SIAM J. NUMER. ANAL.
Vol. 42, No. 1, pp. 383–408

EXCLUSION REGIONS FOR SYSTEMS OF EQUATIONS*

HERMANN SCHICHL† AND ARNOLD NEUMAIER†

Abstract. Branch and bound methods for finding all zeros of a nonlinear system of equations in a box frequently have the difficulty that subboxes containing no solution cannot be easily eliminated if there is a nearby zero outside the box. This has the effect that near each zero many small boxes are created by repeated splitting, whose processing may dominate the total work spent on the global search.

This paper discusses the reasons for the occurrence of this so-called cluster effect and how to reduce the cluster effect by defining exclusion regions around each zero found that are guaranteed to contain no other zero and hence can safely be discarded.

Such exclusion regions are traditionally constructed using uniqueness tests based on the Krawczyk operator or the Kantorovich theorem. These results are reviewed; moreover, refinements are proved that significantly enlarge the size of the exclusion region. Existence and uniqueness tests are also given.

Key words. zeros, system of equations, validated enclosure, existence test, uniqueness test, inclusion region, exclusion region, branch and bound, cluster effect, Krawczyk operator, Kantorovich theorem, backboxing, affine invariant

AMS subject classifications. Primary, 65H20; Secondary, 65G30

DOI. 10.1137/S0036142902418898

1. Introduction.

Branch and bound methods for finding all zeros of a nonlinear system of equations in a box [10, 23] frequently have the difficulty that subboxes containing no solution cannot be easily eliminated if there is a nearby zero outside the box. This has the effect that near each zero many small boxes are created by repeated splitting, whose processing may dominate the total work spent on the global search.

This paper discusses in section 3 the reasons for the occurrence of this so-called cluster effect and how to reduce the cluster effect by defining exclusion regions around each zero found that are guaranteed to contain no other zero and hence can safely be discarded. Such exclusion boxes (possibly first used by Jansson [4]) are the basis for the backboxing strategy by van Iwaarden [24] (see also Kearfott [8, 9]) that eliminates the cluster effect near well-conditioned zeros.

Exclusion regions are traditionally constructed using uniqueness tests based on the Krawczyk operator (see, e.g., Neumaier [16, Chapter 5]) or the Kantorovich theorem (see, e.g., Ortega and Rheinboldt [19, Theorem 12.6.1]); both provide existence and uniqueness regions for zeros of systems of equations. Shen and Neumaier [22] proved that the Krawczyk operator with slopes always provides an existence region which is at least as large as that computed by Kantorovich's theorem. Deuflhard and Heindl [2] proved an affine invariant version of the Kantorovich theorem.

In section 2, these results are reviewed, together with recent works on improved preconditioning by Hansen [3] and on Taylor models by Berz and Hoefkens [1] that are related to our present work. In sections 4–7, we discuss componentwise and affine invariant existence, uniqueness, and nonexistence regions given a zero or any other

*Received by the editors December 3, 2002; accepted for publication (in revised form) May 13, 2003; published electronically March 3, 2004.

http://www.siam.org/journals/sinum/42-1/41889.html

†Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria (Hermann.Schichl@esi.ac.at, Arnold.Neumaier@univie.ac.at, http://www.mat.univie.ac.at/∼neum/).

point of the search region. They arise from a more detailed analysis of the properties of the Krawczyk operator with slopes used in [22].

Numerical examples given in section 8 show that the refinements introduced in this paper significantly enlarge the sizes of the exclusion regions.

In the following, the notation is as in the book [17]. In particular, inequalities are interpreted componentwise, I denotes the identity matrix, intervals and boxes (= interval vectors) are in boldface, and $\operatorname{rad}\mathbf{x} = \frac{1}{2}(\overline{x} - \underline{x})$ denotes the radius of a box $\mathbf{x} = [\underline{x}, \overline{x}] \in \mathbb{IR}^n$. The interior of a set $S \subseteq \mathbb{R}^n$ is denoted by $\operatorname{int}(S)$ and the interval hull by $\square S$.

We consider the nonlinear system of equations

$$F(x) = 0, \tag{1}$$

where $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is twice continuously differentiable in a convex domain D. (For some results, weaker conditions suffice; it will be clear from the arguments used that continuity and the existence of the quantities in the hypothesis of the theorems are sufficient.)

Since F is twice continuously differentiable, we can always (e.g., using the mean value theorem) write

$$F(x) - F(z) = F[z, x](x - z) \tag{2}$$

for any two points x and z with a suitable matrix $F[z, x] \in \mathbb{R}^{n\times n}$, continuously differentiable in x and z; any such $F[z, x]$ is called a *slope matrix* for F. While (in dimension $n > 1$) $F[z, x]$ is not uniquely determined, we always have (by continuity)

$$F[z, z] = F'(z). \tag{3}$$

Thus $F[z, x]$ is a slope version of the Jacobian. There are recursive procedures to calculate a slope $F[z, x]$, given x and z; see Krawczyk and Neumaier [14], Rump [20], and Kolev [13]; a Matlab implementation is in Intlab [21].

Since the slope matrix $F[z, x]$ is continuously differentiable, we can write similarly

$$F[z, x] = F[z, z'] + \sum (x_k - z'_k) F_k[z, z', x] \tag{4}$$

with *second order slope matrices* $F_k[z, z', x]$, continuous in z, z', x. Here, as throughout this paper, the summation extends over $k = 1, \dots, n$. Second order slope matrices can also be computed recursively; see Kolev [13]. Moreover, if F is quadratic, the slope is linear in x and z, and the coefficients of x determine constant second order slope matrices without any work.

If $z = z'$ the formula above somewhat simplifies, because of (3), to

$$F[z, x] = F'(z) + \sum (x_k - z_k) F_k[z, z, x]. \tag{5}$$

Throughout the paper we shall make the following assumption, without mentioning it explicitly.

Assumption A. The point z and the convex subset X lie in the domain of definition of F. The center, $z \in X$, and the second order slope (5) are fixed. Moreover, for a fixed preconditioning matrix $C \in \mathbb{R}^{m\times n}$, the componentwise bounds

$$\begin{aligned} \overline{b} &\geq |CF(z)| \geq \underline{b}, \\ B_0 &\geq |CF'(z) - I|, \\ B'_0 &\geq |CF'(z)|, \\ B_k(x) &\geq |CF_k[z, z, x]| \quad (k = 1, \dots, n) \end{aligned} \tag{6}$$

are valid for all $x \in X$.

Example 1.1. We consider the system of equations

$$x_1^2 + x_2^2 = 25, \quad x_1 x_2 = 12. \tag{7}$$

The system has the form (1) with

$$F(x) = \begin{pmatrix} x_1^2 + x_2^2 - 25 \\ x_1 x_2 - 12 \end{pmatrix}. \tag{8}$$

With respect to the center $z = \binom{3}{4}$, we have

$$F(x) - F(z) = \begin{pmatrix} x_1^2 - 3^2 + x_2^2 - 4^2 \\ x_1 x_2 - 3 \cdot 4 \end{pmatrix} = \begin{pmatrix} (x_1+3)(x_1-3) + (x_2+4)(x_2-4) \\ x_2(x_1-3) + 3(x_2-4) \end{pmatrix}$$

so that we can take

$$F[z,x] = \begin{pmatrix} x_1+3 & x_2+4 \\ x_2 & 3 \end{pmatrix}$$

as a slope. (Note that other choices would be possible.) The interval slope $F[z,\mathbf{x}]$ in the box $\mathbf{x} = [2,4] \times [3,5]$ is then

$$F[z,x] = \begin{pmatrix} [5,7] & [7,9] \\ [3,5] & 3 \end{pmatrix}.$$

The slope can be put in form (5) with

$$F'(z) = \begin{pmatrix} 6 & 8 \\ 4 & 3 \end{pmatrix}, \quad F_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad F_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and we obtain

$$B_1 = \frac{1}{14}\begin{pmatrix} 3 & 0 \\ 4 & 0 \end{pmatrix}, \qquad B_2 = \frac{1}{14}\begin{pmatrix} 8 & 3 \\ 6 & 4 \end{pmatrix}.$$

Since we calculated without rounding errors and z happens to be a zero of F, both B_0 and $\overline{b}$ vanish.

2. Known results. The oldest semilocal existence theorem for zeros of systems of equations is due to Kantorovich [7], who obtained as a by-product of a convergence guarantee for Newton's method (which is not of interest in our context) the following result.

