3. Exception Handling in Derivative Computation with Nonarchimedean Calculus*

(Chapter of "Computational Differentiation: Techniques, Applications, and Tools", Martin Berz, Christian Bischof, George Corliss, and Andreas Griewank, eds., SIAM, 1996.)

Khodr Shamseddine[†] Martin Berz[†]

Abstract

While conventional computational differentiation based on the forward or reverse modes allows highly accurate computation of derivatives, there are situations where these modes fail to produce the values of derivatives, although the underlying function is differentiable. Typical examples of this phenomenon are connected to the occurrence of branch points in coding as in IF-ELSE structures as well as the occurrence of some non-differentiable parts that do not affect the differentiability of the end result.

We show that based on ideas of nonarchimedean calculus on Levi-Civita fields, these problems can be avoided. It is possible to rigorously decide whether a function is differentiable or not at any given point, and if it is, to determine its derivatives to any order, even if the coding exhibits branch points or non-differentiable pieces.

We give details of an implementation of the method and examples for its use for typical pathological problems.

Keywords: Exception handling, nonarchimedean calculus, nonarchimedean field $\mathcal{R}$, Heaviside function, smoothness properties of computer functions, standard form of computer functions, derivatives are differential quotients, differentiability of computer functions, COSY INFINITY.

1 Introduction

The goal of computational differentiation is the fast and accurate computation of derivatives of complicated functions of one or many variables in a computer environment. However, the conventional approaches based on the forward or reverse mode fail to find the derivatives of certain functions at given points, even though the functions are differentiable at the respective points. For example, the functions

$$f_1(x) = \begin{cases} (\sin x)/x & \text{if } x \neq 0 \\ 1 & \text{if } x = 0 \end{cases} \quad \text{and } f_2(x) = x^2\sqrt{|x|} + \exp(x) \tag{1}$$

are both differentiable at 0; but the attempt to compute their derivatives using automatic differentiation (AD) fails. This is particularly unsatisfying since conventional numerical differentiation based on divided differences is able to find approximate values for the derivatives. On the other hand, depending on the implementation of the precompiler, sometimes automated code conversion fails to recognize points where the function under consideration is in fact not differentiable. As an example, consider the piece of code

```
If(cos(x)=1) ; f = cos(x) ; else ; f = 0 ; endif ;
```

*This research was financially supported by the Department of Energy, Grant No. DE-FG02-95ER40931, and the Alfred P. Sloan Foundation.

†Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA, (shamseddine@nscl.msu.edu, berz@pa.msu.edu).

which in order to recognize non-differentiability at $x = 0$ requires a careful treatment of the appearance of variables in If statements, which can become even much more subtle than in our simple example.

In this paper, we show that implementation of the nonarchimedean field $\mathcal{R}$ on a computer provides a remedy to the defects of computational differentiation concerning functions of one variable; see [Berz1992b], [Berz1994a], [Berz1996a] for a detailed study of the field $\mathcal{R}$. Using the calculus on $\mathcal{R}$, we formulate a necessary and sufficient condition for the derivatives of functions from R into R representable on a computer to exist, and show how to find these derivatives whenever they exist.

We start with a study of computer environment functions of one variable, and their properties of smoothness. This class of functions includes all intrinsic functions and the Heaviside function, which is the tool to account for branching in a computer code, as well as any finite combination thereof.

2 Computer Environment Functions

At the machine level, a function $f : R \to R$ is characterized by what it does to the original set of memory locations. So f induces a function $\vec{F}(f) : R^m \to R^m$, where m is the number of memory locations affected in the process of computing f. We note here that, without optimization, $\vec{F}(f)$ is unique up to flipping of the memory locations. On the other hand, with optimization, $\vec{F}(f)$ is unique in the subspace describing the true variables. Moreover, at the machine level, any code constitutes solely of intrinsic functions, arithmetic operations, and branches. In the following, we formally define the machine level representations of intrinsic functions, the Heaviside function, and the arithmetic operations.

DEFINITION 2.1. *Let* $\mathcal{I}= \{H, \sin, \cos, \tan, \exp, \ldots\}$ *be the set consisting of the Heaviside function* H *and all intrinsic functions on a computer, which for the sake of convenience are assumed to include the reciprocal function; and let* $\mathcal{O}= \{+, \cdot\}$.

DEFINITION 2.2. *For* $f \in \mathcal{I}$, *define* $\vec{F}_{i,k,f} : R^m \to R^m$ *by*

$$\vec{F}_{i,k,f}(x_1, x_2, \ldots, x_m) = (x_1, \ldots, x_{k-1}, \underbrace{f(x_i)}_{k}, x_{k+1}, \ldots, x_m);$$

so the kth memory location is replaced by $f(x_i)$. *Then* $\vec{F}_{i,k,f}$ *is the machine level representation of* f. *For* $\otimes \in \mathcal{O}$, *define* $\vec{F}_{i,j,k,\otimes} : R^m \to R^m$ *by*

$$\vec{F}_{i,j,k,\otimes}(x_1, x_2, \ldots, x_m) = (x_1, \ldots, x_{k-1}, \underbrace{x_i \otimes x_j}_{k}, x_{k+1}, \ldots, x_m),$$

so the kth memory location is replaced by $x_i \otimes x_j$. *Then* $\vec{F}_{i,j,k,\otimes}$ *is the machine level representation of* $\otimes$. *Finally, let*

$$\mathcal{F} = \{\vec{F}_{i,k,f} : f \in \mathcal{I}\} \cup \{\vec{F}_{i,j,k,\otimes} : \otimes \in \mathcal{O}\}.$$

DEFINITION 2.3. *A function* $f : R \to R$ *is called a computer function if it can be obtained from intrinsic functions and the Heaviside function through a finite number of arithmetic operations and compositions. In this case, there are some* $\vec{F}_1, \vec{F}_2, \ldots, \vec{F}_M \in \mathcal{F}$ *such that* $\vec{F}(f) = \vec{F}_M \circ \vec{F}_{M-1} \circ \cdots \circ \vec{F}_2 \circ \vec{F}_1$, *and we call* $\vec{F}(f) : R^m \to R^m$, *already mentioned above, the machine level representation of* f.

3 Smoothness Properties of Computer Functions

In this section, we show that within the bounds of the real numbers that can be represented on a computer, the domain of definition as well as the domain of continuity of a computer function is a finite union of intervals. We also show that, for any fixed $n \in N$ and any computer function f, the domain of definition (continuity) of the nth derivative $f^{(n)}$ is again a finite union of intervals.

DEFINITION 3.1. *Let l and L denote, respectively, the lower and upper bounds of the positive real numbers that can be represented on a computer. It follows that the domain of the computer numbers is $D_c = [-L, -l] \cup [l, L]$.*

LEMMA 3.1. *Let $f \in \mathcal{I}$ be given. Let D be the domain of definition (continuity) of f in D_c. Then D is a finite union of intervals. Furthermore, the system $\{f(x) \in I;\ x \in D_c\}$, where $I \subset D_c$ is an interval, has as a solution a finite union of intervals in D_c.*

Proof. We check the statement for each function $f \in \mathcal{I}$, and the arguments are quite straightforward. For reasons of space, we will refrain from writing down the details. As the example of $f(x) = \sin(x)$ and the interval $I = [0, 1]$ shows, the existence of bounds for the real numbers representable on a computer is essential for this lemma to hold. $\square$

We will show that the result of the previous lemma is indeed true for any computer function.

