A Note On the Minimality Problem in Indefinite Summation of Rational Functions

Roberto Pirastu[*]
Research Institute for Symbolic Computation, J. Kepler University, A-4040 Linz

1 Introduction

Consider a field $\mathcal{K}$ of characteristic zero. The shift operator E on $\mathcal{K}(x)$ is defined by $Ef(x) := f(x+1)$ for all $f \in \mathcal{K}(x)$. The (forward) difference operator is defined by $\Delta := E - \mathbf{1}$, where $\mathbf{1}$ operates identically. Then the problem of indefinite summation of rational functions can be stated as follows.

Problem: *Given a proper rational function $f \in \mathcal{K}(x)$, find $h, r \in \mathcal{K}(x)$ such that*

$$f = \Delta h + r \tag{1}$$

and the denominator of r has minimal degree among all such decompositions. We call the pair (h, r) a solution of the indefinite summation problem for f and call r a **bound** *for f.*

Each solution (h, r) in (1) corresponds to the following decomposition

$$g(a,b) := \sum_{k=a}^{b} f(k) = h(b+1) - h(a) + \sum_{k=a}^{b} r(k) \tag{2}$$

In particular, a solution of (1) with remainder $r = 0$ provides a *closed form* for g as a rational expression in a and b, namely $g(a,b) = h(b+1) - h(a)$.

Several algorithms are known for computing solutions of the indefinite rational summation problem (see [2, 1, 3]). Note that giving an "Ansatz" for the denominators of h and r reduces equation (1) to a polynomial equation, which can be solved by coefficient comparison.

Here we describe such candidates for the denominators of h and r which depend only on the denominator of f. For a certain class of rational summands f the given estimates are precise.

Solutions (h, r) are not uniquely determined by f. For instance, the decompositions

$$\Delta\left(-\frac{1}{4}\frac{2x+1}{x(x+1)}\right) - \frac{1}{2(x^2+4x+4)} \qquad \text{and} \qquad \Delta\left(-\frac{1}{4}\frac{2x^3+7x^2+5x+2}{x^2(x+1)^2}\right) - \frac{1}{2x^2}$$

are different solutions for the same rational summand $f = \frac{1}{x(x+2)^2}$. As one can see, the degree of the denominator of h can vary considerably among solutions. For this reason we are interested

[*] Supported by the Commission of the European Communities in the frame of the program "Human Capital and Mobility", contract Nr. ERBCHBICT930501.

in solutions (h, r) with both h and r of minimal degree in the denominator. Our observations prove that the modification of Abramow's algorithm proposed in [4, 5] produces such *minimal* solutions for a certain class of rational summands.

In the following we make use of known results due to Abramow and Paule (see [1, 3]), to which we refer for an extensive treatment of the subject.

We say that two polynomials p_1 and p_2 are *shift equivalent* if $p_1 = E^k p_2$ for some integer k. In this case we write $p_1 \sim p_2$. Consider, for a polynomial $p \in \mathcal{K}[x]$, the complete factorization (over $\mathcal{K}$) $p = \alpha p_1^{e_1} p_2^{e_2} \cdots p_m^{e_m}$, where all p_i are irreducible and monic and $e_1 \cdots e_m \neq 0$, $\alpha \in \mathcal{K}$. Then an equivalence class of the set of irreducible factors $\{p_1, \ldots, p_m\}$ of p under the relation $\sim$ is called a *shift class* of p. A factor $\tilde{p} = p_{i_1}^{e_{i_1}} \cdots p_{i_l}^{e_{i_l}}$ of p is called a *shift component* of p if the set $\{p_{i_1}, \ldots, p_{i_l}\}$ is a shift class of p.

Example 1 *Consider the polynomial $p \in Q[x]$ given by the following factorization*

$$p(x) = (x^2+3)(x^2+2x+4)x(x+1)^2(x+3)(x+5)^2$$

then the set $\{x, x+1, x+3, x+5\}$ is a shift class of p and $\tilde{p} = x(x+1)^2(x+3)(x+5)^2$ is the corresponding shift component.

Define on the polynomials in a shift class the order $<$ by: $q < q'$ if there exists a positive integer k such that $E^k q = q'$. We can represent graphically the shift structure of a shift class. For instance Fig. 1 represents the situation for $\tilde{p}$ of Example 1.

