Reliable Computing 4: 1-5, 1998

A NOTE ON EPSILON-INFLATION

S. M. RUMP*

Abstract. The epsilon-inflation proved to be useful and necessary in many verification algorithms. Different definitions of an epsilon-inflation are possible, depending on the context. Recently, certain theoretical justifications and optimality results were proved for an epsilon-inflation without absolute term. In this note we show that in currently used interval iterations the epsilon-inflation without absolute term does not serve the purpose it is defined for. A new epsilon-inflation is proposed.

Many verification algorithms for calculating an inclusion of the solution of a given problem use Banach's or Brouwer's Fixed Point theorem. The main point of those algorithms is to verify that a certain interval is mapped into itself or into its interior.

We assume the reader is familiar with the fact that this self-mapping is the central part of many verification algorithms for systems of linear or nonlinear equations, algebraic eigenproblems, polynomial zeros and others. References include [2], [9], [12], [13] and many more. For an overview see e.g. [7], commercial implementations include [1], [3], [8], [16].

If this self-mapping cannot be verified for the initial test interval, an interval iteration is started. To the author's knowledge, it was first noted by Caprani and Madsen [5] that it is useful to enlarge the computed iterates prior to the next iteration in order to increase chances for a self-mapping.

The term epsilon-inflation was introduced in [13]. For a real interval X the original definition is [13, Definition 2.6],

$$X \circ \varepsilon := \begin{cases} X + d(X) \cdot [-\varepsilon, \varepsilon] & \text{for } d(X) \neq 0 \\ X + [-\eta, +\eta] & \text{otherwise,} \end{cases}$$

where d denotes the diameter and η denotes the smallest representable positive machine number.

In later papers an analysis of the benefits of the epsilon-inflation was given (cf. [14]). These results can be summarized as follows. Let Z, $X^0 \in \mathbb{IK}^n$ be interval vectors, and let $C \in \mathbb{I}M_n(\mathbb{K})$ be an $n \times n$ interval matrix for $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Define the interval iteration

$$(1.1) \qquad Y^k := X^k \circ \varepsilon \quad \text{and} \quad X^{k+1} := Z + C \cdot Y^k \quad \text{for } k \geq 0 \,.$$

Using the simplified definition

$$(1.2) \qquad X \circ \varepsilon := X + d(X) \cdot [-\varepsilon, \varepsilon] + [-\eta, +\eta]$$

of the epsilon-inflation in (1.1), the following is true ($|C|$ is the matrix of entrywise absolute values of C; ρ denotes the spectral radius):

I) If interval operations are used in the iteration (1.1) and $\rho(|C|) < 1/(1 + 2\varepsilon)$, then the inclusion

$$X^{k+1} \subseteq \text{int}\,(Y^k)$$

is satisfied for some k. If interval operations are used in the iteration (1.1) and $X^{k+1} \subseteq \text{int}(Y^k)$ for some k, then $\rho(|C|) < 1$.

*Inst. f. Informatik III, Technical University Hamburg-Harburg, Eißendorfer Str. 38, 21071 Hamburg, Germany, Tel. (49)40-7718-3027, Fax (49)40-7718-2573, (rump@tu-harburg.de).

II) If power set operations are used in the iteration (1.1) and $\rho(C) < 1/(1+2\varepsilon)$, then the inclusion

$$X^{k+1} \subseteq \mathrm{int}\,(Y^k)$$

is satisfied for some k. If power set operations are used in the iteration (1.1) and $X^{k+1} \subseteq \mathrm{int}(Y^k)$ for some k, then $\rho(C) < 1$.

For $\varepsilon = 0$, which means that the epsilon-inflation contains only an absolute term, we have the beautiful equivalence that $X^{k+1} \subseteq \mathrm{int}(Y^k)$ will be satisfied for some k *if and only if* $\rho(|C|) < 1$ in case of interval operations, and *if and only if* $\rho(C) < 1$ in case of power set operations, respectively.

The results are, in fact, more general; for details see [14]. The results have been extended for P-contractions by Mayer in [11]. We mention that the results I) and II) are valid for arbitrary positive η.

A number of different definitions of the epsilon-inflation can be found in the literature. Recently, Kreinovich, Starks and Mayer gave in [10] theoretical justifications for the following epsilon-inflation used in PASCAL-XSC (see, e.g., [6]):

$$X \circ \varepsilon := X + d(X) \cdot [-\varepsilon, \varepsilon] \; . \tag{1.3}$$

For this type of epsilon-inflation they show certain optimality results.

