CLOSED FORM (pun intended!)

Doron Zeilberger *

Appeared in: Contemporary Mathematics 143 (1993), 579-608.

In fond memory of Emil Grosswald, Great Number Theorist and Wonderful Mensch.

0. Introduction

Mathematics is infinitely wide, while the language that describes it is finite. It follows from the pigeonhole principle that there exist distinct concepts that are referred to by the same name. Mathematics is also infinitely deep and sometimes entirely different concepts turn out to be intimately and profoundly related. When the two phenomena coincide, one has a *mathematical pun.*

The phrase *closed form* has at least two distinct meanings. The first meaning is that of "explicit", "nice", or "in finite terms". For example the definite sum of the binomial coefficients

$$\sum_k \binom{n}{k}$$

can be expressed in closed form, namely 2^n, as can the sum of their squares

$$\sum_k \binom{n}{k}^2 ,$$

which is known to be $(2n)!/n!^2$. On the other hand, the sum of the cubes

$$\sum_k \binom{n}{k}^3 ,$$

is not expressible in closed form. Similarly, $\int xexp(x^2)dx$ can be expressed in closed form while $\int exp(x^2)dx$ cannot.

The other meaning of *closed form* is that of *closed (exterior) differential form*is closed if it is annihilated by the exterior derivative d. I will define a natural discrete-continuous analog of d, also to be denoted by d, and show that any "nice" identity

$$\sum_{\mathbf{k}} \int F(\mathbf{n},\mathbf{k},\mathbf{x},\mathbf{y})d\mathbf{y} \; = \; 1 \;\; ,$$

where **n**,**k** are discrete multi-variables and **x**, **y** are continuous multi-variables, owes its existence to the fact that the integrand-summand $F(n,k,x,y)dy\delta k$ is one term of a so-called "closed holonomic exterior differential-difference form", and if in luck, that form will have all its other components nice too, in which case we have what I will call a *WZ form.* Conversely, I will show that any WZ form gives rise to several "Closed Form" (in the previous sense of the phrase) identities.

* Department of Mathematics, Temple University, Philadelphia, PA19122. Supported in part by NSF grant DMS8800663.

The theory of WZ forms is a direct and natural generalization of that of "WZ pair" ([WZ1],[WZ2]). The pair $(F(n,k), G(n,k))$ is a WZ pair iff $F(n,k)\delta k + G(n,k)\delta n$ is a WZ form. I am very grateful to W for discovering the notion of WZ pair, and for countless conversations.

A cheap way to manufacture closed differential (-recurrence) forms of degree r is to take a form of degree $r-1$, and apply d to it, getting the class of so-called *exact forms.* As we all know, since $d^2 = 0$, every exact form is closed, and thus we have a quick way of generating many (in fact an infinite number of) "nice" identities. Luckily, this embarrassment of riches is deceptive, since I will show that identities obtained in this way are trivial (I will explain what I mean by *trivial* later on). The *nice and interesting* identities are precisely those that come from closed forms that are not exact. The problem of classifying all "nice and interesting" identities thus boils down to computing what I will call the "WZ cohomology".

1. A Short Review Of The Calculus of Exterior Differential Forms

The calculus of exterior differential forms is of fundamental importance in Differential Geometry and elsewhere. Both Flanders's book[F] and Spivak's book [S] are excellent introductions. Let $(x_1, ..., x_n)$ be continuous variables and let $(dx_1, ..., dx_n)$ satisfy the (anti-) commutation relations : $dx_i dx_j = -dx_j dx_i$. An (exterior) *differential form* is a linear combination, with coefficients that are functions (or distributions, or whatever you can differentiate) of "words" in the alphabet $dx_1, ..., dx_n$. For a subset $I = \{i_1, ..., i_k\}$ of $1, 2, ..., n$, let

$$dx_I := dx_{i_1} ... dx_{i_k} \quad .$$

A general differential form in $(x_1, ..., x_n)$ can be written

$$\omega = \sum_{I c \{1,2,\dots,n\}} f_I dx_I \ , \tag{1.1}$$

where the coefficients f_I are functions in $(x_1, ..., x_n)$.

A *homogeneous r-form* is an expression of the form (1.1) where all the subsets I have the same cardinality r. Of course a $0-form$ is just a scalar function.

Recall that the *exterior derivative* d of a form ω is defined as follows. For a $0-form\ f$,

$$df := \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i \ , \tag{1.2}$$

and for a general form ω, given by (1.1),

$$d\omega = \sum_{I c \{1,2,\dots,n\}} (df_I) dx_I \ . \tag{1.3}$$

For example

$$d(f dx + g dy + h dz) = (g_x - f_y) dx dy + (h_x - f_z) dx dz + (h_y - g_z) dy dz \ ,$$

(the coefficients of which are components of *curl* (f, g, h), and

$$d(f dydz + gdzdx + hdxdy) = (f_x + g_y + h_z)dxdydz,$$

producing the divergence.

Recall the fundamental Stokes's Theorem, that for any form ω, and any oriented manifold Ω, which for our purposes can be taken to be a subset of R^n, one has

$$\int_{\partial\Omega} \omega = \int_{\Omega} d\omega \ . \qquad (STOKES)$$

It is well known, and easily checked, that $d^2 = 0$ and $d(\omega\theta) = (d\omega)\theta + (-1)^r \omega(d\theta)$, where ω is homogeneous of degree r.

A differential form ω is said to be a *closed form* if $d\omega = 0$, and it is said to be *exact* if it can be written as $\omega = d\theta$, for some form θ. Of course every exact form is automatically closed, and the quotient space of closed modulo exact, for a given manifold, is the celebrated *de Rham cohomology.* It is well known (e.g. [F]) that for "flat" space R^n, the de Rham cohomology is trivial, i.e. every closed form is exact.

2. The Calculus of Exterior Difference Forms

The discrete analog of the calculus of differential forms is much simpler than the continuous case, since we don't have to fuss about such analytical notions as convergence. We consider functions $F(m_1, ..., m_n)$ defined on the rectangular lattice Z^n. The analogs of partial derivatives are partial forward difference operators:

$$\Delta_i F(m_1, ..., m_i, ..., m_n) := F(m_1, ..., m_i + 1, ..., m_n) - F(m_1, ..., m_i, ..., m_n) \ . \qquad (2.1)$$

A unit r-cube in Z^n is a set of points of the following form: for some subset I of $1, 2, 3, ..., n$, of cardinality r, m_i is fixed (say a_i) if i is not in I, and can take one of two consecutive values (say a_i and $a_i + 1$) if i belongs to I. Of course a unit $r - cube$ has 2^r points. The *interior* of a unit $r-$cube is its "lowest, leftmost" corner, $(a_1, ..., a_n)$. The boundary of an r-cube consists of the union of the $2r$ $(r-1)$-cubes obtained by freezing one of the m_i, for i in I, to be a_i or $a_i + 1$.

The beauty of the notation for integration is that the variable with respect to which things are being integrated is explicitly indicated:

$$\int f dx$$

means that the function f is being integrated with respect to the variable x, regardless of whatever other variables or parameters f may depend upon. For a definite integral, we write

$$\int_a^b f dx \ ,$$

rather than

$$\int_{x=a}^{b} f \ .$$

This is unlike the corresponding convention for summation in which the summation variable has to be indicated below the *SIGMA*, and does not appear within the sum itself. To correct this inequity, Knuth introduced a useful notation (e.g. [GKP], p.48) which we shall adopt. In that notation the sum

$$\sum_{k=a}^{b} f(k)$$

is re-expressed as

$$\sum_{a}^{b} f(k)\delta k \ .$$

Similarly, a multiple sum

$$\sum_{k_1,\ldots,k_r} f(k_1,...,k_r)$$

is written

$$\sum f(k_1,...,k_r)\delta k_1 ... \delta k_r \ .$$

The mark of a good notation is that it leads to new theory. For example, matrices started out as shorthand for writing systems of linear equations, and what emerged was linear algebra. In our case, what comes out of Knuth's notation is the discrete analog of the calculus of differential forms.

Let $(m_1,...,m_n)$ be discrete variables and let $\delta m_1,...,\delta m_n$ be "indeterminates" satisfying $\delta m_i \delta m_j = -\delta m_j \delta m_i$. An (exterior) *difference form* is a linear combination, with coefficients that are discrete functions, of "words" in the alphabet $\delta m_1,...,\delta m_n$. For a subset $I = \{i_1,...,i_k\}$ of $1,2,...,n$, let

$$\delta m_I := \delta m_{i_1} ... \delta m_{i_k} \ .$$

A general difference form in $(m_1,...,m_n)$ can be written

$$\omega = \sum_{I c \{1,2,\ldots,n\}} f_I \delta m_I \ , \tag{2.2}$$

where the coefficients f_I are functions of $(m_1,...,m_n)$.