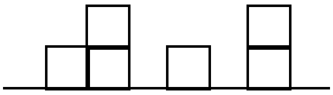

Figure 1: Shift structure of $x(x+1)^2(x+3)(x+5)^2$

We draw d squares at the i-th position on a line when the polynomial $E^i q_0$ arises with multiplicity d as a factor of $\tilde{p}$, where q_0 denotes the smallest polynomial w.r.t. the order $<$ in the shift class.

Definition 1 *The dispersion of a polynomial p is the maximal integer distance between roots of p and is denoted by dis(p).*

For a proper rational function given by a *reduced representation* $f = p/q$, i.e. p and q are relatively prime polynomials, we define $\text{dis}(f) = \text{dis}(q)$. In Example 1 we have $\text{dis}(p) = 5$, viz. the maximal distance between stacks.

Lemma 1 *For $f \in \mathcal{K}(x)$ let $h, r \in \mathcal{K}(x)$ be such that $f = \Delta h + r$ holds. Then r is a bound for f if and only if dis$(r) = 0$.*

This means that the problem of indefinite summation is solved for decompositions $f = \Delta h + r$ where $\text{dis}(r) = 0$, i.e. where each shift component of the denominator of r has only one stack of boxes in the corresponding diagram.

As we saw, solutions of the indefinite summation are not uniquely determined. In the next theorem, due to Paule, uniqueness *up to integer shifts* of the denominator of the remainders is stated.

Theorem 1 *Let $r, r' \in \mathcal{K}(x)$ be bounds for $f \in \mathcal{K}(x)$, given by the reduced representations $r = p/q, r' = p'/q'$. If $q = q_0^{e_0} \cdots q_m^{e_m}$ is the complete factorization of q over $\mathcal{K}$, then $q' = (E^{l_0} q_0^{e_0}) \cdots (E^{l_m} q_m^{e_m})$ for some integers $l_0, \ldots, l_m$.*

2 The Shift-Structure of the Denominator of Δh

In the next section we will discuss the shift-structure of the denominators of (h, r) in equation (1). In anticipation of this, we will now explicitly make some remarks on the denominator of the rational function Δh for a given h.

In the following let $h \in \mathcal{K}(x)$ be given by a reduced representation $h = \gamma/\delta$. Then we have

$$\Delta h = E\frac{\gamma}{\delta} - \frac{\gamma}{\delta} = \frac{\dfrac{\dfrac{\delta}{\tau}E\gamma - \gamma\dfrac{E\delta}{\tau}}{\tau_0}}{\dfrac{\delta E\delta}{\tau\tau_0}} = \frac{\dfrac{\dfrac{\delta}{\tau}E\gamma - \gamma\dfrac{E\delta}{\tau}}{\tau_0}}{\dfrac{\operatorname{lcm}(\delta, E\delta)}{\tau_0}} \tag{3}$$

where $\tau = \gcd(\delta, E\delta)$ and $\tau_0 = \gcd(\frac{\delta}{\tau}E\gamma - \gamma\frac{E\delta}{\tau}, \operatorname{lcm}(\delta, E\delta))$ and the right hand side is in reduced form.

Remark 1 *In equation (3) we have $\tau_0 | \tau$.*

Proof. We know $\tau_0 | \delta E\delta$. Assume now that there exists a non trivial factor τ_1 of τ_0 such that $\tau_1 | \delta$ and $\tau_1 \nmid E\delta$. Then $\tau_1 | \frac{\delta}{\tau}$ and $\tau_1 \nmid \frac{E\delta}{\tau}$. This is a contradiction to (3), for $\tau_0 | (\frac{\delta}{\tau}E\gamma - \gamma\frac{E\delta}{\tau})$ must hold and γ, δ are relatively prime. The case $\tau_1 | E\delta$ and $\tau_1 \nmid \delta$ is analogous. □

From this remark one easily obtains the following property.

