimination de t dans l'idéal $I = \langle y_1 - t f_1, \dots, y_m - t f_m, y - t g \rangle \subset \mathbb{K}[t, x_1, \dots, x_n, y_1, \dots, y_m, y]$ permet de calculer une relation de dépendance, c'est-à-dire des $h_i \in \mathbb{K}[x_1, \dots, x_n]$ tels que $g = \sum h_i f_i$.

4. Radicaux et Nullstellensatz

DÉFINITION 5. *Soit $\mathcal{I}$ un idéal de $\mathbb{A}$. Son* radical *est l'idéal*

$$\sqrt{\mathcal{I}} := \{ f \in \mathbb{A} | \exists p \in \mathbb{N}, f^p \in \mathbb{A} \}.$$

Cette définition vise à éliminer les multiplicités dans les polynômes sur lesquels on travaille. Par exemple, si $\mathbb{A} = \mathbb{K}[X]$ et $\mathcal{I} = \langle X^2 \rangle$, alors $\sqrt{\mathcal{I}} = \langle X \rangle$.

On arrive au théorème fondamental de cette partie.

THÉORÈME 18 (Nullstellensatz de Hilbert). *On suppose que $\mathbb{K}$ est algébriquement clos. Soient $f, f_1, \ldots, f_r$ des éléments de $\mathbb{K}[x_1, \ldots, x_n]$. Alors $f \in \sqrt{\langle f_1, \ldots, f_r \rangle}$ si et seulement si f s'annule sur le lieu des zéros communs des f_i dans $\mathbb{K}^n$.*

Ainsi, la géométrie de l'ensemble algébrique défini par les f_i est codée dans le radical de l'idéal qu'ils engendrent. De plus, on peut tester l'appartenance au radical par un simple calcul de base de Gröbner, grâce au résultat suivant.

PROPOSITION 1 (Astuce de Rabinowitsch, 1929). *Soient $f, f_1, \ldots, f_r$ des éléments de $\mathbb{K}[x_1, \ldots, x_n]$ et t une nouvelle indéterminée. Alors*

$$f \in \sqrt{\langle f_1, \ldots, f_r \rangle} \iff \langle f_1, \ldots, f_r, 1 - tf \rangle = \langle 1 \rangle = \mathbb{K}[x_1, \ldots, x_n, t].$$

Ce résultat donne un algorithme pour le test d'appartenance au radical par le calcul d'une base de Gröbner de l'idéal à droite.

DÉMONSTRATION. Soit $\mathcal{I} = \langle f_1, \ldots, f_r \rangle \subset \mathbb{A}$ et $\tilde{\mathcal{I}} = \langle f_1, \ldots, f_r, 1 - tf \rangle \subset \mathbb{A}[t]$. Supposons que f^p soit dans $\mathcal{I}$. Alors f^p et $t^p f^p$ appartiennent aussi à $\tilde{\mathcal{I}}$. Mais $1 - t^p f^p$ est un multiple de $1 - tf$, donc la différence 1 appartient aussi à $\tilde{\mathcal{I}}$.

Réciproquement, si 1 appartient à $\tilde{\mathcal{I}}$, alors en posant $X = x_1, \ldots, x_n$, il s'écrit

$$1 = g_1(X, t) f_1(X) + \cdots + g_r(X, t) f_r(X) + g(X, t)(1 - tf(X)).$$

En injectant $t = \frac{1}{f(X)}$, et en réduisant au même dénominateur, on obtient explicitement la décomposition de f^m en termes des f_i, où m est le maximum des degrés des g_i en t, ce qui montre $f \in \sqrt{\mathcal{I}}$. □

PREUVE DU NULLSTELLENSATZ. Si $f^p = \sum g_i f_i$, et si $\mathbf{x} \in \mathbb{K}^n$ est un zéro commun aux f_i, alors $f^p(\mathbf{x})$ donc $f(\mathbf{x}) = 0$.

