J Autom Reasoning (2015) 55:223–243
DOI 10.1007/s10817-015-9338-0

Formal Proofs of Hypergeometric Sums

Dedicated to the memory of Andrzej Trybulec

John Harrison[1]

Received: 23 August 2014 / Accepted: 10 July 2015 / Published online: 15 September 2015

Abstract Algorithmic methods can successfully automate the proof, and even the discovery, of a large class of identities involving sums of hypergeometric terms. In particular, the Wilf-Zeilberger (WZ) algorithm is a uniform framework for a substantial class of hypergeometric summation problems. This algorithm can produce a rational function *certificate* that can, on the face of it, be used to verify the result by routine algebraic manipulations, independently of the working of the algorithm that discovered it. It is therefore very natural to consider using this certificate to produce, by automated means, a rigorous deductive proof in an interactive theorem prover. However, naive presentations of the WZ method tend to gloss over trivial-looking but rather knotty questions about zero denominators, which makes their rigorous formalization tricky and their ultimate logical justification somewhat obscure. We describe how we have handled these difficulties to produce rigorous WZ proofs inside the HOL Light theorem prover.

Keywords Formal · Proof · Hypergeometric · Summation

1 Background

Symbolic algorithms have proven to be remarkably effective for verifying, or even finding, the solutions of summation problems involving factorials, binomial coefficients and rational functions, e.g. $\sum_{k=0}^{n} \binom{n}{k} = 2^n$. Simple examples like this one, which is easy enough anyway by considering the binomial expansion of $(1 + 1)^n$, are just the tip of the iceberg of a wide class of problems that can be tackled automatically or with very limited human intervention, including many where any resolution at all (by a human mathematician or machine) was originally considered challenging.

✉ John Harrison
johnh@ichips.intel.com

[1] Intel Corporation RA2-451, 2501 NW 229th Avenue, Hillsboro, OR 97124, USA

One of the more mathematically significant applications is verifying Apéry's proof of the irrationality of $\zeta(3) = 1/1^3 + 1/2^3 + 1/3^3 + 1/4^3 + \cdots$, which involves showing that the sequence $a_n = \sum_{k=0}^{n} \binom{n+k}{k}^2 \binom{n}{k}^2$ and another more complex variant both satisfy the second-order recurrence

$$(n+2)^3 a_{n+2} - (2n+3)(17n^2 + 51n + 39)a_{n+1} + (n+1)^3 a_n = 0$$

When Apéry originally presented these claims [1], the justification was obscure. Apéry quipped that such identities "grow in my garden", and there was widespread skepticism about his proof [9], until work by Cohen, Lenstra, van der Poorten and Zagier eventually justified the strange identities [20]. While Beukers [3] found a much more transparent proof using double integrals, perhaps the simplest way to verify these recurrences is to use symbolic methods along the lines we discuss here [16].[1]

There are actually several somewhat different algorithms in this general area with a long and interesting history [15]. One very general result (see [18] for a proof) is the closure under summation, arithmetic operations etc. of so-called *holonomic* sequences, those that satisfy a recurrence with polynomial coefficients (as with a_n above, which satisfies a second-order recurrence). The use of such results in explicitly algorithmic form to tackle summation problems was originally proposed by Zeilberger [22]. However, we will focus on a more constrained algorithm [21] commonly called the WZ (for Wilf-Zeilberger) method, which has two advantages. First of all, it is much more efficient in its domain of applicability, and implementations are relatively straightforward and widely available. Second, its workings can produce a simple algebraic 'certificate' that can be checked for correctness routinely. *Or so it seems ...*

2 The WZ Method

In what follows we use the notion of a *hypergeometric* sequence (or term, or series) a_n, which is simply one where the ratio a_{n+1}/a_n of successive terms is a *rational function of* n, i.e. can be represented as $a_{n+1}/a_n = r(n) = p(n)/q(n)$ for some non-zero univariate polynomials p and q. (As the name suggests, this generalizes the notion of a *geometric* sequence where the ratio is constant $a_{n+1}/a_n = c$.) For example, the factorial function is hypergeometric because $(n+1)!/n! = n+1$, as is the 'power of 2' function because $2^{n+1}/2^n = 2$. We call a function of several variables hypergeometric if it is hypergeometric for each argument separately. For instance, binomial coefficients $\binom{n}{k}$ are hypergeometric because we have:[2]

$$\binom{n+1}{k} = \frac{n+1}{n-k+1}\binom{n}{k}$$

and

$$\binom{n}{k+1} = \frac{n-k}{k+1}\binom{n}{k}$$

[1] At http://algo.inria.fr/libraries/autocomb/Apery2-html/apery.html, Bruno Salvy gives an elegant checking of the entire logic of Apéry's argument making extensive use of computerized symbolic tools.

[2] We will ignore the question of zero denominators in the rational functions for now, though it will be discussed extensively later. One often sees a function defined as hypergeometric if $a_{n+1}/a_n = r(n)$ whenever the rational function $r(n)$ is well-defined. Alternative one may define it as hypergeometric if it is holonomic with a recurrence of order 1, $q(n)a_{n+1} - p(n)a_n = 0$.

Since rational functions form a field, rational functions are themselves hypergeometric, and the set of hypergeometric sequences is closed under multiplication and division. The hypergeometric sequences are *not* in general closed under addition though.

2.1 Gosper's Algorithm

A landmark in symbolic computation was Gosper's algorithm [10] for indefinite hypergeometric summation. Given a hypergeometric term t_k, it is able either to find a hypergeometric 'antidifference' term s_k such that $s_{k+1} - s_k = t_k$, or to conclude that no such hypergeometric term exists. Note that if s_k is hypergeometric, say $s_{k+1}/s_k = r(k)$ for some rational function $r(k)$, then the antidifference property implies that $(r(k) - 1)s_k = t_k$, so s_k and t_k are rational-function multiples of each other, and in particular the original t_k must also be hypergeometric. However, the converse is false: there are hypergeometric t_k with no hypergeometric antidifference. In such cases Gosper's algorithm will fail, but failure does at least definitively show that *there is no* hypergeometric antidifference.

Such an antidifference s_k plays a role for summations analogous to an indefinite integral or antiderivative $F'(x) = f(x)$ for definite integrals, which allow us to conclude by the Fundamental Theorem of Calculus that $\int_a^b f(x)dx = F(b) - F(a)$. In the case of summations we get a 'telescoping' sum where most of the terms cancel:

$$\sum_{k=a}^{b} t_k = \sum_{k=a}^{b} (s_{k+1} - s_k) = s_{b+1} - s_a$$

Thus, when Gosper's algorithm solves the 'indefinite summation' problem by finding an antidifference, it also gives a solution to the corresponding 'definite summation' problem.

