INRIA
INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE

Algorithms seminar, 1995-1996

Bruno SALVY, éditeur scientifique

N ° 2992
Septembre 1996

THÈME 2

Rapport de recherche

ISSN 0249-6399

ALGORITHMS SEMINAR,

1995-1996

Bruno Salvy
(Editor)

Abstract

These seminar notes represent the proceedings of a seminar devoted to the analysis of algorithms and related topics. The subjects covered include combinatorics, symbolic computation, asymptotic analysis and average-case analysis of algorithms and data structures.

SÉMINAIRE ALGORITHMES,

1995-1996

Abstract

Ces notes de séminaires représentent les actes, en anglais, d'un séminaire consacré à l'analyse d'algorithmes et aux domaines connexes. Les thèmes abordés comprennent : combinatoire, calcul formel, analyse asymptotique et analyse en moyenne d'algorithmes et de structures de données.

ALGORITHMS SEMINAR
1995–1996

Bruno Salvy[1]
(Editor)

Abstract

These seminar notes represent the proceedings of a seminar devoted to the analysis of algorithms and related topics. The subjects covered include combinatorics, symbolic computation, asymptotic analysis and average-case analysis of algorithms and data structures.

This is the fifth of our series of seminar proceedings. The previous ones have appeared as INRIA Research Reports numbers 1779, 2130, 2381 and 2669. The content of these proceedings consists of English summaries of the talks, usually written by a reporter from the audience[2].

The primary goal of this seminar is to cover the major methods of the average-case analysis of algorithms and data structures. Neighbouring topics of study are combinatorics, symbolic computation and asymptotic analysis.

The study of combinatorial objects—their description, their enumeration according to various parameters, or their random generation—arises naturally in the process of analyzing algorithms that often involve classical combinatorial structures like strings, trees, graphs, and permutations.

Computer algebra plays an increasingly important rôle in this area. It provides a collection of tools that allows one to attack complex models of combinatorics and the analysis of algorithms via *generating functions*; at the same time, it inspires the quest for developing ever more systematic solutions and decision procedures for the analysis of well-characterized classes of problems.

Asymptotic analysis is an essential ingredient in the interpretation of quantitative results supplied by the resolution of combinatorial models. Various asymptotic methods are found to be relevant to the analysis of particular algorithms.

The thirty-two articles included in this book represent snapshots of current research in these areas. A tentative organization of their contents is given below.

PART I. COMBINATORICS

The enumeration of self-avoiding walks in dimension d is a very old open problem of combinatorics. In **[1]**, a related simpler problem is solved. A class of partitions of integers having nice and surprising generating functions is studied in **[2]**. An introduction to symmetric functions, together with work involving q-analogues of the Catalan numbers is given in **[3]**. Sums of powers of harmonic numbers divided by powers of the variable are related to special values of Riemann's ζ function. A uniform approach to the computation of these sums is given in **[4]**. A logics viewpoint on some combinatorial objects is taken in **[5]**. The last summary **[6]** takes a formal language approach to problems related with the study of DNA sequences.

[1]This work was supported in part by the Long Term Research Project Alcom-IT (#20244) of the European Union.

[2]The summaries for the past five years are available on the web at the URL `http://www-rocq.inria.fr/algo/seminars`.

[**1**] Three-Dimensional Convex Polygons. *Mireille Bousquet-Mélou*
[**2**] Lecture Hall Partitions. *Mireille Bousquet-Mélou*
[**3**] Déterminants, nombres de Catalan et fonctions symétriques de Macdonald. *Dominique Gouyou-Beauchamps*
[**4**] Sommes d'Euler. *Philippe Flajolet*
[**5**] A 0-1 Law for Planar Maps. *Kevin Compton*
[**6**] Grammaires et séquences biologiques. *Fabrice Lefebvre*

PART II. SYMBOLIC COMPUTATION

This part starts with a survey [**7**] of numerous algorithms related to linear recurrences and linear differential equations, mostly in the univariate case. New algorithms for the multivariate case are described in [**8**] and [**9**]. The numerical resolution of systems of polynomials is studied from different viewpoints in [**10**] and [**11**]. An algorithm from computational number theory is developed in [**12**]. The next two summaries study specific problems: [**13**] answers the question of describing the functions satisfying all the differential equations satisfied by a given function; [**14**] describes polynomials analogous to the Chebyshev polynomials, but much harder to compute. This part ends with a short presentation of the computation of Padé approximants of various kinds [**15**].

[**7**] Linear Recurrences, Linear Differential Equations and Fast Computation. *Bruno Salvy*
[**8**] Creative Telescoping and Applications. *Frédéric Chyzak*
[**9**] ∂-Finite Functions. *Frédéric Chyzak*
[**10**] Computing the Distance of a Point to an Algebraic Hypersurface and Application to Exclusion Methods. *Xavier Gourdon*
[**11**] Méthodes matricielles pour la résolution de systèmes algébriques. *Ioannis Z. Emiris*
[**12**] Le calcul de grandes valeurs de la fonction $\pi(x)$. *Marc Deléglise*
[**13**] Sur un problème de Rubel. *John Shackell*
[**14**] Les polynômes de Tchébychev entiers. *Bruno Salvy*
[**15**] Algebraic Computation of Matrix-like Padé Approximants. *George Labahn*

PART III. ASYMPTOTIC ANALYSIS

The asymptotic analysis of a class of staircase polygons is studied in [**16**]. It involves a nonlinear q-equation for the generating function, and asymptotics where the Airy function arises. The relevance of this and similar problems to statistical mechanics is the topic of [**17**]. Partitions of integers give rise to very subtle asymptotic analyses. A historical survey of the litterature in that area is given in [**18**]; and [**19**] studies a specific problem. Asymptotic techniques from probability theory are used in [**20**] to study a network that models self-service electrical car pools.

[**16**] The Tricritical Scaling Function of Partially Directed Vesicles. *Thomas Prellberg*
[**17**] The Statistical Mechanics of Vesicles. *Thomas Prellberg*
[**18**] Partitions d'entiers et méthode de Meinardus. *Philippe Dumas*
[**19**] Measures of distinctness for partitions and compositions. *Hsien-Kuei Hwang*
[**20**] Étude asymptotique et charge critique pour les grands réseaux fermés à forme produit via le théorème de la limite centrale. *Jean-Marc Lasgouttes*

PART IV. ANALYSIS OF ALGORITHMS AND DATA STRUCTURES

The Quickselect algorithm uses the partitioning process of Quicksort to find the k-th element among n without sorting them. Its average-case analysis is described in [**21**], as well as analyses of variants of Quickselect. Next, [**22**] shows the relevance of basic hypergeometric series to the analysis of digital search trees and of an approximate counting algorithm. Tools from probability theory are used in the analysis of bin-packing **[23]**. In **[24]**, a problem from computational learning theory is attacked with urn models and involves modified Bessel functions. The last four papers **[25–28]** are related to pattern-matching and strings.

[21] Analysis of Quickselect. *Helmut Prodinger*
[22] Basic hypergeometric series, digital search trees, approximate counting. *Helmut Prodinger*
[23] Biased Random Walks, Lyapunov Functions, and Stochastic Analysis of Best Fit Bin Packing. *Claire Kenyon*
[24] Un modèle d'urnes pour l'apprentissage. *Danièle Gardy*
[25] Pattern Matching Image Compression: Theory, Algorithms and Experiments. *Wojciech Szpankowski*
[26] Fast Approximate Pattern Matching. *Ricardo Baeza-Yates*
[27] Rotation of Periodic Strings and Short Superstring. *Dany Breslauer*
[28] Recherche de motifs : combinatoire et probabilités. *Mireille Régnier*

PART V. MISCELLANY

Worst-case analyses of algorithms give rise to equations with a max operator. An algebraic framework for such equations is surveyed in **[29]**. Next, **[30]** gives an introduction to DNA computers and the biology involved in them. Applications of the Mellin transform in signal processing are listed in **[31]**. The last summary **[32]** is concerned with random boolean formulæ.

[29] Le semi-anneau (max,+) : une introduction. *Stéphane Gaubert*
[30] Computation with DNA. *Alain Hénaut and Didier Contamine*
[31] Utilisation de la transformée de Mellin en traitement de signaux fractals. *Jacques Lévy-Vehel*
[32] Évolution de la satisfiabilité et de la difficulté de formules booléennes aléatoires. Applications pour la résolution. *Olivier Dubois*

Acknowledgements. The lectures summarized here emanate from a seminar attended by a community of researchers in the analysis of algorithms, from the Algorithms Project at INRIA (the organizers are Philippe Flajolet, François Morain and Bruno Salvy) and the greater Paris area—especially École Polytechnique (Jean-Marc Steyaert), University of Paris Sud at Orsay (Dominique Gouyou-Beauchamps) and LITP (Michèle Soria).

The editor expresses his gratitude to the various persons who actively supported this joint enterprise, most notably: Philippe Dumas for his careful rereading, Eithne Murray for straightening out some English texts, and Frédéric Chyzak, Pierre Nicodème and Philippe Dumas for writing more than their share of summaries. Thanks are also due to the speakers and to the authors of summaries. Many of them have come from far away to attend one seminar and kindly accepted to write the summary.

We are also greatly indebted to Virginie Collette for making all the organization work smoothly.

The Editor
B. Salvy

Part 1

Combinatorics

Three-Dimensional Convex Polygons

Mireille Bousquet-Mélou
LaBRI, Université Bordeaux 1

February 26, 1996

[summary by Eithne Murray]

Abstract

A method to enumerate self-avoiding convex polygons, which in theory will work for all dimensions, is presented. The generating series for polygons of dimensions 2 (already known) and 3 are given. They are both the quotients of two D-finite series, and it appears that this property might hold for higher dimensions.

1. Introduction

A very old open problem is to enumerate self-avoiding walks (self-avoiding polygons) in dimension d. This talk answers a slightly more restricted problem by presenting a method of enumerating convex self-avoiding polygons. The 2-dimensional case has already been solved in [3] and [6], but this method works in higher dimensions, and provides a combinatorial interpretation of the 2-dimensional result.

Some basic definitions are required. An *(oriented) polygon* of perimeter $2n$ is a closed path $(s_1, s_2, \ldots, s_{2n})$ of vertices on $\mathbb{Z}^d$ such that s_i and s_{i+1} are neighbours for $1 \leq i \leq 2n$ and $s_{2n+1} = s_1$. It is defined up to cyclic permutations of its vertices. The *rooted* polygon $(s_1, s_1, \ldots, s_{2n})$ *represents* all the polygons formed by the cyclic permutations. A *self-avoiding polygon* is such that $s_i \neq s_j$ for $1 \leq i \neq j \leq 2n$; in other words, it never crosses itself except at the start/end point. A non-empty self-avoiding polygon is also called a *loop*. Note that the polygon (s_1, s_2) is a loop.

Polygons are often represented as words over an alphabet. This representation means the polygons are defined up to a translation in $\mathbb{Z}^d$, which is a requirement for counting them, and also gives a convenient method to define additional properties of the polygons. Thus a rooted polygon of perimeter $2n$ will often be regarded as a word $u = u_1 u_2 \cdots u_{2n}$ on the alphabet $\mathcal{A} = \{1, 2, \ldots, d\} \cup \{\bar{1}, \bar{2}, \ldots, \bar{d}\}$. Then if $(e_1, \ldots, e_d)$ is the canonical basis of $\mathbb{Z}^d$, and $u_i = k$ (resp. $\bar{k}$), then u_i is a unitary step from the vertex s_i to s_{i+1} along e_k (resp. $-e_k$). Note that for all $k \leq d$, the number of occurrences of k in u, denoted $|u|_k$, is equal to the number of occurrences of $\bar{k}$ in u. Conversely, any word u on $\mathcal{A}$ that satisfies $|u|_k = |u|_{\bar{k}}$ for $1 \leq k \leq d$ is a rooted polygon. For example, the polygon $12\bar{1}\bar{2}$ would be a unit square. More examples can be seen in figure 1.

This representation is used to define *dimension*, *unimodal polygon* and *convex polygon* (see below). These concepts are important since the method to count the convex polygons involves decomposing them into their unimodal parts, and counting their loops of each dimension.

The *dimension* of a polygon is the dimension of its convex hull, which is equal to the number of k such that $|u|_k > 0$. For example, the loop (s_1, s_2), represented by $u = k\bar{k}$, has dimension 1.

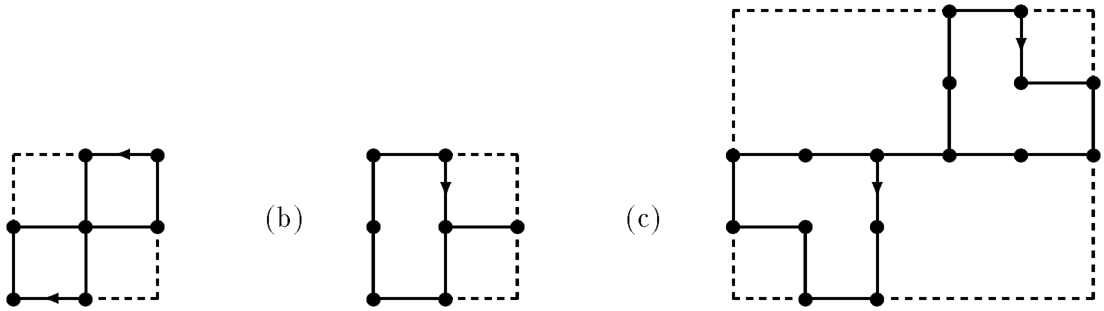

FIGURE 1. (a) staircase (b) unimodal (c) convex polygons

A polygon is *unimodal* if, for each direction k, the polygon can be written $u = vw$ with $|v|_{\bar{k}} = |w|_k = 0$. In other words, all the k's come before all the $\bar{k}$'s in its representative word u, and so all the steps taken in a given direction occur before all the steps taken to return from that direction.

A polygon is *convex* if for each k there is a cyclic permutation of the polygon such that all the k's come before all the $\bar{k}$'s. More intuitively, for each k, and each $a \in \mathbb{R}$, the intersection of a convex polygon with the half-space $\{(a_1, \ldots, a_d) : a_k \leq a\}$ is connected. Another characteristic is that the length of the perimeter of a convex polygon is equal to the length of the perimeter of the smallest bounding box of the polygon. A unimodal polygon is a convex polygon that contains the vertex of minimal coordinates of its smallest bounding box. See figure 1.

2. Enumeration Method

To count the self-avoiding convex polygons, the idea is to count all convex polygons and then remove those that are not self-avoiding. Let P represent the number of all convex polygons of dimension d, and P_k be the number of convex polygons of dimension d with a k-dimensional loop but no loops of dimension $< k$. Then

$$P = P_1 + P_2 + \cdots + P_d. \tag{1}$$

Polygons will be enumerated by using a generating function based on their perimeters. If $\mathcal{P}$ is a set of polygons, then the *perimeter generating function* for the elements of $\mathcal{P}$ is

$$\sum_{u \in \mathcal{P}} t^{|u|/2},$$

where $|u|$ stands for the number of letters of u; and the *multi-perimeter generating function* is

$$\sum_{u \in \mathcal{P}} x_1^{|u|_1} \cdots x_d^{|u|_d}.$$

A *staircase polygon* is a pair of directed paths having the same end-points, so all the steps taken in positive directions (words on $\{1, \ldots, d\}$) occur before all the steps taken in negative directions (words on $\{\bar{1}, \ldots, \bar{d}\}$). The multi-perimeter generating function for staircase polygons, where n_i is the number of steps taken in direction e_i in $\mathbb{Z}^d$, is

$$Z_d(x_1, \ldots, x_d) = \sum_{n_1, \ldots, n_d} \binom{n_1 + \cdots + n_d}{n_1, \ldots, n_d}^2 x_1^{n_1} \cdots x_d^{n_d}$$

(see [4]). This series is D-finite, that is, it satisfies a linear differential equation with polynomial coefficients [7]. Moreover,

$$Z_2(x_1,x_2) = \sum_{n_1,n_2} \binom{n_1+n_2}{n_1,n_2}^2 x_1^{n_1}x_d^{n_2} = \frac{1}{\sqrt{1-2x_1-2x_2-2x_1x_2+x_1^2+x_2^2}} \tag{2}$$

is algebraic. This series has a generalization to Z_λ where λ is a partition [4].

THEOREM 1. *The multi-perimeter generating function of the number of d-dimensional convex polygons that have no 1-dimensional loops is*

$$P - P_1 = E\left[\frac{(d-1)!x_1\cdots x_d(1-x_1)^2\cdots(1-x_d)^2}{(1-x_1-\cdots-x_d)^d}\right]$$

where if $f(x_1,\ldots,x_d) = \sum_{n_1,\ldots,n_d} a_{n_1,\ldots,n_d} x_1^{n_1}\cdots x_d^{n_d}$, *then the* even part *of f is*

$$E\left[f(x_1,\ldots,x_d)\right] = \sum_{n_1,\ldots,n_d} a_{2n_1,\ldots,2n_d} x_1^{2n_1}\cdots x_d^{2n_d}.$$

The proof of this theorem uses the inclusion/exclusion principle and a decomposition of the word-representations of the polygons.

The following gives the formula which will be applied to count convex loops. The idea is that for a convex polygon having loops of dimension d, two cases can occur: either it has only one loop (it itself is a d-dimensional loop), or it can have two loops. There are 2^d possible loop structures, and the loops are unimodal. If the polygon is represented by ul_1vl_2, where the l_i are loops, then uv is essentially a staircase polygon, and so counted by Z_d. Details are presented in [2].

THEOREM 2. *In dimension d, let P_d and Z_d be defined as above, and let U_d be the multi-perimeter generating function for unimodal polygons having only loops of dimension d, and C_d be the generating function for convex polygons having only loops of dimension d. Then*

$$P_d = C_d + 2^{d-1}Z_dU_d^2.$$

Since a convex polygon of dimension d which has only loops of dimension d is self-avoiding, C_d counts the d-dimensional self-avoiding convex polygons. Now U_d can be calculated for all d by rewriting it in terms of Z_d using induction. An important element of the proof is that a loop of a rooted unimodal polygon is unimodal, and hence if a rooted unimodal polygon $u_0l_1u_1l_2u_2$ has loops l_i in $I_i \subset \{1,\ldots,d\}$, then $I_1 \cap I_2 = \emptyset$. Thus a unimodal polygon is made up of a sequence of unimodal loops separated by staircase polygons where the structure of the distribution of the loops can be described by a partition of d. The generating function for unimodal polygons having loops corresponding to this partition can be expressed in terms of Z_λ, λ the partition of d, and U_k, $k \le d$. Then this result, together with equation (1) and theorem 2 gives a means of calculating the number of self-avoiding convex polygons.

3. 2-D Polygons

In dimension $d = 2$, $P - P_1 = P_2$, so combining theorems 1 and 2 gives

$$E\left[\frac{x_1x_2(1-x_1)^2(1-x_2)^2}{(1-x_1-x_2)^2}\right] = C_2 + 2Z_2U_2^2$$

Setting $\Delta = 1 - 2x_1 - 2x_2 - 2x_1x_2 + x_1^2 + x_2^2$, and solving for C_2 using $U_2 = 2\frac{x_1x_2}{\sqrt{\Delta}}$ and (2) gives

$$C_2 = \frac{2x_1x_2A}{\Delta^2} - \frac{8x_1^2x_2^2}{\Delta^{3/2}}$$

where

$$A = 1 - 3x_1 - 3x_2 + 3x_1^2 + 3x_2^2 + 5x_1x_2 - x_1^3 - x_2^3 - x_1^2x_2 - x_1x_2^2 - x_1x_2(x_1 - x_2)^2.$$

This was first proved by Lin and Chang [6], and is a refinement of a result by Delest and Viennot [3]. Alternate proofs are found in [1] and [5]. This work gives a nice combinatorial interpretation of each of the two parts of C_2 in terms of convex polygons having no one-dimensional loops, thereby solving an open problem due to Viennot.

4. 3-D Polygons

This time, the situation is more complicated. Given $P - P_1 = P_2 + P_3$, where $P - P_1$ is calculated using theorem 1, and $P_3 = C_3 + 4Z_3U_3^2$ by theorem 2, it remains to find a way to count P_2, the number of polygons in $\mathbb{Z}^3$ having 2-dimensional loops but no 1-dimension loops. This can be done by a case-by-case analysis of the 7 possible loop structures. The result is

$$C_3 = A(t) + \frac{B(t)}{Z_3}$$

where $A(t)$ and $B(t)$ are algebraic in t, and Z_3 is D-finite. $A(t)$ is of degree 16, and $B(t)$ has degree 8. (The exact value of C_3 would take up a quarter of the page.)

5. Conclusion

This method works because the loops of unimodal polygons are non-overlapping. In theory this method is extensible to higher dimensions, though of course in practice the calculation of the P_i's for $i < d$ would become difficult. Since for each d the series Z_d is D-finite and the series U_d can be written in terms of Z_d, is seems reasonable from the formula to believe that the result will continue to be a quotient of two D-finite series. There may be generalizations to polygons that are convex along $d - 1$ directions, and 3-choice polygons.

Bibliography

[1] Bousquet-Mélou (M.). – Codage des polyominos convexes et équations pour l'énumération suivant l'aire. *Discrete Applied Mathematics*, vol. 48, 1994, pp. 21–43.

[2] Bousquet-Mélou (Mireille) and Guttmann (Anthony J.). – *Enumeration of three-dimensional convex polygons.* – Technical Report n° 1132, Laboratoire Bordelais de Recherche en Informatique, Bordeaux, France, July 1996.

[3] Delest (M.-P.) and Viennot (G.). – Algebraic languages and polyominoes enumeration. *Theoretical Computer Science*, vol. 34, 1984, pp. 169–206.