L'autre sens est plus difficile. Soit f s'annulant en tous les zéros communs des f_i. D'après l'astuce de Rabinowitsch, il suffit de montrer que 1 est dans l'idéal engendré par $f_1, \ldots, f_r, 1 - tf$. Tout d'abord, on observe que ces polynômes n'ont pas de zéros communs. En effet, s'il en existait un, disons $\mathbf{a} = (a_1, \ldots, a_n)$, alors par hypothèse $f(\mathbf{a}) = 0$, donc $(1 - tf)(\mathbf{a}) = 1$, ce qui est contradictoire. La conclusion découle alors de la forme faible du Nullstellensatz ci-dessous. □

PROPOSITION 2 (Nullstellensatz faible). *Si $\mathbb{K}$ est algébriquement clos, et $\mathcal{I}$ est un idéal strict de $\mathbb{K}[x_1, \ldots, x_n]$, alors $\exists \mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{K}^n$ tel que*

$$\forall f \in \mathcal{I}, \qquad f(\mathbf{a}) = 0.$$

La preuve procède par récurrence sur le nombre de variables, et pour $n \geq 2$ utilise un argument de projection. Le lemme suivant permet d'assurer que cette projection se passe bien. (Notons que si $\mathbb{K}$ est algébriquement clos, il est infini.)

LEMME 1 (Lemme de normalisation de Noether). *Supposons $n \geq 2$, et $\mathbb{K}$ infini. Soit $f \in \mathbb{K}[x_1, \ldots, x_n]$ de degré $d > 0$. Alors il existe $(\lambda_1, \ldots, \lambda_{n-1}) \in \mathbb{K}^{n-1}$ tels que le coefficient de x_n^d dans*

$$f(x_1 + \lambda_1 x_n, \ldots, x_{n-1} + \lambda_{n-1} x_n, x_n)$$

soit non nul.

DÉMONSTRATION. Soit $a_{i_1,\dots,i_n}$ le coefficient de $x_1^{i_1}\cdots x_n^{i_n}$ dans f. Puisque f est de degré d, le polynôme homogène

$$g(x_1,\dots,x_n) := \sum_{i_1+\cdots+i_n=d} a_{i_1,\dots,i_n} x_1^{i_1}\cdots x_n^{i_n}$$

est non nul. Le coefficient de x_n^d dans $f(x_1+\lambda_1 x_n,\dots,x_{n-1}+\lambda_{n-1}x_n,x_n)$ vaut précisément $g(\lambda_1,\dots,\lambda_{n-1},1)$. Comme $\mathbb{K}$ est infini, on est assuré de l'existence de scalaires $\lambda_1,\dots,\lambda_{n-1}$ qui lui donnent une valeur non nulle (encore par récurrence sur le nombre de variables : pour une variable, le nombre de racines est borné par le degré, ensuite on regarde le coefficient de tête, on prend un point où il ne s'annule pas et on conclut sur le polynôme en une variable ainsi obtenu). □

DÉMONSTRATION (DU NULLSTELLENSATZ FAIBLE). On procède par récurrence sur n. Le cas $n=1$ est assuré par le fait que $\mathbb{K}$ est algébriquement clos.

Si $n \geq 2$, on va projeter pour se ramener à une variable de moins. Soit $g \in \mathcal{I}$. D'après le lemme de normalisation, on peut supposer quitte à renormaliser g qu'il existe $\lambda_1,\dots,\lambda_{n-1}$ tels que $g(x_1+\lambda_1 x_n,\dots,x_{n-1}+\lambda_{n-1}x_n,x_n) = x_n^d + r(x_1,\dots,x_n)$, r étant de degré au plus $n-1$ en x_n.

Alors en posant $\mathcal{J} := \{f(x_1+\lambda_1 x_n,\dots,x_{n-1}+\lambda_{n-1}x_n,x_n), f\in\mathcal{I}\}$, on obtient un autre idéal strict de $\mathbb{K}[x_1,\dots,x_n]$. Donc, quitte à remplacer $\mathcal{I}$ par $\mathcal{J}$, on suppose que $g(x_1,\dots,x_n) = x_n^d + g_{d-1}x_n^{d-1} + \cdots + g_0$, $g_i \in \mathbb{K}[x_1,\dots x_{n-1}]$.

On pose alors $\mathcal{I}' = \mathcal{I}\cap\mathbb{K}[x_1,\dots,x_{n-1}]$. C'est un idéal strict de $\mathbb{K}[x_1,\dots,x_{n-1}]$. Par hypothèse de récurrence, il existe donc $\mathbf{a} = a_1,\dots,a_{n-1}$ qui annule tous les polynômes de $\mathcal{I}'$.