However, the epsilon-inflation as defined in (1.3) does not serve the purpose it has been introduced for, because it lacks an absolute term. The most trivial example is

$$Z = 0,\ C = 0,\ X^0 = 0 \; ,$$

corresponding to a linear system $1 \cdot x = 0$. Obviously, $X^{k+1} \subseteq \mathrm{int}\,(Y^k)$ will never be satisfied in this case. The problem is that the initial interval vector X^0 has diameter zero.

But even if this is not the case, the absolute term in $X \circ \varepsilon := X + d(X) \cdot [-\varepsilon, \varepsilon] + [-\eta, +\eta]$ is, in general, necessary. More precisely, the following is true.

Theorem 1.1. *For all $\delta > 0$, there is an iteration matrix $C \in M_2(\mathbb{R})$ and an interval vector $X^0 \in \mathbb{IR}^2$ satisfying the following properties:*

i) $\rho(|C|) < \delta$,
ii) $d(X_i^0) > 0$ *for* $i \in \{1,2\}$,
iii) *For $k \geq 0$, let X^{k+1}, Y^k be defined by the iteration (1.1) with the epsilon-inflation (1.3). Then for all $\varepsilon > 0$ it is always*

$$X^{k+1} \not\subseteq Y^k \quad \textit{for all } k \geq 0 \; .$$

Proof. For arbitrary $0 < a < 1$ define

$$C = \begin{pmatrix} 0 & 1 \\ a & 0 \end{pmatrix},\ X^0 = \begin{pmatrix} [0,4] \\ [0,2a] \end{pmatrix} \quad \text{and } Z = \begin{pmatrix} 0 \\ 0 \end{pmatrix} . \tag{1.4}$$

For the analysis it is more convenient to use a midpoint-radius representation. Define

$$mX \pm rX := [mX - rX,\ mX + rX] \quad \text{for } mX,\ rX \in \mathbb{R}^2 \; .$$

Then

$$X^0 = \begin{pmatrix} 2 \\ a \end{pmatrix} \pm \begin{pmatrix} 2 \\ a \end{pmatrix} ,$$

and a short computation using

$$C \cdot (mX \pm rX) = C \cdot mX \pm |C| \cdot rX$$

yields

$$X^{2k} = a^k \cdot \left(\begin{pmatrix} 2 \\ a \end{pmatrix} \pm \begin{pmatrix} 2 \\ a \end{pmatrix} \cdot (1+2\varepsilon)^{2k} \right), \; X^{2k+1} = a^{k+1} \cdot \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} \pm \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot (1+2\varepsilon)^{2k+1} \right).$$

Then

$$\begin{aligned} X^{2k+1} &= a^k \cdot \left(\begin{pmatrix} a \\ 2a \end{pmatrix} \pm \begin{pmatrix} a \\ 2a \end{pmatrix} \cdot (1+2\varepsilon)^{2k+1} \right) \not\subseteq a^k \cdot \left(\begin{pmatrix} 2 \\ a \end{pmatrix} \pm \begin{pmatrix} 2 \\ a \end{pmatrix} \cdot (1+2\varepsilon)^{2k+1} \right) \\ &= Y^{2k} \;, \\ X^{2k+2} &= a^{k+1} \cdot \left(\begin{pmatrix} 2 \\ a \end{pmatrix} \pm \begin{pmatrix} 2 \\ a \end{pmatrix} \cdot (1+2\varepsilon)^{2k+2} \right) \not\subseteq a^{k+1} \cdot \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} \pm \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot (1+2\varepsilon)^{2k+2} \right) \\ &= Y^{2k+1} \;. \end{aligned}$$

Choosing any a with $0 < a < \delta^2$ proves the theorem. □

The reason for the observed behaviour is that the matrix C is not primitive, and therefore the power iteration for the Perron root of C does not necessarily converge for every starting vector (see [4] or [15]). This is the reason why (1.1) contains an absolute term like $[-\eta, \eta]$. An alternative is to replace C by some perturbed C' in order to force the new C' to be primitive.

As has been mentioned before, the results I) and II) are true for any choice of positive η. Choosing η too small increases the number of iterations, a large value of the absolute term η increases the diameter of the computed solution set. The choice of ε in (1.2) is critical: Choosing ε a little too large may make an inclusion impossible if $\rho(|C|) \cdot (1 + 2\varepsilon) \geq 1$.