LEMMA 3.2. *Let f_1 and f_2 be two computer functions with domains of definition (continuity) D_1 and D_2 in D_c, respectively. Assume that, for $j = 1, 2$, D_j is a finite union of intervals, and the system $\{f_j(x) \in I_j;\ x \in D_c\}$, where $I_j \subset D_c$ is an interval, has as a solution a finite union of intervals in D_c. If $F = f_2 \circ f_1$ and D is the domain of definition (continuity) of F in D_c, then D is a finite union of intervals. Furthermore, the system $\{F(x) \in I;\ x \in D_c\}$, where $I \subset D_c$ is an interval, has as a solution a finite union of intervals in D_c.*

Proof. F is defined (continuous) at x whenever $x \in D_1$ and $f_1(x) \in D_2$. Moreover, F could possibly be continuous at only finitely many points $\{x_1, x_2, \ldots, x_M\}$ where, for $k = 1, 2, \ldots, M$, f_1 is not continuous at x_k or f_2 is not continuous at $f_1(x_k)$. This is so because the domains of continuity of f_1 and f_2 are both finite unions of intervals. By assumption, D_2 is a finite union of intervals; so $D_2 = \bigcup_{i=1}^{N_2} I_{2,i}$, where the $I_{2,i}$'s are intervals in D_c. But for each i, $f_1(x) \in I_{2,i} \Leftrightarrow x \in \bigcup_{j=1}^{j_i} J_{i,j}$, where the $J_{i,j}$'s are again intervals in D_c. Altogether, we have that

$$D = D_1 \cap \left(\bigcup_{i=1}^{N_2} \bigcup_{j=1}^{j_i} J_{i,j} \right) (\cup \{x_1, \ldots, x_M\}) = \bigcup_{i=1}^{N_2} \bigcup_{j=1}^{j_i} (D_1 \cap J_{i,j}) (\cup \{x_1, \ldots, x_M\}).$$

Since D_1 is a finite union of intervals by assumption, each $(D_1 \cap J_{i,j})$ with $1 \le i \le N_2, 1 \le j \le j_i$, is a finite union of intervals. Hence D itself is a finite union of intervals.

To prove the second statement in the lemma, we note that $F(x) \in I$ is equivalent to $f_2(f_1(x)) \in I$. By the assumption of the lemma, $f_2(X) \in I$ has as a solution a finite union of intervals $\bigcup_{i=1}^{M_2} A_{2,i}$ in D_c. Thus, $F(x) \in I \Leftrightarrow f_2(f_1(x)) \in I \Leftrightarrow f_1(x) \in \bigcup_{i=1}^{M_2} A_{2,i}$. But for each i, $f_1(x) \in A_{2,i} \Leftrightarrow x \in \bigcup_{k=1}^{k_i} B_{i,k}$, where the $B_{i,k}$'s are again intervals in D_c. Altogether, we have that

$$F(x) \in I \Leftrightarrow x \in \bigcup_{i=1}^{M_2} \bigcup_{k=1}^{k_i} B_{i,k},$$

a finite union of intervals in D_c. $\square$

LEMMA 3.3. *Let f_1 and f_2 be two computer functions with domains of definition (continuity) D_1 and D_2 in D_c, respectively. Assume that, for $j = 1, 2$, D_j is a finite* *union of intervals, and the system* $\{f_j(x) \in I_j;\ x \in D_c\}$, *where* $I_j \subset D_c$ *is an interval, has as a solution a finite union of intervals in* D_c. *If* $\otimes \in \mathcal{O}$, $F = f_2 \otimes f_1$, *and* D *is the domain of definition (continuity) of* F *in* D_c, *then* D *is a finite union of intervals. Furthermore, the system* $\{F(x) \in I;\ x \in D_c\}$, *where* $I \subset D_c$ *is an interval, has as a solution a finite union of intervals in* D_c.

Proof. First we note that F is defined (continuous) at a point x whenever f_1 and f_2 are both defined (continuous) at x. Moreover, F could possibly be continuous at only finitely many points $\{x_1, x_2, \ldots, x_M\}$ where, for $k = 1, 2, \ldots, M$, f_1 is not continuous at x_k or f_2 is not continuous at x_k. Thus F is defined (continuous) at x if and only if $x \in D_1$ and $x \in D_2$ (or $x \in \{x_1, x_2, \ldots, x_M\}$). By assumption, D_1 and D_2 are finite unions of intervals; so $D_1 = \bigcup_{i=1}^{N_1} I_{1,i}$ and $D_2 = \bigcup_{j=1}^{N_2} I_{2,j}$, where the $I_{1,i}$'s and the $I_{2,j}$'s are intervals in D_c. Altogether, we have that F is defined (continuous) at x if and only if

$$x \in \bigcup_{i=1}^{N_1} I_{1,i} \cap \bigcup_{j=1}^{N_2} I_{2,j} (\cup\{x_1, x_2, \ldots, x_M\}) = \bigcup_{i=1}^{N_1} \bigcup_{j=1}^{N_2} (I_{1,i} \cap I_{2,j}) \, (\cup\{x_1, x_2, \ldots, x_M\}),$$

a finite union of intervals in D_c.

To prove the second statement in the lemma, we note that $F(x) \in I \Leftrightarrow f_1(x) \in I_1$ and $f_2(x) \in I_2$, where I_1 and I_2 are both intervals in D_c. Hence, the solution of $F(x) \in I$ is the intersection of two finite unions of intervals in D_c and is itself a finite union of intervals in D_c. □

THEOREM 3.1. *Let* f *be a computer function, and let* D *be the domain of definition (continuity) of* f *in* D_c. *Then* D *is a finite union of intervals. Furthermore, the system* $\{f(x) \in I;\ x \in D_c\}$, *where* $I \subset D_c$ *is an interval, has as a solution a finite union of intervals in* D_c.

Proof. Since f is a computer function, f is obtained in finitely many steps from functions in $\mathcal{I}$ via compositions and arithmetic operations. Since functions in $\mathcal{I}$ satisfy the statement of the theorem and are themselves computer functions, we can apply the previous two lemmas to assert that at each step of "constructing" f from the functions in $\mathcal{I}$, we obtain a computer function that satisfies the statement of the theorem. Hence, f itself satisfies the statement of the theorem. □

The last theorem can immediately be extended to all derivatives of computer functions. Applying the rules of differentiation to the formula describing the function simply yields another (usually more complicated) formula that is obviously again a computer function. Hence, the theorem we have just proved holds as well for derivatives of computer functions:

COROLLARY 3.1. *Let* f *be a computer function. Then, for a fixed* $n \in N$, *the domain of definition (continuity) of* $f^{(n)}$ *in* D_c *is a finite union of intervals. Furthermore, the system* $\{f^{(n)}(x) \in I;\ x \in D_c\}$, *where* $I \subset D_c$ *is an interval, has as a solution a finite union of intervals in* D_c.

In the following, we derive a standard representation of any computer function f around any fixed point x_0 of its domain of definition.

LEMMA 3.4. *Let* $f \in \mathcal{I}$ *be given. Then there exist mutually disjoint intervals* $I_1, \ldots, I_M$ *in* D, *the domain of definition of* f *in* D_c, *such that* $\bigcup_{k=1}^{M} I_k = D$, *and if* x_0 *is an interior point of* I_k *then there exists a positive real number* σ *such that, for* $0 < x < \sigma$,

$$f(x_0 + x) = A_0^+(x) + \sum_{i=1}^{i_k^+} x^{q_i^+} A_i^+(x) \ \text{and} \tag{2}$$

$$f(x_0 - x) \quad = \quad A_0^-(x) + \sum_{i=1}^{i_k^-} x^{q_i^-} A_i^-(x), \tag{3}$$

where $A_i^{\pm}(x), 0 \le i \le i_k^{\pm}$, is a power series in x with positive radius of convergence, $A_i^{\pm}(0) \ne 0$ for $i = 1, \dots, i_k^{\pm}$; and where the $q_i^{\pm}$'s are nonzero rational numbers that are not positive integers.

