The equality problem for holonomic functions is decidable

Journal of Computational and Applied Mathematics 32 (1990) 321–368
North-Holland

A holonomic systems approach to special functions identities *

Doron ZEILBERGER
Department of Mathematics, Temple University, Philadelphia, PA 19122, USA

Received 14 November 1989

Abstract: We observe that many special functions are solutions of so-called holonomic systems. Bernstein's deep theory of holonomic systems is then invoked to show that any identity involving sums and integrals of products of these special functions can be verified in a finite number of steps. This is partially substantiated by an algorithm that proves terminating hypergeometric series identities, and that is given both in English and in MAPLE.

Keywords: Elimination, Weyl algebra, partial difference operators, computer algebra, hypergeometric series.

Certain functions appear so often that it is convenient to give them names. These functions are collectively called special functions. There are many examples and no single way of looking at them can illuminate all examples or even all the important properties of a single example of a special function.

Richard Askey (1984)

1. Introduction

We are nowadays witnessing a spectacular comeback of the theory of special functions. This is expressed both by the various approaches [6,18,20,41] that are successfully used to *explain* and give insight to large families of previously unrelated results, and by the dramatic applications of special function identities to the solution of longstanding open problems in pure mathematics [7,16,55].

This paper initiates yet another approach to special functions that is based on Bernstein's theory of holonomic systems. Unlike the other approaches, the present approach does not give insight on any one particular identity. Instead it gives a kind of *universal insight* on many

* Supported in part by NSF grant DMS 8800663.

identities at once. It implies that a large class of special function identities, that includes all terminating hypergeometric (alias binomial coefficients) identities, is verifiable in a *finite* number of steps.

One of the most salient features of the classical sequences of orthogonal polynomials $\{p_n(x)\}$ is that they satisfy both a second-order linear *differential* equation and a second-order linear *recurrence* equation, both with coefficients that are polynomials in n and x. For example, the Legendre polynomials $\{P_n(x)\}$ satisfy [45, Chapter 10]:

$$(1-x^2)P_n''(x) - 2xP_n'(x) + n(n+1)P_n(x) = 0, \tag{1.1a}$$

$$(n+2)P_{n+2}(x) - (2n+3)xP_{n+1}(x) + (n+1)P_n(x) = 0. \tag{1.1b}$$

Viewing the sequence $P_n(x)$ as one function on $\mathbb{N}\times\mathbb{R}$: $F(n, x) := P_n(x)$ and introducing the differentiation and shift operators

$$Df(n, x) := \frac{\mathrm{d}}{\mathrm{d}x}f(n, x), \qquad Ef(n, x) := f(n+1, x),$$

(1.1) can be rewritten as

$$\left[(1-x^2)D^2 - 2xD + n(n+1)\right]F(n, x) = 0, \tag{1.2a}$$

$$\left[(n+2)E^2 - (2n+3)xE + (n+1)\right]F(n, x) = 0. \tag{1.2b}$$

A system like (1.2) is called "a maximally overdetermined system of linear differential-recurrence equations with polynomial coefficients on $\mathbb{N}\times\mathbb{R}$", or a *holonomic system* on $\mathbb{N}\times\mathbb{R}$, for short. Since (1.2a) and (1.2b) are independent (in a certain technical sense to be made precise later), they *uniquely* determine $F(n, x)$, given a *finite* number of initial conditions. (In this example these are: $F(0, 0) = 1$, $F'(0, 0) = 0$, $F(1, 0) = 0$, $F'(1, 0) = 1$.) A function like $F(n, x)$ that is a solution of a holonomic system is called a *holonomic function* on $\mathbb{N}\times\mathbb{R}$.

If we allow linear equations of any order, and generalize from one discrete and one continuous variable to several variables of each kind, we are led to consider the class of *holonomic functions* $f(x_1, \ldots, x_{n_1}, m_1, \ldots, m_{n_2})$ on $\mathbb{R}^{n_1}\times\mathbb{N}^{n_2}$. These are functions that satisfy "as many (homogeneous) linear (differential-recurrence) equations as possible (with polynomial coefficients)". They are uniquely determined as solutions of a system of such equations subject to a *finite* number of initial conditions. The notion of "satisfying as many equations as possible" is made precise in Bernstein's theory of holonomic systems described in Section 2.

The following two crucial facts make holonomic functions an ideal framework for special function identities.

(i) Every holonomic function can be described by a *finite* amount of information, and has a "canonical holonomic representation".

(ii) The product, addition, sum, and integral of holonomic functions is again holonomic. Furthermore, if one has canonical holonomic representations of f and g, then it is possible to find canonical holonomic representations for $f+g$, $f-g$, fg, $\int f\,\mathrm{d}x_1$ and $\sum_{m_1} f$, where x_1 and m_1 are any continuous and discrete variable, respectively.

It follows that an expression like

$$P_{n_1}(x)P_{n_2}(x)P_{n_3}(x)$$

is holonomic in (n_1, n_2, n_3, x), and that its (definite) integral is holonomic in (n_1, n_2, n_3). Since all the special functions that appear in the famous "tableau d'Askey" [8,37] are given by hypergeometric summation, whose summands are obviously holonomic in all their variables and parameters, it follows that they are all holonomic, both in their variable and their parameters, as are all expressions obtained from them by adding, multiplying, summing and integrating. For example, the Jacobi polynomials $P_n^{(\alpha,\beta)}$ are not only holonomic in (n, x) for fixed (α, β), but are in fact holonomic in (n, x, α, β). It follows that an expression like

$$\int \mathrm{e}^{-\alpha} \sum_n L_n^{(\alpha)}(x) P_n^{(\alpha,\beta)}(x)\, \mathrm{d}\alpha,$$

where $L_n^{(\alpha)}(x)$ are the generalized Laguerre polynomials, is holonomic in x and β.

The above two properties of holonomic functions enable us, at least in principle, to *prove*, or *refute*, any identity involving sums and integrals of products of holonomic functions. All one has to do is to bring everything to the left side and leave only 0 on the right side. Then a canonical holonomic representation of the expression on the left is evaluated step by step, and then it is determined whether such a representation is equivalent to 0 (see Section 4). Note that canonical holonomic representations are not unique, but it is always possible to know when such a representation is equivalent to 0, and thus it is possible to know when two different canonical representations represent the same function. This is analogous to the fact that among the rational numbers a/b, with a and b integers, 0 has an infinite number of representations: $0/b$, and we know that whenever the numerator is 0, the fraction represents 0. Because of this we can tell when two fractions a/b and c/d represent the same rational number: $ad - bc = 0$, since $ad - bc$ is the numerator of $a/b - c/d$. Since the sum and product of rational numbers are again rational, and we know how to find "canonical representations" for sums and products, it follows that any identity involving sums and products of (specific) rational numbers is provable. The same holds for polynomials and algebraic numbers.

Let us see how to prove an identity like

$$\sum_{n=0}^{\infty} P_n(x) t^n - (1 - 2xt + t^2)^{-1/2} = 0. \tag{$*$}$$

$P_n(x)$ is holonomic in n and x, t^n is holonomic in t and n, thus the summand $P_n(x)t^n$ is holonomic in (x, t, n). Summation with respect to n gets rid of the dependence on n, and the sum is holonomic in the surviving variables (x, t). Now $(1 - 2xt + t^2)^{-1/2}$ is clearly holonomic in x, t, as you can easily see by logarithmic differentiation with respect to x and t. Thus the left side of $(*)$ is holonomic in (x, t), and it is possible to find a canonical holonomic representation for it, and determine whether it is indeed equivalent to 0.

An important special case of special functions identities is that of terminating hypergeometric summation. Thanks to the preaching of Askey and his disciples [48], it is nowadays well known that the theory of combinatorial sums, traditionally pursued by combinatorialists, is just a special case of hypergeometric summation, and that many apparently distinct combinatorial identities are really equivalent *qua* hypergeometric sums.

I hope that more professional programmers and algorithmic designers will soon expand the rudimentary ideas in this paper and develop a symbolic software package to prove general special function identities. As a modest first step, I present, in Sections 5 and 6, an explicit algorithm to verify any identity of the form

$$\sum_k F(n, k) = a(n),$$

where $F(n, k)$ is a quotient of products of factorial expressions of the form $(an+bk+c)!$, where a and b are integers and c is a complex number or parameter. Since both $(an+bk+c)!$ and its reciprocal are obviously holonomic in n and k, it follows that $F(n, k)$ is holonomic, and thus that its sum is holonomic in n. But being holonomic in n means that it satisfies a certain (homogeneous) linear recurrence equation with polynomial coefficients. The algorithm of Section 6 finds such a recurrence and then checks whether the right side $a(n)$ also satisfies the same recurrence, and whether the initial conditions match.

A MAPLE program implementing this algorithm is given in the Appendix. This program can be used by all those who have MAPLE, and who are willing to type it in, or who have e-mail, so that I can send it to them. In order to use it all you have to do is read the simple instructions in the Appendix. Readers who are only interested in binomial coefficients identities should go directly to Section 5. [1]

I should warn the reader that my program requires a lot of memory, and it still remains to be seen whether it can prove all known identities with the memory available on a big computer. However, even on my AT&T 3B1 PC, the program was able to prove many nontrivial identities, some of which are given in Section A.2 of the Appendix.

The most time- and space-consuming part of the algorithms for proving special function identities is elimination. This elimination takes place in the noncommutative Weyl algebra of linear differential-recurrence operators with polynomial coefficients. The algorithm that I describe is an adaptation of Sylvester's classical dialytic elimination [56] to the noncommutative context of the Weyl algebra. Sylvester's method, while elegant and simple theoretically, is known to be computationally inefficient, especially when more than one variable is being eliminated. We are nowadays witnessing an ongoing revolution in computational commutative algebra that is centered around Buchberger's concept [14] of Grobner bases. It turns out that, using Grobner bases, it is possible to perform elimination much faster.

Galligo [21] showed how to adapt the method of Grobner bases to the non commutative Weyl algebra. This should result in a considerable improvement in the verification algorithms.

The connection between special functions and holonomic systems has been recently used by Takayama [52,53], who also realized the importance of Grobner bases. Gelfand and his Moscow school [25–28] are currently developing a general theory of hypergeometric functions in several variables, and they showed that they are holonomic functions.

Apology

In [60] I attempted to develop a theory of special function identities by generalizing Sister Celine Fasenmayer's [19] method. While her method is absolutely correct, my attempt to base a general theory on it was wrong. In addition, the definitions of "multi-P-finite", "multi-P-recursive" and "special" given there are inaccurate. I wish to thank Lipshitz, Stanley and Knuth for pointing out the shortcomings of [60]. I am hereby retracting [60] from my list of publications and beg that it be only considered as a historical document.

[1] *Note added in proof.* See [61,62] for a much faster algorithm for this case. However, the present algorithm is much more general, $F(n, k)$ can be any holonomic function.

Limitations of the present approach

Besides the practical limitations of the verification algorithms, there is also a theoretical barrier. The present approach only shows that an identity with a *fixed* number of variables is verifiable in a finite number of steps. For example, the binomial theorem

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

is trivially finitely verifiable. On the other hand, the multinomial theorem

$$(x_1 + \cdots + x_m)^n = \sum_{k_1 + \cdots + k_m = n} \frac{n!}{k_1! \cdots k_m!} x_1^{k_1} \cdots x_m^{k_m}$$

that has an *indefinite* number of variables, is beyond the scope of the present theory, and is *not* finitely verifiable, even not in principle. Of course, for every specific m, say $m = 1000$, the identity is finitely verifiable, but not for an arbitrary number of variables m.

Other examples are Macdonald's root-system conjectures [40]. For every *fixed* root-system, the conjecture is verifiable in a finite number of steps, for example for all the exceptional root systems. But for the infinite families, not even zillion years will suffice.

Milne's deep theory of hypergeometric SU(N) identities [42] is also not covered by the verification algorithms, since it involves an indefinite number of variables.

The theory and algorithms of this paper are easily extendible to the q-analogs of hypergeometric series and special functions. On the other hand, it is not extendible to bibasic and multibasic identities that are presently vigorously studied by Gasper and Rahman (see [22–24]).

Yet another example of the limitation of the present approach is furnished by the multivariate generalizations of the finite form of the Rogers–Ramanujan identities found by Andrews and Bressoud (see [4, p.30]). The finite form of the Rogers–Ramanujan identity is provable by my algorithm, but its multivariate generalization is not. Going in the other direction, toward the less general, the present method is incapable of proving the Rogers–Ramanujan identities directly, and we need a Schur or an Andrews to come up with a *conjectured* finite form that tends to the identities as $n \to \infty$. It would be interesting if one could develop heuristics, in the spirit of [4, Chapter 10], of conjecturing a finite form of a given q-series identity, that could then be proved by the present method.

2. A short course on holonomic systems

2.1. C-finite functions and sequences

2.1.1. C-finite functions

We all know that solutions of (homogeneous) linear ordinary differential equations with constant coefficients

$$P(D)f \equiv 0, \quad P \text{ a polynomial in } D, \tag{2.1}$$

can be expressed as a finite linear combination of exponential polynomial solutions. Indeed, any solution of (2.1) can be expressed as

$$f(x) = \sum_{r=1}^{R} p_r(x)\, e^{\lambda_r x}, \tag{2.2}$$

where $\{\lambda_r\}$ are the roots of the characteristic equation $P(z) = 0$, and the degree of $p_r(x)$, $r = 1, \ldots, R$, is one less than the multiplicity of the root λ_r.

Let us call *C-finite functions* such solutions of constant-coefficients linear ordinary differential equations. It is easily seen that conversely, any function that can be written in the form (2.2) is a C-finite function: $P(D)f = 0$, where

$$P(D) = \prod_{r=1}^{R} (D - \lambda_r)^{\deg(p_r)+1}. \tag{2.3}$$

In order to specify a C-finite function, one only needs a finite number of parameters. One way of coding such functions is by (2.2): specifying the complex numbers λ_r and the polynomials $p_r(x)$. Another way is directly in terms of (2.1); specify the operator $P(D)$, of order N, say, and the N initial conditions $f(0), f'(0), \ldots, f^{(N-1)}(0)$.

In the foregoing we took $P(D)$ to be with arbitrary complex coefficients, and we made the tacit assumption that it is always possible to "write down" an arbitrary complex number. This is of course wrong, since real and complex numbers are merely fictitious entities, and infinite storage space is needed to specify an arbitrary real or complex number. In order to be rigorous we must restrict the coefficients of $P(D)$ to be integer, rational, or at least to belong to some finite algebraic extension field of the rationals. Then the $\{\lambda_r\}$ are algebraic numbers, and algebraic numbers *are* specifiable by a finite number of bits.

We will no longer dwell on this subtlety. All the results will be true for an arbitrary field of characteristic zero. However, whenever we talk about something being "finite" we will tacitly assume that the members of the field are finitely specifiable.

The class of C-finite functions is an algebra: the sum and product of two expressions of the form (2.2) are again of the same form. This can also be proved directly from (2.1). It is easily seen that f is C-finite if and only if the vector space spanned by $\{D^i f;\ i \geqslant 0\}$ is a finite-dimensional vector space. The vector spaces spanned by $\{D^i(f+g);\ i \geqslant 0\}$ and $\{D^i(fg);\ i \geqslant 0\}$ are subspaces of the direct sum and (the homomorphic image of the) tensor product, respectively, of the finite-dimensional vector spaces $\{D^i f;\ i \geqslant 0\}$ and $\{D^i g;\ i \geqslant 0\}$. Thus if f and g are C-finite so are $f+g$ and fg.

Let us summarize the three definitions of C-finite functions.

Definition 0. f is *C-finite* if there exists a polynomial P such that $P(D)f \equiv 0$.

Definition 1. f is *C-finite* if the vector space $C[D]f = \mathrm{span}\{D^i f;\ i \geqslant 0\}$ is finite dimensional.

Definition 2. f is *C-finite* if it can be expressed as a linear combination of exponential polynomial functions, as in (2.2).

In order to motivate holonomic systems, we will introduce some notions that will seem like "abstract nonsense" in the present context.

$C[D]$ is the ring of polynomials in D, (that is the ring of constant-coefficients ordinary linear differential operators). Consider the following left ideal in $C[D]$:

$$I_f = \{P \in C[D];\ P(D)f = 0\}. \tag{2.4}$$

Of course since $C[D]$ is a principal ideal ring, we can write $I_f = P_0(D)C[D]$, where $P_0(D)$ is the minimal order operator that satisfies $P_0(D)f = 0$. Both $C[D]f$ and $C[D]/I_f$ have obvious structure as "$C[D]$ modules": $Q(D)(P(D)f) = (Q(D)P(D))f$ and $Q(D)\overline{P}(D) = \overline{QP}(D)$, where $\overline{P}$ is the image of P under the natural mapping $C[D] \to C[D]/I_f$. It is readily seen that these two $C[D]$ modules are in fact isomorphic, since $P_1(D)f = P_2(D)f$ if and only if $P_1 - P_2 \in I_f$.

We thus have the following definition.

Definition 1′. f is *C-finite* if $C[D]/I_f$ is finite dimensional.

For every function f, let us introduce the "variety associated with the ideal I_f":

$$V_f := \{z \in \mathbb{C};\ P(z) = 0 \text{ for every } P \in I_f\}. \tag{2.5}$$

Let us consider an arbitrary function f on the real line. Most likely it is not a solution of any constant-coefficients linear differential equation, in which case the ideal I_f is the trivial ideal consisting of 0 alone, and $C[D]/I_f$ is equal to $C[D]$, and V_f is the whole of C, and therefore it is a "one-dimensional variety". The other extreme is $f \equiv 0$, in which case $I_f = C[D]$ and $C[D]/I_f$ is the zero module, and V_f is empty, i.e., it is a "$-\infty$-dimensional variety".

