JOURNAL OF APPROXIMATION THEORY 37, 197–211 (1983)

# An Adaptive Differential-Correction Algorithm*

E. H. KAUFMAN, JR.

*Central Michigan University, Mt. Pleasant, Michigan 48859, U.S.A.*

S. F. MCCORMICK

AND

G. D. TAYLOR

*Colorado State University, Fort Collins, Colorado 80523, U.S.A.*

*Communicated by E. W. Cheney*

Received May 19, 1981

A combined first Remes differential-correction algorithm for uniform generalized rational approximation with restricted range constraints is presented. This algorithm can be applied when the data sets are too large to allow for the direct use of differential correction and when the second Remes algorithm does not apply because of the lack of an alternating theory. Under the assumption that differential correction produces a good (though not necessarily best) approximation on each (small) subset to which it is applied, it is proven that the algorithm terminates in a finite number of steps at a good approximation on the entire data set. This is established even though, unlike the standard first Remes, the algorithm sometimes discards points in passing from one subset to the next. This theory also allows for the set to be infinite. Also, a discretization theorem is presented and the algorithm is illustrated with a numerical example.

## 1. INTRODUCTION

The differential-correction algorithm introduced by Cheney and Loeb [4], shown to possess desirable global convergence properties by Barrodale *et al.* [1], and touted as the method of choice by Lee and Roberts [14], enjoys widespread use today. This algorithm uses a linear programming approach to calculate a best uniform rational fit to given values on a finite set. Various

* Supported in part by the National Science Foundation under Grant MCS-80-17056 and the Air Force Office of Scientific Research, Air Force System Command, USAF, under Contract F-49620-79-C-0124.



0021-9045/83 $3.00

codings of this algorithm are available [2, 7, 8] and, recently, a more robust version [13] has been given that allows for the inclusion of a multiplicative weight function and restrictions on the values of the approximating functions (i.e., restricted range constraints). The weight function and restricted-range features are useful, for example, when one wishes to approximate the magnitude squared response of a digital filter (see [5, 9, 15, 16]). The code in [13] also contains subroutines for combining differential correction with the second algorithm of Remes. This combination has proved quite effective in situations where it can be applied (primarily ordinary rational approximation of functions of one variable (see [10, 11, 13])). In the present paper, the algorithm of [13] is combined with an adaptive exchange procedure to extend the domain of application of the differential-correction code to very large (possibly infinite) data sets. This exchange procedure, which is similar to the first algorithm of Remes but allows points to be dropped sometimes and does not always require bringing in the point of maximum error, allows for savings in storage and time requirements. (Dunham [6] has considered rational approximation using a more standard version of the Remes algorithm.) These improvements are due to the fact that the major cost of the differential-correction algorithm in terms of storage and speed is in its linear programming subroutine with these costs increasing rapidly with the size of the data set. Also, this particular subroutine is that part of the procedure where failure, though infrequent, most often occurs. Since our adaptive strategy uses the differential-correction algorithm on (small) subsets of the full data set, the time needed for this subroutine to run is decreased and its chances of a successful run are improved. More importantly, it allows for the solution of problems which are even too large for differential correction alone to be applied; for example, if one wished to ue straight differential correction on a grid formed by subdividing $[0, 1] \times [0, 1] \times [0, 1]$ with spacing 0.01 in each direction, one would have to solve a sequence of nonsparse linear programming problems with more than 2,000,000 constraints!

In what follows, we shall describe our adaptive application of the differential-correction algorithm and prove the convergence of this procedure. The convergence proof is somewhat unusual in that we show that if, at each step of the algorithm, an acceptable approximation is found that comes within some fxed distance of being best on its particular subset of the data, then the algorithm will converge in a finite number of steps to an approximation which comes within some fixed tolerance of being best for the full data set. This sort of a convergence result is intuitively appealing since, in practice, the differential-correction algorithm will only calculate a rational approximation which has error of approximation close to the minimum error norm.

## 2. Notation and Description of the Algorithm

Let $X$ be a finite set of points in $\mathbb{R}^k$ and let $f: X \to \mathbb{R}$ be a given real-valued function. Thus, if we are given a data set $\{(x_i, y_i)\}_{i=1}^M$ with $x_i \in \mathbb{R}^k$ and $y_i \in \mathbb{R}$ for all $i$, we set $X = \{x_i\}_{i=1}^M$ and define $f: X \to \mathbb{R}$ by $f(x_i) = y_i$ for all $i$. Furthermore, let $G$, $L$, and $U$ be fixed subsets of $X$ with $X = G \cup L \cup U$, let $W: X \to \mathbb{R}$ be a given positive real-valued (multiplicative weight) function and let $l: L \to \mathbb{R}$ and $u: U \to \mathbb{R}$ be given real-valued functions satisfying $l(x) < u(x)$ for all $x \in L \cap U$. Finally, let $\mathscr{P} = \langle \phi_1, \ldots, \phi_n \rangle$ and $\mathscr{Q} = \langle \psi_1, \ldots, \psi_m \rangle$ be two finite-dimensional linear subspaces (dimensions $n$ and $m$, respectively) of real-valued functions defined on $X$. Then the class of generalized rational functions on $X$ with respect to $\mathscr{P}$ and $\mathscr{Q}$ is defined to be

$$\mathscr{R} = \Bigg\{ R = P/Q : P = \sum_{i=1}^{n} p_i \phi_i,\ Q = \sum_{i=1}^{m} q_i \psi_i,\ Q(x) \geqslant \eta \text{ for all } x \in X \text{ and } |q_i| \leqslant 1 \text{ for all } i \text{ with equality holding at least once} \Bigg\}.$$