Proof. The statement of the lemma can easily be verified for each $f \in \mathcal{I}$. □

REMARK 3.1. *Noninteger rational powers may appear in Eqs. (2) or (3) as a result of the root function.*

REMARK 3.2. *If in the above lemma x_0 were a lower bound of I_k, then $f(x_0 + x) = A_0(x) + \sum_{i=1}^{i_r} x^{q_i} A_i(x)$ for $0 < x < \sigma_r$, where σ_r is a fixed positive real number.*

On the other hand, if in the above lemma x_0 were an upper bound of I_k, then $f(x_0 - x) = B_0(x) + \sum_{i=1}^{i_l} x^{t_i} B_i(x)$ for $0 < x < \sigma_l$, where σ_l is a fixed positive real number.

LEMMA 3.5. *Let f_1 and f_2 be two computer functions that satisfy the requirements of the previous lemma; then so do $F_\otimes = f_2 \otimes f_1$, where $\otimes \in \{+, \cdot\}$; and $F_\circ = f_2 \circ f_1$.*

Proof. Let D_1, D_2, and D be the domains of definition of f_1, f_2, and $F_\otimes$ in D_c, respectively. Without loss of generality, we may assume that $D = D_1 \cap D_2$. By the assumption of the lemma, there exist mutually disjoint intervals $I_1, \dots, I_{M_1}$ in D_1, and mutually disjoint intervals $J_1, \dots, J_{M_2}$ in D_2 such that

$$\begin{aligned}
&\bigcup_{k=1}^{M_1} I_k = D_1, \ \bigcup_{m=1}^{M_2} J_m = D_2, \\
&f_1(x_0 \pm x) = A_0^{\pm}(x) + \sum_{i=1}^{i_k^{\pm}} x^{q_i^{\pm}} A_i^{\pm}(x) \text{ for } x_0 \text{ inside } I_k, \text{ and } 0 < x < \sigma_1 \\
&f_2(y_0 \pm y) = B_0^{\pm}(y) + \sum_{j=1}^{j_m^{\pm}} y^{t_j^{\pm}} B_j^{\pm}(y) \text{ for } y_0 \text{ inside } J_m, \text{ and } 0 < y < \sigma_2;
\end{aligned} \tag{4}$$

where σ_1 and σ_2 are both positive real numbers; $A_i^{\pm}(x), 0 \le i \le i_k^{\pm}$, and $B_j^{\pm}(y), 0 \le j \le j_m^{\pm}$, are power series in x and y with positive radii of convergence; $A_i^{\pm}(0) \ne 0$ for $i \in \{1, \dots, i_k^{\pm}\}$ and $B_j^{\pm}(0) \ne 0$ for $j \in \{1, \dots, j_m^{\pm}\}$; and the $q_i^{\pm}$'s and the $t_j^{\pm}$'s are nonzero rational numbers that are not positive integers. As a reminder, we note that σ_1, the $A_i^{\pm}$'s, and the $q_i^{\pm}$'s depend on x_0. Similarly, σ_2, the $B_j^{\pm}$'s, and the $t_j^{\pm}$'s depend on y_0.

For $1 \le k \le M_1$ and $1 \le m \le M_2$, let $E_{k,m} = I_k \cap J_m$. Then $E_{k_1,m_1} \cap E_{k_2,m_2} = \emptyset$ if $k_1 \ne k_2$ or $m_1 \ne m_2$; so $\{E_{k,m}; 1 \le k \le M_1, 1 \le m \le M_2\}$ are mutually disjoint intervals in $D = D_1 \cap D_2$. Moreover,

$$\begin{aligned}
\bigcup_{\substack{1 \le k \le M_1 \\ 1 \le m \le M_2}} E_{k,m} \quad &= \quad \bigcup_{\substack{1 \le k \le M_1 \\ 1 \le m \le M_2}} (I_k \cap J_m) = \bigcup_{k=1}^{M_1} \left(I_k \cap \bigcup_{m=1}^{M_2} J_m \right) \\
&= \left(\bigcup_{k=1}^{M_1} I_k \right) \cap \left(\bigcup_{m=1}^{M_2} J_m \right) = D_1 \cap D_2 = D.
\end{aligned}$$

Now let x_0 be a point inside $E_{k,m}$. Then x_0 is simultaneously inside I_k and J_m; hence for x smaller than the minimum of σ_1 and σ_2 in (4), we have that

$$F_\otimes(x_0 \pm x) = f_2(x_0 \pm x) \otimes f_1(x_0 \pm x) = \left(\sum_{i=0}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x)\right) \otimes \left(\sum_{j=0}^{j_m^\pm} x^{t_j^\pm} B_j^\pm(x)\right),$$

where $q_0^\pm = t_0^\pm = 0$. It is easy to check that the sum or product of two expressions of the form (2) or (3) will again yield an expression of the same form. This finishes the proof of the first part of the lemma.

The proof of the second part of the lemma is more involved, and we will only include a sketch of the proof here. Let x_0 be an interior point of the domain of definition of $F_\circ$. Then

$$\begin{aligned}
f_1(x_0 \pm x) &= A_0^\pm(x) + \sum_{i=1}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x) \text{ for } 0 < x < \sigma_1 \\
f_2(A_0^+(0) \pm y) &= B_0^\pm(y) + \sum_{j=1}^{j_m^\pm} y^{t_j^\pm} B_j^\pm(y) \text{ for } 0 < y < \sigma_2 \\
f_2(A_0^-(0) \pm y) &= C_0^\pm(y) + \sum_{j=1}^{j_n^\pm} y^{p_j^\pm} C_j^\pm(y) \text{ for } 0 < y < \sigma_3,
\end{aligned}$$

where σ_1, σ_2 and σ_3 are all positive real numbers; $A_i^\pm(x), 0 \le i \le i_k^\pm$, $B_j^\pm(y), 0 \le j \le j_m^\pm$, and $C_j^\pm(y), 0 \le j \le j_n^\pm$, are power series in x and y with positive radii of convergence; $A_i^\pm(0) \neq 0$ for $i \in \{1, \ldots, i_k^\pm\}$, $B_j^\pm(0) \neq 0$ for $j \in \{1, \ldots, j_m^\pm\}$, and $C_j^\pm(0) \neq 0$ for $j \in \{1, \ldots, j_n^\pm\}$; and the $q_i^\pm$'s, the $t_j^\pm$'s and the $p_j^\pm$'s are nonzero rational numbers that are not positive integers. Let $A_{00}^\pm(x) = A_0^\pm(x) - A_0^\pm(0)$. Then $A_{00}^\pm(x)$ has no constant term, and we have, for $0 < x < \sigma_1$, that $f_1(x_0 \pm x) = A_0^\pm(0) + A_{00}^\pm(x) + \sum_{i=1}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x)$. Since x_0 is an interior point of the domain of definition of $F_\circ = f_2 \circ f_1$ and $A_{00}^\pm(x) + \sum_{i=1}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x)$ has no constant term, there exists a real σ, $0 < \sigma \le \sigma_1$, such that $|A_{00}^\pm(x) + \sum_{i=1}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x)| < \min(\sigma_2, \sigma_3)$ and $A_{00}^\pm(x) + \sum_{i=1}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x)$ has the same sign for all x satisfying $0 < x < \sigma$. Therefore, for $0 < x < \sigma$, we have that

$$\begin{aligned}
F_\circ(x_0 \pm x) &= f_2\left(f_1(x_0 \pm x)\right) = f_2\left(A_0^\pm(0) + A_{00}^\pm(x) + \sum_{i=1}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x)\right) \\
&= E_0\left(A_{00}^\pm(x) + \sum_{i=1}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x)\right) \qquad (5) \\
&\quad + \sum_{j=1}^{J} \left|A_{00}^\pm(x) + \sum_{i=1}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x)\right|^{s_j} E_j\left(A_{00}^\pm(x) + \sum_{i=1}^{i_k^\pm} x^{q_i^\pm} A_i^\pm(x)\right),
\end{aligned}$$

where $E_j, 0 \le j \le J$, are power series; $E_j(0) \neq 0$ for $1 \le j \le J$; and the s_j's are nonzero rational numbers that are not positive integers.

