Applied Numerical Mathematics 8 (1991) 217–223
North-Holland

Solution of an integer programming problem related to convergence of rows of Padé approximants

Michael Kaminski and Avram Sidi

Computer Science Department, Technion — Israel Institute of Technology, Haifa 32000, Israel

Abstract

Kaminski, M. and A. Sidi, Solution of an integer programming problem related to convergence of rows of Padé approximants, Applied Numerical Mathematics 8 (1991) 217–223.

The following integer programming problem arises in the convergence analysis of rows of Padé approximants for meromorphic functions: maximize $\sum_{i=1}^{r}(\omega_i\sigma_i - \sigma_i^2)$, subject to the constraints $\sum_{i=1}^{r}\sigma_i = \tau$, $0 \leqslant \sigma_i \leqslant \omega_i$, $1 \leqslant i \leqslant r$. Here the ω_i and τ are given positive integers, and the σ_i are the integer unknowns. An algorithm is developed, by which all possible solutions can be constructed, and conditions for uniqueness are provided. Examples are appended.

1. Introduction

Let ω_i, $i = 1, \ldots, r$, be given positive integers ordered such that $\omega_1 \geqslant \omega_2 \geqslant \cdots \geqslant \omega_r$, and let τ be a given integer satisfying $0 < \tau < \sum_{i=1}^{r}\omega_i \equiv W$. Denote by IP($\tau$) the following nonlinear integer programming problem:

$$\begin{aligned} &\text{maximize} && \sum_{i=1}^{r}\left(\omega_i\sigma_i - \sigma_i^2\right), \\ &\text{subject to} && \sum_{i=1}^{r}\sigma_i = \tau, \\ & && 0 \leqslant \sigma_i \leqslant \omega_i, \quad i = 1, \ldots, r, \\ & && \sigma_i \text{ integers.} \end{aligned} \tag{1.1}$$

Notice that since there are only finitely many σ_i satisfying the constraints of (1.1), IP(τ) always has a solution.

This problem arose in [2] in the convergence analysis of the so-called intermediate rows of the Padé table for meromorphic functions, and its solution is crucial in determining whether any given such row of the Padé table converges or not.

Specifically, [2, Theorem 6.1], that is the relevant result, says the following: Let the function $f(z)$ by analytic at $z = 0$ and meromorphic in the open disk $K = \{z: |z| < R\}$. Denote the

number of its poles in K, including their multiplicities, by p. Assume further that $f(z)$ has only polar singularities on the boundary of K, $\partial K = \{ z: |z| = R \}$, and denote these poles and their respective multiplicities by z_i and ω_i, $i = 1, 2, \ldots, r$. Let $f_{m,k}(z)$ denote the (m/k) Padé approximant associated with the Maclaurin series of $f(z)$. Assume that for some τ satisfying $0 < \tau < \sum_{i=1}^{r} \omega_i \equiv W$, IP($\tau$) has a unique solution. Then the sequence of Padé approximants $\{ f_{m,p+\tau}(z) \}_{m=0}^{\infty}$, converges to $f(z)$ uniformly in any compact subset of K not including the poles of $f(z)$. Denote by $G(\tau)$ and $G(\tau+1)$ the values of the objective function $\sum_{i=1}^{r}(\omega_i\sigma_i - \sigma_i^2)$ at the solutions of IP(τ) and IP($\tau+1$) respectively. Actually,

$$f(z) - f_{m,p+\tau}(z) = \mathrm{O}\big(m^{G(\tau+1)-G(\tau)} |z/R|^m\big) \quad \text{as } m \to \infty$$

for $z \in K$ but z not a pole of $f(z)$. Furthermore, as $m \to \infty$ the denominator polynomial of $f_{m,p+\tau}(z)$ has p zeros that converge to the p poles of $f(z)$ that are in K, and σ_i zeros that converge to z_i, $1 \leqslant i \leqslant r$, on ∂K. If IP(τ) does not have a unique solution, it can be shown that there exist a set of points $A = \{ z_1', \ldots, z_\tau' \}$ and a subsequence of $\{ f_{m,p+\tau}(z) \}_{m=0}^{\infty}$ that converges uniformly to $f(z)$ in any compact subset of K not including the poles of $f(z)$ and points of A.

