SIAM J. NUMER. ANAL.
Vol. 9, No. 3, September 1972

# THE DIFFERENTIAL CORRECTION ALGORITHM FOR RATIONAL $\ell_\infty$-APPROXIMATION*

I. BARRODALE,† M. J. D. POWELL‡ AND F. D. K. ROBERTS†

**Abstract.** The version of the "differential correction algorithm" that is most used at the present time is a modification of the original version, perhaps because it has been proved that the modified version has sure convergence properties. However, the purpose of this paper is to direct attention back to the original version. It is now proved that the original version also has sure convergence properties. Furthermore, we prove that its rate of convergence is quadratic. This makes it superior to the more popular, modified version of the algorithm. Some numerical examples are given to compare the two versions, and these leave little doubt that the original algorithm is much better in practice.

**1. Introduction.** We consider the problem of approximating a given real-valued function $f(x)$, on a discrete point set $X = \{x_1, x_2, \cdots, x_N\}$, by a rational function

$$R(x) = P(x)/Q(x) = \sum_{i=0}^{m} p_i x^i \Big/ \sum_{j=0}^{n} q_j x^j, \tag{1.1}$$

where the integers $m$ and $n$ are given. We wish to calculate the values of the coefficients $p_i$, $i = 0, 1, \cdots, m$, and $q_j$, $j = 0, 1, \cdots, n$, that, subject to the condition

$$Q(x_t) > 0, \qquad t = 1, 2, \cdots, N, \tag{1.2}$$

minimize the quantity

$$\max_{1 \leqq t \leqq N} |f(x_t) - R(x_t)| = \|f - R\|_\infty. \tag{1.3}$$

This problem need not have a solution, but where necessary we assume that the function $f(x)$ is such that a solution exists.

In this paper we study and compare two versions of the differential correction algorithm for minimizing the expression (1.3). Both versions of the algorithm generate a sequence of approximations, $R_k(x) = P_k(x)/Q_k(x)$, $k = 1, 2, \cdots$. Of course, the two versions generate different sequences, but we prefer not to use a notation that distinguishes the two sequences. For either version we let $\Delta_k$ be the current maximum error

$$\Delta_k = \max_t |f(x_t) - P_k(x_t)/Q_k(x_t)|, \tag{1.4}$$

and in both cases $\Delta_k$, $k = 1, 2, \cdots$, converges to the minimax error

$$\Delta^* = \inf_{P,Q} \max_t |f(x_t) - P(x_t)/Q(x_t)|, \tag{1.5}$$

* Received by the editors April 22, 1971. This work was supported in part by the National Research Council of Canada under Grants A5251 and A7143.

† Department of Mathematics, University of Victoria, Victoria, British Columbia, Canada.
‡ Mathematics Branch, Atomic Energy Research Establishment, Harwell, Berkshire, England.

where $Q(x_t) > 0$, $t = 1, 2, \cdots, N$. Even when the given problem has no solution, $\Delta^*$ is well-defined, and $\Delta_k$ still converges to $\Delta^*$.

The popular version of the differential correction algorithm (DC say) is described by Cheney and Loeb [3], Cheney and Southard [4], Cheney [1] and by Rice [6]. It calculates $P_{k+1}(x)$ and $Q_{k+1}(x)$ by minimizing the expression

$$\max_t \{|f(x_t)Q(x_t) - P(x_t)| - \Delta_k Q(x_t)\}, \tag{1.6}$$

where $\Delta_k$ is defined by (1.4). However, in the original version (ODC say) $P_{k+1}(x)$ and $Q_{k+1}(x)$ are calculated to minimize the expression

$$\max_t \left\{ \frac{|f(x_t)Q(x_t) - P(x_t)| - \Delta_k Q(x_t)}{Q_k(x_t)} \right\} \tag{1.7}$$

(Cheney and Loeb [2]).

Unless $\Delta_k = \Delta^*$, it happens that $P(x)$ and $Q(x)$ can be found such that expressions (1.6) and (1.7) are negative. Therefore, because these expressions are homogeneous in the required coefficients $p_i$, $i = 0, 1, \cdots, m$, and $q_j$, $j = 0, 1, \cdots, n$, a normalization condition is necessary. In this paper we impose the condition

$$\max_j |q_j| = 1, \qquad j = 0, 1, \cdots, n. \tag{1.8}$$

There are three main advantages of the differential correction algorithm over the Remes algorithm (see Rice [6] for instance) for computing rational $\ell_\infty$-approximations. The first is rather weak, and it is that the subproblem of minimizing expression (1.6) or (1.7) is a finite calculation, because it yields to linear programming methods. The second advantage is guaranteed convergence from any starting approximation subject to condition (1.2). Thirdly, on every iteration the condition (1.2) is maintained *automatically* by the definition of $P_{k+1}(x)/Q_{k+1}(x)$.