Note that for $0 \leq j \leq J$, we could factor the leading term $\alpha_q x^q$ in $\left|A_{00}^{\pm}(x) + \sum_{i=1}^{i_k^{\pm}} x^{q_i^{\pm}} A_i^{\pm}(x)\right|^{s_j}$ and obtain $\alpha_q x^q$ multiplied by a power series of an expression of the form (2) or (3). Using an argument similar to that of the treatment of series of series in [Osgood1938a, pages 205-208], we obtain that $E_j\left(A_{00}^{\pm}(x) + \sum_{i=1}^{i_k^{\pm}} x^{q_i^{\pm}} A_i^{\pm}(x)\right), 0 \leq j \leq J$, and $\left|A_{00}^{\pm}(x) + \sum_{i=1}^{i_k^{\pm}} x^{q_i^{\pm}} A_i^{\pm}(x)\right|^{s_j}, 1 \leq j \leq J$, are all of the form (2) or (3). Hence, $F_\circ(x_0 \pm x)$ in (5) is itself of the form (2) or (3). □

THEOREM 3.2. *Let f be a computer function. Then there exist mutually disjoint intervals $I_1, \ldots, I_M$ in D, the domain of definition of f in $D_c = [-L, -l] \cup [l, L]$, such that $\bigcup_{k=1}^{M} I_k = D$, and if x_0 is an interior point of I_k then there exists a positive real number σ such that, for $0 < x < \sigma$,*

$$f(x_0 \pm x) = A_0^{\pm}(x) + \sum_{i=1}^{i_k^{\pm}} x^{q_i^{\pm}} A_i^{\pm}(x),$$

where $A_i^{\pm}(x), 0 \leq i \leq i_k^{\pm}$, is a power series in x with a positive radius of convergence, $A_i^{\pm}(0) \neq 0$ for $i = 1, \ldots, i_k^{\pm}$, and the $q_i^{\pm}$'s are nonzero rational numbers that are not positive integers.

Proof. Since f is a computer function, f is obtained in finitely many steps from functions in $\mathcal{I}$ via compositions and arithmetic operations. Using induction, we obtain the result immediately from Lemmas (3.4) and (3.5). □

Since the family of computer functions is closed under differentiation to any order n, the theorem we have just proved holds as well for derivatives of computer functions.

In the following sections, we extend real computer functions to the nonarchimedean field $\mathcal{R}$ and study the calculus of the resulting class of functions. We show how implementing this calculus on a computer can be used to accurately compute the derivatives of the original real functions at given real points whenever the derivatives exist.

4 Theoretical Tools about $\mathcal{R}$

In this section, we discuss some new theoretical results about $\mathcal{R}$, which will prove useful for computing derivatives of real computer functions.

DEFINITION 4.1. *(k-Equidifferentiable Functions on $\mathcal{R}$) Let $k > 0$ in Q be given. A function $f : D \subset \mathcal{R} \to \mathcal{R}$ is said to be k-equidifferentiable with derivative g at the point $x_0 \in D$ if, for any at most finite positive $\epsilon \in \mathcal{R}$, we can find a positive $\delta \in \mathcal{R}$ satisfying $\delta^k \sim \epsilon$ such that*

$$\left|\frac{f(x) - f(x_0)}{x - x_0} - g\right| < \epsilon \text{ for any } x \in D \backslash \{x_0\} \text{ with } |x - x_0| < \delta.$$

If this is the case, we write $g = f'(x_0)$.

REMARK 4.1. *If $k = 1$, we simply say that f is equidifferentiable at x_0.*

THEOREM 4.1. *(Derivatives are Differential Quotients) Let $f : D \subset \mathcal{R} \to \mathcal{R}$ be a function that is k-equidifferentiable at the point $x_0 \in D$ for some $k > 0$ in Q. Let h be such that $|h| \ll d^r$, and $x_0 + h \in D$, where d is the positive infinitely small number introduced in [Berz1992b], [Berz1994a], [Berz1996a], and r is a given rational number. Then the derivative of f satisfies*

$$f'(x_0) =_{k \cdot r} \frac{f(x_0 + h) - f(x_0)}{h},$$

which means that the difference between the derivative and the differential quotient is at most infinitely smaller in absolute value than $d^{k\cdot r}$. In particular, the real part of the derivative can be calculated exactly from the differential quotient for any infinitely small h.

Proof. Let h be as in the requirements, then $h = h_0 d^{r_h}(1+h_1)$, where $h_0 \in R$, $|h_1|$ at most infinitely small, and $r_h > r$. Choose now $\epsilon = d^{k\cdot(r+r_h)/2}$; since f is k-equidifferentiable at x_0, we can find a positive $\delta \sim \epsilon^{1/k} = d^{(r+r_h)/2}$ such that for any Δx with $|\Delta x| < \delta$, the differential quotient differs by less than ϵ from the derivative, and hence $|\{f(x_0+\Delta x) - f(x_0)\}/\Delta x - f'(x_0)|$ is infinitely smaller than $d^{k\cdot r}$. But the above h clearly satisfies $|h| < \delta$. □

DEFINITION 4.2. *(Continuation of Real Computer Functions) Let f be a real computer function. Then f is given around any given real point of its domain of definition in D_c by a finite combination of roots and power series. Since roots and power series have already been extended to $\mathcal{R}$ [Berz1992b], [Berz1994a], [Berz1996a], f is extended to $\mathcal{R}$ in a natural way similar to that of the extension of power series from R to C.*

THEOREM 4.2. *Let f be a computer function that is differentiable at the point $x_0 \in D_c$. Then the continued function $\bar{f}$ is k-equidifferentiable at x_0 for some positive rational number k; and the derivatives of f and $\bar{f}$ at x_0 agree.*

Proof. Since f is differentiable at x_0, there exists a positive real number σ such that, for $x \in R$ and $0 < x < \sigma$, $f(x_0 \pm x) = f(x_0) \pm f'(x_0)x + \sum_{i=2}^{\infty} \alpha_i^{\pm} x^i + \sum_{j=1}^{J^{\pm}} x^{q_j^{\pm}} A_j^{\pm}(x)$; where $q_1^{\pm}, \ldots, q_{J^{\pm}}^{\pm}$ are noninteger rational numbers greater than 1, and $A_0^{\pm}, A_1^{\pm}, \ldots, A_{J^{\pm}}^{\pm}$ are power series in x. Let

$$q^{\pm} = \begin{cases} \min\{q_j^{\pm}; 1 \le j \le J^{\pm}\} & \text{if } \{q_j^{\pm}; 1 \le j \le J^{\pm}\} \neq \emptyset \\ \infty & \text{if } \{q_j^{\pm}; 1 \le j \le J^{\pm}\} = \emptyset \end{cases},$$

let $q = \min(q^+, q^-)$, and let $k = \min\{1, q-1\}$. Then $0 < k \le 1$. We show that the continued function $\bar{f}$ is k-equidifferentiable at x_0, with derivative $\bar{f}'(x_0) = f'(x_0)$.