Gosper's algorithm works over an arbitrary field of characteristic zero, which we may consider as the field of rational functions in other variables, so it can tackle problems containing other variables as parameters without any difficulty. For example [14], let us prove the following that was originally proposed as Problem E 3088 in the "American Mathematical Monthly".

$$\sum_{k=1}^{n} \frac{k \cdot k!}{n^k} \binom{n}{k} = n$$

Setting $t_k = \frac{k \cdot k!}{n^k}\binom{n}{k}$, Gosper's algorithm finds that $s_k = -\frac{n}{k} t_k = -\frac{n \cdot k!}{n^k}\binom{n}{k}$ is an antidifference, i.e. $s_{k+1} - s_k = t_k$, and therefore

$$\sum_{k=1}^{n} \frac{k \cdot k!}{n^k} \binom{n}{k} = s_{n+1} - s_1 = -\frac{n \cdot (n+1)!}{n^{\underline{n+1}}} \binom{n}{n+1} - -\frac{n \cdot 1!}{n^1} \binom{n}{1} = 0 - -n = n$$

2.2 WZ Pairs

Appealing though that example was, there are many cases where Gosper's algorithm finds there is no hypergeometric antidifference, even though the *definite* sum has a hypergeometric answer. This applies even to our initial example of $\sum_k \binom{n}{k}$, which certainly has a hypergeometric *definite* sum (namely 2^n), but where Gosper's algorithm shows there is no hypergeometric antidifference. The WZ method uses a variant of Gosper's algorithm at its core but can handle a much wider class of definite sums.

We will now make the parametrization by n explicit, considering summing series of the form $\sum_k F(n, k)$ where the summand is a hypergeometric function of two variables. We

will be concerned with summing functions with *finite support* w.r.t. k, i.e. those that for any fixed n are zero for k outside some finite set of integers; for example $\binom{n}{k} = 0$ except for $0 \leqslant k \leqslant n$. When we write sums without explicit ranges, we mean they are summed over all integers, which is a well-defined notion because of this finite support property.

The key idea in the basic WZ algorithm is to find a hypergeometric antidifference not directly for $F(n, k)$ but for the difference $F(n+1, k) - F(n, k)$. The antidifference $G(n, k)$, when it exists, yields a *WZ pair*, meaning that

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

Such a pair is interesting from the point of view of finding $\sum_k F(n, k)$ because of a similar telescoping phenomenon:

$$\begin{aligned}\sum_{k=a}^{b} F(n+1, k) - \sum_{k=a}^{b} F(n, k) &= \sum_{k=a}^{b} (F(n+1, k) - F(n, k)) \\ &= \sum_{k=a}^{b} (G(n, k+1) - G(n, k)) \\ &= G(n, b+1) - G(n, a)\end{aligned}$$

Since we are assuming that our F and G are hypergeometric, we have $F(n+1, k)/F(n, k) = r(n, k)$ and $G(n, k+1)/G(n, k) = s(n, k)$ for some rational functions $r(n, k)$ and $s(n, k)$. Therefore from the WZ-pair property we have $(r(n, k) - 1)F(n, k) = (s(n, k) - 1)G(n, k)$ and so

$$R(n, k) = G(n, k)/F(n, k) = (r(n, k) - 1)/(s(n, k) - 1)$$

is a rational function. Thus, we may assume that $G(n, k) = R(n, k)F(n, k)$ for some rational function $R(n, k)$.

This observation also implies that each $G(n, k)$ too (as a function of k for fixed n) has finite support. So by making the limits of summation $a, \ldots, b$ sufficiently wide, we can assume $G(n, b+1) = G(n, a) = 0$ and so conclude that $\sum_k F(n+1, k) - \sum_k F(n, k) = 0$, i.e. that $\sum_k F(n, k)$ is *independent of* n. So we can simply evaluate the sum $\sum_k F(n, k)$ at some convenient value like $n = 0$, giving $\sum_k F(0, k)$, which in practice we expect to be easy. In particular, in the common case where the actual support of the function is contained in $\{(k, n) \mid 0 \leqslant k \leqslant n\}$ (which occurs whenever the summand is a multiple of $\binom{n}{k}$), the sum is just $F(0, 0)$.

The class of problems to which this technique is applicable is surprisingly wide. In some cases one needs a further generalization to finding recurrences of finite order of the following form, for polynomials $a_0(n), \ldots, a_m(n)$:

$$a_0(n)F(n, k) + a_1(n)F(n+1, k) + \cdots + a_m(n)F(n+m, k) = G(n, k+1) - G(n, k)$$

Summing over k (since each $a_i(n)$ is independent of k) we get a similar recurrence where the terms F are replaced by $\sum_k F$. While in general this may not give a hypergeometric closed form, finding this sort of recurrence is exactly the kind of thing we wanted for Apéry's recurrences, for example. Although in general we may need to use such recurrences, the basic method (after some initial normalization we discuss next) succeeds remarkably often; this is sometimes referred to as the 'WZ phenomenon' [15].

2.3 The WZ Algorithm

The basic WZ algorithm, as presented in [15], is as follows, to verify a summation result of the form

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

1. Divide through by $S(n)$ so that we may thereafter assume that we're trying to verify just the following special case (if the r.h.s. is 0 we skip this, but the rest of the algorithm is essentially the same).
$$\sum_k F(n,k) = 1$$
2. Apply a Gosper-like algorithm to find a WZ-mate $G(n,k)$ (which as noted above we can suppose to be of the form $R(n,k)F(n,k)$ for some rational function $R(n,k)$) such that
$$F(n+1,k) - F(n,k) = G(n,k+1) - G(n,k)$$
3. Conclude as noted above that $\sum_k F(n,k)$ is independent of n and so we just need to check the following, which we expect to be easy
$$\sum_k F(0,k) = 1$$

2.4 An Example

Let us consider the sum of squares of binomial coefficients, following closely the presentation in [15]. Our goal here is to prove the following:

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

1. We divide through by the right-hand side and so reduce the problem to verifying $\sum_k F(n,k) = 1$ where
$$F(n,k) = \frac{\binom{n}{k}^2}{\binom{2n}{n}} = \frac{n!^4}{k!^2(n-k)!^2(2n)!}$$
2. We obtain the magic rational function
$$R(n,k) = \frac{-k^2(3n-2k+3)}{2(n-k+1)^2(2n+1)}$$
such that $G(n,k) = R(n,k)F(n,k)$ satisfies the key property
$$F(n+1,k) - F(n,k) = G(n,k+1) - G(n,k)$$
Verifying this is just a routine algebraic manipulation.
3. We have therefore concluded that $\sum_k F(n,k)$ is independent of n so we just need to verify some convenient special case such as $n = 0$. Since then $\binom{2n}{n} = \binom{0}{0} = 1$ and $\binom{0}{k}^2 = 0$ for $k \neq 0$, we get the result.

3 Formally Certifying WZ Results

From now on, we will be showing a number of explicit theorems in HOL Light's ASCII syntax, and to read some of the notations the following little glossary may be helpful. First, these are the ASCII representations of logical connectives etc.

Standard symbol	ASCII version	Meaning
$\bot$	`F`	falsity
$\top$	`T`	truth
$\neg p$	`~p`	not p
$p \wedge q$	`p/\ q`	p and q
$p \vee q$	`p \/ q`	p or q
$p \Rightarrow q$	`p ==> q`	if p then q
$p \Leftrightarrow q$	`p= q`	p if and only if q
$\forall x.\ p$	`!x. p`	for all x, p
$\exists x.\ p$	`?x. p`	there exists x such that p
$\lambda x.\ t$	`\x. t`	the function $x \mapsto t$

while these are some other mathematical notations that might otherwise be obscure. (For example '`Cx(&0)`' is the rather verbose representation of the complex constant 0.)