In spite of these last two properties, the Remes algorithm is in more frequent use at the present time, probably because the iteration (1.6) converges only linearly. However, we show in § 2 that the iteration (1.7) converges quadratically. Therefore it is likely that the ODC algorithm will be especially useful for solving some of the problems that are awkward for the Remes algorithm.

As well as the quadratic convergence theorem, the guaranteed convergence from any starting approximation subject to condition (1.2) is proved in § 2. Then in § 3 some numerical examples are given that show very clearly that the ODC algorithm is better than the DC algorithm.

**2. Theory.** The theorems given in this section apply to the ODC algorithm, which defines $P_{k+1}(x)/Q_{k+1}(x)$ by minimizing expression (1.7), subject to condition (1.8). Theorem 1 states that each iteration improves the current approximation, Theorem 2 shows that the sequence $\Delta_k$, $k = 1, 2, \cdots$, converges to $\Delta^*$, and Theorem 3 proves that the rate of convergence is at least quadratic. (For the DC algorithm we can also show that Theorems 1 and 2 hold, but that now the rate of convergence is at least linear. Cheney [1] proves these results for the DC algorithm in the case of approximation on an interval.)

Before proving these convergence theorems for the ODC algorithm, we shall first provide some insight into the derivation of expression (1.7). This motivation, which comes about by expressing a product of two quantities by the linear terms in a Taylor series expansion, also suggests that the rate of convergence of the algorithm is quadratic.

For any choice of $P(x)$ and $Q(x)$ satisfying (1.2) and (1.8), let $\Delta$ be a number for which the inequality

$$|f(x_t) - P(x_t)/Q(x_t)| \leqq \Delta \tag{2.1}$$

holds for all $t$. Expression (2.1) can be rewritten as

$$\begin{aligned} |f(x_t)Q(x_t) - P(x_t)| &\leqq \Delta Q(x_t) \\ &\lesssim \Delta_k Q_k(x_t) + (\Delta - \Delta_k)Q_k(x_t) + (Q(x_t) - Q_k(x_t))\Delta_k \end{aligned} \tag{2.2}$$

for each value of $t$. Finally, canceling the first and last products on the right side of (2.2) and rearranging the remaining terms, we observe that

$$\max_t \left\{ \frac{|f(x_t)Q(x_t) - P(x_t)| - \Delta_k Q(x_t)}{Q_k(x_t)} \right\} + \Delta_k \lesssim \Delta. \tag{2.3}$$

Thus, from (2.3), the quantity $\Delta$ is minimized approximately by computing the minimum of (1.7).

THEOREM 1.[1] *If $Q_k(x)$ satisfies conditions* (1.2) *and* (1.8), *and if $\Delta_k \neq \Delta^*$, then $\Delta_{k+1} < \Delta_k$ and $Q_{k+1}(x_t) > 0$, $t = 1, 2, \cdots, N$.*

*Proof*. Let $P^+(x)/Q^+(x)$ be any approximation, satisfying conditions (1.2) and (1.8), whose maximum error on $\{x_1, x_2, \cdots, x_N\}$ is less than $\Delta_k$, and let this maximum error be $\Delta^+$. We let $\eta^+$ be the number

$$\eta^+ = \min_t Q^+(x_t), \tag{2.4}$$

and we note that it is positive. Moreover we require the definition

$$M = \max_t \sum_{j=0}^{n} |x_t|^j, \tag{2.5}$$

because $M$ is the greatest number that can be attained by a polynomial $Q(x)$ of degree $n$ on the point set $\{x_1, x_2, \cdots, x_N\}$, subject to the normalization condition (1.8).

The proof of the theorem depends on the remark that, because of the definition of $P_{k+1}(x)/Q_{k+1}(x)$, the inequality

$$\begin{aligned} &\max_t \left\{ \frac{|f(x_t)Q_{k+1}(x_t) - P_{k+1}(x_t)| - \Delta_k Q_{k+1}(x_t)}{Q_k(x_t)} \right\} \\ &\leqq \max_t \left\{ \frac{|f(x_t)Q^+(x_t) - P^+(x_t)| - \Delta_k Q^+(x_t)}{Q_k(x_t)} \right\} \end{aligned} \tag{2.6}$$

[1] The proof of this theorem is similar to that given in [2] for the DC algorithm.