Let $x \in \mathcal{R}$ satisfy $0 < x < \sigma$ and $x \not\approx \sigma$. Then $\bar{f}(x_0 \pm x) = f(x_0) \pm f'(x_0)x + \sum_{i=2}^{\infty} \alpha_i^{\pm} x^i + \sum_{j=1}^{J^{\pm}} x^{q_j^{\pm}} A_j^{\pm}(x)$. We have that

$$\left| \frac{\bar{f}(x_0 \pm x) - f(x_0)}{(\pm x)} - f'(x_0) \right| = \left| \pm \sum_{i=2}^{\infty} \alpha_i^{\pm} x^{i-1} \pm \sum_{j=1}^{J^{\pm}} x^{q_j^{\pm}-1} A_j^{\pm}(x) \right|.$$

Let $\epsilon \in \mathcal{R}$ be positive and at most finite. As a first case, assume ϵ is finite, and let $\epsilon_r = \Re(\epsilon)$, the real part of ϵ. Since the limit of $\left| \pm \sum_{i=2}^{\infty} \alpha_i^{\pm} y^{i-1} \pm \sum_{j=1}^{J^{\pm}} y^{q_j^{\pm}-1} A_j^{\pm}(y) \right|$, as $y \to 0^+, y \in R$, is equal to zero, there exists a real δ, $0 < \delta < \sigma/2$, such that

$$\left| \frac{f(x_0 \pm y) - f(x_0)}{(\pm y)} - f'(x_0) \right| < \frac{\epsilon_r}{2} \text{ whenever } y \in R \text{ and } 0 < y < 2\delta.$$

Now let $x \in \mathcal{R}$ be such that $0 < x < \delta$, and let $x_r = \Re(x)$. If $x_r = 0$, then x is infinitely small. Thus $|\{\bar{f}(x_0 \pm x) - f(x_0)\}/(\pm x) - f'(x_0)|$ is infinitely small, and hence smaller than the finite ϵ. If $x_r \neq 0$, then $0 < x_r < 2\delta$. Therefore,

$$\left| \frac{\bar{f}(x_0 \pm x) - f(x_0)}{(\pm x)} - f'(x_0) \right| =_0 \left| \frac{f(x_0 \pm x_r) - f(x_0)}{(\pm x_r)} - f'(x_0) \right| < \frac{\epsilon_r}{2}.$$

Hence, $|\{\bar{f}(x_0 + x) - f(x_0)\}/x - f'(x_0)| < \epsilon$ whenever $0 < |x| < \delta$. Note that, since ϵ and δ are both finite, $\delta^k \sim \epsilon$.

As a second case, assume ϵ is infinitely small. Let

$$m^{\pm} = \begin{cases} \min\{i \geq 2 : \alpha_i^{\pm} \neq 0\} & \text{if } \{i \geq 2 : \alpha_i^{\pm} \neq 0\} \neq \emptyset \\ \infty & \text{if } \{i \geq 2 : \alpha_i^{\pm} \neq 0\} = \emptyset \end{cases}.$$

If $m^{\pm} = \infty$, let $\alpha_{m^{\pm}}^{\pm} = 0$. With the convention $1/0 = \infty$, let

$$\delta = \min\left\{ \left(\epsilon/|A_1^+(0)|\right)^{1/k}, \left(\epsilon/|A_1^-(0)|\right)^{1/k}, \left(\epsilon/|\alpha_{m^+}^+|\right)^{1/k}, \left(\epsilon/|\alpha_{m^-}^-|\right)^{1/k} \right\}.$$

Then $\delta^k \sim \epsilon$, and if $0 < |x| < \delta$, then $|\{\bar{f}(x_0 + x) - f(x_0)\}/x - f'(x_0)| < \epsilon$. Thus $\bar{f}$ is k-equidifferentiable at 0, and $\bar{f}'(x_0) = f'(x_0)$. □

COROLLARY 4.1. *Let f be a real computer function that is differentiable at $x_0 \in D_c$, and let $\bar{f}$ be the continued function . Then we have that*

$$f'(x_0) =_0 \frac{\bar{f}(x_0 + d) - f(x_0)}{d}.$$

Having built the necessary theoretical tools, we next try to use the results of this section to compute derivatives of real functions. In the rest of this paper, we will use f instead of $\bar{f}$ to represent the continuation of a real computer function f.

5 Computation of Derivatives

In this section, we develop a criterion that will allow us not only to check the continuity and the differentiability of a real computer function f at a real point x_0, but also to obtain all existing derivatives of f at x_0.

LEMMA 5.1. *Let f be a computer function. Then f is defined at x_0 if and only if $f(x_0)$ can be computed on a computer.*

This lemma hinges on a careful implementation of the intrinsic functions and operations, in particular in the sense that they should be executable for any floating point number in the domain of definition that produces a result within the range of allowed floating point numbers.

LEMMA 5.2. *Let f be a computer function, and let x_0 be such that $f(x_0 - d)$, $f(x_0)$, and $f(x_0 + d)$ are all defined. Then f is continuous at x_0 if and only if*

$$f(x_0 - d) =_0 f(x_0) =_0 f(x_0 + d).$$

If $f(x_0)$ and $f(x_0 + d)$ are defined, but $f(x_0 - d)$ is not, then f is continuous at x_0 if and only if $f(x_0 + d) =_0 f(x_0)$.

Finally, if $f(x_0)$ and $f(x_0 - d)$ are defined, but $f(x_0 + d)$ is not, then f is continuous at x_0 if and only if $f(x_0 - d) =_0 f(x_0)$.

Proof. We prove the first part of the lemma; the proofs of the two other parts follow similar arguments. Since f is a computer function and $f(x_0 - d)$ and $f(x_0 + d)$ are defined, we have that

$$f(x_0 + x) = A_0(x) + \sum_{j=1}^{J_r} x^{q_j} A_j(x) \text{ and } f(x_0 - x) = B_0(x) + \sum_{j=1}^{J_l} x^{t_j} B_j(x)$$

for $0 < x < \sigma$, where σ is a positive real number; where the A_j's and the B_j's are power series in x, where $A_j(0) \neq 0$ for $1 \leq j \leq J_r$ and $B_j(0) \neq 0$ for $1 \leq j \leq J_l$; and where the q_j's and the t_j's are nonzero rational numbers that are not positive integers. Let $A_0(x) = \sum_{i=0}^{\infty} \alpha_i x^i$ and $B_0(x) = \sum_{i=0}^{\infty} \beta_i x^i$. Then f is continuous at x_0 if and only if $q_j > 0$ for all $j \in \{1, \ldots, J_r\}$, $t_j > 0$ for all $j \in \{1, \ldots, J_l\}$, and $\alpha_0 = \beta_0 = f(x_0)$; that is, if and only if $f(x_0 + d) =_0 f(x_0) =_0 f(x_0 - d)$. □