Between these two extremes are the C-finite functions for which V_f is a finite set of points, i.e., a "zero-dimensional variety". Thus we have the next definition.

Definition 3. f is *C-finite* if V_f is a zero-dimensional variety (i.e., it is a finite set of points).

Since the class of C-finite functions is an algebra, and every C-finite function is finitely specifiable, it follows that every identity involving sums and products of C-finite functions is routinely verifiable.

Trivial example. Prove that $\sin(x+a) = \sin(x)\cos(a) + \cos(x)\sin(a)$. Set $f(x) := \sin(x+a) - \sin(x)\cos(a) - \cos(x)\sin(a)$, we have to show that $f(x) \equiv 0$.

First solution.

$$f(x) = \frac{e^{i(x+a)} - e^{-i(x+a)}}{2i} - \frac{e^{ix} - e^{ix}}{2i}\frac{e^{ia} + e^{-ia}}{2} - \frac{e^{ix} + e^{ix}}{2}\frac{e^{ia} - e^{ia}}{2i}$$
$$= \cdots = 0.$$

Second solution. $f(x)$ satisfies $(D^2+1)f = 0$, and $f(0) = 0$, $f'(0) = 0$, so $f(x) \equiv 0$.

2.1.2. C-finite sequences

Analogous considerations apply to the class of *sequences* that are solutions of some (homogeneous) ordinary linear *recurrence* equation with constant coefficients:

$$P(E)a \equiv 0, \quad P \text{ a polynomial}, \tag{2.6}$$

where E is the shift operator: $Ea_n := a_{n+1}$. For example, the Fibonacci numbers $\{F_n\}$ constitute such a sequence since $(E^2 - E - 1)F \equiv 0$.

Let us call such sequences *C-finite sequences*. It is well known that a sequence is C-finite if and only if it can be written in the form

$$a_n = \sum_{r=1}^{l} p_r(n) z_r^n, \tag{2.7}$$

where the z_r are the roots of $P(z) = 0$ and $\deg p_r = \operatorname{mult}(z_r) - 1$.

Everything we said before has its discrete analog. In particular, we can define the ideal in $C[E]$:

$$I_a := \{P \in C[E];\ P(E)a = 0\},$$

and the notion of C-finite sequences can be defined by either one of the following four equivalent definitions.

Definition 0. A sequence a is *C-finite* if there exists a polynomial P such that $P(E)a = 0$.

Definition 1. A sequence a is *C-finite* if the vector space $C[E]a = \operatorname{span}\{E^i a;\ i \geqslant 0\}$ is finite dimensional.

Definition 1′. A sequence a is *C-finite* if the $C[E]$ module $C[E]/I_a$ is a finite-dimensional vector space.

Definition 2. A sequence a is *C-finite* if it is of the form (2.7).

Definition 3. A sequence a is *C-finite* if

$$V_a := \{z \in \mathbb{C};\ P(z) = 0 \text{ for every } P \text{ in } I_a\}$$

is a finite set of points (i.e., a "zero-dimensional variety").

It follows similarly that the class of C-finite sequences is an algebra and that every identity involving sums and products of C-finite sequences is routinely verifiable.

Trivial example. Prove Cassini's identity

$$F_{n+1}F_{n-1} - F_n^2 = (-1)^n,$$

where F_n are the Fibonacci numbers defined by $F_0 = 0$, $F_1 = 1$ and $F_{n+2} = F_{n+1} + F_n$. Set $a(n) := F_{n+1}F_{n-1} - F_n^2 - (-1)^n$, we have to prove that $a(n) \equiv 0$.

First solution. Express F_n in the form (2.7), in terms of the "golden ratio", and do the trivial, but tedious algebra.

Second solution. It can be easily seen that both F_n^2 and $F_{n+1}F_{n-1}$ satisfy the same third-order recurrence equation and therefore $a(n)$ satisfies a certain fourth-order linear recurrence equation with constant coefficients, whose exact form is irrelevant. It follows that it is enough to check the four initial conditions $a(0) = 0$, $a(1) = 0$, $a(2) = 0$, $a(3) = 0$, which is easily done by inspection.

2.2. Multi-C-finite functions and sequences

2.2.1. Multi-C-finite functions

What is the several-variable analog of C-finite functions? The straightforward analog of Definition 0 is "a solution of a linear partial differential equation with constant coefficients". However, it is soon realized that all the other definitions do *not* go through. Furthermore, the "finiteness property" is not preserved. In order to specify, say, a solution of $(D_1^2 + D_2^2)f = 0$, one needs to state the "Dirichlet boundary conditions" on the boundary of a closed region, i.e., one has to furnish an infinite amount of information to specify such a function.

The fact that every solution of an ordinary linear differential equation with constant coefficients is a linear combination of exponential polynomial solutions has a celebrated analog in several variables. It is called the *Ehrenpreis–Palamadov* [17,43] *theorem*. This theorem implies that with an appropriate definition of convergence, every solution of a constant-coefficients linear partial differential equation can be expressed as an (infinite) "linear combination" of exponential polynomial solutions. This theorem generalizes to so-called *overdetermined systems*

$$P_i(D_1,\dots,D_n)f \equiv 0, \quad i = 1,\dots,L, \tag{2.8}$$

where $P_i(D_1,\dots,D_n)$ are polynomials in $D_1,\dots,D_n$. The Ehrenpreis–Palamadov theorem implies that every solution of (2.8) can be expressed as a (usually infinite) "linear combination" of exponential polynomial solutions.

An exponential polynomial solution $f(x) = p_\lambda(x)e^{\lambda x} = p_\lambda(x)e^{\lambda_1 x_1 + \cdots + \lambda_n x_n}$ is a solution of (2.8) only if

$$P_i(\lambda_1,\dots,\lambda_n) = 0, \quad i = 1,\dots,L.$$

Thus the Ehrenpreis–Palamadov theorem asserts that every solution of the system (2.8) is a (usually infinite) "linear combination" of exponential polynomial functions $p_\lambda(x)\, e^{\lambda x}$, where λ ranges over the *algebraic variety*

$$V := \{\lambda \in \mathbb{C}^n;\ P_i(\lambda) = 0,\ i = 1,\dots,L\}.$$

Guided by Definition 2 of C-finite functions, we want the "usually infinite linear combination" to be a good old-fashioned finite linear combination, and thus make our entities "specifiable by a finite amount of information". This prompts us to define multi-C-finiteness as follows.

Definition 2. f is *multi-C-finite* if it can be written as

$$f(x) = \sum_{r=1}^{R} p_r(x)\, e^{\lambda^{(r)}x}. \tag{2.9}$$

In this case the variety V consist of a finite set of points $\{\lambda^{(1)},\dots,\lambda^{(R)}\}$, and so is a "zero-dimensional variety". Thus we have a natural analog of Definition 3: f is C-finite if it is a solution of a system (2.8) whose associated variety V_f is zero-dimensional. More precisely, let us define, as before,

$$I_f := \{P \in C[D_1,\dots,D_n];\ Pf = 0\}, \tag{2.10}$$

and

$$V_f := \{\lambda \in \mathbb{C}^n;\ P(\lambda) = 0 \text{ for every } P \in I_f\}; \tag{2.11}$$

then we have the next definition.

Definition 3. $f: \mathbb{R}^n \to \mathbb{C}$ is *multi-C-finite* if the algebraic variety V_f is zero-dimensional, i.e., is a finite set of points.

Note that for an arbitrary f, I_f is usually the 0 ideal and then V_f is the whole of $\mathbb{C}^n$, i.e., it is an "n-dimensional variety". In general, the dimension of V_f is between 0 and n, and being multi-C-finite means that the function satisfies as many independent constant-coefficients linear partial differential equations as possible. Thus being multi-C-finite is the same as being a solution of a "maximally overdetermined system". Of course, we have again excluded the trivial case of $f \equiv 0$, in which case $I_f = C[D]$ and V_f is the empty set.

What about the analog of Definition 0? It is easily seen that now there are n differential operators $P_1, \ldots, P_n$, each of which is "ordinary" in $D_1, \ldots, D_n$, respectively, such that

$$P_i(D_i)f = 0, \quad i = 1, \ldots, n. \tag{2.12}$$

They are given by

$$P_i(D_i) := \prod_{r=1}^{R} \left(D_i - \lambda_i^{(r)}\right)^{\deg(p_r)+1}, \tag{2.13}$$

where

$$\lambda^{(r)} = \left(\lambda_1^{(r)}, \ldots, \lambda_n^{(r)}\right).$$

Let $P_i(D_i)$ be of order α_i. It follows from (2.12) that $D_i^{\alpha_i}f$ can be expressed as a linear combination of $f, D_i f, \ldots, D_i^{\alpha_i - 1}f$, and, by iterating, $D_i^j f$ can be thus expressed for any $j \geq \alpha_i$. Similarly, it is readily seen that every $D_1^{i_1} \cdots D_n^{i_n} f$ can be expressed as a linear combination of such entities for which $i_1 < \alpha_1, \ldots, i_n < \alpha_n$. We thus have the following definitions.

Definition 1. f is *multi-C-finite* if $C[D_1, \ldots, D_n]f$ is finite-dimensional.

Definition 1′. f is *multi-C-finite* if the $C[D_1, \ldots, D_n]$ module $C[D_1, \ldots, D_n]/I_f$ is a finite-dimensional vector space.

When we talked before about the "dimension of the variety", we meant the usual, analytical dimension, as a complex manifold. However, there are several other, algebraic, notions of dimension, for example the Krull dimension. The notion of dimension that was used by Bernstein, and which is the one that we will need, is that of the *Hilbert dimension of a graded* $C[z_1, \ldots, z_n]$ *module*, which we shall now introduce.

The ring of polynomials $C[z_1, \ldots, z_n]$ with coefficients in $\mathbb{C}$ has a natural "filtration":

$$C[z_1, \ldots, z_n] = \bigcup_{\nu} F_\nu, \tag{2.14}$$

where F_ν is the vector space of all polynomials of (total) degree $\leq \nu$. Since a basis for this vector space is $\{z_1^{\alpha_1} \cdots z_n^{\alpha_n};\ \alpha_i \geq 0,\ \alpha_1 + \cdots + \alpha_n \leq \nu\}$, its dimension is the number of solutions in nonnegative integers of $\alpha_1 + \cdots + \alpha_n \leqslant \nu$, which is well known, and easily seen, to be $(n+\nu)!/n!\nu!$. This is a polynomial in ν of degree n with leading coefficient $1/n!$.

A *graded* $C[z_1,\ldots, z_n]$ *module* is a $C[z_1,\ldots, z_n]$ module S which has the decomposition

$$S = \oplus S(\nu),$$

where $S(\nu)$ are $\mathbb{C}$-subspaces of S and $x_j S(\nu) \subset S(\nu+1)$ for $1 \leqslant j \leqslant n$ and all ν. A famous theorem of Hilbert (e.g., [12, p.7]) states that for $\nu \gg 0$, $\Sigma_{j \leqslant \nu} \dim_C(S(j))$ is always a polynomial in ν with rational coefficients (called the *Hilbert polynomial*). The degree of this polynomial is called the *Hilbert dimension* of S and is a measure of "how big" S is. Of course $C[z_1,\ldots, z_n]$ itself is a graded $C[z_1,\ldots, z_n]$ module with

$$S(j) = \operatorname{span}\{ z_1^{\alpha_1} \cdots z_n^{\alpha_n};\ \alpha_1 + \cdots + \alpha_n = j \}.$$

We saw that

$$\sum_{j \leqslant \nu} \dim(S(j)) = \dim(F_\nu) = \binom{n+\nu}{n},$$

a polynomial of degree n in ν. So the Hilbert dimension of $C[z_1,\ldots, z_n]$ coincides with its natural dimension n.

More generally, for any ideal I in $C[z_1,\ldots, z_n]$, $C[z_1,\ldots, z_n]/I$ has a natural structure as a graded $C[z_1,\ldots, z_n]$ module, inherited from (2.14) by modding out by I. If the ideal I is the 0-ideal, then the Hilbert dimension is n. Excluding this trivial case, the dimension is between 0 and $n-1$. The Hilbert dimension of $C[z_1,\ldots, z_n]/I$ is zero if and only if it is a finite-dimensional $\mathbb{C}$-vector space, because then $\Sigma_{j \leqslant \nu} \dim_C(S(j))$ is eventually constant. We thus have the next definition.

Definition 1″. f is *multi-C-finite* if the graded $C[z_1,\ldots, z_n]$ module $C[z_1,\ldots, z_n]/I_f$ has Hilbert dimension 0, which is the smallest possible.

Of course we have excluded $f \equiv 0$ for which the Hilbert polynomial is identically 0, which is a "polynomial of degree $-\infty$".

It can be seen in a variety of ways that the class of multi-C-finite functions is an algebra. The easiest way is by using Definition 2. However, it is also possible to use Definition 1, like we did with ordinary C-finite functions. $C[D_1,\ldots, D_n](f+g)$ is a subset of $C[D_1,\ldots, D_n]f \oplus C[D_1,\ldots, D_n]g$ and $C[D_1,\ldots, D_n](fg)$ is a subspace of a homomorphic image of $C[D_1,\ldots, D_n]f \otimes C[D_1,\ldots, D_n]g$ and so are finite-dimensional vector spaces.

Another obvious property, which we will generalize later on, is that the *diagonal* $\bar{f}(\bar{x})$ of a multi-C-finite f: $\bar{f}(\bar{x}) := f(\bar{x},\ldots, \bar{x})$ is an ordinary C-finite function in the single variable $\bar{x}$. This is immediate from Definition 2 but is also not hard to deduce using any of the other definitions. For example, from Definition 1 it follows that the vector space $\operatorname{span}\{ D_1^r \cdots D_n^r f;\ r \geqslant 0\}$ is finite-dimensional and hence there is an operator $\bar{P}(\bar{D})$ in $\bar{D} := D_1 \cdots D_n$ that annihilates $f(x_1,\ldots, x_n)$, and thus $P(D_{\bar{x}})\bar{f}(\bar{x}) = 0$.

2.2.2. *Multi-C-finite sequences*

Everything in the previous section has an obvious discrete analog. Now we have *multisequences* that are nothing but functions from $\mathbb{Z}^n$ to $\mathbb{C}$, or $\mathbb{N}^n$ to $\mathbb{C}$, or for that matter one can consider functions from any subset of $\mathbb{Z}^n$ to $\mathbb{C}$. Given a function $a:\mathbb{Z}^n \to \mathbb{C}$, we define the fundamental shift operators

$$E_i a(m_1,\dots,m_i,\dots,m_n) := a(m_1,\dots,m_i+1,\dots,m_n). \tag{2.15}$$

A *linear partial recurrence operator with constant coefficients* is a polynomial $P(E_1,\dots,E_n)$ in the fundamental shift operators. For example, the operator $a(m_1, m_2) \to a(m_1, m_2) + 3a(m_1 + 1, m_2 + 2)$ is written $I + 3E_1E_2^2$. The ring of linear partial recurrence operators with constant coefficients is in fact $C[E_1,\dots, E_n]$, the commutative polynomial algebra in the n indeterminates $E_1,\dots, E_n$ over the field $\mathbb{C}$. We leave it to the readers to do the obvious discrete analog of the previous subsection, and give the definitions of the notion "multi-C-finite" as applied to discrete functions. For example, the next definition.

Definition 1. $a:\mathbb{Z}^n \to \mathbb{C}$ is *multi-C-finite* if $C[E_1,\dots, E_n]a$ is a finite-dimensional vector space.

2.2.3. Mixed continuous and discrete functions

Consider functions $f:\mathbb{R}^{n_1} \times \mathbb{Z}^{n_2} \to \mathbb{C}$. Then $C[D_1,\dots, D_{n_1}, E_1,\dots, E_{n_2}]$ acts in a natural way on these functions, and everything goes through (except the definition of the diagonal).

2.3. P-finite functions and sequences

If, in Definition 0 of C-finite functions, we replace "a solution of a *constant*-coefficients linear ordinary differential equation" by "a solution of a *polynomial*-coefficients linear ordinary differential equation" we get the class of *P-finite functions*. Similarly, a *P-finite sequence* is a sequence that is a solution of a (homogeneous) linear ordinary *recurrence* equation with polynomial coefficients. For example, the Apery numbers

$$a_n := \sum_{k=0}^{n} \binom{n}{k}^2 \binom{n+k}{k}^2$$

of $\zeta(3)$ fame [55] is a P-finite sequence since it satisfies

$$n^3 a_n - (34n^3 - 51n^2 + 27n - 5)a_{n-1} + (n-1)^3 a_{n-2} = 0.$$

P-finite functions and sequences were introduced and studied by Stanley [50]. Please note that our P-finite functions are called by him "D-finite functions", and our P-finite sequences are called by him "P-recursive sequences". In the interest of thawing the cold war between the discrete and the continuous, I have decided to combine these two names into one.

So far we have introduced P-finite functions through the analog of Definition 0. Do the other definitions go through? Unfortunately, Definition 2 does *not* go through, but all the other definitions do. For example, the following definition.

Definition 1. f is *P-finite* if $C[D]f$ is a finite-dimensional vector space over the field of rational functions $C(x)$.

The analogs of Definitions 1′, 1″ and 3 will be discussed in the next subsection, in the more general context of holonomic functions.

Both P-finite functions and P-finite sequences are closed under addition and multiplication [50]. This is easily seen from Definition 1, since the $C(x)$ vector space $C[D](f+g)$ is a subspace of $C[D]f \oplus_{C(x)} C[D]g$ and the $C(x)$ vector space $C[D](fg)$ is a subspace of a homomorphic image of $C[D]f \otimes_{C(x)} C[D]g$, and thus are finite-dimensional over $C(x)$.

Once again, the discrete analog of this is immediate.

2.4. Holonomic systems and holonomic functions

After this somewhat lengthy introduction we are finally ready to introduce holonomic systems, that form the foundation of our approach to special function identities.