holds. Now, from the notation given in the last paragraph, the right-hand side of this inequality satisfies the condition

$$\begin{aligned}\max_t \left\{\left[\left|f(x_t) - \frac{P^+(x_t)}{Q^+(x_t)}\right| - \Delta_k\right]\frac{Q^+(x_t)}{Q_k(x_t)}\right\} &\leqq \max_t \left\{[\Delta^+ - \Delta_k]\frac{Q^+(x_t)}{Q_k(x_t)}\right\} \\ &= -[\Delta_k - \Delta^+]\min_t\left\{\frac{Q^+(x_t)}{Q_k(x_t)}\right\} \\ &\leqq -\eta^+[\Delta_k - \Delta^+]/M. \qquad (2.7)\end{aligned}$$

Therefore, from expressions (2.6) and (2.7), it follows that for all $t$ the inequality

$$\Delta_k Q_{k+1}(x_t)/Q_k(x_t) \geqq \eta^+[\Delta_k - \Delta^+]/M \tag{2.8}$$

holds, so we have proved that $Q_{k+1}(x_t) > 0$, $t = 1, 2, \cdots, N$.

To prove the other half of Theorem 1, we note that inequalities (2.6) and (2.7) also imply that the condition

$$\{|f(x_t)Q_{k+1}(x_t) - P_{k+1}(x_t)| - \Delta_k Q_{k+1}(x_t)\}/Q_k(x_t) < 0 \tag{2.9}$$

holds for all values of $t$. Therefore, because $Q_k(x_t)$ and $Q_{k+1}(x_t)$ are positive, we deduce the inequality

$$|f(x_t) - P_{k+1}(x_t)/Q_{k+1}(x_t)| < \Delta_k, \qquad t = 1, 2, \cdots, N. \tag{2.10}$$

The statement of the theorem, $\Delta_{k+1} < \Delta_k$, is now an immediate consequence of the definition of $\Delta_{k+1}$ (see (1.4)).

THEOREM 2. *The sequence of maximum errors* $\Delta_k$, $k = 1, 2, \cdots$, *converges to* $\Delta^*$ (*see* (1.5)).

*Proof*. Theorem 1 shows that the sequence $\Delta_k$, $k = 1, 2, \cdots$, decreases monotonically, and it is bounded below, so it has a limit $\bar{\Delta}$ say. To prove the theorem we suppose that $\bar{\Delta} > \Delta^*$, and we arrive at a contradiction. Following the notation of Theorem 1, let $P^+(x)/Q^+(x)$ be any approximation, satisfying conditions (1.2) and (1.8), whose maximum error in $X$, $\Delta^+$ say, is less than $\bar{\Delta}$.

Inequalities (2.6) and (2.7) imply that, for $t = 1, 2, \cdots, N$, the condition

$$\left|f(x_t) - \frac{P_{k+1}(x_t)}{Q_{k+1}(x_t)}\right| - \Delta_k \leqq \frac{-\eta^+[\Delta_k - \Delta^+]}{M}\frac{Q_k(x_t)}{Q_{k+1}(x_t)} \tag{2.11}$$

holds. It is convenient to introduce a notation, $\xi_k$ say, for the member of the set $\{x_1, x_2, \cdots, x_N\}$ for which $Q_k(x_t)/Q_{k+1}(x_t)$ is least. Therefore expression (2.11) implies the inequality

$$\Delta_{k+1} - \Delta_k \leqq \frac{-\eta^+[\bar{\Delta} - \Delta^+]}{M}\frac{Q_k(\xi_k)}{Q_{k+1}(\xi_k)}. \tag{2.12}$$

Since the sequence $\Delta_k$, $k = 1, 2, \cdots$, converges and $\bar{\Delta} > \Delta^+$, we have that

$$\lim_{k\to\infty}\frac{Q_k(\xi_k)}{Q_{k+1}(\xi_k)} = 0. \tag{2.13}$$

We now show that this statement leads to a contradiction.

From inequality (2.8) we deduce the bound

$$Q_{k+1}(x_t) \geqq Q_k(x_t)\eta^+[\Delta_k - \Delta^+]/M\Delta_k$$
$$\geqq cQ_k(x_t), \qquad t = 1, 2, \cdots, N, \tag{2.14}$$

where $c$ is the positive constant

$$c = \min\,[\tfrac{1}{2}, \eta^+(\bar{\Delta} - \Delta^+)/M\bar{\Delta}], \tag{2.15}$$

and, believing statement (2.13) for the moment, we let $K$ be an integer such that the inequality

$$Q_k(\xi_k) \leqq c^N Q_{k+1}(\xi_k) \tag{2.16}$$

holds for all $k \geqq K$. Then from expressions (2.14) and (2.16) we deduce the inequality

$$\prod_{t=1}^{N} Q_{k+1}(x_t) \geqq \frac{c^{N-1}}{c^N} \prod_{t=1}^{N} Q_k(x_t)$$
$$\geqq 2 \prod_{t=1}^{N} Q_k(x_t). \tag{2.17}$$

It follows that $\prod Q_k(x_t)$ diverges as $k$ tends to infinity, but this statement contradicts the normalization condition (1.8). Theorem 2 is proved.