THEOREM 2.1 (Kantorovich). *Let z be a vector such that $F'(z)$ is invertible, and let α and β be constants with*

$$\|F'(z)^{-1}\|_\infty \le \alpha, \qquad \|F'(z)^{-1}F(z)\|_\infty \le \beta. \tag{9}$$

Suppose further that $z \in \mathbf{x}$ and that there exists a constant $\gamma > 0$ such that for all $x \in \mathbf{x}$,

$$\max_i \sum_{j,k} \left| \frac{\partial^2 F_i(x)}{\partial x_j \partial x_k} \right| \le \gamma. \tag{10}$$

If $2\alpha\beta\gamma < 1$, then $\Delta := \sqrt{1 - 2\alpha\beta\gamma}$ is real and we have the following:

1. *There is no zero $x \in \mathbf{x}$ with*

$$\underline{r} < \|x - z\|_\infty < \overline{r},$$

where

$$\underline{r} = \frac{2\beta}{1+\Delta}, \qquad \overline{r} = \frac{1+\Delta}{\alpha\gamma}.$$

2. *At most one zero x is contained in $\mathbf{x}$ with*

$$\|x - z\|_\infty < \frac{2}{\alpha\gamma}.$$

3. *If*

$$\max_{x \in \mathbf{x}} \|x - z\|_\infty < \overline{r},$$

then there is a unique zero $x \in \mathbf{x}$, and this zero satisfies

$$\|x - z\|_\infty \le \underline{r}.$$

The affine invariant version of the Kantorovich theorem given in Deuflhard and Heindl [2] essentially amounts to applying the theorem to $F'(z)^{-1}F(x)$ in place of $F(x)$. In practice, rounding errors in computing $F'(z)^{-1}$ are made, which requires the use of a preconditioning matrix $C \approx F'(z)^{-1}$ and $CF(x)$ in place of $F(x)$ to get the benefits of affine invariance in floating point computations.

Kahan [5] used the Krawczyk operator, which needs only first order slopes, to make existence statements. Together with later improvements using slopes, his result is contained in the following statement.

THEOREM 2.2 (Kahan). *Let $z \in \mathbf{z} \subseteq \mathbf{x}$. If there is a matrix $C \in \mathbb{R}^{n\times n}$ such that the Krawczyk operator*

$$K(\mathbf{z}, \mathbf{x}) := z - CF(z) - (CF[\mathbf{z}, \mathbf{x}] - I)(\mathbf{x} - z) \tag{11}$$

satisfies $K(\mathbf{z}, \mathbf{x}) \subseteq \mathbf{x}$, then $\mathbf{x}$ contains a zero of (1). *Moreover, if $K(\mathbf{x}, \mathbf{x}) \subseteq \operatorname{int}(\mathbf{x})$, then $\mathbf{x}$ contains a unique zero of* (1).

Shen and Neumaier [22] proved that the Krawczyk operator with slopes always provides existence regions which are at least as large as those computed by Kantorovich's theorem, and, since the Krawczyk operator is affine invariant, this also covers the affine invariant Kantorovich theorem.

Recent work by Hansen [3] shows that there is scope for gain in Krawczyk's method by improved preconditioning; but he gives only heuristic recipes for how to proceed. For quadratic problems, where the slope is linear in x, his recipe suggests evaluating $CF[z, x]$ term by term before substituting intervals. Indeed, by subdistributivity, we always have

$$CA_0 + \sum CA_k(\mathbf{x}_k - \mathbf{z}_k) \subseteq C\Big(A_0 + \sum A_k(\mathbf{x}_k - \mathbf{z}_k)\Big)$$

so that, for quadratic functions, Hansen's recipe is never worse than the traditional recipe. We adapt it as follows to general functions, using second order slopes; in the general case, the preconditioned slope takes the form

$$CF[z, x] = CF[z, z'] + \sum (x_k - z'_k) CF_k[z, z', x] \tag{12}$$

or, with $z = z'$, as we use it most of the time,

$$CF[z,x] = CF'(z) + \sum (x_k - z_k) CF_k[z,z,x]. \tag{13}$$

In the following, the consequences of this formulation, combined with ideas from Shen and Neumaier [22], are investigated in detail.

Recent work on Taylor models by Berz and Hoefkens [1] (see also Neumaier [18]) uses expansions to even higher than second order, although at a significantly higher cost. This may be of interest for systems suffering a lot from cancellation, where using low order methods may incur much overestimation, leading to tiny inclusion regions. Another recent paper on exclusion boxes is Kalovics [6].

3. The cluster effect. As explained by Kearfott and Du [11], many branch and bound methods used for global optimization suffer from the so-called *cluster effect*. As is apparent from the discussion below, this effect is also present for branch and bound methods using constraint propagation methods to find and verify *all* solutions of nonlinear systems of equations. (See, e.g., Van Hentenryck, Michel, and Deville [23] for constraint propagation methods.)

The cluster effect consists of excessive splitting of boxes close to a solution and failure to remove many boxes not containing the solution. As a consequence, these methods slow down considerably once they reach regions close to the solutions. The mathematical reason for the cluster effect and how to avoid it will be investigated in this section.

Let us assume that for arbitrary boxes $\mathbf{x}$ of maximal width ε the computed expression $F(\mathbf{x})$ overestimates the range of F over $\mathbf{x}$ by $O(\varepsilon^k)$:

$$F(\mathbf{x}) \in (1 + C\varepsilon^k) \square \{F(x) \mid x \in \mathbf{x}\} \tag{14}$$

for $k \leq 2$ and ε sufficiently small. The exponent k depends on the method used for the computation of $F(\mathbf{x})$.

Let x^* be a regular solution of (1) (so that $F'(x^*)$ is nonsingular), and assume (14). Then any box of diameter ε that contains a point x with

$$\|F'(x^*)(x - x^*)\|_\infty \leq \Delta = C\varepsilon^k \tag{15}$$

might contain a solution. Therefore, independent of the pruning scheme used in a branch and bound method, no box of diameter ε can be eliminated. The inequality (15) describes a parallelepiped of volume

$$V = \frac{\Delta^n}{\det F'(x^*)}.$$

Thus, any covering of this region by boxes of diameter ε contains at least V/ε^n boxes.

The number of boxes of diameter ε which cannot be eliminated is therefore proportional to at least

$$\frac{C^n}{\det F'(x^*)} \quad \text{if } k = 1,$$

$$\frac{(C\varepsilon)^n}{\det F'(x^*)} \quad \text{if } k = 2.$$

For $k = 1$ this number grows exponentially with the dimension, with a growth rate determined by the relative overestimation C and a proportionality factor related to the condition of the Jacobian.

In contrast, for $k = 2$ the number is guaranteed to be small for sufficiently small ε. The size of ε, the diameter of the boxes most efficient for covering the solution, is essentially determined by the nth root of the determinant, which, for a well-scaled problem, reflects the condition of the zero. However, for ill-conditioned zeros (with a tiny determinant in naturally scaled coordinates), one already needs quite narrow boxes before the cluster effect subsides.

So, to avoid the cluster effect, we need at least the quadratic approximation property $k = 2$. Hence, Jacobian information is essential, as well as techniques to discover the shape of the uncertainty region.

A comparison of the typical techniques used for box elimination shows that constraint propagation techniques lead to overestimation of order $k = 1$; hence they suffer from the cluster effect. Centered forms using first order information (Jacobians) as in Krawczyk's method provide estimates with $k = 2$ and are therefore sufficient to avoid the cluster effect, except near ill-conditioned or singular zeros. Second order information as used, e.g., in the theorem of Kantorovich still provides only $k = 2$ in estimate (15); the cluster effect is avoided under the same conditions.

For singular (and hence for sufficiently ill-conditioned) zeros, the argument does not apply, and no technique is known to remove the cluster effect in this case. A heuristic that limits the work in this case by retaining a single but *larger* box around an ill-conditioned approximate zero is described in Algorithm 7 (Step 4(c)) of Kearfott [10].

4. Componentwise exclusion regions close to a zero. Suppose that x^* is a solution of the nonlinear system of equations (1). We want to find an *exclusion region* around x^* with the property that in the interior of this region x^* is the only solution of (1). Such an exclusion region need not be further explored in a branch and bound method for finding all solutions of (1); hence we get the name.

In this section we take an approximate zero z of F, and we choose C to be an approximation of $F'(z)^{-1}$. Suitable candidates for z can easily be found within a branch and bound algorithm by trying Newton steps from the midpoint of each box, iterating while x^ℓ remains in a somewhat enlarged box and either $\|x^{\ell+1} - x^\ell\|$ or $\|F(x^\ell)\|$ decreases by a factor of, say, 1.5 below the best previous value in the iteration. This works locally well even at nearly singular zeros and gives a convenient stop in case no nearby solution exists.