On pose alors $\mathcal{J} = \{f(\mathbf{a},x_n), f\in\mathcal{I}\}$. C'est un idéal de $\mathbb{K}[x_n]$. S'il est strict alors il est engendré par un polynôme non constant en x_n, dont n'importe quelle racine a_n donne la réponse $(\mathbf{a},a_n)$ à la question. Si on suppose le contraire, alors $\exists f\in\mathcal{I}$ tel que $f(\mathbf{a},x_n)=1$. Ce polynôme peut s'écrire

$$f(x_1,\dots,x_n) = f_0 + f_1 x_n + \cdots + f_k x_n^k,$$

où les f_i sont dans $\mathbb{K}[x_1,\dots,x_{n-1}]$. L'hypothèse implique que $f_0(\mathbf{a})=1$ et $f_1(\mathbf{a}) = \cdots = f_k(\mathbf{a}) = 0$.

Le résultant R de f et g par rapport à la variable x_n est à la fois un élément de $\mathbb{K}[x_1,\dots,x_{n-1}]$ et de $\langle f,g\rangle \subset \mathcal{I}$. Donc $R\in\mathcal{I}'$, ce qui implique $R(\mathbf{a})=0$. Pourtant, en évaluant la matrice de Sylvester de (f,g) en $\mathbf{a}$, on obtient des 1 sur la diagonale et des 0 au-dessus, ce qui entraîne $R(\mathbf{a})=1$, une contradiction. Donc $\mathcal{J}$ est strict et la preuve est terminée. □

5. Calcul de bases de Gröbner

Nous allons enfin prouver l'existence en exhibant un algorithme qui les produit.

5.1. S-polynômes. On se place toujours dans $\mathbb{K}[x_1,\dots,x_n]$, et on fixe un ordre monomial.

DÉFINITION 6. *Soient f et g deux polynômes. On appelle* S-polynôme *de f et g et on note $S(f,g)$ le polynôme suivant :*

$$S(f,g) := \operatorname{ppcm}(\mathrm{LM}(f),\mathrm{LM}(g))\left(\frac{f}{\mathrm{LT}(f)} - \frac{g}{\mathrm{LT}(g)}\right).$$

Par construction, $S(f,g)$ est un polynôme, et il appartient à $\langle f,g\rangle$. La notation S vient du mot *syzygie*, qui vient lui-même du grec et signifie « sous le même joug ».

La proposition suivante donne la clé de la construction de bases de Gröbner.

PROPOSITION 3. *L'ensemble $G=\{g_1,\dots,g_m\}$ est une base de Gröbner de $\langle G\rangle$ si et seulement si*

$$\forall 1\le i<j\le m,\qquad \overline{S(g_i,g_j)}^{G}=0.$$

DÉMONSTRATION. Le sens direct est clair.

Soit $f\in\mathcal{I}:=\langle G\rangle$. Il s'agit de montrer que $\mathrm{LT}(f)\in\langle\mathrm{LT}(G)\rangle$. Parmi toutes les décompositions

$$f=\sum h_ig_i,$$

on en choisit une qui minimise la quantité

$$\delta:=\max_i \mathrm{LM}(h_ig_i),$$

ce qui est possible puisqu'il n'y a pas de chaîne infinie décroissante. Il suffit de montrer que $\mathrm{LM}(f)=\delta$. Soit $S:=\{i|\,\mathrm{LM}(h_ig_i)=\delta\}$ et quitte à renuméroter les h_i et les g_i, on peut supposer $S=\{1,\dots,k\}$ pour un certain k. La décomposition de f se récrit

$$f=\sum_{i\in S}h_ig_i+\sum_{i\notin S}h_ig_i=\sum_{i\in S}\mathrm{LT}(h_i)g_i+\underbrace{\sum_{i\in S}(h_i-\mathrm{LT}(h_i))g_i+\sum_{i\notin S}h_ig_i}_{\text{chaque terme a un LM}<\delta}.$$

On note $\mathrm{LT}(h_i)=c_iX^{\alpha_i}$ avec la notation $X=(x_1,\dots,x_n)$, c'est-à-dire $\mathrm{LT}(h_i)=c_ix_1^{\alpha_{i_1}}\cdots x_n^{\alpha_{i_n}}$ et on observe que $S(X^{\alpha_i}g_i,X^{\alpha_j}g_j)$ est un multiple de $S(g_i,g_j)$:

$$S(X^{\alpha_i}g_i,X^{\alpha_j}g_j)=\delta\times\left(\frac{X^{\alpha_i}g_i}{\mathrm{LC}(g_i)\delta}-\frac{X^{\alpha_j}g_j}{\mathrm{LC}(g_j)\delta}\right)=\frac{X^{\alpha_i}g_i}{\mathrm{LC}(g_i)}-\frac{X^{\alpha_j}g_j}{\mathrm{LC}(g_j)}.$$