THEOREM 3. *If $N \geqq n + m + 1$, if a best approximation exists, and if the best approximation is "normal," then the rate of convergence of the ODC algorithm is at least quadratic.*

*Proof.* We let the best approximation be the function

$$R^*(x) = P^*(x)/Q^*(x) = \sum_{i=0}^{m} p_i^* x^i \Big/ \sum_{j=0}^{n} q_j^* x^j, \tag{2.18}$$

and, following condition (1.8), we let its normalization be

$$\max_j |q_j^*| = 1. \tag{2.19}$$

The minimax error, $\Delta^*$, is defined by (1.5). The statement that the best approximation is normal means that the polynomials $P^*(x)$ and $Q^*(x)$ have no roots in common, and that the coefficients $p_m^*$ and $q_n^*$ are not both zero. In this case, because $N \geqq m + n + 1$, Cheney's [1, p. 165] "strong unicity theorem" holds.[2] It states that there exists a positive constant $\gamma$ such that the inequality

$$\|f - R\|_\infty \geqq \|f - R^*\|_\infty + \gamma\|R - R^*\|_\infty \tag{2.20}$$

is satisfied, where $R$ is any rational function of the form (1.1) satisfying the condition (1.2), and where the norm is defined by (1.3). To use expression (2.20) we require the following lemma.

[2] Although Cheney [1] states this theorem for approximation on an interval, it is only necessary that the domain $X$ be compact.

LEMMA. *There exists a constant* $\theta$ *such that the inequality*

$$\|Q - Q^*\|_\infty \leqq \theta \|R - R^*\|_\infty \tag{2.21}$$

*holds, for all rational functions of the form* (1.1) *that satisfy condition* (1.2), *and that are normalized to give the equation*

$$\max_j |q_j| = 1. \tag{2.22}$$

*Proof of lemma.* Let $\delta$ be the quantity

$$\delta = \|R - R^*\|_\infty. \tag{2.23}$$

The definition of $M$ (see (2.5)) and (2.23) imply the inequality

$$\max_t |P(x_t)Q^*(x_t) - Q(x_t)P^*(x_t)| \leqq M^2\delta. \tag{2.24}$$

For the remainder of the proof we assume that $p_m^* \neq 0$. However if $p_m^* = 0$, then $q_n^* \neq 0$, and an argument that is analogous to the one that follows can be used.

We let $\alpha$ be the ratio $p_m/p_m^*$, we define the polynomials

$$\begin{aligned} \bar{P}(x) &= P(x) - \alpha P^*(x) = \sum_{i=0}^{m-1} \bar{p}_i x^i, \\ \bar{Q}(x) &= Q(x) - \alpha Q^*(x) = \sum_{j=0}^{n} \bar{q}_j x^j, \end{aligned} \tag{2.25}$$

and we obtain from condition (2.24) the inequality

$$\max_t \left| \sum_{i=0}^{m-1} \bar{p}_i x_t^i Q^*(x_t) - \sum_{j=0}^{n} \bar{q}_j x_t^j P^*(x_t) \right| \leqq M^2\delta. \tag{2.26}$$

Next we show that the functions $x^iQ^*(x)$, $i = 0, 1, \cdots, m-1$, and $x^jP^*(x)$, $j = 0, 1, \cdots, n$, are linearly independent on $X$. If they were dependent, then there would exist nonzero polynomials $A(x)$ and $B(x)$, of degrees at most $m-1$ and $n$ respectively, such that the equation

$$A(x_t)Q^*(x_t) - B(x_t)P^*(x_t) = 0, \qquad t = 1, 2, \cdots, N, \tag{2.27}$$

holds. Since $N \geqq m + n + 1$, the polynomial $\{A(x)Q^*(x) - B(x)P^*(x)\}$ would be identically zero, and therefore $A(x)Q^*(x)$ would be zero at the $m$ zeros of $P^*(x)$. Now the degree of $A(x)$ is at most $m-1$, and therefore $P^*(x)$ and $Q^*(x)$ would have at least one zero in common, which is contrary to the conditions of Theorem 3. It follows that the functions of expression (2.26) are independent on $X$, so we deduce that there exists a constant, $d$ say, such that the bounds

$$\begin{aligned} \bar{p}_i &\leqq d\delta, \qquad i = 0, 1, \cdots, m-1, \\ \bar{q}_j &\leqq d\delta, \qquad j = 0, 1, \cdots, n, \end{aligned} \tag{2.28}$$

are obtained for all $P(x)/Q(x)$.