An oriented unit cube is a unit cube with a choice of sign ±. An *oriented r-manifold* in Z^n is any union of oriented unit r-cubes. Every $r-manifold$ can be considered as an oriented one by taking all the signs of its constituent cubes to be +. The *boundary* of an oriented unit r-cube B, denoted by ∂B, is the union of all the $2r$ $(r-1)$-faces, with the sign of each face determined in the usual way ([S], p. 98, fig. 4-4): the faces adjacent to the lower-left corner alternate in sign and opposite faces have opposite signs. The boundary of an oriented discrete manifold is the "sum"of all the boundaries of its constituent unit cubes. (Of course for a connected manifold, of a fixed sign, all interior walls cancel each other in that sum, and only external walls survive.)

We are now ready to define "integration" of an $r-form$ over an oriented $r-$manifold. By additivity, it is enough to define what is

$$\sum_B f_I \delta m_I \quad , \tag{2.3}$$

where B is an oriented unit $r-$cube whose lower left corner is a typical point $(m_1, ..., m_n)$. If the subset of "active" coordinates of B coincides with I, we define the above to be $\pm f(m_1, ..., m_n)$, where we assume that $i_1 < ... < i_r$, and ± is the sign of B. If the set of active coordinates of B does not coincide with I then (2.3) is defined to be 0.

The discrete analog of the exterior derivative d, the *exterior difference*, δ, is defined in analogy with (1.1). For a $0-form$ f,

$$\delta f := \sum_{i=1}^{n} (\Delta_i f) \delta m_i \ , \tag{2.4}$$

and for a general form ω of (2.2) :

$$\delta \omega = \sum_{I c\{1,2,...,n\}} (\delta f_I) \delta m_I \ . \tag{2.5}$$

Let ω be an $r-form$, and Ω any oriented discrete $(r+1)-manifold$, i.e. a union of oriented unit $(r+1)-cubes$. We have

The DISCRETE STOKES Theorem

$$\sum_{\partial\Omega} \omega = \sum_{\Omega} \delta\omega \ .$$

The proof is even more trivial than its continuous namesake (see [S], p. 104). By additivity, it is enough to prove it for forms with one term and manifolds Ω consisting of a single $r-cube$, and this last task is immediate.

3. Closed Difference Forms

As an immediate corollary of the DISCRETE STOKES theorem, we have

Theorem 1: Let ω be a discrete *closed* difference $r-form$:

$$\omega = \sum_{|I|=r} f_I \delta m_I \ , \tag{3.1}$$

and assume that for some subset I_0 of cardinality r of $1, 2, ..., n$, its complement J_0 has the property that all the coefficients of ω have finite support as functions of the variables m_{I_0}, for any fixed choice of the variables m_{J_0}. The function

$$g(m_{J_0}) := \sum f_{I_0}(m_{I_0}, m_{J_0}) \delta m_{I_0}$$

is identically constant in m_{J_0}.

Proof: Let Ω be the region between two parallel hyperplanes $m_{J_0} = a_{J_0}$ and $m_{J_0} = a'_{J_0}$. The difference $g(a_{J_0}) - g(a'_{J_0})$ is the integral of ω over the boundary of Ω, since the contribution is zero from the other terms, thanks to the assumption that ω vanishes when at least one of the m_{I_0} coordinates is infinite. By the DISCRETE STOKES theorem, that sum is the sum over Ω of $\delta\omega$, but the latter is 0, since ω is a closed form.

4. CLOSED FORM

Theorem 1 by itself is almost a tautology. It is easy to see that the "de Rham cohomology" for arbitrary discrete functions on Z^n is trivial, i.e. every closed form is exact, and so to obtain all closed forms of a certain degree, all one has to do is take all forms of degree one less and apply the exterior difference operator δ to them.

However, we are not interested in *all* identities, only in the *nice*ones! For us "nice" will mean Closed Form.

Definition: A function $f(m_1, ..., m_n)$ is *Closed Form* (CF) if for each of its variables m_i $(i = 1, ..., n)$, the quotient

$$\frac{f(m_1, ..., m_{i-1}, m_i + 1, m_{i+1}, ..., m_n)}{f(m_1, ..., m_{i-1}, m_i, m_{i+1}, ..., m_n)} \tag{4.1}$$

is a rational function of $m_1, ..., m_n$, i.e. a quotient of two polynomials in $m_1, ..., m_n$.

A typical example of a Closed Form function is the factorial expression $(c + a_1 m_1 + a_2 m_2 + ... + a_n m_n)!$, where $a_1, ..., a_n$ are (positive or negative) integers, and c is any constant or indeterminate. Similarly, the reciprocal of such an expression is CF. Since the product of CF functions is again CF, any quotient of products of such expression, times powers $x_1^{msub1} \cdots x_n^{msubn}$ is CF.

Next we define formally what we mean by a *CF identity*

Definition: A multi-sum identity is CF if both its summand and sum are CF.

The general form of a CF identity is

$$\sum_{m_1, ..., m_a} f(m_1, ..., m_n) = g(m_{a+1}, ..., m_n) \tag{4.2}$$

where f is CF in $m_1, ..., m_n$, and g is CF in $m_{a+1}, ..., m_n$. Since the quotient of CF functions is again CF, every CF identity is equivalent to one in which the right side is 1, simply divide through by the right side! This "innocent observation" of Herb Wilf was the starting point of the theory of WZ pairs and consequently WZ forms.

We would like to study "all" CF identities

$$\sum f(m_1, ..., m_a, m_{a+1}, ..., m_n)\delta m_1 ... \delta m_a = 1 \quad . \tag{4.3}$$

The big surprise is that the converse of the tautological theorem 1 is true, for CF identities!, and not in a tautological manner! Whenever we have a CF identity, it is so because the summand is one term in a closed difference form, that can be algorithmically constructed, and all whose terms can be described explicitly. Furthermore, in most cases in practice, all the terms of that closed difference form are CF themselves. At any rate, at the very worst, they belong to the class of so-called holonomic functions, and can be actually exhibited. In the WZ-pairs miracle [WZ1], we set out to prove one identity and we also got the discovery and proof of another one, often more interesting, as a bonus. Now, whenever we have an identity like (4.3), we can always reconstruct the whole closed difference form ω of which the summand $f(m_1, ..., m_a, m_{a+1}, ..., m_n)\ \delta m_1 ... \delta m_a$ is just one term, with all its $\binom{n}{a}$ terms, and we get *free of charge* $\binom{n}{a} - 1$ identities as bonuses, complete with proofs! In fact the same proof that demonstrated the original identity, works for all the bonuses simultaneously, and merely consists of the purely mechanical task of verifying that the proposed form ω is indeed closed. I will also describe an algorithm for constructing ω.

5. Holonomic Functions Identities in Z^n

Holonomic functions are natural generalizations of CF functions. Another way of saying that a function f is CF is to say that it satisfies *homogeneous* ordinary linear recurrences with polynomial coefficients of the *first order* in each of its variables:

$$P^{(i)}(m_1, ..., m_n)f(m_1, ..., m_i + 1, ..., m_n) + Q^{(i)}(m_1, ..., m_n)f(m_1, ..., m_n) = 0 \ , i = 1, ..., n \tag{5.1}$$

for some polynomials $P^{(i)}$, $Q^{(i)}$, $(i = 1, ..., n)$. A *holonomic function*, roughly speaking, is a function on Z^n that satisfies ordinary recurrence equations, not necessarily of the first order, in each of its variables:

$$\sum_{k=0}^{L_i} P_k^{(i)} f(m_1, ..., m_i + k, ..., m_n) = 0 \, . \tag{5.2}$$

To guarantee that f is genuinely holonomic we must insist that the leading coefficients ${P_{L_i}}^{(i)}$ do not have common zeroes. There are also holonomic functions that do not have the above form: for example the Kronecker $\delta_{m,n}$. The full definition of holonomicity is that the set of linear difference equations with polynomial coefficients satisfied by the function forms a so-called "maximally over-determined system" [Z1].

The ring of linear partial difference operators with polynomial coefficients in Z^n is generated by $m_1, ..., m_n$ and $\Delta_1, ..., \Delta_n$. However, it is more convenient and congenial to take the generators to be $m_1, ..., m_n$ and $E_1, ..., E_n$, where E_i are the *shift operators*:

$$E_i f(m_1, ..., m_i, ..., m_n) := f(m_1, ..., m_i + 1, ..., m_n) \ . \tag{5.3}$$

Of course, we have

$$\Delta_i \ = \ E_i - 1 \ , \tag{5.4}$$

where "1" is the identity operator.

Any linear difference operator with polynomial coefficients can be expressed as $P(m_1, ..., m_n, E_1, ..., E_n)$, where P is a polynomial in the indeterminates $(m_1, ..., m_n, E_1, ..., E_n)$. It is readily seen that they satisfy the "commutation relations":

$$E_i m_j \ = \ m_j E_{i \ i} \neq j \ ; \ \ E_i E_j \ = \ E_j E_i \ , \ \ m_i m_j \ = \ m_j m_i \ \ , \tag{5.5}$$

and

$$E_i m_i \ = \ m_i E_i + E_i \ \ . \tag{5.6}$$

The following is a useful mnemonic. The m's are men and the E's are women, m_i is married to E_i. All men pass each other with no problem and all women pass each other without a trace, as do a man and a woman that are not married to each other. However whenever a woman wants to pass to the right of her husband, a new baby girl is born, that is a clone of her mother.