[4] Guttmann (A. J.) and Prellberg (T.). – Staircase polygons, elliptic integrals, Heun functions and lattice Green functions. *Phys. Rev. E*, vol. 47, 1993, pp. R2233–R2236.

[5] Kim (D.). – The number of convex polyominoes with given perimeter. *Discrete Mathematics*, vol. 70, 1988, pp. 47–51.

[6] Lin (K. Y.) and Chang (S. J.). – Rigorous results for the number of convex polygons on the square and honeycomb lattices. *Journal of Physics Series A: Math. Gen.*, vol. 21, 1988, pp. 2635–2642.

[7] Stanley (R. P.). – Differentiably finite power series. *European Journal of Combinatorics*, vol. 1, n° 2, 1980, pp. 175–188.

Lecture Hall Partitions

Mireille Bousquet-Mélou
LaBRI, Unversité de Bordeaux 1

February 26, 1996

[summary by Dominique Gouyou-Beauchamps]

Abstract

A well-known theorem of Euler [2, Chap. 16] says that the number of partitions of an integer N into distinct parts is equal to the number of partitions of N into odd parts. The talk gives a finite version of this theorem that says that the number of "lecture hall partitions of length n" of N equals the number of partitions of N into small odd parts: $1, 3, 5, \ldots, 2n-1$. This work is a common work with Kimmo Eriksson [1].

1. Lecture hall partitions

Let $\mathcal{D}$ be the set of integer partitions with distinct parts. For $n \geq 1$, let $\mathcal{L}_n$ be the following set of partitions (having possibly some empty parts):

$$\mathcal{L}_n = \{(\lambda_1, \ldots, \lambda_n) : 0 \leq \lambda_1/1 \leq \lambda_2/2 \leq \cdots \leq \lambda_n/n\}.$$

We call the members of $\mathcal{L}_n$ *lecture hall partitions of length n*, since they describe all possible ways of designing a lecture hall with space for up to n rows of seats placed on integer heights, such that at every seat there is a clear view of the speaker without obstruction from the seats in front (Figure 1).

Removing the empty parts puts $\mathcal{L}_n$ in one-to-one correspondence with the following subset of $\mathcal{D}$:

$$\mathcal{D}_n = \left\{(\mu_1, \mu_2, \ldots, \mu_m) : m \leq n \quad \text{and} \quad 0 < \frac{\mu_1}{n-m+1} \leq \frac{\mu_2}{n-m+2} \leq \cdots \leq \frac{\mu_m}{n}\right\}.$$

We will prove the following remarkable theorem.

Theorem 1 (Lecture Hall Theorem). *The generating function for lecture hall partitions of length n is*

$$L_n(q) = \sum_{\lambda \in \mathcal{L}_n} q^{|\lambda|} = \prod_{i=0}^{n-1} \frac{1}{1-q^{2i+1}},$$

where the weight $|\lambda|$ of a partition $\lambda = (\lambda_1, \ldots, \lambda_m)$ is $\lambda_1 + \cdots + \lambda_m$.

Equivalently, the generating function for the partitions of $\mathcal{D}_n$ is $L_n(q)$. Observe that $\mathcal{D}_n \subset \mathcal{D}_{n+1}$ and $\mathcal{D} = \lim_{n\to\infty} \mathcal{D}_n$, so in the limit this theorem yields the familiar Euler identity [2, Chap. 16]: the generating function for the elements of $\mathcal{D}$ is equal to the generating function for the elements of $\mathcal{O}$, the set of integer partitions with odd parts:

$$\sum_{\mu \in \mathcal{D}} q^{|\mu|} = \prod_{i \geq 1}(1+q^i) = \prod_{i \geq 1} \frac{1-q^{2i}}{1-q^i} = \prod_{i \geq 0} \frac{1}{1-q^{2i+1}} = \sum_{\mu \in \mathcal{O}} q^{|\mu|}.$$

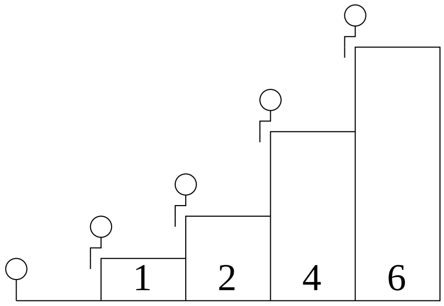

FIGURE 1. The design of a lecture hall of four rows corresponding to the lecture hall partition (1,2,4,6).

We will prove a refinement of the Lecture Hall Theorem. We define the *even* and *odd* weights $|\lambda|_e$ and $|\lambda|_o$ of a partition $\lambda = (\lambda_1, \ldots, \lambda_n)$ by

$$|\lambda|_e = \sum_{0 \le k \le \lfloor (n-1)/2 \rfloor} \lambda_{n-2k} \qquad \text{and} \qquad |\lambda|_o = \sum_{0 \le k \le \lfloor n/2 \rfloor - 1} \lambda_{n-2k-1}.$$

Of course, $|\lambda| = |\lambda|_e + |\lambda|_o$. We will prove the bivariate identity

$$\sum_{\lambda \in \mathcal{L}_n} x^{|\lambda|_e} y^{|\lambda|_o} = \prod_{i=0}^{n-1} \frac{1}{1 - x^{i+1} y^i}.$$

This identity is a corollary of Theorem 3 in section 4, taking $k = l = 2$.

We will in fact discuss a generalization to other sets of partitions of the form $\{(\lambda_1, \lambda_2, \ldots, \lambda_n) : 0 \le \lambda_1/a_1 \le \lambda_2/a_2 \le \cdots \le \lambda_n/a_n\}$ where $(a_1, a_2, \ldots, a_n)$ is a given non-decreasing sequence of integers. We define now $\mathcal{L}_n$ and $S_{(a_1, a_2, \ldots, a_n)}$ as:

$$\mathcal{L}_n = \{(\lambda_1, \ldots, \lambda_n) : 0 \le \lambda_1/a_1 \le \lambda_2/a_2 \le \cdots \le \lambda_n/a_n\} \qquad \text{and} \qquad S_{(a_1, a_2, \ldots, a_n)} = \sum_{\lambda \in \mathcal{L}_n} q^{|\lambda|}.$$

Here are surprisingly simple values of $S_{(a_1, a_2, \ldots, a_n)}$:

$$S_{1,2,5,8} = \frac{1}{(1-q)(1-q^3)(1-q^8)(1-q^{13})},$$
$$S_{1,2,5,8,19} = \frac{1}{(1-q)(1-q^4)(1-q^7)(1-q^{11})(1-q^{27})},$$
$$S_{1,2,5,8,19,30} = \frac{1}{(1-q)(1-q^3)(1-q^8)(1-q^{13})(1-q^{31})(1-q^{49})},$$
$$S_{1,2,7,12,41} = \frac{1}{(1-q)(1-q^5)(1-q^9)(1-q^{31})(1-q^{53})}.$$

2. Reduction of lecture hall partitions

Fix a non-decreasing sequence $a = (a_i)_{i \ge 1}$ of positive integers, and fix a positive integer n. An n-tuple $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in \mathbb{N}^n$ is a lecture Hall partition if and only if $\lambda_i \ge \lceil \lambda_{i-1} a_i / a_{i-1} \rceil$ for $2 \le i \le n$. For $1 \le i \le n$, let $\lambda^{(i)} = (0, \ldots, 0 a_i, a_{i+1}, \ldots, a_n) \in \mathbb{N}^n$. If λ belongs to $\mathcal{L}_n$, then the sum $\lambda + \lambda^{(i)}$ also belongs to $\mathcal{L}_n$.

LEMMA 1. *Let λ be a lecture hall partition belonging to $\mathcal{L}_n$. Then $\lambda - \lambda^{(i)}$ belongs to $\mathcal{L}_n$ if and only if $\lambda_i - \lceil \lambda_{i-1} a_i / a_{i-1} \rceil \ge a_i$ for $1 \le i \le n$.*

DEFINITION 1. A lecture hall partition of length n is said to be *reduced* if $0 \le \lambda_i - \lceil \lambda_{i-1} a_i / a_{i-1} \rceil < a_i$ for $1 \le i \le n$. The set of reduced partitions of $\mathcal{L}_n$ will be denoted by $\mathcal{R}_n$.

LEMMA 2. *Let λ be a lecture hall partition of length n. Then there exists a unique reduced lecture hall partition μ and a unique sequence of integers $(k_i)_{1\le i\le n}$ such that $\lambda = \mu + \sum_{i=1}^n k_i \lambda^{(i)}$.*

Consequently, the generating function for lecture hall partitions of length n is

$$S_n = \sum_{\lambda\in\mathcal{L}_n} x^{|\lambda|_e} y^{|\lambda|_o} = \frac{P_n(x,y)}{\prod_{i=1}^n (1 - x^{|\lambda^{(i)}|_e} y^{|\lambda^{(i)}|_o})}$$

where the polynomial $P_n(x,y) = \sum_{\mu\in\mathcal{R}_n} x^{|\mu|_e} y^{|\mu|_o}$ enumerates reduced lecture hall partitions.

3. An involution on $\mathcal{R}_n$

For $\mu \in \mathcal{R}_n$, let $\mu^* = (\mu_1^*, \ldots, \mu_n^*)$ be the unique n-tuple such that

$$\begin{cases} \mu_{n-2k}^* = \mu_{n-2k} & \text{for } n-2k \ge 1 \\ \mu_{n-2k-1}^* - \left\lceil \frac{a_{n-2k-1}}{a_{n-2k-2}} \mu_{n-2k-2}^* \right\rceil = \left\lfloor \frac{a_{n-2k-1}}{a_{n-2k}} \mu_{n-2k} \right\rfloor - \mu_{n-2k-1} & \text{for } n-2k-1 \ge 1. \end{cases}$$

THEOREM 2. *The correspondence $\mu \mapsto \mu^*$ defines an involution on the set $\mathcal{R}_n$.*

We can extend the involution $\mu \mapsto \mu^*$ into a bijection f from $\mathcal{R}_n \times [0, a_{n+1}]$ onto $\mathcal{R}_{n+1}$, by defining

$$f(\mu_1, \ldots, \mu_n; i) = \left(\mu_1^*, \ldots, \mu_n^*, \left\lceil \frac{a_{n+1}}{a_n} \mu_n^* \right\rceil + i \right).$$

It is clear that:

$$\begin{aligned} |f(\mu,i)|_o &= |\mu^*|_e = |\mu|_e, \\ |f(\mu,i)|_e &= i - |\mu|_o + \sum_k \left(\left\lceil \frac{a_{n-2k+1}}{a_{n-2k}} \mu_{n-2k} \right\rceil + \left\lfloor \frac{a_{n-2k-1}}{a_{n-2k}} \mu_{n-2k} \right\rfloor \right). \end{aligned}$$

4. The $(k-l)$-sequences

By a $(k-l)$-sequence we shall mean a sequence a defined by the initial values $a_1 = 1$ and $a_2 = l$ and the following recurrence relations:

$$\begin{cases} a_{2n} = l a_{2n-1} - a_{2n-2} & \text{for } n \ge 2 \\ a_{2n+1} = k a_{2n} - a_{2n-1} & \text{for } n \ge 1 \end{cases}$$

where $k, l \ge 2$ are two integers. We obtain

$$\begin{aligned} |f(\mu,i)|_o &= |\mu|_e, \\ |f(\mu,i)|_e &= i - |\mu|_o + \begin{cases} k|\mu|_e & \text{if } n \text{ is even}, \\ l|\mu|_e & \text{if } n \text{ is odd}. \end{cases} \end{aligned}$$

This implies that the generating functions $P_n(x,y) = \sum_{\mu\in\mathcal{R}_n} x^{|\mu|_e} y^{|\mu|_o}$ for reduced lecture hall partitions can be computed inductively via the following recurrence relations:

$$P_{2n+1}(x,y) = \frac{1 - x^{a_{2n+1}}}{1-x} P_{2n}(x^k y, x^{-1}) \qquad \text{and} \qquad P_{2n}(x,y) = \frac{1 - x^{a_{2n}}}{1-x} P_{2n-1}(x^l y, x^{-1})$$

with the initial condition $P_0 = 1$.

The sequence a^* is defined by $a_0^* = 0$, $a_1^* = 1$ and the recurrence relations:

$$\begin{cases} a_{2n} = l a_{2n-1} - a_{2n-2} & \text{for } n \geq 2 \\ a_{2n+1} = k a_{2n} - a_{2n-1} & \text{for } n \geq 1. \end{cases}$$

THEOREM 3. *Given a (k,l)-sequence a, the generating functions $S_n = \sum_{\lambda \in \mathcal{L}_n} x^{|\lambda|_e} y^{|\lambda|_o}$ for lecture hall partitions of even and odd length are given by:*

$$S_{2n} = \prod_{i=1}^{2n} \frac{1}{1 - x^{a_i} y^{a_i^*}} \qquad \text{and} \qquad S_{2n+1} = \prod_{i=1}^{2n+1} \frac{1}{1 - x^{a_{i+1}^*} y^{a_{i-1}}}.$$

5. Limit theorems

Taking the limit $n \to \infty$ in Theorem 3 leads to the following results:

THEOREM 4. *For $k \in \mathbb{N}$ and $k \geq 2$, the bivariate generating function of partitions $(\mu_1, \ldots, \mu_n)$ such that $\frac{\mu_{i+1}}{\mu_i} > \frac{k+\sqrt{k^2-4}}{2}$ is:*

$$\sum_{\mu} x^{|\mu|_e} y^{|\mu|_o} = \prod_{i \geq 1} \frac{1}{1 - x^{a_i} y^{a_{i-1}}}$$

with $a_0 = 0$, $a_1 = 1$ and $a_{i+1} = k a_i - a_{i-1}$.

THEOREM 5. *For $k \in \mathbb{N}$ and $k \geq 2$, the generating function of partitions $(\mu_1, \ldots, \mu_n)$ such that $\frac{\mu_{i+1}}{\mu_i} > \frac{k+\sqrt{k^2-4}}{2}$ is:*

$$\sum_{\mu} q^{|\mu|} = \prod_{i \geq 1} \frac{1}{1 - q^{e_i}}$$

with $e_1 = 1$, $e_2 = k+1$ and $e_{i+1} = k e_i - e_{i-1}$.

EXAMPLE. $k = 2$. In that case $\mu_{i+1} > \mu_i$ and we obtain the Euler identity [2, Chap. 16]:

$$\sum_{\mu \in \mathcal{D}} q^{|\mu|} = \prod_{i \geq 0} \frac{1}{1 - q^{2i+1}}.$$

EXAMPLE. $k = 3$. In that case $\mu_{i+1} > \frac{3+\sqrt{5}}{2}\mu_i$ and:

$$\sum_{\mu \in \mathcal{L}_n} q^{|\mu|} = \frac{1}{(1-q)(1-q^4)(1-q^{11})(1-q^{29})(1-q^{76})\cdots} = \prod_{i \geq 1} \frac{1}{(1 - q^{e_i})},$$

with $e_1 = 1$, $e_2 = 4$ and $e_{i+1} = 3e_i - e_{i-1}$. In fact $e_i = F_{2i-3} + F_{2i-1}$ where F_i is the ith Fibonnaci number.

6. Questions

(1) Give a characterization of the sequences $(a_1, \ldots, a_n)$ that have a simple expression for the corresponding generating functions.
(2) Find finite version of other theorems like the Rogers-Ramanujan theorem for instance.

Bibliography

[1] Bousquet-Mélou (M.) and Eriksson (K.). – *Lecture Hall Partitions.* – Rapport LaBRI n° 1107-96, Laboratoire Bordelais de Recherche en Informatique, Université de Bordeaux I, 1996.
[2] Euler (L.). – *Introductio in analysin infinitorum.* – Marcum-Michælem Bousquet, Lausannæ, 1748.

Determinants, Catalan numbers and Macdonald's symmetric functions

Dominique Gouyou-Beauchamps
LRI, Orsay

March 25, 1996

[summary by Bruno Salvy]

Abstract

A famous conjecture in the theory of symmetric functions states that the coefficients of Macdonald's polynomials in the basis of Schur's symmetric functions are positive. F. Bergeron, A. M. Garsia and M. Haiman have introduced a linear operator ∇ whose eigenvalues are related to Macdonald's polynomials. Properties of this operator in a special case are related to combinatorial determinants which can be evaluated by the Gessel-Viennot technique relating them to non-intersecting paths.

1. Introduction to symmetric functions

This section and the following one are based on [4].

Partitions and symmetric functions are strongly related. A *partition* is an infinite decreasing sequence of positive integers $\lambda = (\lambda_1, \lambda_2, \dots)$, with finitely many non-zero elements. The index of the last non-zero element in the partition is called its *length* and is denoted $\ell(\lambda)$; the sum of the λ_i's is called the *weight* of the partition and is denoted $|\lambda|$. For $n \geq \ell(\lambda)$, λ is identified with the n-tuple of its first elements. Then if $x = (x_1, \dots, x_n)$ is a n-tuple of indeterminates, x^λ denotes the monomial $x_1^{\lambda_1} \cdots x_n^{\lambda_n}$ and S_n^λ denotes a maximal set of distinct permutations of λ.

A fundamental basis of symmetric functions is constituted by the *monomial* symmetric functions, indexed by the partitions: for $n \geq \ell(\lambda)$,

$$m_\lambda(x_1, \dots, x_n) = \sum_{\sigma \in S_n^\lambda} x^{\sigma(\lambda)}.$$

Clearly, the set of m_λ's, when λ runs through all partitions of length at most n is a basis of the symmetric polynomials in n variables. The set Λ of symmetric functions is *defined* as the vector space generated by the m_λ's.

Three important sets of symmetric functions, $e_r = m_{(1^r)}$ (elementary), $h_r = \sum_{|\lambda|=r} m_\lambda$ (complete) and $p_r = m_{(r)}$ (power sum), have simple generating functions:

$$\begin{aligned}
E(t) &= \sum_{r\geq 0} e_r t^r = 1 + t\sum_i x_i + t^2 \sum_{i<j} x_i x_j + \cdots = \prod_{i>0} (1 + x_i t),\\
H(t) &= \sum_{r\geq 0} h_r t^r = 1 + t\sum_i x_i + t^2 \sum_{i\leq j} x_i x_j + \cdots = \prod_{i>0} \frac{1}{1 - x_i t},\\
P(t) &= \sum_{r\geq 0} p_r t^r = \sum_i x_i + t\sum_i x_i^2 + \cdots = \sum_{i>0} \frac{x_i}{1 - x_i t}.
\end{aligned}$$

Each of these three sets of symmetric functions generates Λ as a ring. In all three cases, defining for a partition λ a function $f_\lambda = f_{\lambda_1} f_{\lambda_2} \cdots$, where f is e, h or p yields a basis of Λ as a vector space, when λ runs through the set of partitions.

Formulæ giving the coordinates of one of these functions in terms of the other families are obtained by extracting the coefficient of t^n in the following straightforward relations between the generating functions:

$$(1) \qquad E(t)H(-t) = 1, \qquad P(t) = \frac{H'(t)}{H(t)}, \qquad P(-t) = \frac{E'(t)}{E(t)}.$$

The last two equations yield the classical Newton formulæ between power sums and elementary symmetric functions. Integrating these equation also yields

$$H(t) = \exp \sum_{r>0} p_r \frac{t^r}{r} = \sum_\lambda p_\lambda \frac{t^{|\lambda|}}{z_\lambda}, \qquad E(t) = \sum_\lambda (-1)^{|\lambda|-\ell(\lambda)} p_\lambda \frac{t^{|\lambda|}}{z_\lambda}, \qquad \text{with} \qquad z_\lambda = \prod_{i>0} i^{m_i} m_i!,$$

where m_i is the number of occurrences of the part i in λ.

Another family of symmetric functions, the *Schur functions*, is defined for $n \geq \ell(\lambda)$ by

$$s_\lambda(x_1, \ldots, x_n) = \frac{\det(x_i^{\lambda_j+n-j})_{1\leq i,j\leq n}}{\det(x_i^{n-j})_{1\leq i,j\leq n}}.$$

The s_λ's are indeed polynomials, since the numerator is a polynomial in the x_i's which vanishes whenever $x_i = x_j$ with $i \neq j$, and thus is a multiple of the Vandermonde determinant in the denominator. The s_λ form another basis of Λ. They are related to the complete and elementary symmetric functions by the Jacobi-Trudi identities:

$$(2) \qquad s_\lambda = \det(h_{\lambda_i-i+j})_{1\leq i,j\leq n}, \qquad s_\lambda = \det(e_{\lambda'_i-i+j})_{1\leq i,j\leq m},$$

where λ' is the *conjugate* of λ, i.e. the partition whose Ferrers diagram is the reflexion of that of λ with respect to the diagonal.

Recall that a *Young tableau* of shape λ is a Ferrers diagram of shape λ with squares numbered by consecutive positive integers $1, 2, \ldots, r$, the numbers increasing strictly in each column and weakly along each row. The *weight* $w(T)$ of a tableau T is the r-tuple $(m_1, \ldots, m_r)$, m_i counting the number of occurrences of i. The tableau is called *standard* when it contains each number $1, 2, \ldots, |\lambda|$ exactly once, i.e. its weight is $(1^{|\lambda|})$. The Schur functions are related to tableaux by

$$s_\lambda = \sum_T x^{w(T)},$$

summed over all tableaux T of shape λ. From this follows that the coordinates $K_{\lambda\mu}$ of s_λ in the basis m_μ are positive integers counting the number of tableaux of shape λ and weight μ and thus are positive integers. Macdonald's conjecture is a generalization of this property.