Standard symbol	ASCII version	Meaning
$-x$	`--x`	Unary negation
$n!$	`FACT(n)`	Factorial function
$\lvert x \rvert$	`abs(x)`	absolute value function
$\circ$	`o`	function composition
N/A	`&`	type cast $\mathbb{N} \to \mathbb{R}$
N/A	`Cx`	type cast $\mathbb{R} \to \mathbb{C}$
(juxtaposition)	`%%`	scalar-matrix multiplication

In general, results given by powerful tools like computer algebra systems can be difficult to prove in a formal, rigorous way inside a foundational theorem prover. However, it may be very easy when those tools can produce some kind of easily checkable *certificate*. For example in [12], Maple is used to perform polynomial factorization and transcendental function integration. In each case the checking process (respectively multiplying polynomials and taking derivatives) is substantially easier than the process of finding the certificate, and rather easy to implement foundationally. Similarly, some certificates of primality can be checked quite easily even though the process of *finding* them requires complicated and difficult processes like factorization of large integers [5, 11, 19]. On the face of it the WZ algorithm also belongs in this 'easy' category, because we can rely on an existing implementation to provide the magical rational function $R(n, k)$, and all we need to do is check some routine algebraic manipulations. Indeed, there are readily available implementations of Gosper's algorithm and the WZ method. We'll use the one developed by Fabrizio Caruso in `maxima`:

```
Maxima 5.20.1 http://maxima.sourceforge.net
using Lisp GNU Common Lisp (GCL) GCL 2.6.7 (a.k.a. GCL)
Distributed under the GNU Public License. See the file
COPYING.
Dedicated to the memory of William Schelter.
The function bug_report() provides bug reporting
information.
```

We start by loading the `zeilberger` package:

```
(%i1) load(zeilberger);
```

We can easily use this to find the antidifference in our Gosper example $t_k = \frac{k \cdot k!}{n^k}\binom{n}{k}$, which gets presented as $s_k = -k!n^{1-k}\binom{n}{k}$:

```
(%i2) AntiDifference(k * k! * binomial(n,k) / n^k,k);
                                    1 - k
(%o2)                        - k! n      binomial(n, k)
```

and we can also solve our WZ problem, resulting in exactly the rational function certificate $R(n,k) = \frac{-k^2(3n-2k+3)}{2(n-k+1)^2(2n+1)}$ we gave above. (The additional $[-1, 1]$ in the output means that this gives an antidifference for $-1 \cdot F(n,k) + 1 \cdot F(n+1,k) = F(n+1,k) - F(n,k)$; in general the algorithm can find antidifferences for more complicated recurrence combinations as mentioned above.)

```
(%i3) Zeilberger(binomial(n,k)^2 / binomial(2 * n,n),k,n);
                              2
                             k  (3 n - 2 k + 3)
(%o3)                [[- ------------------------, [- 1, 1]]]
                         2 2  (n - k + 1)  (2 n + 1)
```

Thus, it seems we just need to solve the straightforward engineering issues about linking `maxima` and HOL Light, and we can obtain formal proofs pretty easily. The basic 'telescoping' argument is simple, and there are already corresponding formal theorems, e.g.

```
SUM_DIFFS_ALT =
  |- !m n.
         sum (m..n) (\k. f (k + 1) - f k) =
         (if m <= n then f (n + 1) - f m else &0)
```

The only other requirement is to formalize the algebraic simplifications involved in checking the antidifference or WZ-pair properties, as well as the equivalence between the end result and the purported answer. In the case of the Gosper example the antidifference property is

$$-(k+1)!n^{1-(k+1)}\binom{n}{k+1} - -k!n^{1-k}\binom{n}{k} = \frac{k \cdot k!}{n^k}\binom{n}{k}$$

and the end result to check is

$$-(n+1)!n^{1-(n+1)}\binom{n}{n+1} - -1!n^{1-1}\binom{n}{1} = n$$

while for the WZ example, we need to check the WZ-pair property

$$\frac{\binom{n+1}{k}^2}{\binom{2(n+1)}{n+1}} - \frac{\binom{n}{k}^2}{\binom{2n}{n}} = \frac{-[k+1]^2(3n-2[k+1]+3)}{2(n-[k+1]+1)^2(2n+1)}\frac{\binom{n}{k+1}^2}{\binom{2n}{n}} - \frac{-k^2(3n-2k+3)}{2(n-k+1)^2(2n+1)}\frac{\binom{n}{k}^2}{\binom{2n}{n}}$$

and the end result

$$\sum_k \frac{\binom{0}{k}^2}{\binom{2\cdot 0}{0}} = 1$$

verifications that are described by [15] in the following terms:

> Well, at this point we have arrived at a situation that will be referred to throughout this book as a "routinely verifiable" identity. That phrase means roughly that your pet chimpanzee could check out the equation. More precisely it means this. First cancel out all factors that look like c^n or c^k [...] that can be cancelled. Then replace every binomial coefficient in sight by the quotient of factorials that it represents. Finally, cancel out all of the factorials by suitable divisions, leaving only a polynomial identity that involves n and k. After a few more strokes of the pen, or keys on the keyboard, this identity will reduce to the indisputable form $0 = 0$, and you'll be finished with the "routine verification".

However, if we examine our examples in more detail — which an attempt to formalize them forces one to do — this starts to look somewhat too glib. For a start, we have assumed that we are summing over *all* integers, relying on the finite support property. In this case, we need a priori to justify the WZ-pair property for arbitrary k, and it's not at all obvious that the equation that was nonchalantly asserted by [15] in their presentation of the WZ example (assumed implicitly in 'replace every binomial coefficient in sight by the quotient of factorials')

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

is true in the case $k > n$ when the LHS collapses to zero, or indeed even meaningful (how do we define factorials of negative integers?) Similarly, when $k = n$ or $k = n+1$ one of the denominators in the expression on the right of the WZ-pair property becomes zero, which makes the whole interpretation of the formula questionable — at the very least we seem to be cancelling terms of the form $0/0$. (Ignoring division by zero is a staple trick in 'proofs' of $1 = 2$, so concern doesn't seem entirely unwarranted.) And even if we are less casual about summing over 'all integers' and just consider the k we really need, we do still apparently need to prove the WZ-pair property in the problematic case $k = n+1$ to justify the reasoning. All in all, it is difficult to take these plausible-looking presentations at face value.

One solution to this problem is to explicitly exclude the troublesome points k from our sum, and add those values of the summand back in separately. This approach is workable, and we have managed to formalize a few examples in HOL Light in this way. However in more complicated instances, the troublesome singularities occur not just at the upper limit of summation, but at multiple points in between. This makes automation tricky in many ways, not only in the way the size of the formula expands, but because we actually need to be able to determine where the singularities are, which for a general bivariate polynomial is not a trivial matter. In a similar way, the formalization of Apéry's proof reported in [4], which uses related techniques, required considerable semi-manual intervention to handle such special cases. Moreover, it is unsatisfying to depart so radically from the plausible-looking informal counterpart, even if it is not obvious how to make it completely rigorous.

4 Limits to the Rescue?

If we seek a really rigorous proof, even some of the other casual assumptions that we made earlier need to be re-examined. For example, we observed that a hypergeometric antidifference must be a rational function of the original term, in our example $-\frac{n}{k}\frac{k\cdot k!}{n^k}\binom{n}{k} = -\frac{n\cdot k!}{n^k}\binom{n}{k}$. On the face of it, this is problematic when $k = 0$. (As it happens we didn't really need this for $k = 0$ in our example, but it would have arisen if we'd started the sum at $k = 0$, which we could perfectly well have done as the summand vanishes.)