It is easy to see that for $\tau = 0$ and $\tau = \sum_{i=1}^{r} \omega_i \equiv W$ the problem IP(τ) has unique solutions, namely, $\sigma_i = 0$, $1 \leqslant i \leqslant r$, for the former, and $\sigma_i = \omega_i$, $1 \leqslant i \leqslant r$, for the latter. The uniform convergence of the sequences $\{ f_{m,p}(z) \}_{m=0}^{\infty}$ and $\{ f_{m,p+W}(z) \}_{m=0}^{\infty}$ is guaranteed by the well-known de Montessus' theorem.

For details and references pertaining to the above see [2].

The same problem also arises in [1] in the convergence analysis of the basic QR method on Hessenberg matrices, and an algorithm for its solution is given there. This algorithm is iterative in nature, in the sense that the solutions to IP($\tau+1$) are obtained, by testing r possibilities at most, from those of IP(τ), provided the latter are already known.

The primary purpose of this note is to present yet a different algorithm, by which the solution(s) to IP(τ) can be constructed in a noniterative fashion. This new algorithm is based on a sequence of at most $r-1$ reductions, which decrease the dimension of the problem, and it also enables us to decide whether the solution to IP(τ) is unique, and in case of nonuniqueness it provides all of the possible solutions.

Some of the properties of the solutions to IP(τ) are discussed in [2, Section 6]. We give these below as some of them will be of use in the sequel.

Let σ_j, $1 \leqslant j \leqslant r$, be a solution of IP($\tau$).

(1) $\sigma_j' = \omega_j - \sigma_j$, $1 \leqslant j \leqslant r$, is a solution of IP($\tau'$) with $\tau' = \sum_{j=1}^{r} \omega_j - \tau = W - \tau$. This implies that IP($\tau$) and IP($W-\tau$) simultaneously have unique (or nonunique) solutions and their solution sets are in one-to-one correspondence. Thus it is sufficient to treat IP(τ) for $0 < \tau \leqslant [\frac{1}{2}W]$.

(2) If $\omega_{j'} = \omega_{j''}$ for some j' and j'', $1 \leqslant j', j'' \leqslant r$, and if $\sigma_{j'} = \gamma_1$ and $\sigma_{j''} = \gamma_2$ is a solution to IP(τ), $\gamma_1 \neq \gamma_2$, then there is another solution to IP(τ) with $\sigma_{j'} = \gamma_2$ and $\sigma_{j''} = \gamma_1$. Consequently, a solution to IP(τ) cannot be unique unless $\sigma_{j'} = \sigma_{j''}$. One implication of this is that for $\omega_1 = \cdots = \omega_r = \bar{\omega} > 1$ the problem IP(τ) has a unique solution only for $\tau = qr$, $q = 1, \ldots, \bar{\omega} - 1$, and in this solution $\sigma_j = q$, $1 \leqslant j \leqslant r$. For $\omega_1 = \cdots = \omega_r = 1$ no unique solution to IP(τ) exists with $1 \leqslant \tau \leqslant r-1$. Another implication is that for $\omega_1 = \cdots = \omega_\mu > \omega_{\mu+1} \geqslant \cdots \geqslant \omega_r$, $\mu < r$, no unique solution to IP(τ) exists for $\tau = 1, \ldots, \mu - 1$, and a unique solution exists for $\tau = \mu$, this solution being $\sigma_1 = \cdots = \sigma_\mu = 1$, $\sigma_j = 0$, $\mu + 1 \leqslant j \leqslant r$.

(3) A unique solution to IP(τ) exists when ω_j, $1 \leqslant j \leqslant r$, are all even or all odd, and

$$\tau = qr + \tfrac{1}{2}\sum_{j=1}^{r}(\omega_j - \omega_r), \quad 0 \leqslant q \leqslant \omega_r.$$

This solution is given by

$$\sigma_j = q + \tfrac{1}{2}(\omega_j - \omega_r), \quad 1 \leqslant j \leqslant r.$$

The solution to IP(τ) can be discussed more conveniently by transforming the variables σ_i and the constant τ by

$$\delta_i = \tfrac{1}{2}\omega_i - \sigma_i, \quad 1 \leqslant i \leqslant r, \tag{1.2}$$
$$\eta = \tfrac{1}{2}W - \tau.$$

By (1.2) it is obvious that δ_i takes on integer (half-integer) values if ω_i is an even (odd) integer. Similarly, η takes on integer (half-integer) values if W is an even (odd) integer. Also, we have $-\frac{1}{2}W < \eta < \frac{1}{2}W$.