Here $\eta$ is a small positive number ($\eta = 10^{-11}$ in our code on a CYBER 172, which has roughly 15 digits of accuracy in single precision). We use the denominator restriction $Q \geqslant \eta$ instead of the usual weaker restriction $Q > 0$ for several reasons. First, rational approximations with very small denominators tend not to be useful in applications. Second, imposing the condition $Q \geqslant \eta$ on the small subsets of $X$ on which the differential correction algorithm is applied does not require much extra computing time and decreases the possibility of failure; it turns out that requiring this condition on the small subsets plus some mild hypotheses guarantees that the condition will be satisfied on all of $X$ when our algorithm terminates. Finally, this condition is needed to prove our convergence and discretization results. In most examples, if $\eta$ is small enough, the constraint $Q \geqslant \eta$ will never actually come into play. Hence, our algorithm could in fact be implemented without this condition. The best uniform restricted range generalized rational approximation problem is to determine

$$\rho \equiv \inf \{ \max_{x \in G} |W(x)(f(x) - R(x))| : R \in \mathscr{R},\ R(x) \geqslant l(x),\ \forall x \in L \text{ and } R(x) \leqslant u(x),\ \forall x \in U \}.$$

In what follows, we assume that there exists an $R \in \mathscr{R}$ satisfying $R(x) \geqslant l(x)$ $\forall x \in L$ and $R(x) \leqslant u(x)$ $\forall x \in U$.

In order to describe our algorithm, we need to extend the above notation

to subsets of $X$. Thus, let $X_k$ be a subset of $X$ and set $G_k = X_k \cap G$, $L_k = X_k \cap L$, and $U_k = X_k \cap U$. Further, set

$$\mathscr{R}_k = \Big\{ R = P/Q : P = \sum_{i=1}^{n} p_i \phi_i,\ Q = \sum_{i=1}^{m} q_i \psi_i,\ Q(x) \geqslant n \text{ for all } x \in X_k$$
$$\text{and } |q_i| \leqslant 1 \text{ for all } i \text{ with equality holding at least once} \Big\},$$

and

$$\rho_k = \inf \{ \max_{x \in G_k} |W(x)(f(x) - R(x))| : R \in \mathscr{R}_k, R(x) \geqslant l(x), \forall x \in L_k$$
$$\text{and } R(x) \leqslant u(x), \forall x \in U_k \}.$$

For any $R_k = P_k/Q_k \in \mathbb{R}_k$ we define

$$\Delta_k = \max_{x \in G_k} |W(x)(f(x) - R_k(x))|.$$

Our method of handling constraints is to include them in the error function (a similar strategy is used in [17]). In addition, points where the denominator is very small in absolute value or negative are handled by assigning to them a large error. To be precise, let $x$ be an arbitrary point in $X$ and define the error $e_k(x)$ to be $(\Delta_k + 1)\, 10^6$ if $Q_k(x) < \eta$. For $Q_k(x) \geqslant \eta$ define

$$\begin{aligned} e_k^G(x) &= W(x)(f(x) - R_k(x)), && x \in G, \\ &= 0, && x \notin G; \\ e_k^L &= l(x) + \Delta_k - R_k(x), && x \in L, \\ &= 0, && x \notin L; \\ e_k^U(x) &= u(x) - \Delta_k - R_k(x), && x \in U, \\ &= 0, && x \notin U. \end{aligned}$$

Finally, take $e_k(x)$ to be $e_k^G(x)$, $e_k^L(x)$ or $e_k^U(x)$, or $e_k^U(x)$ according to if $|e_k^G(x)|$, $e_k^L(x)$, or $-e_k^U(x)$ is largest, with the first chosen in case of a tie. Observe that an error $e_k(x)$ arising from a constraint is greater than $\Delta_k$ in absolute value iff the constraint is violated; observe also that $e_k(x)$ could depend discontinuously on $R_k(x)$, but if $R_k(x)$ is a point where such a discontinuity occurs and $f(x)$ does not itself violate a constraint, then $|e_k(x)| < \Delta_k$. Such points will have no effect on our algorithm. Finally, we set $E_k(x) = |e_k(x)|$, and define the set of "extreme points" on $X_k$ by $T_k(R_k) = \{x \in X_k : E_k(x) \geqslant \Delta_k - TOL\}$, where $TOL$ is a small positive number ($TOL = 10^{-8}$ in our code).

We can now describe the algorithm. Initially, let $X_0$ be a subset of $X$ containing at least $m+n$ points. Apply the differential-correction algorithm [13] to compute $R_0 \in \mathscr{R}_0$ that satisfies

$$\rho_0 - \delta_1 \leqslant \Delta_0 \equiv \max_{x \in G_0} |W(x)(f(x) - R_0(x))| \leqslant \rho_0 + \delta,$$

$$R_0(x) \geqslant l(x) - \delta_1, \quad \forall x \in L_0, \qquad R_0(x) \leqslant u(x) + \delta_1, \quad \forall x \in U_0,$$

where $\delta$ and $\delta_1$ are small positive constants. Theoretically, the algorithm in [13] will produce an approximation $R_0$ for which $\delta_1 = 0$, but round-off error may allow a small violation of the constraints, which in turn may allow $\Delta_0$ to be slightly smaller than $\rho_0$. On the other hand, $\Delta_0$ is normally larger than $\rho_0$ even without round-off error because the differential-correction algorithm usually must be terminated before a best approximation is found, even if one exists. Thus $\delta$ and $\delta_1$ are measures of how far $R_0$ can deviate from being best on $X_0$.

The next step is to construct a set $A_0 \subseteq X$ which contains at least one $x \in X$ with $E_0(x) > \Delta_0 + \beta$ ($\beta = 10^{-11}$ in our code); if no such $x$ exists, the algorithm is terminated and $R_0$ is accepted as the "best" approximation on $X$. The exact choice of $A_0$ does not affect the proof of convergence but does affect the convergence rate; the procedure is to search for relative extrema of $E_0(x)$ in $X$ by "walking uphill" from each point of $T_0(R_0)$ in turn. This procedure requires that $X$ be a cross product of finite sets of real numbers. (As we show below, the more general case of scattered data can be reduced to this special case quite easily.) To walk from a point $x_0 \in T_0(R_0)$, we examine all neighbors of $x_0$ (which are defined as points whose corresponding indices differ from those of $x_0$ by at most one) to find a "feasible" direction, that is, we seek a point $x_1$ where $e_0(x_1)\, e_0(x_0) > 0$, $E_0(x_1) > E_0(x_0)$, and $E_0(x_1)$ is maximized. If such a point is found, the walk continues in this direction as long as $e_0(x)$ does not change sign and $E_0(x)$ increases. When a point is reached that prevents further such progress in this direction, a new feasible direction is sought by examining all neighbors of this last point. Eventually, this process terminates at a relative extremum of $e_0(x)$. If for at least one such relative extremum we have $E_0(x) > \Delta_0 + \beta$, define $A_0$ to be the set of all these relative extrema, together with up to $2k$ other points turned up in the searching where $E_0(x)$ is largest. (We, however, exclude any points where $E_0(x) < \Delta_0\text{-}TOL$). If no extremum satisfies $E_0(x) > \Delta_0 + \beta$, a two-stage scan is performed on all points of $X$ (a coarse scan followed by a scan of the remaining points). If no $x$ is found with $E_0(x) > \Delta_0 + \beta$, the algorithm is terminated. If such an $x$ is found, we reconstruct $A_0$ by walking uphill from the single point $x$.