If we could work systematically not over *values* with free parameters but rather *rational functions* as syntactic objects, then manipulations like $x/x = 1$ become unobjectionable. However, doing so in a formal context adds a lot of complications, and it's not at all clear how to bring binomial coefficients and factorials into this worldview. Instead, we will consider a slightly different approach using limits. For motivation, consider the formal proof in HOL Light of Sylvester's determinant identity,[3] which states that for an $m \times n$ matrix A and an $n \times m$ matrix B, we have the following, I representing any identity matrix of the appropriate dimension:

$$\det(I + AB) = \det(I + BA)$$

A nice simple intuitive proof is as follows. By the multiplication law for determinants

$$\begin{aligned}\det(I + AB)\det(A) &= \det([I + AB]A)\\ &= \det(A + ABA)\\ &= \det(A[I + BA])\\ &= \det(A)\det(I + BA)\end{aligned}$$

and the result then follows by cancelling $\det(A)$ from both sides. Delightfully simple as this is, there are two problems. First, if A is not square, $\det(A)$ has no obvious meaning — however it is rather easy and dull to get round this objection by padding the matrices out so we can assume them square. More interestingly, the proof collapses if $\det(A) = 0$, i.e. if A is singular (non-invertible).

One approach is to work over a 'generalized' polynomial ring.[4] A somewhat different approach that we take, which has a similar net effect but seems simpler, is to argue by *continuity*; often this style of argumentation is referred to as 'generic'. It is not too hard to prove that for every square matrix A (whether itself invertible or not) there is some $\epsilon > 0$ such that for $0 < |x| < \epsilon$ the perturbed matrix $A + xI$ is invertible.

```
NEARBY_INVERTIBLE_MATRIX =
 |- !A:real^N^N.
       ?e. &0 < e /\
           !x. ~(x = &0) /\ abs x < e ==> invertible(A + x %% mat 1)
```

Since $f(x) = \det(I+(A+xI)B)-\det(I+B(A+xI))$ is a continuous function (indeed, it's a polynomial in x), it suffices to prove that $f(x) = 0$ in a neighbourhood of 0 (which the proof for the invertible case does) and we can also conclude $f(0) = 0$. This, wrapped up in a more general 'assume without loss of generality that a matrix is invertible' theorem, is how the HOL Light proof of Sylvester's determinant identity works.

It is using reasoning of this form that we propose to understand the WZ method, interpreting the various formulas involved as *limits*. In order to interpret binomial conditions and

[3] In http://code.google.com/p/hol-light/source/browse/trunk/Multivariate/topology.ml with the name `SYLVESTER_DETERMINANT_IDENTITY`.

[4] See http://math.stackexchange.com/questions/17831/sylvesters-determinant-identity

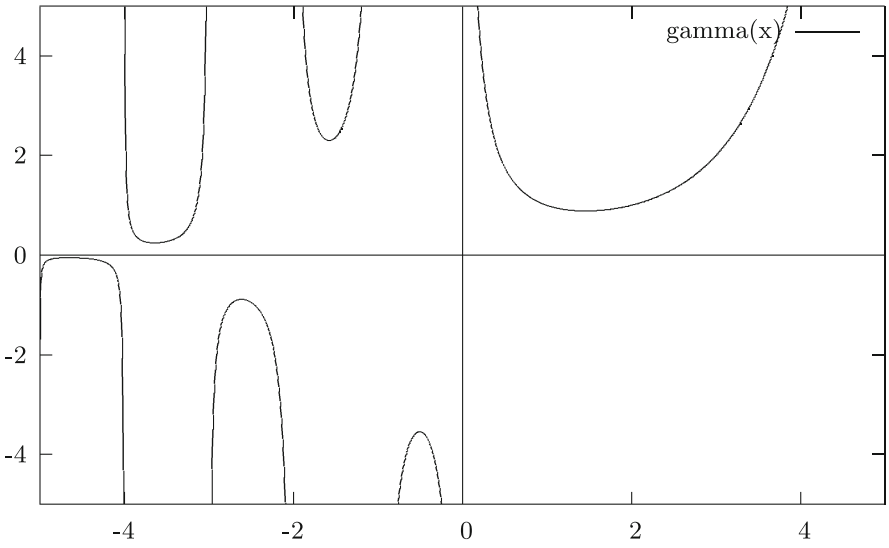
factorials as limits, we will use the *gamma function*, which extends the factorial in a natural way into a function of a real variable, or indeed a complex one. We will take up the formalization of the gamma function in HOL Light in the next section and then describe how it can be used to formalize the kind of reasoning we were struggling with above. Note that this is not really an original idea: if one looks carefully, one can find presentations of WZ using gamma function limits [8]

> This enables us to verify [...] without having to worry about the domain of definition of F and G. We may then take a limit, where necessary, to deal with cases in which the formulas for F and G may be ambiguous (and these cases often arise in interesting identities).

This may even be exactly how the people using the WZ method actually think intuitively about factorials. Still, one struggles to find any such explanation or justification in the popular expositions, and as far as we know it has never been machine-formalized before.

5 The Gamma Function

The gamma function $\Gamma(z)$ is a natural extension of the factorial function to real and complex arguments. In fact, the functions are offset by 1: for natural numbers n we have $n! = \Gamma(n+1)$. The complex gamma function is holomorphic (complex differentiable) in the entire complex plane with the exception of the nonpositive integers $\{0, -1, -2, -3, \ldots\}$ where it has poles. (Actually the reciprocal $\Gamma^{-1}(z) = 1/\Gamma(z)$ of the gamma function, setting $\Gamma^{-1}(-n) = 0$ at Γ's poles, is entire, i.e. holomorphic in the whole complex plane.) Likewise, the real gamma function is differentiable everywhere except at the nonnegative integers where it shoots off to $\pm\infty$, as shown in the following picture:



The gamma function can be defined in various different ways. The first formalization in a mechanical theorem prover was the development of the real gamma function in HOL4 by Siddique and Hasan [17] using Euler's integral $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. We elect to start by developing the complex gamma function and then we can fairly straightforwardly derive

properties of the real function from it. As it happens, for present purposes the real gamma function is entirely adequate, but the definitions and proofs for the complex gamma function are a fairly straightforward generalization of their real counterparts, and it seemed a good investment for the future. One can just as well use the Euler integral $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ to define the complex function, where the integral is over the reals but the parameter z and hence the integrand is complex. We eventually derive that as a consequence of our definition; the disadvantage of taking it *as* the definition is that it's only valid for $\Re(z) > 0$ (or $x > 0$ in the real case), so some additional measures are needed to extend it to the negative half of its domain. As we shall see, we *do* make essential use of Γ's behaviour close to negative integers, so we find it more convenient to define $\Gamma(z)$ once and for all by the limit

$$\Gamma(z) = \lim_{n\to\infty} \frac{n^z n!}{\Pi_{m=0}^{n}(z+m)}$$

or in HOL Light:

```
|- cgamma(z) =
   lim sequentially (\n. (Cx(&n) cpow z * Cx(&(FACT n))) /
                         cproduct(0..n) (\m. z + Cx(&m)))
```