All these symmetric functions can also be related by expanding in several ways the doubly infinite product $P(x, y) = \prod(1 - x_i y_j)^{-1}$. Thus one gets

$$(3) \qquad \prod_{i,j} \frac{1}{1 - x_i y_j} = \sum_\lambda z_\lambda^{-1} p_\lambda(x) p_\lambda(y) = \sum_\lambda h_\lambda(x) m_\lambda(y) = \sum_\lambda m_\lambda(x) h_\lambda(y) = \sum_\lambda s_\lambda(x) s_\lambda(y).$$

This motivates the definition of a scalar product by $\langle h_\lambda, m_\mu \rangle = \delta_{\lambda\mu}$ for all partitions λ, μ, where $\delta_{\lambda\mu}$ is the Kronecker delta. The relations (3) show that the p_λ's form an orthogonal basis, while the s_λ's form an orthonormal basis of Λ. This property characterizes the Schur functions.

The next step is to consider the Hall-Littlewood symmetric functions with one parameter

$$P_\lambda(x_1, \ldots, x_n; t) = \sum_{\sigma \in S_n^\lambda} \sigma\left(x^\lambda \prod_{i<j} \frac{x_i - tx_j}{x_i - x_j} \right).$$

These functions interpolate between the monomial symmetric functions—obtained when $t = 1$—and the Schur symmetric functions—obtained when $t = 0$. They form a $\mathbb{Z}[t]$-basis of $\Lambda[t]$. Therefore, one may consider the polynomials $K_{\lambda\mu}(t)$ defined by

$$s_\lambda(x) = \sum_\mu K_{\lambda\mu}(t) P_\mu(x; t).$$

The polynomials $K_{\lambda\mu}(t)$ turn out to have positive coefficients, and this has been proved by Lascoux and Schützenberger who gave an expression of the form

$$K_{\lambda\mu}(t) = \sum_T t^{c(T)},$$

summed over all tableaux T of shape λ and weight μ, where $c(T)$ is a certain combinatorial function of the tableau (its *charge*). Several expansions of the product $P(x, y; t) = \prod_{i,j} (1 - tx_iy_j)/(1 - x_iy_j)$ lead to results very similar to those obtained above and to the definition of a scalar product on $\Lambda[t]$ with values in $\mathbb{Q}(t)$ with respect to which the $P_\lambda(x; t)$ are orthogonal. Also $\langle P_\lambda, m_\mu \rangle = 0$ when $\mu \not\subseteq \lambda$ (the Ferrers diagram of μ is not included in that of λ), and together with their orthogonality this characterizes the P_λ. The basis which is dual to the Schur functions $s_\lambda(x)$ with respect to this scalar product is denoted $S_\lambda(x; t)$, i.e., $\langle S_\lambda(x; t), s_\mu(x) \rangle = \delta_{\lambda\mu}$.

2. Macdonald's conjecture

Macdonald's conjecture concerns the Macdonald symmetric functions, which have two parameters. The doubly infinite product

$$\Pi(x, y; q, t) = \prod_{i,j} \frac{(tx_iy_j; q)_\infty}{(x_iy_j; q)_\infty}, \qquad \text{where} \qquad (a; q)_\infty = \prod_{r=0}^{\infty} (1 - aq^r),$$

can be expanded as

$$\Pi(x, y; q, t) = \sum_\lambda \frac{1}{z_\lambda(q, t)} p_\lambda(x) p_\lambda(y), \qquad \text{with} \qquad z_\lambda(q, t) = z_\lambda \prod_{i=1}^{\ell(\lambda)} \frac{1 - q^{\lambda_i}}{1 - t^{\lambda_i}}.$$

This motivates the definition of a scalar product by

$$\langle p_\lambda, p_\mu \rangle_{q,t} = \delta_{\lambda\mu} z_\lambda(q, t).$$

The Macdonald symmetric functions are defined uniquely by two properties: they are orthogonal with respect to this scalar product and they decompose in the basis of the monomial symmetric functions as

$$P_\lambda(x; q, t) = m_\lambda + \sum_{\mu < \lambda} u_{\lambda\mu} m_\mu.$$

When $q = t$, they reduce to the Schur functions s_λ, and when $q = 0$ to the Hall-Littlewood functions $P_\lambda(x; t)$.

For a partition λ and a cell $c = (i, j)$ of its Ferrers diagram, one defines the *arm* of c to be $a(c) = \lambda_i - j$ and its *leg* to be $l(c) = \lambda'_j - i$. Now we can state Macdonald's conjecture.

Conjecture 1 (Macdonald). *The coefficients $K_{\lambda\mu}(q,t)$ of the following decomposition are polynomials with positive coefficients:*

$$(4)\ \tilde{H}_\lambda(x;t) := c_\lambda(q,t)P_\lambda(x;q,t) = \sum_\lambda K_{\lambda\mu}(q,t)S_\lambda(x;t), \quad \textit{where} \quad c_\lambda(q,t) = \prod_{c\in\lambda}(1-q^{a(c)}t^{l(c)+1}).$$

These coefficients possess a lot of structure. For instance, for $\lambda = (3,1)$, Eq. (4) becomes

$$\tilde{H}_{(3,1)} = S_{(4)} + (q^2+t+q)S_{(3,1)} + (t+q)qS_{(2,2)} + (tq+q^2+t)qS_{(2,1,1)} + tq^3S_{(1,1,1,1)}.$$

Only special cases of Macdonald's conjecture have been proved.

3. Combinatorial properties of ∇ when $t = 1$

In order to study the polynomials $\tilde{H}_\lambda$, Bergeron, Garsia and Haiman have introduced a linear operator ∇ which is diagonal in the basis $\tilde{H}_\lambda$, with eigenvalues $T_\lambda(q,t) = q^{n(\lambda')}t^{n(\lambda)}$, where $n(\lambda) = \sum(i-1)\lambda_i$. The matrix of ∇ in the Schur basis turns out to have a fascinating structure of which much is still only conjectured [2].

The aim of [1] is to study this operator in more detail in the special case $t = 1$. Then the basis $\tilde{H}_\lambda(x;q) := \tilde{H}_\lambda(x;q,1)$ becomes multiplicative: $\tilde{H}_\lambda(x;q) = \tilde{H}_{(\lambda_1)}(x;q)\tilde{H}_{(\lambda_2)}(x;q)\cdots$ and ∇ becomes multiplicative too. Thus any identity involving symmetric functions gives rise to a similar identity for its image by ∇. In particular, from (2) follows $\nabla(s_\lambda) = \det(\nabla e_{\lambda'_i+j-i})_{1\le i,j\le m}$. Moreover, still when $t = 1$, the coordinate $\nabla(e_n)|_{e_n}$ of $\nabla(e_n)$ on e_n is a q-Catalan number C_n, with generating function $C(x)$ defined by $C(x) = 1 + xC(x)C(xq)$. Hence $D(\lambda) := \nabla(s_\lambda)|_{e_n} = \det(C_{\lambda'_i+j-i})_{1\le i,j\le m}$, and the idea of [1] is to use the Gessel-Viennot technique [3] to evaluate determinants of this type for various classes of partitions λ. Typical results are summarised in the following theorem.

Theorem 1.

$$D((k^k)) = (-1)^{\binom{k}{2}}q^{\frac{k(k-1)(4k+1)}{6}}, \qquad D((k^{k+1})) = (-1)^{\binom{k+1}{2}}q^{\frac{k(k+1)(4k-1)}{6}},$$

$$D((k^{k+2})) = (-1)^{\binom{k+2}{2}+1}q^{\frac{k(k+1)(4k-1)}{6}+k^2}[k+1], \quad D((k+1)^k) = (-1)^{\binom{k}{2}}q^{\frac{k(k-1)(4k+7)}{6}}[k+1],$$

$$D((k+2)^k) = (-1)^{\binom{k}{2}}q^{\frac{k(k-1)(4k+13)}{6}}\frac{[k+1][k+2]([k+1]+q[k+2])}{[2][3]},$$

where $[k] = 1 + q + q^2 + \cdots + q^{k-1}$.

Another linear operator diagonal in the basis $\tilde{H}_\lambda$ is also studied in [1]. Similar techniques apply and results of a similar kind are obtained.

Bibliography

[1] Bergeron (François), Bousquet-Mélou (Mireille), and Gouyou-Beauchamps (Dominique). – Preprint, 1996.

[2] Bergeron (François), Garsia (Adriano), and Haiman (Mark). – New identities and conjectures for Macdonald's $\tilde{H}_\mu[x;q,t]$ polynomials. – Preprint, 1995.

[3] Gessel (I.) and Viennot (G.). – Binomial determinants, paths and hook length formulae. *Advances in Mathematics*, vol. 58, 1985, pp. 300–321.

[4] Macdonald (Ian Grant). – *Symmetric functions and Hall polynomials.* – Oxford University Press, 1995, 2nd edition, *Oxford Mathematical Monographs.*

Euler sums

Philippe Flajolet
INRIA Rocquencourt

January 29, 1996

[summary by Jean-Paul Allouche]

In 1742 Goldbach wrote a letter to Euler proposing the study of the sums

$$S_{p,q} := \sum_{n=1}^{\infty} \left(\frac{1}{1^p} + \frac{1}{2^p} + \cdots + \frac{1}{n^p} \right) \frac{1}{n^q} = \sum_{n=1}^{\infty} \frac{H_n^{(p)}}{n^q},$$

where $H_n^{(r)}$ and $H_n = H_n^{(1)}$ are the *harmonic numbers* defined by

$$H_n^{(r)} := \sum_{j=1}^{n} \frac{1}{j^r}.$$

Euler was able to compute all the sums $S_{p,q}$ for $p + q \leq 13$, for example

$$\sum_{n=1}^{\infty} \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n} \right) \frac{1}{n^2} = 2\zeta(3).$$

Then, in 1906, Nielsen gave relations linking the sums $S_{p,q}$ having the same weight $w = p + q$. Hence the $S_{p,q}$ of odd weight are polynomials in the values of zeta, for example

$$S_{2,5} = \sum_{n=1}^{\infty} \frac{H_n^{(2)}}{n^5} = 5\zeta(2)\zeta(5) + 2\zeta(3)\zeta(4) - 10\zeta(7).$$

Many similar identities have then been found or conjectured. Some of them involve multiple zeta functions; see the papers of Bayley, D. and J. Borwein, De Doelder, Don Zagier, Girgensohn, Hoffman, Markett.

The authors [1] propose a simple and unifying method that gives most of the known results about these identities. Furthermore they are able to prove some conjectures. The key idea is to use a contour integral *with a well-chosen kernel.*

1. The idea of the authors: a simple case

Let us denote by $I(p,q)$ the integral

$$I(p,q) = \frac{1}{2i\pi} \int_C (\psi(-s) + \gamma)^2 \frac{ds}{s^q},$$

where C is a circle whose radius goes to infinity, and where ψ is the logarithmic derivative of the Γ function. Denoting by γ the Euler constant, we have

$$\psi(z) = \frac{d}{dz} \log \Gamma(z) = -\gamma - \frac{1}{z} + \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+z} \right).$$

Hence, when s tends to a positive integer m, then

$$(\psi(-s)+\gamma)^2 \underset{s\to m}{=} \frac{1}{(s-m)^2} + 2H_m \frac{1}{s-m} + \cdots .$$

If s tends to 0, we use the relation $\psi(s)+\gamma = -1/s + \zeta(2)s - \zeta(3)s^2 + \cdots$. Hence, by residue computation the Euler sum $S_{1,q}$ can be expressed as an explicit quantity which is "homogeneous" of degree 2 in the zeta values.

In the general case the authors consider integrals

$$\frac{1}{2i\pi}\int_C r(s)\xi(s)\,ds,$$

where r is a rational function, and ξ a *suitable* kernel. They then obtain numerous results: some of them were already known, but some of them were only conjectures.

2. A zoo of beautiful identities

The authors obtain the following results.

THEOREM 1 (EULER). *Let q be an integer ≥ 2. Then*

$$S_{1,q} = \sum_{n=1}^{\infty} \frac{H_n}{n^q} = \left(1+\frac{1}{2}q\right)\zeta(q+1) - \frac{1}{2}\sum_{k=1}^{q-2}\zeta(k+1)\zeta(q-k).$$

For example

$$\sum_{n=1}^{\infty}\frac{H_n}{n^2} = 2\zeta(3), \quad \sum_{n=1}^{\infty}\frac{H_n}{n^3} = \frac{5}{4}\zeta(4), \quad \sum_{n=1}^{\infty}\frac{H_n}{n^4} = 3\zeta(5)-\zeta(2)\zeta(3).$$

THEOREM 2 (EULER, BORWEIN ET AL.). *If the weight $m=p+q$ is odd, then*

$$\sum_{n=1}^{\infty}\frac{H^{(p)}(n)}{n^q} = \zeta(m)\left[\frac{1}{2} - \frac{(-1)^p}{2}\binom{m-1}{p} - \frac{(-1)^p}{2}\binom{m-1}{q}\right] + \frac{1-(-1)^p}{2}\zeta(p)\zeta(q)$$
$$+ (-1)^p \sum_{k=1}^{\lfloor \frac{p}{2}\rfloor}\binom{m-2k-1}{q-1}\zeta(2k)\zeta(m-2k) + (-1)^p\sum_{k=1}^{\lfloor \frac{q}{2}\rfloor}\binom{m-2k-1}{p-1}\zeta(2k)\zeta(m-2k),$$

where any occurrence of $\zeta(1)$ has to be replaced by 0.

If we then use the symmetry $S_{p,q}+S_{q,p} = \zeta(p)\zeta(q)+\zeta(p+q)$, we obtain

$$5\sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^6} + 2\sum_{n=1}^{\infty}\frac{H_n^{(3)}}{n^5} = -\frac{21}{2}\zeta(8) + 10\zeta(3)\zeta(5) + \frac{9}{2}\zeta(4)^2,$$
$$7\sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^8} + 2\sum_{n=1}^{\infty}\frac{H_n^{(3)}}{n^7} = -33\zeta(10) + 14\zeta(3)\zeta(7) + 15\zeta(4)\zeta(6) + 8\zeta(5)^2,$$
$$7\sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^8} - 2\sum_{n=1}^{\infty}\frac{H_n^{(4)}}{n^6} = -\frac{229}{5}\zeta(10) + 14\zeta(3)\zeta(7) + 21\zeta(4)\zeta(6) + 10\zeta(5)^2,$$
$$9\sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^{10}} + 2\sum_{n=1}^{\infty}\frac{H_n^{(3)}}{n^9} = -\frac{143}{2}\zeta(12) + 18\zeta(3)\zeta(9) + 21\zeta(4)\zeta(8) + 24\zeta(5)\zeta(7) + \frac{25}{2}\zeta(6)^2,$$

$$8\sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^{10}}-\sum_{n=1}^{\infty}\frac{H_n^{(4)}}{n^8}=-\frac{575}{7}\zeta(12)+16\zeta(3)\zeta(9)+24\zeta(4)\zeta(8)+28\zeta(5)\zeta(7)+\frac{295}{21}\zeta(6)^2,$$

$$7\sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^{10}}+\sum_{n=1}^{\infty}\frac{H_n^{(5)}}{n^7}=-73\zeta(12)+28\zeta(5)\zeta(7)+21\zeta(4)\zeta(8)+14\zeta(3)\zeta(9)+\frac{35}{3}\zeta(6)^2.$$

Then

$$\sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^2}=\frac{7}{4}\zeta(4),\quad \sum_{n=1}^{\infty}\frac{H_n^{(3)}}{n^3}=\frac{1}{2}\zeta(3)^2+\frac{1}{2}\zeta(6),\quad \sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^4}=\zeta(3)^2-\frac{1}{3}\zeta(6).$$

Theorem 3 (Borwein et al.). *The following relations hold.*

$$\sum_{n=1}^{\infty}\frac{(H_n)^2}{n^q}-S_{2,q}=qS_{1,q+1}-\frac{q(q+1)}{6}\zeta(q+2)+\zeta(2)\zeta(q).$$

For example

$$\sum_{n=1}^{\infty}\frac{(H_n)^2}{n^3}=\frac{7}{2}\zeta(5)-\zeta(2)\zeta(3),$$

$$\sum_{n=1}^{\infty}\frac{(H_n)^2}{n^5}=6\zeta(7)-\zeta(2)\zeta(5)-\frac{5}{2}\zeta(3)\zeta(4),$$

$$\sum_{n=1}^{\infty}\frac{(H_n)^2}{n^7}=\frac{55}{6}\zeta(9)-\zeta(2)\zeta(7)-\frac{7}{2}\zeta(3)\zeta(6)-\frac{5}{2}\zeta(4)\zeta(5)+\frac{1}{3}\zeta(3)^3,$$

and

$$\sum_{n=1}^{\infty}\frac{H_n^2}{n^6}-\sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^6}=\frac{91}{12}\zeta(8)-8\zeta(3)\zeta(5)+\zeta(2)\zeta(3)^2,$$

$$\sum_{n=1}^{\infty}\frac{H_n^2}{n^8}-\sum_{n=1}^{\infty}\frac{H_n^{(2)}}{n^8}=\frac{473}{40}\zeta(10)-10\zeta(3)\zeta(7)-5\zeta(5)^2+\zeta(4)\zeta(3)^2+2\zeta(2)\zeta(3)\zeta(5),$$

$$\sum_{n=1}^{\infty}\frac{(H_n)^2}{n^2}=\frac{17}{4}\zeta(4),$$

$$\sum_{n=1}^{\infty}\frac{(H_n)^2}{n^4}=\frac{307}{24}\zeta(6)-5\zeta(2)\zeta(4)-2\zeta(3)^2.$$

Theorem 4. *If $i+j+k$ is odd, with $i>1$, $j>1$, $k>1$, then*

$$[(-1)^k+(-1)^{i+j}]\sum_{n\geq 1}\frac{H_n^{(i)}H_n^{(j)}}{n^k}+A+B+C+D+E+F=0,$$

where

$$A=(-1)^{i+j+k}\zeta(i)\zeta(j)\zeta(k)+(-1)^{i+k}\zeta(i)S_{j,k}+(-1)^{j+k}\zeta(j)S_{i,k},$$

$$B=-2(-1)^k\sum_{q+2r+t=i}\binom{j+q-1}{q}\binom{k+t-1}{k-1}[(-1)^qS_{j+q,k+t}+(-1)^j\zeta(j+q)\zeta(k+t)]\zeta(2r),$$

$$C=-2(-1)^k\sum_{p+2r+t=j}\binom{i+p-1}{p}\binom{k+t-1}{k-1}[(-1)^pS_{i+p,k+t}+(-1)^i\zeta(i+p)\zeta(k+t)]\zeta(2r),$$

$$
\begin{aligned}
D &= -2(-1)^k \sum_{2r+t=i+j} \binom{k+t-1}{k-1} \zeta(2r)\zeta(k+t),\\
E &= (-1)^{i+j}\,[-S_{i,j+k} - S_{j,i+k} - \zeta(j)S_{i,k} - \zeta(i)S_{j,k}\\
&\qquad + \zeta(i+j+k) + \zeta(i+k)\zeta(j) + \zeta(j+k)\zeta(i) + \zeta(i)\zeta(j)\zeta(k)],\\
F &= \sum_{p+q+r=i+j+k} \zeta(2r)\lambda_p^{(i)}\lambda_q^{(j)},
\end{aligned}
$$

and

$$
\lambda_0^{(i)} = 1, \quad \lambda_1^{(i)} = \lambda_2^{(i)} = \cdots = \lambda_{i-1}^{(i)} = 0, \quad \lambda_{i+t}^{(i)} = (-1)^i \zeta(t) \binom{t+i-1}{i-1}.
$$

The summations are over the indices ≥ 0. One has to replace $\zeta(0)$ by $-\frac{1}{2}$, and $\zeta(1)$ by 0.

Corollary 1 (Borwein and Girgensohn). *Let $c > 1$. If the weight $a+b+c$ is even, the triple zeta function $\zeta(a,b,c) = \sum_{0<n_1<n_2<n_3} \frac{1}{n_1^a n_2^b n_3^c}$ can be reduced to linear Euler sums.*

Theorem 5. *(i) The cubic expression $\sum_{n=1}^{\infty} \frac{(H_n)^3}{n^q} - 3\sum_{n\geq 1} \frac{H_n H_n^{(2)}}{n^q}$ can be expressed in terms of the zeta values, for any weight.*

(ii) For even weights, $\sum_{n=1}^{\infty} \frac{(H_n)^3}{n^q}$ can be computed in terms of $S_{2,q+1}$ and polynomials in the zeta values.

As a consequence, this gives a proof of conjectures of Bailey, Borwein and Girgensohn:

Corollary 2. *We have*

$$
\begin{aligned}
\sum_{n=1}^{\infty} \frac{(H_n)^3}{(n+1)^2} &= \frac{15}{2}\zeta(5) + \zeta(2)\zeta(3)\\
\sum_{n=1}^{\infty} \frac{(H_n)^3}{(n+1)^3} &= -\frac{33}{16}\zeta(6) + 2\zeta(3)^2\\
\sum_{n=1}^{\infty} \frac{(H_n)^3}{(n+1)^4} &= \frac{119}{16}\zeta(7) - \frac{33}{4}\zeta(3)\zeta(4) + 2\zeta(2)\zeta(5)\\
\sum_{n=1}^{\infty} \frac{(H_n)^3}{(n+1)^6} &= \frac{197}{24}\zeta(9) - \frac{33}{4}\zeta(4)\zeta(5) - \frac{37}{8}\zeta(3)\zeta(6) + \zeta(3)^3 + 3\zeta(2)\zeta(7).
\end{aligned}
$$

3. Other relations?

If the reader wants to discover other relations, including relations on alternating Euler sums, read the details of the proofs, check that he was able to discover tricky integration contours, or know where some of these relations naturally occur in theoretical computer science, he should read this very nice paper. He will certainly enjoy it.

Bibliography

[1] Flajolet (Philippe) and Salvy (Bruno). – *Euler Sums and Contour Integral Representations.* – Research Report n° 2917, Institut National de Recherche en Informatique et en Automatique, June 1996.

