J. Symbolic Computation (1994) **11**, 1–000

Improvement of the Degree Setting in Gosper's Algorithm

PETR LISONĚK[†], PETER PAULE[†] AND VOLKER STREHL[‡]

[†] *Research Institute for Symbolic Computation, J. Kepler University, A–4040 Linz, Austria*
[‡] *IMMD I, University of Erlangen-Nürnberg, D–8520 Erlangen, F. R. Germany*

(Received 2 December 1992)

A detailed study of the degree setting for Gosper's algorithm for indefinite hypergeometric summation is presented. In particular, we discriminate between rational and proper hypergeometric input. As a result, the critical degree bound can be improved in the former case.

1. Introduction

Gosper's algorithm for *indefinite* hypergeometric summation (see Gosper (1978), Lafon (1983) or Graham, Knuth and Patashnik (1989)) belongs to the standard methods implemented in most computer algebra systems. Current interest in this algorithm is mainly due to the fact that it can also be used for *definite* hypergeometric summation (e.g. verifying binomial identities "automatically", finding recurrence operators annihilating hypergeometric sums) in a non-obvious and non-trivial way (see Zeilberger (1990*a*), (1990*b*), Wilf and Zeilberger (1992) and the references given in the latter).

One of the steps in Gosper's algorithm, crucial for its running time and memory requirement, is the determination of a degree bound for a possible polynomial solution of a certain difference equation - the so-called "key equation", see (GE) in section 4. In this paper a detailed analysis of this degree setting is given. It turns out that the situation for rational sequences is different from that for proper, i.e. non-rational hypergeometric input. Besides several theoretical results one practical implication of our discussion is an improvement for the degree setting in Gosper's algorithm in the rational case. At first glance, this improvement might seem to be of minor interest since Gosper's algorithm is not primarily intended for the special case of rational summation. But we have to stress that in many computer algebra systems it is *the only* summation algorithm available. (The only exception from this situation is probably Maple providing a variety of summation algorithms and choosing the appropriate one depending on the particular form of the input.) This motivates a study of the behavior of Gosper's algorithm for different classes of inputs in order to make it *input-sensitive* as a balance to having more algorithms at hand.

After the basic definitions, in sections 2 and 3 algebraic relations between rational and hypergeometric sequences are discussed. Two representations (Gosper form and

0747–7171/90/000000 + 00 $03.00/0 © 1994 Academic Press Limited

Petkovšek's normal form) of rational functions are introduced which are crucial for our investigation. In section 4 a brief outline of Gosper's algorithm is given, including information on the solution space of the key equation (GE). Section 5 presents the careful analysis of the degree setting for polynomial solutions of (GE). The difference between rational and hypergeometric input sequences is made explicit. For example, if an indefinite sum over a regular rational sequence again is rational then there exist at least two polynomial solutions of the key equation with different degrees. The one with the higher degree corresponds to the "K_0-case" in Gosper's original degree setting. This is different from the situation for proper hypergeometric input. Based on the degree setting analysis, a suggestion for a corresponding improvement in Gosper's algorithm is made.

In section 6 two examples illustrating the difference between rational and proper hypergeometric situation are given. One of them is related to the famous Apéry recurrence.

In section 7 we conclude by a brief comment on other methods for rational sequence summation.

2. Rational and Hypergeometric Sequences

Let $\mathbf{N}$ denote the set of nonnegative integers. Let Q be a field of characteristic 0. A sequence $(a_k)_{k\geq 0}$ in Q is called

rational, if there exist relatively prime polynomials $s, t \in Q[x]$ such that

$$a_k = \frac{s(k)}{t(k)} \qquad (k \in \mathbf{N}) \tag{2.1}$$

(in particular: $t(k) \neq 0$ for all $k \in \mathbf{N}$)

hypergeometric, if there exist relatively prime polynomials $\sigma, \tau \in Q[x]$ such that

$$a_k = \frac{\sigma(k)}{\tau(k)} \cdot a_{k-1} \qquad (k \geq 1) \tag{2.2}$$

where $\tau(k) \neq 0$ for all $k \geq 1$.

A rational sequence $(a_k)_{k\geq 0}$ is called *regular rational* iff $\deg(s) < \deg(t)$ in equation (2.1) holds.

Note that once a term a_n of some hypergeometric sequence vanishes, all the subsequent terms a_{n+k} $(k \geq 0)$ will automatically vanish too, i.e. $(a_k)_{k\geq 0}$ has only a finite number of nonzero terms in this case. This degenerate situation is obviously not of much interest as far as *indefinite* hypergeometric summation is concerned. On the other hand, rational sequences can only have a finite number of vanishing terms, hence rational sequences with at least one vanishing term cannot be hypergeometric. Again, since we are interested in indefinite summation, we can always dispense with a finite initial segment of a sequence to be summed by shifting indices.

Hence, for the remainder of this article *rational sequence* will always mean "rational sequence without vanishing terms" and *hypergeometric sequence* will always mean "hypergeometric sequence without vanishing terms".

Under this convention, every rational sequence is a hypergeometric one, since

$$a_k = \frac{s(k)}{t(k)} \cdot \frac{t(k-1)}{s(k-1)} \cdot a_{k-1} = \frac{\sigma(k)}{\tau(k)} \cdot a_{k-1}$$

with

$$\begin{aligned}\sigma(x) &= s(x)\cdot t(x-1)/d(x)\\ \tau(x) &= t(x)\cdot s(x-1)/d(x)\end{aligned}$$

where $d(x) = \gcd(s(x)\cdot t(x-1), t(x)\cdot s(x-1))$. Thus it makes sense to introduce the concept of *proper hypergeometric sequence* which means hypergeometric sequence that is not a rational one.