In [Z1] I showed how to adjust Sylvester's dialytic elimination to the algebra of linear difference operators. Given two operators $P(m_1, ..., m_n, E_1, ..., E_n)$ and $Q(m_1, ..., m_n, E_1, ..., E_n)$ it is possible, in general, to eliminate one of the variables, say m_1. This means that one can find linear difference operators with polynomial coefficients $A(m_1, ..., m_n, E_1, ..., E_n)$ and $B(m_1, ..., m_n, E_1, ..., E_n)$ such that $AP+BQ$ does not involve m_1. Of course this is only possible when P and Q are "independent" in a certain technical sense. For example if both of them are left multiples of m_1 then every combination $AP + BQ$ will still have that property, and there is no getting rid of m_1. More generally, if P_1, P_2, ..., P_r are operators, it is possible, if all goes well, to eliminate $r - 1$ out of the $2n$ indeterminates $m_1, ..., m_n, E_1, ..., E_n$.

Given a discrete function f, the set of linear difference operators with polynomial coefficients annihilating it is a left ideal in the whole ring, and is denoted by $I(f)$. It is a left ideal since if P annihilates f, so does obviously AP, for *any* operator A. If f is holonomic, then the ideal $I(f)$ is "as big as possible" (which means that the Hilbert dimension of the quotient ring $K < m_1, ..., m_n, E_1, ..., E_n > /I(f)$ viewed as a left $K < m_1, ..., m_n, E_1, ..., E_n >$ module, is n, which is the least that it can possibly be (thanks to the celebrated Bernstein inequality [Ber], [Bj]). It follows from the general theory of holonomic systems ([Ber],[Bj], see also [Z1]) that if f is holonomic, then $I(f)$ contains an operator that does not involve $m_2, ..., m_n$. Let that operator be $A(m_1, E_1, E_2, ..., E_n)$. Now write

$$A(m_1, E_1, ..., E_n) \ = \ A_1(m_1, E_1) \ - \ (E_2 - 1)A_2 \ - \ ... \ - \ (E_n - 1)A_n \ . \tag{5.7}$$

Applying both sides of (5.7) on our holonomic f, the left side is zero and we have proved the following

Theorem 2: Let $f(m_1, ..., m_n)$ be holonomic in all its variables. For each of its variables m_i, there exists an operator $A^{(i)}(m_i, E_i)$, and operators $A_j^{(i)}$ $(j \ = \ 1, ..., i-1, i+1, ..., n)$ such that

$$A^{(i)}(m_i, E_i)f \ = \ \sum_{\substack{1<=j<=n \\ j \neq i}} \Delta_j A_j^{(i)} f \ \ . \tag{5.8}$$

Note that we have proved *more* than the theorem. We have proved that the operators $A_j^{(i)}$ can be taken to be independent of $m_2, ..., m_n$, which is an unnecessary extravagance. However, to prove the mere existence of such operators we can be big spenders, since window-shopping does not cost anything. Once we know that our shopping expedition is guaranteed to succeed we should do a better job of shopping, and get the cheapest possible operators.

Now define

$$G_j^{(i)}(m_1, ..., m_n) := A_j^{(i)} f(m_1, ..., m_n) \tag{5.9}$$

$G_j^{(i)}$ are obviously holonomic, since any linear difference operator with polynomial coefficients applied to any holonomic function yields a holonomic function. Furthermore if $f(-, m_i, -)$ has finite support in $(m_1, ..., m_{i-1}, m_{i+1}, m_n)$, for any fixed m_i, so do the $G_j^{(i)}$. We thus have

Theorem 3: For any holonomic $f(m_1, ..., m_n)$, and any of its variables m_i, there exists a non-zero operator $A^{(i)}(m_i, E_i)$, and holonomic $G_j^{(i)}$ (that have the form (5.9)) such that

$$A^{(i)}(m_i, E_i) f(m_1, ..., m_n) \ = \ \sum_{\substack{1<=j<=n \\ j \neq i}} \Delta_j G_j^{(i)} \ \ . \tag{5.10}$$

Furthermore, if $f(-, m_i, -)$ has finite support for any fixed m_i, so do the $G_j^{(i)}$.

An immediate consequence is

Corolary 4: Let $f(m_1, ..., m_n)$ be holonomic, and let m_i be a variable such that for any specific $m_i \ = \ c$, there are only finitely many $(m_1, ..., m_{i-1}, m_{i+1}, ..., m_n)$ for which $f(m_1, ..., m_{i-1}, c, m_{i+1}, ..., m_n)$ is non-zero. Let

$$a_i(m_i) := \sum_{m_1, ..., m_{i-1}, m_{i+1}, ..., m_n} f(m_1, ..., m_n) \ \ , \tag{5.11}$$

then $a_i(m_i)$ satisfies an ordinary linear recurrence

$$A^{(i)}(m_i, E_i)a_i(m_i) \; = \; 0 \; .$$

Proof: Sum (5.10) w.r.t to $m_1, ..., m_{i-1}, m_{i+1}, ..., m_n$. The sum on the right is a sum of $n-1$ telescoping series, each of finite support.

Let's consider now identities of the form

$$\sum_{m_1, ..., m_{i-1}, m_{i+1}, ..., m_n} f(m_1, ..., m_n) \; \equiv \; 1 \; .$$

How to prove them? Let's call the left side $a_i(m_i)$. Corollary 4 manufactures an operator $A^{(i)}(m_i, E_i)$ which annihilates $a_i(m_i)$. To prove that $a_i(m_i)$ is indeed identically 1, we must show that 1 is also annihilated by same, and that $a_i(m_i) \; = \; 1$ for the first L_i values of m_i, where L_i is the order (i.e. degree in E_i) of $A^{(i)}$. Alternatively, Euclid-dividing $A^{(i)}$ by Δ_i (from the right!) gives

$$A^{(i)}(m_i, E_i) \; = \; \bar{A}^{(i)}(m_i, E_i)\Delta_i + b(m_i) \; ,$$

for some operator $\bar{A}^{(i)}$, and for some function $b(m_i)$ that must be identically zero, since $A^{(i)} 1 \; == \; 0$, so we can write

$$A^{(i)}(m_i, E_i) \; = \; \bar{A}^{(i)}(m_i, E_i)\Delta_i \; . \tag{5.12}$$

Now for each $1 <= j <= n, j \neq i$, let's find new functions $\bar{G}_j^{(i)}$ such that

$$\bar{A}^{(i)}(m_i, E_i)\bar{G}_j^{(i)} \; = \; - G_j^{(i)} \; , \tag{5.13}$$

subject to appropriate initial conditions to be specified momentarily. It is well known, and easy to see, that this is always possible, although the $\bar{G}_j^{(i)}$ usually don't inherit the "finite support" property of the $G_j^{(i)}$. Since $G_j^{(i)}$ are holonomic, so are the $\bar{G}_j^{(i)}$.

Substituting (5.12) and (5.13) into (5.10), and extracting $\bar{A}_i(m_i, E_i)$ out, we get

$$\bar{A}^{(i)}(m_i, E_i) \, (\, \Delta_i f(m_1, ..., m_n) + \sum_{1<=j<=n, j \neq i} \Delta_j \bar{G}_j^{(i)} \,) \; = \; 0 \; . \tag{5.14}$$

Now we are ready to specify the initial conditions promised earlier, that determine the solutions of (5.13): we choose them in such a way that

$$\Delta_i f(m_1, ..., m_n) + \sum_{1<=j<=n, j \neq i} \Delta_j \bar{G}_j^{(i)} = 0 \;\; , for \, m_i = 0, 1, ..., L_i - 1 \; . \tag{5.15}$$

It is easy to see that this is always possible, and for this choice of $\bar{G}_j^{(i)}$, it follows from (5.14) that

$$\Delta_i f(m_1, ..., m_n) + \sum_{1<=j<=n, j \neq i} \Delta_j \bar{G}_j^{(i)} = 0 \quad , for\, all\, m_i \ . \tag{5.16}$$

We have just proved (renaming $\bar{G}_j^{(i)} = H_j$, $f = H_i$, in (5.16))

Theorem 5: Let $f(m_1, ..., m_n)$ be holonomic.

$$\sum f(m_1, ..., m_n) \delta m_1 ... \delta \hat{m_i} ... \delta m_n == constant \tag{5.17}$$

if and only if there exists a *closed difference* form ω:

$$\omega := \sum_{j=1}^{n} H_j \, \delta m_1 ... \delta \hat{m}_j ... \delta m_n \ , \tag{5.18}$$

such that the components H_j are all holonomic, and $H_i = f$.

A similar analysis holds when the number of variables that are being summed over is bigger than one. We have

Theorem 6: Let $f(m_1, ..., m_n)$ be holonomic. Let I be any subset of $1, 2, ..., n$. An identity of the form

$$\sum f(m_1, ..., m_n) \delta m_I == constant \ , \tag{5.19}$$

where the constant is independent of the variables in the complement of I, is possible if and only if there exists a *closed difference* form ω, of degree $k := |I|$,

$$\omega := \sum_{J c \{1, ..., n\}, |J| = k} H_J \, \delta m_J \tag{5.20}$$

such that the components H_J are all holonomic, and $H_I = f$.