A Zero-One Law for Maps

Kevin Compton

University of Michigan, Ann Arbor, U.S.A.

June 10, 1996

[summary by Frédéric Chyzak]

Abstract

A class of structures has a 0–1 law when any property expressible in a certain logic has limiting probability 0 or 1 as the size of the structures tends to infinity. We prove 0–1 laws for classes of maps of a given genus. This is a joint work with E. Bender and B. Richmond [1].

1. Definition of the problem

Let S be a set of primitive elements called *sorts*. A *vocabulary* Σ consists of a collection of constant and relation symbols, together with a mapping from each constant symbol to a sort, and a mapping from each relation symbol to a sequence of sorts, the *arity* of the relation (see [4] for an introduction to model theory). A *multi-sorted structure* $\mathcal{A}$ over Σ then consists of

- a collection of disjoints sets (or *universes*) A_s, one for each sort s;
- elements $c^{\mathcal{A}} \in A_s$, one for each constant symbol c of sort s;
- relations $R^{\mathcal{A}} \subset A_{s_1} \times \cdots \times A_{s_p}$, one for each relation symbol R of arity $(s_1, \ldots, s_p)$.

A *class* of structures is a set of structures defined on the same vocabulary. In the study of random structures, one says that a class of finite structures *has a 0–1 law* when any property expressible in a certain logic has limiting probability 0 or 1 as the size of the structures tends to infinity. The *relational signature* of a class of structures over Σ is the common set of relation symbols in the vocabulary Σ, together with their arities. A famous theorem by Glebskiĭ, Kogan, Liogon'kiĭ and Talanov [9], and proved independently by Fagin [7], states that if $\mathcal{C}$ is the class of all structures for a given relational signature, then $\mathcal{C}$ has a first-order 0–1 law. However, deciding the limiting probability of a given property is a difficult problem, as formalized by a theorem by Grandjean: when a class $\mathcal{C}$ has a 0–1 law, the set of first-order sentences of limiting probability 1 is PSPACE-complete.

A *map* $\mathcal{M}$ is an embedding of a connected graph $\mathcal{G}$ into a closed surface $\mathcal{S}$ such that all connected components of $\mathcal{S} \setminus \mathcal{G}$, the *faces* of $\mathcal{M}$, are homeomorphic to a disc. Let $t = 1 - (v - e + f)/2$ be the *genus* of $\mathcal{M}$, with v, e and f its number of vertices, edges and faces respectively. When t is an integer, the map is called *orientable*. The *size* $|\mathcal{M}|$ of a map is e. The purpose of this exposition is to provide similar results to the theorems mentioned above for maps, even in the non-orientable case. Our main result is the following theorem [1].

THEOREM 1. *The class of all maps on surfaces of fixed genus has a 0–1 law. The set of first-order sentences of limiting probability 1 for this class has lower bound complexity of* $\mathrm{DTIME}(\exp_\infty(cn))$, *for some* $c > 0$.

(Recall that $\exp_\infty(n) = 2^{2^{\cdot^{\cdot^{2}}}}$, with n nested exponentiations.)

The 0–1 law theorem for structures cannot be applied to maps, since the latter do not form a full class of structures of a given relational since. Besides, we have to explain how maps can be represented as structures.

2. Representation of maps as structures

Any naive attempt of representing a map $\mathcal{M}$ on a surface $\mathcal{S}$ by its graph, i.e., by its set of edges, is bound to fail. Indeed, this representation would not encapsulate any information about the embedding of $\mathcal{M}$ into $\mathcal{S}$: easy examples show that isomorphic graphs need not correspond to homeomorphic maps, and that the order of edges around a vertex has to be taken into account. However, on a non-orientable surface there is no consistent way to choose an edge order around each vertex.

A solution stems from an idea of Edmonds [5], later elaborated by Tutte [10] as a basis for a combinatorial theory of maps: to each edge, one associates a pair of *darts*, pointing in opposite directions. On orientable surfaces, a possible representation of maps is then given by an involution α on the set of darts, mapping a dart to its opposite dart, together with a permutation β whose cycles consist of all darts out of a vertex, listed clockwise. Then, $\alpha\beta$ is a permutation whose cycles consist of all darts around a face, listed counter-clockwise. One is thus able to determine the embedding using α and β. In the context of possibly non-orientable surfaces, a map is analogously described as a structure by the sets U_v, U_d and U_f of its vertices, darts and faces, together with incidence relations $I(x_v, x_d)$ and $J(x_f, x_d)$ of darts with vertices and faces, a co-dart relation $C(x_d, x_{d'})$ and a dart adjacency relation $A(x_d, x_{d'}, x_f)$. The co-dart relation is an analogue for α, while the dart adjacency relation encapsulates the information formerly supplied by β, specifying a face to supply the orientable information.

3. Ehrenfeucht-Fraïssé games

The 0–1 law theorem for structures still does not apply to maps: not all structures of signature (I, J, C, A) are maps. We overcome this difficulty in the case of the class of all maps on surfaces of a fixed genus by determining subclasses of limiting probability 1.

The sentences of first-order logic under consideration for our 0–1 laws can all be written in the form $S = \theta_1 x_1 \ldots \theta_r x_r \phi(x_1, \ldots, x_r)$, where the θ_i's are quantifiers, either $\forall$ or $\exists$, the x_i's are variables and ϕ is a boolean expression free from quantifiers built on the x_i's using conjunctions and disjunctions. The *rank* of the sentence S is the integer r. Let $\mathcal{A}$ and $\mathcal{B}$ be two structures with same relational signature. We write $\mathcal{A} \equiv_m \mathcal{B}$ when both structures satisfy exactly the same sentences of rank m. This defines an equivalence relation between structures. The next paragraph describes this equivalence relation by a game-theoretic approach.

The Ehrenfeucht-Fraïssé game is an m-round game between two players called *Spoiler* and *Duplicator* and played on a pair of structures $\mathcal{A}$ and $\mathcal{B}$ of same relational signature. In each round, Spoiler picks any element from either structure and Duplicator responds by picking any element from the other structure. This yields two substructures $\mathcal{A}' = \{a_1, \ldots, a_m\} \subset \mathcal{A}$ and $\mathcal{B}' = \{b_1, \ldots, b_m\} \subset \mathcal{B}$, with relations induced in a natural way. Duplicator wins if he is able to choose his responses so as to make $\mathcal{A}'$ and $\mathcal{B}'$ isomorphic; if not, Spoiler wins. Duplicator *has a winning strategy* if and only if he is capable of winning for any choices made by Spoiler. A fundamental result used in the sequel is the Ehrenfeucht-Fraïssé theorem [6, 8] which states that Duplicator has a winning strategy in the m-round first-order game played on two structures $\mathcal{A}$ and $\mathcal{B}$ if and only if $\mathcal{A} \equiv_m \mathcal{B}$.

Now, the relation $\equiv_m$ defines a finite number of (possibly infinite) equivalence classes on the ambient class. It can be proved that one of these classes has limiting probability 1, and this suffices to prove our theorem. For the sake of clarity, we present the idea of the proof on a simplified example only.

4. A 0–1 law by a 3^{r-k} strategy for a simplified problem

For this example, the class of structures under consideration is the set of square toroidal grids with a unary relation (we simply tag some vertices). We play r-round Ehrenfeucht-Fraïssé games on pairs of grids. The crucial fact we use is that any fixed square plane grid with vertices tagged at random appears in a toroidal grid with limiting probability 1.

It follows that Duplicator has a strategy to win *almost surely*, i.e., with limiting probability 1. Define a *distance* between two vertices of a grid by the minimum number of edges in a connecting path. The *ball* $N(c_1, \ldots, c_p; d)$ is the set of vertices at distance at most d from any c_i. Let $\mathcal{A}$ and $\mathcal{B}$ be two structures. Assume we are in round $k+1$ and that $a_1, \ldots, a_k$ have already been picked out of $\mathcal{A}$, $b_1, \ldots, b_k$ out of $\mathcal{B}$ in a way such that $N(a_1, \ldots, a_k; 3^{r-k})$ and $N(b_1, \ldots, b_k; 3^{r-k})$ are isomorphic, when viewed as substructures with naturally induced relations. Now, Spoiler picks an element out of either structure, say a_{k+1} out of $\mathcal{A}$—the case b_{k+1} out of $\mathcal{B}$ is symmetric. If $N(a_1, \ldots, a_{k+1}; 3^{r-k-1}) \subset N(a_1, \ldots, a_k; 3^{r-k})$, then Duplicator can trivially choose b_{k+1} in $N(b_1, \ldots, b_k; 3^{r-k})$ so that $N(a_1, \ldots, a_{k+1}; 3^{r-k-1})$ and $N(b_1, \ldots, b_{k+1}; 3^{r-k-1})$ are isomorphic. Otherwise, there is almost surely a ball in the complement of $N(b_1, \ldots, b_k; 3^{r-k-1})$ in $\mathcal{B}$ which is isomorphic to $N(a_{k+1}; 3^{r-k-1})$. Duplicator then chooses b_{k+1} to be its center. After r rounds, the balls $N(a_1, \ldots, a_r; 1)$ and $N(b_1, \ldots, b_r; 1)$ are almost surely isomorphic. Thus, Duplicator wins almost surely by following the strategy that we have just described. By the Ehrenfeucht-Fraïssé theorem, $\mathcal{A} \equiv_r \mathcal{B}$ almost surely. Therefore, one of the (finitely many) equivalence classes of $\equiv_r$ has limiting probability 1. Call it $\mathcal{E}_r$.

Consider now a first-order sentence S of rank r on toroidal grids. By the Ehrenfeucht-Fraïssé theorem, the set of all grids satisfying S is either contained in $\mathcal{E}_r$, or disjoint from $\mathcal{E}_r$. In the former case S has limiting probability 1, in the latter 0. We have thus proved a 0–1 law for the class of toroidal grids with a unary relation.

5. A 0–1 law for maps of a given genus

We first recall two difficult results on maps.

The first result [2, Sec. 5] plays the rôle of the crucial fact we used in the previous section, namely the limiting probability 1 of the appearance of a fixed plane grid in a toroidal grid. It states that for a class $\mathcal{C}$ of maps of fixed genus, there is a $c > 0$ such that for any given planar map $\mathcal{P}$, the property for maps in $\mathcal{C}$ to contain more than cn disjoint copies of $\mathcal{P}$ has limiting probability 1.

The second result [3] is about *representativity* of maps. The representativity of a map $\mathcal{M}$ on a surface $\mathcal{S}$ is the smallest number of intersections a non-contractible curve in $\mathcal{S}$ has with $\mathcal{M}$. The result is that for a class of maps of fixed genus, there is a $c > 0$ such that the property for maps to have representativity more than $c \ln n$ has limiting probability 1. This result is used in the proof of Theorem 1 to ensure the planarity of certain submaps built on balls playing a rôle similar to the $N(a_1, \ldots, a_k; 3^{r-k})$ of the previous section.

Next, the proof of Theorem 1 runs as for the example of the previous section: we prove a first-order 0–1 law for the class of all maps of a given genus by showing that for each r, Duplicator has an almost surely winning strategy in r-round Ehrenfeucht-Fraïssé games. More specifically, this strategy is a 3^{r-k} strategy using balls around elements picked by Spoiler and Duplicator. However, the notion of distance used is not that of the previous section. The proper distance to prove the result is by means of *quadrangulations* of maps. For a given map $\mathcal{M}$ on a surface, add a new point on each edge and a point in each face. Next add new edges from the new points on the edges to the new points in the faces. The quadrangulation of $\mathcal{M}$ is then the new map on the same surface built in this way. This construction induces a natural mapping from a map $\mathcal{M}$ to its quadrangulation $\mathcal{Q}$. We extend this map to the dart representation of $\mathcal{M}$ by mapping both co-darts defined by an edge to the image of this edge in $\mathcal{Q}$. A distance is then defined on the set $U_v \cup U_d \cup U_f$ of all vertices, darts and faces of the dart representation, as the distance between the images in $\mathcal{Q}$. This distance is not a metric, since two co-darts are at distance 0 for each other. However, the concept of balls it induces is sufficient for the proof of Theorem 1.

6. Conclusions

Theorem 1 has been refined for several classes of maps on a surface of fixed genus [1] (see this reference for missing definitions): the class of all maps; the class of smooth maps; the class of k-connected maps where k is 2 or 3; the class of k-connected triangulations where k is 1, 2 or 3. However, the question of a 0–1 law for planar graphs remains open, though we believe it should be true.

As for complexity results, we proved an $\exp_\infty(cn)$ *lower* bound for the complexity of the set of first-order sentences of limiting probability 1 in the case of the dart representation. Another result holds for an *extended* dart representation (see [1] for the definition): in this extended representation, we proved undecidability. What we have not been able to prove is an *upper* bound in the case of the dart representation, though we feel $\exp_\infty(dn)$ is a good candidate for such an upper bound.

Finally, all results presented here concern sentences of first-order logic. An extension to other logics seems reasonable, in particular to MSO (monadic second-order) logic, with application to the theory of databases.

Bibliography

[1] Bender (Edward A.), Compton (Kevin J.), and Richmond (L. Bruce). – Zero-one laws for maps. – In preparation, 1996.

[2] Bender (Edward A.), Gao (Zhi-Cheng), McCuaig (William D.), and Richmond (L. Bruce). – Submaps of maps I: General 0-1 laws. *Journal of Combinatorial Theory*, vol. 55, n° B, 1992, pp. 104–117.

[3] Bender (Edward A.), Gao (Zhi-Cheng), and Richmond (L. Bruce). – Almost all rooted maps have large representativity. *Journal of Graph Theory*, vol. 18, 1994, pp. 545–555.

[4] Chang (Chen Chung) and Keisler (H. Jerome). – *Model theory.* – North-Holland, Amsterdam, 1990, third edition, *Studies in Logic and the Foundations of Mathematics*, vol. 73.

[5] Edmonds (Jack R.). – A combinatorial representation for polyhedral surfaces. *Notices of the American Mathematical Society*, vol. 7, 1960, p. 646.

[6] Ehrenfeucht (A.). – An application of games to the completeness problem for formalized theories. *Fundamenta Mathematicae*, vol. 49, 1961, pp. 129–141.

[7] Fagin (Ronald). – Probabilities on finite models. *Journal of Symbolic Logic*, vol. 41, 1976, pp. 50–58.

[8] Fraïssé (Roland). – *Sur quelques classifications des systèmes de relations.* – Technical report, Université d'Alger, 1954. English summary.

[9] Glebskiĭ (Y. V.), Kogan (D. I.), Liogon'kiĭ (M. I.), and Talanov (V. A.). – Range and degree of realizability of formulas in the restricted predicate calculus. *Cybernetics*, vol. 5, 1969, pp. 142–154. – English translation.

[10] Tutte (William T.). – Combinatorial oriented maps. *Canadian Journal of Mathematics*, vol. 31, 1979, pp. 986–1004.

A grammar-based unification of several alignment and folding algorithms

Fabrice Lefebvre
LIX - École Polytechnique

June 24, 1996

[summary by Pierre Nicodème]

Abstract

We show that many popular models of folding and/or alignment may be described by a new formalism: multi-tape S-attribute grammars (MTSAGs). This formalism relieves the designer of biological models of implementation details. We present also a tool which, given a MTSAG, will output an efficient parser for this grammar and show that MTSAGs offer a new, efficient and useful way to handle stochastic context-free grammars. This summary is an extended abstract of [7].

1. Introduction

We shall see here that most popular models of alignment and/or folding of DNAs, RNAs or proteins, HMMs (Hidden Markov Models) [5], SCFGs (Stochastic Context-Free Grammars) [8] and CMs (Covariance Models) [3] share a common representation in terms of a new formalism: Multi-Tape S-Attribute Grammars (MTSAGs). This formalism is not only a help for the description of old or new methods. We designed and implemented a tool which, from the high-level description given by a MTSAG, will automatically generate the C source of an efficient C parser which is able to compute alignments and foldings. The speed and memory requirements of such generated parsers stand the comparison with programs manually written from dynamic programming relations. As a consequence, we show how to automatically build SCFGs from sets of unaligned, unfolded RNAs.

2. Definitions

We define a special "m-tape" alphabet which will handle sequence alignments, and then a "m-tape" extension of context-free grammars which will handle structures of alignments.

Definition 1. A *m-tape alphabet* Σ is a product of m alphabets $\Sigma^{(i)}$ augmented with the empty string: $\Sigma = \bigotimes_{i=1\ldots m}(\Sigma^{(i)} \cup \{\epsilon\})$. An element $a_1 \cdots a_l$ of the free monoid Σ^*, generated by formal concatenation of m-tape elements of Σ, is called a *m-tape alignment* of length l. The empty alignment of Σ^* is denoted by ϵ.

Example. $(abba, dcd)$ is a 2-tape input string on $\Sigma^{(1)} = \{a, b\}$ and $\Sigma^{(2)} = \{c, d\}$. We shall also write this 2-tape input string as $\substack{abba \\ dcd}$, which is a somewhat more natural notation in the context of alignments. This 2-tape input string has a 2-tape input substring $\substack{bb \\ dc}$.

Definition 2. Given any m-tape alignment $a_1 \cdots a_l$, we get a m-tape input string by concatenation, or *ϵ-deletion*, of symbols of the projection of $a_1 \cdots a_l$ on every tape.

$$\Sigma^* \longrightarrow \langle \Sigma^* \rangle, a_1 \cdots a_l \longrightarrow \langle a_1 \cdots a_l \rangle.$$

$$
\begin{array}{rlr}
start \longrightarrow & frame0 & (0) \\
frame0 \longrightarrow & frame0 \begin{bmatrix} X \\ X \end{bmatrix} \mid \begin{bmatrix} - \\ - \end{bmatrix} & (0) \\
\mid & frame0 \begin{bmatrix} X \\ Y \end{bmatrix} & (2) \\
\mid & frame1 \begin{bmatrix} - \\ X \end{bmatrix} \mid frame2 \begin{bmatrix} X \\ - \end{bmatrix} & (1) \\
frame1 \longrightarrow & frame1 \begin{bmatrix} X \\ X \end{bmatrix} & (1) \\
\mid & frame1 \begin{bmatrix} X \\ Y \end{bmatrix} & (3) \\
\mid & frame0 \begin{bmatrix} X \\ - \end{bmatrix} \mid frame2 \begin{bmatrix} - \\ X \end{bmatrix} & (3) \\
frame2 \longrightarrow & frame2 \begin{bmatrix} X \\ X \end{bmatrix} & (1) \\
\mid & frame2 \begin{bmatrix} X \\ Y \end{bmatrix} & (3) \\
\mid & frame0 \begin{bmatrix} - \\ X \end{bmatrix} \mid frame1 \begin{bmatrix} X \\ - \end{bmatrix} & (3)
\end{array}
$$

FIGURE 1. In this weighted left-regular grammar, weights are written in parentheses after each group of productions having the same weight. Later on, weights will be turned into attribute evaluation functions.

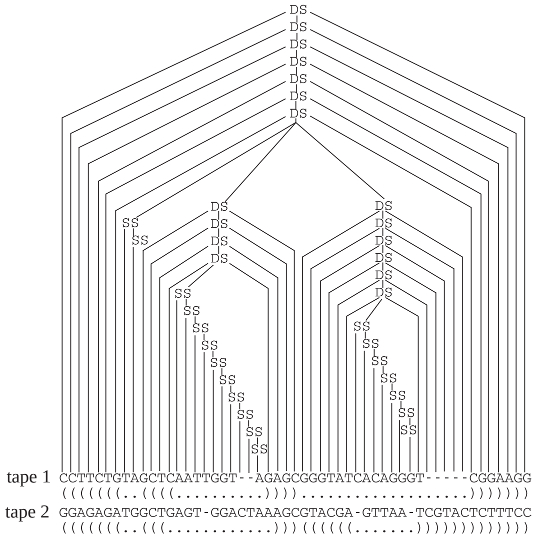

FIGURE 2. Derivation tree of an alignment of two RNAs.

EXAMPLE. Our 2-tape input string $\substack{abba \\ dcd}$ may be defined as an ϵ-deletion of the alignments $\langle \begin{bmatrix} \epsilon \\ d \end{bmatrix} \begin{bmatrix} a \\ \epsilon \end{bmatrix} \begin{bmatrix} b \\ \epsilon \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix} \begin{bmatrix} a \\ d \end{bmatrix} \rangle$ or $\langle \begin{bmatrix} a \\ \epsilon \end{bmatrix} \begin{bmatrix} b \\ d \end{bmatrix} \begin{bmatrix} \epsilon \\ c \end{bmatrix} \begin{bmatrix} b \\ \epsilon \end{bmatrix} \begin{bmatrix} a \\ d \end{bmatrix} \rangle$.

Searls did show that the alignment of two strings according to some edit-distance may be carried out by some simple 2-tape nondeterministic finite automaton (NFA) with weighted transitions [9]. The sought alignment has a minimal total weight. The set of alignments recognized by a Searls' NFA is a regular language over our 2-tape terminal alphabet, and may be described by a regular grammar with weighted productions (see figure 1).

As regular grammars are a proper subset of context-free grammars, we found natural to generalize this idea of alignment to m-tape (i.e. the terminal alphabet is a subset of a m-tape alphabet) context-free grammars (MTCFGs) and their recognizing devices, namely m-tape nondeterministic pushdown automata (NPDA). Weighted transitions of NFA are easily translated into weighted pop-transitions of NPDA. The sought alignment is obtained from a sequence of pop-transitions of the NPDA which has an optimal (minimal for some problems, maximal for others, etc...) total weight.

Figure 2 shows how alignments and structures may be deduced from a single m-tape derivation. The underlying grammar may be easily recovered. Base pairings are inferred from derivations of DS (Double-Strand) and they are given below each tape. Notice that a double-strand has been defined as a substructure whose ends must be paired on at least one tape, whereas a single-strand (SS) may only have unpaired bases on both tapes.