Conversely, if $(a_k)_{k\geq 0}$ is a hypergeometric sequence as in (2.2) such that the rational function $\sigma(x)/\tau(x)$ can be written as

$$\frac{\sigma(x)}{\tau(x)} = \frac{p_1(x)}{p_1(x-1)}\cdot\frac{p_2(x-1)}{p_2(x)}$$

for (relatively prime, w.l.o.g.) polynomials $p_1, p_2 \in Q[x]$, then $(a_k)_{k\geq 0}$ is rational because

$$a_k = \frac{\prod_{i=1}^{k}\sigma(i)}{\prod_{j=1}^{k}\tau(j)}\cdot a_0 = \frac{p_1(k)\cdot p_2(0)}{p_1(0)\cdot p_2(k)}\cdot a_0$$

i.e. we have (2.1) with $s(x) = a_0\cdot p_2(0)\cdot p_1(x)$ and $t(x) = p_1(0)\cdot p_2(x)$.

We may summarize this discussion in

Proposition 2.1. *Let $(a_k)_{k\geq 0}$ be a hypergeometric sequence with rational function certificate $\lambda(x) = \sigma(x)/\tau(x) \in Q(x)$, i.e.*

$$a_k = \lambda(k)\cdot a_{k-1} \quad \textit{for all } k \geq 1.$$

The sequence $(a_k)_{k\geq 0}$ is rational if and only if there exist polynomials $p_1,\ p_2 \in Q[x]$ such that

$$\lambda(x) = \frac{p_1(x)}{p_1(x-1)}\cdot\frac{p_2(x-1)}{p_2(x)}.$$

3. Gosper- and Petkovšek Representations of Rational Functions

Gosper's algorithm makes essential use of the following fact about rational functions:

Proposition 3.1. (Gosper) *Every nonzero rational function $\lambda(x) \in Q(x)$ can be written as*

$$\lambda(x) = \frac{p(x)}{p(x-1)}\cdot\frac{q(x)}{r(x)}, \tag{G1}$$

where $p, q, r \in Q[x]$ are polynomials such that

$$\gcd(q(x), r(x+j)) = 1 \qquad \textit{for all } j \in \mathbf{N}. \tag{G2}$$

A triple (p, q, r) satisfying (G1) and (G2) will be called a G-form of $\lambda(x)$. Gosper (1978) outlines an algorithm for the computation of a G-form. Note that such a form is not unique. As a simple example: in

$$\lambda(x) = \frac{(x+1)^2}{x} = \frac{x+1}{x}\cdot\frac{x+1}{1} = \frac{(x+1)^2}{x^2}\cdot\frac{x}{1}$$

both the third and the fourth term are G-forms with

$$p(x) = x+1,\ q(x) = x+1,\ r(x) = 1$$

and

$$p(x) = (x+1)^2,\ q(x) = x,\ r(x) = 1,$$

respectively.

It has been shown recently by Petkovšek (1992) that uniqueness for this kind of form can be enforced by imposing two more conditions.

Proposition 3.2. (Petkovšek) *Every nonzero rational function $\lambda(x) \in Q(x)$ can be written uniquely as*

$$\lambda(x) = c \cdot \frac{p(x)}{p(x-1)} \cdot \frac{q(x)}{r(x)}, \tag{P1}$$

where $0 \neq c \in Q$ and where $p, q, r \in Q[x]$ are monic polynomials such that

$$\gcd(q(x), r(x+j)) = 1 \qquad \textit{for all } j \in \mathbf{N}, \tag{P2}$$

$$\gcd(p(x), r(x)) = 1, \tag{P3a}$$

$$\gcd(p(x-1), q(x)) = 1. \tag{P3b}$$

Petkovšek also gives an algorithm for computing what we will call the P-form (p, q, r) of a rational function.

As an immediate simple consequence of Petkovšek's representation we note

Proposition 3.3. *Let $\alpha, \beta \in Q[x]$. If the equation*

$$\beta(x) \cdot y(x) - \alpha(x) \cdot y(x-1) = 0 \tag{3.1}$$

admits a nontrivial polynomial solution $y \in Q[x]$, then all polynomial solutions of (3.1) are precisely given by the scalar multiples $c \cdot y(x),\ c \in Q$, of $y(x)$.

Proof. If $y(x)$ is any monic solution of the equation, then view the r.h.s. of

$$\frac{\alpha(x)}{\beta(x)} = \frac{y(x)}{y(x-1)}$$

as the P-form $(p(x) = y(x), q(x) = r(x) = 1)$ of the l.h.s. By the uniqueness assertion of proposition 3.2 any polynomial solution of (3.1) must be a scalar multiple of $y(x)$. □

As a further consequence of Petkovšek's result we get information about the possible G-forms of rational sequence certificates. (Cf. proposition 2.1 for the notion of the certificate.)

Proposition 3.4. *Let $\rho(x) = \sigma(x)/\tau(x) \in Q(x)$ be a rational function with $\gcd(\sigma(x), \tau(x)) = 1$, and let $(p(x), q(x), r(x))$ be any G-form of $\rho(x)/\rho(x-1)$. Then*

1 if $q(x) = r(x) = 1$, then $\rho(x)$ is a polynomial, i.e. $\tau(x) = 1$;

2 if $p(x) = 1$, then $\rho(x)$ is the reciprocal of a polynomial, i.e. $\sigma(x) = 1$;

3 in the general situation: $\sigma(x) \mid p(x)$.

Proof. 1. For $q = r = 1$

$$\frac{\sigma(x)}{\sigma(x-1)} = \frac{(\tau \cdot p)(x)}{(\tau \cdot p)(x-1)}.$$

Both sides are in P-form, thus $\sigma = \tau \cdot p$, which implies $\tau = 1$ by $\gcd(\sigma, \tau) = 1$.