We remark that every closed holonomic form is exact, since it is always possible to take "anti-differences" in the class of holonomic functions, and the proof is analogous to the proof that the de Rham cohomology in flat space is trivial. It follows that the totality of identities of the form (5.19) is obtained by taking arbitrary holonomic forms of degree $k - 1$ and applying the exterior difference operator δ. This is not very exciting, since most identities produced that way are very boring. We are interested in "nice" identities, in which $f(m_1, ..., m_n)$ is a Closed Form function rather than just a plain holonomic function.

6. Closed Form Identities and WZ forms

Let's return to our main object of interest: Closed Form identities. Since any Closed Form function $f(m_1, ..., m_n)$ is holonomic, theorem 6 tells us that whenever an identity of the from (5.19) holds, it is so because $f \delta m_I$ is one term of a closed holonomic form. However, more is true in that case.

The $G's$ of (5.9) have special form now! It is easy to see that if f is CF then applying to it any linear difference operator with polynomial coefficients yields a multiple of f by a rational function. Indeed, by iterating (4.1),

$$E_1^{i_1}...E_n^{i_n}f/f = f(m_1+i_1,...,m_n+i_n)/f \tag{6.1}$$

is a rational function, and since for any linear difference operator $P(E_1,...,E_n,m_1,...,m_n)$, Pf/f is a linear combination, with polynomial coefficients, of such expressions, Pf/f itself is a rational function. In general, there is no guarantee that the operator $A^{(i)}$ of (5.10) is Δ_i, so the terms of the closed form ω do not all have to be CF. In practice, however, it happens, *very often*, that this is the case, and then the form ω has all CF components, and moreover, they are all multiples of f by rational functions. By taking common denominator, we are lead to the following definition of a *WZ form.*

Definition: A WZ form of degree k is a *closed difference form*that looks as follows: for some CF function $f(m_1,...,m_n)$, and polynomials $P_I(m_1,...,m_n)$:

$$\omega = f \cdot [\sum_{Ic\{1,2,...,n\},|I|=k} P_I \delta m_I] \ . \tag{6.2}$$

7. Examples of WZ forms

I have already mentioned that any WZ pair $(F(n,k), G(n,k))$ gives rise to the WZ form $F(n,k)\delta k + G(n,k)\delta n$. So we can take all the examples of [WZ1].

The way to obtain non-trivial WZ pairs is to start with a well known identity, or a specialization thereof:

$$\sum_k F(n,k) = 1 \ , \tag{7.1}$$

and apply Gosper's [G] algorithm for *indefinite hypergeometric summation*, w.r.t to k, to $\Delta_n F(n,k)$, thus hopefully getting a CF $G(n,k)$ such that

$$\Delta_n F(n,k) = \Delta_k G(n,k) \quad . \tag{7.2}$$

As was narrated above, one is always guaranteed, thanks to the theory of holonomic systems, to find a CF $G(n,k)$ and an operator $A(N,n)$ such that

$$A(N,n)\Delta_n F(n,k) = \Delta_k G(n,k) \quad , \tag{7.3}$$

which at any rate "certifies", i.e. proves, the identity. Being lucky means that the operator $A(N,n)$ turns out to be the identity operator, and the good news is that we are lucky in the vast majority of cases.

Brilliant Discovery Of Herb Wilf: Even in the "unlucky" cases, there is still hope of getting a WZ pair. If there exists an operator $B(N,n)$ such that $A(N,n)\Delta_n = \Delta_n B(N,n)$, then $(B(N,n)F(n,k), G(n,k))$ is a WZ pair.

Many times an identity has several auxiliary parameters, and hitherto we just picked one of them, called it n and left the other passive. That was arbitrary and unfair, so let's rectify is. Suppose we have a known identity:

$$\sum_k f(n_1,...,n_r,k) = 1 \ . \tag{7.4}$$

For each of the parameters that is not being summed over, $n_1,...,n_r$, do the above, getting, hopefully, CF G_i s.t

$$\Delta_{n_i} F(n_1,...,n_r,k) = \Delta_k G_i(n_1,...,n_r,k), \quad i = 1,...,r \ . \tag{7.5}$$

The above r identities are equivalent to the single statement that

$$\omega := F\delta k + \sum_{i=1}^{r} G_i \, \delta n_i \tag{7.6}$$

is a closed difference form, and hence a WZ form. It follows that we have r "companion" identities

$$\sum_{n_i} G_i(n_1,...,n_r,k) = constant \ , \tag{7.7}$$

whenever the sum converges. In practice they would usually not converge, but one can perform the operation of "shadowing" described in section 4 of [WZ1], by which for any given $i = 1,...,r$, we can always find a shadow $\tilde{\omega}^{(i)}$ of ω, such that the coefficient $\tilde{G}_i$ of δn_i in $\tilde{\omega}^{(i)}$ has the property that for any fixed $n_1,...,n_{i-1}$, $n_{i+1},...,n_r,k$, there are only finitely many n_i for which $\tilde{G}_i$ is non-zero.

Using the above procedure for Dixon's identity (compare [WZ1], p. 153)

$$\sum_k (-1)^k \frac{(a+b)!(a+c)!(b+c)!a!b!c!}{(a+k)!(a-k)!(b+k)!(b-k)!(c+k)!(c-k)!(a+b+c)!} = 1 \tag{7.8}$$

we get the following WZ form

$$\omega_{DIXON} := (-1)^k \frac{(a+b)!(a+c)!(b+c)!a!b!c!}{2(a+k)!(a-k+1)!(b+k)!(b-k+1)!(c+k)!(c-k+1)!(a+b+c+1)!} \cdot \tag{7.9}$$

$$[2(a-k+1)(b-k+1)(c-k+1)(a+b+c+1)\delta k \ + \ (b-k+1)(b+k)(c-k+1)(c+k)\delta a +$$

$$(a-k+1)(a+k)(c-k+1)(c+k)\delta b \;+\; (a-k+1)(a+k)(b-k+1)(b+k)\delta c] \;.$$

Saalschu:tz's identity

$$\sum_k \frac{(m-r+s)!(n+r-s)!(r+k)!m!n!(r-m)!(s-n)!}{k!(m-r+s-k)!(n-k)!(r-s+k)!(m+n)!(r+k-m-n)!r!s!} = 1 \;, \tag{7.10}$$

leads to, and is proved by, the following WZ form

$$\omega_{SAALSCHUTZ} := \frac{(m-r+s)!(n+r-s)!(r+k)!m!n!(r-m)!(s-n)!}{k!(m-r+s-k)!(n-k)!(r-s+k)!(m+n)!(r+k-m-n)!r!s!} \cdot \tag{7.11}$$

$$[\delta k \;+\; \frac{-k(r-s+k)(-r-k+m+n)}{(n-k+1)(m+n+1)(n-s)}\delta n \;+\; \frac{-2r-m+1-s+k)k}{(r+1)(-m+r-s)}\delta r +$$

$$\frac{-k(-r+s-k)((r+k-m-n)}{(m-r+s-k+1)(s+1)(-n-r+s)}\delta s \;+\; \frac{-k(r-s+k)(-r-k+m+n)}{(m-r+s-k+1)(m+n+1)(m-r)}\delta m] \;.$$

Moving right along, Vandermonde's identity, involving two free parameters, after some shadowing, yields the WZ form

$$\omega_{VANDERMONDE} := \tag{7.12}$$

$$\frac{n!^2 k!^2 a!^2}{(a+k+1)!(n+k+1)!(n+a+1)!}[(n+a+1)\delta k \;+\; (k+a+1)\delta n \;+\; (n+k+1)\delta a] \;.$$

So far all our examples were 1-forms, having been obtained from single sum identities. To find higher degree forms, we must start with a well known multi-sum CF identity, and we would need a multivariate analog of Gosper's algorithm. I am presently working on developing such an algorithm, but until I succeed, all I can present is the r-form arising out of the multinomial identity

$$\sum \frac{n!}{k_1!...k_r!(n-k_1-...-k_r)!(r+1)^n}\delta k_1 ... \delta k_{r=1} \;, \tag{7.13}$$

which produces

$$\omega_{MULTINOMIAL} := \frac{n!}{k_1!...k_r!(n-k_1-...-k_r+1)!(r+1)^n} \cdot \tag{7.14}$$

$$[(n - k_1 - ... - k_r \, + \, 1)\delta k_1 ... \delta k_r \, + \, \sum_{i=1}^{r} k_i \, \delta n \delta k_1 ... \delta \hat{k_i} ... \delta k_r] \ .$$

8. WZ COHOMOLOGY

Since every holonomic closed form is exact, it follows that every WZ $r - form \ \omega$ can be written as $\delta theta$, for some *holonomic* $(r - 1) - form \ theta$. However, this is not an effective way of cranking out WZ forms, since there is no way of knowing beforehand which holonomic $(r - 1)$-forms *theta* are such that $\delta theta$ are WZ forms, i.e. all their components are CF. Of course, if *theta* is CF to begin with, then $\delta theta$ is a WZ-form: it is closed, and all its components are CF, as can be easily verified. We will call such WZ forms *exact.* I will now explain why exact WZ forms lead to trivial identities, but before we must define *trivial.*

Let's look at the following two definite integrals

$$\int_0^\infty xexp(-x^2)dx \, = \, 1/2, \quad \int_0^\infty exp(-x^2)dx \, = \, \sqrt{\pi}/2 \ . \tag{8.1}$$

The first of these is trivial, since the indefinite integral $\int^y xexp(-x^2)dx$ can be expressed in finite terms (in the sense of Liouville, see [DST]), and equals $(-1/2)exp(-y^2) + C$, from which the definite integral can be obtained by plugging in $y = 0$ and $y \, = \, \infty$. On the other hand, the indefinite integral $\int_0^y exp(-x^2)dx$ cannot be expressed in finite terms (so a special name, *"erf"*, had to be coined for it), and hence the definite integral has a nice formula for a deeper reason. Since according to Lord Kelvin, the second definite integral above was to Liouville what $2 + 2 = 4$ is to an ordinary mortal, the first integral must have been as clear to him as $0 + 0 = 0$ is to the man in the street.