Assuming the algorithm does not terminate, define $X_1 = T_0(R_0) \cup A_0$.

Differential correction is then used to compute an approximation $R_1 \in \mathscr{R}_1$ satisfying

$$\rho_1 - \delta_1 \leqslant \Delta_1 = \max_{x \in G_1} |W(x)(f(x) - R_1(x))| \leqslant \rho_1 + \delta,$$

$$R_1(x) \geqslant l(x) - \delta_1, \quad \forall x \in L_1, \qquad R_1(x) \leqslant u(x) + \delta_1, \quad \forall x \in U_1.$$

We are now ready to describe the general exchange procedure of the algorithm. Suppose, for $k \geqslant 2$, that sets $X_{k-2}$, $X_{k-1}$ have been found with corresponding rational approximations $R_{k-2} \in \mathscr{R}_{k-2}$ and $R_{k-1} \in \mathscr{R}_{k-1}$ and with computed errors $\Delta_{k-2}$ and $\Delta_{k-1}$, respectively. We also assume $\Delta_{k-3}$ has been computed from $R_{k-3}$ on $X_{k-3}$ unless $k = 2$, in which case we set $\Delta_{-1} = 0$. Then the set $S_{k-1}$ is found where

$$\begin{aligned} S_{k-1} &= X_{k-1}, && \Delta_{k-1} \leqslant \max(\Delta_{k-2}, \Delta_{k-3}) + \beta, \\ &= T_{k-1}(R_{k-1}), && \text{otherwise.} \end{aligned}$$

The set $A_{k-1}$ is constructed in the same way as $A_0$ was, and if the algorithm does not terminate we define

$$\begin{aligned} X_k &= S_{k-1} \cup A_{k-1}, && \Delta_{k-1} \geqslant \Delta_{k-2}, \\ &= S_{k-1} \cup A_{k-1} \cup X_{k-2}, && \text{otherwise.} \end{aligned}$$

Intuitively, we allow ourselves to drop the nonextreme points from $X_{k-1}$ if $\Delta_{k-1}$ has increased significantly over both $\Delta_{k-2}$ and $\Delta_{k-3}$, while we put in the points from $X_{k-2}$ if $\Delta_{k-1}$ is actually smaller than $\Delta_{k-2}$. We then continue in this fashion. In the next section we shall show that under certain mild assumptions the algorithm must eventually terminate.

In practice, we have found time is saved by first computing a nested sequence of subsets of $X$ by essentially removing at each stage alternate points in each direction (with boundary points of $X$ immune to removal); the smallest subset containing at least $m + n$ points is taken to be $X_0$. The algorithm as described above is then run to get an approximation on the next larger subset (which is playing the role of $X$). The final $X_j$ becomes the $X_0$ for the net larger subset, and so forth until an approximation has been computed on $X$. Our code does this automatically. For example, if $m + n = 5$ and $X$ is a $5 \times 201$ grid, the sequence of subsets has dimensions $2 \times 3 \to 2 \times 5 \to 2 \times 8 \to 2 \times 14 \to 2 \times 26 \to 2 \times 51 \to 3 \times 101 \to 5 \times 201$.

If $X$ is not a cross product of finite sets of real numbers (i.e., we have scattered data), we put $k$-dimensional boxes about the points of $X$ with at most one point per box, then treat the boxes like points (the main algorithm will ignore empty boxes). The boxing procedure involves initially placing borders between each two adjacent points in each direction, then successively

removing borders (in increasing order of the directional separation distance between points) whenever this can be done without putting two points in the same box. The idea is that, as far as possible, points which are close together in some direction should be on the same level in that direction. As an example, if $X = \{(0.9, 1.1), (0, 0.8), (1.1, 0.7), (1.9, 0), (0.4, 1.7)\}$, then the result will be two rows of three boxes each. If the boxes are numbered from left to right and bottom to top, the first point listed above will be in box 5, the second point will be in box 1, the third point will be in box 2, the fourth point will be in box 3, the fifth point will be in box 4, and box 6 will be empty.

## 3. Convergence and Related Results

In this section, convergence results are developed for the algorithm defined above. In addition, some discretization-type results are noted. Because of the different exchange procedures employed, the following notation and lemma are needed.

The notation $X_k \downarrow X_{k+1}$ shall mean that $\Delta_k > \max(\Delta_{k-1}, \Delta_{k-2}) + \beta$ has occurred; that is, the nonextreme points of $X_k$ have been discarded. Also, the notation $X_k \to X_{k+1}$ means that $\Delta_k \leqslant \max(\Delta_{k-1}, \Delta_{k-2}) + \beta$ has occurred; note that $X_k \subseteq X_{k+1}$ in this case.

Lemma 1. *Suppose* $\varepsilon = \beta - \delta - \delta_1 > 0$ *and let* $3 \leqslant k_1 < k_2 \cdots < k_n$ *be an increasing sequence of positive integers for which*

$$X_{k_1-1} \downarrow X_{k_1} \to X_{k_1+1} \to \cdots \to X_{k_2-1} \downarrow X_{k_2} \to \cdots \to X_{k_n-1} \downarrow X_{k_n} \to X_{k_n+1}.$$

*Then* $\rho_{k_n+1} \geqslant \max(\rho_{k_1-2}, \rho_{k_1-3}) + n\varepsilon$.