2. Representing $\tau(x-1)/\tau(x)$ as

$$\frac{\tau(x-1)}{\tau(x)} = \frac{u(x)}{u(x-1)} \cdot \frac{v(x)}{w(x)}$$

implies $v(x) \mid \tau(x-1)$ and $w(x) \mid \tau(x)$ by considering

$$\tau(x-1) \cdot u(x-1) \cdot w(x) = \tau(x) \cdot u(x) \cdot v(x)$$

together with the Petkovšek conditions. But then both sides of

$$\frac{(\sigma \cdot u)(x)}{(\sigma \cdot u)(x-1)} \cdot \frac{v(x)}{w(x)} = \frac{q(x)}{r(x)}$$

are in P-form. E.g., $\gcd(w, \sigma \cdot u) = 1$, the Petkovšek condition (P3a), holds because of $w \mid \tau$, $\gcd(\sigma, \tau) = 1$, and $\gcd(u, w) = 1$. Analogously the other "diagonal" Petkovšek condition (P3b) is verified using $v(x) \mid \tau(x-1)$.
Thus we have $v = q, w = r, \sigma \cdot u = 1$, and thus $u = \sigma = 1$.

3. For $\tilde{\rho} = \sigma/(p \cdot \tau)$ the G-form of $\tilde{\rho}(x)/\tilde{\rho}(x-1)$ is

$$\frac{\tilde{\rho}(x)}{\tilde{\rho}(x-1)} = \frac{q(x)}{r(x)}.$$

It follows from 2. that $\tilde{\rho}(x)$ is the reciprocal of a polynomial. This, together with $\gcd(\sigma, \tau) = 1$, implies $\sigma \mid p$. □

We use this assertion in the following result which is crucial for discussing the behavior of Gosper's algorithm on rational sequences.

Proposition 3.5. *Let $\lambda(x) \in Q(x)$ be a rational function, and let (p, q, r) be any G-form of $\lambda(x)$. Then the following assertions are equivalent:*

1 We have

$$\lambda(x) = \frac{\rho(x)}{\rho(x-1)}$$

for some rational function $\rho(x) \in Q(x)$.

2 The equation

$$q(x+1)\ y(x) - r(x)\ y(x-1) = 0$$

admits a nontrivial polynomial solution $y \in Q[x]$.

Proof. Let $\rho(x) = \sigma(x)/\tau(x) \in Q(x)$ with $\gcd(\sigma(x), \tau(x)) = 1$. Then $\sigma(x) \mid p(x)$ by the general part of the previous proposition. We may thus rewrite the G-form of

$$\lambda(x) = \frac{\rho(x)}{\rho(x-1)} = \frac{\sigma(x)}{\sigma(x-1)} \cdot \frac{\tau(x-1)}{\tau(x)}$$

as

$$\frac{\tau(x-1)}{\tau(x)} = \frac{(p/\sigma)(x)}{(p/\sigma)(x-1)} \cdot \frac{q(x)}{r(x)}$$

or

$$q(x) \cdot \left(\frac{p \cdot \tau}{\sigma}\right)(x) - r(x) \cdot \left(\frac{p \cdot \tau}{\sigma}\right)(x-1) = 0. \tag{3.2}$$

Now $\gcd(q(x), r(x)) = 1$ by property (G2), hence $q(x) \mid (p \cdot \tau/\sigma)(x-1)$, i.e.

$$\left(\frac{p \cdot \tau}{\sigma}\right)\sigma(x) = q(x+1) \cdot y(x)$$

for some nonzero polynomial $y \in Q[x]$. Dividing both sides of (3.2) by $q(x)$ then gives

$$q(x+1)\ y(x) - r(x)\ y(x-1) = 0.$$

For the other direction, let $y(x)$ be a nontrivial solution of this previous equation, then

$$\frac{q(x)}{r(x)} = \frac{q(x)}{q(x+1)} \cdot \frac{y(x-1)}{y(x)},$$

and

$$\lambda(x) = \frac{p(x)}{p(x-1)} \cdot \frac{q(x)}{q(x+1)} \cdot \frac{y(x-1)}{y(x)},$$

i.e. we have $\lambda(x) = \rho(x)/\rho(x-1)$ with

$$\rho(x) = \frac{p(x)}{q(x+1) \cdot y(x)}.$$

□

4. Gosper's Algorithm: Uniqueness of Solutions

The essence of Gosper's algorithm (see Gosper (1978) or Graham, Knuth and Patashnik (1989)) can be shortly described as follows:

Given a hypergeometric sequence $(a_k)_{k\geq 0}$ with values from the field Q. Let us assume that the sequence $(s_n)_{n\geq 0}$ defined as

$$s_n = \sum_{k=0}^{n} a_k,$$

for all nonnegative integers n, again is hypergeometric. Then to solve the summation problem is equivalent to find the hypergeometric solution $(s_k)_{k\geq 0}$ of the difference equation

$$s_k - s_{k-1} = a_k \qquad k \geq 1 \tag{DE}$$

with the initial condition $s_0 = a_0$. If exists, this solution can be expressed as

$$s_n = \frac{q(n+1)}{p(n)}\ f(n)\ a_n,$$

where $f(x)$ is a polynomial satisfying the *key equation*

$$p(x) = q(x+1)\ f(x) - r(x)\ f(x-1) \tag{GE}$$

and where (p, q, r) is a G-form of the rational function certificate determined by a_k/a_{k-1} $(k \geq 1)$. In order to discuss the set of all possible polynomial solutions $f \in Q[x]$ of the key equation (GE) we make use of the following proposition which is evident:

PROPOSITION 4.1. *Given polynomials $\alpha, \beta, \gamma \in Q[x]$ with $\gamma \neq 0$, then the set of all polynomial solutions of*

$$\gamma(x) = \alpha(x)\ y(x) - \beta(x)\ y(x-1) \tag{4.1}$$

consists precisely of all expressions of the form

$$y\ +\ z,$$

where $y \in Q[x]$ is a solution of (4.1) and $z \in Q[x]$ runs through all polynomial solutions of the homogeneous equation

$$0 = \alpha(x)\ z(x) - \beta(x)\ z(x-1). \tag{4.2}$$

Just before finishing our paper, we learned that a similar statement is proven as Lemma 3.7 in Koornwinder (1992). However, no further investigations appear there. On the contrary, here we proceed by showing that there is an intimate connection between the situation described in the previous proposition and the two principal classes of input sequences:

Let us assume that a polynomial solution $f \in Q[x]$ of (GE) exists. Then, by proposition 4.1 above we have to consider two different cases, *(A)* and *(B)*, induced by the structure of the corresponding homogeneous equation

$$0 = q(x+1)\ y(x) - r(x)\ y(x-1). \tag{4.3}$$

(A) If (4.3) admits no nontrivial solution, then $f(x)$ is the only solution of the key equation (GE).