Then IP(τ) becomes equivalent to:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{r}\delta_i^2, \\ &\text{subject to} && \sum_{i=1}^{r}\delta_i = \eta, \\ & && -\tfrac{1}{2}\omega_i \leqslant \delta_i \leqslant \tfrac{1}{2}\omega_i, \quad 1 \leqslant i \leqslant r. \end{aligned} \tag{1.3}$$

In the next section we derive certain properties of the solution to (1.3), on which we base our constructive algorithm.

2. Theory

By what has been said about the correspondence between IP(τ) and IP($W - \tau$), we conclude that it is sufficient to treat the cases in which $0 \leqslant \frac{1}{2}W - [\frac{1}{2}W] \leqslant \eta < \frac{1}{2}W$, and this is done in the remainder of this section.

We also introduce the notation

$$\boldsymbol{\delta} = (\delta_1, \ldots, \delta_r) \quad \text{and} \quad F(\boldsymbol{\delta}) = \sum_{i=1}^{r}\delta_i^2.$$

Lemma 1. *If $\delta_k \neq -\frac{1}{2}\omega_k$ and $\delta_j \neq \frac{1}{2}\omega_j$, and if $\delta_k - \delta_j \geqslant \frac{3}{2}$, then $F(\boldsymbol{\delta}') < F(\boldsymbol{\delta})$, where $\boldsymbol{\delta}'$ is obtained from $\boldsymbol{\delta}$ by replacing δ_k and δ_j by $\delta_k - 1$ and $\delta_j + 1$ respectively.*

Proof. The assertion follows by noting that both $\boldsymbol{\delta}$ and $\boldsymbol{\delta}'$ satisfy the constraints in (1.3), and that

$$\begin{aligned} F(\boldsymbol{\delta}') &= \sum_{\substack{i=1 \\ i\neq j,k}}^{r}\delta_i^2 + (\delta_j + 1)^2 + (\delta_k - 1)^2 \\ &= F(\boldsymbol{\delta}) + 2 - 2(\delta_k - \delta_j) \leqslant F(\boldsymbol{\delta}) - 1. \end{aligned}$$

□

Lemma 2. *If* $\eta/r \geqslant \frac{1}{2}\omega_j$ *for some* j, *an* (*optimal*) *solution* $\boldsymbol{\delta}$ *to* (1.3) *must have* $\delta_j = \frac{1}{2}\omega_j$.

Proof. Suppose to the contrary that $\delta_j < \frac{1}{2}\omega_j$. Thus, $\delta_j < \frac{1}{2}\omega_j \leqslant \eta/r$. Now η/r is the average of $\delta_1, \ldots, \delta_r$. Consequently, if $\delta_j < \eta/r$, then there exists δ_k that satisfies $\delta_k > \eta/r$. Combining all this we have

$$\delta_j < \tfrac{1}{2}\omega_j \leqslant \eta/r < \delta_k \leqslant \tfrac{1}{2}\omega_k.$$

We now note that if $\delta_j < \frac{1}{2}\omega_j$, then $\delta_j \leqslant \frac{1}{2}\omega_j - 1$, and if $\delta_k > \frac{1}{2}\omega_j$, then $\delta_k \geqslant \frac{1}{2}\omega_j + \frac{1}{2} > 0$. Therefore, $\delta_k - \delta_j \geqslant \frac{3}{2}$, and δ_k and δ_j are as in Lemma 1, so that $F(\boldsymbol{\delta})$ is not minimal, contrary to the assumption. Thus, $\delta_j = \frac{1}{2}\omega_j$ must hold. □