Holonomic systems were introduced and studied by Bernstein [10,11] and were used by him to give an elementary proof of a famous conjecture of Gelfand concerning the existence of a meromorphic extension of the distribution valued complex function $\lambda \to P^{\lambda}$, where P is a polynomial of several variables in $\mathbb{R}^n$. This theory is very deep and stands at the forefront of current research in analysis. Fortunately, we will only need the most basic notions from this imposing theory, and everything we need will be stated and explained. An excellent introduction to the theory of holonomic systems can be found in the introduction and Chapters 1 and 7 of Björk's monograph [12]. Much of the following follows closely Björk's readable account on [12, pp. ix and x].

Holonomic systems are obtained when the linear partial differential operators with *constant* coefficients that feature in the definition of multi-C-finite functions are replaced by linear partial differential operators with *polynomial* coefficients. Thus instead of the commutative algebra $C[D_1, \ldots, D_n]$ we will have to consider the so-called *Weyl algebra* $A_n(C) = C\langle D_1, \ldots, D_n, x_1, \ldots, x_n \rangle$ which is the (noncommutative) algebra generated by the indeterminates $D_1, \ldots, D_n$, $x_1, \ldots, x_n$. Here x_i denotes the operator of multiplication by $x_i : f \to x_i f$. These generators satisfy the following commutation relations: $[x_i, x_j] = 0$, $[D_i, D_j] = 0$, and from the product rule of differentiation $[D_i, x_j] = \delta_{i,j}$, where $\delta_{i,j}$ is Kronecker's delta function that equals 1, when $i = j$, and equals 0, otherwise; (recall that "1" means the identity operator: $f \to f$).

The noncommutative structure is already so strong that the ring $A_n(C)$ has no two-sided ideals, except the zero-ideal and the whole ring $A_n(C)$. But of course there are one-sided ideals, left or right, and more generally left and right $A_n(C)$-modules.

A remarkable result, due to Stafford [49], [12, (1.7)], asserts that every left ideal in $A_n(C)$ can be generated by 2 elements.

In particular $A_n(C)$ is a left and right noetherian ring. The main idea is to use *filtrations* on the ring $A_n(C)$ and its modules.

The ring $A_n(C)$ is equipped with the F-filtration $= \{F_\nu\}$, where F_ν is the finite-dimensional vector space generated by the (x, D)-monomials $x^{\alpha} D^{\beta}$ whose total weight $|\alpha| + |\beta| \leqslant \nu$.

Thus $F_0 \subset F_1 \subset F_2 \subset \cdots$ is an increasing sequence of subspaces of $A_n(C)$. Of course $\bigcup F_\nu = A_n(C)$ and the inclusion $F_\nu F_k \subset F_{\nu+k}$ for all pairs of nonnegative integers ν and k, follows from the definition of the ring product of $A_n(C)$.

Let us now consider a left ideal L in the ring $A_n(C)$. If $\nu \geqslant 0$, we can consider the dimensions of the complex vector spaces $F_\nu/(F_\nu \cap L)$, which form the induced filtration on the left $A_n(C)$-module $A_n(C)/L$. Let $H(\nu)$ be the dimension of $F_\nu/F_\nu \cap L$. Then $H(\nu) = \dim_C(F_\nu) -$

$\dim_C(F_\nu \cap L)$. It turns out [12, (1.3)] that $H(\nu)$ is a *polynomial function* $a_d\nu^d + \cdots + a_0$ when $\nu \gg 0$. Here $a_0, \dots, a_d$ are rational numbers and $(d!)a_d$ is a positive integer that is denoted by e.

The degree d of the Hilbert polynomial $a_d\nu^d + \cdots a_0$, which computes $H(\nu)$ when $\nu \gg 0$, is called the F-dimension of the left $A_n(C)$-module $A_n(C)/L$ and is denoted by $d_F(A_n/L)$.

The celebrated *Bernstein inequality* asserts that for every left ideal L, $d_F(A_n/L) \geqslant n$, unless, of course L is the whole ring A_n.

The above notion of dimension can be extended to general finitely-generated left $A_n(C)$-modules of which modules of the form $A_n(C)/L$ are special cases. One considers a certain class of "good filtrations", and shows that the notion of dimension makes sense for each such good filtration. Then it is shown that the dimension is independent of the particular filtration, and thus one is able to talk about the *dimension of the $A_n(C)$-module M, $d(M)$*, without reference to any particular filtration. Bernstein [10], [12, (1.4)] proved the more general statement that $d(M) \geqslant n$ for every nonzero and finitely-generated left $A_n(C)$-module. An extremely short and elegant proof of this important inequality was given by Antony Joseph, and can be found in [13], and in [21, p.417].

Let f be either a C^∞ function, or a distribution, or a formal power series, or what have you (anything on which it is possible to differentiate and multiply by x_i), in n variables. Then one can consider all the elements of $A_n(C)$ that annihilate f:

$$I_f := \{P \in A_n(C);\ Pf = 0\}. \tag{2.16}$$

Obviously I_f is a left ideal. By Bernstein's inequality the dimension of the $A_n(C)$-module is $\geqslant n$, unless f is identically zero. We will call functions f for which the dimension of $A_n(C)/I_f$ is the smallest possible, namely *n, holonomic functions* (or distributions, formal power series, etc.). It can be easily seen that multi-C-finite functions are ipso facto holonomic functions. This makes sense since in the passage from $C[D_1, \dots, D_n]/I_f$ to $A_n(C)/I_f$ we have acquired n "extra dimensions", corresponding to the generators $x_1, \dots, x_n$, and the "gap" between numerator and denominator is still n.

Because of the analogy to multi-C-finite functions and P-finite functions it would make sense to call holonomic functions "multi-P-finite functions". Björk [12] calls holonomic functions "members of the Bernstein class".

The Weyl algebra $A_n(C)$ acts just as well on discrete functions $f: \mathbb{Z}^n \to \mathbb{C}$, via $x_i \to E_i^{-1}$ and $D_i \to (n_i + 1)E_i$, as can be easily seen by examining the action of x_i and D_i on a generic formal power series, and see how the coefficients get transformed. Of course this extends to mixed continuous-discrete functions $f(x_1, \dots, x_n, m_1, \dots, m_{n'})$ on which the Weyl algebra $A_{n+n'}(C)$ acts naturally.

So let us state the formal definition of *holonomic functions (or distribution, formal power series, etc.)*.

Definition 1″. Let f be a nonzero member of a family on which the Weyl algebra acts naturally. Let the ideal I_f be defined by (2.16) above. Then f is *holonomic* if the left $A_n(C)$-module $A_n(C)/I_f$ has its smallest possible dimension, namely n.

It is readily seen that this definition is the natural extension of Definition 1″ of multi-C-finite functions, in Section 2.2.1. Since $A_n(C)/I_f$ is naturally isomorphic, as $A_n(C)$ modules, to the $A_n(C)$-module $A_n(C)f$. We can also state the next definition.

Definition 1′. f is *holonomic* if the $A_n(C)$-module $A_n(C)f$ has the smallest possible dimension, namely n.

Definition 3 can also be extended to the present context. But we first need some notation. A typical element of the Weyl algebra $A_n(C)$ may be written

$$P = \sum_{|\alpha| \leqslant N} a_\alpha(x) D^\alpha,$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)$, $D^\alpha := D_1^{\alpha_1} \cdots D_n^{\alpha_n}$ and $a_\alpha(x) = a_\alpha(x_1, \ldots, x_n)$ are polynomials. Its *symbol* is obtained by replacing D_i by the complex variables ζ_i, and replacing x_i by the complex variables z_i, obtaining a polynomial in the $2n$ complex variables $z_1, \ldots, z_n, \zeta_1, \ldots, \zeta_n$. The *principal symbol* $\sigma(P)$ is obtained by only retaining the leading terms for which $|\alpha| = N$:

$$\sigma(P) = \sum_{|\alpha| = N} a_\alpha(z) \zeta^\alpha.$$

For any left ideal L in $A_n(C)$ let us consider the following variety in $\mathbb{C}^n \times \mathbb{C}^n = \mathbb{C}^{2n}$:

$$V_L = \{(z, \zeta) \in \mathbb{C}^n \times \mathbb{C}^n;\ \sigma(Q)(z, \zeta) = 0, \text{ for every } Q \text{ in } L\}$$
$$= \bigcap \{\sigma(Q)^{-1}(0);\ Q \in L\}.$$

It is a deep result (see [12, p. x and Chapter 2]) that the (complex analytic) dimension of the variety V_L coincides with the previously defined dimension of the $A_n(C)$-module $A_n(C)/L$, for any left ideal L in $A_n(C)$. In particular we can define the following.

Definition 3. Let f be a nonzero member of a family on which the Weyl algebra acts naturally. Then f is *holonomic* if the variety

$$V_f := \{(z, \zeta) \in \mathbb{C}^n \times \mathbb{C}^n;\ \sigma(Q)(z, \zeta) = 0, \text{ for every } Q \text{ such that } Q(f) = 0\}$$
$$= \bigcap \{\sigma(Q)^{-1}(0);\ Q(f) = 0\}$$

has dimension n.

2.5. Examples of holonomic functions

In the next section we will show that the class of holonomic functions is an algebra: The sum and product of holonomic functions are also holonomic. We will also prove that it is closed under integration (or summation) with respect to a variable: If $f(x_1, \ldots, x_n)$ is a holonomic function of n variables, then $\int_{-\infty}^{\infty} f \, dx_n$ is holonomic in the surviving $n-1$ variables $x_1, \ldots, x_{n-1}$, if it is defined. Similarly if $f = f(m_1, \ldots, m_n) : \mathbb{Z}^n \to \mathbb{C}$ is holonomic, then $\Sigma_{m_n} f$ is holonomic in the surviving variables $m_1, \ldots, m_{n-1}$.

All polynomials and exponential-polynomial functions, being multi-C-finite, are automatically holonomic. The simplest discrete function of a single variable that is P-finite but not C-finite is $f(n) := n!$, that satisfies $(E - n - 1)f = 0$. Similarly $f(n) := 1/n!$ is P-finite since it satisfies $((n+1)E - 1)f = 0$.

The function $f := m_1!$ is a holonomic function of $m_1, \ldots, m_n$. One way of seeing this is noting that I_f contains the operators $I - E_i$, $2 \leqslant i \leqslant n$, as well as $E_1 - m_1 - 1$. When translating, this becomes $1-z_i$, $2 \leqslant i \leqslant m$, and $z_1 - z_1 D_1 - 1$, whose principal symbols are $-z_i$ and $-z_1\zeta_1$. The variety V_f is a subset of the set of common zeros of $\{-z_i, 2 \leqslant i \leqslant n\}$ and $z_1\zeta_1$, which is a union of the two n-dimensional varieties in C^{2n}: $\{z_1 = \cdots = z_n = 0\}$ and $\{\zeta_1 = 0, z_2 = \cdots = z_n = 0\}$, and thus its dimension must be $\leqslant n$, and, by Bernstein's inequality, it should be equal to n, and thus is holonomic.

In a similar way we can prove that $1/m_1!$ is holonomic and of course this is true for all $m_i!$ and $1/m_i!$. Similarly $(m_1 + \cdots + m_n)!$ is holonomic and thus the multinomial coefficient $(m_1 + \cdots + m_n)!/m_1! \cdots m_n!$. This would also follow from the following general result, upon taking $P = 1 - x_1 - \cdots - x_n$, and looking at the coefficients of $1/P$, viewed as a formal power series.

Proposition 2.1. *Let $P = P(x_1, \ldots, x_n)$ be a polynomial in n variables; then $1/P$ is holonomic, whenever it is defined.*

First proof. Let us try and find as many linear differential operators with polynomial coefficients that annihilate $f := 1/P$ as possible. Among the many such operators are $PD_1 + P^{(1)}$ and $P^{(i)}D_1 - P^{(1)}D_i$, for $2 \leqslant i \leqslant n$. ($P^{(i)}$ is the partial derivative of P with respect to x_i: $\partial P/\partial x_i$.) The principal symbols of these are $P\xi_1$ and $P^{(j)}\xi_1 - P^{(1)}\xi_j$. The variety V_f is a subset of the set of common zeros of these n polynomials in $x_1, \ldots, x_n$, $\xi_1, \ldots, \xi_n$, which is easily seen to comprise an n-dimensional variety. Thus the dimension of the variety V_f is $\leqslant n$, and by Bernstein's inequality it must be equal to n. It follows that $1/P$ is holonomic. □

Second proof. $A_n(C)(1/P)$ is a submodule of $C[x_1, \ldots, x_n, P^{-1}]$, and this last $A_n(C)$-module is shown to be a holonomic $A_n(C)$-module in [12, Theorem 1.5.5, p.13]. □

Since every polynomial Q is obviously holonomic it follows (from the closure under product to be proved in Section 3) that all rational functions Q/P are holonomic. If P has a nonzero constant term, then Q/P makes sense as a formal power series, and its multisequence of coefficients f, defined by

$$Q/P = \Sigma f(m_1, \ldots, m_n) x_1^{m_1} \cdots x_n^{m_n},$$

is a holonomic discrete function from $\mathbb{N}^n$ to $\mathbb{C}$.

3. Operations that preserve holonomicity

3.1. Continuous functions and distributions

We will now show that the sum and product of holonomic functions are themselves holonomic, and that if $f(x_1, \ldots, x_n)$ is holonomic on $\mathbb{R}^n$, then $f(x_1, \ldots, x_{n-1}, c)$ (c a constant) and $\int f(x_1, \ldots, x_n)\, dx_n$ are holonomic functions of $(x_1, \ldots, x_{n-1})$.

Proposition 3.1. *Let $f: \mathbb{R}^n \to \mathbb{C}$, $g: \mathbb{R}^n \to \mathbb{C}$ be holonomic; then $f+g$ is also holonomic.*

Proof. Let $A_n(C) = \bigcup_{\nu=0}^{\infty} F_\nu$ be the natural filtration of $A_n(C)$, described above. Then the natural filtrations for $A_n(C)f$, $A_n(C)g$ and $A_n(C)(f+g)$ are $\bigcup_{\nu=0}^{\infty} F_\nu f$, $\bigcup_{\nu=0}^{\infty} F_\nu g$ and $\bigcup_{\nu=0}^{\infty} F_\nu (f+g)$.

Now for every ν we have $F_\nu(f+g) \subset F_\nu f \oplus F_\nu g$. Thus $\dim F_\nu(f+g) \leqslant \dim(F_\nu f) + \dim(F_\nu g)$ are polynomials in ν of degree n, and the same is true for their sum. We do know a priori that for $\nu \gg 0$, $\dim F_\nu(f+g)$ is a polynomial, and so it follows that its degree must be $\leqslant n$, and therefore, by Bernstein's inequality it is equal to n, and $f+g$ is holonomic. □

Proposition 3.2. *If f and g are holonomic, so is their product (if it is defined).*

Proof. Let M_1 and M_2 be $A_n(C)$-modules. Bernstein [11, p.277] defines the $A_n(C)$-module $M_1 \times M_2$ as follows: as an $C[x_1, \ldots, x_n]$ module $M_1 \boxtimes M_2 = M_1 \otimes_{C[x]} M_2$ and the operators $\partial/\partial x_i$ acts as follows:

$$\frac{\partial}{\partial x_i}(f_1 \otimes f_2) = \left(\frac{\partial}{\partial x_i} f_1\right) \otimes f_2 + f_1 \otimes \left(\frac{\partial}{\partial x_i} f_1\right), \quad f_1 \in m_1, \ f_2 \in M_2.$$

Bernstein proved ([11, p.278, Theorem 3.2, Part 3) that if M_1 and M_2 are holonomic $A_n(C)$ modules, so is $M_1 \boxtimes M_2$.

Now, by Leibnitz's rule

$$A_n(C)(fg) \subset (A_n(C)f) \otimes (A_n(C)g).$$

It follows that $A_n(C)(fg)$ is a holonomic $A_n(C)$-module, since it is a nonzero submodule of a holonomic $A_n(C)$-module. □

Proposition 3.2 has some far-reaching consequences. It is obvious that the Dirac delta measure δ_a, concentrated at $x = a$ (i.e., $\delta_a(f) = f(a)$, for every test function) is a holonomic distribution in a single variable, since it satisfies the equations (of order zero) $(x-a)\delta_a = 0$. When δ_a is embedded in $\mathbb{R}^n$ as the Lebesgue measure that is supported at the hyperplane $x_n = a$, it is still holonomic, but now in $\mathbb{R}^n$. Multiplication by δ_a from the right is nothing but the evaluation $x_n = a$:

$$f(x_1, \ldots, x_{n-1}, x_n)\delta_a = f(x_1, \ldots, x_{n-1}, a),$$

when $x_n = a$, and 0 otherwise.

It follows from Proposition 3.2 that if f is holonomic and $f\delta_a$ is defined (which is the case, for example, when f is continuous), then it is holonomic. In other words, if $f(x_1, \ldots, x_{n-1}, x_n)$ is holonomic in $\mathbb{R}^n$, then $f(x_1, \ldots, x_{n-1}, a)$ is holonomic in $\mathbb{R}^{n-1}$.

It turns out [12, (7.4.1)] that it is always possible to define the evaluation $x_n = a$ on any holonomic distribution. This is certainly not possible for an arbitrary distribution, so the class of holonomic distributions is indeed privileged.

The differentiation operators $(1/i)(\partial/\partial x_j)$ and the "multiplication by x_j" operators x_j are dual to each other under the Fourier transform, so it is easily seen [12, p.293] that the Fourier transform of a holonomic entity on $\mathbb{R}^n$, if it exists (and it does if it is a tempered distribution), is also holonomic, and vice versa. The Fourier transform sends the act of multiplication into the act of convolution: $(\widehat{fg}) = \hat{f} * \hat{g}$, so we have as an immediate consequence of Proposition 3.2, the next proposition.