Next we show that the value of $\alpha$ is close to one when $\delta$ is small. Equation (2.25) gives the identity

$$\bar{q}_j = q_j - \alpha q_j^*, \qquad j = 0, 1, \cdots, n, \tag{2.29}$$

and therefore, if we let $j$ have the value for which $|q_j| = 1$, we find $|\bar{q}_j + \alpha q_j^*| = 1$, and from expressions (2.19) and (2.28) it follows that $|\alpha| \geqq 1 - d\delta$. Further if in (2.29) we let $j$ have the value for which $|q_j^*| = 1$, then from expressions (2.22) and (2.28) we deduce that $|\alpha| \leqq 1 - d\delta$. Thus we obtain the inequality

$$-d\delta \leqq 1 - |\alpha| \leqq d\delta, \tag{2.30}$$

which implies that when $\delta$ is small, then either $\alpha$ is close to $+1$ or it is close to $-1$. We expect $\alpha$ to be positive when $\delta$ is small, because condition (1.2) has to hold. Specifically we deduce from expressions (1.2) and (2.25) the bound

$$0 < Q(x_1) = \bar{Q}(x_1) + \alpha Q^*(x_1), \tag{2.31}$$

and statements (2.5) and (2.28) give the inequality

$$\|\bar{Q}\|_\infty \leqq Md\delta, \tag{2.32}$$

so it follows that $\alpha > -Md\delta/Q^*(x_1)$. This inequality and expression (2.30) imply that the bounds

$$-d\delta \leqq 1 - \alpha \leqq d\delta[3 + 2M/Q^*(x_1)] = \bar{d}\delta, \tag{2.33}$$

say, hold for all $\delta > 0$.

To complete the proof of the lemma, we note that expressions (2.25), (2.32) and (2.33) imply the inequality

$$\begin{aligned} \|Q - Q^*\|_\infty &\leqq \|\bar{Q}\|_\infty + \|(\alpha - 1)Q^*\|_\infty \\ &\leqq Md\delta + \bar{d}\delta\|Q^*\|_\infty, \end{aligned} \tag{2.34}$$

and therefore the truth of the lemma follows from the definition (2.23).

In (2.20) we let $R(x) = R_k(x)$, and we obtain the condition

$$\|R_k - R^*\|_\infty \leqq (\Delta_k - \Delta^*)/\gamma. \tag{2.35}$$

Therefore (2.21) implies the bound

$$\|Q_k - Q^*\|_\infty \leqq \theta(\Delta_k - \Delta^*)/\gamma, \tag{2.36}$$

so, letting $\eta^*$ be the number

$$\eta^* = \min_t Q^*(x_t), \tag{2.37}$$

we deduce the inequalities

$$\min_t \left\{\frac{Q^*(x_t)}{Q_k(x_t)}\right\} \geqq \frac{\eta^*}{\eta^* + (\Delta_k - \Delta^*)\theta/\gamma} \tag{2.38}$$

and

$$\begin{aligned} \min_t \left\{\frac{Q_k(x_t)}{Q_{k+1}(x_t)}\right\} &\geqq \frac{\eta^* - (\Delta_k - \Delta^*)\theta/\gamma}{\eta^* + (\Delta_{k+1} - \Delta^*)\theta/\gamma} \\ &> \frac{\eta^* - (\Delta_k - \Delta^*)\theta/\gamma}{\eta^* + (\Delta_k - \Delta^*)\theta/\gamma}. \end{aligned} \tag{2.39}$$

We now substitute (2.38) in the middle line of (2.7), and obtain from (2.6) the bound

$$\left\{\left|f(x_t) - \frac{P_{k+1}(x_t)}{Q_{k+1}(x_t)}\right| - \Delta_k\right\}\frac{Q_{k+1}(x_t)}{Q_k(x_t)} \leqq \frac{-\eta^*(\Delta_k - \Delta^*)}{\eta^* + (\Delta_k - \Delta^*)\theta/\gamma} \tag{2.40}$$

which holds for $t = 1, 2, \cdots, N$. By multiplying both sides of this expression by $Q_k(x_t)/Q_{k+1}(x_t)$, and by using statements (1.4) and (2.39), we deduce the bound

$$\Delta_{k+1} - \Delta_k \leqq \frac{-\eta^*(\Delta_k - \Delta^*)\{\eta^* - (\Delta_k - \Delta^*)\theta/\gamma\}}{\{\eta^* + (\Delta_k - \Delta^*)\theta/\gamma\}^2} \tag{2.41}$$