*Proof.* The proof is by induction. Suppose $n = 1$ so that $X_{k_1-1} \downarrow X_{k_1} \to X_{k_1+1}$. We know then that $\Delta_{k_1-1} > \max(\Delta_{k_1-2}, \Delta_{k_1-3}) + \beta$ as $X_{k_1-1} \downarrow X_{k_1}$. Now, if $\Delta_{k_1} < \Delta_{k_1-1}$, then $X_{k_1-1} \subseteq X_{k_1+1}$ by construction, so that

$$\begin{aligned}\rho_{k_1+1} &\geqslant \rho_{k_1-1} \geqslant \Delta_{k_1-1} - \delta > \max(\Delta_{k_1-2}, \Delta_{k_1-3}) + \beta - \delta \\ &\geqslant \max(\rho_{k_1-2}, \rho_{k_1-3}) + \varepsilon.\end{aligned}$$

On the other hand, if $\Delta_{k_1} \geqslant \Delta_{k_1-1}$, then since $X_{k_1} \subseteq X_{k_1+1}$ as $X_{k_1} \to X_{k_1+1}$, we have that

$$\begin{aligned}\rho_{k_1+1} &\geqslant \rho_{k_1} \geqslant \Delta_{k_1} - \delta \geqslant \Delta_{k_1-1} - \delta > \max(\Delta_{k_1-2}, \Delta_{k_1-3}) + \beta - \delta \\ &\geqslant \max(\rho_{k_1-2}, \rho_{k_1-3}) + \varepsilon.\end{aligned}$$

So, in either case, $\rho_{k_1+1} \geqslant \max(\rho_{k_1-2}, \rho_{k_1-3}) + 1 \cdot \varepsilon$ as desired.

Now suppose the lemma is true for $n \leqslant l$, and suppose

$$X_{k_1-1} \downarrow X_{k_1} \to \cdots \to X_{k_2-1} \downarrow X_{k_2} \to \cdots \to X_{k_{l+1}-1} \downarrow X_{k_{l+1}} \to X_{k_{l+1}+1}.$$

The proof involves three cases.

*Case* 1. $k_2 = k_1 + 1$, so that $X_{k_1-1} \downarrow X_{k_2-1} \downarrow X_{k_2}$. By the induction assumption, we have that $\rho_{k_{l+1}+1} \geqslant \max(\rho_{k_2-2}, \rho_{k_2-3}) + l\varepsilon$ so that

$$\begin{aligned}\rho_{k_{l+1}+1} &\geqslant \rho_{k_2-2} + l\varepsilon = \rho_{k_1-1} + l\varepsilon \geqslant \Delta_{k_1-1} - \delta + l\varepsilon \\ &> \max(\Delta_{k_1-2}, \Delta_{k_1-3}) + \beta - \delta + l\varepsilon \\ &\geqslant \max(\rho_{k_1-2}, \rho_{k_1-3}) + (l+1)\,\varepsilon.\end{aligned}$$

*Case* 2. $k_2 = k_1 + 2$ so that $X_{k_1-1} \downarrow X_{k_1} \to X_{k_2-1} \downarrow X_{k_2}$. Then by the induction assumption, we have $\rho_{k_{l+1}+1} \geqslant \max(\rho_{k_2-2}, \rho_{k_2-3}) + l\varepsilon$ so that

$$\begin{aligned}\rho_{k_{l+1}+1} &\geqslant \rho_{k_2-3} + l\varepsilon = \rho_{k_1-1} + l\varepsilon \geqslant \Delta_{k_1-1} - \delta + l\varepsilon \\ &> \max(\Delta_{k_1-2}, \Delta_{k_1-3}) + \beta - \delta + l\varepsilon \\ &\geqslant \max(\rho_{k_1-2}, \rho_{k_1-3}) + (l+1)\,\varepsilon.\end{aligned}$$

*Case* 3. $k_2 > k_1 + 2$, so that $X_{k_1-1} \downarrow X_{k_1} \to X_{k_1+1} \to \cdots \to X_{k_2-1} \downarrow X_{k_2}$. Then, by the induction assumption once again, we have that $\rho_{k_{l+1}+1} \geqslant \max(\rho_{k_2-2}, \rho_{k_2-3}) + l\varepsilon$. In addition, since $X_{k_1+1} \subseteq X_{k_2-2}$ implies that $\rho_{k_2-2} \geqslant \rho_{k_1+1}$, we have that

$$\begin{aligned}\rho_{k_{l+1}+1} &\geqslant \rho_{k_2-2} + l\varepsilon \geqslant \rho_{k_1+1} + l\varepsilon \geqslant \max(\rho_{k_1-2}, \rho_{k_1-3}) + \varepsilon + l\varepsilon \\ &= \max(\rho_{k_1-2}, \rho_{k_1-3}) + (l+1)\,\varepsilon,\end{aligned}$$

where the last inequality is simply the result of the $n = 1$ case. The three cases are thus established and the lemma is proved. ∎

THEOREM 2. *Assume that $X$ is finite and that each application of the differential-correction algorithm produces an approximation $R_k$ in $\mathscr{R}_k$ that satisfies*

$$\rho_k - \delta_1 \leqslant \Delta_k \equiv \max_{x \in G_k} |W(x)(f(x) - R_k(x))| \leqslant \rho_k + \delta,$$

$$R_k(x) \geqslant l(x) - \delta_1, \quad \forall x \in L_k, \qquad R_k(x) \leqslant u(x) + \delta_1, \quad \forall x \in U_k,$$

*where $\varepsilon = \beta - \delta - \delta_1 > 0$ and $\beta < 10^6$. Then the algorithm terminates at an $R^* \in \mathscr{R}$ satisfying*

$$\Delta \equiv \max_{x \in G} |W(x)(f(x) - R^*(x))| \leqslant \rho + \beta + \delta,$$

$$R^*(x) \geqslant l(x) - \beta, \quad \forall x \in L, \qquad R^*(x) \leqslant u(x) + \beta, \quad \forall x \in U.$$