Proposition 3.2*. *If f and g are holonomic, so is their convolution, if it is defined.*

This proposition can be also easily proved directly, by mimicking the proof of Proposition 3.2.

Let $\delta_n^{(n-1)}$ be the Lebesgue measure of the "line" $x_1 = \cdots = x_{n-1} = 0$ in $\mathbb{R}^n$, i.e., for any test function ϕ:

$$\delta_n^{(n-1)}(\phi(x_1,\ldots,x_n)) = \int \phi(0,\ldots,0,\ x_n)\ \mathrm{d}x_n.$$

It is holonomic because it satisfies the system:

$$x_i \delta_n^{(n-1)} = 0, \quad i = 1,\ldots,n-1, \qquad D_n \delta_n^{(n-1)} = 0.$$

Now

$$\bar{f}(x_1,\ldots,x_{n-1}) := \int f(x_1,\ldots,x_{n-1},\ x_n)\ \mathrm{d}x_n \tag{3.1}$$

is nothing but the convolution $f * \delta_n^{(n-1)}$, and so it follows from Proposition 3.2* that if f is holonomic in $\mathbb{R}^n$, then $\bar{f}$, as given in (3.1) is holonomic in $\mathbb{R}^{n-1}$.

Summarizing we have:

Proposition 3.3. *Let $f = f(x_1,\ldots,x_n)$ be holonomic in $\mathbb{R}^n$, and let a be any constant; then $f(x_1,\ldots,x_{n-1},\ a)$ is holonomic in $\mathbb{R}^{n-1}$.*

Proposition 3.4. *Let $f(x_1,\ldots,x_n)$ be holonomic in $\mathbb{R}^n$; then $\int f(x_1,\ldots,x_n)\ \mathrm{d}x_n$ is holonomic in $\mathbb{R}^{n-1}$, if it is defined.*

It is obvious that the class of holonomic functions (or distributions) is closed under differentiation and indefinite integration. Furthermore, the closure under indefinite integration and evaluation implies closure under *definite integration*, so we have the following proposition.

Proposition 3.5. *If $f(x_1,\ldots,x_n)$ is holonomic, then, for any constants a and b,*

$$\int_a^b f(x_1,\ldots,x_{n-1},\ x_n)\ \mathrm{d}x_n$$

is also holonomic, if it is defined.

3.2. Discrete functions

Everything that we said about the continuous realm goes over smoothly to the discrete realm. Given a discrete function $f : \mathbb{Z}^n \to \mathbb{C}$, we can talk about its Fourier transform

$$\hat{f}(t_1,\ldots,t_n) = \sum_{m \in \mathbb{Z}^n} f(m)\ \mathrm{e}^{\mathrm{i}m_1t_1 + \cdots + \mathrm{i}m_nt_n},$$

that lives on the torus $T^n = (-\pi,\ \pi)^n$. Alternatively, by taking $\mathrm{e}^{\mathrm{i}t_j} \to z_j$, $j = 1,\ldots,n$, we get the "z-transform"

$$\hat{f}(z_1,\ldots,z_n) = \sum_{m \in \mathbb{Z}^n} f(m) z_1^{m_1} \cdots z_n^{m_n}. \tag{3.2}$$

Equation (3.2) always makes sense as a "formal Laurent series", which may be viewed as "distributions" with respect to the Laurent polynomials as "test functions". If $f(m)$ is of polynomial growth, then $\hat{f}$ is a distribution on T^n; if $f(m)$ is of exponential growth, then $\hat{f}$ is analytic in some multi-annulus that contains T^n, and if $f(m)$ is supported in $\mathbb{N}^n$, then $\hat{f}$ is a nice and honest formal power series.

Let us apply the operators $z(\mathrm{d}/\mathrm{d}z)$ and z on a single-variable formal Laurent series

$$\hat{f}(z) = \sum_{-\infty}^{\infty} f(m) z^m.$$

We get

$$z\frac{\mathrm{d}}{\mathrm{d}z}\hat{f}(z) = \sum_{-\infty}^{\infty} mf(m) z^m$$

and

$$z\hat{f}(z) = \sum_{-\infty}^{\infty} f(m) z^{m+1} = \sum_{-\infty}^{\infty} f(m-1) z^m,$$

so the z-transform sends the operator $z(\mathrm{d}/\mathrm{d}z)$ to the operator "multiplication by m", and it sends the operator "z" into the "backward shift" E^{-1} ($E^{-1}f(m) = f(m-1)$).

In several variables we have

$$z_j \leftrightarrow E_j^{-1}, \qquad z_j D_j \leftrightarrow m_j,$$

where, as always, m_j is shorthand for the operator "multiplication by m_j". A linear partial recurrence operator with polynomial coefficients is written

$$P\left(m_1,\ldots, m_n, E_1^{-1},\ldots, E_n^{-1}\right),$$

and this corresponds to the differential operator

$$P(z_1 D_1,\ldots, z_n D_n, z_1,\ldots, z_n).$$

Conversely any linear differential operator with polynomial coefficients can be written as

$$z_1^{-a_1} \cdots z_n^{-a_n} P\left(z_1 D_1,\ldots, z_n D_n, z_1,\ldots, z_n\right),$$

for some nonnegative integers $a_1,\ldots, a_n$. So it is obvious that $f(\boldsymbol{m})$ is holonomic if and only if $\hat{f}(z)$ is holonomic. In fact this last sentence is a tautology since we have basically defined $f(\boldsymbol{m})$ to be holonomic when $\hat{f}(z)$ is holonomic.

The same reasoning as in the continuous case yields.

Proposition 3.1′. *The sum of two holonomic discrete functions from (a subset of) $\mathbb{Z}^n$ to $\mathbb{C}$ is also holonomic.*

Proposition 3.2′. *The product of two holonomic discrete functions from (a subset of) $\mathbb{Z}^n$ to $\mathbb{C}$ is also holonomic.*

Proposition 3.2*′. *If f and g are holonomic discrete functions, so is their convolution, if it is defined.*

The characteristic function of the discrete hyperplane $m_n = a$ is clearly holonomic, so we have the following proposition.

Proposition 3.3′. *Let $f = f(m_1,\ldots, m_n)$ be holonomic in $\mathbb{Z}^n$, and let a be any constant; then $f(m_1,\ldots, m_{n-1}, a)$ is holonomic in $\mathbb{Z}^{n-1}$.*

The characteristic function of the discrete line $m_1 = \cdots = m_{n-1} = 0$ is obviously holonomic, so we have the next proposition.

Proposition 3.4′. *Let $f(m_1, \ldots, m_n)$ be holonomic in $\mathbb{Z}^n$; then*

$$\sum_{m_n} f(m_1, \ldots, m_{n-1}, m_n)$$

is holonomic in $\mathbb{Z}^{n-1}$, if it is defined.

Since the property of being holonomic is preserved under addition we have the next proposition.

Proposition 3.5′. *If $f(m_1, \ldots, m_n)$ is holonomic in $\mathbb{Z}^n$, then, for any constants a and b,*

$$\sum_{m_n=a}^{b} f(m_1, \ldots, m_{n-1}, m_n)$$

is holonomic in $\mathbb{Z}^{n-1}$ (in the surviving variables $m_1, \ldots, m_{n-1}$).

3.2.1. The diagonal of a holonomic formal power series is D-finite

If

$$\hat{f}(z_1, \ldots, z_n) = \sum_{m \in \mathbb{N}^n} f(m_1, \ldots, m_n) z_1^{m_1} \cdots z_n^{m_n}$$

is a formal power series, we define the *diagonal* as

$$\hat{f}_{\mathrm{D}}(w) = \sum_{m} f(m, \ldots, m) w^m.$$

In other words, we restrict $f(m_1, \ldots, m_n)$ to the diagonal and replace $z_1 \cdots z_n$ by a new variable w.

The z-transform of the characteristic set of the diagonal $\{(m, \ldots, m),\ m \geqslant 0\}$ is the rational formal power series $1/(1-(z_1 \cdots z_n))$, which, by Proposition 2.1 (Section 2.5) is holonomic (this can be also checked directly). If $\hat{f}$ is a holonomic formal power series, then f itself must be a holonomic discrete function and it follows by Proposition 3.2′ that the product of f with the characteristic function of the diagonal $\{(m, \ldots, m),\ m \geqslant 0\}$ is holonomic. But this product is the restriction $f(m, \ldots, m)$ of f to the diagonal, and so $f(m, \ldots, m)$ is a single variable holonomic discrete function (what Stanley [50] calls P-recursive), and its z-transform $\hat{f}_{\mathrm{D}}(w)$ must be a holonomic formal power series in one variable, what Stanley [50] calls D-finite.

By Proposition 2.1, a rational formal power series $Q(z_1, \ldots, z_n)/P(z_1, \ldots, z_n)$ ($P(0, \ldots, 0) \neq 0$) is holonomic, so it follows that the diagonal of a rational formal power series is D-finite. This fact was conjectured by Stanley [50], and incompletely proved in [60] and [29]. An elementary proof was given by Lipshitz [38].

3.2.2. Mixed continuous and discrete functions

Our discussion extends immediately to mixed continuous-discrete functions $f : \mathbb{R}^{n_1} \times \mathbb{Z}^{n_2} \to \mathbb{C}$ (in fact the continuous component may be a distribution, i.e., $f(\cdot, m)$ may be a distribution on $\mathbb{R}^{n_1}$ for every m in $\mathbb{Z}^{n_2}$).

Holonomic entities on $\mathbb{R}^{n_1} \times \mathbb{Z}^{n_2}$ are closed under addition, and if well-defined, also under multiplication and convolution.

The same argument as above shows that if $f: \mathbb{R}^{n_1} \times \mathbb{Z}^{n_2} \to \mathbb{C}$ is holonomic, then $f(x_1, \ldots, x_{n_1-1}, b, m_1, \ldots, m_{n_2})$ and $\int f \, dx_{n_1}$ are holonomic in $\mathbb{R}^{n_1-1} \times \mathbb{Z}^{n_2}$ and $f(x_1, \ldots, x_{n_1}, m_1, \ldots, m_{n_2-1}, a)$ and $\Sigma_{m_{n_2}} f$ are holonomic as functions on $\mathbb{R}^{n_1} \times \mathbb{Z}^{n_2-1}$.

3.3. Applications to combinatories

The general problem in enumerative combinatorics is that of *counting* the number of elements $a(n)$ in a family $A(n)$ of combinatorial objects, parametrized by the (single or multi) discrete variable n. For example, the number of subsets of an n-element set is 2^n. Of course it is nice to get a simple explicit expression for $a(n)$, but this is very rarely possible. But what is an *explicit* answer [57]? One narrow definition is that $a(n+1)/a(n)$ be a rational function in n. In other words, we know the answer explicitly if $a(n)$ satisfies a first-order recurrence equation with polynomial coefficients. One way of broadening our definition of "explicit" is not to insist on the recurrence being first order. In other words, given a combinatorial sequence $a(n)$, to which there is no simple explicit form, it is of interest to know whether it is *P-finite*, i.e., is a solution of a linear recurrence equation with polynomial coefficients (see [50]).

We saw in Section 2.5 that the factorial and multinomial coefficients are holonomic in their variables. Many times in combinatorics, there is an explicit expression for $a(n)$ in terms of a huge sum of summands that are obviously holonomic. The general scenario is that the summand depends on, say, k parameters, and the sum contains $k-1$ sigmas. It follows then from an iteration of Propositions 3.4$'$ and/or 3.5$'$ that $a(n)$ is holonomic in the surviving variable n. But being holonomic in a single variable is nothing but being P-finite! We have thus a powerful way of proving that sequences of combinatorial interest are P-finite.

One example is $t_k(n)$, the number of standard Young tableaux with n cells and height $\leqslant k$. For every fixed k, this is a sum of f_λ over the set

$$\{(m_1, \ldots, m_k);\ m_1 + \cdots + m_k = n,\ m_1 \geqslant m_2 \geqslant \cdots \geqslant m_k \geqslant 0\}.$$

f_λ is the number of standard Young tableaux of shape $\lambda = (m_1, \ldots, m_k)$, and is given explicitly by the famous Young–Frobenius formula:

$$f_\lambda = \left[\prod_{1 \leqslant i < j \leqslant k} (m_i - m_j + j - i)\right] \frac{(m_1 + \cdots + m_k)!}{(m_1 + k - 1)! \cdots (m_k)!}.$$

Since f_λ is obviously holonomic in its variables, it follows that for every fixed k, $t_k(n)$ is P-finite. (Explicit expressions for $t_k(n)$ for small values of k were given by Regev [46] ($k = 3$) and Gouyou-Beauchamps [35] ($k = 4,5$), and Goulden [34] developed an interesting symmetric functions method for handling these and related sums.)

Many other combinatorial applications were found by Gessel [30,31]. Lipshitz [38,39] gave elegant elementary proofs of many of the results in this section for D-finite formal power series.

4. How to "write down" holonomic functions and how to verify identities

4.1. Continuous holonomic functions

In order to actually work with holonomic functions we should have some way to "write them down". A holonomic function is completely determined by a left ideal J of $A_n(C)$ that annihilates it, and by "initial conditions". We will now show that every holonomic function in n variables has n "ordinary" operators $P_i(D_i, x_1, \ldots, x_n)$, $i = 1, \ldots, n$, that annihilate it. This would follow from the following lemma, whose proof was kindly shown to me by Joseph Bernstein.

Lemma 4.1. *Let L be a left ideal in $A_n(C)$ such that $A_n(C)/L$ is a holonomic $A_n(C)$-module. For every $n+1$ generators out of the $2n$ generators $\langle x_1, \ldots, x_n, D_1, \ldots, D_n\rangle$ of $A_n(C)$ there is a nonzero member of L that only depends on these $n+1$ generators.*

Proof. For the sake of definiteness let us take the $n+1$ generators to be $x_1, \ldots, x_n, D_1$. The proof in general is similar. Consider the mapping

$$\phi : C\langle x_1, \ldots, x_n, D_1\rangle \to A_n(C)/J,$$

given by $\phi(f) = f \pmod J$. Let $\bigcup F_\nu^{(0)}$ and $\bigcup F_\nu/J$ be the natural filtrations of $C\langle x_1, \ldots, x_n, D_1\rangle$ and $A_n(C)/J$, respectively. (Recall that F_ν is the vector space of all members of $A_n(C)$ of degree $\leqslant \nu$.) Since $A_n(C)/J$ is holonomic, we have that $\dim(F_\nu/J)$ is a polynomial in ν of degree n for $\nu \gg 0$. Obviously $\dim(F_\nu^{(0)})$ is a polynomial of degree $n+1$, so it follows that there is a ν such that $\dim(F_\nu^{(0)}) > \dim(F_\nu/J)$. The restriction of the linear map ϕ to the finite-dimensional vector space $F_\nu^{(0)}$ is therefore a linear transformation from a higher-dimensional vector space to a lower-dimensional vector space, and its kernel must therefore be nonzero. But this kernel is precisely $J \cap C\langle x_1, \ldots, x_n, D_1\rangle$. $\square$

The above lemma shows that it is always possible to eliminate any $n-1$ of the generators. The above proof is an *existence* proof. In Section 5, I will present an algorithm that actually constructs the eliminated operators.

In order to write down a holonomic function f in "holonomic notation", we can give any set of generators of its corresponding ideal, with the appropriate initial conditions. However, in general it is not clear how many initial conditions are required to uniquely specify the function f. It is therefore necessary to introduce what I call "canonical holonomic representation" as follows. We find the n "ordinary" operators guaranteed by the above lemma

$$P_i(D_i, x_1, \ldots, x_n), \quad i = 1, \ldots, n, \tag{4.1a}$$

of order α_i (in D_i), say, that annihilate f, and give the "initial conditions"

$$D_1^{i_1} \cdots D_n^{i_n} f(x_0), \quad 0 \leqslant i_1 < \alpha_1, \ldots, 0 \leqslant i_n < \alpha_n, \tag{4.1b}$$

where x_0 is any point that is not on the "characteristic set" of the system (4.1a) (the *characteristic set* of a system is the set of common zeros of the leading coefficients of the operators P_i). The discussion below assumes, for the sake of simplicity, that the characteristic set of (4.1a) is finite. If it is not, we have to take some other operators out of the annihilation ideal I_f, until the characteristic set is finite, and then give the appropriate initial conditions for the enlarged set of equations. A good "canonical holonomic representation" that always works, is in terms of Grobner bases.

In the last section we proved that if f and g are holonomic, then so are their sum and product. In order to manipulate concrete holonomic functions we should know how to find the "canonical holonomic representation" (4.1) of $f+g$ and fg given the canonical representations of f and g.

Suppose $P(D_1, x_1, \ldots, x_n)f = 0$, $Q(D_1, x_1, \ldots, x_n)g = 0$, where P has order α and Q has order β (in D_1). We can find an operator $R = R(D_1, x_1, \ldots, x_n)$ of order $\alpha + \beta$ that annihilates

$f+g$ as follows. The equation $Pf=0$ can be used to express $D_1^\alpha f$ as a linear combination, with coefficients that are rational functions in $x_1,\ldots,x_n$, of f, $D_1 f$, $D_1^2 f,\ldots, D_1^{\alpha-1}f$. By successively applying D_1, using the product rule of differentiation, and replacing $D_1^\alpha f$ by the above-mentioned linear combination, we can express $D_1^i f$, for $0 \leqslant i \leqslant \alpha+\beta$ in terms of $D_1^i f$ for $0 \leqslant i < \alpha$. Similarly we can express $D_1^i g$ for $0 \leqslant i \leqslant \alpha+\beta$ in terms of $D_1^i g$ for $0 \leqslant i < \beta$. It follows that we can express each of the $\alpha+\beta+1$ quantities $D_1^i(f+g)$, $0 \leqslant i \leqslant \alpha+\beta$, as linear combinations of the $\alpha+\beta$ quantities $D_1^i f$, $0 \leqslant i < \alpha$, and $D_1^j g$, $0 \leqslant j < \beta$. It follows that $D^i(f+g)$, $0 \leqslant i \leqslant \alpha+\beta$, are linearly dependent (over the rational functions in $x_1,\ldots,x_n$). The linear relation between them, after clearing denominators, is exactly a differential equation in D_1, satisfied by $f+g$. Now we repeat the process for D_i and get a system of the form (4.1) satisfied by $f+g$. To get the initial conditions we just use the information about f and g. To get the higher-order initial conditions for f and g, out of those that are given, we iteratively use the differential equations and plug in the point where the initial conditions are taken.