PROPOSITION 4.1. *For every solution $x \in X$ of* (1)*, the deviation*

$$s := |x - z|$$

satisfies

$$0 \le s \le \Big(B_0 + \sum s_k B_k(x)\Big)s + \overline{b}. \tag{16}$$

Proof. By (2) we have $F[z,x](x-z) = F(x) - F(z) = -F(z)$, because x is a zero. Hence, using (5), we compute

$$\begin{aligned} -(x-z) &= -(x-z) + C(F[z,x](x-z) + F(z) + F'(z)(x-z) - F'(z)(x-z)) \\ &= C(F[z,x] - F'(z))(x-z) + (CF'(z) - I)(x-z) + CF(z) \\ &= \Big(CF'(z) - I + \sum (x_k - z_k) CF_k[z,z,x]\Big)(x-z) + CF(z). \end{aligned}$$

Now we take absolute values, use (6), and get

$$\begin{aligned} s = |x - z| &\le \Big(|CF'(z) - I| + \sum |x_k - z_k|\,|CF_k[z, z, x]|\Big)\,|x - z| + |CF(z)| \\ &\le \Big(B_0 + \sum s_k B_k(x)\Big)\, s + \bar{b}. \qquad \square \end{aligned}$$

Using this result we can give a first criterion for existence regions.

THEOREM 4.2. *Let* $0 < u \in \mathbb{R}^n$ *be such that*

$$\Big(B_0 + \sum u_k \overline{B}_k\Big) u + \bar{b} \le u \tag{17}$$

with $B_k(x) \le \overline{B}_k$ *for all* $x \in M_u$, *where*

$$M_u := \{x \mid |x - z| \le u\} \subseteq X. \tag{18}$$

Then (1) *has a solution* $x \in M_u$.

Proof. For arbitrary x in the domain of definition of F we define

$$K(x) := x - CF(x).$$

Now take any $x \in M_u$. We get

$$\begin{aligned} K(x) = x - CF(x) &= z - CF(z) - (CF[z, x] - I)(x - z) \\ &= z - CF(z) - \Big(C\Big(F'(z) + \sum F_k[z, z, x](x_k - z_k)\Big) - I\Big)\,(x - z); \end{aligned}$$

hence

$$K(x) = z - CF(z) - \Big(CF'(z) - I + \sum CF_k[z, z, x](x_k - z_k)\Big)\,(x - z). \tag{19}$$

Taking absolute values we find

$$\begin{aligned} |K(x) - z| &= \Big| -CF(z) - \Big(CF'(z) - I + \sum CF_k[z, z, x](x_k - z_k)\Big)\,(x - z) \Big| \\ &\le |CF(z)| + \Big(|CF'(z) - I| + \sum |CF_k[z, z, x]|\,|x_k - z_k|\Big)\,|x - z| \\ &\le \bar{b} + \Big(B_0 + \sum u_k \overline{B}_k\Big)\, u. \end{aligned} \tag{20}$$

Now assume (17). Then (20) gives

$$|K(x) - z| \le u,$$

which implies by Theorem 2.2 that there exists a solution of (1) which lies in M_u. $\square$

Note that (17) implies $B_0 u \le u$; thus the spectral radius $\rho(B_0) \le 1$. In the applications, we can make both B_0 and $\bar{b}$ very small by choosing z as an approximate zero and C as an approximate inverse of $F'(z)$.

Now the only thing that remains is the construction of a suitable vector u for Theorem 4.2.

THEOREM 4.3. *Let* $S \subseteq X$ *be any set containing* z, *and take*

$$\overline{B}_k \ge B_k(x) \quad \textit{for all } x \in S. \tag{21}$$

For $0 < v \in \mathbb{R}^n$, *set*

$$w := (I - B_0)v, \qquad a := \sum v_k \overline{B}_k v. \tag{22}$$

We suppose that

$$D_j = w_j^2 - 4a_j \overline{b}_j > 0 \tag{23}$$

for all $j = 1, \dots, n$, *and we define*

$$\lambda_j^e := \frac{w_j + \sqrt{D_j}}{2a_j}, \qquad \lambda_j^i := \frac{\overline{b}_j}{a_j \lambda_j^e}, \tag{24}$$

$$\lambda^e := \min_{j=1,\dots,n} \lambda_j^e, \qquad \lambda^i := \max_{j=1,\dots,n} \lambda_j^i. \tag{25}$$

If $\lambda^e > \lambda^i$, *then there is at least one zero* x^* *of* (1) *in the (inclusion) region*

$$R^i := [z - \lambda^i v, z + \lambda^i v] \cap S. \tag{26}$$

The zeros in this region are the only zeros of F *in the* interior *of the (exclusion) region*

$$R^e := [z - \lambda^e v, z + \lambda^e v] \cap S. \tag{27}$$

Proof. Let $0 < v \in \mathbb{R}^n$ be arbitrary, and set $u = \lambda v$. We check for which λ the vector u satisfies property (17) of Theorem 4.2. The requirement

$$\begin{aligned}\lambda v &\geq \Big(B_0 + \sum u_k \overline{B}_k\Big) u + \overline{b} = \Big(B_0 + \sum \lambda v_k \overline{B}_k\Big) \lambda v + \overline{b} \\ &= \overline{b} + \lambda B_0 v + \lambda^2 \sum v_k \overline{B}_k v = \overline{b} + \lambda(v - w) + \lambda^2 a\end{aligned}$$

leads to the sufficient condition $\lambda^2 a - \lambda w + \overline{b} \leq 0$. The jth component of this inequality requires that λ lies between the solutions of the quadratic equation $\lambda^2 a_j - \lambda w_j + \overline{b}_j = 0$, which are λ_j^i and λ_j^e. Hence, for every $\lambda \in [\lambda^i, \lambda^e]$ (this interval is nonempty by assumption), the vector u satisfies (17).

Now assume that x is a solution of (1) in $\operatorname{int}(R^e) \setminus R^i$. Let λ be minimal with $|x - z| \leq \lambda v$. By construction, $\lambda^i < \lambda < \lambda^e$. By the properties of the Krawczyk operator, we know that $x = K(z, x)$; hence

$$\begin{aligned}|x - z| &\leq |CF(z)| + \Big(|CF'(z) - I| + \sum |CF_k[z, z, x]|\, |x_k - z_k|\Big)|x - z| \\ &\leq \overline{b} + \lambda B_0 v + \lambda^2 \sum v_k \overline{B}_k v < \lambda v,\end{aligned} \tag{28}$$

since $\lambda > \lambda^i$. But this contradicts the minimality of λ. So there are indeed no solutions of (1) in $\operatorname{int}(R^e) \setminus R^i$. ∎

This is a componentwise analogue of the Kantorovich theorem. We show in Example 8.1 that it is best possible in some cases.

We observe that the inclusion region from Theorem 4.3 can usually be further improved by noting that $x^* = K(z, x^*)$ and (19) imply

$$\begin{aligned}x^* &\in K(z, \mathbf{x}^i) \\ &= z - CF(z) - \Big(CF'(z) - I + \sum CF_k[z, z, \mathbf{x}^i](\mathbf{x}_k^i - z_k)\Big)(\mathbf{x}^i - z) \\ &\subset \operatorname{int}(\mathbf{x}^i).\end{aligned}$$

An important special case is when $F(x)$ is quadratic in x. For such a function $F[z,x]$ is linear in x, and therefore all $F_k[z,z,x]$ are constant in x. This, in turn, means that $B_k(x) = B_k$ is constant as well. So we can set $\overline{B}_k = B_k$, and the estimate (21) becomes valid everywhere.

COROLLARY 4.4. *Let F be a quadratic function. For arbitrary $0 < v \in \mathbb{R}^n$ define*

$$w := (I - B_0)v, \qquad a := \sum v_k B_k v. \tag{29}$$

We suppose that

$$D_j = w_j^2 - 4a_j\overline{b}_j > 0 \tag{30}$$

for all $j = 1, \ldots, n$, and we set

$$\lambda_j^e := \frac{w_j + \sqrt{D_j}}{2a_j}, \qquad \lambda_j^i := \frac{\overline{b}_j}{a_j \lambda_j^e}, \tag{31}$$

$$\lambda^e := \min_{j=1,\ldots,n} \lambda_j^e, \qquad \lambda^i := \max_{j=1,\ldots,n} \lambda_j^i. \tag{32}$$

If $\lambda^e > \lambda^i$, then there is at least one zero x^* of (1) in the (inclusion) box

$$\mathbf{x}^i := [z - \lambda^i v, z + \lambda^i v]. \tag{33}$$

The zeros in this region are the only zeros of F in the interior *of the (exclusion) box*

$$\mathbf{x}^e := [z - \lambda^e v, z + \lambda^e v]. \tag{34}$$

The examples later will show that the choice of v greatly influences the quality of the inclusion and exclusion regions. The main difficulty for choosing v is the positivity requirement for every D_j. In principle, a vector v, if it exists, could be found by local optimization. A method worth trying could be to choose v as a local optimizer of the problem

$$\begin{aligned} &\max\, n \log \lambda^e + \sum_{j=1}^n \log v_j \\ &\text{s.t.}\;\; D_j \ge \eta \quad (j = 1, \ldots, n), \end{aligned}$$

where η is the smallest positive machine number. This maximizes locally the volume of the excluded box. However, since λ^e is nonsmooth, solving this needs a nonsmooth optimizer (such as SolvOpt [15]).