from which it follows that the inequality

$$\begin{aligned}(\Delta_{k+1} - \Delta^*) &\leqq (\Delta_k - \Delta^*)\left[1 - \frac{\eta^*\{\eta^* - (\Delta_k - \Delta^*)\theta/\gamma\}}{\{\eta^* + (\Delta_k - \Delta^*)\theta/\gamma\}^2}\right] \\ &= \frac{(\Delta_k - \Delta^*)^2\{3\eta^* + (\Delta_k - \Delta^*)\theta/\gamma\}\theta/\gamma}{\{\eta^* + (\Delta_k - \Delta^*)\theta/\gamma\}^2}\end{aligned} \tag{2.42}$$

is satisfied. Now $\theta$, $\gamma$ and $\eta^*$ are all positive constants, and therefore this expression shows that the rate of convergence of the sequence $\Delta_k$, $k = 1, 2, \cdots$, to $\Delta^*$ is at least quadratic. Theorem 3 is proved.

Note that from the quadratic convergence of the sequence $\Delta_k$, $k = 1, 2, \cdots$, and from inequalities (2.20) and (2.21), it may be proved that, if the conditions of Theorem 3 hold, then the coefficients of the polynomials $P_k(x)$ and $Q_k(x)$ converge quadratically to the coefficients of the polynomials $P^*(x)$ and $Q^*(x)$, in the sense that the differences in the coefficients are bounded by quadratically convergent sequences.

**3. Numerical results.** The implementation of the ODC algorithm is straightforward. We have to calculate the coefficients of $P(x)$ and $Q(x)$ that minimize expression (1.7), and therefore we require the least value of $w$ subject to the constraints

$$\begin{aligned}[f(x_t) + \Delta_k]Q(x_t) - P(x_t) + Q_k(x_t)w &\geqq 0, \qquad t = 1, 2, \cdots, N, \\ [-f(x_t) + \Delta_k]Q(x_t) + P(x_t) + Q_k(x_t)w &\geqq 0, \qquad t = 1, 2, \cdots, N,\end{aligned} \tag{3.1}$$

and

$$\max_j |q_j| = 1, \tag{3.2}$$

where the coefficients of $P(x)$ and $Q(x)$ are variables. Since the constraints (3.1) and (3.2) are linear in $w$, $p_i$, $i = 0, 1, \cdots, m$, and $q_j$, $j = 0, 1, \cdots, n$, linear programming methods are applicable.

In practice it is more convenient to replace (3.2) by the inequality

$$-1 \leqq q_j \leqq 1, \qquad j = 0, 1, \cdots, n. \tag{3.3}$$

Expressions (3.2) and (3.3) are equivalent in the present calculation, because the inequalities (3.1) are homogeneous in the variables. Moreover it is more convenient in practice to solve the dual of this linear programming problem.

To compare the ODC and the DC versions of the differential correction algorithm, six sets of data are used, each set being a table of values of a function $f(x)$. These functions and data points are specified in Table 1. The data is fitted by the rational polynomials $P_1(x)/Q_1(x)$, $P_2(x)/Q_2(x)$, $P_1(x)/Q_3(x)$ and $P_4(x)/Q_2(x)$, where now the subscripts denote the degrees of the polynomials. Thus twenty-four different approximations are calculated.

TABLE 1
*The six data sets for the numerical examples*

| Function $f(x)$ | Abscissas $\{x_1, x_2, \cdots, x_N\}$ | Number of points $N$ |
|---|---|---|
| A. $e^x$ | $-1.0(0.1)1.0$ | 21 |
| B. $\sin x$ | $-3.0(0.3)3.0$ | 21 |
| C. $\sqrt{x}$ | $0.0(0.05)1.0$ | 21 |
| D. $\begin{cases} 1 \\ 0 \\ -1 \end{cases}$ | $\left.\begin{matrix} 0.0(0.05)0.45 \\ 0.5 \\ 0.55(0.05)1.0 \end{matrix}\right\}$ | 21 |
| E. $\begin{cases} x \\ 0.5x + 0.4 \end{cases}$ | $\left.\begin{matrix} 0.0(0.1)1.0 \\ 1.1(0.1)2.0 \end{matrix}\right\}$ | 21 |
| F. $\cos(\frac{1}{2}x)$ | $\left\{\begin{matrix} 0.0(1/7)6/7 \\ 1.0(0.2)1.8 \\ 2.0(1/8)3.0 \end{matrix}\right\}$ | 21 |

To begin the iterations the starting values $p_0 = q_0 = 1$ and $p_i = q_j = 0$, $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, n$, are used. However, to calculate $P_2(x)/Q_2(x)$ two other sets of starting coefficients are also used. In one set $q_0 = 1$ and the remaining coefficients are zero, and in the other set $q_0 = 1$ and the remaining coefficients are obtained by a random number generator. The different starting approximations do not influence the final approximation, but they do affect the number of iterations that are required by the ODC and DC algorithms. (Consequently, in practice we would normally supply either algorithm with starting values based on good initial rational approximations.)