Trivial example. The holonomic functions (in the single variable x) $f=\mathrm{e}^{-x}$ and $g=\mathrm{e}^{-x^2}$ have the following canonical holonomic representations:

$$f=[D+1;\ f(0)=1], \qquad g=[D+2x;\ g(0)=1].$$

Now

$$(f+g)=f+g, \qquad D(f+g)=-f-2xg, \qquad D^2(f+g)=f+(4x^2-2)g.$$

Eliminating f and g yields the following differential equations for $(f+g)$:

$$(-2x+1)(f+g)''+(-4x^2+3)(f+g)'+(-4x^2+2x+2)(f+g)=0.$$

Since $(f+g)$ satisfies a second-order differential equation, we need to know $(f+g)(0)$ and $(f+g)'(0)$. Of course $(f+g)(0)=f(0)+g(0)=2$, but we do not know $(f+g)'(0)$ right away, since $f'(0)$ and $g'(0)$ are not part of the canonical representation of f and g, respectively. However, we can find $f'(0)$ and $g'(0)$ from their respective differential equations: $f'(0)=-f(0)=-1$ and $g'(0)=-2(0)g(0)=0$, so $(f+g)'(0)=-1+0=-1$. It follows that the canonical holonomic representation of $h:=(f+g)$ is

$$h=\left[(-2x+1)D^2+(-4x^2+3)D+(-4x^2+2x+2);\ h(0)=2,\ h'(0)=-1\right].$$

A similar process applies to fg. If P has order α and Q has order β, then it is possible to find a differential equation of order $\alpha\beta$ satisfied by fg. Now we use Leibnitz's rule to express the $\alpha\beta+1$ quantities $D_1^i(fg)$, $0 \leqslant i \leqslant \alpha\beta$, in terms of the $\alpha\beta$ quantities $(D_1^i f)(D_1^j g)$, for $0 \leqslant i \leqslant \alpha$ and $0 \leqslant j < \beta$.

The canonical representation (4.1), with its accompanying initial conditions is not unique. Given such a representation, we can left-multiply by any operator in D_i, get higher-order equations, and add the appropriate number of initial conditions. It is therefore necessary to know when two holonomic functions, given in terms of their canonical representation (4.1), are in fact the same. If f and g are holonomic functions suspected to be the same, which are given in terms of canonical representations, we can use the above method to find a canonical representation of $f-g$. Then we find the appropriate initial values of $f-g$ and check that they are all 0. Unfortunately, we actually have to find a concrete canonical representation, because otherwise we would not know the orders of the P_i, and the characteristic set of the system, both of which are needed to find how many initial conditions are required.

4.2. Discrete holonomic systems

Analogous considerations apply in the discrete realm. Every holonomic function f of the n discrete variables $m_1, \ldots, m_n$ can be specified by giving n "ordinary" recurrence equations:

$$P_i(E_i, m_1, \ldots, m_n)f = 0, \quad i = 1, \ldots, n, \tag{4.2}$$

with appropriate initial conditions. If we write

$$P_i = \sum_{j=0}^{\alpha_i} p_j^{(i)} E_i^j,$$

where α_i is the order of P_i as a recurrence operator in E_i, then the initial conditions needed are

$$f(a_1 + b_1, \ldots, a_n + b_n), \quad 0 \leqslant b_i < \alpha_i, \quad i = 1, \ldots, n,$$

where $(a_1, \ldots, a_n)$ is any point in $\mathbb{Z}^n$. In addition we need the values of f in the *characteristic set* of all common zeroes, on $\mathbb{Z}^n$, of the leading coefficients $\{p_{\alpha_i}(m_1, \ldots, m_n)\}$. If this set is not finite, then we must add to the "ordinary" equations in (4.2) some more "partial" equations out of its annihilation ideal I_f, with the appropriate initial conditions.

4.3. Continuous-discrete functions

Everything applies equally well to mixed continuous-discrete functions on $\mathbb{R}^N \times \mathbb{Z}^M$, with the obvious analogy.

4.4. How to compute indefinite integrals and sums

We know from Propositions 3.5 and 3.5′ that if f is holonomic in $\mathbb{R}^N \times \mathbb{Z}^M$, then

$$F_1 := \int f \, \mathrm{d}x_N \quad \text{and} \quad F_2 := \sum_{m_M} f$$

are holonomic in $\mathbb{R}^{N-1} \times \mathbb{Z}^M$ and $\mathbb{R}^N \times \mathbb{Z}^{M-1}$, respectively. However, in order to *compute* with holonomic functions we would need a way to find canonical holonomic representations of F_1 and F_2 out of a canonical holonomic representation of f.

It can be seen that the ideal I_{F_1} in $A_{N-1,M}(C)$ that annihilates F_1 is nothing but the ideal obtained by setting $D_N = 0$ in the "elimination ideal" $I_f \cap C\langle x_1, \ldots, x_{N-1}, D_1, \ldots \rangle$. Similarly I_{F_2} is obtained by setting $E_M = 1$ in the elimination ideal $I_f \cap C\langle \ldots, \ldots, \ \ldots, m_1, \ldots, m_{M-1} \rangle$.

In the next section we will see how to adapt Sylvester's classical dialytic elimination method [56, Volume II] to the Weyl algebra. Galligo [21] showed how to adapt Buchberger's powerful method of Grobner bases to the Weyl algebra which implies, in particular, efficient algorithms for elimination.

4.5. How to verify holonomic functions identities

We are now capable, at least in principle, of verifying any identity that involves *a fixed number of* sums, differences, products, integrals and sigmas of holonomic functions in *a fixed set of variables*. Simply bring it to the form in which the right side is zero, and compute a canonical holonomic representation of the right. Finally, compute the necessary initial conditions and verify that they are all zero.

4.6. *Special functions identities*

Many special functions, in particular all those in the famous "tableau d'Askey" [8,37], are holonomic in all their variables and parameters. For example, the Legendre polynomial $P_n(x)$, when viewed as a function of n and x, is holonomic, and the Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$, when viewed as a function of n, x, α and β is a holonomic function of $\mathbb{N} \times \mathbb{R}^3$. In general, any function that is given by hypergeometric summation, whose entries and arguments are linear combinations of the variables and parameters, is holonomic. This follows from the fact the summand (with respect to k, say) is clearly holonomic in all its variables and parameters, and it follows from Proposition 3.5′ that the sum itself is a holonomic function.

It follows from the above comments that, at least in principle, we can verify any identity involving a finite number of sums, products, sigmas and integrals of these special functions. However, at present I am only able to implement successfully the verification of single sum terminating hypergeometric identities, which is carried out in Sections 5 and 6 and in the Appendix. I am afraid that it is a long way before the Askey–Gasper identity [7, (2.30)] can be machine-proved. On the other hand, Apery's recurrence can be verified using the program. Of course, we still need the human to conjecture the identities, but once conjectured, the machine can save us the chore of proving them. [2]

5. Sylvester's dialytic elimination in the Weyl algebra

5.0.

In the previous section we saw that if $P(n, k, E_n, E_k)$ and $Q(n, k, E_n, E_k)$ are independent in the sense that they generate a holonomic system, then we can find an operator $R(n, E_n, E_k)$, *independent* of k, in the ideal I generated by P and Q. The proof given there, however, was an existence proof (using the pigeonhole principle). In this section we will present an algorithm that inputs $P(n, k, E_n, E_k)$ and $Q(n, k, E_n, E_k)$ and outputs an operator $R(n, E_n, E_k)$, with k missing, and two operators A and B such that $R = AP + BQ$. We will then show how this implies an effective algorithm for verifying any binomial coefficients identity ($\equiv$ terminating hypergeometric sum).

Given $F(n, k)$ that is holonomic, Proposition 3.4′ guarantees that

$$a(n) := \sum_k F(n, k)$$

is holonomic in the single variable n, wherever $a(n)$ is defined (i.e., the sum converges). But being holonomic in the single variable n means that $a(n)$ satisfies a linear recurrence equation with polynomial coefficients. Once we found an operator $R(n, E_n, E_k)$ that annihilates $F(n, k)$, i.e., $RF \equiv 0$, we can find an equation satisfied by $a(n)$ as follows.

[2] *Note added in proof*. The method of WZ pairs [58,59] is capable of *discovering* (and at the same time proving) new identities.

Write

$$R(n, E_n, E_k) = (1 - E_k)^{\alpha} \tilde{R}(n, E_n, E_k),$$

where α is maximal, i.e., $\tilde{R}(n, E_n, I) \neq 0$. We get

$$(1 - E_k)^{\alpha} \{ \tilde{R}(n, E_n, E_k) F(n, k) \} = 0.$$

But this means that

$$G(n, k) := \tilde{R}(n, E_n, E_k) F(n, k)$$

is a polynomial in k for every fixed n. If $G(n, k)$ were a nonzero polynomial in k, then $\Sigma_k G(n, k)$ would have diverged, a contradiction, since the convergence of $\Sigma_k F(n, k)$ implies the convergence of $\Sigma_k G(n, k)$. So $G(n, k) \equiv 0$ and we have

$$\tilde{R}(n, E_n, E_k) F(n, k) = 0.$$

Now write

$$\tilde{R}(n, E_n, E_k) = S(n, E_n) + (1 - E_k) \tilde{R}'(n, E_n, E_k),$$

where $S(n, E_n) := \tilde{R}(n, E_n, I)$ $(\neq 0)$.

We have

$$0 = \sum_k \tilde{R}(n, E_n, E_k) F(n, k)$$

$$= \sum_k S(n, E_n) F(n, k) + \sum_k (I - E_k) [\tilde{R}'(n, E_n, E_k) F(n, k)].$$

The second term vanishes by telescoping, so this is equal to

$$S(n, E_n) \sum_k F(n, k) = S(n, E_n) a(n).$$

We have just proved the next theorem.

Theorem 5.1. *Let $f : \mathbb{Z}^2 \to \mathbb{C}$ be such that $\Sigma_k F(n, k)$ converges for every n (this happens, in particular, if $F(n, \cdot)$ has finite support, like in all terminating hypergeometric series). If $F(n, k)$ is a solution of a linear partial recurrence equation with polynomial coefficients of the form*

$$R(n, E_n, E_k) F(n, k) \equiv 0,$$

with k missing from R, then $a(n) := \Sigma_k F(n, k)$ satisfies the following ordinary linear recurrence equation with polynomial coefficients:

$$S(n, E_n) a(n) \equiv 0,$$

where $S(n, E_n) = \tilde{R}(n, E_n, I)$ and $\tilde{R}$ is obtained from R by dividing by the highest possible power of $(I - E_k)$.

5.1. Sylvester's dialytic elimination in commutative algebra

We will now review Sylvester's classic dialytic elimination method [51] (see also [56]). In the next section we will describe how to modify it to the context of the (noncommutative) algebra of linear partial recurrence operators with polynomial coefficients. Sylvester's method basically consists of reducing algebraic elimination to linear elimination, so we will start with linear elimination.

5.1.1. Linear (Gaussian) elimination

Let R be any commutative ring, and let x be an indeterminate that does not necessarily commute with the elements of R. Consider the two linear affine forms $P = ax + b$ and $Q = cx + d$. In order to eliminate x, we multiply P by c, Q by a, and subtract: $cP - aQ = cb - ad$. So the ring element $cb - ad$ is in the "ideal generated by P and Q". All we needed was that a and c commute with each other. The procedure is still valid if b and d do not commute between themselves or with a and c.

More generally, consider the n affine-linear forms in the indeterminates $x_1, \ldots, x_{n-1}$:

$$P_j = \sum_{j=1}^{n-1} a_{i,j} x_j + \boldsymbol{b}_j, \quad j = 1, \ldots, n,$$

and consider the matrix of coefficients

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n-1} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{n,1} & \cdots & a_{n,n-1} & b_n \end{pmatrix}.$$

Let Δ_i be the cofactor of b_i in the above matrix A. Then obviously

$$\Delta_1 P_1 + \cdots + \Delta_n P_n = \det(A),$$

and $\det(A)$, *independent of the* x_i, is in the "ideal generated by $\{P_1, \ldots, P_n\}$". It is important for the sequel to note that all we need is that the $a_{i,j}$ be pairwise commutative, but the b_i neither have to commute with the $a_{i,j}$, nor among themselves. Of course, in that case we must be careful and evaluate the determinant with respect to the last column.

5.1.2. Sylvester's dialytic elimination

Following Sylvester [51], we will first do an example. Let

$$P = ax^2 + bx + c, \qquad Q = a'x^2 + b'x + c.$$

Form xP and xQ, and consider P, Q, xP, xQ as affine-linear forms in x, x^2, x^3:

$$\begin{pmatrix} P \\ xP \\ Q \\ xQ \end{pmatrix} = \begin{pmatrix} 0 & a & b & c \\ a & b & c & 0 \\ 0 & a' & b' & c' \\ a' & b' & c' & 0 \end{pmatrix} \begin{pmatrix} x^3 \\ x^2 \\ x \\ 1 \end{pmatrix}. \tag{5.1}$$

Let A be the matrix on the right side of (5.1) and let Δ_1, Δ_2, Δ_3 and Δ_4 be the cofactors of the entries of the last column in the above matrix; then

$$(\Delta_1 + \Delta_2 x)P + (\Delta_3 + \Delta_4 x)Q = \det(A).$$

It follows that $\det(A)$ belongs to the ideal in $R[x]$ generated by the two polynomials $P(x)$ and $Q(x)$, and is independent of x, i.e., an element of the ring of coefficients R. This is called *the resultant of P and Q*. If $\det(A)$ is equal to zero, then it means that P and Q have a common factor [56, Volume I, p.83]. So if P and Q are "independent" in the sense that they do not have a common factor (or, equivalently, a common zero), then the resultant must be nonzero.

More generally, to eliminate x out of

$$P = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0 \quad \text{and} \quad Q = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_0,$$

we form $P, xP, \ldots, x^{n-1}P$ and $Q_j xQ, \ldots, x^{m-1}Q$ and get $m+n$ linear forms in the indeterminates $1, x, \ldots, x^{m+n-1}$:

$$(P, \ldots, x^{n-1}P, Q, \ldots, x^{m-1}Q)^{\mathrm{T}} = A(x^{m+n-1}, \ldots, 1)^{\mathrm{T}}, \tag{5.2}$$

where A is the appropriate matrix. Let Δ_i, $i = 1, \ldots, n$, and Δ'_j, $j = 1, \ldots, m$, be the $m+n$ cofactors of the entries of the last column of A; then we have

$$(\Delta_1 + \Delta_2 x + \cdots + \Delta_n x^{n-1})P + (\Delta'_1 + \Delta'_2 x + \cdots + \Delta'_m x^{m-1})Q = \det(A).$$

$\det(A)$ is called the *resultant* since it is an expression in the coefficients of P and Q that determines whether they have a common factor: $\det(A)$ is zero if and only if P and Q have a common factor [56, Volume I, Section 27, p.83]. $\det(A)$ belongs to "the ideal generated by P and Q", and is independent of x.

5.2. *Dialytic elimination in the Weyl algebra: two simple examples*

Given two partial linear recurrence operators with polynomial coefficients $P(N, K, n, k)$ and $Q(N, K, n, k)$ we would like to "eliminate k", i.e., find two operators A and B such that $R := AP + BQ$ is independent of k, and, of course, nonzero. Trying to emulate Sylvester's dialytic method, the first thought that comes to mind is to multiply P and Q by powers of k. However, we cannot take the determinant, since the "ring of coefficients" is $C\langle N, K, n\rangle$, which is not commutative, so we do not get rid of k. Our twist on Sylvester's method is to multiply P and Q by various monomials $n^i k^j$, get many more "affine-linear forms" in the "indeterminates" $n^\alpha k^\beta$, $\beta > 0$, with coefficients in the *commutative* ring $C[N, K]$, with the constant part, i.e., the coefficient of "1", belonging to $C\langle N, K, n\rangle$. Only now can we take the determinant of the resulting system, and get the eliminated operator R. I will first present two examples.

Let us define the shift operators N and K by $Nf(n, k) := f(n+1, k)$, $Kf(n, k) := f(n, k+1)$. A *partial recurrence operator with polynomial coefficients* is any polynomial in the four indeterminates n, k, N, K. They satisfy the commutation relations: $Nn = (n+1)N$, $Kk = (k+1)K$, $NK = KN$, $Nk = kN$, $nK = Kn$, $nk = kn$.

By using the commutation relations it is easy to see that every operator can be written in the form

$$\sum_{r,s} p_{r,s}(N, K) n^r k^s. \tag{1}$$

Trivial example. $F(n, k) := n!/(k!(n-k)!)$, find $R(n, N, K)$ such that $RF = 0$.