Note that this HOL formulation inherits the effects of our definition $0^{-1} = 0$, so at the poles the function receives the concrete value 0. These limit definitions can be reformulated in various equivalent ways, e.g.[5]

$$\Gamma(z+1) = \lim_{n\to\infty} n^z \frac{1}{(z+1)}\frac{2}{(z+2)}\frac{3}{(z+3)}\cdots\frac{n}{(z+n)}$$

We find it convenient to perform the convergence and holomorphy proofs first on the *logarithm* of the Gamma function using the additive analog:

$$\log\Gamma(z) = \lim_{n\to\infty} z\log(n) - \log(z) - \Sigma_{m=1}^{n}\log((m+z)/m)$$

Although we write this function informally as $\log\Gamma(z)$, it is *not* just the result of applying the usual (principal value) complex logarithm to the Gamma function,[6] In the formal HOL development it is a distinct function `lgamma` defined as follows:[7]

```
|- lgamma z = lim sequentially
              (\n. z * clog(Cx(&n)) - clog z -
                   vsum(1..n) (\m. clog((Cx(&m) + z) / Cx(&m))))
```

By expanding $\log(n)$ as a telescoping sum $\sum_{m=2}^{n}(\log(m) - \log(m-1))$, one can express the limit mainly as a sum of terms like $z\log(1 - 1/m) - \log(1 + z/m)$, and since the linear terms of those Taylor series cancel nicely, it's easy to show convergence, even locally uniform convergence for z bounded away from negative integers. This not only implies

[5]The n^{th} elements of these sequences differ by a factor of $n/(n+1)$ so they have the same limit.

[6]Lang's complex analysis book [13] confusingly seems to imply on p423 that it is, though presumably it is not intended to. Exercise 35 of Chapter 9 (p. 125) of [6] gives an explicit estimate for the imaginary part of $\log\Gamma(z)$.

[7]This function has applications of its own, which is one motivation for giving it a distinct definition, given the difficulties in recovering it from the gamma function with the right imaginary part. For example, it is implemented in Mathematica as `LogGamma`, in MATLAB as `gammaln`.

that the limit does indeed exist except at negative integers, but that it defines a holomorphic function there:

```
|- !z. (Im z = &0 ==> &0 < Re z)
       ==> lgamma complex_differentiable at z
```

Now it is easy to demonstrate the key properties of the gamma function that we need, that the limit does indeed exist for all z, the function is zero only in the degenerate case where z is a negative integer, and except at these points it is the exponential of the `lgamma` function.

```
|- !z. cgamma z = Cx(&0) <=> (?n. z + Cx(&n) = Cx(&0))

|- !z. cgamma(z) = if ?n. z + Cx(&n) = Cx(&0) then Cx(&0)
                   else cexp(lgamma z))
```

This last property easily implies that Γ is holomorphic except at negative integers.

```
|- !z. (!n. ~(z + Cx(&n) = Cx(&0)))
       ==> cgamma complex_differentiable at z
```

We derive numerous elementary theorems relatively easily, such as $\Gamma(1) = 1$ and the fact that Γ commutes with complex conjugation

```
|- !z. cnj(cgamma z) = cgamma(cnj z)
```

That implies in particular that when applied to real arguments it gives real results, and hence the corresponding real function defined as

```
|- gamma(x) = Re(cgamma(Cx x))
```

has the following key property:

```
|- !x. Cx(gamma x) = cgamma(Cx x)
```

This allows all the required properties of the real gamma function to be derived straightforwardly from those of the complex function. The most important properties for us in what follows are the recurrence formulas, which extend the basic property $(n + 1)! = (n + 1)n!$ of factorials. Note that the second one needs no case analysis because $0^{-1} = 0$ and so it works degenerately in that case.

```
|- !x. gamma(x + &1) = if x = &0 then &1 else x * gamma(x)

|- !x. gamma(x) = gamma(x + &1) / x
```

The proof (for the complex function first) is almost immediate just by rearranging the limit that defines Γ. Legendre's duplication formula $\Gamma(z)\Gamma(z + 1/2) = 2^{1-2z}\Gamma(1/2)\Gamma(2z)$ is also relatively easy just by following one's nose:

```
|- !x. gamma(x) * gamma(x + &1 / &2) =
       &2 rpow (&1 - &2 * x) * gamma(&1 / &2) * gamma(&2 * x)
```

The only really non-trivial proof is involved in the so-called reflection formula $\Gamma(z)\Gamma(1-z) = \pi/\sin \pi z$; we use a complex-analytic variant of a nice proof for the real gamma function due to Artin [2])

```
|- !x. gamma(x) * gamma(&1 - x) = pi / sin(pi * x)
```

This implies easily that $\Gamma(1/2) = \sqrt{\pi}$, which we can use to rewrite the duplication formula as $\Gamma(z)\Gamma(z+1/2) = 2^{1-2z}\sqrt{\pi}\Gamma(2x)$. Another interesting property of the *real* gamma function (considered as a function with domain $\{x \in \mathbb{R} \mid x > 0\}$) is that it is log-convex (i.e. its logarithm is a convex function)

```
|- !s. (!x. x IN s ==> &0 < x) ==> gamma real_log_convex_on s
```

and is in fact *uniquely characterized* by just a few properties including the recurrence and log-convexity:

```
|- !f. f(&1) = &1 /\
       (!x. &0 < x ==> f(x + &1) = x * f(x)) /\
       (!x. &0 < x ==> &0 < f x) /\
       f real_log_convex_on {x | &0 < x}
       ==> !x. &0 < x ==> f x = gamma x
```

We are also finally able to justify the definition via the Euler integral in a suitable domain. Interestingly, Euler originally seems to have used $\Gamma(x) = \int_0^1 (-\log t)^{x-1} dt$, which can be related to the more common integral by a change of variables. Another change of variables and the fact that $\Gamma(1/2) = \sqrt{\pi}$ yields as a nice piece of mathematical collateral the value of the Gaussian integral $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$.

```
|- ((\x. exp(--(x pow 2))) has_real_integral sqrt pi) (:real)
```

We also derive various forms of Stirling's approximation, but since the formalization of the gamma function is not the main topic of this paper we will not dwell on this development any more.

6 Real Generalizations of Operations

In order to realize the idea of a limit-based formulation of the WZ method, we need two pieces of mathematical infrastructure:

1. Real-number generalizations of the basic functions involved in the problem
2. A proof that one can approach the limit avoiding 'troublesome' values.