Lemma 2 *Each shift-component $\pi^{m_0} \cdots (E^{l-1}\pi)^{m_{l-1}}$ of length l in δ corresponds to a shift-component $\pi^{m'_0} \cdots (E^l\pi)^{m'_l}$ of length $l+1$ in the denominator of Δh. Furthermore $m_0 = m'_0$ and $m_{l-1} = m'_l$ holds.*

Proof. Consider a shift-component of δ, viz. $\pi^{m_0}(E\pi)^{m_1} \cdots (E^{l-1}\pi)^{m_{l-1}}$ for some irreducible polynomial π and non-negative integers $m_0, \ldots, m_{l-1}$ with $m_0 m_{l-1} \neq 0$. Then obviously $(E\pi)^{m_0}(E^2\pi)^{m_1} \cdots (E^l\pi)^{m_{l-1}}$ is a shift-component of $E\delta$ and

$$(E\pi)^{\min(m_0,m_1)}(E^2\pi)^{\min(m_1,m_2)} \cdots (E^{l-1}\pi)^{\min(m_{l-2},m_{l-1})}$$

is a shift-component of $\gcd(\delta, E\delta)$, while

$$\pi^{m_0}(E\pi)^{\max(m_0,m_1)} \cdots (E^{l-1}\pi)^{\max(m_{l-2},m_{l-1})}(E^l\pi)^{m_{l-1}}$$

is a shift-component of $\operatorname{lcm}(\delta, E\delta)$. From Remark 1 we know $\tau_0 | \gcd(\delta, E\delta)$, so $\pi \nmid \tau_0$ and $E^l\pi \nmid \tau_0$. This implies the statement. □

Now we give a more precise description of the factors of τ_0.

Lemma 3 *For all non-trivial factors σ of $\gcd(\delta, E\delta)$ we have: $\sigma|\tau_0 \Rightarrow \sigma \nmid \frac{\delta}{\tau}$ and $\sigma \nmid \frac{E\delta}{\tau}$*

Proof. Assume $\sigma|\tau_0$ and $\sigma|\frac{\delta}{\tau}$. Notice that then $\sigma \nmid \frac{E\delta}{\tau}$ immediately follows. From $\sigma|\tau_0$ we know $\sigma|\frac{\delta}{\tau}E\gamma - \gamma\frac{E\delta}{\tau}$, so $\sigma|\gamma\frac{E\delta}{\tau}$. As δ and γ are relatively prime, we have $\sigma|\frac{E\delta}{\tau}$. This implies $\gcd(\frac{E\delta}{\tau}, \frac{\delta}{\tau}) \neq 1$, a contradiction to $\tau = \gcd(\delta, E\delta)$. Similarly for the assumption $\delta|\tau_0$ and $\sigma|\frac{E\delta}{\tau}$. □

In other words, if π is a factor of τ_0, then π and $E^{-1}\pi$ arise with same multiplicity in δ. This fact implies that the reduced denominator of Δh might differ from $\operatorname{lcm}(\delta, E\delta)$ only at those places where we have *repeated multiplicities* in the corresponding shift-component of δ.

Example 2 Consider a rational function, whose denominator has the shift-structure as represented in the left part of Fig. 2. Then the shift-structure of τ_0 will be contained in the diagram in the right part of Fig. 2.

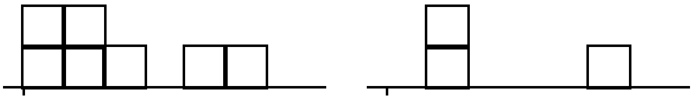

Figure 2: *Upper bound* for τ_0

Let us define a particular kind of shift structure which will play an important role in the following section.

Definition 2 *Let the shift structure of the polynomial q be given by one shift component*

$$q = \pi^{m_0}(E\pi)^{m_1}\cdots(E^l\pi)^{m_l}$$

where $l > 0$ and $m_0 m_l > 0$. We call such a shift structure **safe** *if for all $0 \leq i < j \leq l$ such that $m_i = m_j$ one of the following conditions holds*

1. *$i = 0$ and $j = l$*
2. *$\exists k : i < k < j$ and $m_k > m_i$*
3. *$\exists k_1, k_2 : k_1 < i < j < k_2$ and $m_{k_1} > m_i$ and $m_{k_2} > m_j$*

This means that if two stacks in the shift structure have the same number of boxes, then there must be an higher stack between them or they have to be enclosed between two higher stacks (or they are the endpoints). The definition naturally generalizes to polynomials with several shift components, i.e. each shift component is supposed to be safe.