DEFINITION 3. A ***m*-tape *context-free grammar*** $G = (V_T, V_N, P, S)$ consists of a finite set of terminals V_T such that V_T *is a subset of a* ***m*****-*tape alphabet***, a finite set of nonterminals V_N such that $V_N \cap V_T = \emptyset$, a finite set of productions (rewriting rules) P and a start symbol $S \in V_N$. Let $V = V_T \cup V_N$ denote the vocabulary of the grammar. Each production in P has the form $A \rightarrow \alpha$, where $A \in V_N$ and $\alpha \in V^*$. A is the left-hand side of the production and α its right-hand side.

A *derivation tree* is a planar representation of a sequence of derivations (replacements of a nonterminals A in a string of V^* by strings α such that $A \rightarrow \alpha$) and it is a result of *parsing*. The language $L(G)$ is the set of m-tape input strings generated by G: $L(G) = \{\langle u \rangle \in \langle V_T^* \rangle \mid S \rightarrow^* u\}$.

EXAMPLE. The following toy MTCFG will align two properly parenthesized strings interspersed with a:

$$S \to \left[\begin{smallmatrix} (\\ (\end{smallmatrix}\right] S \left[\begin{smallmatrix}) \\ ) \end{smallmatrix}\right] \mid \left[\begin{smallmatrix} a \\ a \end{smallmatrix}\right] \mid \left[\begin{smallmatrix} a \\ - \end{smallmatrix}\right] \mid \left[\begin{smallmatrix} - \\ a \end{smallmatrix}\right] \mid \left[\begin{smallmatrix} - \\ - \end{smallmatrix}\right] \mid SS$$

In this MTSAG, the structure defined by parentheses must be the same on both tapes, but substrings of a may be aligned with gaps (denoted by $-$ in terminals instead of ϵ, because a dash is appropriate, and even expected, in the context of alignments).

DEFINITION 4. Let $G = (V_T, V_N, P, S)$ be a proper m-tape context-free grammar. For every tape i $(1 \leq i \leq m)$, define the *projected grammar* $G^{(i)}$ as the conversion of the grammar $(V_T^{(i)}, V_N, P^{(i)}, S)$ into a proper grammar, where $V_T^{(i)}$ and $P^{(i)}$ are the sets of values on tape i of all the elements of V_T and P respectively.

EXAMPLE. The MTCFG of the preceding example has the same projected grammar on both tapes :

$$S \to (S) \mid () \mid a \mid SS$$

Projected grammars are useful for the study of the complexity of our parsing algorithm as a function of the ambiguity of MTCFGs.

We said earlier that we could assign a cost to each alignment or folding produced by a NPDA, thanks to weights on pop-transitions. This cost-evaluation step is essential for the determination of an optimal cost alignment or folding.

We use the general mechanism of synthesized attributes, or S-attributes which, together with MTCFGs, give us m-tape S-attribute context-free grammars, or MTSAGs. S-attributes are attributes which are assigned to every vertex of a derivation tree and which are computed from the bottom of a derivation tree (i.e. every terminal has a known S-attribute) to its root by means of attribute evaluation functions associated to grammar productions. Thanks to these functions, the computation of the final attribute of the derivation tree does not have to rely on a fixed, predetermined, operation (summation, multiplication, ...), as it would have been the case with weighted productions. In our implementation, attribute evaluation functions are C functions. We have already shown the effectiveness of S-attributes with our adaptation of the thermodynamic model of folding to context-free grammars [6]. This algorithm uses a parse table to store the shared forest of derivation trees of a m-tape input string.

DEFINITION 5. A *m-tape S-attribute grammar* is denoted by $G = (V_T, V_N, P, S, \mathcal{A}, S_{\mathcal{A}}, F_P)$. It is an extension of a m-tape context-free grammar $G = (V_T, V_N, P, S)$, where an attribute $x \in \mathcal{A}$ is attached to each symbol $X \in V$ and a string of attributes $\lambda \in \mathcal{A}^*$ to each string $\alpha \in V^*$. $S_{\mathcal{A}}$ is a function from V_T to $\mathcal{A}$ assigning attributes to terminals. F_P is a set of functions from $\mathcal{A}^*$ to $\mathcal{A}$. A function $f_{A \to \alpha}$ is in F_P iff $A \to \alpha$ is in P.

The attribute λ of a string α is the concatenation of the attributes of the symbols in α. When a function $f_{A \to \alpha}$ is applied to the attribute λ of a string α derived from A, it returns the attribute x of A (hence the bottom-up computation of attributes).

3. Syntax analysis for MTSAGs

A generalization of Cocke-Younger-Kasami's algorithm (CYK) would be an easy algorithm to parse m-tape input strings. This algorithm has a time complexity of $O(n^3)$ and a space complexity of $O(n^2)$ when only one tape of size n is considered [1]. A generalization to m tapes, each of size n, would lead to an algorithm having a complexity of $O(n^{3m})$ in time and $O(n^{2m})$ in space.

To overcome the limitations of CYK's algorithm, we generalized our parsing algorithm for 1-tape MTSAGs [6].

When constructing the parse table, a minimum condition of usefulness is applied. This condition means that an item is never add to an entry if it has no chance of being used in a derivation tree, up to the already parsed part of the m-tape input string. This condition is akin to a condition verified by Earley's parsing algorithm and it is the key to the lower parsing complexities of our algorithm when some projections of the underlying MTSAG are unambiguous.

In fact, out algorithm may be considered as an improvement of Earley's algorithm, where Earley's items $[\Delta \to \alpha \cdot \beta, i], (\alpha, \beta \in V^*)$ which share the same α and i are factorized into a single item $[\Delta \rightharpoonup \alpha, i]$. Also, non-kernel items of Earley's algorithm are replaced by much smaller sets of expected non-terminals.

Proposition 1 (1-tape complexity). *Let G be a proper 1-tape MTSAG and let $r \geq 1$ be the maximum number of nonterminals appearing at the right-hand side of a production of G. For a tape of length n, the time and space complexities of the previous parsing algorithm are, in order of decreasing constraints on G:*

- *Equal and at most $O(n)$ if G is LR(k) and not right-recursive (this encompasses left-regular grammars);*
- *equal and at most $O(n^2)$ if G is unambiguous;*
- *$O(n^{r+1})$ and $O(n^r)$ for a generic proper MTSAG.*

Proposition 2 (m-tape complexity). *Let G be a proper m-tape MTSAG. The time complexity of our parsing algorithm on G is equal to the product of the parsing complexities of the same algorithm applied on each tape i with each projected grammar $G^{(i)}$. The same kind of result holds for space complexities. Hence the time complexity is at most $O(n^{m(r+1)})$, and the space complexity is at most $O(n^{mr})$, for m-tapes of size n*

In practice, MTSAGs that we used verified $r \leq 2$ and thus the time and space complexities of our parsers for those grammars were respectively $O(n^{3m})$ and $O(n^{2m})$ at most, but were sometimes much better.

4. Stochastic Context-Free Grammars

An essential aspect of MTSAGs is the ability to easily generate efficient parsers from grammars. On the basis of the tool we had already written to generate parsers from 1-tape S-attribute grammars, we designed a new tool, MTSAG2C, which automatically generates the C source of a parser from a given MTSAG. The generated parser is able to read tapes (thanks to a lexical analyzer provided by the user), parse tapes, and then output a single derivation tree which satisfies constraints given in the MTSAG.

When using 2-tape MTSAGs for SCFGs, we transfer on the first tape the high-level description of a family of RNAs usually used with SCFGs, and on the second tape the RNA to be folded and aligned. All rules used by the traditional SCFG generating tool to generate a SCFG from its high level description are then written down as a single, fixed, MTSAG. This has the additional benefit of shortening the development cycle (see figure 3b).

We compared the parser generated from a 1-tape version of a 97 nonterminals SCFG (this parser already proved to be quite fast [6]) to the parser generated from the 2-tape version of this SCFG (figure 4(a)).

Tests done on an Alpha 2100-500MP give the results:

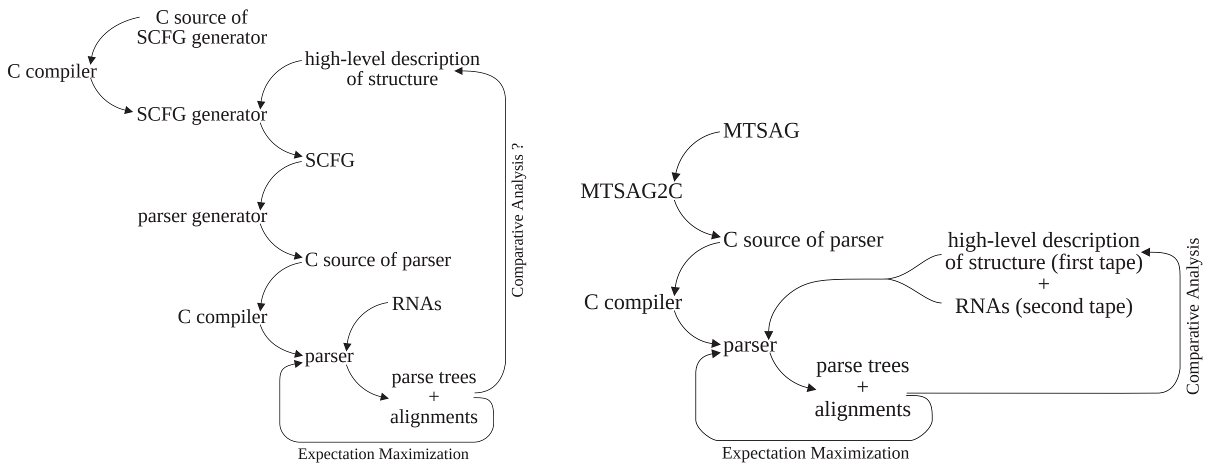

FIGURE 3. (left) 1-tape MTSAG; (right) 2-tape MTSAG. Development cycle of a MTSAG implementation of SCFGs. It has been suggested [4] that a comparative analysis of alignments resulting from parsing may be used to build a new SCFG or a new high-level description of it. With 2-tape MTSAGs, this kind of feedback is as easy to implement as the feedback designed for CMs by Eddy and Durbin [3].

	1-tape	2-tape
83 bases tRNA		
time in seconds:	0.45	0.33
space in Mbytes:	1.8	0.9

With MTSAGs you do not have to generate and compile another parser every time you modify the high-level description of your family (figure 3 (left)). Instead, we may use the following adaptation of the procedure of Eddy and Durbin to learn their CMs from initially unaligned and unfolded RNAs:

(1) Use a MTSAG adaptation of any folding algorithm (Sankoff, Zuker) to get a rough (and even wrong) initial folding. Convert this folding to a suitable first tape (by replacing all single strands by `'*'` for instance);
(2) Use Dirichlet priors to estimate probabilities;
(3) Align and fold all RNAs with a 2-tape MTSAG;
(4) Optimize probabilities from results of the previous step and repeat the previous step until probabilities converge;
(5) Use a comparative analysis algorithm on alignments of step 3 to get a new approximation of the common structural features of all RNAs. Then convert this approximation to a suitable first tape;
(6) Repeat steps 2 to 6 until the first tape converges.

5. Conclusion

We introduce a new way to describe SCGFs in the form of 2-tape MTSAGs and special first tapes. This new way alleviates the need for specialized SCFG building tools and for recompilations of parsers every time the model is changed (only the first tape has to be changed).

MTSAGs may also be applied to most useful sequence analysis methods which were usually expressed with dynamic programming relations (Smith-Waterman alignment model, global alignment, HMMs, simultaneous alignment and folding of RNAs). We believe that MTSAGs should be

```
(((((((..((((********)))).(((((.......)))))********(((((.......)))))))))))).
CCUUCUGUAGCUCAAUUGGUAGAGCAUGUGACUGUAGAGUAUGCGGGUAUCACAGGGUCGCUGGUUCGAUUCCGGCCGGAAGG
```

(a) unaligned 2-tape input string.

```
(((((((..((((********)))).(((((.......)))))*******----------*(((((.......)))))))))))).
CCUUCUGUAGCUCAAUUGGUAGAGCAUGUGACUGUAGAGUAUGC--GG-GUAUCACAGGGUCGCUGGUUCGAUUCCGGCCGGAAGG
```

(b) 2-tape alignment of the previous 2-tape input string.

FIGURE 4. Unaligned and aligned version of a 2-tape input string. The first tape of this 2-tape input-string has a cloverleaf-like structure. This structure has two single strands which may have a variable length around 8 bases. The second tape is the RNA DY6050 extracted from a well known freely available compilation of tRNAs [10].

used instead of dynamic programming relations because these relations hinder the inventiveness of designers of new sequence analysis models.

We also gave a sketch of a method to build stochastic models from unaligned, unfolded RNAs. However, divide and conquer methods may be a prerequisite for long RNAs [2, 4]. We will try to apply MTSAGs to these methods.

Bibliography

[1] Aho (Alfred V.) and Ullman (Jeffrey D.). – *The Theory of Parsing, Translation, and Compiling.* – Prentice-Hall, 1972, vol. 1.

[2] Corpet (Florence) and Michot (Bernard). – RNAlign program: alignment of RNA sequences using both primary and secondary structures. *Computing Applications in the Biosciences*, vol. 10, n° 4, 1994, pp. 389–399.

[3] Eddy (Sean R.) and Durbin (Richard). – RNA sequence analysis using covariance models. *Nucleic Acids Research*, vol. 22, n° 11, 1994, pp. 2079–2088.

[4] Grate (Leslie). – Automatic RNA secondary structure determination with stochastic context-free grammars. In *Third International Conference on Intelligent Systems for Molecular Biology.* pp. 136–144. – AAAI Press, 1995.

[5] Krogh (A.), Brown (M.), Mian (I. S.), Sjölander (K.), and Haussler (D.). – Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, vol. 235, 1994, pp. 1501–1531.

[6] Lefebvre (Fabrice). – An optimized parsing algorithm well suited to RNA folding. In *Third International Conference on Intelligent Systems for Molecular Biology.* pp. 222–230. – AAAI Press, 1995.

[7] Lefebvre (Fabrice). – A grammar-based unification of several alignment and folding algorithms. In *Fourth International Conference on Intelligent Systems for Molecular Biology.* pp. 143–154. – AAAI Press, 1996.

[8] Sakakibara (Yasubumi), Brown (Michael), Hughey (Richard), Mian (I. Saira), Sjölander (Kimmen), Underwood (Rebecca C.), and Haussler (David). – Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, vol. 22, 1994, pp. 5112–5120.

[9] Searls (David B.) and Murphy (Kevin P.). – Automata – theoretic models of mutation and alignment. In *Third International Conference on Intelligent Systems for Molecular Biology.* pp. 341–349. – AAAI Press, 1995.

[10] Steinberg (S.), Misch (A.), and Sprinzl (M.). – Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, vol. 21, 1993, pp. 3011–3015.

Part 2

Symbolic Computation

Linear recurrences, linear differential equations, and fast computation

Bruno Salvy
INRIA Rocquencourt

November 13, 1995

[summary by Philippe Dumas]

Linear recurrences and linear differential equations with polynomial coefficients provide a finite representation of special functions or special sequences. Many algorithms are at our disposal; some give a way to automate the computation of recurrences or differential equations; some provide solutions to recurrences or differential equations; and some give the asymptotic behaviour of these solutions, directly from the recurrence or differential equation. All of this provides a method to efficiently compute special functions and special sequences.

1. Classical algorithms concerning formal power series

In the sequel, we use the ring $\mathbb{A}[[x]]$ of formal power series

$$F(x) = \sum_{n=0}^{+\infty} f_n x^n$$

with coefficients f_n in a commutative ring $\mathbb{A}$; this ring is assumed to contain the field $\mathbb{Q}$ of rational numbers, even though it is possible to consider a more general situation. Practically, one deals with truncated series

$$F(x) = \sum_{n=0}^{N} f_n x^n + O(x^{N+1}),$$

that is to say essentially polynomials. It must be noted that there exist lazy algorithms to deal with truncated series of arbitrary order, but their cost is generally excessive. We indicate how to deal with basic operations [7, Chap. 4].

Product of polynomials. The naive method to obtain the product of two polynomials of degree N has complexity $O(N^2)$ arithmetic operations. A better way is Karatsuba's algorithm, which has complexity $O(N^{\log_2 3}) = O(N^{1.59})$. The idea behind the algorithm resides in writing

$$P(x) = P_0(x) + x^k P_1(x), \qquad Q(x) = Q_0(x) + x^k Q_1(x),$$

$$P(x)Q(x) = P_0(x)Q_0(x) + R(x)x^k + P_1(x)Q_1(x)x^{2k},$$

$$R(x) = (P_0(x) + P_1(x))(Q_0(x) + Q_1(x)) - (P_0(x)Q_0(x) + P_1(x)Q_1(x)),$$

with $k \simeq N/2$; this formula needs only three multiplications of polynomials of degree less than k instead of four multiplications, and this leads to an efficient recursive computation.

From a practical standpoint, Karatsuba's method becomes efficient in Maple for an N greater than about a hundred. The fast Fourier transform algorithm needs a much larger value of N to be useful.

Composition. Here the goal is the computation of the first N coefficients of the series $F(G(x))$, where $g_0 = 0$. The naive method leads to a computation with $O(N)$ series multiplications. Brent and Kung's algorithm [2] has a better behaviour. It consists of three steps; first write $F(x)$ as

$$F(x) = F_0(x) + F_1(x)x^k + F_2(x)x^{2k} \cdots + F_{k-1}(x)x^{k(k-1)},$$

where $F_0(x)$, $F_1(x)$, …, $F_{k-1}(x)$ are the series obtained by factoring out the powers of x^k, where $k = \lceil (N+1)^{1/2} \rceil$; next compute the powers $G^i(x)$ for i from 2 to $k-1$, and the series $F_i(G(x))$; finally, compute $T(x) = G^k(x)$ and $F(G(x))$ using a Horner scheme.

The algorithm uses $3k$ series multiplications and $O(N)$ coefficient multiplications, hence it has cost $O(N^{1/2})$ if the unit cost is series multiplication. Via Karatsuba's algorithm, this gives a cost of $O(N^{2.09})$ expressed in terms of coefficient multiplications, while the naive method has cost $O(N^3)$.

Powering and simple functions. Powering and simple functions are a particular case of composition, but in this case it is possible to be more efficient. We show the idea for the case of powering. If $H(x) = F^\alpha(x)$, then $H(x)$ satisfies the equation

$$H'(x)F(x) = \alpha F'(x)H(x),$$

therefore the coefficients of $H(x)$ are provided by the following recurrence

$$\sum_{k=0}^{n} k h_k f_{n-k} = \alpha \sum_{k=0}^{n} (n-k) h_k f_{n-k}.$$

This makes it possible to compute the first N coefficients at a cost of $O(N^2)$ coefficient multiplications, instead of $O(N^{2.09})$.

Newton iteration. An ever better way to compute elementary functions is Newton's method. If we search for a series $y(x)$ such that $\Phi(x, y(x)) = 0$, we use the recurrence

$$y_{k+1}(x) = y_k(x) - \frac{\Phi(x, y_k(x))}{\partial\Phi/\partial y(x, y_k(x))} \mod x^{2k+2}.$$

We start from $y_0(x) = 0$ and the formula is iterated until $2k + 2 > N$. The number of correct coefficients is doubled at each step. For example, the reciprocal $y(x) = 1/F(x)$ satisfies the equation $\Phi(x, y) = 1/y - F(x) = 0$ and the recursion is $y_{k+1} = 2y_k - F(x)y_k^2$. In the same way one can compute the logarithm $\ln(F(x))$, the exponential $\exp(F(x))$, and solutions of simple differential equations. In all these cases the complexity of the computation is the complexity of the multiplication, that is $O(N^{1.59})$.

For the reversion of series the same method can be used. Given $F(x)$ with $f_0 = 0$, $f_1 \neq 0$, one looks for a series $y(x)$ such that $F(y(x)) = x$. This is carried out by Newton's method applied to the equation $F(y) - x = 0$; hence the recurrence is

$$y_{k+1}(x) = y_k(x) - \frac{F(y_k(x)) - x}{F'(y_k(x))} \mod x^{2k+2}.$$

The cost is of the same order as the composition cost, because of the terms $F(y_k(x))$ and $F'(y_k(x))$.

Linear differential equations. Assume that the power series $F(x)$ satisfies a linear differential equation

$$a_0(x)y^{(k)} + a_1(x)y^{(k-1)} + \cdots + a_k(x)y = 0,$$

whose coefficients are polynomials. If 0 is an ordinary point, this differential equation translates into a linear recurrence for the coefficients of $F(x)$. This leads to an algorithm whose cost is $O(N)$, while the preceding ones use at best $O(N^{1.59})$ basic operations.

Obviously, the complexity $O(N)$ is optimal, therefore for large N there is great interest in finding a linear differential equation with $F(x)$ as a solution, if possible. In the sequel, we will focus our attention on such power series.

2. Univariate holonomic series

A power series is said to be holonomic if it is a solution of a linear differential equation with polynomial coefficients. In the same manner, a sequence is said to be holonomic if it is a solution of a linear recurrence with polynomial coefficients.

It is easy to see that rational series, $\exp(x)$, $\sin x$, $\cos x$, $\log(1+x)$, and the Bessel functions $J_\nu(x)$ are all holonomic. Rational, factorial, Fibonacci, and hypergeometric sequences are all holonomic sequences. Recall that a sequence is hypergeometric if the sequence of quotients of consecutive terms is a rational sequence.