(B) If there exists a nontrivial solution h of (4.3), then due to propositions 4.1 and 3.3 the polynomial solution set of the key equation (GE) consists precisely of all polynomials of the form

$$f(x) + c \cdot h(x),$$

where c is running through all the elements of Q.

We show that the cases *(A)* and *(B)* correspond to $(a_k)_{k\geq 0}$ being either a *proper* hypergeometric sequence (i.e. not a rational one), or being a rational sequence:

PROPOSITION 4.2. *Let $(a_k)_{k\geq 0}$ be a hypergeometric sequence with rational function certificate $\lambda(x) \in Q(x)$, i.e. $a_k = \lambda(k) \cdot a_{k-1}$ for all $k \geq 1$, and let (p,q,r) be a G-form of λ. Then the key equation*

$$p(x) = q(x+1)\ f(x) - r(x)\ f(x-1)$$

arising in Gosper's algorithm admits

1 *at most one polynomial solution, if $(a_k)_{k\geq 0}$ is a proper hypergeometric sequence;*

2 *none or a one-parameter family of polynomial solutions, if $(a_k)_{k\geq 0}$ is a rational sequence.*

PROOF. By proposition 3.5 the homogeneous form of the key equation

$$0 = q(x+1)\ f(x) - r(x)\ f(x-1)$$

admits a nontrivial solution if and only if $\lambda(x) = \rho(x)/\rho(x-1)$ for some rational function

$\rho \in Q(x)$. By proposition 2.1 this representation of λ is possible if and only if $(a_k)_{k\geq 0}$ is rational. The rest of the proposition is implied by the analysis of the cases *(A)* and *(B)* above. □

We conclude this section by a proposition describing how, in the rational input case, a polynomial solution of the homogeneous form of the key equation (GE) can be computed from the corresponding G-form. The special form of this solution implies a degree relation which turns out to be fundamental for the analysis of the degree setting (see subsection 5.2.2).

We define the *degree of a rational function* $F(x) = f_1(x)/f_2(x)$ as $\mathrm{Deg}(F(x)) := \deg(f_1(x)) - \deg(f_2(x))$.

PROPOSITION 4.3. *Let* $(F(k))_{k\geq 0}$ *be a rational sequence where* $F(x) = f_1(x)/f_2(x)$ *with* $f_1, f_2 \in Q[x]$ *and* $\gcd(f_1, f_2) = 1$, *and let* (p, q, r) *be a G-form of* $F(x)/F(x-1)$. *Suppose that the key equation (GE) admits a one-parameter family of polynomial solutions (cf. 4.2.2). Then the following holds:*

1 We have that $P(x) := p(x)/f_1(x)$ *and* $z(x) := f_2(x)P(x)/q(x+1)$ *are polynomials in* $Q[x]$.
2 The polynomial $z \in Q[x]$ *is a solution of the homogeneous form of the key equation (GE), i.e.*

$$0 = q(x+1)\ z(x) - r(x)\ z(x-1) \qquad (4.4)$$

holds.
3 For $\mathrm{Deg}(F(x)) = \deg(f_1(x)) - \deg(f_2(x))$ *we have*

$$\deg(p(x)) - \deg(q(x)) - \mathrm{Deg}(F(x)) = \deg(z(x)).$$

PROOF. By proposition 3.4.3 we know that $f_1(x) \,|\, p(x)$, hence from the G-form representation

$$\frac{f_2(x-1)}{f_2(x)} = \frac{P(x)}{P(x-1)} \cdot \frac{q(x)}{r(x)}.$$

¿From proposition 3.5 we know there must exist a non-trivial solution of (4.4). Suppose $Z \in Q[x]$ is such a solution. Then by rewriting (4.4) as

$$\frac{q(x)}{r(x)} = \frac{Z(x-1)}{Z(x)} \cdot \frac{q(x)}{q(x+1)}$$

and corresponding replacement of $q(x)/r(x)$ in the equation above, after some rearrangements we obtain

$$\frac{f_2(x)P(x)}{Z(x)q(x+1)} = \frac{f_2(x-1)P(x-1)}{Z(x-1)q(x)}.$$

This equation implies that for some non-zero constant $c \in Q$

$$f_2(x)\ P(x) = c\ Z(x)\ q(x+1).$$

Consequently $z(x) := f_2(x)P(x)/q(x+1)$ must be a polynomial and a solution of (4.4), too.

The assertion on degrees follows immediately from 1. □

Later (in proposition 5.1) we shall see that the critical value of K_0 is just the degree of the non-zero homogeneous solution $z \in Q[x]$.

5. Gosper's Algorithm: The Degree Setting

We resume the discussion of Gosper's algorithm at the point where a G-form has been computed. Then the remaining task in Gosper's algorithm is to solve the key equation (GE).

To be specific, let $(a_k)_{k\geq 0}$ be a hypergeometric sequence with rational function certificate $\lambda(x) \in Q(x)$, i.e. $a_k = \lambda(k) \cdot a_{k-1}$ for all $k \geq 1$, and let (p, q, r) be a G-form of λ. One possibility to compute a polynomial solution $f(x)$ of Gosper's key equation (GE) is by coefficient comparison. This can be carried out algorithmically once an upper bound K for the degree of $f(x)$ is known. As Gosper (1978) showed, K can be derived from an analysis of the following equation, which is equivalent to (GE):

$$p(x) = (q(x+1) - r(x))\frac{f(x) + f(x-1)}{2} + (q(x+1) + r(x))\frac{f(x) - f(x-1)}{2}.$$

The following two cases may arise:

Case 1: If $\deg(q(x+1) + r(x)) \leq \deg(q(x+1) - r(x)) =: M$, then K is uniquely determined as $K := \deg(p) - M$.