THEOREM 5.1. *Let f be a computer function that is continuous at x_0, and let $f(x_0 - d)$ and $f(x_0 + d)$ be both defined. Then f is differentiable at x_0 if and only if*

$$\frac{f(x_0 + d) - f(x_0)}{d} \text{ and } \frac{f(x_0) - f(x_0 - d)}{d}$$

are both at most finite in absolute value, and their real parts agree. In this case,

$$\frac{f(x_0 + d) - f(x_0)}{d} =_0 f'(x_0) =_0 \frac{f(x_0) - f(x_0 - d)}{d}.$$

If f is differentiable at x_0, then f is twice differentiable at x_0 if and only if

$$\frac{f(x_0 + 2d) - 2f(x_0 + d) + f(x_0)}{d^2} \text{ and } \frac{f(x_0) - 2f(x_0 - d) + f(x_0 - 2d)}{d^2}$$

are both at most finite in absolute value, and their real parts agree. In this case

$$\frac{f(x_0 + 2d) - 2f(x_0 + d) + f(x_0)}{d^2} =_0 f^{(2)}(x_0) =_0 \frac{f(x_0) - 2f(x_0 - d) + f(x_0 - 2d)}{d^2}.$$

In general, if f is $(n-1)$ times differentiable at x_0, then f is n times differentiable at x_0 if and only if

$$d^{-n} \left(\sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} f(x_0 + jd) \right) \text{ and } d^{-n} \left(\sum_{j=0}^{n} (-1)^{j} \binom{n}{j} f(x_0 - jd) \right)$$

are both at most finite in absolute value, and their real parts agree. In this case,

$$d^{-n} \left(\sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} f(x_0 + jd) \right) =_0 f^{(n)}(x_0) =_0 d^{-n} \left(\sum_{j=0}^{n} (-1)^{j} \binom{n}{j} f(x_0 - jd) \right).$$

Proof. Since f is continuous at x_0, we have that

$$\begin{aligned} f(x_0 + x) &= f(x_0) + \sum_{i=1}^{\infty} \alpha_i x^i + \sum_{j=1}^{J_r} x^{q_j} A_j(x) \\ f(x_0 - x) &= f(x_0) + \sum_{i=1}^{\infty} \beta_i x^i + \sum_{j=1}^{J_l} x^{t_j} B_j(x) \end{aligned} \tag{6}$$

for $0 < x < \sigma$, where σ is a positive real number, where the A_j's and the B_j's are power series in x that do not vanish at $x = 0$, and where the q_j's and the t_j's are noninteger positive rational numbers. Observe that f is n times differentiable at x_0 if and only if

$$q_j > n \text{ for } 1 \le j \le J_r,\ t_j > n \text{ for } 1 \le j \le J_l, \text{ and } \alpha_j = (-1)^j \beta_j \text{ for } 1 \le j \le n. \tag{7}$$

Assume f is differentiable at x_0. Then, using (7), we have that

$$q_j > 1\ \forall\, j \in \{1, \ldots, J_r\},\ t_j > 1\ \forall\, j \in \{1, \ldots, J_l\}, \text{ and } \alpha_1 = -\beta_1 = f'(x_0).$$

Hence,

$$\frac{f(x_0+d)-f(x_0)}{d} = \sum_{i=1}^{\infty} \alpha_i d^{i-1} + \sum_{j=1}^{J_r} d^{q_j-1} A_j(d) =_0 \alpha_1 = f'(x_0).$$

Similarly,

$$\frac{f(x_0)-f(x_0-d)}{d} = -\sum_{i=1}^{\infty} \beta_i d^{i-1} - \sum_{j=1}^{J_l} d^{t_j-1} B_j(d) =_0 -\beta_1 = f'(x_0).$$

Combining the above two equations, we obtain that

$$\frac{f(x_0+d)-f(x_0)}{d} =_0 f'(x_0) =_0 \frac{f(x_0)-f(x_0-d)}{d}.$$

Now assume that $(f(x_0+d)-f(x_0))/d$ and $(f(x_0)-f(x_0-d))/d$ are both at most finite in absolute value, and their real parts agree. Then, using (6), $|\sum_{i=1}^{\infty} \alpha_i d^{i-1} + \sum_{j=1}^{J_r} d^{q_j-1} A_j(d)|$ and $|-\sum_{i=1}^{\infty} \beta_i d^{i-1} - \sum_{j=1}^{J_l} d^{t_j-1} B_j(d)|$ are both at most finite, and

$$\sum_{i=1}^{\infty} \alpha_i d^{i-1} + \sum_{j=1}^{J_r} d^{q_j-1} A_j(d) =_0 -\sum_{i=1}^{\infty} \beta_i d^{i-1} - \sum_{j=1}^{J_l} d^{t_j-1} B_j(d).$$

Hence,

$$q_j > 1 \ \forall\, j \in \{1,\ldots,J_r\},\ t_j > 1 \ \forall\, j \in \{1,\ldots,J_l\}, \text{ and } \alpha_1 = -\beta_1,$$

from which we infer, using (7), that f is differentiable at x_0 with

$$f'(x_0) = \alpha_1 = -\beta_1 =_0 \frac{f(x_0+d)-f(x_0)}{d} =_0 \frac{f(x_0)-f(x_0-d)}{d}.$$

This finishes the proof of the first part of the theorem.

Since the second part of the theorem is only a special case of the last one, with $n=2$, we will go directly to proving the last part of the theorem. Note that since f is $(n-1)$ times differentiable at x_0,

$$\begin{aligned}
f(x_0+x) &= \sum_{i=0}^{n-1} \frac{f^{(i)}(x_0)}{i!} x^i + \sum_{i=n}^{\infty} \alpha_i x^i + \sum_{j=1}^{J_r} x^{q_j} A_j(x) \\
f(x_0-x) &= \sum_{i=0}^{n-1} (-1)^i \frac{f^{(i)}(x_0)}{i!} x^i + \sum_{i=n}^{\infty} \beta_i x^i + \sum_{j=1}^{J_l} x^{t_j} B_j(x)
\end{aligned}$$

for $0 < x < \sigma$, where σ is a positive real number, where the A_j's and the B_j's are as before, and where the q_j's and the t_j's are noninteger rational numbers greater than $n-1$.

Assume f is n times differentiable at x_0. Then

$$q_j > n \ \forall\, j \in \{1,\ldots,J_r\},\ t_j > n \ \forall\, j \in \{1,\ldots,J_l\},\ n!\,\alpha_n = (-1)^n n!\,\beta_n = f^{(n)}(x_0).$$

It can be shown by induction on n that

$$\begin{aligned}
d^{-n}\left(\sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} f(x_0+jd)\right) &=_0 \quad n!\,\alpha_n, \text{ and} \\
d^{-n}\left(\sum_{j=0}^{n} (-1)^{j} \binom{n}{j} f(x_0-jd)\right) &=_0 \quad (-1)^n n!\,\beta_n.
\end{aligned}$$

Therefore,

$$\begin{aligned} d^{-n}\left(\sum_{j=0}^{n}(-1)^{n-j}\begin{pmatrix} n \\ j \end{pmatrix} f\left(x_0+jd\right)\right) &=_0 f^{(n)}(x_0) \\ &=_0 d^{-n}\left(\sum_{j=0}^{n}(-1)^{j}\begin{pmatrix} n \\ j \end{pmatrix} f\left(x_0-jd\right)\right). \end{aligned}$$