*Proof.* First observe that $\rho_k \leqslant \rho$ for all $k$ and $\rho$ finite (since we have assumed the existence of an $R \in \mathscr{R}$ satisfying the constraints) implies that an exchange of the form $X_k \downarrow X_{k+1}$ can occur only a finite number of times by Lemma 1. Thus, once the algorithm is past the last index at which this sort of exchange occurs, then we must have $X_k \subseteq X_{k+1}$ if the algorithm does not terminate at stage $k$. This containment must be proper, since by construction there exists an $x \in A_k \subseteq X_{k+1}$ with $E_k(x) > \Delta_k + \beta > \Delta_k + \delta_1$, and this together with $Q_k \geqslant \eta$ on $X_k$ implies $x \notin X_k$. Since $X$ is finite, it follows that the algorithm will eventually terminate in a finite number of steps. If $X_N$ is the final subset of the algorithm and $R_N$ is the approximation computed on it, then we first observe that $Q_N \geqslant \eta$ on $X$ so that $R_N \in \mathscr{R}$. If not, then for some $x \in X$ we have $Q_N(x) < \eta$, but then $E_N(x) = (\Delta_N + 1)\, 10^6$ by construction so $(\Delta_N + 1)\, 10^6 \leqslant \Delta_N + \beta$ (else the algorithm would not have terminated). Thus, $\Delta_N 10^6 + 10^6 \leqslant \Delta_N + \beta < \Delta_N + 10^6$, so $\Delta_N(10^6 - 1) < 0$, which is a contradiction. We also have

$$\max_{x \in X} |W(x)(f(x) - R_N(x))| \leqslant \max_{x \in X} E_N(x) \leqslant \Delta_N + \beta$$

$$\leqslant \rho_N + \delta + \beta \leqslant \rho + \beta + \delta.$$

Finally, if for some $x \in X$ we had $R_N(x) < l(x) - \beta$ or $R_N(x) > u(x) + \beta$, then by definition $E_N(x) > \Delta_N + \beta$, which contradicts the second inequality above. ∎

The assumption that $X$ be finite can be weakened to assume more generally that $X$ and each $X_k$ be only compact, although to implement the algorithm one would want each $X_k$ to be finite. To insure termination of the algorithm and leave some room for errors due to incomplete searching on an infinite set, we need a stronger restriction for $Q_k$ on $X_k$. We thus define $\mathscr{R}^*$, $\rho^*$, $\mathscr{R}_k^*$, and $\rho_k^*$ by replacing $\eta$ by $4\eta$ in the definitions of $\mathscr{R}$, $\rho$, $\mathscr{R}_k$, and $\rho_k$, respectively. We also assume that an element of $\mathscr{R}^*$ satisfying the constraints exists, and define $E_k^*(x)$ by replacing $\eta$ by $2\eta$ in the definition of $E_k(x)$. As before, we assume that $A_k$ is constructed to contain a point $x$, where $E_k^*(x) > \Delta_k + \beta$, and the algorithm is terminated if no such $x$ is found. Note that an alternate search procedure for actually constructing $A_k$ must be divised. For example, a procedure involving Newton's method is suggested in [18] for linear (i.e., $m = 1$) approximation on a rectangle using a more standard Remes first algorithm where points are never dropped. Assuming we have some kind of search procedure, we can prove the following theorem,

which takes into account the fact that the search procedure may be unable to find a point where $E_k^*(x) > \Delta_k + \beta$ even if one exists.

THEOREM 3. *Suppose that $X$ and each $X_k$ are compact, and at stage $k$ an approximation $R_k \in \mathscr{R}_k^*$ is produced that satisfies*

$$\rho_k^* - \delta_1 \leqslant \Delta_k \equiv \max_{x \in G_k} |W(x)(f(x) - R_k(x))| \leqslant \rho_k^* + \delta,$$

$$R_k(x) \geqslant l(x) - \delta_1, \quad \forall x \in L_k, \qquad R_k(x) \leqslant u(x) + \delta_1, \quad \forall x \in U_k,$$

*where $\varepsilon = \beta - \delta - \delta_1 > 0$. Also assume that $f$, $W$, $\phi_1, \ldots, \phi_n$, $\psi_1, \ldots, \psi_m$, $l$ and $u$ are continuous and bounded on their domains of definition with $W$ bounded away from zero on $G$, and suppose that either $L \cup U \subseteq G$ or $\{\phi_1, \ldots, \phi_n\}$ is linearly independent on $G \cap X_k$ for all but a finite number of indices $k$. Finally, assume that if at stage $k$ there exists $x \in X$ with $E_k(x) > \Delta_k + \beta + \gamma$, where $\gamma$ is a positive constant with $\beta + \gamma < 10^6$, then the search procedure will find at least one $x \in X$ with $E_k^*(x) > \Delta_k + \beta$, so the algorithm will not terminate at stage $k$. Then the algorithm will eventually terminate at an $R^* \in \mathscr{R}$ satisfying*

$$\Delta \equiv \max_{x \in G} |W(x)(f(x) - R^*(x))| \leqslant \rho^* + \beta + \gamma + \delta,$$

$$R^*(x) \geqslant l(x) - \beta - \gamma, \quad \forall x \in L, \qquad R^*(x) \leqslant u(x) + \beta + \gamma, \quad \forall x \in U.$$

*Proof.* We shall show only that the algorithm must terminate in a finite number of steps. Once this has been done, the remaining conclusions follow by the same arguments used in Theorem 2. Thus, assume the algorithm does not terminate in a finite number of steps. Lemma 1 applies in this situation to show that there exists $k_L$ such that $k \geqslant k_L$ implies that only $X_k \to X_{k+1}$ (and thus $X_k \subseteq X_{k+1}$) is possible. We now show that there exists $k_D \geqslant k_L$ such that if $k \geqslant k_D$, then $Q_k(x) \geqslant 2\eta$ for all $x \in \bar{X} = \text{closure}(\bigcup_{j=k_D}^{\infty} X_j) \subseteq X$.