3 The j-Shift-Saturation

In order to say more about the shift structure of the denominator polynomials of the solutions, we need the following definition.

Definition 3 *Let $q \in \mathcal{K}[x]$ be given by one shift component $q = \pi^{m_0}(E\pi)^{m_1} \cdots (E^l\pi)^{m_l}$. For all nonnegative integers $j \leq l$ we call the polynomial*

$$ShiftSat_j(q) := \pi^{t_0}(E\pi)^{t_1} \cdot \cdots \cdot (E^{l-1}\pi)^{t_{l-1}}$$

j-shift-saturation of q, where $t_i = \max\{m_0, \ldots, m_i\}$ for $i = 0, \ldots, j$ and $t_i = \max\{m_{i+1}, \ldots, m_l\}$ for $i = j, \ldots, l-1$.

The definition can be extended to $j < 0$ or $j > l$ assuming $m_i = 0$ for all $i < 0$ and $i > l$.

One can visualize the last definition as follows. Fix the j-th stack in the diagram corresponding to the shift structure of the polynomial, then do a saturation from the left of the stacks left of j and a saturation from the right of the stack right of j. After this, one simply erases the j-th stack and shifts the right boxes one step to the left.

In the following theorem the importance of the $ShiftSat_j(q)$ for our problem is explained.

Theorem 2 *Let $f = p/q \in \mathcal{K}(x)$ be such that the shift structure of q consists of a unique shift component $q = \pi^{m_0}(E\pi)^{m_1} \cdots (E^l\pi)^{m_l}$ with $m_0 m_l \neq 0$ and $l > 0$. Then for all integers j there exist solutions $h = \gamma/\delta$, $r = \varepsilon/\eta$ of $f = \Delta h + r$ such that*

$$\delta = Shift_j(q) \qquad \text{and} \qquad \eta = E^j\pi^m$$

holds, where $m = \max\{m_0, \ldots, m_l\}$. Furthermore, if the shift structure of q is safe, then γ/δ is already a reduced representation.

Proof. The existence of a solution for $\eta = (E^l\pi)^m$ is consequence of Theorem 3 in [3]. Consider now the decomposition

$$\frac{\varepsilon}{(E^l\pi)^m} = -\frac{E\varepsilon}{(E^{l+1}\pi)^m} + \frac{\varepsilon}{(E^l\pi)^m} + \frac{E\varepsilon}{(E^{l+1}\pi)^m}$$

This means that also $h' = h - r$ and $r' = Er$ form a solution to the indefinite summation problem, so a solution with $\eta = (E^j\pi)^m$ exists for all j (and with Theorem 1 all possible η have this form).

Let (h, r) be a solution with $\eta = (E^j\pi)^m$ for a certain j. For simplicity we consider only $0 < j < l$, but the proof can be easily extended to the remaining cases. As a consequence of Lemma 2 we know that the denominator δ of the rational part consists of a class of length $l-1$, starting with the same factor as q, i.e.

$$\delta = \pi^{d_0}(E\pi)^{d_1} \cdots (E^{l-1}\pi)^{d_{l-1}}$$

and $d_0 = m_0$. We now consider the multiplicities of the factors $E\pi, E^2\pi, \ldots, E^{j-1}\pi$ of δ. The multiplicity of $E\pi$ in $E\delta$ is $d_0 = m_0$, while the multiplicity in q is m_1. One of the following must hold.

1. $m_1 < d_0$. From the considerations after Remark 3 this can happen only if we have repeated multiplicities, so $d_1 = d_0$ must hold.

2. $m_1 > d_0$. This means that $d_1 = m_1$, for this is the only way to increase the multiplicity.

3. $m_1 = d_0$. This may happen only if $d_1 \leq d_0$.

In the example in Fig. 3 we have $d_0 = m_0 = 1$ and $m_1 = 2$, so it follows that $d_1 = 2$. In any case we have $d_1 \leq \max\{d_0, m_1\} = \max\{m_0, m_1\}$. These comparisons go step by step until we reach the $j - th$ position. From this follows that for all $0 \leq k < j$ it holds $d_k \leq \max\{m_0, \ldots, m_k\}$.