Solution. $NF/F = (n+1)/(n-k+1)$, $KF/F = (n-k)/(k+1)$, so $P = (n-k+1)N - (n+1)$, $Q = (k+1)K - (n-k)$. Converting to the format of (1) we have:

$$P = -N(k) + (Nn - n - 1)(1), \qquad Q = (K+1)(k) - n(1).$$

In this case the only indeterminates are "k" and "1", and we already have as many equations as unknowns, so we do not have to multiply by any monomials $n^r k^s$. The matrix of coefficients is

$$\begin{pmatrix} -N & Nn - n - 1 \\ K+1 & -n \end{pmatrix},$$

whose determinant is $R(n, N, K) := (-N)(-n) - (K+1)(Nn - n - 1) = (n+1)(-NK + K + 1)$. Thus if $n \neq -1$, $F(n, k) = \binom{n}{k}$ satisfies

$$(-NK + K + 1)\binom{n}{k} = 0,$$

and we have rediscovered Pascal's triangle. In particular, $R(n, N, I) = (n+1)(-N+2I)$, and it follows that if

$$a(n) := \sum_k F(n, k),$$

then $(n+1)(-N+2I)a = 0$ which in everyday language means $(n+1)(-a(n+1) + 2a(n)) = 0$, whose solution is $C2^n$, for some constant C, which is easily seen to be 1 by plugging in $n = 0$. We have just proved the deep binomial coefficients identity

$$\sum_k \binom{n}{k} = 2^n.$$

A simple but not quite so trivial example. Let

$$F(n, k) = \binom{n}{k}\binom{n+k}{k} = \frac{(n+k)!}{k!^2(n-k)!}.$$

Here $NF/F = (n+k+1)/(n-k+1)$, $NKF/F = (n+k+1)(n+k+2)/(k+1)^2$. The relevant operators, in standard form are

$$P = (NK-1)(k^2) - 2(nk) - 3(k) - (n+1)(n+2),$$
$$Q = -(N+1)(k) + (N-1)n - 1.$$

Now we have four indeterminates: k^2, k, nk and 1, where the indeterminate "1" is allowed to have coefficients in $C\langle n, N, K\rangle$. However, we only have two equations. Is it possible to get more equations without introducing more indeterminates? Usually not, because creating more equations usually brings in more indeterminates and all you can do is reduce the excess "# indeterminates − #equations", until you get 0. However, whenever one of the operators is linear in k and n, like Q is in this example, then we can get away without increasing the number of indeterminates.

Now multiply Q by n and k, respectively, convert to the format (1), and you get

$$kQ = -(N+1)k^2 + (N-1)nk - k,$$
$$nQ = -(N+1)nk + Nk + N(n^2 - n) - (n^2 + n).$$

So

$$\begin{pmatrix} P \\ Q \\ kQ \\ nQ \end{pmatrix} = \begin{pmatrix} NK-1 & -2 & -3 & -(n+1)(n+2) \\ 0 & 0 & -N-1 & (N-1)n-1 \\ -N-1 & N-1 & -1 & 0 \\ 0 & -N-1 & N & N(n^2-n) - n^2 - n \end{pmatrix} \begin{pmatrix} k^2 \\ nk \\ k \\ 1 \end{pmatrix}.$$

Now take the determinant of this, and get $R(n, N, K)$. Plugging in $K = I$ gives a linear recurrence operator that annihilates

$$a(n) := \sum_k F(n, k) = \sum_k \binom{n}{k}\binom{n+k}{k}.$$

It turns out to be

$$n^2N^4 + (-5n^2 - 10n - 3)N^3 + (2 - 5n^2)N^2 + (n^2 + 2n + 1)N.$$

This operator is not the minimal operator annihilating $a(n)$, so this method only produces *an* operator annihilating $\Sigma_k F(n, k)$. However, a minimal operator can be found empirically and then one can use the Euclidean algorithm (adapted to linear ordinary recurrence operators with polynomial coefficients) to show that the conjectured operator is a right factor of the operator that the method outputed.

5.3. Sylvester dialytic elimination in the Weyl algebra: the general method

Consider two operators $P(N, K, n, k)$ and $Q(N, K, n, k)$ of degrees b_1 and b_2, respectively, in k, and maximal degree c in n, and let us write them in the standard form, and with the k-free part set apart at the end:

$$P = \sum_{i=1}^{b_1} \sum_{j=0}^{c} p_{i,j}(N, K)k^i n^j + R_p(n, N, K)$$

and

$$Q = \sum_{i=1}^{b_2} \sum_{j=0}^{c} q_{i,j}(N, K)k^i n^j + R_Q(n, N, K).$$

We consider all operators as polynomials in the variables n and k with coefficients that belong to the commutative ring $C[N, K]$. The coefficients mutually commute and the only noncommutativity that arises is from the interaction between the coefficients and the variables. As in the commutative case, let us form

$$P, kP, \ldots, k^{b_2-1}P, \quad \text{and} \quad Q, kQ, \ldots, k^{b_1-1}Q.$$

We must bring them to the standard form as above, using $k^iN^\alpha K^\beta = N^\alpha K^\beta(k-\beta)^i$.

We now consider these as linear forms in the "indeterminates"

$$k^\alpha n^\beta, \quad 1 \leqslant \alpha \leqslant b_1 + b_2 - 1, \ 0 \leqslant \beta \leqslant c.$$

Alas, now we have far too few linear forms, or equivalently, far too many indeterminates. Namely, we have $b_1 + b_2$ linear forms and $(b_1 + b_2 - 1)(c + 1)$ indeterminates, and the parameter "#linear forms − #indeterminates" is $1 - c(b_1 + b_2 - 1)$. We need this parameter to be 1.

To set the balance right, we will multiply by successive powers of n. At the rth stage, when we multiply the $b_1 + b_2$ linear forms

$$\{P, kP, \ldots, k^{b_2-1}P, Q, \ldots, k^{b_1-1}Q\} \tag{5.3}$$

by the power n^r, we gain $b_1 + b_2$ new linear forms, but are only burdened with $b_1 + b_2 - 1$ new indeterminates: $k^\alpha n^{r+c}$, for $\alpha = 1, \ldots, b_1 + b_2 - 1$. It follows that the budgetary deficit "#indeterminates − #linear forms" gets reduced by one for every power of n by which we multiply. So if we multiply (5.3) by n^r for $r = 1, \ldots, c(b_1 + b_2 - 1)$, we will finally have one more linear form than we have indeterminates.

Let $A_f(i, j)$ be the "k-free" part of k^in^jP, $i = 0, \ldots, b_2 - 1$, $j = 0, \ldots, c(b_1 + b_2 - 1)$, and $B_f(i, j)$ be the "k-free" part of k^in^jQ, $i = 0, \ldots, b_1 - 1$, $j = 0, \ldots, c(b_1 + b_2 - 1)$. Also let

$A(i, j) = k^i n^j P - A_f(i, j)$ and $B(i, j) = k^i n^j Q - B_f(i, j)$. By linear algebra, $A(i, j) \cup B(i, j)$ are linearly dependent, so there exist $a_{i,j}(N, K)$ and $b_{i,j}(N, K)$, *not all zero*, such that

$$\sum a_{i,j}(N, K) A(i, j) + \sum b_{i,j}(N, K) B(i, j) = 0. \tag{5.4}$$

If follows that

$$\begin{aligned} R &:= \left(\sum a_{i,j}(N, K) k^i n^j\right) P + \left(\sum b_{i,j}(N, K) k^i n^j\right) Q \\ &= \sum a_{i,j}(N, K) A_f(i, j) + \sum b_{i,j}(N, K) B_f(i, j) \end{aligned}$$

is independent of k, i.e., $R = R(n, N, K)$. Obviously R is an operator that belongs to the left ideal generated by P and Q and so if $F(n, k)$ is a solution of the two linear partial recurrence equations $P(n, k, N, K)F \equiv 0$ and $Q(n, k, N, K)F \equiv 0$, then F automatically satisfies $R(n, N, K)F \equiv 0$. In other words, we have succeeded in eliminating k from P and Q.

We have to make sure that R is not the zero operator. Of course, if P and Q are dependent, for example if they are both left multiples of the same operator: $P = P'A$, $Q = Q'A$, then R *will* be zero. The technical condition of "independence" is exactly that $\{P, Q\}$ generate a holonomic system, i.e., that the ideal I generated by P and Q is such that $A_2(C)/I$ is a holonomic $A_2(C)$-module.

Suppose that R was the zero operator. This means that we can find two operators $A(k, n, K, N)$ of degree $< b_2$ in k and $B(k, n, K, N)$ of degree $< b_1$ in k such that $AP = BQ$. Now taking the "principal symbols" as operators in $A_2(C)$ will show that the images of P and Q in the commutative graded ring are dependent, which is impossible if P and Q generate a holonomic system.

To find the $a_{i,j}$ and $b_{i,j}$ of (5.4) we can take the cofactors of the last column in the determinant formed by the matrix of coefficients of the $k^i n^j P$, $k^i n^j Q$ expressed as linear combinations of the monomials $n^\alpha k^\beta$, for $\beta > 0$ and the "monomial" k^0. If all these cofactors are zero, then we would have to take the cofactors of an appropriate nonsingular subdeterminant.

5.4. *Implementation*

In the preceding discussion we worked in terms of *the k-degrees in k, b_1 and b_2, of P and Q, respectively, and the highest n-degree, c. Thus it applies to operators whose "supports"*

$$\{(i, j);\ p_{i,j} \neq 0\}, \qquad \{(i, j);\ q_{i,j} \neq 0\}$$

are included in the rectangles

$$\{(i, j);\ 0 \leqslant i \leqslant b_1, 0 \leqslant j \leqslant c\} \quad \text{and} \quad \{(i, j);\ 0 \leqslant i \leqslant b_2, 0 \leqslant j \leqslant c\}.$$

It turns out that most operators that occur "in nature" have their "support" in the shape of a triangle:

$$\{(i, j);\ 0 \leqslant i, 0 \leqslant j, i + j \leqslant d\},$$

so it is a waste to consider the smallest rectangle containing their support, since half of it will be zeros. In short, it is more natural to consider the *total degree* in k and n. This reduces the size of the determinant considerably. In Section 6 I will give an explicit algorithm that uses the total degree.

5.5. Elimination in general

A similar method works for the continuous case. Just bring all operators to the form

$$P=\sum_{i=1}^{b_1}\sum_{j=0}^{c} p_{i,j}(D_x,\ D_y)y^ix^j+R_P(n,\ N,\ K),$$

multiply from the left by monomials $y^\alpha x^\beta$, and convert back to that form using

$$xD_x^iD_y^j=D_x^iD_y^jx-iD_x^{i-1}D_y^j \quad\text{and}\quad yD_x^iD_y^j=D_x^iD_y^jy-jD_x^iD_y^{j-1}.$$

The q-analog is just as straightforward. For the discrete q-analog you replace n by q^n, k by q^k, and the commutation rules $nN=N(n-1)$, $kK=K(k-1)$ by $q^nN=Nq^n/q$, $q^kK=Kq^k/q$. For the continuous q-analog, featuring the q-dilation operators $Q_xf(x,\ y)=f(qx,\ y)$ and $Q_yf(x,\ y)=f(x,\ qy)$, we have the commutation rules $xQ_x=Q_xx/q$ and $yQ_y=Q_yy/q$.

Everything extends to the elimination of several variables from a general annihilation ideal, as in the classical case [56], but in this case Buchberger's method of Grobner bases, as adapted by Galligo [21], is far superior.

6. An algorithm for proving binomial coefficients identities

When someone finds a new identity,
there are not many people who get
excited about it any more, except
the discoverer.

Donald E. Knuth [36, p.53]

6.1. The algorithm

We saw that, *in principle*, it is possible to verify any (holonomic) special function identity involving sums, integrals and products. However, as Askey said so aptly [5]: "Many things that can be done in theory cannot be done in practice". In this section I will show that at least some things can be done in practice, by giving an explicit algorithm for *proving*, or refuting, binomial coefficients identities ($\equiv$ terminating hypergeometric identities) of the form

$$\sum_k F(n,\ k)=\mathrm{rhs}(n). \tag{6.1}$$

A MAPLE implementation of this algorithm is given in the Appendix.

In (6.1) it is assumed that $F(n,\ k)$ has the form

$$F(n,\ k)=\frac{\prod_{i=1}^{A}(a_in+a_i'k+a_i'')!}{\prod_{i=1}^{B}(b_in+b_i'k+b_i'')!}z^k, \tag{6.2a}$$

where the a_i, a_i', b_i, b_i' have to be *constant*, specific, (positive or negative) integers, but z, a_i'' and b_i'' can be any complex numbers or parameters, and $x!$ means $\Gamma(x+1)$.

The right side of (6.1), rhs(n), may be given explicitly, in the form

$$\mathrm{rhs}(n) = C \frac{\prod_{i=1}^{A} (\bar{a}_i n + \bar{a}_i')!}{\prod_{i=1}^{B} (\bar{b}_i n + \bar{b}_i')!} \bar{x}^n, \tag{6.2b}$$

where $\bar{a}_i$ and $\bar{b}_i$ are specific (positive or negative) integers, and $\bar{a}_i'$, $\bar{b}_i'$, C and $\bar{x}$, are complex numbers or parameters. Another possibility is that the right side is given implicitly in terms of a minimal (ordinary) linear recurrence operator with polynomial coefficients, conj(n, N), that annihilates rhs(n), together with the appropriate initial conditions.

The algorithm consists of 10 steps.

Step I: (a) Find operators $p(k, n, N, K)$ and $q(k, n, N, K)$ that annihilate $F(n, k)$. [In a more general situation p and q may be given from the outset, in which case you go to Step 2 directly.]

You do this by computing the rational functions $F(n+1, k)/F(n, k)$ and $F(n, k+1)/F(n, k)$:

$$\frac{F(n+1, k)}{F(n, k)} = \frac{P(n, k)}{Q(n, k)}, \qquad \frac{F(n, k+1)}{F(n, k)} = \frac{P'(n, k)}{Q'(n, k)}, \tag{6.3}$$

where P, Q, P', Q' are polynomials in n and k and, of course, $\gcd(P, Q) = 1$ and $\gcd(P', Q') = 1$. The operators p and q are given by

$$p(k, n, N, K) := NQ(n-1, k) - P(n, k),$$
$$q(k, n, N, K) := KQ'(n, k-1) - P'(n, k).$$

[p and q both annihilate $F(n, k)$.]

(b) If the right side of (6.1) is given explicitly, find the (minimal) linear recurrence operator with polynomial coefficients that annihilates rhs(n). You do this by computing rhs$(n+1)/$rhs$(n) = \bar{P}(n)/\bar{Q}(n)$, say, and set

$$\mathrm{conj}(n, N) := N\bar{Q}(n-1) - \bar{P}(n).$$

Step II: Find the degree, in k, of $p(k, n, N, K)$ and $q(k, n, N, K)$, say α and β, respectively. [It is readily seen that because the form (6.2a) of $F(n, k)$ the (generic) degree of p in n equals to the degree in k, which equals to the total degree in (n, k), and similarly for q. We will view all operators as polynomials in n and k with coefficients that belong to $C[N, K]$.]

Step III: Let $I1$ and $I2$ be the following sets of pairs of integers (i, j):

$$I1 := \{(i, j);\ 0 \leqslant i \leqslant \beta - 1,\ j \geqslant 0,\ i + j \leqslant \alpha(\beta - 1)\},$$
$$I2 := \{(i, j);\ 0 \leqslant i \leqslant \alpha - 1,\ j \geqslant 0,\ i + j \leqslant (\alpha - 1)\beta\}.$$

For each pair (i, j) in $I1$, form the operator $k^i n^j p$, convert it to standard form, and expand in terms of the monomials $k^{i'} n^{j'}$. Similarly, for q and $I2$. So let

$$k^i n^j p = \sum A_{i'j'}^{(i,j)}(K, N) k^{i'} n^{j'}, \quad (i, j) \in I1, \tag{6.4a}$$

$$k^i n^j q = \sum B_{i'j'}^{(i,j)}(K, N) k^{i'} n^{j'}, \quad (i, j) \in I2. \tag{6.4b}$$

[A simple subroutine left-multiplies any operator by a monomial and converts it to standard form, using $k^i n^j K^a N^b = K^a N^b (k-a)^i (n-b)^j$.]

Step IV: [We now consider (6.4) as affine-linear forms in the "indeterminates" $k^{i'} n^{j'}$. It is readily seen that the monomials that feature in (6.4) are those belonging to the following set

$$l3 := \{(i',\ j');\ 1 \leqslant i' \leqslant \alpha + \beta - 1,\ j' \geqslant 0,\ i' + j' \leqslant \alpha\beta\},$$

with coefficients that are in $C[N,\ K]$. The "affine part" (i.e., "the coefficient of k^0") is in $C\langle N,\ K,\ n\rangle$. Note that $|l1| + |l2| = |l3| + 1 = \frac{1}{2}(2\alpha\beta - \alpha - \beta + 2)(\alpha + \beta - 1) + 1$.]

Let

$$\text{SIZE} := \tfrac{1}{2}(2\alpha\beta - \alpha - \beta + 2)(\alpha + \beta - 1) + 1.$$

Form the SIZE $\times$ SIZE matrix M as follows. First convert the sets $l1$, $l2$, $l3$, into lists:

$$l1 := l1[r], \quad r = 1, \ldots, |l1|,$$
$$l2 := l2[r], \quad r = 1, \ldots, |l2|,$$
$$l3 := l3[r], \quad r = 1, \ldots, |l3|.$$

Now, for $1 \leqslant r \leqslant |l1|$, and $1 \leqslant s \leqslant \text{SIZE} - 1$, set

$$M[r,\ s] := A_{i'j'}^{(i,j)}(N,\ K) \quad \text{(of (6.4a))},$$

where $(i,\ j) := l1[r]$ and $(i',\ j') := l3[s]$.

For $|l1| + 1 \leqslant r \leqslant \text{SIZE}$, and $1 \leqslant s \leqslant \text{SIZE} - 1$, set

$$M[r,\ s] := B_{i'j'}^{(i,j)}(N,\ K) \quad \text{(of (6.4b))},$$

where $(i,\ j) := l2[r - |l1|]$ and $(i',\ j') := l3[s]$.