Now assume that

$$\begin{aligned} &d^{-n}\left(\sum_{j=0}^{n}(-1)^{n-j}\begin{pmatrix} n \\ j \end{pmatrix} f\left(x_0+jd\right)\right) \text{ and} \\ &d^{-n}\left(\sum_{j=0}^{n}(-1)^{j}\begin{pmatrix} n \\ j \end{pmatrix} f\left(x_0-jd\right)\right) \end{aligned}$$

are both at most finite in absolute value, and their real parts agree. Then

$$q_j > n \; \forall \; j \in \{1,\ldots,J_r\}, \; t_j > n \; \forall \; j \in \{1,\ldots,J_l\}, \text{ and } n! \; \alpha_n = (-1)^n n! \; \beta_n,$$

from which we infer, again using (7), that f is n times differentiable at x_0 with

$$\begin{aligned} f^{(n)}(x_0) = n! \; \alpha_n = (-1)^n n! \; \beta_n &=_0 d^{-n}\left(\sum_{j=0}^{n}(-1)^{n-j}\begin{pmatrix} n \\ j \end{pmatrix} f\left(x_0+jd\right)\right) \\ &=_0 d^{-n}\left(\sum_{j=0}^{n}(-1)^{j}\begin{pmatrix} n \\ j \end{pmatrix} f\left(x_0-jd\right)\right). \end{aligned}$$

This finishes the proof of the theorem. □

Since knowledge of $f(x_0-d)$ and $f(x_0+d)$ gives us all the information about a computer function f in a real positive radius σ around x_0, we have the following result which states that, from the mere knowledge of $f(x_0-d)$ and $f(x_0+d)$, we can find at once the order of differentiability of f at x_0 and the accurate values of all existing derivatives.

THEOREM 5.2. *Let f be a computer function that is continuous at x_0. Then f is n times differentiable at x_0 if and only if $f(x_0-d)$ and $f(x_0+d)$ are both defined and can be written as*

$$f(x_0-d) =_n f(x_0) + \sum_{j=1}^{n}(-1)^j \alpha_j d^j \text{ and } f(x_0+d) =_n f(x_0) + \sum_{j=1}^{n} \alpha_j d^j,$$

where the α_j's are real numbers. Moreover, in this case $f^{(j)}(x_0) = j! \; \alpha_j$ for $1 \le j \le n$.

REMARK 5.1. *The theorem above is similar in flavor to the Pointformula à la Cauchy [Berz1992b], [Berz1994a], [Berz1996a], which holds for the continuations of power series around a real point x_0. In the latter case, the continued function is completely determined by its value at x_0+h for any arbitrary nonzero h infinitely small in absolute value.*

In the following section, we apply our theory to find the order of differentiability and all existing derivatives (at zero) of two functions for which the traditional methods of AD fail.

6 Examples

As a first example, we consider a function mentioned in the introduction and study its differentiability at 0.

Example 1: Consider the function $f(x) = x^2\sqrt{|x|} + \exp(x)$. It is easy to show that f is twice differentiable at 0 with $f(0) = f'(0) = f^{(2)}(0) = 1$ and that f is not three times differentiable at 0. We will show now how using the result of Theorem (5.1) will lead us to the same conclusion. First we note that $f(-d)$, $f(0)$, and $f(d)$ are all defined.

It is useful to look at what goes on inside the computer for this simple example. Altogether, we need seven memory locations to store the variable, the intermediate values, and the function value. These seven memory locations are

$$\begin{array}{llll} x, & S_1 = \mathrm{abs}(x), & S_2 = \mathrm{sqrt}(S_1), & S_3 = x * x, \\ S_4 = S_2 * S_3, & S_5 = \exp(x), & a = S_4 + S_5. & \end{array}$$

Hence, we can look at $\vec{F}(f)$ as a function from R^7 into R^7. Let

$$\begin{cases} \vec{E} : R \to R^7; & \vec{E}(x) = (x, 0, 0, 0, 0, 0, 0) \\ \vec{F} : R^7 \to R^7; & \vec{F}(x, p_2, p_3, p_4, p_5, p_6, p_7) = (x, S_1, S_2, S_3, S_4, S_5, a) \\ P : R^7 \to R; & P(x, S_1, S_2, S_3, S_4, S_5, a) = a \\ G : R \to R; & G(x) = P \circ \vec{F} \circ \vec{E}(x). \end{cases}$$

Then $G(x) = a =_M f(x)$, where M is an upper bound of the support points that can be obtained on the computer.

If we enter the value $x = -d$, then the seven memory locations will be filled as follows:

$$x = -d, \quad S_1 = d, \quad S_2 = d^{1/2}, \quad S_3 = d^2,$$

$$S_4 = d^{5/2}, \quad S_5 = \textstyle\sum_{j=0}^{M} (-1)^j d^j / j!, \quad a = d^{5/2} + \sum_{j=0}^{M} (-1)^j d^j / j!.$$

Hence, the output is $G(-d) = 1 - d + d^2/2! + d^{5/2} + \sum_{j=3}^{M} (-1)^j d^j / j! =_M f(-d)$.

Similarly, we find that $G(0) = 1 = f(0)$, and

$$\begin{aligned} G(d) &= 1 + d + d^2/2! + d^{5/2} + \sum_{j=3}^{M} d^j / j! =_M f(d) \\ G(-2d) &= 1 - 2d + 2d^2 + 2^{5/2} d^{5/2} + \sum_{j=3}^{M} (-2)^j d^j / j! =_M f(-2d) \\ G(2d) &= 1 + 2d + 2d^2 + 2^{5/2} d^{5/2} + \sum_{j=3}^{M} 2^j d^j / j! =_M f(2d) \\ G(-3d) &= 1 - 3d + 9d^2/2 + 3^{5/2} d^{5/2} + \sum_{j=3}^{M} (-3)^j d^j / j! =_M f(-3d) \\ G(3d) &= 1 + 3d + 9d^2/2 + 3^{5/2} d^{5/2} + \sum_{j=3}^{M} 3^j d^j / j! =_M f(3d). \end{aligned}$$

Since $f(d) =_0 1 = f(0) =_0 f(-d)$, f is continuous at 0. A simple computation shows that

$$\frac{f(d) - f(0)}{d} =_0 1 =_0 \frac{f(0) - f(-d)}{d},$$

from which we infer that f is differentiable at 0, with $f'(0) = 1$. Also

$$\frac{f(2d) - 2f(d) + f(0)}{d^2} =_0 1 =_0 \frac{f(0) - 2f(-d) + f(-2d)}{d^2},$$

from which we conclude that f is twice differentiable at 0, with $f^{(2)}(0) = 1$. On the other hand,

$$\frac{f(3d) - 3f(2d) + 3f(d) - f(0)}{d^3} =_0 d^{-1/2}\left(3^{5/2} - 2^{5/2}3 + 3\right) + 1,$$

from which we readily obtain that $|(f(3d) - 3f(2d) + 3f(d) - f(0))/d^3|$ is infinitely large. Hence, f is not three times differentiable at 0.

In the following example, we study a function which appears in many physics problems and which is infinitely often differentiable everywhere, including at 0.

Example 2: The electric field of a spherical Gaussian charge is given, up to a normalizing constant, by

$$f(x) = \begin{cases} \{1 - \exp(-x^2)\}/x & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}, \tag{8}$$

where x is the radial distance from the origin of the charge.

It is easy to check that $f(x) = \sum_{j=0}^{\infty}(-1)^j x^{2j+1}/(j+1)!$ for all $x \in R$, where the infinite series converges for all $x \in R$. Hence, f is infinitely often differentiable at 0.