The $\overline{B}_k$ can be constructed using interval arithmetic for a given reference box $\mathbf{x}$ around z. Alternatively, they could be calculated once in a bigger reference box $\mathbf{x}_{\text{ref}}$ and later reused on all subboxes of $\mathbf{x}_{\text{ref}}$. Saving the $\overline{B}_k$ (which needs the storage of n^3 numbers per zero) provides a simple exclusion test for other boxes. This takes $O(n^3)$ operations, while recomputing the $\overline{B}_k$ costs $O(n^4)$ operations.

5. Exclusion polytopes. Instead of boxes, we can use more general polytopes to describe exclusion and inclusion regions. With the notation as in the introduction, we assume the upper bounds

$$\overline{B}_k \ge |B_k(x)| \qquad \text{for all } x \in X. \tag{35}$$

THEOREM 5.1. *For $0 \le v \le w \in \mathbb{R}^n$, define*

$$P(w) = (\overline{B}_1{}^T w, \ldots, \overline{B}_n{}^T w) \in \mathbb{R}^{n\times n}, \tag{36}$$

$$\Pi^i = \{x \in \mathbb{R}^n \mid (w-v)^T |x-z| \le \overline{b}^{\,T} w\}. \tag{37}$$

Then any zero $x \in X$ of (1) *contained in the polytope*

$$\Pi^e = \{x \in \mathbb{R}^n \mid P(w)|x-z| + B_0^T w \le v\} \tag{38}$$

lies already in Π^i.

Proof. Suppose $x \in \Pi^e$ satisfies $F(x) = 0$. By Proposition 4.1, $s = |x-z|$ satisfies

$$\begin{aligned} s^T w &\le s^T \Big(B_0^T w + \sum s_k \overline{B}_k{}^T w\Big) + \overline{b}^{\,T} w \\ &= s^T (B_0^T w + P(w)s) + \overline{b}^{\,T} w \\ &\le s^T v + \overline{b}^{\,T} w. \end{aligned}$$

Hence $s^T(w-v) \le \overline{b}^{\,T} w$, giving

$$(w-v)^T |x-z| \le \overline{b}^{\,T} w; \tag{39}$$

hence $x \in \Pi^i$. ◻

COROLLARY 5.2. *Let $\mathbf{x} \subseteq X$ be a box and $z \in \mathbf{x}$ be an approximate zero. If there is a vector $0 \le w \in \mathbb{R}^n$ with*

$$v := P(w)u + B_0^T w \le w, \tag{40}$$

where $u := |\mathbf{x} - z|$, then all solutions $x \in \mathbf{x}$ of (1) *satisfy* (39)*, and, in particular,*

$$|x-z|_i \le \overline{b}^{\,T} w \, (w_i - v_i)^{-1} \quad \textit{for all } i \textit{ with } w_i > v_i. \tag{41}$$

Proof. Let $x \in \mathbf{x}$ be a solution of (1). Then $x \in \Pi^e$ by (40), and, due to Theorem 5.1, $\mathbf{x} \in \Pi^i$. Therefore (39) holds. In particular, $(w-v)_i |x-z|_i \le \overline{b}^{\,T} w$. This implies the result. ◻

In contrast to (32), the test (40) needs only $O(n^2)$ operations (once $P(w)$ is computed) and the storage of $n^2 + n$ numbers per zero. Since $P(w)$ can be calculated columnwise, it is not even necessary to keep all $\overline{B}_k$ in store.

Since B_0 and $\overline{b}$ usually are very tiny (they contain only roundoff errors), this is a powerful box reduction technique if we can find a suitable vector w.

The result is most useful, of course, if $w > v$, but in some cases this is not possible. In these cases boxes are at least reduced in some components.

A suitable choice for w may be an approximation $w > 0$ to a Perron eigenvector [16, section 3.2] of the nonnegative matrix

$$M = \sum_k u_k \overline{B}_k{}^T,$$

where $u > 0$ is proportional to the width of the box of interest. Then

$$\lambda w = Mw = \sum u_k \overline{B}_k{}^T w = P(w)u.$$

If

$$\frac{\max(B_0^T w)_i}{w_i} < \alpha < 1, \qquad \mu := (1-\alpha)\lambda^{-1},$$

we can conclude from Corollary 5.2 (with μu in place of u) that the box $[z-\mu u, z+\mu u]$ can be reduced to $[z-\hat{u}, z+\hat{u}]$, where (with $c/0 = \infty$)

$$\hat{u}_i := \min\left(\mu u_i, \frac{\overline{b}^T w}{\max(0, \alpha w_i - (B_0^T w))_i}\right).$$

6. Uniqueness regions. Regions in which there is a unique zero can be found most efficiently as follows. First, one verifies as in the previous sections an exclusion box $\mathbf{x}^e$ which contains no zero except in a much smaller inclusion box $\mathbf{x}^i$. The inclusion box can usually be refined further by some iterations with Krawczyk's method, which generally converges quickly if the initial inclusion box is already verified. Thus we may assume that $\mathbf{x}^i$ is really tiny, with width determined by rounding errors only.

Clearly, $\operatorname{int}(\mathbf{x}^e)$ contains a unique zero iff $\mathbf{x}^i$ contains at most one zero. Thus it suffices to have a condition under which a tiny box contains at most one zero. This can be done even in fairly ill-conditioned cases by the following test.

THEOREM 6.1. *Take an approximate solution $z \in X$ of* (1)*, and let B be a matrix such that*

$$|CF[z,\mathbf{x}] - I| + \sum |\mathbf{x}_k - z_k|\,|CF_k[\mathbf{x}, z, \mathbf{x}]| \le B. \tag{42}$$

If $\|B\| < 1$ for some monotone norm, then $\mathbf{x}$ *contains at most one solution x of* (1).

Proof. Assume that x and x' are two solutions. Then we have

(43)
$$0 = F(x') - F(x) = F[x,x'](x'-x) = \Big(F[x,z] + \sum (x'_k - z_k)F_k[x,z,x']\Big)(x'-x).$$

Using an approximate inverse C of $F'(z)$ we further get

$$x - x' = \Big((CF[z,x] - I) + \sum (x'_k - z_k)CF_k[x,z,x']\Big)(x'-x). \tag{44}$$

Applying absolute values, and using (42), we find

$$|x'-x| \le \Big(|CF[z,x] - I| + \sum |CF_k[x,z,x']|\,|x'_k - z_k|\Big)|x'-x| \le B|x'-x|. \tag{45}$$

This, in turn, implies $\|x'-x\| \le \|B\|\,\|x'-x\|$. If $\|B\| < 1$ we immediately conclude $\|x'-x\| \le 0$; hence $x = x'$. □

Since B is nonnegative, $\|B\| < 1$ holds for some norm iff the spectral radius of B is less than one (see, e.g., Neumaier [16, Corollary 3.2.3]); a necessary condition for this is that $\max B_{kk} < 1$, and a sufficient condition is that $|B|u < u$ for some vector $u > 0$.

So one first checks whether $\max B_{kk} < 1$. If this holds, one checks whether $\|B\|_\infty < 1$; if this fails, one computes an approximate solution u of $(I-B)u = e$, where e is the all-one vector, and checks whether $u > 0$ and $|B|u < u$. If this fails, the spectral radius of B is very close to 1 or larger. (Essentially, this amounts to testing $I - B$ for being an H-matrix; cf. [16, Proposition 3.2.3].)

We can find a matrix B satisfying (42) by computing $\hat{B}_k \geq |CF_k[\mathbf{x}, z, \mathbf{x}]|$, for example by interval evaluation, using (5), and observing

$$\begin{aligned}|CF[z,\mathbf{x}] - I| &\leq |CF'(z) - I| + \sum |\mathbf{x}_k - z_k|\, |CF_k[z,z,\mathbf{x}]| \\ &\leq |CF'(z) - I| + \sum |\mathbf{x}_k - z_k|\, |CF_k[\mathbf{x},z,\mathbf{x}]|.\end{aligned}$$

Then, using (6), we get

$$|CF[z,\mathbf{x}] - I| + \sum |\mathbf{x}_k - z_k|\, |CF_k[\mathbf{x},z,\mathbf{x}]| \leq B_0 + 2\sum |\mathbf{x}_k - z_k|\, \hat{B}_k =: B, \tag{46}$$

where B can be computed using rounding towards $+\infty$.