The major part of the answer to (1) is the gamma function that we described above. To keep a convenient relationship between the real generalizations and their original natural number counterparts we define real variants of the factorial function and binomial coefficient; see [7] for some discussion of the latter. Note that we use the names n and k but these are variables of type $\mathbb{R}$ not $\mathbb{Z}$ and the subtraction is true subtraction over $\mathbb{R}$.

```
|- rfact x = gamma(x + &1)

|- rbinom(n,k) = rfact n / (rfact k * rfact (n - k))
```

These functions generalize their natural number counterparts and behave quite well with respect to limits. For example, the following expresses the fact that if some function `nn` tends to a natural number limit (according to some convergence net '`net`'), then the corresponding factorial tends to the expected limit:

```
|- !net nn n.
       (nn ---> &n) net
       ==> ((\a. rfact (nn a)) ---> &(FACT n)) net;
```

The real factorial function, for example, is continuous except at negative integers, and its reciprocal is actually continuous everywhere (with value 0 at negative integers). (Here '`o`' is function composition and '`inv`' is the multiplicative inverse function on reals.)

```
|- !x. ~(integer x /\ x <= -- &1)
       ==> rfact real_continuous atreal x

|- !x. inv o rfact real_continuous atreal x
```

As a little thought shows, the real binomial function has the correct limit as the argument tends to (n, k) for *any* nonnegative integers n and k, so in some sense it 'builds in' the fact that $\binom{n}{k} = 0$ for $k > n$.

```
|- !net nn kk n k.
       (nn ---> &n) net /\ (kk ---> &k) net
       ==> ((\a. rbinom (nn a,kk a)) ---> &(binom (n,k))) net
```

Moreover, if $k \in \mathbb{Z}$ with $n \geqslant 0$ and $k < 0$ it tends to zero at (n, k). This means that we can safely sum (in our limit-based formulation) $\binom{n}{k}$ over *all* integers k, positive or negative, provided n is nonnegative (integer or not).

```
|- !net nn kk n k.
       (nn ---> n) net /\ (kk ---> k) net /\
       integer k /\ &0 <= n /\ k < &0
       ==> ((\a. rbinom (nn a,kk a)) ---> &0) net
```

So much for factorials and binomial coefficients. All the basic arithmetic operations like addition and multiplication are mapped to their real counterparts, and the power function is mapped to the standard 'real to the power real' power function in the HOL Light library `rpow`. In fact, this last identification proves a little troublesome since the behavior of x^y when $x < 0$ is not trivial and is often left undefined. The HOL Light `rpow` function is defined for negative x in a rather intricate way to make at least some intuitive laws of exponents work for arbitrary *rational* exponents (for example $(-2)^{\frac{1}{3}} = -\sqrt[3]{2}$). However, this definition results in a function that is not continuous, making it quite unsuitable for use in our limit-based formulation because $y_n \to y$ does not in general imply $x^{y_n} \to x^y$ for $x < 0$. For the most part, as we shall see, we can assume $x > 0$ in common examples, but one outstanding exception is that many problems involve alternating signs expressed as $(-1)^k$,

and we do need to generalize these. Our solution[8] is to use the function $x \mapsto \cos(\pi x)$, which has all the required properties:

- For x an integer, it gives the intuitively expected answer $\cos(\pi x) = (-1)^x$, so the original problem mapping works correctly.
- It is continuous everywhere, so it can be used in a limit-based argument
- It satisfies $\cos(\pi(x+1)) = -\cos(\pi x)$ for all real x, which is the key property needed in the algebraic identities.

7 Safely Approaching the Limit

In order to justify the 'naive' rearrangements we want to be able to approach the limit (n, k) via pairs of reals $(n + \delta, k + \epsilon)$ that avoid various problematic issues where the simple rearrangements break down. For example, we certainly want to avoid any poles of the rational function, any points where other denominators are zero, and places where the naive expansion formulas like $(x+1)! = (x+1)x!$ cease to be valid. All the difficult pairs of values have something in common: they are of the form (x, y) where for some polynomial $p \in \mathbb{Q}[x, y]$ (i.e. one with rational coefficients in two variables x and y) the pair satisfies $p(x, y) = 0$. Given any polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$, its set of roots, the algebraic variety it defines in the language of algebraic geometry, $\{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid p(x_1, \ldots, x_n) = 0\}$ is a 'small' subset of $\mathbb{R}^n$ except in the degenerate case where p is the zero polynomial. More precisely, it has Lebesgue measure zero:

```
|- !f c. real_polynomial_function f /\ ~(!x. f x = c)
        ==> negligible {x | f x = c}
```

and (consequently, since it is also closed, being the preimage of a singleton set under a continuous function) is nowhere dense:

```
|- !f c. real_polynomial_function f /\ ~(!x. f x = c)
        ==> interior (closure {x | f x = c}) = {}
```

Because these two notions of 'small' are preserved under finite unions, it is easy to see that we can make $(x, y) \to (n, k)$ avoiding any particular finite set of polynomial constraints. In fact, we can do what for most purposes is even more convenient: we can avoid *all possible* polynomial constraints with rational coefficients, because our 'smallness' properties (measure zero, or closed set with empty interior) are also preserved by *countable* unions. The latter is a Baire-type result, which in HOL looks like this:

```
|- !g. COUNTABLE g /\ (!s. s IN g ==> closed s /\ interior s = {})
      ==> interior(UNIONS g) = {}
```

[8] In an earlier version of this work we used an ad hoc triangle wave function, but changed to $\cos(\pi x)$ at the suggestion of one of the referees.

We make a somewhat ad hoc definition of the set of bivariate rational polynomial functions, where '`(:num#num) CROSS rational`' just means $\mathbb{N} \times \mathbb{N} \times \mathbb{Q}$:

```
|- ratpolyfun p <=>
       ?s. FINITE s /\
           s SUBSET (:num#num) CROSS rational /\
           p = \(x,y). sum s
                (\((i,j),c). c * x pow i * y pow j)
```

and prove without difficulty that it has various expected closure properties, so it's easy to check that a specific expression is indeed such a function. For example, this is the HOL theorem that the set is closed under pointwise addition:

```
|- !p q. ratpolyfun (\(n,k). p n k) /\ ratpolyfun (\(n,k). q n k)
         ==> ratpolyfun (\(n,k). p n k + q n k));
```

Critical for us is that there are only countably many polynomials with rational coefficients (this is almost immediate from the definition), and hence we will be able to avoid them all in the passage to the limit.

```
|- COUNTABLE ratpolyfun
```

We define a pair of real numbers $t = (x, y)$ to be *ratty* (the name is meant to suggest rational numbers as well as undesirability) if it is a zero of a *non-trivial* rational polynomial function

```
|- ratty t <=> ?p. ratpolyfun p /\ p t = &0 /\ ~(!w. p w = &0)
```

It is now easy to see that we can approach a pair (n, k) arbitrarily closely by non-ratty pairs (x, y). In fact, we can choose the approach to satisfy various other constraints, in particular so that it approaches from above ($x > n$ and $y > k$), which turns out to be somewhat convenient. We are now ready to prove the main theorems underlying our implementation.

Our first version, inspired directly by the way the WZ method is usually presented, used summation of the limit function over all *integers*. In some sense this is the most convenient approach, because we can assume the range of summation to be wide enough that it properly includes the support of the function *and* that the endpoints are not poles of the rational function (since a non-trivial polynomial only has finitely many roots). However, this approach requires us (at least prima facie) to show that the limit as (n, k) is approached is always zero for $k < 0$ and $n \geqslant 0$. As we noted, for the binomial coefficient $\binom{x}{y}$ itself, this is indeed the case. But for composite expressions in general it becomes more complicated. For example, to show that the limit form of the Apéry's summand $\binom{n+k}{k}^2\binom{n}{k}^2$ tends to 0 for $n, k \in \mathbb{Z}, k < 0 \leqslant n$ is not quite trivial: although we know $\binom{x}{y}^2 \to 0$ as $(x, y) \to (n, k)$, we need to show that $\binom{x+y}{y}$ doesn't approach $\pm\infty$ so fast that it overwhelms it. It so happens that in this case it doesn't, but that is only by virtue of the fact that we approach the limit from above, and in more complicated examples it's not in general the case at all. On the other hand, we noticed that cases where the rational function has a pole at $k = 0$ are almost unknown, and it therefore seemed simpler to just consider summation over all *natural numbers*, which completely obviates these difficulties. Thus, our main theorem is as follows:

```
|- !P f h ff hh p q.
     ((!n. FINITE {k | ~(f n k = &0)}) /\
      (!n. FINITE {k | ~(h n k = &0)})) /\
     ((!n k. P n ==> (ff ---> f n k) (at (complex(&n,&k)))) /\
      (!n k. P n ==> (hh ---> h n k) (at (complex(&n,&k))))) /\
     (ratpolyfun p /\ ratpolyfun q) /\
     (!n. P n ==> ~(q (&n,&0) = &0)) /\
     (!n k.
        ~ratty (n,k) /\ &0 < n
        ==> hh (complex (n,k)) =
            p (n,k + &1) / q (n,k + &1) * ff (complex(n,k + &1)) -
            p (n,k) / q (n,k) * ff (complex(n,k)))
     ==> (!n. P n
              ==> sum (:num) (\k. h n k) =
                  --(p (&n,&0) / q (&n,&0) * f n 0))
```