Next we show that application of Theorem (5.2) to the function in (8) not only proves the differentiability of f at 0 up to a very high order, but also allows us to obtain all derivatives at once. Evaluating $f(-d)$ and $f(d)$ on the computer yields

$$f(-d) =_M \sum_{j=0}^{[(M-1)/2]} (-1)^{j+1}\frac{d^{2j+1}}{(j+1)!} \text{ and } f(d) =_M \sum_{j=0}^{[(M-1)/2]} (-1)^j \frac{d^{2j+1}}{(j+1)!},$$

where $[(M-1)/2]$ is the largest integer that does not exceed $(M-1)/2$. Applying Theorem (5.2), we obtain for all k, $0 \leq k \leq [(M-1)/2]$, that

$$f^{(2k)}(0) = 0 \text{ and } f^{(2k+1)}(0) = (-1)^k \frac{(2k+1)!}{(k+1)!}.$$

The two methods discussed above for computing derivatives of real computer functions can be of practical use only if we can implement the $\mathcal{R}$ numbers on a computer. We do have a first version of the implementation using COSY INFINITY [Berz1995a], [Berz1996b], and in the following section we show briefly how this is done.

7 Implementation

Besides allowing illuminating theoretical conclusions, the strength of the $\mathcal{R}$ numbers is that they can be used in practice, and even in a computer environment. In this respect, they differ from the non-constructive structures in Non-Standard Analysis [Laugwitz1973a], [Robinson1974a].

An implementation of the $\mathcal{R}$ numbers is not as direct as one of the Differential Algebras [Berz1989a] since $\mathcal{R}$ is infinite dimensional. However, as we shall see now, it is still possible to implement the structure in a very useful way. Since there are only finitely many support points below every bound, it is possible to pick any such bound and store all the values of a function to the left of it. Hence, each $\mathcal{R}$ number is represented by these values as well as the value of the bound.

The sum of two such numbers can then be computed for all values to the left of the minimum of the two bounds. Hence, the minimum of the bounds is the bound of the sum. In a similar way, it is possible to find a bound below which the product of two such numbers can be computed from the bounds of the two numbers. Altogether, the bound to which each individual number is known is carried along through all arithmetic.

8 Computer Functions of Many Variables

Since we know now how to compute the nth order derivative of a real computer function of one variable at a given real point x_0 whenever the nth order derivative exists, the following lemma shows how to find all nth order partial derivatives at a given real point $\vec{p}_0$ of a function $f : R^m \to R$ which can be represented on a computer whenever all the nth order partial derivatives exist and are continuous in a neighborhood of $\vec{p}_0$.

Lemma 8.1. *Let $f : R^m \to R$ be a function representable on a computer whose nth order partial derivatives exist and are continuous in the neighborhood of the point $\vec{p}_0 = (x_{01}, x_{02}, \ldots, x_{0m})$. Then the nth order partial derivatives of f at $\vec{p}_0$ can always be computed in terms of nth order derivatives of real computer functions of one variable.*

Proof. Let l be the number of nth order partial derivatives of f. We note in passing that it can be shown [Berz1989a] by induction on n and m that $l = (n+m-1)!/(n!\,(m-1)!)$. Let $k = l \cdot m$, and let $p_1, p_2, \ldots, p_k$ denote the first k prime numbers. For $j = 1, \ldots, k$, let $\alpha_j = \sqrt[n+1]{p_j}$. For $i = 1, \ldots, l$, let

$$f_i(x) = f(x_{01} + \alpha_{(i-1)m+1}x, x_{02} + \alpha_{(i-1)m+2}x, \ldots, x_{0m} + \alpha_{im}x).$$

Then $f_i, i = 1, \ldots, l$, are l real computer functions of x, n times differentiable at 0. Evaluating $(d^n f_i/dx^n)|_{x=0}$ for $i = 1, \ldots, l$ yields l equations in the l unknowns

$$\left.\frac{\partial^n f}{\partial x_1^{n_1} \partial x_2^{n_2} \ldots \partial x_m^{n_m}}\right|_{\vec{p}=\vec{p}_0}, \text{ with } \left\{ \begin{array}{l} n_1, n_2, \ldots, n_m \in \{0, 1, \ldots, n\}, \text{ and} \\ n_1 + n_2 + \cdots + n_m = n \end{array} \right. .$$

The matrix $\widehat{M}$ of the coefficients has as entries products of integers with the different α's raised to exponents between 0 and n. In the ith row, we have only products of the form $c_{n;n_1,n_2,\ldots,n_m}\alpha_{(i-1)m+1}^{n_1}\alpha_{(i-1)m+2}^{n_2}, \ldots, \alpha_{im}^{n_m}$, where $c_{n;n_1,n_2,\ldots,n_m}$ is a positive integer. The determinant of $\widehat{M}$ is the sum of $l!$ terms, each of which is the product of a positive integer and the α's raised to exponents less than or equal to n, and such that not all the exponents in any one term agree with those in any of the remaining $(l!-1)$ terms. By our choice of the α's, no cancellation in the evaluation of the determinant can occur. Hence, $\det \widehat{M} \neq 0$. □

It is worth noting that the choice of the α's above is far from being the only one possible. Let $\alpha_1, \alpha_2, \ldots, \alpha_k$ be any set of k real numbers. We look at $\det \widehat{M}$ as a function from R^k into R. A purely statistical argument shows that it is very unlikely that $\det \widehat{M}$ be zero for a given choice of numbers. We are led to believe that there exist even uncountably many choices of $(\alpha_1, \alpha_2, \ldots, \alpha_k) \in R^k$ that give a nonvanishing determinant. Here we provide simpler choices of the α's only in the case $m = 2$: For $m = 2$, we have that $l = n+1$ and $k = 2(n+1)$. For $i = 1, 2, \ldots, n+1$, let $\alpha_{2i-1} = 1$ and $\alpha_{2i} = \beta_{i-1}$, where $\beta_0 = 0$ and $\beta_{j_1} \neq \beta_{j_2}$ if $j_1 \neq j_2$ in $\{0, 1, \ldots, n\}$.

References

[Berz1989a] M. BERZ, *Differential algebraic description of beam dynamics to very high orders*, Particle Accelerators, 24 (1989), p. 109.

[Berz1992b] ——, *Automatic differentiation as nonarchimedean analysis*, in Computer Arithmetic and Enclosure Methods, Amsterdam, 1992, Elsevier Science Publishers.

[Berz1994a] ——, *Analysis on a nonarchimedean extension of the real numbers*, Lecture Notes, Studienstiftung Summer School, September 1992 MSUCL–933, National Superconducting Cyclotron Laboratory, Michigan State University, East Lansing, Mich., 1994.

[Berz1995a] ——, *COSY INFINITY Version 7 reference manual*, Tech. Report MSUCL–977, National Superconducting Cyclotron Laboratory, Michigan State University, East Lansing, Mich., 1995.

[Berz1996a] ——, *Calculus and numerics on Levi-Civita fields*, in Computational Differentiation: Techniques, Applications, and Tools, M. Berz, C. Bischof, G. Corliss, and A. Griewank, eds., Philadelphia, Penn., 1996, SIAM, pp. 19–35.

[Berz1996b] M. BERZ, K. MAKINO, K. SHAMSEDDINE, G. H. HOFFSTÄTTER, AND W. WAN, *COSY INFINITY and its applications to nonlinear dynamics*, in Computational Differentiation: Techniques, Applications, and Tools, M. Berz, C. Bischof, G. Corliss, and A. Griewank, eds., SIAM, Philadelphia, Penn., 1996, pp. 363–365.

[Laugwitz1973a] D. LAUGWITZ, *Ein Weg zur Nonstandard-Analysis*, Jahresberichte der Deutschen Mathematischen Vereinigung, 75 (1973), pp. 66–93.

[Osgood1938a] W. F. OSGOOD, *Functions of Real Variables*, G. E. Stechert, New York, 1938.

[Robinson1974a] A. ROBINSON, *Non-Standard Analysis*, North-Holland, 1974.