Case 2: $\deg(q(x+1) - r(x)) < \deg(q(x+1) + r(x)) =: m$. This case appears exactly if $\deg(q) = \deg(r)$ and, moreover, the leading coefficients of q and r are equal. Thus by the Gosper-type representation (G1) we may assume that these leading coefficients are equal to 1. Let $f(x) = f_K x^K + O(x^{K-1})$, $f_K \in Q \setminus \{0\}$, be a polynomial solution f of (GE). Then the rest of the degree analysis can be read off the observation that

$$p(x) = f_K \; L(K) \; x^{K+m-1} + O(x^{K+m-2}), \tag{5.1}$$

with $L(K)$ being a linear polynomial of the form $L(K) = K - K_0$, where K_0, the root of $L(K)$, is determined as the coefficient of x^{m-1} in $r(x) - q(x+1)$, in usual notation

$$K_0 := \langle \, x^{m-1} \, \rangle \, (r(x) - q(x+1)). \tag{5.2}$$

According to the degree comparison of both sides of (5.1) the set of polynomial solutions f of (GE) splits into two classes: those solutions f with $\deg(f) = K_0$, which is just possible for K_0 being an integer greater than $\deg(p) - m + 1$, and those solutions f with $\deg(f) \neq K_0$, which corresponds to $K := \deg(p) - m + 1$. Recalling that $m = \deg(q)$, one has

Case 2a: if K_0 is not an integer, then K is uniquely determined as $K := \deg(p) - \deg(q) + 1$,

Case 2b: if K_0 is an integer, take $K := \max(K_0, \deg(p) - \deg(q) + 1)$.

It may happen that K is determined to be a negative integer. This means that no hypergeometric sequence $(s_n)_{n\geq 0}$ solving the difference equation (DE) exists and Gosper's algorithm terminates.

5.1. K_0-cases

In his survey on indefinite summation algorithms, Lafon (1983) writes about Gosper's algorithm: "We have never observed that (the degree) ... was set to K_0; here some improvements may be possible."

This remark is a bit confusing. Actually such a "K_0-example" is provided by Lafon himself on the same page (Lafon, 1983, p. 75): For the (regular rational) input $a_n = 1/(n(n+2))$, $K := K_0$ $(= 2)$ is set by Gosper's algorithm as the degree for the polynomial $f(n)$.

Moreover, there are prominent *proper hypergeometric* sequences in which exactly the K_0-setting yields a solution. One of such sequences arises from the famous Apéry recurrence, see section 6.3.

5.2. rational sequence summation

Suppose we run Gosper's algorithm on the rational sequence input $(F(n))_{n\geq 0}$ of the form $F(x) = f_1(x)/f_2(x)$, f_1, $f_2 \in Q[x]$.

Let (5.3) be Gosper's representation of the quotient $F(x)/F(x-1)$:

$$\frac{f_1(x)f_2(x-1)}{f_2(x)f_1(x-1)} = \frac{p(x)}{p(x-1)} \cdot \frac{q(x)}{r(x)}. \tag{5.3}$$

¿From (5.3) we have

$$f_1(x)f_2(x-1)p(x-1)r(x) = f_2(x)f_1(x-1)p(x)q(x). \tag{5.4}$$

We see that $\deg(q(x)) = \deg(r(x)) =: m$ and $\langle\, x^m \,\rangle\, q(x) = \langle\, x^m \,\rangle\, r(x)$. Thus Case 2 of Gosper's degree analysis applies. We have to look for the value of K_0:

Proposition 5.1. *For each nonzero rational function $F(x)$ we have*

$$K_0 = \deg(p(x)) - \deg(q(x)) - \mathrm{Deg}(F(x)), \tag{5.5}$$

where K_0 is the value computed by Gosper's algorithm in the Case 2 and $p(x)$, $q(x)$, $r(x)$ are the polynomials arising in Gosper's representation (5.3).

Proof. Denote $f_1(x) = \sum_{i=0}^{s} a_i x^i$, $f_2(x) = \sum_{i=0}^{t} b_i x^i$, $p(x) = \sum_{i=0}^{d} p_i x^i$, $q(x) = \sum_{i=0}^{m} q_i x^i$, $r(x) = \sum_{i=0}^{m} r_i x^i$ with $a_s\, b_t\, p_d\, q_m\, r_m \neq 0$. Note that $q_m = r_m$.

By coefficients comparison at $x^{s+t+d+m-1}$ in (5.4) we obtain (remember that $q_m = r_m$)

$$a_{s-1}b_tp_dr_m + a_s(-tb_t + b_{t-1})p_dr_m + a_sb_t(-dp_d + p_{d-1})r_m + a_sb_tp_dr_{m-1} =$$
$$b_{t-1}a_sp_dr_m + b_t(-sa_s + a_{s-1})p_dr_m + b_ta_sp_{d-1}r_m + b_ta_sp_dq_{m-1},$$

hence

$$(s-t-d)a_s\, b_t\, p_d\, r_m = a_s\, b_t\, p_d\, (q_{m-1} - r_{m-1})$$

and

$$(-d+m-t+s)r_m = mr_m + q_{m-1} - r_{m-1}.$$

Thus

$$d - m + t - s = -\frac{mr_m + q_{m-1} - r_{m-1}}{r_m},$$

which together with $q_m = r_m$ yields

$$\deg(p(x)) - \deg(q(x)) - \mathrm{Deg}(F(x)) = -2\frac{mq_m + q_{m-1} - r_{m-1}}{q_m + r_m}.$$