Lemma 2 is very useful in that if $\omega_r \leqslant \cdots \leqslant \omega_{r'+1} \leqslant 2\eta/r < \omega_{r'} \leqslant \cdots \leqslant \omega_1$, then it *uniquely* fixes $\delta_i = \frac{1}{2}\omega_i$, $r' + 1 \leqslant i \leqslant r$, and allows us to complete the solution by solving the reduced problem:

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{r'} \delta_i^2, \\ &\text{subject to} && \sum_{i=1}^{r'} \delta_i = \eta - \sum_{i=r'+1}^{r} \tfrac{1}{2}\omega_i \equiv \eta', \\ & && -\tfrac{1}{2}\omega_i \leqslant \delta_i \leqslant \tfrac{1}{2}\omega_i, \quad 1 \leqslant i \leqslant r'. \end{aligned} \tag{2.1}$$

Note that if $\eta \geqslant 0$, then $\eta' \geqslant 0$ too. It may be possible to reduce (2.1) further by applying Lemma 2 again. Finally, when Lemma 2 can no longer be applied, the solution and an algorithm for it can be completed by applying Lemmas 3 and 4 below.

Lemma 3. *If* $\eta/r < \frac{1}{2}\omega_i$, $1 \leqslant i \leqslant r$, *then an* (*optimal*) *solution* $\boldsymbol{\delta}$ *must satisfy* $|\delta_k - \delta_j| \leqslant 1$ *for any two indices* j *and* k.

Proof. Suppose to the contrary that for some j and k $|\delta_k - \delta_j| \geqslant \frac{3}{2}$. Without loss of generality we can write $\delta_k - \delta_j \geqslant \frac{3}{2}$, or equivalently $\delta_k \geqslant \delta_j + \frac{3}{2}$. We can let

$$\delta_k = \max_i \delta_i, \quad \text{and} \quad \delta_j = \min_i \delta_i.$$

Now since $\delta_j < \delta_k$ and since η/r is the average of $\delta_1, \ldots, \delta_r$, we must have $\delta_j < \eta/r < \delta_k$. By assumption, $\eta/r < \frac{1}{2}\omega_i$ for all i, so that $\delta_j < \eta/r < \frac{1}{2}\omega_j$, hence $\delta_j \neq \frac{1}{2}\omega_j$. Similarly, $0 < \eta/r < \delta_k$, hence $\delta_k \neq -\frac{1}{2}\omega_k$. Therefore, Lemma 1 applies, and $\boldsymbol{\delta}$ cannot be a minimal solution contrary to the assumption. Thus we must have $|\delta_k - \delta_j| \leqslant 1$ for all j and k. □

Lemma 3 has some very powerful consequences that will be exploited in the development of the algorithm. First, if we set $a \equiv \max_i \delta_i$ in an optimal solution $\boldsymbol{\delta}$, and a is integer (half-integer), then the δ_i corresponding to the even (odd) ω_i take on the values a or $a - 1$, while the δ_i corresponding to the odd (even) ω_i *all* take on the value $a - \frac{1}{2}$. Let us denote the numbers of δ_i that assume the values a, $a - \frac{1}{2}$, and $a - 1$, by x, y, and z respectively. Let us also denote by p_e (p_o) the number of the even (odd) ω_i. Then $x + z = p_e$ and $y = p_o$ in case a is an integer, $x + z = p_o$ and $y = p_e$ in case a is a half-integer. By the definition of a, $x > 0$. Obviously, $y = 0$ if ω_i are all even or all odd, and $z \geqslant 0$ always.

Lemma 4. *Under the conditions of Lemma* 3, a, x, y, *and* z *are unique. Consequently, if* $z = 0$, *the solution to* IP(τ) *is unique. If* $z \neq 0$, IP(τ) *does not have a unique solution, but the solutions are obtained from each other by permutation of the* δ_i *that assume the values* a *and* $a - 1$.