Maintenant,

$$\begin{aligned}\sum_{i=1}^{k}c_iX^{\alpha_i}g_i=\;&c_1\,\mathrm{LC}(g_1)S(X^{\alpha_1}g_1,X^{\alpha_2}g_2)\\&+(c_1\,\mathrm{LC}(g_1)+c_2\,\mathrm{LC}(g_2))\,S(X^{\alpha_2}g_2,X^{\alpha_3}g_3)+\cdots\\&+(c_1\,\mathrm{LC}(g_1)+\cdots+c_{k-1}\,\mathrm{LC}(g_{k-1}))\,S(X^{\alpha_{k-1}}g_{k-1},X^{\alpha_k}g_k)\\&+(c_1\,\mathrm{LC}(g_1)+\cdots+c_k\,\mathrm{LC}(g_k))\,\frac{X^{\alpha_k}g_k}{\mathrm{LC}(g_k)}.\end{aligned}$$

Tous les termes de la somme sauf le dernier ont, par construction des S-polynômes, un monôme de tête strictement inférieur à δ et par hypothèse peuvent être récrits par l'algorithme de division comme combinaison des g_i, dont le terme de tête n'atteint pas δ. Par mimimalité de δ, le dernier terme doit donc avoir pour monôme de tête δ et on a récrit f comme une somme avec un seul monôme égal à δ et tous les autres inférieurs, donc $\mathrm{LM}(f)=\delta$. □

5.2. L'algorithme de Buchberger. L'algorithme de Buchberger est donné en Algorithme 3. Il se base sur des calculs de S-polynômes que l'on réduit puis que l'on ajoute à la base que l'on a déjà.

Algorithme 3 Algorithme de Buchberger

Entrées: $f_1, \dots, f_m \in \mathbb{K}[x_1, \dots, x_n]$, muni d'un ordre monomial.
Sorties: Une base de Gröbner de l'idéal engendré par les f_i,
pour l'ordre monomial donné.

```
G := {f_1, ..., f_m}
S := {S(f_i, f_j), i < j}
tant que S ≠ ∅ faire
  Choisir un p ∈ S
  S := S \ {p}
  g := p̄^G
  si g ≠ 0 alors
    S := S ∪ {S(g, h), h ∈ G}
    G := G ∪ {g}
  fin si
fin tant que
renvoyer G
```

Preuve de l'algorithme. À chaque étape de l'algorithme, l'idéal engendré par G est $\langle f_1, \dots, f_m \rangle$. L'inclusion provient de l'initialisation de G et ensuite G ne s'accroît que de S-polynômes d'éléments de G. Enfin, lorsque l'algorithme termine, tous les S-polynômes de G sont bien réduits à 0 par G, ce qui prouve qu'il s'agit d'une base de Gröbner. Le seul point délicat à prouver est donc la terminaison.

À chaque étape, soit le cardinal de S décroît, soit $\langle \mathrm{LT}(G) \rangle$ croît. Il suffit de montrer que la deuxième possibilité ne peut se produire qu'à un nombre fini d'étapes. L'union de tous ces idéaux $\langle \mathrm{LT}(G) \rangle$ est un idéal. Le résultat est alors une conséquence du lemme de Dickson ci-dessous. □

Lemme 2 (Lemme de Dickson). *Soit A un ensemble de multi-indices en n variables. Alors tout idéal monomial $\mathcal{I} = \langle (x_1, \dots, x_n)^\alpha, \alpha \in A \rangle$ admet une base monomiale finie.*

Démonstration. On procède par récurrence sur le nombre de variables. Pour $n = 1$, on a $\mathcal{I} = \langle X^\beta \rangle$ où $\beta = \min A$.

Supposons $n > 1$. On note $X = x_1, \dots, x_{n-1}$ et $Y = x_n$. Posons

$$\mathcal{J} := \{X^\alpha | \exists m, X^\alpha Y^m \in \mathcal{I}\}$$

Où α est un multi-indice en $n-1$ variables.