On the r.h.s. of the last equation we have the value

$$-2\frac{\langle\, x^{m-1}\,\rangle\,(q(x+1)-r(x))}{\langle\, x^{m}\,\rangle\,(q(x)+r(x))}. \tag{5.6}$$

W.l.o.g. we can assume q and r to be monic. Then (5.6) is precisely equal to the value K_0 (cf. (5.2)) in Gosper's degree analysis. Thus we have proved that (5.5) holds for each rational input sequence. □

5.2.1. DESCRIPTION OF THE K_0-CASE IN RATIONAL SUMMATION

With respect to our result, the "$K \leftarrow K_0$"-case in rational summation occurs if and only if

$$\deg(p(x)) - \deg(q(x)) - \operatorname{Deg}(F(x)) \geq \deg(p(x)) - \deg(q(x)) + 1 \tag{5.7}$$

iff

$$\operatorname{Deg}(F(x)) \leq -1$$

iff the summation input $(F(k))_{k\geq 0}$ is a regular rational sequence.

If this is the case, the solution to the summation problem is given by

$$R(x) = \frac{q(x+1)f(x)}{p(x)}F(x)$$

with $\deg(f(x)) := K_0 = \deg(p(x)) - \deg(q(x)) - \operatorname{Deg}(F(x))$. (It follows from Gosper's precise analysis that this degree bound is accurate.) Thus

$$\begin{aligned}\operatorname{Deg}(R(x)) \quad = \quad & \deg(q(x)) + \big(\deg(p(x)) - \deg(q(x)) - \operatorname{Deg}(F(x))\big) \\ & - \deg(p(x)) + \operatorname{Deg}(F(x)).\end{aligned}$$

Hence, $\operatorname{Deg}(R(x)) = 0$.

5.2.2. A BETTER DEGREE SETTING FOR REGULAR RATIONAL INPUT

We learned that for regular rational function inputs $F(x)$, the solution function $R(x)$ of (DE) computed by Gosper's algorithm arises from the K_0-case and, moreover, $\operatorname{Deg}(R(x)) = 0$ holds. Let $R(x) = r_1(x)/r_2(x)$, $\deg(r_1(x)) = \deg(r_2(x))$ and thus $r_1(x)/r_2(x) = c + r_3(x)/r_2(x)$, $c \in Q \setminus \{0\}$ with $r_3 = 0$ or $\deg(r_3(x)) < \deg(r_2(x))$. ¿From this we get another solution of the difference equation (DE), namely

$$\frac{r_3(x)}{r_2(x)} - \frac{r_3(x-1)}{r_2(x-1)} = F(x). \tag{5.8}$$

Since $\operatorname{Deg}(r_3(x)/r_2(x)) < 0$, we see that this solution cannot correspond to the K_0-case. Moreover, (5.8) implies $\operatorname{Deg}(r_3(x)/r_2(x)) = \operatorname{Deg}(F(x)) + 1$ for the regular rational solution of (DE) from which we calculate the degree of the respective polynomial $f(x)$ to be

$$\operatorname{Deg}(F(x)) + 1 - \operatorname{Deg}(F(x)) + \deg(p(x)) - \deg(q(x))$$

which is the second alternative of Case 2b of Gosper's algorithm.

For the practical applications we note that the degree of $f(x)$ (and so the order of the linear system for coefficients of $f(x)$) decreases by the same value as the rational function degree of the resulting sum does, i.e. by $-\operatorname{Deg}(F(x)) - 1$.

5.2.3. NON-REGULAR RATIONAL INPUT

The last class of inputs that was not treated yet is the set of non-regular rational sequences. We show that no improvement of degree setting is possible here:

¿From proposition 4.3.3 we have that

$$\mathrm{Deg}(F(x)) \geq 0 \qquad \Longleftrightarrow \qquad \deg(z(x)) < \deg(p(x)) - \deg(q(x)) + 1$$

for any non-zero solution $z(x) \in Q[x]$ of the homogeneous form of the key equation (GE). Due to Gosper's degree analysis this also implies

$$\deg(z(x)) < \deg(y(x))$$

where $y \in Q[x]$ solves (GE). This means that in the case of the non-regular rational input, all solutions of (GE) are of the same degree. In particular, we see that this degree is $\deg(p(x)) - \deg(q(x)) + 1$ since proposition 4.3.3 and proposition 5.1 yield together

$$K_0 = \deg(z(x))$$

and so K_0 is less than $\deg(p(x)) - \deg(q(x)) + 1$ here.

5.3. "PLAIN" AND "HIDDEN" RATIONAL SEQUENCES

We should be aware of the fact that the input sequence actually might be a rational one but in a disguised form. For example,

$$a_k = k!/(k+6)! \tag{5.9}$$

is a rational sequence.

Such cases are recognized easily when processed by humans but need more care when we implement summation in a computer algebra system. Success with the rationality test allows us to reduce computation time by taking the better degree setting instead of the maximum in Case 2b.

Here we meet central issues of symbolic computation - simplification and canonical forms.

However, even if we do not simplify the input completely, there is a guideline that can help us:

PROPOSITION 5.2. *Let m be the value computed in Case 2 of Gosper's algorithm (cf. section 5). If we get into Case 2b with $m = 1$ (q and r are linear polynomials), then the input sequence is rational.*

PROOF. We can make q and r monic. Suppose $q(x) = x + q_0$, $r(x) = x + r_0$. Then $K_0 = r_0 - q_0 - 1$ must be a nonnegative integer. Denote the summation input by $(a_k)_{k \geq 0}$. Then the G-representation is

$$\frac{a_k}{a_{k-1}} = \frac{p(k)}{p(k-1)} \cdot \frac{k+q_0}{k+r_0}$$

for some $p(x) \in Q[x]$. Now the result follows directly from proposition 2.1 applied with $p_1(x) := p(x)$ and $p_2(x) := \prod_{i=q_0+1}^{r_0}(x+i)$. □

Based on our results given up to now, *we suggest the following improvement of the degree setting in Gosper's algorithm by regrouping the two subcases of Case 2:*

```
If $K_0$ is not an integer or the input sequence is rational or $m = 1$
    then $K := \deg(p(n)) - m + 1$                {Case 2a}
    else  $K := \max(K_0\,,\ \deg(p(n)) - m + 1)$   {Case 2b}
```

6. Other K_0-Examples

We have shown how to improve the degree reasoning for rational inputs. In this section we present some proper hypergeometric K_0-cases documenting that no general improvement can be made here.