If F is quadratic, the results simplify again. In this case all $F_k[x', z, x] =: F_k$ are constant, and we can replace $\hat{B}_k$ by $B_k := |CF_k|$. Hence (46) becomes

$$B = B_0 + 2\sum |\mathbf{x}_k - z_k| B_k.$$

7. Componentwise exclusion regions around arbitrary points. In a branch–and–bound-based method for finding all solutions to (1), we not only need to exclude regions close to zeros but also boxes far away from all solutions. This is usually done by interval analysis on the range of F, by constraint propagation methods (see, e.g., Van Hentenryck, Michel, and Deville [23]), or by Krawczyk's method or preconditioned Gauss–Seidel iteration (see, e.g., [16]). An affine invariant, componentwise version of the latter is presented in this section.

Let z be an arbitrary point in the region of definition of F. Throughout this section, $C \in \mathbb{R}^{m\times n}$ denotes an arbitrary rectangular matrix. M_u is as in (18).

THEOREM 7.1. *Let $0 < u \in \mathbb{R}^n$, and take $\overline{B}_k \geq B_k(x)$ for all $x \in M_u$. If there is an index $i \in \{1, \dots, n\}$ such that the inequality*

$$\underline{b}_i - (B_0' u)_i - \sum u_k (\overline{B}_k u)_i > 0 \tag{47}$$

is valid, then (1) *has no solution $x \in M_u$.*

Proof. We set $\mathbf{x} = [z - u, z + u]$. For a zero $x \in M_u$ of F, we calculate, using (5), similar to the proof of Theorem 4.2,

(48)
$$\begin{aligned}0 = |K(x) - x| &= \Big| - CF(z) - \Big(CF'(z) - \sum CF_k[z,z,x](x_k - z_k)\Big)(x - z)\Big| \\ &\geq |CF(z)| - \Big|(CF'(z) - I)(x - z) + \sum (x_k - z_k) CF_k[z,z,x](x - z)\Big|.\end{aligned}$$

Now we use (6) and (47) to compute

$$\begin{aligned}|CF(z)|_i &\geq \underline{b}_i > (B_0' u)_i + \sum (u_k \overline{B}_k u)_i \\ &\geq \Big(|CF'(z)| u\Big)_i + \sum \Big(u_k |CF_k[z,z,x]|\, u\Big)_i \\ &\geq \Big|CF'(z)(x - z)\Big|_i + \sum \Big|(x_k - z_k) CF_k[z,z,x](x - z)\Big|_i \\ &\geq \Big|(CF'(z) - I)(x - z) + \sum (x_k - z_k) CF_k[z,z,x](x - z)\Big|_i .\end{aligned}$$

This calculation and (47) imply

$$\begin{aligned}|CF(z)|_i &- \Big|CF'(z)(x - z) + \sum (x_k - z_k) CF_k[z,z,x](x - z)\Big|_i \\ &\geq \underline{b}_i - (B_0' u)_i - \sum (u_k \overline{B}_k u)_i > 0,\end{aligned}$$

contradicting (48). □

Again, we need a method to find good vectors u satisfying (47). The following theorem provides that.

THEOREM 7.2. *Let $S \subseteq X$ be a set containing z, and take $\overline{B}_k \geq B_k(x)$ for all $x \in S$. If for any $0 < v \in \mathbb{R}^n$ we define*

$$\begin{aligned} w^\times &:= B_0' v, \\ a^\times &:= \sum v_k \overline{B}_k v, \\ D_i^\times &:= {w_i^\times}^2 + 4\underline{b}_i a_i^\times, \\ \lambda_i^\times &:= \frac{\underline{b}_i}{w_i^\times + \sqrt{D_i^\times}}, \\ \lambda^\times &:= \max_{i=1,\dots,n} \lambda_i^\times, \end{aligned} \tag{49}$$

then F has no zero in the interior of the exclusion region

$$R^\times := [z - \lambda^\times v, z + \lambda^\times v] \cap S. \tag{50}$$

Proof. We set $u = \lambda v$ and check the result (47) of Theorem 7.1:

$$0 < \underline{b}_i - (B_0' u)_i - \sum (u_k \overline{B}_k u)_i = \underline{b}_i - \lambda (B_0' v)_i - \lambda^2 \sum (v_k \overline{B}_k v)_i.$$

This quadratic inequality has to be satisfied for some $i \in \{1, \dots, n\}$. The ith inequality is true for all $\lambda \in [0, \lambda_i^\times[$, so we can take the maximum of all these numbers and still have the inequality satisfied for at least one i. Bearing in mind that the estimates are only true in the set S, the result follows from Theorem 7.1. □

As in the last section, a vector v could be calculated by local optimization, e.g., as a local optimizer of the problem

$$\max n \log \lambda^\times + \sum_{j=1}^n \log v_j.$$

This maximizes locally the volume of the excluded box. Solving this also needs a nonsmooth optimizer since $\lambda^\times$ is nonsmooth like λ^e. However, in contrast to the v needed in Theorem 4.3, there is no positivity requirement which has to be satisfied. In principle, every choice of v leads to some exclusion region.

Finding a good choice for C is a subtle problem and could be attacked by methods similar to Kearfott, Hu, and Novoa [12]. Example 8.3 below shows that a pseudoinverse of $F'(z)$ usually yields reasonable results. However, improving the choice of C sometimes widens the exclusion box by a considerable amount.

Again, for quadratic F the result can be made global, due to the fact that the $F_k[z, z, x]$ are independent of x.

COROLLARY 7.3. *Let F be quadratic, and $0 < v \in \mathbb{R}^n$. Choose $\overline{B}_k \geq |CF_k|$, $w_i^\times$, $a_i^\times$, $D_i^\times$, $\lambda_i^\times$, and $\lambda^\times$ as in Theorem 7.2. Then F has no zero in the interior of the exclusion box*

$$\mathbf{x}^\times := [z - \lambda^\times v, z + \lambda^\times]. \tag{51}$$

Proof. This is a direct consequence of Theorem 7.2 and the fact that all $F_k[z, z, x]$ are constant in x. □

Results analogous to Theorems 4.3, 5.1, 6.1, and 7.2 can be obtained for exclusion regions in global optimization problems by applying the above techniques to the first order optimality conditions. Since nothing new happens mathematically, we refrain from giving details.

8. Examples. We illustrate the theory with a few examples.

Example 8.1. We continue Example 1.1, doing all calculations symbolically, hence free of rounding errors, assuming a known zero. (This idealizes the practically relevant case where a good approximation of a zero is available from a standard zero-finder.)

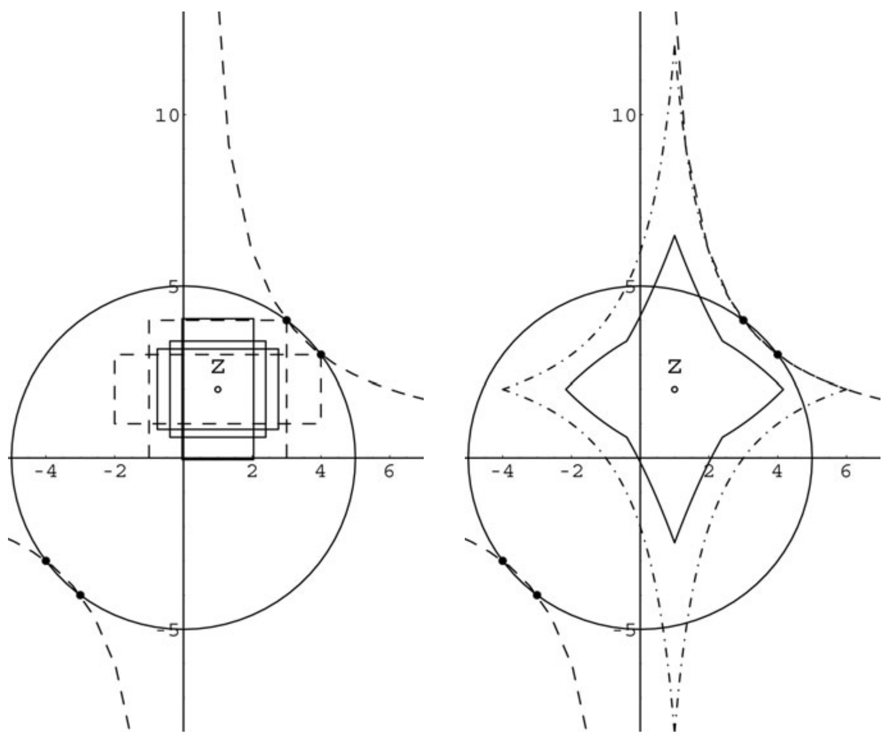

FIG. 1. *Maximal exclusion boxes around* $\binom{1}{2}$ *and the total excluded region for Example* 8.1.