Indeed, if this is false, then there is a subsequence $\{Q_\mu\}$ and a sequence $\{x_\mu\} \subseteq \bar{\bar{X}} = \text{closure}(\bigcup_{j=k_L}^{\infty} X_k)$ such that $Q_\mu(x_\mu) < 2\eta$; going to further subsequences we may assume $x_\mu \to \bar{x} \in \bar{\bar{X}}$ and $Q_\mu \to \bar{Q}$ (uniformly on $\bar{\bar{X}}$). Thus $\bar{Q}(\bar{x}) \leqslant 2\eta$. Now choose a sequence $\{x'_\mu\}$ with $x'_\mu \in X_\mu$ and $x'_\mu \to \bar{x}$; we have $Q_\mu(x'_\mu) \geqslant 4\eta$ and $Q_\mu(x'_\mu) \to \bar{Q}(\bar{x})$, so $\bar{Q}(\bar{x}) \geqslant 4\eta$, which is a contradiction.

We next wish to show that the numerator coefficients of $R_k$ can be chosen to be bounded as $k \to \infty$. This follows from standard arguments under the hypothesis that there is some $k_N \geqslant k_D$ such that $\{\phi_1, \ldots, \phi_n\}$ is linearly independent on $G \cap X_k$ for $k \geqslant k_N$. If, on the other hand, we assume that $L \cup U \subseteq G$, then $X_k = G_k$ for all $k$, and the nestedness of $\{X_k\}$ for $k \geqslant k_D$ implies the existence of some $k_N \geqslant k_D$ and a fixed subset of $\{\phi_1, \ldots, \phi_n\}$ which is a maximal linearly independent subset on each $X_k$ with $k \geqslant k_N$. For each

corresponding $R_k$, one can rewrite the numerator in terms of the basis functions in this subset, with zero coefficients for the other basis functions. Then standard arguments again imply the boundedness of the coefficients. From this, further standard arguments (using a subsequence of $\{R_k\}$ is necessary) show that $R_k \to$ some $\bar{R}$ (uniformly on $\bar{X}$), implying the existence of a $k_p \geqslant k_N$ such that if $i,j \geqslant k_p$, then $\max_{x\in\bar{X}} |R_i(x) - R_j(x)| \leqslant (\beta - \delta_1)/2$ and $\max_{x\in\bar{X}} |W(x)(R_i(x) - R_j(x))| \leqslant (\beta - \delta_1)/2$.

Now define $\bar{E}_k = \sup_{x\in X_k} E_k^*(x)$. By the definition of $E_k^*(x)$ and our assumptions on $R_k$ we have $\Delta_k \leqslant \bar{E}_k \leqslant \Delta_k + \delta_1$. Further let $\alpha_k = \bar{E}_k - \rho_k^*$. We then have $-\delta_1 \leqslant \alpha_k \leqslant \delta + \delta_1$. For each $k$, let $x_k$ be a point brought into $X_{k+1}$ which satisfies $E_k^*(x_k) > \Delta_k + \beta$. Then for $k \geqslant k_p$ we have $Q_k(x_k) \geqslant 2\eta$, and

$$\begin{aligned}
\bar{E}_{k+1} &\geqslant \max(|e_{k+1}^G(x_k)|, e_{k+1}^L(x_k), -e_{k+1}^U(x_k)) \\
&\geqslant \max(|e_k^G(x_k)|, e_k^L(x_k), -e_k^U(x_k)) - \tfrac{1}{2}(\beta - \delta_1) \\
&= E_k^*(x_k) - \tfrac{1}{2}(\beta - \delta_1) > \Delta_k + \beta - \tfrac{1}{2}(\beta - \delta_1) \\
&\geqslant \bar{E}_k + \beta - \delta_1 - \tfrac{1}{2}(\beta - \delta_1) = \rho_k^* + \alpha_k + \tfrac{1}{2}(\beta - \delta_1).
\end{aligned}$$

Thus, we have

$$\rho_{k+1}^* = \bar{E}_{k+1} - \alpha_{k+1} \geqslant \rho_k^* + \alpha_k - \alpha_{k+1} + \tfrac{1}{2}(\beta - \delta_1).$$

Arguing similarly, we have

$$\rho_{k+2}^* \geqslant \rho_{k+1}^* + \alpha_{k+1} - \alpha_{k+2} + \tfrac{1}{2}(\beta - \delta_1) \geqslant \rho_k^* + \alpha_k - \alpha_{k+2} + 2 \cdot \tfrac{1}{2}(\beta - \delta_1)$$

and by induction, for any $n$ we have

$$\rho_{k+n}^* \geqslant \rho_k^* + \alpha_k - \alpha_{k+n} + n \cdot \tfrac{1}{2}(\beta - \delta_1)$$

implying that

$$\rho_{k+n}^* \geqslant \rho_k^* - \delta - 2\delta_1 + n \cdot \tfrac{1}{2}(\beta - \delta_1).$$

This, in turn, implies that the algorithm must terminate in a finite number of steps since $\rho_\mu^* \leqslant \rho^*$ for all $\mu$. The theorem is proved. ∎

We have assumed that $\{\phi_1, \ldots, \phi_n\}$ is linearly independent on $G \cap X_k$ for all but a finite number of $k$'s if $L \cup U \not\subseteq G$. This will normally be satisfied in practice. Even in cases where it is not, one could obtain it by including some fixed subset of $G$ on which $\{\phi_1, \ldots, \phi_n\}$ is independent in each $A_k$.