Going back to sums, a definite sum

$$\sum_{k=-\infty}^{\infty} F(n,k) \, = \, 1 \ , \tag{8.2}$$

with $F(n,k)$ of CF, is *trivial* if the indefinite sum

$$\Phi(n,m) := \sum_{k=-\infty}^{m} F(n,k) \ , \tag{8.3}$$

is CF in m and n, so that the definite sum follows from it by taking $\Phi(n, \infty)$ and checking that it is indeed 1. As we saw above, the way we proved an identity like (8.2), was to apply Gosper's algorithm to $\Delta_n F(n,k)$, but if the identity (8.2) follows from the indefinite version (8.3), then Gosper's algorithm is already successful on $F(n,k)$ itself. In other words, there exists a CF function $PHI(n,k)$ such that

$$F(n,k) \, = \, \Delta_k \Phi(n,k) \ . \tag{8.4}$$

Let

$$G(n,k) \;=\; \Delta_n \Phi(n,k) \;\; . \tag{8.5}$$

It follows that $G(n,k)$ is the WZ-mate of $F(n,k)$ and that

$$\delta\Phi(n,k) \;=\; F(n,k)\delta k \;+\; G(n,k)\delta n \;\; , \tag{8.6}$$

and the WZ form $F(n,k)\delta k \;+\; G(n,k)\delta n$ is thus an exact form.

Similar considerations hold for identities with r free parameters, leading to $1-forms$ with $r+1$ variables, and to multi-sum identities. The interesting identities are precisely those arising out of WZ forms that are not exact, so loosely speaking, finding all interesting identities amounts to computing the *WZ cohomology.* we must be careful to take the word "cohomology" with a grain of salt, since the set of WZ forms is not a vector space. One way out is to consider the vector space of linear combinations, but a better way is as follows.

Research Problem: Characterize those CF functions f such that there exist (closed), non-exact, WZ forms $f\omega_P$, with ω_P a polynomial difference form, and for those successful f, compute the cohomology of closed modulo exact.

9. Continuous WZ forms

The theory of holonomic systems makes sense, and in fact was initiated([Ber]), in R^n. The notion of CF is defined naturally as follows. (See [AZ], where it is called "hyperexponential".)

Definition: A function $f(x_1,...,x_n)$ is CF if all its logarithmic partial derivatives $\frac{\partial f}{\partial x_i}/f$ are rational functions of $x_1,...,x_n$.

In analogy with the discrete case, we define

Definition: A continuous WZ form on R^n is a closed differential form that is the product a CF function by a polynomial differential form.

It is very easy to manufacture new continuous WZ forms out of old ones. If ω is any WZ form, so is $\omega(g)$, for any rational transformation g, and if ω and *theta* are WZ forms, so is their wedge product ω*theta*. But let's not get too exited: I don't know of any continuous closed WZ form that is not exact! and in fact I am almost sure that the following conjecture is true:

Conjecture: The continuous WZ cohomology is trivial, i.e. every closed continuous WZ form is exact.

Although I do know of a few CF identities

$$\int_{-\infty}^{\infty} F(x,y)dy \;=\; 1 \;\; , \tag{9.1}$$

with $F(x,y)$ CF (see [AZ]), none of them come from WZ forms. What we do have in this case, in analogy to the discrete case, is a CF $G(x,y)$, and an operator $A(D_x,x)$ such that

$$A(D_x,x)D_xF(x,y) \;=\; D_yG(x,y) \;\; , \tag{9.2}$$

but $A(D_x, x)$ has never, in my experience, turned out to be the identity operator!

It seems that the reason that the discrete case leads to so many more interesting things is that finite differences are not derivations, and you can't compose with a transformation, so WZ forms are hard to come by, and hence are more interesting. However we should not write off the continuous realm altogether, all we have to do is interface it with the discrete, which brings us to the next section.

10. Holonomic and WZ forms in $R^r \times Z^s$

The full beauty and power of the theory of holonomic functions, as developed in [Z1], is on functions of several discrete and continuous variables. Everything we said before extends naturally. Since "continuous came first" we will denote the exterior derivative-difference by d. It is defined on scalar functions by:

$$df(x_1, ..., x_r, m_1, ..., m_s) := \sum_{i=1}^{r} \frac{\partial f}{\partial x_i} dx_i + \sum_{j=1}^{s} (\Delta_{m_j} f) \delta m_j \ , \tag{10.1}$$

and on general forms as before, where all the "letters" $dx_1, ..., dx_n, \delta m_1, ..., \delta m_n$ anti-commute. Once again we can raise the question of WZ cohomology, and trying to find "all of them". The Macdonald conjectures (see [H] for a nice review and references), offer many examples of possible WZ forms, and for A_2 and G_2 they were found by Shalosh B. Ekhad [E].

11. Examples of Discrete-Continuous WZ forms

The customary proof of Euler's integral

$$\int_0^\infty e^{-x} x^k \ = \ k! \tag{11.1}$$

is by integration by parts. This proof can be recast by saying that the form

$$\omega_{"\Gamma"} := e^{-x} x^k over (k+1)! [(k+1) dx \ + \ x \delta k\,] \tag{11.2}$$

is closed, as is easily verified. The "companion identity" is obtained by summing $\omega_{"GAMMA"}$ w.r.t to k, and, surprise!, it turns out to be good old

$$e^x \ = \sum_{k=0}^{\infty} \frac{x^k}{k!} \ \ .$$

So Euler's integral for the Gamma function could have been discovered by starting with the series expansion for e^x and finding its companion identity.

Similarly, Euler's beta integral could have been discovered by taking the companion identity of the binomial theorem for $(1-y)^{-m}$. Euler's Beta integral gives rise to the following WZ form, whose (easily verified) closedness proves it.

$$\omega_{BETA} := \frac{y^n(1-y)^m(n+m+1)!}{(m+1)!(n+1)!}\,[\,((n+1)(m+1))dy-(y(1-y)(m+1))\delta n-(y(1-y)(n+1))\delta m]\ . \tag{11.4}$$

It is possible to get many more non-trivial examples as follows. Take a known evaluable ""${}_2F_1(x)$, (many of which were found by Gosper and Gessel & Stanton, ([PBM], pp. 491-497,[GS]), convert them to an integral formula, via the well known integral formula (e.g. [Ba] (1.5),(1), p. 4)

$${}_2F_1(a,b;c;z) \;=\; \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}dt\ , \tag{11.5}$$

and find the corresponding WZ proof and form. A MAPLE program, based on the algorithm of [AZ], that does just that, is available upon requests, as is an extensive list of successful outputs. For example, the identity ([GS],(5.23))

$${}_2F_1(-2n-1/3\,,\,-n\,;\,2/3\,;\,-8) \;=\; (-27)^n\ , \tag{11.6}$$

yields the following continuous-discrete WZ form:

$$\frac{y^{-n-1}(1-y)^{n-1/3}(1+8y)^{2n+1/3}n!}{27^n\,(5/3)_n}\cdot[(27n+18)dy\;+\;(4y-1)(1-y)(1+8y)\delta n]\ . \tag{11.7}$$

12. A WZ approach to Hypergeometric Convergence Acceleration Formulas

Some WZ forms can be useful for convergence acceleration. If ω is a WZ $r-form$, then its integral is zero over any r-manifold that is a boundary of some $(r+1)-$manifold. By partitioning such a manifold into two subsets, we get that two different summations are equal. If one of them converges faster then the other we have a convergence acceleration formula.

Let's specialize to two variables (n,k). Let $\omega := F(n,k)\delta k\;+\;G(n,k)\delta n$ be a WZ form, and let's sum it over the discrete contour

$$\partial\Omega := \{((n,0)\leftarrow(n+1,0);n>=0\}\,\cup\,\{(n,n)\to(n+1,n)\to(n+1,n+1)\,;\,n>=0\}\cup$$
$$\{(\infty,k+1)\to(\infty,k);k>=0\},$$

which is the "boundary" of the region $\Omega=(n,k);n>=k$. Since

$$0=\sum_{\partial\Omega}\omega=\sum_{n=0}^{\infty}G(n,0)-\sum_{n=1}^{\infty}(F(n,n-1)+G(n-1,n-1))\ , \tag{12.2}$$

we have

Theorem 7: For any WZ pair (F,G)

$$\sum_{n=0}^{\infty} G(n,0) \;=\; \sum_{n=1}^{\infty}(F(n,n-1)+G(n-1,n-1)), \tag{12.3}$$

whenever both sums converge.