Now consider the right endpoint of the class. Similarly $d_{l-1} = m_l$, as $E^l\pi$ does not arise as factor of δ. Now we have to compare the $(l-1)$-st column, i.e. d_{l-2}, d_{l_1} and m_{l-1}. The only possible cases are

1. $m_{l-1} < d_{l-2}$. Again, after Remark 3 we need the same multiplicity of $E^{l-1}\pi$ in $E\delta$, so we have $d_{l-2} = d_{l-1}$.

2. $m_{l-1} > d_{l-2}$. This implies that $d_{l-2} = m_{l-1}$

3. $m_{l-1} = d_{l-2}$. This means that $d_{l-2} \leq d_{l-1}$.

In the example in Fig. 3 we have $d_5 = m_5 = 1$ and $m_4 = 0$, so it follows that $d_4 = 1$. Again we can iterate until the multiplicity d_j is determined. For all $j \leq k < l$, $d_k \leq \max\{m_{k+1}, \ldots, m_{l-1}\}$ holds.

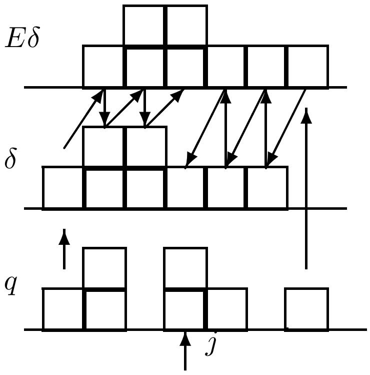

Figure 3: Determining δ

The steps on the multiplicities can be explained as follows. Going from left to the right we always take the maximal multiplicity ariseng so far. Similarly from the right to left. So δ is a divisor of the j-shift-saturation of q. On the other side, two of the cases above give no freedom of choice (the first and the second). One can easily see that if the denominator of f has a safe shift structure the third case never arises for any choice of j. For such f the rational part γ/δ is already reduced.

In fact, the third case arises only when in q two multiplicities with same value arise as “maxima” in the saturation. In other words, in order to have say $m_k = d_{k-1}$ for a $k < j$ there must be $m_{k+i} = m_k$ for some positive i but no $m_{k+h} > m_k$ with $h > 0$. Such a structure is unsafe. □

The fact that for safe shift structures the denominator δ of the rational part h is precisely $ShiftSat_j(q)$ allows us to compute a solution, where also δ has minimal degree. It is sufficient to choose j such that the corresponding j-shift-saturation is minimal in the degree, i.e. take j where the maximal multiplicity arises.

4 Concluding Remarks

In this note we discussed the shift structure of the denominators of solutions $h, r \in \mathcal{K}(x)$ to the problem of indefinite rational summation for a given $f \in \mathcal{K}(x)$.

In Theorem 2 we gave polynomials which were multiples of the denominators of h and r, respectively. We proved that for a certain class of summands the given estimates are precise. This description depends only on the structure of the denominator of the summand f and it is not reasonable to expect lower estimates without considering the numerator of the summand.

Provided $ShiftSat_j(q)$ can be computed by some effective algorithm, Theorem 2 leads to an algorithmic solution of the problem. In fact, knowing the denominators of h and r corresponds to reducing equation (1) to a polynomial equation, which can be solved by coefficient comparison.

In addition, our observations prove that the modification of Abramow's algorithm (see [1]) that we proposed in [4, 5] often yields a solution where both h and r are minimal in the degree of the denominator.

References

[1] S. A. Abramow, *The Rational Component of The Solution of a First-Order Linear Recurrence Relation with a Rational Right Side*, Zh. vychisl. Mat. mat. fis., 15, N. 4, 1035-1039, 1975

[2] R. Moenck, *On computing closed forms for summations*, Proceedings of MACSYMA users' conference, Berkley, pp. 225-236, 1977

[3] P. Paule, *Greatest Factorial Factorization and Symbolic Summation I*, RISC-Linz Report Series 93-02, J. Kepler University, Linz, 1993, *submitted to J. Symb. Comp.*

[4] R. Pirastu, *Algorithmen zur Summation rationaler Funktionen, (Algorithms for Summation of Rational Functions, in german)*, Diploma Thesis, Univ. Erlangen-Nürnberg, 1992

[5] R. Pirastu, *Algorithms for Indefinite Summation of Rational Functions in Maple*, submitted to The Maple Technical Newsletter