We consider the system of equations (7), which has the four solutions $\pm\binom{3}{4}$ and $\pm\binom{4}{3}$; cf. Figure 1. The system has the form (1) with F given by (8). If we take the solution $x^* = \binom{3}{4}$ as center z, we can use the slope calculations from the introduction. From (29) we get

$$w_j = v_j, \qquad D_j = v_j^2 \quad (j = 1, 2),$$

$$a_1 = \tfrac{1}{14}(3v_1^2 + 8v_1v_2 + 3v_2^2), \qquad a_2 = \tfrac{1}{14}(4v_1^2 + 6v_1v_2 + 4v_2^2),$$

and, for the particular choice $v = \binom{1}{1}$, we get from (31)

$$\lambda^i = 0, \qquad \lambda^e = 1. \tag{52}$$

Thus, Corollary 4.4 implies that the interior of the box

$$[x^* - v, x^* + v] = \begin{pmatrix} [2,4] \\ [3,5] \end{pmatrix}$$

contains no solution apart from $\binom{3}{4}$. This is best possible, since there is another solution $\binom{4}{3}$ at a vertex of this box. The choice $v = \binom{1}{2}$, $\omega(v) = \frac{8}{7}$, gives another exclusion box, neither contained in nor containing the other box.

If we consider the point $z = \binom{1}{2}$, we find

$$F(z) = \begin{pmatrix} -20 \\ -10 \end{pmatrix}, \quad F'(z) = \begin{pmatrix} 2 & 4 \\ 2 & 1 \end{pmatrix}, \quad C = \frac{1}{6}\begin{pmatrix} -1 & 4 \\ 2 & -2 \end{pmatrix},$$

$$\underline{b} = \frac{10}{3}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad B_0 = 0, \quad B_1 = \frac{1}{6}\begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}, \quad B_2 = \frac{1}{6}\begin{pmatrix} 4 & 1 \\ 2 & 2 \end{pmatrix},$$

$$w^\times = v, \qquad a^\times = \frac{1}{6}\begin{pmatrix} v_1^2 + 4v_1v_2 + v_2^2 \\ 2v_1^2 + 2v_1v_2 + 2v_2^2 \end{pmatrix},$$

$$D_1^\times = \frac{1}{9}(29v_1^2 + 80v_1v_2 + 20v_2^2), \qquad D_2^\times = \frac{1}{9}(40v_1^2 + 40v_1v_2 + 49v_2^2).$$

Since everything is affine invariant and $v > 0$, we can set $v = (1, v_2)$, and we compute

$$\lambda^\times = \begin{cases} \frac{20}{3v_2 + \sqrt{40 + 40v_2 + 49v_2^2}} & \text{if } v_2 \le 1, \\ \frac{30}{3 + \sqrt{29 + 80v_2 + 20v_2^2}} & \text{if } v_2 > 1. \end{cases}$$

Depending on the choice of v_2, the volume of the exclusion box varies. There are three locally best choices $v_2 \approx 1.97228$, $v_2 \approx 0.661045$, and $v_2 = 1$, the first providing the globally maximal exclusion box.

For any two different choices of v_2 the resulting boxes are never contained in one another. Selected maximal boxes are depicted in Figure 1 (left) in solid lines; the total region which can be excluded by Corollary 7.3 is shown in solid lines in the right part of the figure.

The optimal preconditioner for exclusion boxes, however, does not need to be an approximate inverse to $F'(z)$. In this case, it turns out that $C = (0\ 1)$ is optimal for every choice of v. Two clearly optimal boxes and the total excluded region for every possible choice of v with $C = (0\ 1)$ can be found in Figure 1 in dashed lines.

Example 8.2. The system of equations (1) with

$$F(x) = \begin{pmatrix} x_1^2 + x_1x_2 + 2x_2^2 - x_1 - x_2 - 2 \\ 2x_1^2 + x_1x_2 + 3x_2^2 - x_1 - x_2 - 4 \end{pmatrix} \tag{53}$$

has the solutions $\binom{1}{1}$, $\binom{1}{-1}$, $\binom{-1}{1}$; cf. Figure 2. It is easily checked that

$$F[z,x] = \begin{pmatrix} x_1 + x_2 + z_1 - 1 & 2x_2 + z_1 + 2z_2 - 1 \\ 2x_1 + x_2 + 2z_1 - 1 & 3x_2 + z_1 + 3z_2 - 1 \end{pmatrix}$$

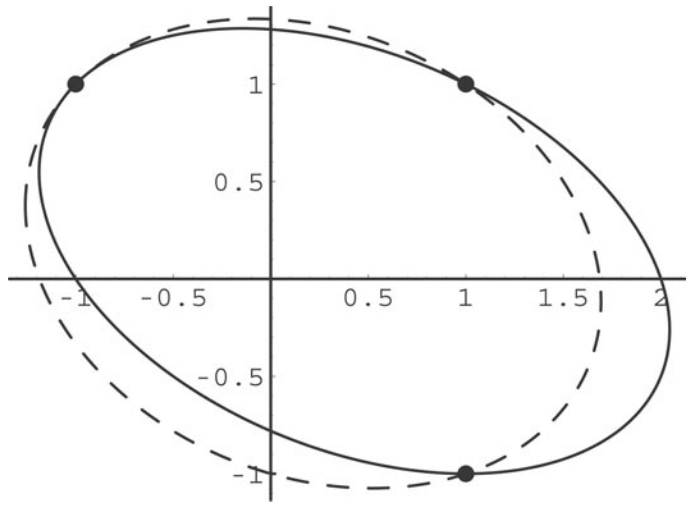

FIG. 2. *Two quadratic equations in two variables; Example* 8.2.

satisfies (2). Thus (5) holds with

$$F'(z) = \begin{pmatrix} 2z_1 + z_2 - 1 & z_1 + 4z_2 - 1 \\ 4z_1 + z_2 - 1 & z_1 + 6z_2 - 1 \end{pmatrix}, \qquad F_1 = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}, \qquad F_2 = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}.$$

We consider boxes centered at the solution $z = x^* = \binom{1}{1}$. For

$$\mathbf{x} = [x^* - \varepsilon u, x^* + \varepsilon u] = \begin{pmatrix} [1-\varepsilon, 1+\varepsilon] \\ [1-\varepsilon, 1+\varepsilon] \end{pmatrix},$$

we find

$$F'[x^*, \mathbf{x}] = \begin{pmatrix} [2-2\varepsilon, 2+2\varepsilon] & [4-2\varepsilon, 4+2\varepsilon] \\ [4-3\varepsilon, 4+3\varepsilon] & [6-3\varepsilon, 6+3\varepsilon] \end{pmatrix},$$

$$F'(\mathbf{x}) = \begin{pmatrix} [2-3\varepsilon, 2+3\varepsilon] & [4-5\varepsilon, 4+5\varepsilon] \\ [4-5\varepsilon, 4+5\varepsilon] & [6-7\varepsilon, 6+7\varepsilon] \end{pmatrix}.$$

The midpoint of $F'(\mathbf{x})$ is here $F'(z)$, and the optimal preconditioner is

$$C := F'(x^*)^{-1} = \begin{pmatrix} -1.5 & 1 \\ 1 & -0.5 \end{pmatrix};$$

from this, we obtain

$$B_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0 \end{pmatrix}, \qquad B_2 = \begin{pmatrix} 0.5 & 0 \\ 0.5 & 0.5 \end{pmatrix}.$$

The standard uniqueness test checks for a given box $\mathbf{x}$ whether the matrix $F'(\mathbf{x})$ is strongly regular (Neumaier [16]). But given the zero x^* (or, in finite precision calculations, a tiny enclosure for it), it suffices to show strong regularity of $F[x^*, \mathbf{x}]$. We find

$$|I - CF'(\mathbf{x})| = \frac{\varepsilon}{2}\begin{pmatrix} 19 & 29 \\ 11 & 17 \end{pmatrix},$$

with spectral radius $\varepsilon(9 + 4\sqrt{5}) \approx 17.944\varepsilon$. Thus $F'(\mathbf{x})$ is strongly regular for $\varepsilon < 1/17.944 = 0.0557$. The exclusion box constructed from slopes is better, since

$$|I - CF[x^*, \mathbf{x}]| = \varepsilon\begin{pmatrix} 6 & 6 \\ 3.5 & 3.5 \end{pmatrix}$$

has spectral radius 9.5ε. Thus $F[x^*, \mathbf{x}]$ is strongly regular for $\varepsilon < 1/9.5$, and we get an exclusion box of radius $1/9.5$.