From a practical point of view the following heuristic may be used:

$$(1.5) \qquad X \circ \varepsilon := X + \mathrm{d}(Y^0) \cdot [-e, +e] + [-\eta, \eta] \quad \text{where} \quad Y^0 := Z + CX^0 \,.$$

Note that the epsilon-inflation in (1.5) contains *only* an absolute term, independent of the current iterate. The reasoning for this heuristic is as follows. First of all, according to I) we have the best possible convergence behaviour:

$$X^{k+1} \subseteq \mathrm{int}(Y^k) \quad \text{for some } k \in \mathbb{N} \quad \text{iff} \quad \rho(|C|) < 1.$$

Furthermore, one might use $\mathrm{d}(X^0) \cdot [-e, +e]$ instead of $d(Y^0) \cdot [-e, +e]$ in the absolute term of (1.5). However, X^0 may consist of zero or small components due to a bad choice of the initial X^0 or bad scaling. In this case, only the very small absolute term $[-\eta, +\eta]$ would be operative, resulting in many iterations. Therefore, we choose the first iterate $Y^0 := Z + CX^0$ to define the absolute term. It is not likely that Y^0 still contains a zero component, and if, the additive term $[-\eta, +\eta]$ will do. Otherwise, the heuristic has the advantage that the components of Y^0 are already "adjusted" to the subsequent iteration, and they are of appropriate magnitude. To the author's experience, $e = 0.1$ or $e = 0.2$ are reasonable values if the error with respect to an approximate solution is to be included.

For $e = 0.1$ or $e = 0.2$, the iteration (1.1) with epsilon-inflation (1.5) stops for the data (1.4) for $a = 0.25$ (corresponding to $\rho(C) = 0.5$) after 5 iterations. For $a = 0.64$

(corresponding to $\rho(C) = 0.8$), 7 iterations for $e = 0.2$ and 9 iterations for $e = 0.1$ are necessary.

Acknowledgement. The author wishes to thank the anonymous referee for helpful comments.

REFERENCES

[1] ACRITH High-Accuracy Arithmetic Subroutine Library, Program Description and User's Guide. *IBM Publications*, (SC 33-6164-3), 1986.

[2] G. Alefeld. Rigorous Error Bounds for Singular Values of a Matrix Using the Precise Scalar Product. In E. Kaucher, U. Kulisch, and Ch. Ullrich, editors, *Computerarithmetic*, pages 9–30. Teubner Stuttgart, 1987.

[3] ARITHMOS, Benutzerhandbuch. Siemens AG, Bibl.-Nr. U 2900-I-Z87-1, 1986.

[4] A. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* SIAM classics in Applied Mathematics, Philadelphia, 1994.

[5] O. Caprani and K. Madsen. Iterative Methods for Interval Inclusion of Fixed Points. *BIT*, 18:42–51, 1978.

[6] M. Hammer, R.Hocks, U. Kulisch, and D. Ratz. *Numerical Toolbox for Verified Computing. I. Basic Numerical Problems.* Springer Verlag, Heidelberg, N.Y., 1993.

[7] J. Herzberger, editor. *Topics in Validated Computations — Studies in Computational Mathematics.* Elsevier, Amsterdam, 1994.

[8] R. Klatte, U. Kulisch, A. Wiethoff, C. Lawo, and M. Rauch. *C-XSC A C++ Class Library for Extended Scientific Computing.* Springer, Berlin, 1993.

[9] R. Krawczyk. Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. *Computing 4*, pages 187–201, 1969.

[10] V. Kreinovich, S. Starks, and G. Mayer. On a Theoretical Justification of the Choice of Epsilon-Inflation in PASCAL-XSC. *Reliable Computing.*, 3(4):437–445, 1997.

[11] G. Mayer. Epsilon-Inflation in Verification Algorithms. *Journal of Computational and Applied Mathematics*, 60:147–169, 1995.

[12] R.E. Moore. A Test for Existence of Solutions for Non-Linear Systems. *SIAM J. Numer. Anal. 4*, pages 611–615, 1977.

[13] S.M. Rump. *Kleine Fehlerschranken bei Matrixproblemen.* Dissertation, Universität Karlsruhe, 1980.

[14] S.M. Rump. On the Solution of Interval Linear Systems. *Computing 47*, pages 337–353, 1992.

[15] R.S. Varga. *Matrix Iterative Analysis.* Prentice-Hall, Englewood Cliffs, N.J., 1962.

[16] W.V. Walter. A Portable Fortran 90 Module Library for Accurate and Reliable Scientific Computing. *Computing*, Suppl.9:265–285, 1993.