6.1. A "SIMPLE" PROPER HYPERGEOMETRIC K_0-CASE

It is hard to find a "nice" example with binomials or even with (integer) factorials and not to fall into the rational case at the same time. This is the reason why we use somewhat cumbersome fractions and raising factorials here:

Let $x^{\overline{n}} = x(x+1)\dots(x+n-1)$ be the raising factorial and let $\tilde{p}(n) = \frac{1}{36}(-35n^2 - 20n + 65)$. We want to sum

$$a_n = \tilde{p}(n)\frac{{(-5/2)^{\overline{n+1}}}^2}{(-1/3)^{\overline{n+1}}(-2/3)^{\overline{n+1}}}.$$

The Gosper's representation here is $p(n) = \tilde{p}(n)$, $q(n) = (n-5/2)^2$, $r(n) = (n-1/3)(n-2/3)$, so $\deg(q(n)) = \deg(r(n))$ and leading coefficient of q is equal to the leading coefficient of r. We have that $m = 2$ and the degree bounds for polynomial $f(n)$ are

$$\deg(p) - m + 1 = 2 - 2 + 1 = 1$$

and

$$K_0 = 2.$$

Thus we are in the K_0-case with a balanced linear system for unknown coefficients c_2, c_1, c_0 of polynomial $f(n)$. The system has exactly one solution because its determinant is different from zero (namely 1225/46656). The solution is $(c_2, c_1, c_0) = (1, 0, 1)$, thus we are in the proper hypergeometric K_0-case with $f(n) = n^2 + 1$. The sum is

$$s_n = (n - 3/2)^2(n^2 + 1)\frac{{(-5/2)^{\overline{n+1}}}^2}{(-1/3)^{\overline{n+1}}(-2/3)^{\overline{n+1}}}.$$

Remark: We note that for $m = 1$ and $K = K_0$, the linear system for coefficients of $f(x) = c_{K_0}x^{K_0} + \ldots + c_0$ is underdetermined. It has got K_0 equations and $K_0 + 1$ unknowns. This fact just supports the claim of proposition 4.2.2.

Generally, in the "$K = K_0$" case the linear system for coefficients of $f(x)$ arises from coefficient comparisons at $\deg(q(x+1)) + \deg(f(x)) + 1 - 1 = K_0 + m$ different powers of x in (GE). (The $+1$ counts the absolute term whereas -1 discounts the vanishing leading term, cf. Gosper's degree analysis!). Hence, the system has $K_0 + m$ equations and $K_0 + 1$ unknowns. Thus it is the value of m that influences whether the system is underdetermined, balanced or overdetermined ($m = 1$, $m = 2$, $m > 2$). The value $m = 1$ means rational input, hence the value $m = 2$ from previous example is minimal for presentation of a *proper hypergeometric K_0-case*.

6.2. THE APÉRY RECURRENCE

Finally we present a “nice” proper hypergeometric K_0-example. The value of m is equal to 4 here.

For nonnegative integers n, k let

$$F_{n,k} = \binom{n}{k}^2 \binom{n+k}{k}^2. \tag{6.1}$$

Let us recall the famous Apéry recurrence

$$\forall n \in \mathbf{N} \quad c_0(n)\ S_n + c_1(n)\ S_{n+1} + c_2(n)\ S_{n+2} = 0, \tag{6.2}$$

where

$$\begin{aligned} &c_0(n) = (n+1)^3, \quad c_1(n) = -(2n+3)(17n^2+51n+39), \\ &c_2(n) = (n+2)^3, \end{aligned} \tag{6.3}$$

and

$$S_n = \sum_{k=0}^{n} F_{n,k}. \tag{6.4}$$

Remark: For an excellent account on how this recurrence is used to prove the irrationality of $\zeta(3)$ see (van der Poorten, 1979).

Note that the double-indexed sequence $(F_{n,k})_{n\geq 0,k\geq 0}$ is hypergeometric in both variables. Under slight side-conditions (see e.g. Zeilberger (1990*a*), (1990*b*) or Wilf and Zeilberger (1992)) for such sequences there exist a nonnegative integer d, polynomials $c_0(n), \ldots, c_d(n)$ being independent of k, and a double-indexed sequence $(G_{n,k})_{n\geq 0,k\geq 0}$, again hypergeometric in both variables, such that

$$\begin{aligned} c_0(n)\ F_{n,k} + c_1(n)\ F_{n+1,k} + \ldots + c_d(n)\ F_{n+d,k} \\ = G_{n,k} - G_{n,k-1}. \end{aligned} \tag{6.5}$$

Due to the fact that the left-hand-side of the equation above can be rewritten as $F_{n,k}$ times a rational function in the two variables n, k, i.e. the resulting expression is hypergeometric in k (actually it is hypergeometric in both variables), it is possible to compute $G_{n,k}$ and the coefficient polynomials $c_i(n)$ by executing Gosper's algorithm once the order d is known.

Now running this procedure in the Apéry situation, i.e. with choosing $F_{n,k}$ as defined in (6.1) and setting $d = 2$, produces exactly the situation of Case 2b described above. In the following we give the details of that computation.