The idea is that we have some functions $\mathtt{f} : \mathbb{N} \to \mathbb{N} \to \mathbb{R}$ and $\mathtt{h} : \mathbb{N} \to \mathbb{N} \to \mathbb{R}$ that are (subject to some general constraint `P` that we carry around) limits of corresponding functions $\mathtt{ff} : \mathbb{C} \to \mathbb{R}$ and $\mathtt{hh} : \mathbb{C} \to \mathbb{R}$. In actual use, the function `h` is always a linear combination of offset versions of `f`, which allows us to handle the basic WZ method where $\mathtt{h}(n,k) = \mathtt{f}(n+1,k) - \mathtt{f}(n,k)$ as well as more general ones, but from the point of view of stating this theorem it is better to keep it generic. We suppose there are bivariate rational polynomials p and q with $q(n,0) \neq 0$ such that, ignoring some type distinctions in the formal counterpart, we have $\mathtt{hh}(x,y) = G(x+1,y) - G(x,y)$ where $G(x,y) = (p(x,y)/q(x,y))\mathtt{ff}(x,y)$ for all non-ratty (x,y) with $x > 0$. Then by a simple telescoping sum and limit argument we can conclude that the sum $\sum_k \mathtt{h}(n,k)$ is given by the limit of $-G(n,0)$.

8 Implementation

We have implemented a derived rule in HOL Light for proving Zeilberger-type results using the rational function certificates provided by the Maxima implementation. The inputs (ignoring some identifying the actual variables used in place of n and k in our generic examples) are as follows:

– A term specifying the main sum to be analyzed, e.g. $S_n = \sum_{k=0}^{n} \binom{n}{k}\binom{m}{k+p}/\binom{n+m}{n+p}$, or as a HOL term:

```
'sum (0..n) (\k. (&(binom(n,k)) * &(binom(m,k + p))) /
                 &(binom(n + m,n + p)))'
```

– The linear combination to be considered, e.g. the list $[1; -1]$ being short for $S_n - S_{n+1}$ as in the WZ method (in general these terms can be polynomials in n)

```
['&1'; '-- &1']
```

– Any additional assumptions about n or the other parameters, here $p \leqslant m$:

```
'&p <= &m'
```

– The rational function certificate, e.g. $k(p+k)/[(n-k+1)(n+m+1)]$:

```
'(&k * (&p + &k)) / ((&n - &k + &1) * (&n + &m + &1))'
```

Note that if one uses an actual interface to Maxima, the linear combination and the rational function are returned automatically by `Zeilberger`; all the user needs to provide is the term to be summed and the assumptions. However, it is convenient to have this 'manual' version to make it independent of Maxima and allow manual tweaking of its results.

The function automatically proves various required conditions on the sum, in particular that it has finite support and that the support is contained in the actual set used for summation, and instantiates the main theorem above appropriately. It then attempts to eliminate the five conjuncts of the antecedent:

- It proves that the functions have finite support. This was already done as part of the setup process so not much is needed at this stage except to lift the result to linear combinations.
- It proves that the real counterpart does indeed have the appropriate limit. This is normally proved by routine backchaining through limit composition theorems, but occasionally (for example when division is involved) this throws up additional proof obligations that certain terms are nonzero. These are themselves disposed of by some fairly simple ad hoc tactics.
- It proves that the numerator and denominator of the rational function are indeed rational polynomial functions. This is trivial by backchaining through composition theorems.
- It proves that $q(n, 0) \neq 0$ for arbitrary n, which collapses to proving various polynomials in n are nonzero. Fortunately, the denominators usually split into (and indeed are presented as) a product of linear factors, which can be analyzed separately. These in turn can usually be disposed of automatically either by linear arithmetic reasoning (e.g. $n + 1 \neq 0$ for $n \in \mathbb{N}$) or divisibility reasoning (e.g. $2m + 2n - 1 \neq 0$ for $m, n \in \mathbb{Z}$ because $2m + 2n$ is divisible by 2 but 1 is not).
- It proves the core algebraic rearrangement. Since it can assume that the real number pairs (n, k) are not ratty and that $n > 0$, this is indeed quite routine. First any arguments to factorials and binomial coefficients, and right-hand arguments of powers, are normalized into 'higher-order-term plus or minus iterated 1', e.g. $2(n + 1)$ becoming $(2n + 1) + 1$. Then 'stepping' theorems are applied to simplify these in the expected way like $(n - 1)! = n!/n$ — any side-conditions arising are trivialized because of non-rattiness while powers of n are easy to handle since we can also assume $n > 0$. The resulting expression is sometimes large and ugly but can invariably be solved quickly by a slightly specialized variant of HOL Light's existing `REAL_FIELD` tactic.

We can then draw the conclusion as in the main theorem, and after a little automatic simplification the RHS often collapses to 0 (as required in the WZ method) or some other constant. For example, in the present case we get the following:

```
|- !n. &1 * sum (0..n)
                 (\k. (&(binom(m,k)) * &(binom(n,k))) /
                      &(binom(m + n,m))) +
       -- &1 * sum (0..n + 1)
                    (\k. (&(binom(m,k)) * &(binom(n + 1,k))) /
                         &(binom(m + n + 1,m))) =
       &0
```

All this is totally automatic and the final result is proved with the usual HOL Light rigour.