The number of iterations is reported in Tables 2 and 3. For both algorithms the test used to terminate the iterative process is the inequality

$$\Delta_k - \Delta_{k+1} < 10^{-4}\Delta_k. \tag{3.4}$$

These tables also give the coefficients of the best approximations and the minimax errors $\Delta^*$ calculated by the ODC algorithm. All the computations were performed in double precision arithmetic on an IBM 360/44.

The numerical examples do confirm that the final convergence rate of the ODC algorithm is very rapid, and in fact we tried changing the constant in inequality (3.4) to $10^{-6}$, and there was no increase in the number of iterations required by the ODC algorithm. However, this fast rate of convergence is not shared by the DC algorithm, and we note that Tables 2 and 3 show that in nearly every case the DC algorithm requires substantially more iterations than the ODC algorithm.

TABLE 2

*Minimax approximations by $P_1(x)/Q_1(x)$ and $P_2(x)/Q_2(x)$*

| | Data Sets | | | | | |
|---|---|---|---|---|---|---|
| | A | B | C | D | E | F |
| $P_1(x)/Q_1(x)$ | | | | | | |
| $p_0$ | 1.01705 | 0.00000 | 0.01814 | 1.81818 | −0.05874 | 1.03061 |
| $p_1$ | 0.51756 | 0.25551 | 1.34288 | −3.63636 | 1.57292 | −0.33749 |
| $q_0$ | 1.00000 | 1.00000 | 0.42214 | 1.00000 | 1.00000 | 1.00000 |
| $q_1$ | −0.43977 | 0.00000 | 1.00000 | 0.00000 | 0.60867 | −0.18266 |
| $\Delta^*$ | $(2.09541)10^{-2}$ | $(6.25422)10^{-1}$ | $(4.29721)10^{-2}$ | $(8.18182)10^{-1}$ | $(5.87394)10^{-2}$ | $(3.06115)10^{-2}$ |
| Number of iterations ODC | 6 | 2 | 6 | 5 | 5 | 8 |
| Number of iterations DC | 15 | 2 | 23 | 3 | 8 | 17 |
| $P_2(x)/Q_2(x)$ | | | | | | |
| $p_0$ | 1.00007 | 0.00000 | 0.00007 | 0.20096 | 0.00239 | 1.00015 |
| $p_1$ | 0.50840 | 1.43537 | 0.32728 | −0.40192 | −0.15684 | −0.04028 |
| $p_2$ | 0.08571 | 0.00000 | 1.24475 | 0.00000 | 2.03176 | −0.08859 |
| $q_0$ | 1.00000 | 0.62909 | 0.03683 | 0.27500 | 0.04407 | 1.00000 |
| $q_1$ | −0.49133 | 0.00000 | 1.00000 | −1.00000 | 1.00000 | −0.03792 |
| $q_2$ | 0.07781 | 1.00000 | 0.53831 | 1.00000 | 0.94090 | 0.03056 |
| $\Delta^*$ | $(8.47766)10^{-5}$ | $(3.06078)10^{-1}$ | $(1.92938)10^{-3}$ | $(2.69231)10^{-1}$ | $(5.42353)10^{-2}$ | $(1.51135)10^{-4}$ |
| Number of iterations ODC | 10/10/11 | 7/6/7 | 10/10/10 | 9/9/9 | 5/6/4 | 10/10/10 |
| Number of iterations DC | 23/24/23 | 19/17/18 | > 50/> 50/> 50 | 45/43/49 | 27/23/20 | 17/17/14 |