If one wishes to compute a good approximation on a compact but infinite set $X$, instead of attempting to do this directly, one frequently chooses a (large) finite subset $Y$ and computes an approximation on $Y$, hoping that a good approximation on $Y$ will not be too bad on $X - Y$. The discretization

theorem below, which is an extension of two theorems in [4, pp. 84–88], indicates the reasonableness of this approach. We shall need the following notation. Suppose $B \subseteq A \subseteq X$ and $g$ is a function defined on $A$. Then, for $d(x, y) =$ Euclidean distance between $x$ and $y$, $Vx, y \in X$, we define

$$|B|_A = \text{density of } B \text{ in } A = 0, \qquad A = \varnothing,$$
$$= +\infty, \qquad A \neq \varnothing \quad \text{and} \quad B = \varnothing,$$
$$= \sup_{x \in A} \inf_{y \in B} d(x, y), \qquad A \neq \varnothing \quad \text{and} \quad B \neq \varnothing,$$

$$\| g \|_A \sup_{x \in A} | g(x)|,$$

$\omega =$ joint modulus of continuity of $f$, $l$, and $u$; that is, for $\delta > 0$,

$$\omega(\delta) = \max\Big( \sup_{\substack{x,y \in G \\ d(x,y) \leqslant \delta}} |f(x) - f(y)|, \sup_{\substack{x,y \in L \\ d(x,y) \leqslant \delta}} |l(x) - l(y)|,$$
$$\sup_{\substack{x,y \in U \\ d(x,y) \leqslant \delta}} |u(x) - u(y)|\Big)$$

(where a supremum over the empty set is 0),

$\Omega =$ joint modulus of continuity of $\phi_1, \ldots, \phi_n, \psi_1, \ldots, \psi_m$,

and

$$\mathscr{R}_Y = \Big\{ R = P/Q : P = \sum_{i=1}^{n} p_i \phi_i, Q = \sum_{i=1}^{m} q_i \psi_i, Q(x) \geqslant \eta, \forall x \in Y,$$
$$|q_i| \leqslant 1 \text{ for all } i \text{ with equality holding at least once},$$
$$R(x) \geqslant l(x), \forall x \in Y \cap L, R(x) \leqslant u(x), \forall x \in Y \cap U \Big\}.$$

For any $Y \subseteq X$, we say $\bar{R} \in \mathscr{R}_Y$ is a best approximation to $f$ on $Y$ if $\| f - \bar{R} \|_{G \cap Y} \leqslant \| f - R \|_{G \cap Y}, \forall R \in \mathscr{R}_Y$.

THEOREM 4. *Suppose $f$, $W$, $\phi_1, \ldots, \phi_n$, $\psi_1, \ldots, \psi_m$, $l$, $u$ are continuous and bounded on their domains of definition with $W$ bounded away from zero on $G$ and $\{\phi_1, \ldots, \phi_n\}$ linearly independent on $G$. Suppose $R^* \in \mathscr{R}_X$ is a best approximation to $f$ on $X$, $\delta$ is a positive number, $Y$ is a subset of $X$ with $|G \cap Y|_G < \delta$, $|L \cap Y|_L < \delta$, $|U \cap Y|_U < \delta$, and $\bar{R} \in \mathscr{R}_Y$ is a best approximation to $f$ on $Y$. Then*

(i) *for $\delta$ sufficiently small, there exists a constant $\gamma$ independent of $Y, R^*$, and $\bar{R}$ such that*

$$\max\{\|f-\bar{R}\|_G - \|f-R^*\|_G, \sup_{x\in L}(l(x)-\bar{R}(x)), \sup_{x\in U}(\bar{R}(x)-u(x)),$$

$$\sup_{x\in X}\{(\eta-\bar{Q}(x))\} \leqslant \omega(\delta)+\gamma\Omega(\delta)\};$$

(ii) *If $R^*$ is unique, then $\bar{R}$ converges uniformly to $R^*$ on $X$ as $\delta \to 0$.*

*Because it is straightforward, we simply sketch the proof.*

*Proof.* One first shows by subsequence arguments that $\bar{Q} \geqslant \eta/2$ on $X$ if $\delta$ is sufficiently small, and the numerator coefficients of $\bar{R}$ are bounded independently of $Y$, $R^*$, and $\bar{R}$ if $\delta$ is sufficiently small. The theorem can then be proved using arguments similar to those in [4, *pp.* 85–88]; in part (ii) we use the quantity $m(\varepsilon) = \inf_{R\in R_Y,\, \|R-R^*\|_X \geqslant \varepsilon}[\|f-R\|_X - \|f-R^*\|_X]$ which by a subsequence argument can be shown to be positive for $\delta$, $\varepsilon$ sufficiently small.

It is possible to have $\|f-\bar{R}\|_G < \|f-R^*\|_G$ since $\bar{R}$ need not satisfy the constraints on $X-Y$. It can be shown by a subsequence argument that $\|f-\bar{R}\|_G \to \|f-R^*\|_G$ as $\delta \to 0$, but there are examples for which $\|f-R^*\|_G - \|f-\bar{R}\|_G > \omega(\delta)+\gamma\Omega(\delta)$ for any constant $\gamma$, for $\delta$ sufficiently small.

We observe that $|Y|_X < \delta$ does not imply the density hypotheses of Theorem 4, and is not sufficient to get the conclusions. For example, if the point $(0, 0)$ is not included in $Y$ in the example presented later in this paper, then $|\|f-\bar{R}\|_X - \|f-R^*\|_X|$ cannot be forced arbitrarily close to 0 by just forcing $|Y|_X$ arbitrarily close to zero. As a general rule, it is wise to include any isolated points of $G$, $L$ and $U$ in $Y$.

## 4. An Example and Conclusions

Consider the function $f$ defined on $[0, 1] \times [0, 1]$ by

$$\begin{aligned} f(x,y) &= 1; && 0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant -1.25x + 1.25, \\ &= \text{undefined}; && 0 \leqslant x \leqslant 1, \quad -1.25x + 1.25 < y < -1.25x + 1.375, \\ &0; && 0 \leqslant x \leqslant 1, \quad -1.25x + 1.375 \leqslant y \leqslant 1. \end{aligned}$$