The Kantorovich theorem, Theorem 2.1, yields the following results:

$$\begin{aligned} &F'' = \left(\begin{pmatrix} 2 & 1 \\ 4 & 1 \end{pmatrix} \quad \begin{pmatrix} 4 & 1 \\ 1 & 6 \end{pmatrix}\right), \\ &\alpha = 2.5, \quad \beta = 0, \quad \gamma = 12, \quad \Delta = 1, \\ &\underline{r} = 0, \qquad \overline{r} = \frac{2}{2.5 \cdot 12} = \frac{1}{15}; \end{aligned}$$

hence it provides an even smaller (i.e., inferior) exclusion box of radius $\frac{1}{15}$.

If we apply Kahan's theorem, Theorem 2.2, with $F'(\mathbf{x})$, we have to check that $K(\mathbf{x}, \mathbf{x}) \subseteq \operatorname{int}(\mathbf{x})$. Now

$$K(\mathbf{x}, \mathbf{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{\varepsilon}{2}\begin{pmatrix} 19 & 29 \\ 11 & 17 \end{pmatrix}\begin{pmatrix} [-\varepsilon, \varepsilon] \\ [-\varepsilon, \varepsilon] \end{pmatrix}$$

is in $\operatorname{int}(\mathbf{x})$ if

$$\begin{pmatrix} [1 - 24\varepsilon^2, 1 + 24\varepsilon^2] \\ [1 - 14\varepsilon^2, 1 + 14\varepsilon^2] \end{pmatrix} \subseteq \begin{pmatrix} [1 - \varepsilon, 1 + \varepsilon] \\ [1 - \varepsilon, 1 + \varepsilon] \end{pmatrix},$$

which holds for $\varepsilon < 1/24$. This result can be improved if we use slopes instead of interval derivatives. Indeed,

$$K(z, \mathbf{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \varepsilon\begin{pmatrix} 6 & 6 \\ 3.5 & 3.5 \end{pmatrix}\begin{pmatrix} [-\varepsilon, \varepsilon] \\ [-\varepsilon, \varepsilon] \end{pmatrix}$$

is in $\operatorname{int}(\mathbf{x})$ if

$$\begin{pmatrix} [1 - 12\varepsilon^2, 1 + 12\varepsilon^2] \\ [1 - 7\varepsilon^2, 1 + 7\varepsilon^2] \end{pmatrix} \subseteq \begin{pmatrix} [1 - \varepsilon, 1 + \varepsilon] \\ [1 - \varepsilon, 1 + \varepsilon] \end{pmatrix},$$

i.e., for $\varepsilon < 1/12$.

Now we consider the new results. From (31) we get

$$\lambda^e = \frac{2}{v_1 + v_2}. \tag{54}$$

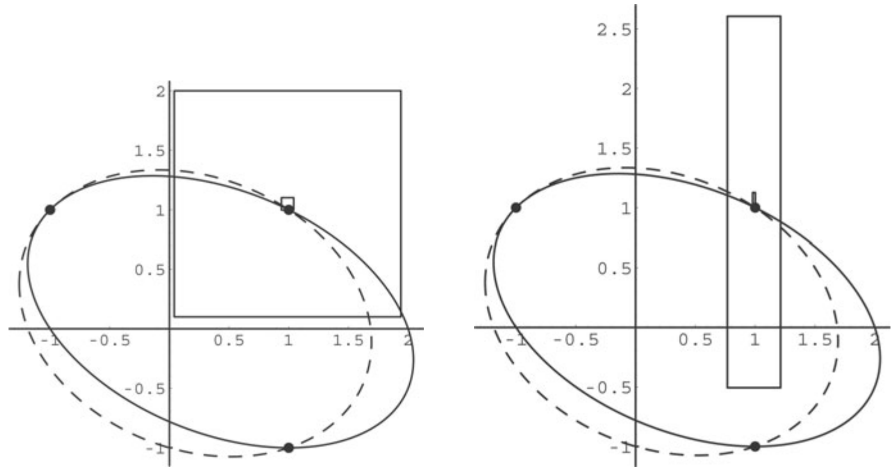

FIG. 3. $\mathbf{x}^e$ *and* $\mathbf{x}^i$ *calculated for Example* 8.2 *with three significant digits for* $v = (1, 1)$ *and* $v = (1, 7)$ *at* $z = (0.99, 1.05)$.

In exact arithmetic, we find $\lambda^e = 1$ so that Corollary 4.4 implies that the interior of the box

$$[x^* - v, x^* + v] = \begin{pmatrix} [0, 2] \\ [0, 2] \end{pmatrix} \tag{55}$$

contains no solution apart from z. In this example, the box is not as large as desirable, since in fact the larger box

$$[x^* - 2v, x^* + 2v] = \begin{pmatrix} [-1, 3] \\ [-1, 3] \end{pmatrix}$$

contains no other solution. However, *the box* (55) *is still one order of magnitude larger than that obtained from the standard uniqueness tests or the Kantorovich theorem.*

If we use inexact arithmetic (we used Mathematica with three significant digits, using this artificially low precision to make the inclusion regions visible in the pictures) and only approximative zeros, the results do not change too much, which can be seen in the pictures of Figure 3.

Corollary 7.3 also gives very promising results. The size of the exclusion boxes again depends on the center z and the vector v. The results for various choices can be found in Figure 4.

To utilize Corollary 5.2 at the exact zero $z = \binom{1}{1}$ we first choose for $u = \binom{1}{1}$ the Perron eigenvector $w_p = \binom{1}{0}$. Its eigenvalue is $\lambda = 1$, and, since $B_0 = 0$ and $\overline{b} = 0$, we conclude that Corollary 5.2 reduces the *first* component of every box $\mathbf{x}$ in the parallelogram P,

$$|x_1 - 1| + |x_2 - 1| < 2, \tag{56}$$

to the thin value $[1, 1]$. That the second component is not reduced is caused by the degeneracy of u. If we choose instead a positive approximation $w = \binom{1}{\varepsilon}$ to w_p and consider any box $\mathbf{x} \subseteq P$, there is $\alpha < 1$ with

$$|x_1 - 1| + |x_2 - 1| < 2\alpha < 2,$$

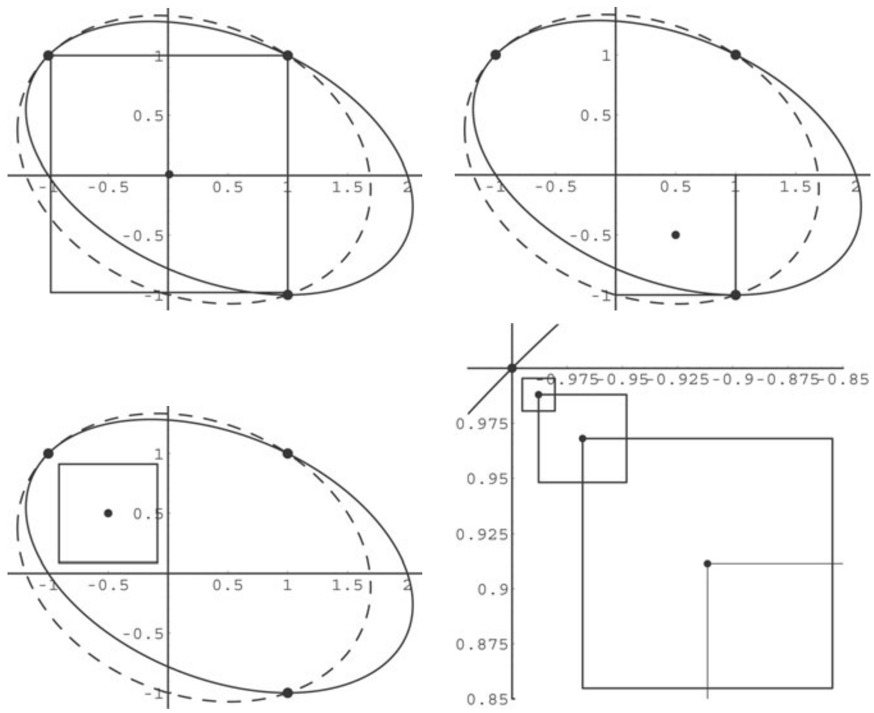

FIG. 4. $\mathbf{x}^\times$ *for Example* 8.2 *and various choices of* z *and* $v = (1, 1)$.

because $\mathbf{x}$ is compact. For $\varepsilon \le 1/\alpha - 1$, we therefore get

$$v = \frac{1}{2}\begin{pmatrix} |x_1 - 1| + (1+\varepsilon)|x_2 - 1| \\ \varepsilon|x_2 - 1| \end{pmatrix} \le \frac{1}{2}\begin{pmatrix} (1+\varepsilon)(|x_1 - 1| + |x_2 - 1|) \\ \varepsilon|x_2 - 1| \end{pmatrix} < w.$$

Then Corollary 5.2 implies that $|x_i - 1| \le 0$ for $i = 1, 2$.