L'idéal $\langle \mathcal{J} \rangle$ est un idéal monomial de $\mathbb{K}[X]$. Il admet donc une base finie notée $X^{\alpha_1}, \dots, X^{\alpha_s}$. Posons alors :

$$m_i := \min \{m \in \mathbb{N} | X^{\alpha_i} Y^m \in \mathcal{I}\}, \qquad m := \max_i m_i.$$

Ensuite, pour $k = 0, \dots, m-1$, on considère les « tranches »

$$\mathcal{J}_k := \{X^\alpha | X^\alpha Y^k \in \mathcal{I}\}.$$

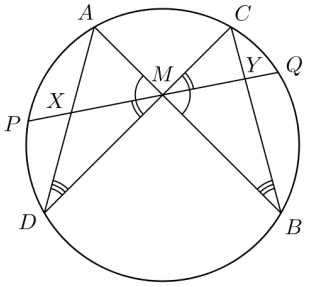
Chaque $\langle \mathcal{J}_k \rangle$ est un idéal monomial de $\mathbb{K}[X]$, et admet par hypothèse de récurrence une base finie $X^{\alpha_1^{(k)}}, \dots, X^{\alpha_{s_k}^{(k)}}$.

Une base finie de $\mathcal{I}$ est donc finalement donnée par

$$\{X^{\alpha_j^{(k)}} Y^k | 0 \le k < m, 1 \le j \le s_k\} \cup \{X^{\alpha_1} Y^m, \dots, X^{\alpha_s} Y^m\}. \qquad \square$$

TP 8

Bases de Gröbner pour la géométrie



Le théorème du papillon est un exercice classique de géométrie Euclidienne. Soit M le milieu d'une corde PQ d'un cercle et soient AB et CD deux autres cordes passant par M ; AD et BC coupent PQ en X et Y respectivement. Il s'agit de montrer que M est aussi le milieu de XY.

Le but de l'exercice est d'utiliser les bases de Gröbner pour parvenir à la preuve sans trop de considérations géométriques. Sans perte de généralité, on pourra choisir de centrer le cercle en 0, de lui donner un rayon 1, de donner à M une abscisse nulle et d'imposer à PQ d'être horizontale. Il s'agit donc de construire un idéal de $\mathbb{K}_{\mathsf{pol}} := \mathbb{Q}[x_A, y_A, x_B, y_B, x_C, y_C, x_D, y_D, x_P, y_P, x_Q, y_Q, x_X, y_X, x_Y, y_Y, x_M, y_M]$ codant la géométrie du problème, puis de tester si le polynôme $x_X + x_Y$ y appartient (ou en toute rigueur s'il appartient à son radical).

1. Écrire une procédure prenant en argument un point et renvoyant un polynôme exprimant que ce point est sur le cercle, une autre prenant trois points et exprimant qu'ils sont alignés ;
2. Former un système mettant en équations le problème à l'aide de ces deux procédures et calculer une base de Gröbner G_1 de ce système pour un ordre du degré lexicographique inverse ;
3. Constater que $x_X + x_Y$ n'appartient pas à l'idéal engendré par G_1, ni même à son radical, et donc que le théorème n'est pas vrai en toute généralité.

Comme souvent en géométrie, l'existence de cas dégénérés rend la propriété fausse ou mal posée. Il est cependant possible de tester la validité *générique* de la propriété en plaçant les paramètres du problème dans le corps de base. Ici, il s'agit donc de calculer la base dans l'anneau

$$\mathbb{K}_{\mathsf{rat}} = \mathbb{Q}(y_M, y_A, y_C)[x_A, x_B, y_B, x_C, x_D, y_D, x_P, y_P, x_Q, y_Q, x_X, y_X, x_Y, y_Y, x_M]$$

plutôt que dans $\mathbb{K}_{\mathsf{pol}}$.

4. Calculer une base de Gröbner G_2 de ce système pour un ordre du degré lexicographique inverse dans $\mathbb{K}_{\mathsf{rat}}$ et vérifier l'appartenance de la condition $x_X + x_Y$ à l'idéal engendré par G_2 dans $\mathbb{K}_{\mathsf{rat}}$.

On peut ensuite identifier puis traiter séparément les cas de dégénérescence.

5. Factoriser les polynômes de G_1 de petit degré et en déduire des cas de dégénérescence à éviter. Poursuivre le calcul jusqu'à avoir identifié tous les cas dégénérés dans $\mathbb{K}_{\mathsf{pol}}$.