9 Status

Our current code has been tested on about 50 examples drawn from various sources with an almost perfect success rate. Of course, sometimes we refined some of the automated tactics as a result of the fact that an interesting example failed, which gives our code a slightly ad hoc flavour. But on the whole it is quite robust. For example, it can easily handle the Apéry example, producing the following theorem

```
|- !n. --((&n + &1) pow 3) *
       sum (0..n)
           (\k. &(binom(n + k,k)) pow 2 *
                &(binom(n,k)) pow 2) +
       ((&2 * &n + &3) * (&17 * &n pow 2 + &51 * &n + &39)) *
       sum (0..n + 1)
           (\k. &(binom((n + 1) + k,k)) pow 2 *
                &(binom(n + 1,k)) pow 2) +
       --((&n + &2) pow 3) *
       sum (0..n + 2)
           (\k. &(binom((n + 2) + k,k)) pow 2 *
                &(binom(n + 2,k)) pow 2) =
       &0
```

as well as the following, which gives a WZ proof that $\sum_k (-1)^{n-k} 4^k \binom{n+k+1}{2k+1} = n+1$

```
|- !n. -- &1 *
       sum (0..n)
       (\k. (cos (pi * (&n - &k)) *
             &4 rpow &k *
             &(binom(n + k + 1,2 * k + 1))) /
             (&n + &1)) +
       &1 *
       sum (0..n + 1)
       (\k. (cos (pi * (&(n + 1) - &k)) *
             &4 rpow &k *
             &(binom((n + 1) + k + 1,2 * k + 1))) /
             (&(n + 1) + &1)) =
       &0
```

We also have convenient front-ends to perform the initial WZ normalization and to transform sums of functions with values in $\mathbb{N}$ to $\mathbb{R}$, which is used in the core implementation. Usually we can handle additional parameters without much difficulty, provided they are integers (this was already done in the running example above with m). But a slight shortcoming is that we cannot currently handle general real parameters, because they don't a priori give us polynomials with rational coefficients. However, it would be easy to generalize our approach to allow rational functions based on any finite set of other parameters. There are also a very few cases where our method fails because of a pole in the rational function at 0, but these can always be fixed by reindexing the sum $k \to k+1$ and adding in the zero term separately.

The only really difficult cases to handle are those where in some sense the function being summed does *not* have finite support. The precise definition of finite support is delicate. One might, from the discrete point of view, consider $\binom{n}{k} k!^2$ to have finite support

because $\binom{n}{k}$ does, but it does not have finite support as a limit of the corresponding real functions. Similarly, we cannot handle cases like $\sum_k (-1)^k \binom{4n}{2k} / \binom{2n}{k}$ because the denominator vanishes, even though the numerator does too. On the whole these are cases where one would not normally consider the WZ method applicable, since even informally it clearly assumes finite support. But informally one often applies such methods to problems using subtraction inside the binomial coefficients, such as $\sum_k \binom{n}{k}^2 \binom{3n-k}{2n} / \binom{2n}{n}^2$. Intuitively, one is only 'really' summing over $0 \leqslant k \leqslant n$ because $\binom{n}{k}$ vanishes elsewhere. But as limits this summand is in general problematic because for $k > 3n$ the $\binom{3n-k}{2n}$ features singularities. One could of course use related techniques with a specific range of summation, or make a more careful analysis of the singularities, but either approach loses the attractively 'routine' nature of the procedure. Another approach is to make the change of variable $k \to n - k$, which since $\binom{n}{n-k} = \binom{n}{k}$ gives us an equivalent $\binom{n}{k}^2 \binom{2n+k}{2n} / \binom{2n}{n}^2$ which we *can* handle with our automated setup:

```
|- !n. -- &1 *
       sum (0..n)
       (\k. (&(binom(n,k)) pow 2 * &(binom(2 * n + k,2 * n))) /
            &(binom(2 * n,n)) pow 2) +
       &1 *
       sum (0..n + 1)
       (\k. (&(binom(n + 1,k)) pow 2 *
            &(binom(2 * (n + 1) + k,2 * (n + 1)))) /
           &(binom(2 * (n + 1),n + 1)) pow 2) =
       &0
```

10 Conclusion

On the face of it, the WZ method is simple and transparent, but considerable subtlety lurks inside it. We believe that our setup represents a reasonable interpretation that is provably effective at producing really rigorous formal proofs with minimal human intervention on a wide range of examples. As noted, it is not perfect and perhaps some future refinements would make it even better. In any case, we think this is an excellent illustration of how formalization may not be merely a routine matter of crossing 't's and dotting 'i's, but can provide really interesting insights. I hope that the great pioneer of formalization Andrzej Trybulec would have found this a worthwhile exercise.

Acknowledgments The author would like to thank the anonymous referees for their helpful comments, which have led to a significant improvement in this version over the initial draft.

References

1. Apéry, R.: Irrationalité de $\zeta(2)$ et $\zeta(3)$. Astérisque **61**, 11–13 (1979)
2. Artin, E.: The Gamma Function. Holt, Ronehart and Winton (1964). English translation by Michael Butler
3. Beukers, F.: A note on the irrationality of $\zeta(2)$ and $\zeta(3)$. Bull. Lond. Math. Soc. **11**, 268–272 (1979)
4. Chyzak, F., Mahboubi, A., Sibu-Pinote, T., Tassi, E.: A computer-algebra-based formal proof of the irrationality of $\zeta(3)$. To appear in ITP 2014 (2014)
5. Caprotti, O., Oostdijk, M.: Formal and efficient primality proofs by the use of computer algebra oracles. J. Symb. Comput. **32**, 55–70 (2001)
6. Cohen, H.: Number Theory volume II: Analytic and Modern Tools, volume 240 of Graduate Texts in Mathematics. Springer (2007)
7. Fowler, D.: The binomial coefficient function. Am. Math. Mon. **103**, 1–17 (1996)
8. Gessel, I.M.: Finding identities with the WZ method. J. Symb. Comput. **20**, 537–566 (1995)
9. Gibbs, P.E.: Crackpots who were right II. Prespacetime J. **1**, 489–497 (2010)
10. Gosper, W.R.: Decision procedure for indefinite hypergeometric summation. Proc. Natl. Acad. Sci. U.S.A. **75**, 40–42 (1978)
11. Grégoire, B., Théry, L., Wener, B.: A computational approach to Pocklington certificates in type theory. In: Proceedings of the 8th International Symposium on Functional and Logic Programming, volume 3945 of Lecture Notes in Computer Science, pp. 97–113. Springer-Verlag (2006)
12. Harrison, J., Théry, L.: A sceptic's approach to combining HOL and Maple. J. Autom. Reason. **21**, 279–294 (1998)
13. Lang, S. Complex Analysis. Graduate Texts in Mathematics, 3rd edition. Springer (1993)
14. Nemes, I., Petkovšek, M., Wilf, H., Zeilberger, D.: How to do your monthly problems with your computer. Am. Math. Mon. **104**, 505–519 (1997)
15. Petkovšek, M., Wilf, H.S., Zeilberger, D.: A = B. A K Peters (1996)
16. Carsten Schneider: Apéry's double sum is plain sailing indeed. Electron. J. Comb., 14 (2007)
17. Siddique, U., Hasan, O.: On the formalization of the gamma function in HOL. J. Autom. Reason., 407–429 (2014)
18. Stanley, R.P.: Enumerative Combinatorics, vol. 2. Cambridge University Press (1999)
19. Théry, L., Hanrot, G.: Primality proving with elliptic curves. In: Schneider, K., Brandt, J. (eds.) Proceedings of the 20th International Conference on Theorem Proving in Higher Order Logics, TPHOLs 2007, volume 4732 of Lecture Notes in Computer Science, pp. 319–333. Springer, Kaiserslautern, Germany (2007)
20. van der Poorten, A.J.: A proof that Euler missed: Apéry's proof of the irrationality of $\zeta(3)$, an informal report. Math. Intell. **1**, 195–303 (1979)
21. Wilf, H.S., Zeilberger, D.: Rational functions certify combinatorial identities. J. Am. Math. Soc. **3**, 147–158 (1990)
22. Zeilberger, D.: A holonomic systems approach to special functions identities. J. Comput. Appl. Math. **32**, 321–368