Finally, for $1 \leqslant r \leqslant |l1|$,

$$M[r,\ \text{SIZE}] := \sum_{j'} A_{0,j'}^{(i,j)}(N,\ K) \quad \text{(of (6.4a))},$$

where $(i,\ j) := l1[r]$. In other words $M[r,\ \text{SIZE}]$ is the coefficient of k^0 in $k^i n^j p$. Similarly, for $|l1| \leqslant r \leqslant \text{SIZE}$,

$$M[r,\ \text{SIZE}] := \sum_{j'} B_{0j'}^{(i,j)}(N,\ K) \quad \text{(of (6.4b))},$$

where $(i,\ j) := l2[r - |l1|]$. In other words $M[r,\ \text{SIZE}]$ is the coefficient of k^0 in $k^i n^j q$.

Step V: Let

$$R(n,\ N,\ K) := \text{determinant}(M).$$

If R is identically zero (a very rare event), then take an appropriate subdeterminant, making sure that the last column gets chosen. [More precisely, find the rank of the matrix A without the last column, pick that many linearly independent rows and one less column and adjoin the (relevant part of the) last column.] [$R(n,\ N,\ K)$ annihilates $F(n,\ k)$.]

Step VI: Find $\overline{R}(n, N, K)$ such that

$$R(n, N, K) = (1-K)^g \overline{R}(n, N, K),$$

where g is as big as possible. [In Section 5 we showed that $\overline{R}(n, N, K)$ also annihilates $F(n, k)$.]

Step VII: Let $S(n, N)$ be the ordinary linear recurrence operator with polynomial coefficients (in n) obtained by substituting $K=1$ in $\overline{R}(n, N, K)$:

$$S(n, N) := \overline{R}(n, N, 1).$$

[$S(n, N)$ annihilates $a(n) := \sum_k F(n, k)$.]

Step VIII: Use the Euclidean algorithm (adapted to the algebra of ordinary linear recurrence operators) to find operators $T(n, N)$ and rem(n, N) such that

$$s(n, N) = T(n, N)\mathrm{conj}(n, N) + \mathrm{rem}(n, N),$$

where the degree of rem(n, N) in N (i.e., its order as a recurrence operator) is smaller than that of conj(n, N).

Step IX: If rem$(n, N) \neq 0$, then the identity is false, i.e., the left side is *not* annihilated by conj(n, N), while the left side is (by definition). This follows from the fact that conj(n, N) was taken to be the lowest-order (nonzero) operator that annihilates the right side. If the left side would have been equal to the right side, then rem(n, N) would also annihilate the right side, contradicting the minimality of conj(n, N).

If rem$(n, N) \equiv 0$, then the identity is true provided it is true at some initial points, which are determined as follows. We have

$$T(n, N)[\mathrm{conj}(n, N)a(n)] \equiv 0.$$

Let the order (alias degree in N) of $T(n, N)$ be t, and let the leading coefficient, of N^t, be mekadem(n). Then it is obvious that conj$(n, N)a(n) \equiv 0$ provided it is 0 at $n = 0, \dots, t-1$, and at the (the usually empty) set of positive integer zeros of mekadem(n).

Step X: Now we know that

$$a(n) := \sum_k F(n, k),$$

and rhs(n) are both solutions of the linear recurrence equation

$$\mathrm{conj}(n, N)s(n) \equiv 0.$$

In order to infer that $a(n)$ and rhs(n) are identically equal we have to check that they match at the first r values $n = 0, \dots, r-1$, where r is the order of rhs(n, N), and at the "characteristic set" of positive integer zeros of the leading coefficient of rhs(n, N).

6.2. Discussion

The above algorithm is capable of proving *any* identity of the form (6.1), with the summand of the form (6.2a). This includes all terminating hypergeometric identities in [9], as well as those of Gosper's list [33], and almost all those of Gessel and Stanton [32]. A MAPLE program implementing the algorithm is given in the Appendix. This program requires a lot of memory. For example, on my AT&T 3B1 PC, I was only able to do those identities for which SIZE, the size of the matrix whose determinant we take, is smaller than 13. In other words, my computer can handle the following values for (α, β): $\{(\alpha, 1);\ \alpha \leqslant 5\}$, $\{(1, \beta);\ \beta \leqslant 5\}$ and (2,2). A sampling of the identities proved by my computer, using the program, is given in the Appendix. For example, my computer can prove the Saalschutz identity, for which $\alpha = 2$, $\beta = 2$, but it ran out of memory for Dixon's identity, for which $\alpha = 2$, $\beta = 3$, and hence SIZE = 19. Among the identities in [32], my program proved (1.1) (originally due to Andrews [3]), (1.5), (5.24) and (5.25). I am sure that with a bigger computer, and even more importantly, with a better programmer, the algorithm would be able to handle much deeper identities. Incidentally, one may use the parameter SIZE as a measure of the depth of an identity. [3]

In any case, the main virtue of the above algorithm is theoretical, because it tells you that you are *guaranteed* to find an operator $R(n, N, K)$ annihilating $F(n, k)$, and from its expression as a determinant, it is very easy to derive a priori bounds for the degrees in K, N and n. Now that we know that there is such an operator, we can easily try out a generic form:

$$R(n, N, K) := \Sigma a_{i,j,l} n^i N^j K^l,$$

require that $R(n, N, K)F(n, k) \equiv 0$, plug in specific values of (n, k) and solve for the coefficients $a_{i,j,l}$. Having conjectured such an operator we can easily verify that indeed $R(n, N, K)$ annihilates $F(n, k)$ and obtain $S(n, N)$ as before. Alternatively, we can use the method of *creative telescoping*.

6.3. The method of creative telescoping

The term *creative telescoping* was coined by van der Poorten in his charming account [55] of Apery's proof of the irrationality of $\zeta(3)$. The method consists in "creating" a certain multiple of $F(n, k)$ by a rational function $P(n, k)/Q(n, k)$:

$$G(n, k) := \frac{P(n, k)}{Q(n, k)} F(n, k), \tag{6.5}$$

with the property that

$$G(n, k+1) - G(n, k) \quad (= (K-1)G(n, k)) = \mathrm{conj}(n, N)F(n, k), \tag{6.6}$$

which immediately implies that $a(n) := \Sigma_k F(n, k)$ satisfies $\mathrm{conj}(n, N)a(n) \equiv 0$.

Perlstadt [44] used creative telescoping and MACSYMA to find recurrences for sums of powers (up to the 6th) of binomial coefficients.

[3] *Note added in proof.* The improved algorithm [61,62] can do all of [32,33].

In order to implement creative telescoping we write $P(n, k)$ and $Q(n, k)$ in generic form (guessing their degrees), plug into (6.6), divide by $F(n, k)$, clear denominators, and compare coefficients, getting equations that can be solved.

Can you *always* use creative telescoping? The algorithm of Section 6.1 guarantees that we have an operator $\bar{R}(n, N, K)$ that annihilates $F(n, k)$, and $S(n, N) := \bar{R}(n, N, 1)$ is nonzero. Then we have

$$\bar{R}(n, N, K) - S(n, N) = (1-K)L(n, N, K).$$

It follows that

$$S(n, N)F(n, k) = (1-K)[L(n, N, K)F(n, k)].$$

We saw above that if the identity is true, then $S(n, N)$ must be a left multiple of $\operatorname{conj}(n, N)$. Taking

$$G(n, k) := L(n, N, K)F(n, k),$$

we see that (6.6) is *always* true, but sometimes with $\operatorname{conj}(n, N)$ replaced by some left multiple of it. So we now have an explanation why creative telescoping works. First try to find $G(n, k)$ in (6.6) with the conjectured operator $\operatorname{conj}(n, N)$; if you fail, try to do it over with some left multiple of it.

Note added at the revised version

The ideas in this section were extended to a much faster algorithm [61,62]. This leads to the notions of "WZ pair" and "rational function certification" [58,59], see also [15].

6.4. *Continuous and discrete-continuous analoges*

The algorithm of (6.1) can be easily adapted to the continuous case of eliminating y out of $p(x, y, D_x, D_y)$ and $q(x, y, D_x, D_y)$, and proving identities of the form

$$\int_{-\infty}^{\infty} F(x, y)\, \mathrm{d}y = a(x),$$

where $F(x, y)$ has the form

$$F(x, y) = \frac{A(x, y)}{B(x, y)} \exp\left(\frac{C(x, y)}{D(x, y)}\right),$$

for some polynomials A, B, C, D in x and y, and $a(x)$ is either given explicitly, or in terms of a linear differential equation with polynomial coefficients that it satisfies.

We can also easily adapt the algorithm to handle discrete-continuous identities like

$$\sum_{n=0}^{\infty} P_n(x)t^n = (1-2xt+t^2)^{-1/2}.$$

Note added at the revised version

The ideas in this section were extended to a much faster algorithm in [2].

6.5. The q-analog

The algorithm of Section 6.1 can be easily adapted to the proof of any q-binomial identity of the form

$$\sum_k F(n, k) = a(n),$$

where $F(n, k)$ has the form

$$F(n, k) = \frac{\prod_{i=1}^{A} (q^{a_i''}; q)_{a_i n + a_i' k}}{\prod_{i=1}^{B} (q^{b_i''}; q)_{b_i n + b_i' k}} z^k,$$

where the a_i, a_i', b_i, b_i' have to be *constant*, specific (positive or negative) integers, but z, a_i'', and b_i'' can be any complex numbers or parameters.

Note added at the revised version
A much faster algorithm is given in [63].

More generally, the whole theory of holonomic systems can be q-analogized, where instead of $C\langle E_1, \ldots, E_n, m_1, \ldots, m_n\rangle$ we have $C\langle E_1, \ldots, E_n, q^{m_1}, \ldots, q^{m_n}\rangle$, and the commutation rule is $E_i(q^{m_i}) = q(q^{m_i}E_i)$. Also, instead of $C\langle D_1, \ldots, D_n, x_1, \ldots, x_n\rangle$, we have $C\langle Q_1, \ldots, Q_n, x_1, \ldots, x_n\rangle$, where Q_i are the q-dilation operators: $Q_i f(x_1, \ldots, x_i, \ldots, x_n) = f(x_1, \ldots, qx_i, \ldots, x_n)$.

Acknowledgements

I have already mentioned my indebtedness to Leonard Lipshitz, Richard Stanley and Donald Knuth for pointing out the inadequacy of [60]. I also wish to thank Richard Askey for his penetrating critique [5] of the philosophical side of [60] and for many helpful discussions.

It was Bernd Sturmfels who told me about Grobner bases and thus gave me hope that my clumsy algorithms may be improved considerably.

Special thanks are due to Joseph Bernstein who spent two hours of his precious time explaining his beautiful theory to me, in a language that I understood and that enabled me to transmit the holonomic fire to my fellow mortals in the provinces of special functions and combinatorics.

I would also like to thank the four developers of MAPLE: Bruce Char, Keith Geddes, Gaston Gonnet and Stephen Watt, for developing such a beautiful system, and Kurt Horton, the local UNIX guru, for informing me that MAPLE is available on the AT&T 3B1.

Many thanks are due to Gert Almkvist for his interest, encouragement and many valuable suggestions. Readers who can read Swedish will enjoy Almkvist's delightful account of the present ideas [1].

Finally, I also benefited from numerous conversations with Leon Ehrenpreis, Dominique Foata, Ira Gessel, Yakar Kannai, Steve Milne, Ron Perline, Marci Perlstadt, Joel Spencer, Dennis Stanton, Jet Wimp and Herbert Wilf.

Appendix: a MAPLE implementation of the algorithm in Section 6.1 that proves binomial coefficients identities

In a world in which the price of calculation
continues to decrease rapidly, but the price of
theorem proving continues to hold steady or increase,
elementary economics indicates that we ought to spend
a larger and larger fraction of our time on
calculation.

John W. Tukey [54, p.74]

A.1. The program

The following is a MAPLE program that implements the algorithm in Section 6.1. In order to use it, either type it into a file or get it from me via electronic mail. Call the program file *verifier* (or any name of your choice). Then in the UNIX operating system, which has MAPLE, you type the following on your terminal:

maple < verifier

The input given in the program below is for proving the following deep binomial coefficients identity

$$\sum_k \frac{1}{k!(n-k)!} = \frac{2^n}{n!}.$$

In order to use this program for other identities

$$\sum_k F(n, k) = a(n),$$

you must first bring it to the form (6.2), i.e., make sure that the left has no terms depending on n only, and express everything in terms of factorials. Then you replace the $1/(k!(n-k)!)$ in the INPUT below by $F(n, k)$ and the $2^n/n!$ by $a(n)$. A sampling of inputs for which this program ran successfully (i.e., did not run out of memory) on my AT&T 3B1 PC (that has 2 Meg of RAM, 67 Meg of hard disk, and 60% of free disk space), is given in Section A.2.

If the right side $a(n)$ is given implicitly, in terms of a recurrence, then you should change "expli := 1" to "expli := 0", erase the line with rhs, and replace it by "conj := ...", where the right side is the conjectured operator conj(n, N). See examples (1), (3), (6)–(10), in Section A.2.

```
#THE IDENTITY VERIFIER
#This program proves binomial coefficients identities (alias
#terminating hypergeometric summation) of the
#form described in Section 6.1. The input consists of
#F(n, k), the summand of the lhs, and the conjectured rhs,
#both of which should be quotients of products of factorials
#of the form (an + bk + c)! for some integers a and b, where b is not 0.
#The output consists of a human readable proof or refutation of the
#identity
#THIS IS THE INPUT
```

```
expli := 1:
summand := 1/(k!*(n - k)!):
rhs := 2^n/n!:
#END OF INPUT
if expli = 1 then
rat := subs(n = n + 1,rhs)/rhs:
rat := expand(rat):
conj := denom(rat)*N - numer(rat):
fi:
rat := subs(n = n + 1,summand)/summand:
rat := expand(rat):
p1 := subs(n = n - 1,denom(rat))*N - numer(rat):
rat := subs(k = k + 1,summand)/summand:
rat := expand(rat):
q1 := subs(k = k - 1,denom(rat))*K - numer(rat):
dp1 := degree(p1,k):
dq1 := degree(q1,k):
p1 := expand(p1):
q1 := expand(q1):
#to generate the lists of exponents mapa, maq
#by which p1 and q1 should be multiplied respectively, and khezka
#of exponents that feature, given the degree of p1,
#dp1, and the degree of q1, dq1
kamap := 0:
for i from 0 to dq1 - 1 do
for j from 0 to (dq1 - 1)*dp 1 - i do
 kamap := kamap + 1:
mapa[kamap,1] := i; mapa[kamap,2] := j:
od:
od:
kamaq := 0:
for i from 0 to dp1 - 1 do
for j from 0 to (dp1 - 1)*dq1 - i do
 kamaq := kamaq + 1:
maq[kamaq,1] := i: maq[kamaq,2] := j:
od:
od:
kamakh := 0:
for i from 1 to dp1 + dq1 - 1 do
for j from 0 to dp1*dq1 - i do
 kamakh := kamakh + 1:
khezka[kamakh,1] := i; khezka[kamakh,2] := j;
od:
od;
with(linalg):
```

```
#The following function converts an operator p(n,x) from
#N^j*n^i form to n^i*N^j form
convert1 := proc(p)
 local i,p1,pa,r,term:
    p1 := expand(p):
     pa := 0:
     r := degree(p,N):
      for i from 0 to r do
      term := coeff(p1,N,i):
         term := expand(term):
         term := subs(n = n + i,term):
         term := expand(term):
         pa := pa + term* N^i:
      od:
     RETURN(expand(pa)):
     end:
timesn := proc(p):
 RETURN(expand(n* p - N *diff(p,N))):
end:
timesk := proc(p):
RETURN(expand(k* p - K*diff(p,K))):
end:
timeskn := proc(p,exk,exn)
 local pa,i
    pa := p:
 for i from 1 to exk do
    pa := timesk(pa):
    pa := expand(pa):
 od:
 for i from 1 to exn do
    pa := timesn(pa):
    pa := expand(pa):
 od:
 RETURN(pa):
end:
p1 := expand(p1):
q1 := expand(q1):
kama := kamap + kamaq:
ma := array(1..kama,1..kama):
 for i from 1 to kamap do
    p11 := timeskn(p1,mapa[i,1],mapa[i,2]):
    p11:expand(p11):
    for j from 1 to kama - 1 do
     tem := coeff(p11,k,khezka[j,1]):
     tem := expand(tem):
```

```
      tem := coeff(tem,n,khezka[j,2]):
      ma[i,j] := expand(tem):
    od:
      p11 := expand(p11):
    tem := coeff(p11,k,0):
    ma[i,kama] := expand(tem):
      od:
for i from kamap + 1 to kamap + kamaq do
  p11:timeskn(q1,maq[i - kamap,1],maq[i - kamap,2]):
  p11 := expand(p11):
    for j from 1 to kama - 1 do
      tem := coeff(p11,k,khezka[j,1]):
      tem := expand(tem):
      tem := coeff(tem,n,khezka[j,2]):
      ma[i,j] := expand(tem):
    od:
    tem := coeff(p11,k,0):
    ma[i,kama] := expand(tem):
  od:
equ := det(ma):
if equ = 0 then
  print('the determinant vanishes, take an appropriate subdeterminant'):
fi:
equ := subs(K = K + 1,equ):
equ := expand(equ):
equ1 := coeff(equ,K,0):
  while equ1 = 0 do
    equ := expand(equ/K):
equ1 := coeff(equ,K,0):
od:
equ := convert1(equ):
equ := subs(K = K - 1,equ):
print('F(n,k) is annihilated by the following, k-free, operator(check!)'):
print('(Warning: it should be in the standard form in which n is in)'):
print('front of N, if it is not, it is still meant that way'):
print(equ):
equ1 := convert1(equ1):
equ1 := expand(equ1):
print('Therefore, putting K = 1, we get an operator A(N,n) annihilating the rhs'):
print(equ1):
timesx := proc(p,nup)
local gu:
  gu := expand(subs(n = n + nup,p)) * N^nup:
  RETURN(gu):
     end:
```

```
  aa := 1:
  sa := 0:
  ra := equ1:
  qaq := conj:
  degq := degree(qaq,N):
  while degree(ra,N) > = degq do
     degr := degree(ra,N):
     nu := degr - degq:
     qaqa := expand(timesx(qaq,nu)):
     leadr := coeff(expand(ra),N,degr):
     leadqa := coeff(qaqa,N,degr):
     gcd(leadr,leadqa,'leadr','leadqa'):
     aa := aa*leadqa:
     sa := sa*leadqa:
     sa := sa + leadr*N^nu:
     sa := expand(sa):
     ra := leadqa*ra - leadr*qaqa:
     ra := expand(ra):
od:
print('the conjectured rhs is annihilated by the following operator B(N,n)'):
print(conj):
if(ra = 0) then
  print('Now A(N,n) = S(N,n)*B(N,n)(check!), where S(N,n) is'):
  print(sa):
  gu := degree(sa,N):
  print('Since S has order',gu,'in N, THE IDENTITY IS TRUE provided'):
  print('it is true for n = 0,1,...,',gu - 1,'and at the positive integer'):
  print('roots of the leading term of S, which is'):
  sac := coeff(sa,N,gu):
  print(factor(sac)):
  else
print('since A(N,n) is not a left multiple of B(N,n), THE IDENTITY'):
print('IS FALSE'):
     fi:
quit
```

A.2. Ten sample inputs to the program

The above program ran successfully (on my PC) with the following ten inputs. To test them yourself just replace the part between "#THIS IS THE INPUT" and "#END OF INPUT" in the program by the ones given below.