Both definitions are related by the following property: a sequence is holonomic if and only if its generating series is holonomic. The proof is easy and uses the simple but basic correspondences which may be summarized as follows,

$$\begin{array}{rcl} F(x) & \longleftrightarrow & f_n, \\ x^k F(x) & \longleftrightarrow & f_{n-k}, \\ xF'(x) & \longleftrightarrow & nf_n. \end{array}$$

Closure properties. The set of holonomic series is closed with respect to sum, Cauchy product, Hadamard product, Borel transform, and Laplace transform [10]. We give a sketch of the proof for the Cauchy product. If $F(x)$ satisfies a differential equation of order s and $G(x)$ satisfies a differential equation of order t, we formally compute the derivatives of $H(x) = F(x)G(x)$ and, using the equations satisfied by $F(x)$ and $G(x)$, we express them as linear combinations of the products $F^{(i)}(x)G^{(j)}(x)$ where the indices i and j vary from 0 to $s-1$ and from 0 to $t-1$ respectively. The space of such combinations has a finite dimension, hence the derivatives of $H(x)$ satisfy a dependance relation, that is a linear differential equation.

There is a similar result for holonomic sequences: the sum, product, and convolution of two holonomic sequences are holonomic; the indefinite summation of a holonomic sequence is holonomic. Both types of closure properties are interrelated, and the proofs use whichever is easier.

Identity proving. An application of holonomy, widely exemplified by D. Zeilberger, is identity proving [12]. The idea is the following: to prove $F(x) = G(x)$, build the equation satisfied by $F(x) - G(x)$, and compute sufficiently many initial conditions to ensure $F(x) - G(x) = 0$.

Here is a simple example. Suppose we want to prove the identity

$$\sqrt{x}J_{1/2}(x) = \sqrt{\frac{2}{\pi}}\sin x,$$

where $J_{1/2}(x)$ is a Bessel function of index 1/2. This function satisfies a second order differential equation, while the square root satisfies a first order equation; hence the product is a solution of an equation of order not greater than 2. On the other hand, sine satisfies a second order equation; therefore the difference of the two sides of the formula satisfies an equation of order not greater than 4. It suffices to verify that the power series of the difference is $O(x^4)$, using the differential equations and the initial conditions defining the components. The alert reader may think we were lucky, because $\sqrt{x}J_{1/2}(x)$ has a power series expansion at 0, while $J_{1/2}(x)$ has not. But, if this had not been the case, we would have used use another point than 0.

Algebraic functions. Algebraic functions are holonomic. Comtet [4] gave an algorithm to compute the differential equation satisfied by a function $F(x)$ solution of $P(x, y) = 0$, where P is an irreducible polynomial. The idea is to find a Bezout relation $UP + VP_y = 1$ by the extended Euclidean algorithm and use $P_x + P_y F' = 0$ to express the successive derivatives of $F(x)$ as polynomials in $F(x)$ of degree less than $d = \deg_y P$. The family of powers 1, $F(x)$, $\ldots$, $F^{d-1}(x)$ is a basis of the space generated by the derivatives of $F(x)$, and there is a dependance relation between $F(x)$, $F'(x)$, $\ldots$, $F^{(d)}(x)$.

Algebraic substitution. If $F(x)$ is holonomic and $G(x)$ is algebraic, then $F(G(x))$ is holonomic by the same kind of technique as above. An immediate application of this result is the following: if f_n is a holonomic sequence, then its Euler transform

$$h_n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} f_k$$

is holonomic too. This is obvious because the two generating functions are connected by $H(x) = F(-x/(1-x))/(1-x)$.

3. Search for solutions

Recurrence relations and differential equations almost never have explicit solutions, but if an explicit solution exists it might be important to recognize it, and find the solution. Above all an explicit solution gives a global information about the equation.

Rational solutions to recurrences. Abramov [1] gives a method to obtain the rational solutions $u_n = P(n)/Q(n)$ of a recurrence

$$a_0(n)u_{n+k} + \cdots + a_k(n)u_n = b(n),$$

where a_0, $\ldots$, a_k, and b are polynomials. The principle which guides the algorithm is: the zeros of the coefficients must match the poles of u_n and its shifts $u_{n+\ell}$. As a consequence, Q must be a multiple of $\gcd(a_0(n-k), \ldots, a_k(n))$ if the roots of Q do not differ by an integer. The last condition is not necessarily fulfilled; to avoid this problem one considers a recurrence satisfied by the sequence $v_n = u_{nh}$, where h is the maximal difference between two roots of Q. It must be noted that the number h is not greater than the maximal difference between the roots of $a_k(n)$ and $a_0(n-k)$.

Indefinite hypergeometric summation. The indefinite sum [6] of f_k is equivalent to finding a closed formula for $F_n = \sum^n f_k$ where f_k is a given sequence. This relation means

$$F_n - F_{n-1} = f_n$$

for all n. If f_n is assumed to be hypergeometric, and we look for a hypergeometric F_n, the relation $1 - F_{n-1}/F_n = f_n/F_n$ shows that the sequence $u_n = F_n/f_n$ must be rational. Hence we are led to search for a rational solution of the equation

$$u_n - \frac{f_{n-1}}{f_n} u_{n-1} = 1.$$

Hypergeometric solutions to recurrences. Petkovšek's algorithm [8] provides the hypergeometric solutions of a linear recurrence

$$a_0(n)u_{n+k} + \cdots + a_k(n)u_n = 0,$$

where $a_0, \ldots, a_k$, and b are polynomials. Writing $u_{n+1}/u_n = P(n)/Q(n)$ and substituting leads to a non-linear equation, which is not tractable. There exists a decomposition

$$\frac{u_{n+1}}{u_n} = \frac{P(n)}{Q(n)} \frac{A(n+1)}{A(n)}$$

in which all pairs $(A(n), P(n))$, $(A(n), P(n))$, $(P(n), Q(n))$, $(P(n), Q(n+1))$, $\ldots$, $(P(n), Q(n+k))$ are relatively prime. With this decomposition a substitution gives

$$a_0(n)A(n+k)P(n+k)\cdots P(n) + a_1(n)A(n+k-1)P(n+k-1)\cdots P(n)Q(n+k) \\ + \cdots + a_k(n)A(n)Q(n+k)\cdots Q(n) = 0.$$

This equation is still non-linear, but it shows that $P(n)$ divides $a_k(n)$, and $Q(n+k)$ divides $a_0(n)$. Finally it suffices to test all pairs of factors of $a_0(n-k)$ and $a_k(n)$.

Note that this algorithm is a powerful tool; it is equivalent to finding factors of order 1 on the right of the recurrence.

Symbolic solutions to differential equations. Searching for generalized hypergeometric solutions is a first approach to a linear differential equation: the recurrence satisfied by the coefficients of the series is computed; the hypergeometric solutions to this recurrence are found; finally the result is translated from sequences to generating functions.

The more general class of Liouvillian functions may be used. Liouvillian functions are obtained from rational functions with rational coefficients by repeated use of the four elementary operations, taking exponentials and logarithms, integration, and algebraic extensions. Singer gives a purely theoretic algorithm to obtain Liouvillian solutions of linear differential equations of arbitrary order. Kovacic's algorithm for equations of order 2 is partially implemented in most computer algebra systems. The theory behind all these algorithms is differential Galois theory. It is difficult to use, because for each order it is necessary to classify the Galois groups which come into play [11].

4. Asymptotic analysis

Even when no explicit solution of a differential equation is known, it is possible to perform an asymptotic analysis. The theory of linear differential equations prescribes the asymptotic behaviour of a solution near a singularity and this asymptotic behaviour is strongly related to the asymptotic behaviour of the Taylor coefficients of the solution.

Singular points. The solutions of a linear differential equation

$$a_0(x)y^{(k)}(x) + \cdots + a_k(x)y(x) = 0$$

may only have singularities at the roots of the dominant coefficient $a_0(x)$, and possibly at infinity. In addition all formal solutions to the equation are known. A logarithmic sum is a formal series

$$\lambda(z) = z^\alpha \sum_{j=1}^{J} \sum_{i\geq 0} c_{ij} z^i \log^j z,$$

and a formal solution in the neighbourhood of the root a of $a_0(x)$ is a finite combination of logarithmic sums

$$y(x) = \sum_{k=0}^{K} \lambda_k(z) \exp(P_k(z)), \qquad z = \frac{1}{(1-x/a)^{1/r}},$$

which formally satisfies the differential equation. All quantities involved in these formulae can be explicitly computed. In the case where the point a is a regular singular point, that is to say $a_\ell(x) = (x-a)^{k-\ell}A(x)$ for $\ell = 0, \ldots, k$, and $A_\ell(x)$ is analytic in the neighbourhood of a, the formal solutions are logarithmic sums and locally define actual solutions, with a possible ramification point at a. Conversely, in the case of an irregular singular point the formal solutions are generally divergent series, but provide asymptotic expansions for actual solutions in a sector with origin a.

The preceding classification demonstrates that the composition of two holonomic functions is not necessarily holonomic. For instance $1/\sin x$, which is the composition of the two holonomic functions $\sin x$ and $1/x$, is not holonomic because it has an infinite number of singularities. The sequence of Bell numbers is not holonomic because its exponential generating function $\exp(e^x - 1)$ does not have the right form, given by the formula above (after changing x into $1/x$).

Singularity analysis. The smallest singularity ρ of a function analytic in a neighbourhood of zero prescribes the behaviour of the Taylor coefficients of the function. This rough correspondence may be strongly refined [5]; indeed an asymptotic expansion in some sufficiently large neighborhood of the singularity a of smallest modulus

$$f(x) \underset{x\to a}{=} c_0(1-x/a)^{\alpha_0} \log^{\beta_0} \frac{1}{1-x/a} + c_1(1-x/a)^{\alpha_1} \log^{\beta_1} \frac{1}{1-x/a} + \cdots$$

translates into an asymptotic expansion for the coefficient of the Taylor expansion of $f(x)$ at 0

$$f_n \underset{n\to\infty}{=} \rho^{-n} \frac{n^{-\alpha_0-1}}{\Gamma(-\alpha_0)} \log^{\beta_0} n \left(c_0 + \frac{d_1}{\log n} + \cdots \right).$$

This result leads to the following idea: to study the asymptotic behaviour of a sequence which satisfies a linear recurrence it suffices to translate the recurrence into a differential equation for the generating function; next a singularity analysis of this function gives the asymptotic behaviour of the sequence. This simple method presents a difficulty. The function is determined as a solution of a differential equation and some initial conditions, which are specified at the point 0. The study of the differential equation provides a basis of formal solutions near the smallest singularity, but there is no direct way to express the generating function with respect to this basis. Obviously if a closed form of the function is available it is possible to realize the connection between the data at 0 and the behaviour at the smallest singularity; but in that case more direct procedures may be used. Generally, it is necessary to use analytic continuation and a resummation method [9]. Note that such a method needs to know about the singularities of the Borel transform of the function; and we have seen that it is possible to compute the differential equation satisfied by the Borel transform of a holonomic function.

5. Multivariate holonomy

The machinery of holonomic sequences or functions is so powerful that it is tempting to generalize holonomy for sequences or functions with more than one variable.

Weyl algebra. The Weyl algebra $A_N(\mathbb{K})$ is an algebra of linear operators which is defined over the space of polynomials $\mathbb{K}[\mathbf{x}] = \mathbb{K}[x_1, \ldots, x_N]$. These operators are the partial derivatives ∂_j, the multiplications by the variable x_i's, and all their combinations. The generators $\partial_1, \ldots, \partial_N, x_1, \ldots, x_N$ satisfy the following commutation rules:

$$\partial_i \partial_j = \partial_j \partial_i, \qquad x_i x_j = x_j x_i$$

$$\partial_i x_j = x_j \partial_i \quad \text{for } i \neq j, \qquad \partial_i x_i = x_i \partial_i + 1.$$

Then, an element f of a module over the Weyl algebra is D-finite if the submodule spanned by f and all its derivatives $\boldsymbol{\partial}^\alpha f$ has a finite dimension over the field of rational functions $\mathbb{K}(\mathbf{x})$. An equivalent definition is obtained as follows: for f from an $A_N(\mathbb{K})$-module, consider the set of all equations $P(\mathbf{x}, \boldsymbol{\partial})f = 0$ satisfied by f; the polynomials $P(\mathbf{x}, \boldsymbol{\partial})$ are elements of the left ideal $\operatorname{Ann}(f)$ in the Weyl algebra; then f is D-finite if the quotient $A_N(\mathbb{K})/\operatorname{Ann}(f)$ of the Weyl algebra by the annihilator ideal $\operatorname{Ann}(f)$ has a finite dimension over $\mathbb{K}(\mathbf{x})$ as a vector space.

A more effective definition uses the idea of a rectangular system. A set of N polynomials $P_k(\mathbf{x}, \boldsymbol{\partial})$ from the Weyl algebra is a rectangular system if each polynomial involves only one partial derivative ∂_i, and each partial derivative appears in exactly one of these polynomials $P_k(\mathbf{x}, \boldsymbol{\partial})$. One proves that f is D-finite if and only if there exists a rectangular system contained in the annihilator ideal $\operatorname{Ann}(f)$. As a consequence a D-finite element f satisfies a special set of equations of the form

$$P_1(\mathbf{x}, \partial_1)f = 0, \qquad P_2(\mathbf{x}, \partial_2)f = 0, \qquad \ldots, \qquad P_N(\mathbf{x}, \partial_N)f = 0.$$

In addition, Bernstein worked out the concept of multivariate holonomy. The Weyl algebra is naturally graded by the degree: the degree of the monomial $\mathbf{x}^\alpha \boldsymbol{\partial}^\beta$ is $|\alpha| + |\beta|$, and the component F_d of the natural filtration is composed of the polynomials of degree not greater than d. For f from a module over the Weyl algebra, this induces a filtration of the submodule $A_N(\mathbb{K})f$; the component Γ_d is merely $F_d f$. It turns out that the dimension of Γ_d over $\mathbb{K}$ is expressed as a polynomial in d for all sufficiently large d. The degree of this polynomial is the Bernstein dimension of the module $A_N(\mathbb{K})f$. Moreover it is shown that the Bernstein dimension of $A_N(\mathbb{K})f$ is greater or equal to N. Now, f is *holonomic* if the Bernstein dimension of $A_N(\mathbb{K})f$ is exactly N.

Kashiwara's theorem proves that D-finiteness and holonomy are the same concept. But each one has its own merits. The D-finiteness property makes it easy to show that sums and products of holonomic functions are holonomic too. On the other hand, definite integration with respect to one of the x_i's preserves holonomy, and this is more easily shown using the definition of holonomy.

The link between sequences and generating functions is not as nice in the multivariate case as in the univariate case. A sequence u_ν, where the index ν is an N-tuple $(n_1, \ldots, n_N)$ is P-finite if the sequence u_ν and all its shifts $u_{\nu+\tau}$ span a finite dimensional space over $\mathbb{K}(\tau_1, \ldots, \tau_N)$. An equivalent formulation of the P-finiteness can be written as follows: there exists a rectangular system

$$P_1(\nu, S_1)u = 0, \qquad P_2(\nu, S_2)u = 0, \qquad \ldots, \qquad P_N(\nu, S_N)u = 0,$$

where S_i is the shift operator defined by $S_i u_\nu = u_{n_1, \ldots, n_i+1, \ldots, n_N}$. One proves that a sequence is P-finite if its multivariate generating function is D-finite. The reciprocal assertion is false.

The study of P-finite sequences shows it is interesting to consider a more general concept than Weyl algebras. This leads to Ore algebras, which are defined as polynomial algebras with some commutation rules for the variables [3]. For instance, the finite difference calculus in one variable is formalized by the algebra $\mathbb{K}\langle n, \Delta\rangle$ with $\Delta n = (n+1)\Delta + 1$.

Creative telescoping. We search for a recurrence relation for the definite sum $U_n = \sum_k u_{n,k}$, where the double sequence $u_{n,k}$ is P-finite. The idea is to find an equation $P(n, S_n, \Delta_k)u = 0$, where the variable k does not occur, S_n is the shift operator with respect to n, and Δ_k is the difference operator with respect to k; then, U satisfies $P(n, S_n, \Delta_k)U = 0$. Contrary to the case of holonomic functions such an equation does not exist a priori; but if it exists, it is possible to find it by a Gröbner basis technique. As an example we want to rederive the Franel relation on the sum

$$U_n = \sum_{k=0}^{n} \binom{n}{k}^3.$$

First we give a rectangular system for the double sequence $u_{n,k} = \binom{n}{k}^3$,

$$[(n-k+1)^3 S_n - (n+1)^3]u = 0, \qquad [(k+1)^3 S_k - (n-k)^3]u = 0.$$

Here the analogue to the Bernstein dimension is 2, hence elimination provides a relation $P(n, S_k, S_n)u = 0$. Next the summation with respect to k, and the substitution $S_k = 1$, or equivalently $\Delta_k = 0$, give the desired formula:

$$\begin{aligned}\big[(n+3)^3(3n+4)S_n^3 - (18n^3 + 114n^2 + 232n + 148)S_n^2 \\ -(3n+5)(15n^2+55n+48)S_n - 8(n+1)^2(3n+7)\big]\, U = 0.\end{aligned}$$

Bibliography

[1] Abramov (S. A.). – Rational solutions of linear differential and difference equations with polynomial coefficients. *USSR Computational Mathematics and Mathematical Physics*, vol. 29, n° 11, 1989, pp. 1611–1620. – Translation of the Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki.

[2] Brent (R. P.) and Kung (H. T.). – Fast algorithms for manipulating formal power series. *Journal of the ACM*, vol. 25, 1978, pp. 581–595.

[3] Chyzak (Frédéric) and Salvy (Bruno). – *Non-commutative Elimination in Ore Algebras Proves Multivariate Holonomic Identities.* – Research Report n° 2799, Institut National de Recherche en Informatique et en Automatique, February 1996.

[4] Comtet (L.). – Calcul pratique des coefficients de Taylor d'une fonction algébrique. *L'Enseignement Mathématique*, vol. 10, 1964, pp. 267–270.

[5] Flajolet (Philippe) and Odlyzko (Andrew M.). – Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, vol. 3, n° 2, 1990, pp. 216–240.

[6] Gosper (R. William). – Decision procedure for indefinite hypergeometric summation. *Proceedings of the National Academy of Sciences USA*, vol. 75, n° 1, January 1978, pp. 40–42.

[7] Knuth (Donald E.). – *The Art of Computer Programming.* – Addison-Wesley, 1981, 2nd edition, vol. 2: Seminumerical Algorithms.

[8] Petkovšek (Marko). – Hypergeometric solutions of linear recurrences with polynomial coefficients. *Journal of Symbolic Computation*, vol. 14, 1992, pp. 243–264.

[9] Ramis (Jean-Pierre). – *Séries divergentes et théories asymptotiques.* – Société Mathématique de France, 1993, *Panoramas et Synthèses*, vol. 121.

[10] Stanley (R. P.). – Differentiably finite power series. *European Journal of Combinatorics*, vol. 1, n° 2, 1980, pp. 175–188.

[11] Tournier (Évelyne) (editor). – *Computer Algebra and Differential Equations.* – Academic Press, 1990. Proceedings of CADE 89.

[12] Zeilberger (Doron). – A holonomic systems approach to special functions identities. *Journal of Computational and Applied Mathematics*, vol. 32, n° 3, 1990, pp. 321–368.

Creative Telescoping and Applications

Frédéric Chyzak
INRIA Rocquencourt

January 15, 1996

[summary by Bruno Salvy]

Abstract

Creative telescoping is a method to compute definite sums and integrals. Numerous examples are given, together with an introduction to algorithmic techniques based on Gröbner bases of linear operators.

Creative telescoping applies to solutions of systems of linear recurrences and linear differential equations. It yields a linear recurrence or differential equation satisfied by the definite sum or integral of the solutions. It can be used to "compute" generating functions, to extract their coefficients, and to prove identities.

1. Examples

A typical example is the sum $S_n = \sum_{k=0}^{n} \binom{n}{k}$. One starts with a system of equations defining the summand:

$$Au := (n+1-k)u_{n+1,k} - (n+1)u_{n,k} = 0, \qquad Bu := (k+1)u_{n,k+1} - (n-k)u_{n,k} = 0.$$

The aim is to derive a recurrence satisfied by S_n from these equations. This is done by first finding an equation satisfied by $u_{n,k}$ where k does not appear in the coefficients. Such an equation is given by Pascal's triangle rule $u_{n+1,k+1} = u_{n,k+1} + u_{n,k}$ which can be deduced from the above equations as $(S_k+1)A + S_nB$, where S_k (resp. S_n) denotes the shift with respect to k (resp. n). This equation is then rewritten in a form suitable for summation with respect to k:

$$(u_{n+1,k+1} - u_{n+1,k}) - (u_{n,k+1} - u_{n,k}) + u_{n+1,k} - 2u_{n,k} = 0.$$

Since the binomial coefficient $\binom{n}{k}$ is 0 when $k < 0$ or $k > n$, summing over k simply yields the desired result $S_{n+1} - 2S_n = 0$ (this is where telescoping takes place). Using the initial condition $S_0 = 1$, any solver of recurrence equations would then produce $S_n = 2^n$.

A similar example is provided by $U_n = \sum_{k=0}^{n} \binom{n}{k}^2$. The system of equations is a simple modification of the former one. Finding an equation which does not involve k in the coefficients is slightly harder. One finds

$$(n+1)u_{n+2,k+2} - (2n+3)u_{n+1,k+2} + (n+1)u_{n,k+2} - (2n+3)u_{n+1,k+1} - 2(n+1)u_{n,k+1} + u_{n,k} = 0.$$

Again, this is rewritten in a form where telescoping will take place by repeatedly expressing $v_{k+1} = (v_{k+1} - v_k) + v_k$. Summing then yields

$$(n+1)U_{n+1} - 2(2n+1)U_n = 0.$$

Again, with the initial condition $U_0 = 1$, it is easy to conclude that $U_n = \binom{2n}{n}$.