Proof. By the constraints in (1.3) we have

$$\eta = \sum_{i=1}^{r} \delta_i = xa + y\left(a - \tfrac{1}{2}\right) + z(a-1) = (x+y+z)a - \tfrac{1}{2}y - z,$$

which can be rewritten in the form

$$a = \frac{\eta}{r} + \frac{\frac{1}{2}y + z}{r}, \tag{2.2}$$

if we recall that

$$x + y + z = r. \tag{2.3}$$

We now look for solutions of (2.2) and (2.3) under the additional constraints $x > 0$, $z \geqslant 0$, and $y = p_o$ or $y = p_e$ depending on whether a is an integer or half-integer. There are two cases to consider:

(1) There are both even and odd ω_i, so that $y > 0$. In this case (2.2) implies

$$\frac{\eta}{r} < a < \frac{\eta}{r} + 1. \tag{2.4}$$

Let us denote $S = (\eta/r,\ \eta/r + 1)$. Now if η/r is an integer (half-integer), then a is the unique half-integer (integer) contained in S, and consequently $y = p_e$ ($y = p_o$) and x and z are uniquely determined from (2.2) and (2.3). If η/r is neither an integer nor a half-integer, then S contains exactly one integer and one half-integer, which we denote by α_1 and $\alpha_2 = \alpha_1 + \frac{1}{2}$. Thus either $a = \alpha_1$ or $a = \alpha_2$, or both. We now need to show that α_1 and α_2 cannot both be solutions for a. Suppose to the contrary that a is not unique. Then (2.2) and (2.3) have two solutions $(\alpha_1, x_1, y_1, z_1)$ and $(\alpha_2, x_2, y_2, z_2)$ for (a, x, y, z). By (2.2) these solutions must satisfy

$$\tfrac{1}{2}x_2 - \tfrac{1}{2}z_2 = -\tfrac{1}{2}y_1 - z_1, \tag{2.5}$$

and, by (2.3) and the fact that y is either p_e or p_o they must also satisfy

$$x_1 + z_1 = y_2 \quad \text{and} \quad x_2 + z_2 = y_1. \tag{2.6}$$

From (2.6), $z_2 = y_1 - x_2$. Substituting this in (2.5), we obtain $x_2 = -z_1 \leqslant 0$, which contradicts $x_2 > 0$. Once a is known, x, y, z can be determined uniquely as explained before.

(2) All ω_i are even or they are all odd, so that $y = 0$. By $x > 0$ and $x + z = r$, we have $0 \leqslant z < r$. Thus (2.2) implies that

$$\frac{\eta}{r} \leqslant a < \frac{\eta}{r} + 1. \tag{2.7}$$

Now if $p_o = 0$ ($p_e = 0$), then a is an integer (half-integer), thus (2.7) can have only one solution for a. Once a is determined, x and z can be determined uniquely as before.

When $z = 0$ the uniqueness of the optimal $\boldsymbol{\delta}$ is obvious. When $0 < z < r - 1$, (2.4) is always satisfied. Recalling that we are dealing with the case $\eta \geqslant 0$, (2.4) implies that $a > 0$. This implies

that $a \geq \frac{1}{2}$ so that $a - 1 \geq -\frac{1}{2} \geq -\frac{1}{2}\omega_i$ for all i. Also by the assumption $\eta/r < \frac{1}{2}\omega_i$ for all i, (2.4) implies $a < \frac{1}{2}\omega_i + 1$ for all i. If a is integer (half-integer), then $a \leq \frac{1}{2}\omega_i$ for ω_i even (odd). Combining all this we see that *all* permutations among the δ_i that assume the values a and $a - 1$ are possible. This completes the proof. □

It is important to note that in case the interval S in the proof of Lemma 4 contains both an integer and a half-integer (α_1 and α_2) only one of them satisfies (2.2) and (2.3).

3. Examples

We demonstrate the application of the results of the previous section with two examples.

Example 1. Consider IP(τ) with $r = 2$ and ω_1 even (odd) and ω_2 odd (even), and $\omega_1 > \omega_2$. For $0 \leq \tau \leq [\frac{1}{2}(\omega_1 - \omega_2)]$ we have $\omega_2 \leq \eta < \omega_1$. Consequently, Lemma 2 applies, and we obtain the unique solution $\sigma_1 = \tau$ and $\sigma_2 = 0$. For

$$[\tfrac{1}{2}(\omega_1 - \omega_2)] < \tau \leq [\tfrac{1}{2}W] = [\tfrac{1}{2}(\omega_1 + \omega_2)],$$

we have

$$\tfrac{1}{2} = \tfrac{1}{2}W - [\tfrac{1}{2}W] \leq \eta < \omega_2 < \omega_1.$$