Suppose we would like to approximate $f$ by a generalized rational function of the form $R(x, y) = (p_1 + p_2 x + p_3 y)/(q_1 + q_2 \sin(x+y))$, with the requirements $R(x, y) \geqslant 0$ for $y > -1.25x + 1.25$ and $R(0, 0) \leqslant 1$. Further, we set $W(x, y) = 0.5$, where $f(x, y) = 1$ and $W(x, y) = 1$, where $f(x, y) = 0$; thus

we are willing to allow a larger error where $f(x, y) = 1$ in order to improve the approximation where $f(x, y) = 0$. Taking $X = \{(0.005i, 0.01j): 0 \leqslant i \leqslant 200,\ 0 \leqslant j \leqslant 100\}$ (i.e., a $101 \times 201$ subdivision) and running this example on a CYBER 172 (roughly 15 digits of accuracy), we obtained $R(x, y) = 1.00000 - 0.55915x - 0.44085y)/(1.00000 + 0.24173 \sin(x + y))$ after 40.5 sec of execution time and 15 applications of differential correction, with no $X_k$ having more than nine points. The error norm was $\Delta = 0.31746$; the extreme points and the weighted errors at them were (0, 0) (0-hit upper constraint), $(0.28, 0.9)(\Delta)$, $(0.24, 0.95)(\Delta)$, $(0.2, 1)(\Delta)$, $(0.3, 1)(-\Delta)$, (1, 1)(0-hit lower constraint). It was not necessary to insert extra constraints in the differential-correction subroutines to force $Q_k \geqslant$ some $\eta$ on $X_k$, since each $Q_k$ was actually greater than 0.84 throughout $[0, 1] \times [0, 1]$.

For comparison, we modified the program so that no points would be dropped (i.e., $S_{k-1} = X_{k-1}$ for all $k$); this time differential correction was applied 12 times, with $X_{11}$ having 23 points and 43.0 sec were required. Although the time difference was not great, we remark here that in some problems this procedure could cost considerable time, or even cause failure of the program due to storage problems when $X_k$ becomes too large.

To illustrate the remark following Theorem 4, we ran the program with the point (0, 0) (and its constraint) deleted; the results were $R(x, y) = (1.49500 - 0.84554x - 0.64945y)/(0.90823 + 1.00000 \sin(x + y))$ with error norm $\bar{\Delta} = 0.31621$ and the extreme point at (0, 0) replaced by one at (0.005, 0) with error $-\bar{\Delta}$. Eighteen applications of differential correction and 41.0 sec were required.

Finally, in order to compare the algorithm with straight differential correction on a set small enough for the latter to be applied, we use a $11 \times 9$ subdivision; our algorithm required 1.9 sec (with differential correction being applied six times to grids with maximum size 10), while straight differential correction required 3.3 sec. In both cases, we obtained $R(x, y) = (1.00000 - 0.55922x - 0.44078y)/(1.00000 + 0.24189 \sin(x + y))$. The error norm was $\bar{\bar{\Delta}} = 0.31746$ (actually, about $4 \times 10^{-7}$ smaller than $\Delta$), and the extreme points were (0, 0)(0-hit upper constraint), $(0.3, 0.875)(\bar{\bar{\Delta}})$, $(0.2, 1)(\bar{\bar{\Delta}})$, $(0.3, 1)(-\bar{\bar{\Delta}})$, (1, 1)(0-hit lower constraint).

A second paper is being prepared with further examples and more discussion of the code; a FORTRAN listing will also be included.

## REFERENCES

1. I. BARRODALE, M. J. D. POWELL, AND F. D. K. ROBERTS, The differential correction algorithm for rational $l_\infty$ approximation, *SIAM J. Numer. Anal.* **9** (1972), 493–504,
2. I. BARRODALE, F. D. K. ROBERTS, AND K. B. WILSON, "An Efficient Computer

Implementation of the Differential Correction Algorithm for Ratioal Approximation," Univ. of Victoria, Dept. of Math. report DM-115-IR (1978), presented at the Manitoba Conf. on Numerical Math. and Computing, Sept. 1977.

3. E. W. CHENEY, "Introduction to Approximation Theory," McGraw–Hill, New York, 1966.
4. E. W. CHENEY AND H. L. LOEB, Two new algorithms for rational approximation, *Numer. Math.* **4** (1962), 124–127.
5. D. E. DUDGEON, Recursive filter design using differential correction, *IEEE Trans. Acoust., Speech, Signal Process.* **ASSP-22** (1974), 443–448.
6. C. B. DUNHAM, Rational Approximation by the First Algorithm of Remez, preprint.
7. BARBARA D. ELDRIDGE AND DANIEL D. WARNER, "An Implementation of the Differential Correction Algorithm," Bell Labs. Computing Science tech. report No. 48, 1976.
8. E. H. KAUFMAN, JR. AND G. D. TAYLOR, Uniform rational approximation of functions of several variables, *Internat. J. Numer. Methods Engrg.* **9** (1975), 297–323.
9. E. H. KAUFMAN, JR. AND G. D. TAYLOR, An application of a restricted range version of the differential correction algorithm to the design of digital systems, *Internat. Ser. Numer. Math.* **30** (1976), 207–232.
10. E. H. KAUFMAN, JR., D. J. LEEMING, AND G. D. TAYLOR, A combined Remes-differential correction algorithm for rational approximation, *Math. Comp.* **32** (1978), 233–242.
11. E. H. KAUFMAN, JR., D. J. Leeming, and G. D. Taylor, A combines Remes-differential correction algorithm for rational approximation: Experimental results, *Comput. Math. Appl.* **6** (1980), 155–160.
12. E. H. KAUFMAN, JR. AND G. D. TAYLOR, Uniform rational approximation with restricted denominators, *J. Approx. Theory* **32** (1981), 9–26.
13. E. H. KAUFMAN, JR., D. J. LEEMING, AND G. D. TAYLOR, Uniform rational approximation by differential correction and Remes-differential correction, *Internat. J. Numer. Methods Engrg.* **17** (1981), 1273–1280.
14. C. M. LEE AND F. D. K. ROBERTS, A comparison of algorithms for rational $l_\infty$ approximation, *Math. Comp.* **27** (1973), 111–121.
15. M. T. MCCALLIG, Sperry Research Center, research report.
16. M. T. MCCALLIG, R. KURTH, AND B. STEEL, Recursive digital filters with low coefficient sensitivity *in* "Proc. 1979 IEEE Int. Conf. on Accoustics, Speech, and Signal Processing," Washington D.C., April 1979.
17. G. A. WATSON, The calculation of best restricted approximations, *SIAM J. Numer. Anal.* **11** (1974), 693–699.
18. G. A. WATSON, A multiple exchange algorithm for multivariate Chebyshev approximation, *SIAM. J. Numer. Anal.* **12** (1975), 46–52.