Exactly the same computation applies to definite integrals. For instance, to compute $F(x) = \int_{-\infty}^{+\infty} \exp(-xy^2)\,dy$, one starts from a system satisfied by the integrand

$$D_x + y^2 = 0, \qquad D_y + 2xy = 0,$$

where D_x denotes differentiation with respect to x (and similarly for D_y). Then we look for an equation satisfied by f without y in the coefficients. It is not difficult to find that such an equation is $(D_y^2 + 4x^2 D_x + 2x)f = 0$. Since for any value of x, $\exp(-xy^2)$ and its derivatives with respect to y tend to 0 at $\pm\infty$, integrating this equation over y yields $4x^2 F'(x) + 2xF(x) = 0$. The initial condition $F(1) = \sqrt{\pi}$ leads to $F(x) = \sqrt{\pi/x}$.

2. Ore algebras

A very natural framework to describe creative telescoping is provided by a special case of skew polynomial rings called Ore algebras. These are algebras of linear operators which generalize the difference and differential operators.

Definition 1. Let $\mathbb{K}$ be a (possibly skew) field. Let $\partial_1, \ldots, \partial_r$ be defined by the following commutation rules with all the elements P in $A = \mathbb{K}(x_1, \ldots, x_p)[y_1, \ldots, y_q]$:

$$\partial_i P = \sigma_i(P)\partial_i + \delta_i(P),$$

where σ_i is a ring endomorphism of A and δ_i is an additive endomorphism which satisfies the following Leibniz rule:

$$\delta_i(ab) = \sigma_i(a)\delta_i(b) + \delta_i(a)b, \qquad \forall a, b \in A.$$

Then $\mathbb{K}(x_1, \ldots, x_p)[y_1, \ldots, y_q]\langle \partial_1, \ldots, \partial_r \rangle$ is called an *Ore algebra.*

Examples of Ore operators are given in Table 1. These can be combined in an algebra where each operator acts on a different variable. For instance, the Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ can be described in $\mathbb{Q}(\alpha, \beta, x, n)\langle S_n, D_x \rangle$ by a linear differential equation and a linear recurrence.

More complicated examples arise when one of the ∂_i has a special commutation rule with several of the commutative variables. For instance, in $\mathbb{Q}(n, q, q^n)\langle S_n^{(q)} \rangle$, the q-shift operator satisfies the following commutation rule:

$$S_n^{(q)} n^i (q^n)^j = q^j (n+1)^i (q^n)^j S_n^{(q)}.$$

In this framework, creative telescoping becomes an elimination process. Given a set of operators generating an ideal of operators which vanish on the function we want to sum or integrate, the main

Operator ∂	$\sigma(a)$	$\delta(a)$	Commutation	Action of ∂
Differentiation	$a(x)$	$a'(x)$	$\partial x = x\partial + 1$	$f(x) \mapsto f'(x)$
Shift	$a(x+1)$	0	$\partial x = (x+1)\partial$	$f(x) \mapsto f(x+1)$
Difference	$a(x+1)$	$a(x+1) - a(x)$	$\partial x = (x+1)\partial + 1$	$f(x) \mapsto f(x+1) - f(x)$
q-Dilation	$a(qx)$	0	$\partial x = qx\partial$	$f(x) \mapsto f(qx)$
q-Difference	$a(qx)$	$a(qx) - a(x)$	$\partial x = qx\partial + (q-1)x$	$f(x) \mapsto f(qx) - f(x)$
q-Differentiation	$a(qx)$	$\frac{a(qx)-a(x)}{(q-1)x}$	$\partial x = qx\partial + 1$	$f(x) \mapsto \frac{f(qx)-f(x)}{(q-1)x}$
Eulerian operator	$a(x)$	$xa(x)$	$\partial x = x\partial + x$	$f(x) \mapsto xf'(x)$
e^t-Differentiation	$a(x)$	$xa(x)$	$\partial x = x\partial + x$	$f(t) \mapsto f'(t) \quad (x = e^t)$
Mahlerian operator	$a(x^p)$	0	$\partial x = x^p\partial$	$f(x) \mapsto f(x^p) \quad (p \geq 2)$

Table 1. Ore operators

step of creative telescoping asks for an operator in the ideal that does not involve the variable with respect to which we want to integrate or sum. It turns out that under mild conditions on the σ_i's and δ_i's, Ore algebras are Noetherian and an extension of Buchberger's algorithm can be used to compute Gröbner bases. The elimination necessary for creative telescoping can thus be performed automatically provided we have a good description of the ideal.

Given an ideal $\mathcal{I}$ and an operator ∂ of the Ore algebra $\mathcal{O} = \mathbb{K}[x_1, \dots, x_n]\langle\partial_1, \dots, \partial_k\rangle$, let $\mathbf{x}$ be those elements of $\{x_1, \dots, x_n\}$ which commute with ∂. The first step of creative telescoping is therefore to find a basis of the ideal $\mathcal{J} = \mathcal{I} \cap \mathbb{K}[\mathbf{x}]\langle\partial_1, \dots, \partial_k\rangle$ by elimination. The elements of $\mathcal{J}$ can be written

$$\partial A + B, \tag{1}$$

where B does not involve ∂. Since this is an element of $\mathcal{I}$, it cancels whatever function f the ideal $\mathcal{I}$ was cancelling. Now assuming Af to be 0 on the "borders" of the domain, multiplying by ∂^{-1} shows that B is the result we are after (see [2] for a more rigourous description and the application to indefinite operations).

3. More examples

The computation of Gröbner bases of Ore algebras has been implemented by F. Chyzak in his *Mgfun* Maple package available at the URL `http://www-rocq.inria.fr/algo/`. We now illustrate some uses of this package.

3.1. Generating Function of the Jacobi Polynomials. The idea is first to define operators annihilating $P_n^{(\alpha,\beta)}(x)y^n$ and then to compute the sum over n by creative telescoping.

We start with two operators in D_x and S_n annihilating $P_n^{(\alpha,\beta)}(x)$ (omitted here for space reasons):

```
G:=[...,...]:
```

We then load the package and define the Ore algebra in which this computation will take place.

```
with(Mgfun):
A:=orealg(diff=[Dx,x],diff=[Dy,y],shift=[Sn,n],comm=[alpha,beta]):
```

This expresses that there are two variables with a differentiation-like commutation rule, one variable with a shift-like commutation rule and two commutative variables. From the operators annihilating $P_n^{(\alpha,\beta)}(x)$, it is easy to derive operators annihilating $P_n^{(\alpha,\beta)}(x)y^n$:

```
G:=map(primpart,map(numer,[op(subs(Sn=Sn/y,G)),y*Dy-n]),[Sn,Dx,Dy]):
```

Then we are ready for elimination: we create an appropriate term order and then compute a Gröbner basis with respect to it:

```
T:=termorder(A,lexdeg=[[n],[Sn,Dx,Dy]]):
GB:=gbasis(G,T,ratpoly(rational,[x,y,alpha,beta])):
```

We finally select those operators in this basis which do not involve n, and sum over n, which is equivalent to taking the remainder of the division by Δ_n:

```
subs(Sn=1,remove(has,GB,n)):
```

The computation has taken 17 seconds (on a Dec Alpha). After a further fast Gröbner basis computation, the result is reduced to a system of two equations, a large one of order 2 in D_y and another one linear in D_x and D_y. It is then possible to interact with a differential equation solver and, using the initial conditions, obtain the closed-form formula

$$F(x,y) = \frac{1}{R(1-y+R)^a(1+y+R)^b}, \qquad R = \sqrt{1-2xy+y^2}.$$

3.2. q-Dixon identity. The aim is to show that

$$\sum_k (-1)^k q^{\frac{k(3k+1)}{2}} \binom{a+b}{a+k}_q \binom{b+c}{b+k}_q \binom{a+c}{c+k}_q = \binom{a+b+c}{a,b,c}_q. \tag{2}$$

The algebra is $\mathbb{Q}(q, q^a, q^b, q^c, q^k)\langle S_a, S_b, S_c, S_k\rangle$ which has only q-shift operators:

```
A:=orealg(comm=[q],qshift=[Sa,qa,q],qshift=[Sb,qb,q],
          qshift=[Sc,qc,q],qshift=[Sk,qk,q]):
```

The operators defining the summand are all of order 1 and can be obtained in *Mgfun* by

```
G:=subs([q^a=qa,q^b=qb,q^c=qc,q^k=qk], hypergeomtoholon((-1)^k*q^(k*(3*k+1)/2)
      *qbinomial(a+b,a+k)*qbinomial(a+c,c+k)*qbinomial(b+c,b+k),A)):
```

Then we eliminate q^k and proceed with the telescoping:

```
T:=termorder(A,lexdeg=[[qk],[Sa,Sb,Sc,Sk]]):
GB:=gbasis(G,T,ratpoly(rational,[q,a,b,c,qa,qb,qc])):
CT:=subs(Sk=1,remove(has,GB,[k,qk])):
```

This yields a system of operators symmetrical in a, b, c. Using one more Gröbner basis computation, one obtains an operator involving only S_a. By symmetry similar operators in S_b and S_c can be found. Then checking that the right-hand side of (2) satisfies these equations and that sufficiently many initial condition coincide proves the identity. It is also possible to use Abramov and Petkovšek's q-version of Petkovšek's algorithm to find the right-hand side.

4. Takayama's algorithm

The computation of A and B in (1) is slightly more than what is strictly necessary. Actually we only need to compute B. N. Takayama gave an algorithm for doing so in the Weyl algebra, and this algorithm generalizes to Ore algebras.

The idea is that it is possible to throw away all the right multiples of ∂ during the computation as long as we know they will only be multiplied by polynomials which commute with ∂ during later computations (so that they will remain right multiples of ∂). This is done by working in increasingly large modules where multiplication by the x_i's which do not commute with ∂ is forbidden. The operator ∂ can then easily be eliminated in a preprocessing phase.

This results in an algorithm which is generally faster than the general one, but which is only guaranteed to terminate when there is an element free of the undesirable variables in the ideal.

Bibliography

[1] Abramov (Sergei A.) and Petkovšek (Marko). – Finding all q-hypergeometric solutions of q-difference equations. In Leclerc (B.) and Thibon (J. Y.) (editors), *Formal power series and algebraic combinatorics*. pp. 1–10. – Université de Marne-la-Vallée, 1995. Proceedings SFCA'95.

[2] Chyzak (Frédéric) and Salvy (Bruno). – *Non-commutative Elimination in Ore Algebras Proves Multivariate Holonomic Identities.* – Research Report n° 2799, Institut National de Recherche en Informatique et en Automatique, February 1996.

[3] Ore (Oystein). – Theory of non-commutative polynomials. *Annals of Mathematics*, vol. 34, 1933, pp. 480–508.

[4] Petkovšek (Marko), Wilf (Herbert), and Zeilberger (Doron). – *A=B*. – A. K. Peters, Wellesley, Mass., 1996.

[5] Takayama (Nobuki). – An algorithm of constructing the integral of a module — an infinite dimensional analog of Gröbner basis. In *Symbolic and algebraic computation*. pp. 206–211. – ACM, 1990. Proceedings of ISSAC'90, Kyoto.

∂-finite functions

Frédéric Chyzak
INRIA Rocquencourt

January 15, 1996

[summary by Bruno Salvy]

Abstract

The algebra of ∂-finite functions and sequences enjoys several closure properties useful when computing a description suitable for creative telescoping. A simple description of ∂-finiteness can be given in the context of Ore algebras. In the special case of the Weyl algebra, a special property called holonomy plays a crucial role.

We consider an Ore algebra $\mathcal{A} = \mathbb{K}(x_1, \dots, x_n)\langle \partial_1, \dots, \partial_k \rangle$ (see previous summary). A function is ∂-finite with respect to $\mathcal{A}$ when its pseudo-derivatives $\boldsymbol{\partial}^{\boldsymbol{\alpha}} f = \partial_1^{\alpha_1} \cdots \partial_k^{\alpha_k} f$ with $\alpha_i \in \mathbb{N}$ for $i = 1, \dots, k$ span a finite-dimensional vector space over $\mathbb{K}(x_1, \dots, x_n)$. Examples of ∂-finite functions in the univariate case are: hypergeometric power series and sequences, solutions of linear recurrences and solutions of linear differential equations. In several variables, it becomes necessary to specify with respect to which operators one considers ∂-finiteness; for instance, all sequences of orthogonal polynomials are ∂-finite with respect to shift of the index and differentiation in the argument.

An equivalent definition is that f is ∂-finite when the module $\mathcal{M} = \mathcal{A} \cdot f$ is finitely generated: $\mathcal{M} = \oplus_{\alpha \in A} \mathbb{K}(\mathbf{x}) \boldsymbol{\partial}^{\boldsymbol{\alpha}} f$, for a finite set of indices A. If $\operatorname{Ann} f$ denotes the ideal of the elements of $\mathcal{A}$ vanishing on f, then $\mathcal{A}/\operatorname{Ann} f$ is isomorphic to $\mathcal{A} \cdot f$, and this yields a purely ideal-theoretic definition of ∂-finiteness which avoids the introduction of functions. An ideal $\mathcal{I}$ of $\mathcal{A}$ is thus called ∂-finite when $\mathcal{A}/\mathcal{I}$ is finitely generated as a $\mathbb{K}(\mathbf{x})$-module.

1. Closure properties

What makes ∂-finite functions so useful is that it is possible to compute with these functions without reference to any sort of "closed-form". Many computations can be performed directly on sets of generators of their annihilating ideal. In particular, sum and product of ∂-finite functions can be obtained this way.

1.1. Rectangular systems. Before giving the algorithms for sum and product we note that a ∂-finite function f is always annihilated by a *rectangular* system of polynomials, which is such that each ∂_i of the algebra is involved in exactly one of the polynomials. Consequently, each of the polynomials involves only one ∂_i. That this is so follows from the finite dimension of $\sum_n \mathbb{K}(\mathbf{x}) \partial_i^n f$, which implies the existence of a linear relation between a finite number of $\partial_i^n f$. Rectangular systems are useful to prove ∂-finiteness of various constructions, or in the case where Gröbner bases are not available. In other cases, they generally describe an ideal which is smaller than the one we would like to work with, and this leads to slower computations.

EXAMPLE. In $\mathcal{A} = \mathbb{Q}(x, y)\langle \partial_x, \partial_y \rangle$, the sum of the Bessel functions $J_\mu(x)$ and $J_\nu(y)$ is annihilated by the rectangular system $\mathcal{S} = \{\partial_x(x^2\partial_x^2 + x\partial_x + x^2 - \mu^2), \partial_y(y^2\partial_y^2 + y\partial_y + y^2 - \nu^2)\}$. If $\phi(x, y)$

is a solution of $\mathcal{S}$, and $\mathcal{I}$ is the ideal generated by $\mathcal{S}$ in $\mathcal{A}$, then it is easily checked that $\mathcal{A}/\mathcal{I}$ is generated by $\{\phi, \partial_x\phi, \partial_x^2\phi, \partial_y\phi, \partial_x\partial_y\phi, \partial_x^2\partial_y\phi, \partial_y^2\phi, \partial_x\partial_y^2\phi, \partial_x^2\partial_y^2\phi\}$ and thus is of dimension 9. However, the annihilating ideal of $f = J_\mu(x) + J_\nu(y)$ also contains $\partial_x\partial_y$. The ideal generated by the adjunction of this polynomial to the rectangular system above is $\operatorname{Ann} f$ and $\mathcal{A}/\operatorname{Ann} f$ is generated by $\{f, \partial_x f, \partial_x^2 f, \partial_y f, \partial_y^2 f\}$ and is only of dimension 5.

1.2. Sum. If f and g are two ∂-finite functions, then by linearity $\boldsymbol{\partial}^{\boldsymbol{\alpha}}(f+g) \in \mathcal{A}f + \mathcal{A}g$ which is finite-dimensional. Hence a sum of ∂-finite functions is ∂-finite.

Given a rectangular system for f and a rectangular system for g a rectangular system for $f+g$ is obtained by reducing $h_n = \partial_i^n f + \partial_i^n g$ for increasing values of n. These reductions use the initial rectangular systems and right Euclidean division, which works in any Ore algebra. All the h_n's are thus rewritten in a finite basis $\{f, \partial f, \ldots, \partial^J f, g, \partial g, \ldots, \partial^K g\}$. The value of n is increased until a linear relation between the h_n's is found by Gaussian elimination.

1.3. Product. We assume that for each ∂_i in the algebra, the morphisms σ_i and δ_i defined by the commutation rule

$$\partial_i p = \sigma_i(p)\partial_i + \delta_i(p)$$

are polynomials in ∂_i over $\mathbb{K}(x_1, \ldots, x_n)$. This is not a severe restriction. Then by the same kind of argument as above, the product of two ∂-finite functions is ∂-finite. The algorithm to produce a rectangular system for the product out of two rectangular systems for the functions being multiplied is exactly the same as above.

1.4. Generalizations. Actually, the same algorithm extends to the direct computation of a rectangular system for any polynomial h in some $\boldsymbol{\partial}^{\alpha_{i,j}} f_i$'s given the rectangular systems defining the f_i's.

The FGLM algorithm [3] provides another generalization: given rectangular systems defining the f_i's and a term order T on the $\boldsymbol{\partial}^{\boldsymbol{\alpha}}$'s, this algorithm returns a Gröbner basis for T. Roughly speaking, this algorithm considers all the monomials $\boldsymbol{\partial}^{\boldsymbol{\alpha}} h$ in the order T and stops when it has found sufficiently many relations. More precisely we start with $F = \{h\}$, the resulting basis is set to $L = \{\}$ and the basis of $\mathcal{A}.h$ is set to $R = \{\}$. At each step the smallest element t of F with respect to T is selected and reduced by the rectangular systems defining the $f_i's$. Gaussian elimination is then performed to detect a linear dependency between t and the elements of R. If no linear dependency is found, t is added to R, removed from F, and all the $\partial_i t$ are added to F. Otherwise, the dependency is added to L. The algorithm stops when F is empty, and returns L.

Note that the Gröbner basis returned by this method is not necessarily a basis of $\operatorname{Ann} f$ since, as we have already seen, the rectangular systems do not necessarily generate a sufficiently large ideal.

Yet another extension consists in using any Gröbner basis for the f_i's instead of a rectangular system. In the reduction step, the Euclidean division is replaced by a reduction using the Gröbner bases.

Once again, when it is available, the advantage of this approach over manipulating only rectangular systems is that it results in modules of a smaller dimension, and therefore lessens the complexity of further computations.

1.5. Example. The following identity between Apéry numbers and Franel numbers was proved by V. Strehl:

$$\sum_{k=0}^{n} \binom{n}{k}^2 \binom{n+k}{k}^2 = \sum_{k=0}^{n} \binom{n}{k} \binom{n+k}{k} \sum_{j=0}^{k} \binom{k}{j}^3. \tag{1}$$

A system is easily found for $\binom{k}{j}^3$ which is hypergeometric:

$$(k+1-j)^3 S_k - (k+1)^3, \qquad (k-j)^3 S_j - (j+1)^3.$$

Then using creative telescoping (see previous summary), one gets an equation for the sum over j:

$$(k+2)^2 S_k^2 - (7k^2+21k+16) S_k - 8(k+1)^2.$$

Again, a system is easily found for $\binom{n}{k}\binom{n+k}{k}$ which is hypergeometric:

$$(n+1-k) S_n - (n+1+k), \qquad (k+1)^2 S_k - (n(n+1) - k(k+1)).$$

The product of this with the previous equation yields a system for the summand of the right-hand side of (1) whose first equation is the first one above (obviously!) and whose second equation is:

$$(k+2)^4 S_k^2 + (n-k-1) S_k + 8(n+k+2)(n+k+1)(n-k) - (7k^2+21k+6)(n+k+2).$$

Now, creative telescoping yields an equation for the right-hand side of (1):

$$(n+2)^3 S_n^2 - (2n+3)(17n^2+51n+39) S_n + (n+1)^3. \tag{2}$$

The same process is then applied to the left-hand side. First, $\binom{n}{k}^2\binom{n+k}{k}^2$ is hypergeometric and satisfies

$$(n+1-k)^2 S_n - (n+1+k)^2, \qquad (k+1)^4 S_k - (n(n+1) - k(k+1))^2.$$

Creative telescoping then yields (2) again. The identity is then proved by checking that two initial conditions coincide, which they do. The whole computation takes less that 10 seconds on a Dec Alpha.

2. Holonomy

The algorithms for creative telescoping which we have described in the previous summary depend on the existence of a polynomial free of one of several variables in the ideal we are working in. It is thus very important to be working in the proper ideal and to be able to check whether such a polynomial exists or not. In the Weyl algebra case, holonomy theory provides such a guarantee. We describe elements of this theory, and give some hints on what remains valid in the more general Ore algebra case.

2.1. Hilbert dimension. Let $\mathcal{A}$ be an Ore algebra: $\mathcal{A} = \mathbb{K}[\mathbf{x}]\langle\boldsymbol{\partial}\rangle$. Let deg denote the total degree with respect to $\mathbf{x}$ and $\boldsymbol{\partial}$. We consider the graduation F_n of $\mathcal{A}$ where F_n contains the elements of $\mathcal{A}$ of degree at most n. Finally, let $h_n = \dim_{\mathbb{K}}(F_n \cdot f)$.

EXAMPLE. For $f = 1$ in the algebra $K[x_1,\dots,x_p]\langle\partial_1,\dots,\partial_p\rangle$, one has $h_n = \binom{n+p}{p} \sim n^p/p!$.

For $f = \exp(x^2)$ in $\mathbb{K}[x]\langle\partial_x\rangle$, it is easy to compute the first few values and be convinced that $h_n = n+1$.