We will give three applications of theorem 7. The WZ form

$$\omega_{log(1+x)} := \frac{(-1)^k n!k!}{(n+k+1)!x^k(1+x)^{n+1}}[\,(1+x)\delta k \;+\; x\delta n] \tag{12.4}$$

leads to the acceleration formula

$$\sum_{n=1}^{\infty} \frac{1}{n(1+x)^n}(=\; log((1+x)/x)) \;=\; (1+2x)\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n\binom{2n}{n}((1+x)x)^n} \quad . \tag{12.5}$$

The WZ form

$$\omega_{\zeta(2)} \;:=\; \frac{(-1)^{(n+k)}k!^2(n-k-1)!}{(n+k+1)!}[\delta k \;+\; (2(n-k)/(n+1))\delta n] \tag{12.6}$$

leads to the well known formula for $\zeta(2)($ [P]):

$$2\sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)^2} \;=\; (\,\zeta(2)\,,\,(why?)\,) \;=\; 3\sum_{n=1}^{\infty} \frac{1}{(\binom{2n}{n})n^2} \quad , \tag{12.7}$$

while the WZ form

$$\omega_{\zeta(3)} := \frac{(-1)^k k!^2(n-k-1)!}{(n+k+1)!}[\,(1/(2(k+1)))\delta k \;+\; ((n-k)/(n+1)^2)\delta n\,] \tag{12.8}$$

leads to the acceleration formula ([P])

$$\sum_{n=1}^{\infty} \frac{1}{n^3} = (5/2)\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{\binom{2n}{n}n^3} \quad , \tag{12.9}$$

which was the starting point of Apéry's wonderful proof, which we will discuss and redo later.

The notion of WZ pairs and forms thus gives a unified setting for proving and discovering convergence acceleration formulas. To discover such new formulas we take any known identity, or *a specialization thereof*, find the corresponding WZ pair, take appropriate shadow ([WZ1], sect. 4), and cross our fingers that one of the sums in (12.3) converge slowly to a well known number, while the other sum converges fast.

The same reasoning can be applied to $1 - forms$ in more variables and also to higher degree forms. We will only state it for $1 - forms$ of 3 variables, since the general statement, though straightforward, is rather cumbersome. we have

Theorem 8: For any WZ 1-form of three variables,

$$\omega := F(n,k,a)\delta k \; + \; G(n,k,a)\delta n \; + \; H(n,k,a)\delta a \;\; , we\, have \; , \tag{12.10}$$

$$\sum_{a=0}^{\infty} H(0,0,a) \; = \; \sum_{n=1}^{\infty}(H(n,n,n-1) + F(n,n-1,n-1) + G(n-1,n-1,n-1)), \tag{12.11}$$

provided both sides converge.

Applying theorem 8 to $\omega_{VANDERMONDE}$ of (7.12) we get an even better acceleration for $\zeta(2)$:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \sum_{n=1}^{\infty} \frac{n!^6(21n+13)}{8(2n+1)!^3} \quad .$$

13. A Conceptual Frame To Apéry's Magic

Apéry ([A1][A2][P][R][M]) used the fact that if α is a rational number, then for any sequence $a_n/b_n (\neq \alpha)$ of rational numbers tending to α, $|\alpha - a_n/b_n| >= C/|b_n|$ (if $\alpha = c/d$ then $|\alpha - a_n/b_n| > (1/d)/b_n$). Thus if one can exhibit a sequence $a_n/b_n (\neq \alpha)$, of rational numbers, with a_n, b_n integers, such that $|\alpha - a_n/b_n| < C/|b_n|^{1+\delta}$, $\delta > 0$, then one has proven the irationality of α. For α defined by a hypergeometric series

$$\alpha = \sum_{n=0}^{\infty} f(n) \;\; , (\, f(n+1)/f(n) \, a\, rational\, function\, in\, n) \; , \tag{13.1}$$

the natural temptation is to take the approximating sequence to be the sequence of partial sums

$$\sum_{k=0}^{n} f(k) \; , \tag{13.2}$$

and this works for e, for example. However, for $\zeta(n)$ the convergence is far too slow, although the denominators are not too bad. To rectify the slow convergence, one may try to find a convergence acceleration formula, like (12.7) for $\zeta(2)$ and (12.9) for $\zeta(3)$. Alas now the denominators of the partial sums grow too fast, and the improved numerical convergence does not make up for it.

Apéry's breakthrough ([P][R][M]) consisted in finding a doubly-indexed sequence $c(n,k)$ (defined for $n >= k >= 0$) that tends to the number of interest α (in his case $\zeta(3)$ and $\zeta(2)$), no matter how you go to (∞,∞) in the region $n >= k >= 0$. The sequence $c(n,0)$ turned out to be the sequence of partial sums of the defining series (whose denominators were decent, but whose convergence rate was too slow), while the diagonal sequence $c(n,n)$ turned out to be the sequence of partial sums of the accelerated series (12.7) and (12.9), whose convergence rate was good, but whose denominators grew too large. Somehow it was necessary to have a tradeoff. This was achieved by introducing a new approximating sequence, by taking a weighted average of the $c(n,k)$, $k=0,...,n$, with explicitly defined weights $b(n,k)$:

$$d_n \; = \; \frac{a(n)}{b(n)} := \frac{\sum_{k=0}^{n} c(n,k) b(n,k)}{\sum_{k=0}^{n} b(n,k)} \quad . \qquad (13.3)$$

(Please be warned that the $c(n,k)$ are rational numbers, so the numerator $a(n)$ in the above expression is not an integer.) It turned out that for a judicious choice of $c(n,k)$ and weights $b(n,k)$, things worked out for Apéry. To wit, for $\zeta(2)$ he gave

$$c(n,k) := 2 \sum_{m=1}^{n} \frac{(-1)^{m-1}}{m^2} + \sum_{m=1}^{k} \frac{(-1)^{n+m-1}}{m^2 \binom{n}{m}\binom{n+m}{m}} \; , \quad b(n,k) = \binom{n}{k}^2 \binom{n+k}{k} \; , \qquad (13.4)$$

while for $\zeta(3)$ Apéry gave

$$c(n,k) := \sum_{m=1}^{n} \frac{1}{m^3} + \sum_{m=1}^{k} \frac{(-1)^{m-1}}{2m^3 \binom{n}{m}\binom{n+m}{m}} \; , \quad b(n,k) = \binom{n}{k}^2 \binom{n+k}{k}^2 \; . \qquad (13.5)$$

The way Apéry proved that d_n had the desired properties was to show that both the top $(a(n))$ and bottom $(b(n))$ of (13.3) were solutions of a certain three term linear recurrence with polynomial coefficients. In the case of $\zeta(2)$ it turned out that both $x_n = a(n)$ and $x_n = b(n)$ were solutions of

$$n^2 x_n - (11n^2 - 11n + 3) x_{n-1} - (n-1)^2 x_{n-2} = 0 \quad , \qquad (13.6)$$

while in the case of $\zeta(3)$ the recurrence was

$$n^3 x_n - (34n^3 - 51n^2 + 27n - 5) x_{n-1} + (n-1)^3 x_{n-2} = 0 \quad . \qquad (13.7)$$

In Alf van der Poorten's delightful account of Apéry's proof([P]) (with details that were filled in by H. Cohen), everything is presented as magic, in tune with Apéry's own personality and style ([A1],[M]). Subsequently, Beukers[Beu] presented a new proof that took away the magic but retained the charm of the original proof. Beukers's elegant proof was extended and generalized by himself, Allady and Robinson[AR], and others, while Apéry's original proof seemed ad-hoc and ungeneralizable.

Ken Wilson once said that today's tricks are tomorrow's theory. I will now demystify Apéry's original proof by placing it in the context of holonomic functions and WZ theory. Hopefully this will open the door for further applications and generalizations. I will also show how all the "hairy" steps in Apéry's proof are now purely routine, and provable by computer, using the fast algorithm [Z3] that is based on Gosper's [G] algorithm. See the appendix for an example.

Let's make two key observations about Apéry's proofs:

Observation 1: The doubly-infinite sequence $c(n,k)$ is the "potential function" of a WZ form! Namely it is the "contour sum" of a WZ form $\omega = F\delta k + G\delta n$ from $(0,0)$ to (n,k), and since ω is closed, the contour is immaterial. In other words, $\omega = \delta c$. Since the holonomic cohomology is trivial, we know that every closed 1-form ω, qua holonomic form, can be expressed as δc, for some holonomic 0-form (i.e. function) $c(n,k)$. Note that $c(n,k)$ is not CF, or else everything would be trivial. The $c(n,k)$ for $\zeta(2)$, (13.4), is the potential function of the WZ form $\omega_{\zeta(2)}$ given in (12.6), and the $c(n,k)$ for $\zeta(3)$, (13.5), is the potential function for the WZ form $\omega_{\zeta(3)}$ given in (12.8). (The way $c(n,k)$ is presented in (13.4) and (13.5) is the "contour sum" of $\omega_{\zeta(2)}$ and $\omega_{\zeta(3)}$ over the contour $(0,0) \to (1,0) \to ... \to (n,0) \to (n,1) \to ... \to (n,k)$.)