(1) An identity of [3], see also [32, (1.1)]):

```
#THIS IS THE INPUT
expli := 0:
```

```
summand := (-1)^k*(n + 3*a + k - 1)!*(a + k - 1)!*3^k/
  ((n - k)!*(3*a/2 + k - 1)!*(3*a/2 + k - 1/2)!*k!*4^k):
conj := (n + 3)*N^3 - (3*a + n):
#END OF INPUT
```

(2) An identity of Gessel and Stanton [32, (5.24)]:

```
#THIS IS THE INPUT
expli := 1:
summand := (-1)^k*(n + k)!*8^k/((2*n - k)!*(1/3 + k)!*k!*9^k):
rhs := (-1)^n*(n - 1/2)!*(1/6)!*n!/((-1/2)!*(1/6 + n)!*(2*n)!*(1/3)!*3^n):
#END OF INPUT
```

(3) With this input the program will prove that [55, p.202]

$$a(n) := \sum_k \binom{n}{k}\binom{n+k}{k}$$

satisfies the recurrence

$$(n+2)a(n+2) - (6n+9)a(n+1) + (n+1)a(n) = 0.$$

```
#THIS IS THE INPUT
expli := 0:
summand := (n + k)!/(k!^2*(n - k)!):
conj := (n + 2)*N^2 - (6*n + 9)*N + (n + 1):
#END OF INPUT
```

(4) Saalschutz's identity (e.g., [36, 1.2.6, Ex. 31]):

```
#THIS IS THE INPUT
expli := 1:
summand := (b + k)!/(k!*(a - b + c - k)!*(n - k)!*(b - c + k)!*(b + k - a - n)!):
rhs := b!*c!*(a + n)!/(a!*(b - a)!*n!*(c - n)!*(a - b + c)!*(n + b - c)!):
#END OF INPUT
```

(5) The Vandermonde–Chu identity ($\equiv$, terminating case of Gauss's ${}_2F_1(1)$):

```
#THIS IS THE INPUT
expli := 1:
summand := 1/(k!*(n - k)!*(c - k)!*(b - c + k)!):
rhs := (n + b)!/(n!*b!*c!*(n + b - c)!):
#END OF INPUT
```

(6) It follows from [47, Section 8.2, formula (21), p.201] that the so-called "reduced straight Menage numbers" V_n satisfy the recurrence

$$V_{n+3} - (n+2)V_{n+2} - (n+2)V_{n+1} - V_n = 0. \qquad (*)$$

The explicit expression for V_n is given by (take $t = 0$, $k \to n - k$ in [47, Section 8.2, formula (6), p.197])

$$V_n = \sum_k \binom{n+k}{2k} k!(-1)^{n-k}. \qquad (**)$$

The program, with the following input, proves that indeed $(**)$ implies $(*)$.

```
#THIS IS THE INPUT
expli := 0:
summand := (-1)^(n-k)*(n+k)!*k!/((2*k)!*(n-k)!):
conj := N^3-(n+2)*N^2-(n+2)*N-1:
#END OF INPUT
```

(7) With the following input the program proves a recurrence obtained by Fasenmayer [45, pp.234–235, (2) and (11)].

```
#THIS IS THE INPUT
expli := 0:
summand := (-x)^k*(n+k)!/(k!^2*(k-1/2)!*(n-k)!):
conj := (n+3)*N^3-(3*n+7-4*x)*N^2+(3*n+4*x+5)*N-(n+1):
#END OF INPUT
```

(8) With the following input the program proves another recurrence of Fasenmayer [45, p.236, (14) and (22)]. Note that "al" stands for α and "be" stands for β.

```
#THIS IS THE INPUT
expli := 0:
summand := (-x)^k*(be+k)!/(k!^2*(al+k)!*(n-k)!):
conj := (n+3)^2*(al+n+3)*N^3-(3*(n+3)^2-3*(n+3)+1+al*(2*n+5)-(be+n+
3)*x)*N^2+(al+3*n+6-x)*N-1:
#END OF INPUT
```

(9) With the following input the program proves that the Jacobi polynomials satisfy their recurrence [45, p.255, (4) and p.263, (1) (Section 137)]. Once again "al" stands for α and "be" stands for β.

```
#THIS IS THE INPUT
#Recurrence or Jacobi
expli := 0:
summand := (al+be+n+k)!*((x-1)/2)^k/(k!*(n-k)!*(al+k)!):
conj := (2*n+4)*(al+be+2*n+2)*(al+n+1)*(al+n+2)*N^2-
  (al+be+2*n+3)*((al^2-be^2)+x*(al+be+2*n+4)*(al+be+2*n+2))*(al+n+
  1)*N+2*(al+n+1)*(be+n+1)*(al+be+2*n+4)*(al+be+n+1):
#END OF INPUT
```

(10) With the following input the program proves that the Hahn polynomials indeed satisfy their three-term recurrence [37]. Here "a" stands for α, "b" stands for β, and "N1" stands for N. (We cannot use "N" since it has a special meaning, namely the shift operator in the n-direction.)

```
#THIS IS THE INPUT
#Recurrence for Hahn
expli := 0:
summand := (-1)^k*(n+a+b+k)!*(N1-k)!/((n-k)!*(x-k)!*(a+k)!*k!):
conj := (n+a+2)*(N1-n-1)*(n+2)*(2*n+a+b+2)*N^2-
  ((n+a+b+2)*(n+a+2)*(N1-n-1)*(2*n+a+b+2)+(n+1)*(n+b+1)*(n+a
  +b+N1+2)*(2*n+a+b+4)-(2*n+a+b+2)*(2*n+a+b+3)*(2*n+a+b+
  4)*x)*N+(n+b+1)*(n+a+b+N1+2)*(n+a+b+1)*(2*n+a+b+4):
#END OF INPUT
```

A.3. A sample output

(1) When the program was run with the original (trivial) input, the output was as follows: (I have deleted the comments and the MAPLE logo)

F(n,k) is annihilated by the following, k-free, operator(check!)
(Warning: it should be in the standard form in which n is in)
front of N, if it is not, it is still meant that way
$-K-1+nN(K-1)+(K-1)N+nN+N$
Therefore, putting K = 1, we get an operator A(N,n) annihilating the rhs
$nN+N-2$
the conjectured rhs is annihilated by the following operator B(N,n)
$(n+1)N-2$
Now A(N,n) = S(N,n) * B(N,n)(check!), where S(N,n) is 1
Since S has order, 0, in N, THE IDENTITY IS TRUE provided
it is true for n = 0,1, . . . , − 1, and at the positive integer
roots of the leading term of S, which is 1
words used = 57169, alloc = 36864, time = 8.82

References

[1] G. Almkvist, Zeilbergers bevismaskin, *Elementa*, to appear (in Swedish).

[2] G. Almkvist and D. Zeilberger, The method of differentiation under the integral sign, *J. Symbolic Comput.*, to appear.

[3] G.E. Andrews, Connection coefficient problems and partitions, in: D. Ray-Chaudhuri, Ed., *AMS Proc. Symposia in Pure Mathematics* **34** (Amer. Mathematical Soc., Providence, RI, 1979) 1–24.

[4] G.E. Andrews, *q-Series: Their Development And Applications in Analysis, Number Theory, Combinatorics, Physics, and Computer Algebra*, CBMS Ser. **66** (Amer. Mathematical Soc., Providence, RI, 1986).

[5] R.A. Askey, Review of [60], *Math. Rev.* **83f:33001** (1983).

[6] R.A. Askey, Preface, in: R.A. Askey, T.H. Koornwinder and W. Schempp, Eds., *Special Functions: Group Theoretical Aspects and Applications* (Reidel, Dordrecht, 1984).

[7] R.A. Askey and G. Gasper, Inequalities for polynomials, in: A. Baernstein et al., Eds., *The Bieberbach Conjecture* (Amer. Mathematical Soc., Providence, RI, 1986).

[8] R.A. Askey and J.A. Wilson, *Some Basic Hypergeometric Orthogonal Polynomials that Generalize Jacobi Polynomials,* Mem. Amer. Math. Soc. **318** (Amer. Mathematical Soc., Providence, RI, 1985).

[9] W.N. Bailey, *Generalized Hypergeometric Series*, Cambridge Math. Tracts **32** (Cambridge Univ. Press, London, 1935); (reprinted: Hafner, New York, 1964).

[10] I.N. Bernstein, Modules over a ring of differential operators, study of the fundamental solutions of equations with constant coefficients, *Functional Anal. Appl.* **5** (2) (1971) 1–16 (in Russian); 89–101 (English translation).

[11] I.N. Bernstein, The analytic continuation of generalized functions with respect to a parameter, *Functional Anal. Appl.* **6** (4) (1972) 26–40 (in Russian); 273–285 (English translation).

[12] J.-E. Björk, *Rings of Differential Operators* (North-Holland, Amsterdam, 1979).

[13] J.-E. Björk, Differential systems on algebraic manifolds, preprint, Univ. Stockholm, 1984.

[14] B. Buchberger, Grobner bases—an algorithmic method in polynomial ideal theory, in: N.K. Bose, Ed., *Multidimensional System Theory* (Reidel, Dordrecht, 1985) Chapter 6.

[15] B. Cipra, How the grinch stole mathematics, *Science* **245** (1989) 595.

[16] L. de Brange, A proof of the Bieberbach conjecture, *Acta Math.* **154** (1985) 137–152.

[17] L. Ehrenpreis, *Fourier Analysis in Several Variables* (Wiley, New York, 1970).

[18] L. Ehrenpreis, Hypergeometric functions, preprint, Temple Univ., PA.

[19] M.C. Fasenmayer, Some generalized hypergeometric polynomials, *Bull. Amer. Math. Soc.* **53** (1947) 806–812.
[20] D. Foata, Combinatoire des identités sur les polynômes orthogonaux, in: *Proc. Internat. Congress of Math.*, Warsaw, 1983 (Varsovie, 1983) 1541–1553.
[21] A. Galligo, Some algorithmic questions on ideals of differential operators, in: B.F. Cavines, Ed., *EUROCAL '85, Vol. 2*, Lecture Notes in Comput. Sci. **204** (Springer, Berlin, 1985).
[22] G. Gasper, Summation, transformation, and expansion formulas for bibasic series, *Trans Amer. Math. Soc.* **312** (1989) 257–277.
[23] G. Gasper and M. Rahman, An indefinite bibasic summation formula and some quadratic, cubic, and quartic summation and transformation formulas, *Canad. J. Math.*, to appear.
[24] G. Gasper and M. Rahman, *Basic Hypergeometric Series* (Cambridge Univ. Press, Cambridge, 1990).
[25] I.M. Gel'fand, General theory of hypergeometric functions, *Dokl. Akad. Nauk USSR* **286** (1) (1986) 14–18; translated in: *Soviet Math. Dokl.* **33** (3) (1986) 573–577.
[26] I.M. Gel'fand and S.I. Gel'fand, Generalized hypergeometric equations, *Dokl. Akad. Nauk USSR* **288** (2) (1986) 279–282; translated in: *Soviet Math. Dokl.* **33** (3) (1986) 643–646.
[27] I.M. Gel'fand and M.I. Graev, A duality theorem for general hypergeometric functions, *Dokl. Akad. Nauk USSR* **289** (1) (1986) 25–31; translated in: *Soviet Math. Dokl.* **34** (1) (1987) 9–13.
[28] I.M. Gel'fand and A.V. Zelevinskii, Algebraic and combinatorial aspects of the general theory of hypergeometric functions, *Functional Anal. Appl.* **20** (3) (1986) 17–34 (in Russian); 183–197 (English translation).
[29] I. Gessel, Two theorems on rational power series, *Utilitas Math.* **19** (1981) 247–254.
[30] I. Gessel, Counting Latin rectangles, *Bull. Amer. Math. Soc. (N.S.)* **16** (1987) 79–82.
[31] I. Gessel, Symmetric functions and P-recursiveness, preprint, Brandeis Univ.
[32] I. Gessel and D. Stanton, Strange evaluations of hypergeometric series, *SIAM J. Math. Anal.* **13** (1982) 295–308.
[33] B. Gosper, Letters to R. Askey and D. Stanton, private communication.
[34] I.P. Goulden, Exact values for degree sums over strips of Young diagrams, preprint, Univ. Waterloo.
[35] D. Gouyou-Beauchamps, Standard Young tableaux of height 4 and 5, preprint.
[36] D.E. Knuth, *Fundamental Algorithms, The Art of Computer Programming, Vol. 1* (Addison-Wesley, Reading, MA, 2nd ed., 1973).
[37] J. Labelle, Tableau d'Askey, in: C. Brezinski et al., Eds., *Polynômes Orthogonaux et Applications*, Bar-le-Duc, 1984, Lecture Notes in Math. **1171** (Springer, Berlin, 1985) p. vii.
[38] L. Lipshitz, The diagonal of a D-finite power series is D-finite, *J. Algebra* **113** (1988) 373–378.
[39] L. Lipshitz, D-finite power series, *J. Algebra*, to appear.
[40] I.G. Macdonald, Some conjectures for root systems, *SIAM J. Math. Anal.* **13** (1982) 91–143.
[41] W. Miller, *Symmetry and Separation of Variables*, Encyclopedia Math. Appl. **4** (Addison-Wesley, London, 1977).
[42] S.C. Milne, Hypergeometric series well poised in SU(n) and a generalization of Biedenharn's G function, *Adv. in Math.* **36** (1980) 169–211.
[43] V.P. Palamadov, *Linear Differential Operators with Constant Coefficients*, Grundlehren Math. Wiss. **168** (Springer, Berlin, 1970).
[44] M. Perlstadt, Recurrences for sums of powers of binomial coefficients, *J. Number Theory* **27** (1987) 304–309.
[45] E.D. Rainville, *Special Functions* (Macmillan, New York, 1960); (reprinted: Chelsea, New York, 1971).
[46] A. Regev, Asymptotic values for degrees associated with strips of Young diagrams, *Adv. in Math.* **41** (1981) 115–136.
[47] J. Riordan, *An Introduction to Combinatorial Analysis* (Princeton Univ. Press, Princeton, NJ, 1980); (originally published by (Wiley, New York, 1958)).
[48] R. Roy, Binomial identities and hypergeometric series, *Amer. Math. Monthly* **94** (1987) 37–46.
[49] J.T. Stafford, Module structure of Weyl algebras, *J. London Math. Soc. (2)* **18** (1978) 429–442.
[50] R. Stanley, Differentiably finite power series, *European J. Combin.* **1** (1980) 175–188.
[51] J.J. Sylvester, A method of determining by mere inspection the derivatives from two equations of any degree, *Philos. Magazine* **16** (1840) 132–135; (*Collected Works I*, 54–57).
[52] N. Takayama, Holonomic solutions of Weisner's operator, preprint, Tokushima Univ.
[53] N. Takayama, Grobner basis and the problem of contiguous relations, *Japan J. Appl. Math.* **6** (1989) 147–160.
[54] J.W. Tukey, *Amer. Statist.* **40** (1986) 74.
[55] A. van der Poorten, A proof that Euler missed... Apery's proof of the irrationality of $\zeta(3)$, *Math. Intelligencer* **1** (1979) 195–203.

[56] B.L. van der Waerden, *Modern Algebra, Volumes I and II* (Frederick Ungar, New York, 1940).
[57] H.S. Wilf, What is an answer?, *Amer. Math. Monthly* **89** (1982) 289–292.
[58] H.S. Wilf and D. Zeilberger, Towards computerized proofs of identities, *Bull. Amer. Math. Soc.*, to appear.
[59] H.S. Wilf and D. Zeilberger, Rational functions certify combinatorial identities, *J. Amer. Math. Soc.*, to appear.
[60] D. Zeilberger, Sister Celine's technique and its generalizations, *J. Math. Anal. Appl.* **85** (1982) 114–145.
[61] D. Zeilberger, A fast algorithm for proving terminating hypergeometric identities, preprint.
[62] D. Zeilberger, The method of creative telescoping, preprint.
[63] D. Zeilberger, The method of creative telescoping for q-series, in preparation.