For $f = (s^3 - s^2 + sx)^{-1/2}$ in $\mathbb{K}[s,x]\langle\partial_x,\partial_y\rangle$, the first values indicate that $h_n = 3n^2+2$.

For $f = \exp(\sin(x))$ in $\mathbb{K}[x]\langle\partial_x\rangle$, one gets $h_n = n^2/2 + 3n/2 + 1$.

Finally, for $f = \binom{n}{k}$ in $\mathbb{K}[n,k]\langle S_n, S_k\rangle$, $h_n = 2n+1$.

A general theorem of Hilbert implies that asymptotically, $h_n \sim cn^d$ with d an integer which is called the Hilbert dimension of the ideal. The relevance of this notion to creative telescoping is of a combinatorial nature: if $\mathcal{B}$ is obtained by forming all the monomials in q of the variables $(\mathbf{x},\boldsymbol{\partial})$, then $F_n \cap \mathcal{B}$ contains $\binom{n+q}{q}$ monomials. As soon as this number grows faster than n^d where d is the Hilbert dimension of the annihilating ideal of some f, then a linear combination of elements of $\mathcal{B}$ has to vanish on f, which means that the ideal contains elements of $\mathcal{B}$.

2.2. Weyl algebra. The Weyl algebra is a special case of a polynomial Ore algebra $\mathcal{A}_p = \mathbb{K}[x_1, \ldots, x_p]\langle \partial_1, \ldots, \partial_p \rangle$ where ∂_i is the differentiation operator with respect to the corresponding x_i, for $i = 1, \ldots, p$. A fundamental theorem of Bernstein states that in this case, the Hilbert dimension of an ideal is always larger than p. Those ideals for which the Hilbert dimension is exactly p are called *holonomic*. By extension, a function whose annihilating ideal in a Weyl algebra is holonomic will be called holonomic too. In the examples above, $\exp(x^2)$ and $(s^3 - s^2 + sx)^{-1/2}$ are holonomic functions, while $\exp(\sin x)$ is not.

Holonomy of functions is preserved under sum and product, algebraic functions are holonomic, algebraic substitution preserves holonomy, the diagonal of a holonomic function is holonomic [4, 5]. In addition, a result due to Kashiwara states that when an ideal $\mathcal{I}$ in the *rational* Ore algebra $\mathbb{K}(x_1, \ldots, x_p)\langle \partial_1, \ldots, \partial_p \rangle$ is ∂-finite, then $\mathcal{I} \cap \mathcal{A}$ is a holonomic ideal. This means that all ∂-finite functions with respect to differentiation are also holonomic. Finally, creative telescoping always works in holonomic ideals.

3. Conclusions

The algorithms we have given work in a very general context of Ore algebras. However, creative telescoping is never guaranteed *a priori* to give an answer in the general case, unless the existence of the result is ensured, for instance by holonomy. An advantage of our approach is that it may well return results in non-holonomic cases.

An important difficulty will be the subject of future work. Even in the Weyl algebra case, the ideals $\mathcal{I}$ we are dealing with have a natural description in *rational* Ore algebras $\mathbb{K}(\mathbf{x})\langle \boldsymbol{\partial} \rangle$. However, for creative telescoping what we need is a basis of $\mathcal{I} \cap \mathbb{K}[\mathbf{x}]\langle \boldsymbol{\partial} \rangle$. At the moment, we do not have any algorithm to produce this basis. However, algorithms exist to deal with the same problem in the commutative case, and they might extend to this framework.

This problem is illustrated by the computation of the diagonal of $1/(1 - x - y)$. This can be obtained via a residue computation as the definite integral of $f = (s^2 - s + x)^{-1}$ which is holonomic. Thus creative telescoping applies and there exists an operator free of s in the ideal. The annihilating ideal $\operatorname{Ann} f$ of f in $\mathbb{K}(s, x)\langle \partial_s, \partial_x \rangle$ is generated by $\mathcal{S} = \{(s^2 - s + x)\partial_s + 2s - 1, (s^2 - s + x)\partial_x + 1\}$. However, the ideal generated by $\mathcal{S}$ in the Weyl algebra $\mathcal{A} = \mathbb{K}[s, x]\langle \partial_s, \partial_x \rangle$ is smaller than $\operatorname{Ann} f \cap \mathcal{A}$ and does not contain any polynomial free of s. To get such a polynomial, it is necessary to augment $\mathcal{S}$, for instance with $(s^2 - s + x)\partial_s\partial_x + 2\partial_s$.

Bibliography

[1] Cartier (Pierre). – Démonstration 'automatique' d'identités et fonctions hypergéométriques. *Astérisque*, vol. 206, 1992, pp. 41–91. – Séminaire Bourbaki.

[2] Chyzak (Frédéric) and Salvy (Bruno). – *Non-commutative Elimination in Ore Algebras Proves Multivariate Holonomic Identities.* – Research Report n° 2799, Institut National de Recherche en Informatique et en Automatique, February 1996.

[3] Faugère (J. C.), Gianni (P.), Lazard (D.), and Mora (T.). – Efficient computation of zero-dimensional Gröbner bases by change of ordering. *Journal of Symbolic Computation*, vol. 16, 1993, pp. 329–344.

[4] Lipshitz (L.). – The diagonal of a D-finite power series is D-finite. *Journal of Algebra*, vol. 113, 1988, pp. 373–378.

[5] Lipshitz (L.). – D-finite power series. *Journal of Algebra*, vol. 122, n° 2, 1989, pp. 353–373.

[6] Petkovšek (Marko), Wilf (Herbert), and Zeilberger (Doron). – *A=B.* – A. K. Peters, Wellesley, Mass., 1996.

Computing the Distance of a Point to an Algebraic Hypersurface and Application to Exclusion Methods

Xavier Gourdon

Algorithms Project, INRIA Rocquencourt

February 12, 1996

[summary by Pierre Nicodème]

Abstract

We compute lower bounds for the distance in $\mathbb{C}^n$ from a point u to an algebraic surface $\mathcal{Z}$. Such lower bounds or proximity tests give an approximation of $\mathcal{Z}$. We present tests based on both Taylor's formula and a generalization of the Dandelin-Graeffe process to the multivariate case, and their application to the exclusion method [2].

1. Introduction

Given a point a in $\mathbb{C}^n$, and an algebraic hypersurface

$$\mathcal{Z}(P) = \{(z_1, \ldots, z_n) \in \mathbb{C}^n \,|\, P(z_1, \ldots, z_n) = 0\},$$

with $P \in \mathbb{C}[z_1, \ldots, z_n]$, we want to evaluate the distance $d(a, \mathcal{Z})$ corresponding to the norm

$$\|z\| = \max_{1\leq k\leq n} |z_i|.$$

By shifting the variable z, we can restrict to the case $a = 0$.

2. Univariate Polynomials

Let $P(z) = \sum_{i=0}^{d} a_i z^i \in \mathbb{C}[z]\,, a_d \neq 0$, and $\mathcal{Z}(P) = \{U_1, \ldots, U_d\}$. We want to evaluate $d(0, \mathcal{Z}) = \min_i |U_i|$. In Henrici [4, vol. 1], Theorems 6.4.d and 6.4.i give the following classical bound for $\mathcal{Z}(P)$:

Proposition 1. *If $\rho(P)$ is the nonnegative root of the equation $|a_0| = \sum_{j=1}^{d} |a_j| \rho^j$, then*

$$\rho(P) \leq d(0, \mathcal{Z}) \leq \frac{1}{2^{1/d} - 1}\rho(P) \approx \frac{d}{\log 2}\rho(P).$$

Graeffe Iteration. With $P(z) = a_d \prod_{i=1}^{d}(z - U_i)$, we consider

$$P(z)P(-z) = (-1)^d a_d^2 \prod_{i=1}^{d}(z^2 - U_i^2) = P^{\langle 1\rangle}(z^2).$$

We note $P^{\langle 1\rangle}$ the classical Graeffe iterate; the roots of $P^{\langle 1\rangle}$ are the squares of those of P, and $d(0, \mathcal{Z}(P^{\langle 1\rangle})) = d(0, \mathcal{Z}(P))^2$; we have

$$\rho(P^{\langle 1\rangle}) \leq d(0, \mathcal{Z}(P^{\langle 1\rangle})) \leq \frac{\rho(P^{\langle 1\rangle})}{2^{1/d} - 1};$$

so with $\rho_1 = \sqrt{\rho(P^{\langle 1 \rangle})}$, we get

$$\rho_1 \le d(0, \mathcal{Z}(P)) \le \frac{\rho_1}{(2^{1/d}-1)^{1/2}}.$$

Generally, we define $P^{\langle k \rangle} = \mathrm{Graeffe}(P^{\langle k-1 \rangle})$; then, we get $d(0, \mathcal{Z}(P^{\langle k \rangle})) = d(0, \mathcal{Z}(P))^{2^k}$; with $\rho_k = \rho(P^{\langle k \rangle})^{1/2^k}$, we have

$$\rho_k \le d(0, \mathcal{Z}(P)) \le \frac{\rho_k}{(2^{1/d}-1)^{1/2^k}}.$$

The upper bound tends rapidly to the lower bound as k increases, thus we have obtained an effective process to compute $d(0, \mathcal{Z})$.

Computing the $P^{\langle k \rangle}$. With $A(z) = \sum_{i \equiv 0 \bmod 2} a_i z^{i/2}$ and $B(z) = \sum_{i \equiv 1 \bmod 2} a_i z^{(i-1)/2}$, we have

$$P(z)P(-z) = A(z^2)^2 - z^2 B(z^2)^2,$$

and therefore,

$$\mathrm{Graeffe}(P) = A(z)^2 - zB(z)^2.$$

A practical problem is that the coefficient size doubles at each Graeffe iteration.

3. Multivariate Polynomials

In the multivariate case, the polynomial $P(z)P(-z)$ can not be written as $Q(z^2)$ where $Q(z)$ is a polynomial, thus we need to modify the definition. We generalize the Graeffe process to the multivariate case as follows:

Definition 1. We call the N-th Graeffe iterate of $P(z) \in \mathbb{C}[z_1, \ldots, z_n]$ the polynomial $P^{[N]}(z)$ defined by

$$P^{[N]}(z) = \prod_{j=0}^{2^N-1} P(\omega^j z), \qquad \omega = \exp\left(\frac{2i\pi}{2^N}\right), \quad i^2 = -1,$$

where $\omega^j z$ denotes the point $(\omega^j z_1, \ldots, \omega^j z_n)$.

Proposition 2. *For all non negative integer N, the N-th Graeffe iterate of $P(z)$ writes as*

$$P^{[N]}(z) = \sum_{j \ge 0} B_j^{[N]}(z),$$

where the $B_j^{[N]}$'s are homogeneous polynomials of degree $2^N j$. The $(N+1)$-st Graeffe iterate can be computed from the N-th thanks to the formula

$$P^{[N+1]}(z) = P_0^{[N]}(z)^2 - P_1^{[N]}(z)^2, \qquad P_k^{[N]}(z) = \sum_{j \equiv k \bmod 2} B_j^{[N]}(z).$$

With the multivariate Graeffe process, we easily generalize the univariate algorithm to compute $d(0, \mathcal{Z})$ in the multivariate case.

Theorem 1. *Let $P(z)$ be a polynomial in $\mathbb{C}[z_1, \ldots, z_n]$ of total degree d. Let $P^{[N]}(z) = \sum_{j \ge 0} B_j^{[N]}(z)$ be its N-th Graeffe iterate and R_N the non-negative solution of the equation in R*

(1) $$|P^{[N]}(0)| = \sum_{j \ge 1} \|B_j^{[N]}\|_\infty R^j,$$

d	r_0/d	r_1/d	r_2/d	r_3/d	r_4/d
2	0.7673	0.9725	0.9996	1.0000	1.0000
5	0.6525	0.9479	0.9973	1.0000	1.0000
7	0.6325	0.9400	0.9960	0.9999	1.0000
15	0.6067	0.9271	0.9938	0.9999	1.0000

TABLE 1. Some values of $r_N/d(0, \mathcal{Z}_{n,d})$ for $n = 2$.

d	r_0/d	r_1/d	r_2/d	r_3/d
2	0.5832	0.6338	0.8108	0.8224
3	0.4802	0.5108	0.6478	0.7561

TABLE 2. Some values of $r_N/d(0, \mathcal{Z}_{n,d})$ for $n = 7$.

where $\|B_j^{[N]}\|_\infty = \sup_{\|z\|=1} \|B_j(z)\|$. *Then we have*

$$r_N \le d(0, \mathcal{Z}) \le \left(\frac{1}{2^{1/d}-1}\right)^{2^{-N}} r_N, \qquad r_N = R_N^{2^{-N}}. \tag{2}$$

Computing $\|B_j^{[N]}\|_\infty$ raises a difficult practical problem; therefore, we make use of the norm $\|\sum_\alpha a_\alpha z^\alpha\| = \sum |a_\alpha|$, easy to compute. Our main result is stated using this norm; one demonstrates the equivalence of the norms $\|\cdot\|_\infty$ and $\|\cdot\|$ by combination of the Parseval identity and of the Cauchy-Schwarz inequality.

THEOREM 2. *Let ρ_N be the unique nonnegative solution of*

$$|P^{[N]}(0)| = \sum_{j=1}^{d} \|B_j^{[N]}\|\rho^j \tag{3}$$

The distance from 0 to $\mathcal{Z}$ satisfies

$$r_N \le d(0, \mathcal{Z}) \le \kappa_N r_N, \tag{4}$$

where

$$r_N = \rho_N^{2^{-N}} \quad \textit{and} \quad \kappa_N = \left(\frac{1}{2^{1/d}-1}\sqrt{\binom{2^N+n-1}{n-1}}\right)^{1/2^N}.$$

Moreover $\lim_{N\to\infty} \kappa_N = 1$, *which implies* $\lim_{N\to\infty} r_N = d(0, \mathcal{Z})$.

4. Examples

We take a polynomial of degree d in n variables: $P_{n,d} = \sum_{j=1}^n (1 - z_j)^d - 1$. With $\mathcal{Z}_{n,d} = \mathcal{Z}(P_{n,d})$, we have $d(0, \mathcal{Z}_{n,d}) = 1 - \frac{1}{n^{1/d}}$.

Tables 1 and 2 give the value of the ratio $r_N/d(0, \mathcal{Z}_{n,d})$ of Theorem 3 for several values of n, d and N. The computations were performed in Maple. These examples show that the bound is quite good for a small value N of Graeffe iterates.

5. Exclusion methods

We give the principle of the method for a polynomial of one variable $P(z) \in \mathbb{C}[z]$.

- Let the *exclusion function* be: $z_0 \mapsto \rho(z_0)$, with ρ given by theorem 2 after a proper shift of the variable, and
  (1) $\rho(z_0) = 0 \iff P(z_0) = 0$,
  (2) P has no zero in $|z - z_0| < \rho(z_0)$, which is equivalent to $\rho(z_0) \le d(z_0, \mathcal{Z})$;
- then, the *exclusion test* is: let C be a square of centre z_0 and half-side $a > 0$. If $\rho(z_0) \ge \sqrt{2}a$, C contains no zero of P.

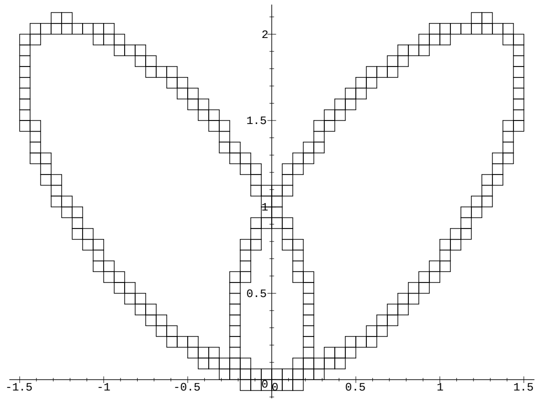

FIGURE 1. Representing by exclusion the curve $y^4-2y^3+y^2-3x^2y+2x^4=0$ (petal).

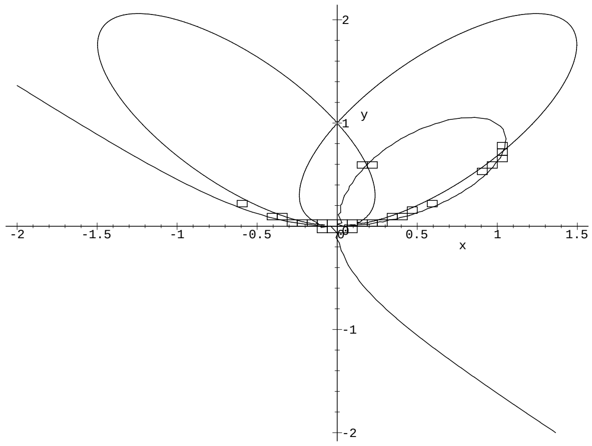

FIGURE 2. Intersection of the curves $x^3+y^3-2xy=0$ (Descartes folium) and $y^4-2y^3+y^2-3x^2y+2x^4=0$ (petal).

Exclusion algorithm.

- Consider the reciprocal polynomial $R(z)$ of $P(z)$; compute by Graeffe a lower bound of the smallest root of $R(z)$, which gives an upper bound b_u of the largest root of $P(z)$;
- Start from a big square centred at the origin, with side $2b_u$, which contains all the roots of $P(z)$;
- Recursively split the square in four squares of equal size, discarding by the exclusion test squares containing no zeros;
- Stop the recursion when the desired precision is reached (the surface of the area covering the zeros decreases exponentially fast to zero).

Figure 1 shows an application of the exclusion method to localize an algebraic curve in $\mathbb{R}^2$.

For an algebraic variety $\mathcal{Z}_i = \mathcal{Z}(P_i)$ and $\mathcal{Z} = \bigcap_i \mathcal{Z}(P_i)$, with $P_1,\ldots,P_m \in \mathbb{C}[z_1,\ldots,z_n]$, let $\rho_i(z_0)$ be an exclusion function defined by theorem 2 for $P_i, (1\le i\le m)$; we can define an exclusion function for the variety as $\rho(z_0)=\sup_{1\le i\le m}\rho_i(z_0)$.

An application of exclusion method to localize the intersection of two curves in $\mathbb{R}^2$ is given in Figure 2.

Bibliography

[1] Bareiss (Erwin H.). – Resultant procedure and the mechanization of the Graeffe process. *Journal of the ACM*, vol. 7, 1960, pp. 346–386.

[2] Dedieu (Jean-Pierre), Gourdon (Xavier), and Yakoubsohn (Jean-Claude). – Computing the distance from a point to an algebraic hypersurface. – July 1996. Seminar of the American Mathematical Society. Park City. 8 pages. In press.

[3] Dedieu (Jean-Pierre) and Yakoubsohn (Jean-Claude). – Localization of an algebraic hypersurface by the exclusion algorithm. *Applicable Algebra in Engineering, Communication and Computing*, vol. 2, 1992, pp. 239–256.

[4] Henrici (Peter). – *Applied and Computational Complex Analysis.* – John Wiley, New York, 1977. 3 volumes.

[5] Pan (V.). – Solving a polynomial equation: some story and recent progress. – 1995. Preprint.

Matrix-based methods for solving polynomial systems

Ioannis Emiris
Projet SAFIR, Inria Sophia-Antipolis

11 mars, 1996

[summary by Frédéric Chyzak]

Abstract

We present a uniform approach to the elimination of variables between polynomials and the construction of matrices that express resultants. Building a matrix whose determinant is a multiple of the resultant reduces the solving of a polynomial system to a generalized eigenvalues/eigenvectors problem for a square matrix. Several such matrices are of interest, in particular the Newton and Bézout/Dixon matrices, which lead to efficient calculations.

1. Classical resultants versus sparse resultants

Classically, the resultant is a single polynomial which characterizes the solvability of a system of dense polynomials [7]. We introduce another concept of resultant which takes the structure of the coefficients into account.

Let $f_1(\mathbf{c},x),\ldots,f_{n+1}(\mathbf{c},x)$ be $n+1$ polynomials in the n indeterminates $x_1,\ldots,x_n$ and with coefficients that are polynomial in $c_1,\ldots,c_N$ over a field $\mathbb{K}$. A *sparse resultant* $R(\mathbf{c})$ with respect to a subfield $\mathbb{L}$ of the algebraic closure $\overline{\mathbb{K}}$ is an irreducible polynomial of $\mathbb{K}[c_1,\ldots,c_N]$ that vanishes at a specialization γ of the c_i's if and only if the corresponding specializations of the f_i's have a common zero. In other words, the resultant satisfies

$$\forall\gamma\in\mathbb{L}^N\quad (R(\gamma)=0\Longleftrightarrow\exists\xi\in\mathbb{L}^n\quad\forall i=1,\ldots,n\quad f_i(\gamma,\xi)=0).$$

For some applications, one requires that the coefficients of the f_i's be generic, i.e., that one c_i be introduced for each coefficient. Special cases are of particular interest. In the case of dense homogenized polynomials

$$f_i(x_0,x_1,\ldots,x_n)=\sum_{a_0+\cdots+a_n=d_i} c_{a_0,\ldots,a_n}x_0^{a_0}\ldots x_n^{a_n},$$

we recover the classical *homogeneous resultant* [7]. In the case of two (dense) univariate polynomials, we recover Sylvester's classical notion of the *univariate resultant* [6], whose expression as a determinant is recalled in the next section. In the case of (possibly sparse) polynomials with generic coefficients, i.e., when

$$f_i(x_1,\ldots,x_n)=\sum_{j=1}^{r_i} c_{i,j}x_1^{a_{i,j,1}}\ldots x_n^{a_{i,j,n}}$$

for non-zero undetermined coefficients $c_{i,j}$ that are transcendental over the field $\mathbb{K}$, the resultant $R(\mathbf{c})$ is called the *sparse resultant* of the f_i's.