Observation 2: The "weighting" function $b(n,k)$ is CF.

Let's recall the "hairy steps" in Apéry's proof ([P],[R]). First a recurrence operator $P(N,n)$ is pulled out of the hat, and it is claimed that both $a(n)$ and $b(n)$ are annihilated by it. The way these claims are proved is to manufacture a $B(n,k)$, again out of the blue, such that

$$P(N,n)b(n,k) = B(n,k) \ - \ B(n,k-1) \ ,$$

and a $D(n,k)$ such that

$$P(N,n)(b(n,k)c(n,k)) = B(n,k)c(n,k) \ - \ B(n,k-1)c(n,k-1) + D(n,k) - D(n,k-1) \ .$$

Furthermore, the $B(n,k)$ and $D(n,k)$ turned out to be CF.

I will now show that for any $c(n,k)$ that is a potential function of a WZ 1-form ω, and any CF function $b(n,k)$ there exists such $P(N,n)$, $B(n,k)$, and $D(n,k)$. In other words:

1)We have the "meta-theorem" that for *every* $c(n,k)$ such that δc is WZ, and for *every* CF $b(n,k)$, we have an Apéry-style proof.

2) The form of the proof is always the same.

I will also show how my computer (or your computer, if you have MAPLE and ask for my program), can generate the proof, in every given case, out of the input $((F(n,k), G(n,k)), b(n,k))$, (where $\delta c = F\delta k + G\delta n$, and the latter is WZ.) The proof consists in presenting the recurrence operator $P(N,n)$ and the two "certificates" $B(n,k)\, , D(n,k)$.

Unfortunately, the recurrence $P(N,n)$ is usually of higher order, and usually does not yield irrationality, and so far I am unable to prove irrationality of new interesting numbers.

After this long introduction, let's go to business.

Theorem 9: Let $c(n,k)$ be the potential function of a WZ 1-form $F(n,k)\delta k + G(n,k)\delta n$ in the two variables (n,k). In other words,

$$F(n,k) = c(n,k+1) - c(n,k) \, , \ \ G(n,k) = c(n+1,k) - c(n,k) \ ,$$

and let $b(n,k)$ be CF. Let $a(n)$ and $b(n)$ be the top and bottom of (13.3):

$$a(n) := \sum_{k=0}^{n} c(n,k)b(n,k) \, , \quad b(n) := \sum_{k=0}^{n} b(n,k) \ .$$

There exist (rapidly exhibitable) linear recurrence operators with polynomial coefficients $R(N,n)$ and $S(N,n)$ such that

$$R(N,n)b(n) \ = \ 0 , \quad S(N,n)R(N,n)a(n) \ = \ 0 \ . \tag{13.9}$$

Furthermore, there exist rapidly exhibitable CF "certificates" $B(n,k)$ and $D(n,k)$ such that the following routinely verifiable identities are true:

$$R(N,n)b(n,k) = B(n,k) - B(n,k-1) \, , \tag{13.10}$$

$$S(N,n)R(N,n)(b(n,k)c(n,k)) = S(N,n)(c(n,k)B(n,k) - c(n,k-1)B(n,k-1)) + D(n,k) - D(n,k-1) \ .$$

In addition, $B(n,k)/b(n,k)$ and $D(n,k)/(b(n,k)F(n,k))$ are both rational functions.

Before proving Theorem 9, let's make a few remarks. Since $R(N,n)b(n) = 0$ obviously implies $S(N,n)R(N,n)b(n) = 0$, it follows that both $a(n)$ and $b(n)$ are annihilated by the same operator $P(N,n) = S(N,n)R(N,n)$, but the latter is in general *not* the minimal order operator that annihilates $b(n)$. $P(N,n)$ is minimal for both $a(n)$ and $b(n)$ whenever $S(N,n)$ is the identity operator. This was what happened in Apéry's cases, and explained why he was successful, since both $a(n)$ and $b(n)$ satisfied *the same minimal second order linear recurrenceS* is the identity operator an *Apéry pair.*

Note also that (13.10) immediately imply (13.9), by summing w.r.t k, and that (13.10) is routinely verifiable, as we will see.

Proof of Theorem 9: The first parts of (13.9) and (13.10) were proved in [Z3], where an algorithm for constructing $R(N,n)$ and $B(n,k)$ was given, and it was shown that $B(n,k)/b(n,k)$ is a rational function of (n,k). Now let's write

$$R(N,n) = \sum_{i=0}^{L} r_i(n)N^i \tag{13.11}$$

and consider $R(N,n) \, (\, b(n,k)c(n,k) \,)$. We have

$$R(N,n) \, (\, b(n,k)c(n,k) \,) \ - c(n,k)R(N,n)b(n,k) = \tag{13.12}$$

$$(\sum_{i=0}^{L} r_i(n)N^i)b(n,k)c(n,k) \ - \ c(n,k)(\sum_{i=0}^{L} r_i(n)N^i)b(n,k)$$

$$= \sum_{i=0}^{L} r_i(n) b(n+i,k)[c(n+i,k) - c(n,k)] \;=\; \sum_{i=0}^{L} r_i(n) b(n+i,k)\,[\sum_{j=0}^{i-1} G(n+j,k)\,] \;\; .$$

It follows from this and the first part of (13.10) that

$$R(N,n)(\,b(n,k)c(n,k)\,) \;= \tag{13.13}$$

$$c(n,k)(B(n,k) - B(n,k-1)) \;+\; \sum_{i=0}^{L} r_i(n) b(n+i,k)\,[\,\sum_{j=0}^{i-1} G(n+j,k)\;]\;\; .$$

But

$$c(n,k)(B(n,k) - B(n,k-1)) = \tag{13.14}$$

$$c(n,k)B(n,k) - c(n,k-1)B(n,k-1) - (c(n,k) - c(n,k-1))B(n,k-1) =$$

$$c(n,k)B(n,k) - c(n,k-1)B(n,k-1) - F(n,k-1)B(n,k-1)\;\; .$$

It follows that

$$R(N,n)\,(\,b(n,k)c(n,k)\,) = c(n,k)B(n,k) - c(n,k-1)B(n,k-1) + H(n,k)\;\;, \tag{13.15}$$

where $H(n,k)$ is given by the expression:

$$H(n,k) = \sum_{i=0}^{L} r_i(n)\, b(n+i,k)\,[\,\sum_{j=0}^{i-1} G(n+j,k)\;]\; - F(n,k-1)B(n,k-1)\;\; . \tag{13.16}$$

The pleasant surprise it that $H(n,k)$ is CF on its own right! Indeed we can write

$$H(n,k) \;=\; \{\sum_{i=0}^{L} r_i(n)(b(n+i,k)/b(n,k))[\sum_{j=0}^{i-1}(\,\frac{G(n+j,k)}{F(n+j,k)}\,)\,(\,\frac{F(n+j,k)}{F(n,k)}\,)\,] \tag{13.17}$$

$$-(\frac{F(n,k-1)}{F(n,k)}\,\frac{B(n,k-1)}{b(n,k-1)}\,\frac{b(n,k-1)}{b(n,k)}\}b(n,k)F(n,k)\;\; .$$

Since $b(n,k)$ and $F(n,k)$ are CF, and $B(n,k)/b(n,k)$ and $G(n,k)/F(n,k)$ are rational functions, it follows that the expression inside the braces in (13.17) is a rational function, since it is a sum of products of them. Obviously $b(n,k)F(n,k)$ is CF, and it follows that $H(n,k)$ is CF.