The left-hand-side of equation (6.5) can be rewritten as the following rational function multiple of $F_{n,k}$:

$$\frac{p_0(n,k)\ c_0(n) + p_1(n,k)\ c_1(n) + p_2(n,k)\ c_2(n)}{(n-k+1)^2(n-k+2)^2}\ F_{n,k} = a_k, \tag{6.6}$$

where

$$\begin{aligned} p_0(n,k) &= (n-k+2)^2(n-k+1)^2, \\ p_1(n,k) &= (n-k+2)^2(n+k+1)^2, \\ p_2(n,k) &= (n+k+2)^2(n+k+1)^2. \end{aligned}$$

The polynomials corresponding to a G-form of the quotient a_k/a_{k-1} are computed as

$$\begin{aligned} p(x) &= c_0(n)p_0(n,x) + c_1(n)p_1(n,x) + c_2(n)p_2(n,x) \\ q(x) &= (n+x)^2(x-n-3)^2, \\ r(x) &= x^4. \end{aligned}$$

In addition, we find that

$$\deg(p(x)) = 4, \quad \deg(q(x+1) - r(x)) = 3, \text{ and } \deg(q(x+1) + r(x)) = 4.$$

One can observe that

$$K_0 = 2 \text{ and } \deg(p(x)) - m + 1 = 1.$$

Thus we are in Case 2b, where now the degree setting K for $f(x)$ has to be set to K_0, i.e. $K := 2$.

Following that pattern, i.e. that of the polynomials p, q, r, it is easy to construct further examples where exactly the same instance of Case 2b occurs.

Remark: For the sake of completeness we want to remark that running through Gosper's algorithm one gets for the coefficient polynomials $c_i(n)$, $i = 0, 1, 2$, the same values (6.3) as in the Apéry recurrence, $f(x) = 4(2n+3)(2x^2 + x - (2n+3)^2)$ and thus

$$G_{n,k} = \frac{q(k+1)}{p(k)} f(k) a_k = \frac{(n+k+1)^2}{(n-k+1)^2} f(k) F_{n,k}.$$

With these substitutions Apéry's recurrence (6.2) follows from (6.5) by "telescoping", i.e. summation w.r.t. k. (If n is fixed then $G_{n,k}$ as a function in k has finite support, as it is a rational function multiple of $F_{n,k}$.)

7. Conclusion

To conclude this article we briefly comment on other methods for rational function summation.

The (probably) first method for rational sequence summation was designed by Abramow (1971). Nowadays it can be viewed as a special version of Gosper's algorithm adjusted for rational sequences. Abramow solves an equation which is similar to Gosper's equation (GE), however, he considers only Case 2a in degree setting since, in his approach, it leads to the solution if there is any. It follows from (5.7) that Case 2a delivers a better setting than Case 2b if and only if the degree of the numerator of the input sequence is less than the degree of the denominator. This can be always done by putting the polynomial part of the input aside. There are considerably easier methods for summing polynomials. (E. g., transformation into the falling factorial base.)

On the other hand, Gosper stuck to the higher degree setting because he wanted to ensure that no solution is lost. Sometimes we pay for this comfort by unnecessary computations.

As far as we know, neither of them considered or discussed the approach of the other one.

Summation analogs of Hermite integration of rational functions have been provided by Abramow (1975) and Moenck (1977). Since both methods are iterative and based on gcd-computations, they cannot be compared to the two mentioned above.

Recently Paule (1992), in an effort to close gaps in Moenck's work, introduced the concept of greatest-factorial factorization. In that paper a new approach to rational sequence summation is given including a summation analog of Horowitz's method for rational function integration.

For a detailed comparison of the last three approaches mentioned see Pirastu (1992).

The first named author was supported by the Austrian Federal Ministry of Science and Research. The work of the second author was partially sponsored by Austrian FWF P7220 and by the Austrian Federal Ministry of Science and Research in the frame of ESPRIT basic research action 3125 "MEDLAR".

References

Abramow, S.A. (1971). On the summation of rational functions. *Zh. vycisl. matem. i matem. fiz.* **11**, 1071–1075. (In Russian.)

Abramow, S.A. (1975). The rational component of the solution of a first-order linear recurrence relation with a rational right-hand side. *Zh. vycisl. matem. i matem. fiz.* **15**, 1035–1039. (In Russian.)

Gosper, R.W. (1978). Decision procedures for indefinite hypergeometric summation. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 40–42.

Graham, R.L., Knuth, D.E., Patashnik, O. (1989). *Concrete Mathematics, A Foundation for Computer Science*. Reading: Addison-Wesley.

Koornwinder, T.H. (1992). On Zeilberger's Algorithm and Its q-analogue: A Rigorous Description. Report AM-R9207, CWI, Amsterdam.

Lafon, J.C. (1983). Summation in finite terms. In *Computer Algebra and Symbolic Computation, 2nd ed.* 71–77. Wien: Springer-Verlag.

Lisoněk, P. (1991). The performance of Gosper's algorithm on rational function inputs. Technical Report, RISC-Linz Series 91-31.0

Moenck, R. (1977). On computing closed forms for summations. In *Proceedings of MACSYMA users' conference*. Berkeley.

Paule, P. (1992). Greatest-Factorial Factorization and Symbolic Summation I. *Submitted to J. Symb. Comp.*

Petkovšek, M. (1992). Hypergeometric solutions of linear recurrences with polynomial coefficients. *J. Symb. Comp.* **14**, 243–264.

Pirastu, R. (1992). *Algorithmen zur Summation rationaler Funktionen*. Diploma Thesis, Univ. Erlangen-Nürnberg. (In German.)

van der Poorten, A. (1979). A Proof that Euler Missed ... Apéry's Proof of the Irrationality of $\zeta(3)$. *Math. Intell.* **1**, 195–203.

Wilf, H.S., Zeilberger, D. (1992). An algorithmic proof theory for hypergeometric (ordinary and "q") multisum/integral identities. *Invent. Math.* **108**, 575–633.

Zeilberger, D. (1990*a*). A holonomic systems approach to special function identities. *J. Comp. Appl. Math.* **32**, 321–368.

Zeilberger, D. (1990*b*). A fast algorithm for proving terminating hypergeometric identities. *Discr. Math.* **80**, 207–211.