TABLE 3

*Minimax approximations by $P_1(x)/Q_3(x)$ and $P_4(x)/Q_2(x)$*

| | Data Sets | | | | | |
|---|---|---|---|---|---|---|
| | A | B | C | D | E | F |
| $P_1(x)/Q_3(x)$ | | | | | | |
| $p_0$ | 0.99988 | 0.00000 | 0.00066 | 0.20096 | −0.01438 | 1.00077 |
| $p_1$ | 0.25359 | 1.43537 | 0.56679 | −0.40192 | 0.47971 | −0.31922 |
| $q_0$ | 1.00000 | 0.62909 | 0.08656 | 0.27500 | 0.67233 | 1.00000 |
| $q_1$ | −0.74661 | 0.00000 | 1.00000 | −1.00000 | −0.90994 | −0.30747 |
| $q_2$ | 0.24520 | 1.00000 | −0.93446 | 1.00000 | 1.00000 | 0.09520 |
| $q_3$ | −0.03749 | 0.00000 | 0.41972 | 0.00000 | −0.27484 | −0.01179 |
| $\Delta^*$ | $(1.22371)10^{-4}$ | $(3.06078)10^{-1}$ | $(7.63026)10^{-3}$ | $(2.69231)10^{-1}$ | $(4.55729)10^{-2}$ | $(7.73421)10^{-4}$ |
| Number of iterations ODC | 8 | 8 | 8 | 10 | 9 | 12 |
| Number of iterations DC | 35 | 18 | 37 | 49 | 13 | 18 |
| $P_4(x)/Q_2(x)$ | | | | | | |
| $p_0$ | 1.00000 | 0.00000 | 0.00000 | 0.27425 | 0.00316 | 1.00000 |
| $p_1$ | 0.67030 | 1.01770 | 0.06422 | −1.33004 | 0.53075 | −0.01197 |
| $p_2$ | 0.20262 | 0.00000 | 0.79795 | 2.34464 | −1.12906 | −0.11473 |
| $p_3$ | 0.03412 | −0.10444 | 0.53757 | −1.56309 | 0.68648 | 0.00158 |
| $p_4$ | 0.00286 | 0.00000 | −0.07241 | 0.00000 | −0.08390 | 0.00124 |
| $q_0$ | 1.00000 | 1.00000 | 0.00497 | 0.25619 | 0.50919 | 1.00000 |
| $q_1$ | −0.32970 | 0.00000 | 0.32243 | −1.00000 | −1.00000 | −0.01198 |
| $q_2$ | 0.03231 | 0.08155 | 1.00000 | 1.00000 | 0.49832 | 0.01031 |
| $\Delta^*$ | $(2.04651)10^{-7}$ | $(6.64822)10^{-3}$ | $(6.36423)10^{-5}$ | $(7.04653)10^{-2}$ | $(1.11768)10^{-2}$ | $(2.64760)10^{-7}$ |
| Number of iterations ODC | 10 | 9 | 12 | 11 | 13 | 11 |
| Number of iterations DC | 20 | 20 | > 50 | > 50 | > 50 | 46 |

We have checked the equioscillation property that characterizes the error of the minimax approximation (Cheney [1], for instance), and we find that for the results obtained by the ODC method, equioscillation is present to ten or more significant figures. However on the average we find about six figures of accuracy in the errors calculated by the DC algorithm, due to the fact that the rate of convergence of the DC algorithm is only linear.

Finally we note that both versions of the differential correction algorithm can be used to calculate rational approximating functions that are more general than the ratio of two algebraic polynomials. It is only necessary that the adjustable parameters appear linearly in the numerator and denominator of the approximating function, and therefore we can approximate functions of several variables.

For example, Fox, Goldstein and Lastman [5] calculate the coefficients $p_i$, $i = 0, 1, \cdots, 5$, and $q_j$, $j = 0, 1, \cdots, 5$, that minimize the error of the approximation

$$\text{Re}\left\{\int_{x+iy}^{\infty} \frac{e^{-z}}{z}\,dz\right\} \approx \frac{p_0 + p_1 y + p_2 x + p_3 xy + p_4 y^2 + p_5 x^2}{q_0 + q_1 y + q_2 x + q_3 xy + q_4 y^2 + q_5 x^2} \tag{3.5}$$

on a set of twenty-five discrete data points. We applied the ODC algorithm to this problem, using the starting approximation and convergence criterion specified above, and we found convergence in ten iterations. For our calculated best approximation the equioscillation property holds to eleven significant figures.

The theory and numerical examples given in this paper show that the differential correction algorithm, *in its original form*, is an excellent method for computing discrete minimax rational approximations.

**Acknowledgments.** It is a pleasure to acknowledge the programming assistance of Mr. K. B. Wilson of Victoria. Also we are grateful to Dr. J. K. Reid, for he studied a draft of this paper and suggested many improvements to the manuscript.

REFERENCES

[1] E. W. CHENEY, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.
[2] E. W. CHENEY AND H. L. LOEB, *Two new algorithms for rational approximation*, Numer. Math., 3 (1961), pp. 72–75.
[3] ———, *On rational Chebyshev approximation*, Ibid., 4 (1962), pp. 124–127.
[4] E. W. CHENEY AND T. H. SOUTHARD, *A survey of methods for rational approximation, with particular reference to a new method based on a formula of Darboux*, SIAM Rev., 5 (1963), pp. 219–231.
[5] P. FOX, A. A. GOLDSTEIN AND G. LASTMAN, *Rational approximations on finite point sets*, Approximation of Functions, H. L. Garabedian, ed., Elsevier, Amsterdam, 1965, pp. 56–67.
[6] J. R. RICE, *The Approximation of Functions*, vol. II, Addison-Wesley, Reading, Mass., 1969.