It follows once again from [Z3] that there exists an operator $S(N,n)$ and a CF function $D(n,k)$ that is a multiple of $H(n,k)$ by a rational function, such that

$$S(N,n)H(n,k) = D(n,k) - D(n,k-1) \ . \tag{13.18}$$

Applying $S(N,n)$ to (13.15), and using (13.18), we get

$$S(N,n)R(N,n)(\,b(n,k)c(n,k)\,) \ = \tag{13.19}$$

$$S(N,n)[c(n,k)B(n,k) - c(n,k-1)B(n,k-1)] + D(n,k) - D(n,k-1) \ \ ,$$

which is the second half of (13.10), which we had to prove, and that immediately implies (13.9). QED

One of the rewards of a general theory is that it points the way to fruitful future generalizations. Instead of a two-variable $c(n,k)$, we can consider a multi-variate $c(n,k_1,...,k_r)$ that is a potential function of a WZ 1-form

$$\omega \ = \ F(n,k_1,...,k_r)\delta n + \sum_{i=1}^{r} G_i(n,k_1,...,k_r)\delta k_i \ \ .$$

Now we consider weighted averages

$$d_n = \frac{\sum_{k_1,...,k_r} b(n,k_1,...,k_r)c(n,k_1,...,k_r)}{\sum_{k_1,...,k_r} b(n,k_1,...,k_r)} \ \ . \tag{13.20}$$

It is possible to show that (13.9) still holds, while (13.10) is replaced by the existence of $B_i(n,k_1,...,k_r)$, $D_i(n,k_1,...,k_r)$, $i=1,...,r$, all CF, such that

$$R(N,n)b(n,k_1,...,k_r) \ = \ \sum_{i=1}^{r} \Delta_{k_i} B_i(n,k_1,...,k_r) \ \ , \tag{13.21}$$

$$S(N,n)R(N,n)(\,b\,c\,) = S(N,n)(\sum_{i=1}^{r} \Delta_{k_i}(cB_i)) \ + \ \sum_{i=1}^{r} \Delta_{k_i} D_i \ \ . \tag{13.22}$$

Furthermore, the B_i's are all multiples of $b(n,k)$ by rational function, and the D_i are all multiples of $b(n,k)F(n,k)$ by rational functions. The only thing different is that at present there is no fast algorithm for finding the recurrence operators R, S and the certificates B_i, D_i, and the slow algorithm is too slow. I hope that soon I or someone else will find such a fast algorithm.

Concluding comments

Another approach to the insight behind the Apéry recurrence for $b(n)$ (but not for $a(n)$) was given by Askey and Wilson[AW]. Askey and Wilson derive recurrences for much more general sequences. It would be interesting to find the analog of $c(n,k)$ and $a(n)$ for these more general sequences, since perhaps they can be used to prove the irrationality of new numbers.

The referee has pointed out that so far we didn't show how the WZ-form is affected by specialization, save for doing it empirically. Herb Wilf and I, succeeded in doing just this, and this will appear in a forthcoming paper. The referee has also pointed out that our definition of "closed form" excludes sequences such as $f(n) = 0, n\, odd, f(n) = (n/2)!, n\, even$. It would be interesting to generalize the present theory to this more general class.

APPENDIX: AN EXAMPLE OF AN Apéry-STYLE COMPUTER-GENERATED PROOF: The Irrationality of log(2)

My program is available upon request. Here is the input file for the irrationality of log(2). (TOPb, BOTb, bPol, bARGn, bARGk) describe b: b is bPol times $(bARGn)^n(bARGk)^k$ times the product of the factorials in the list TOPb divided by the product of the factorials in the list BOTb. Ditto, regarding F, for (TOPF, BOTF, FPol, FARGn, FARGk), GoF is the rational function that is G/F, $SEDER$ is the conjectured order of $R(N,n)$, while SEDER1 is the conjectured order of $S(N,n)$.

```
#Begin Input File read ‘full_path_name_of_program_file‘ : TOPb:=[n+k]: BOTb:=[k,k,n-k]: bPol:=1:

bARGn:=1: bARGk:=1: TOPF:=[n,k]: BOTF:=[n+k+1]: FPol:=1: FARGn:=1/2:

FARGk:=-1: GoF:=1/2: SEDER:=2; SEDER1:=0: NAME:=log(2):

apery(TOPb, BOTb, bPol, bARGn, bARGk, TOPF, BOTF, FPol, FARGn, FARGk, GoF, NOTATION, SEDER, SEDER1, NAME): quit; #End of Input File
```

Placing the input file in a file called, say, inlog2, I instructed my computer, Shalosh B. Ekhad (that runs under UNIX):

```
maple < inlog2
```

After a few seconds, came the output:

AN APERY PAIR THAT IMPLIES THE IRRATIONALITY OF ln(2)

by SHALOSH B. EKHAD, 85 Wilson Road, Princeton, NJ 08540

Theorem : Let W(n,k):=

$$\frac{(-1)^k(1/2)^n n!k!}{(n+k+1)!}$$

let $Z(n,k) := (1/2)W(n,k)$ and let $b(n,k) :=$

$$\frac{(n+k)!}{k!^2(n-k)!}$$

then ((W(n,k),Z(n,k)),b(n,k)) is an Apéry pair and both b(n):=Sum_k b(n,k) and a(n):=Sum_k b(n,k)c(n,k) are solutions of the linear recurrence equation, (c(n,k) is the potential function of (W,Z))

$$(nN^{-1} - 6n - 3 + (n+1)N)u(n) = 0 \ . \tag{1}$$

PROOF: We cleverly construct $B(n,k) :=$

$$\frac{(-4n-2)(n+k)!}{k!^2(n-k)!}$$

with the motive that

$$(nN^{-1} - 6n - 3 + (n+1)N)b(n,k) = B(n,k) - B(n,k-1), \tag{2}$$

(check!), which upon summing w.r.t to k shows that b(n) is a solution of (1). To establish that a(n) also satisfies (1), we cleverly construct

$$A(n,k) := B(n,k)c(n,k) + \frac{(-1)^k (1/2)^n n!}{k!(n-k+1)!} \tag{3}$$

with the motive that

$$(nN^{-1} - 6n - 3 + (n+1)N)b(n,k)c(n,k) = A(n,k) - A(n,k-1)$$

(To get (3) use (2) and the expression for differences of c(n,k) in terms of W and Z) and the result follows upon summing w.r.t to k.

My program also reproduced within seconds the irrationality proofs of $\zeta(2)$ and $\zeta(3)$.

REFERENCES

[A1] Apéry, R., *Irrationalite' de $\zeta(2)$ et $\zeta(3)$*, Asterisque **61**(1979), 11-13.

[A2] Apéry, R., *Interpolation de fractions continues et irrationalite' de certaines constantes*, Bulletin de la section de sciences du C.T.H.S. #3, 37-53.

[AR] Allady, K., and Robinson, M. L., *Legendre polynomials and irrationality*, Crelle's Journal **318**(1980), 137-155.

[AW] Askey, R.,and Wilson, J., *A recurrence relation generalizing those of Apéry*, J. Aust. Math. Soc. **36**(1984), 267-278.

[AZ] Almkvist, G., And Zeilberger, D., *The method of differentiating under the integral sign*, J. Symbolic Computation, to appear.

[Ba] Bailey, W. N. , "*Generalized Hypergeometric Series*", Cambridge Math. Tracts **32**, Cambridge University Press, London, 1935. (Reprinted: Hafner, New York, 1964 .)

[Ber] Bernstein, I. N. , *Modules over a ring of differential operators, study of the fundamental solutions of equations with constant coefficients*, Functional Analysis and Its Applications, Vol **5**, No. 2(1971), Russian original :1-16, English translation:89-101.

[Beu] Beukers, F., *A note on the irrationality of* $\zeta(2)$ *and* $\zeta(3)$*, Bull. London Math. Soc.* **11** *(1979), 268-272.*

[Bj] Bjo:rk, J. -E., "Rings Of Differential Operators", North-Holland, Amsterdam, 1979.

[DST] Davenport, J. H., Siret, Y., Tournier, E., "Computer Algebra", Academic Press, London, 1988.

[E] Ekahd, S. B., A one-line proof of the Habsieger-Zeilberger G_2 *constant term identity*, J. Computational and Applied Mathematics, to appear.

[F] Flanders, H., "*Differential Forms: With Applications to the Physical Sciences*", Academic Press, New York, 1963.

[G] Gosper, R. W., Jr., *Decision Procedure for indefinite hypergeometric summation*, Proc. Natl. Acad. Sci. USA**75**(1978), 40-42.

[GKP] Graham, R., Knuth, D.E., and Patashnik, O., "*Concrete Mathematics*", Addison Wesley, Reading, 1989.

[H] Habsieger, L., *Macdonald conjectures and the Selberg integral*, in: "q-Series and Partitions", edited by D. Stanton, IMA volumes **18**, Springer, New York, 1989.

[M] Mende‘s France, M., *Roger Apéry et l'irrationel*, La Recherche, **10**(1979), 170-172.

[P] van der Poorten, A., *A Proof that Euler missed..., Apéry's proof of the irrationality of*$\zeta(3)$, Math. Intel. **1**(1979), 195-203.

[PBM] Prudnikov, A.P., Brychkov, Yu.A., and Marichev, O.I., *"Integrals and Series"*, vol. 3., Translated from the Russian by G. G. Gould, Gordon and Breach, New York, 1990.

[R] Reyssat, E., *Irrationalite' de* $\zeta(3)$ *selon Apéry*, Seminaire DELANGE-PISOT-POITU (Theorie des nombres) **20**(1978/1979), #6, 6 pp.

[S] Spivak, M., *"Calculus On Manifolds"*, Addison-Wesley, Reading, Mass, 1965.

[WZ1] Wilf, H.S., and Zeilberger, D., *Rational functions certify combinatorial identities*, J. Amer. Math. Soc. **3**(1990), 147-158. ied Math., to appear.

[Z2] Zeilberger, D., *A Fast Algorithm for proving terminating hypergeometric identities*, Discrete Math **80**(1990), 207-211.

[Z3] Zeilberger, D., *The method of creative telescoping*, J. Symbolic Computation, to appear.