Consequently, Lemmas 3 and 4 apply, and we obtain

$$x = 1, \qquad y = 1, \qquad z = 0, \qquad a = \tfrac{1}{4}(2\eta + 1) = \tfrac{1}{4}(W - 2\tau + 1),$$

and the solution to IP(τ) is again unique. The values of σ_1 and σ_2 can now be determined easily. Consider, for instance, the case in which ω_1 is even and ω_2 is odd. If a is an integer, then $\sigma_1 = \frac{1}{2}\omega_1 - a$ and $\sigma_2 = \tau - \sigma_1$, and if a is a half-integer, then $\sigma_2 = \frac{1}{2}\omega_2 - a$ and $\sigma_1 = \tau - \sigma_2$. Also, the solution to IP(τ) for $[\frac{1}{2}W] < \tau < W$ can be obtained from that of IP($W - \tau$) as explained in the introduction. In any case, the solution of IP(τ), for *all* possible τ, is unique.

Example 2. Consider IP(τ) with $r = 10$ and $\omega_i = 10 - i + 1$, $1 \leq i \leq 10$. Let us take $\tau = 5$. Then

$$W = \sum_{i=1}^{r} \omega_i = 55 \quad \text{and} \quad \eta = \tfrac{45}{2} > 0.$$

Now $\eta/r = \frac{45}{20}$ so that $\eta/r \geq \frac{1}{2}\omega_i$, $7 \leq i \leq 10$. By Lemma 2, $\delta_i = \frac{1}{2}\omega_i$, $7 \leq i \leq 10$. Thus we have reduced the problem as in (2.1) with $r' = 6$ and $\eta' = \frac{35}{2}$, and $\omega_i = 10 - i + 1$, $1 \leq i \leq 6$. This time $\eta'/r' = \frac{35}{12}$ so that $\eta'/r' \geq \frac{1}{2}\omega_6$ only. Again by Lemma 2, $\delta_6 = \frac{1}{2}\omega_6$, and the problem is reduced further as in (2.1) with $r'' = 5$ and $\eta'' = 15$, and $\omega_i = 10 - i + 1$, $1 \leq i \leq 5$. This time $\eta''/r'' = 3 \geq \frac{1}{2}\omega_5$, and again $\delta_5 = \frac{1}{2}\omega_5$ by Lemma 2. In the new reduced problem $r''' = 4$ and $\eta''' = 12$ and $\omega_i = 10 - i + 1$, $1 \leq i \leq 4$. Now $\eta'''/r''' = 3 < \frac{1}{2}\omega_i$, $1 \leq i \leq 4$, so that no further reduction is possible, and Lemmas 3 and 4 apply. Since $y \neq 0$ in this reduced problem, (2.4) holds and we have $3 < a < 4$ so that $a = 3.5$ is the only possible solution. Consequently, $y = 2$, and $z = 1$ from (2.2), so that $x = 1$. This means that there are two (nonunique) optimal solutions with $(\delta_1, \delta_2, \delta_3, \delta_4) = (3, 3.5, 3, 2.5)$ and $(3, 2.5, 3, 3.5)$. Invoking $\delta_i = \frac{1}{2}\omega_i$, $5 \leq i \leq 10$, and (1.2), we finally have that IP(5) has two optimal solutions $(\sigma_1, \ldots, \sigma_{10})$ with $\sigma_i = 0$, $5 \leq i \leq 10$, $\sigma_1 = 2$,

$\sigma_3 = 1$ in both solutions. $\sigma_2 = 1$ and $\sigma_4 = 1$ in one of these solutions, while $\sigma_2 = 2$ and $\sigma_4 = 0$ in the other. Finally, the solutions to IP(50) are obtained by replacing σ_i in these solutions by $\omega_i - \sigma_i$.

References

[1] B. Parlett, Global convergence of the basic QR algorithm on Hessenberg matrices, *Math. Comp.* 22 (1968) 803–817.

[2] A. Sidi, Quantitative and constructive aspects of the generalized Koenig's and the Montessus's theorems for Padé approximants, *J. Comput. Appl. Math.* 29 (1990) 257–291.