The parallelogram P is best possible in the sense that it contains the other two solutions on its boundary. (But, for general systems, the corresponding maximal exclusion set need not reach another zero and has no simple geometric shape.)

For a nonquadratic polynomial function, all calculations become more complex, and the exclusion sets found are usually far from optimal, though still much better than those from the traditional methods. The $F_k[z, z, x]$ are no longer independent of x, so Theorems 4.3 and 7.2 have to be applied. This involves the computation of a suitable upper bound $\overline{B}_k$ of $F_k[z, z, x]$ by interval arithmetic.

Example 8.3. Figure 5 displays the following system of equations $F(x) = 0$ in two variables, with two polynomial equations of degree 2 and 8:

(57)

$$\begin{aligned} F_1(x) &= x_1^2 + 2x_1x_2 - 2x_2^2 - 2x_1 - 2x_2 + 3, \\ F_2(x) &= x_1^4x_2^4 + x_1^3x_2^4 + x_1^4x_2^3 + 15x_1^2x_2^4 - 8x_1^3x_2^3 + 10x_1^4x_2^2 + 3x_1x_2^4 + 5x_1^2x_2^3 \\ &\quad + 7x_1^3x_2^2 + x_1^4x_2 - 39x_2^4 + 32x_1x_2^3 - 57x_1^2x_2^2 + 21x_1^3x_2 - 17x_1^4 - 27x_2^3 - 17x_1x_2^2 \\ &\quad - 8x_1^2x_2 - 18x_1^3 - 478x_2^2 + 149x_1x_2 - 320x_1^2 - 158x_2 - 158x_1 + 1062. \end{aligned}$$

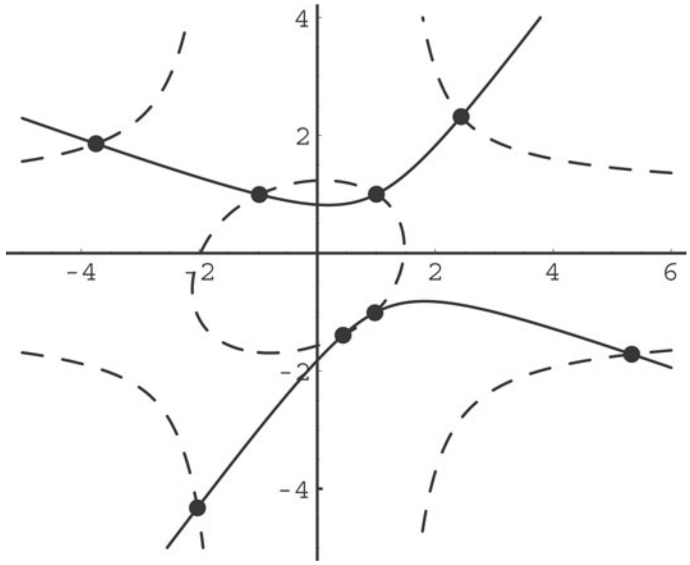

FIG. 5. *Two polynomial equations in two variables; Example* 8.3.

The system (57) has 8 solutions, at approximately

$$\begin{pmatrix}1.0023149901708083\\1.0011595047756938\end{pmatrix},\quad \begin{pmatrix}0.4378266929701329\\-1.3933047617799774\end{pmatrix},\quad \begin{pmatrix}0.9772028387127761\\-1.0115934531170049\end{pmatrix},$$

$$\begin{pmatrix}-0.9818234823156266\\0.9954714636375825\end{pmatrix},\quad \begin{pmatrix}-3.7502535429488344\\1.8585101451403585\end{pmatrix},\quad \begin{pmatrix}2.4390986061035260\\2.3174396617957018\end{pmatrix},$$

$$\begin{pmatrix}5.3305903297000243\\-1.7161362016394848\end{pmatrix},\quad \begin{pmatrix}-2.0307311621763933\\-4.3241016906293375\end{pmatrix}.$$

We consider the approximate solution $z = \binom{0.99}{1.01}$. For the set S we choose the box $[z-u, z+u]$ with $u = \binom{1}{1}$. In this case we have

$$F(z) \approx \begin{pmatrix}-0.0603\\-1.170\end{pmatrix}, \qquad F'(z) \approx \begin{pmatrix}2 & -4.06\\-717.55 & -1147.7\end{pmatrix},$$

$$F_1[z,z,x] = \begin{pmatrix}1 & 0\\f_1 & 0\end{pmatrix}, \qquad F_2[z,z,x] = \begin{pmatrix}2 & -2\\f_2 & f_3\end{pmatrix},$$

where

$$\begin{aligned}f_1 \approx &-405.63 - 51.66x_1 - 17x_1^2 + 36.52x_2 + 23x_1x_2 + x_1^2x_2\\&-13.737x_2^2 + 26.8x_1x_2^2 + 10x_1^2x_2^2 - 7.9x_2^3 - 6.02x_1x_2^3 + x_1^2x_2^3\\&+19.92x_2^4 + 2.98x_1x_2^4 + x_1^2x_2^4,\\f_2 \approx &191.04 - 7.6687x_2 + 62.176x_2^2 + 39.521x_2^3,\\f_3 \approx &-588.05 - 36.404x_2 - 19.398x_2^2.\end{aligned}$$

We further compute

$$C = \begin{pmatrix} 0.22035 & -0.00077947 \\ -0.13776 & -0.00038397 \end{pmatrix},$$

$$B_0 = 10^{-5}\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \overline{B}_1 = \begin{pmatrix} 1.0636 & 0 \\ 0.5027 & 0 \end{pmatrix}, \quad \overline{B}_2 = \begin{pmatrix} 0.3038 & 0.1358 \\ 0.5686 & 0.5596 \end{pmatrix}, \quad \bar{b} = \begin{pmatrix} 0.0124 \\ 0.0088 \end{pmatrix}.$$

If we use Theorem 4.3 for $v = \binom{1}{1}$, we get

$$w = \begin{pmatrix} 0.99999 \\ 0.99998 \end{pmatrix}, \quad a = \begin{pmatrix} 1.5032 \\ 1.6309 \end{pmatrix}, \quad D = \begin{pmatrix} 0.925421 \\ 0.942575 \end{pmatrix},$$

$$\lambda^i = 0.0126403, \qquad \lambda^e = 0.604222,$$

so we may conclude that there is exactly one zero in the box

$$\mathbf{x}^i = \begin{pmatrix} [0.97736, 1.00264] \\ [0.99736, 1.02264] \end{pmatrix},$$

and this zero is the only zero in the interior of the exclusion box

$$\mathbf{x}^e = \begin{pmatrix} [0.385778, 1.59422] \\ [0.405778, 1.61422] \end{pmatrix}.$$

In Figure 6 the two boxes are displayed.

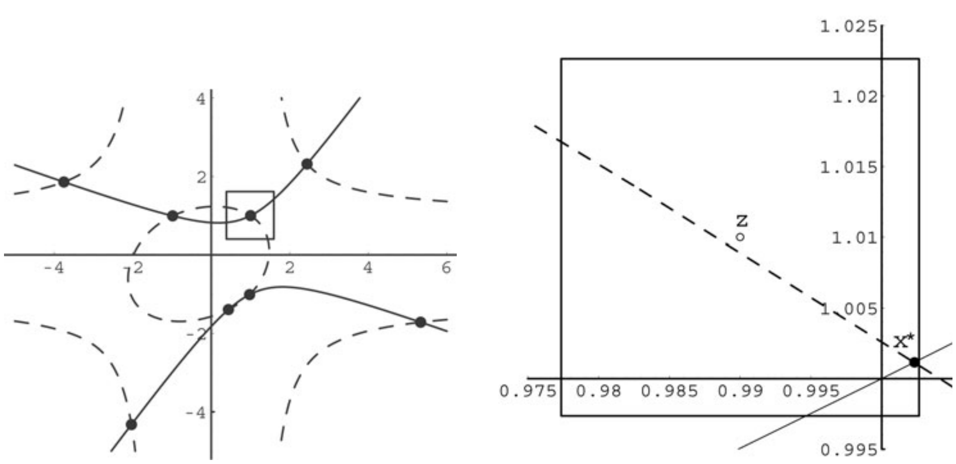

FIG. 6. *Exclusion and inclusion boxes for Example* 8.3 *at* $z = (0.99, 1.01)$.

Next we consider the point $z = \binom{1.5}{-1.5}$ to test Theorem 7.2. We compute

$$F(z) \approx \begin{pmatrix} -3.75 \\ -1477.23 \end{pmatrix}, \qquad F_1[z, z, x] \approx \begin{pmatrix} 1 & 0 \\ g_1 & 0 \end{pmatrix},$$

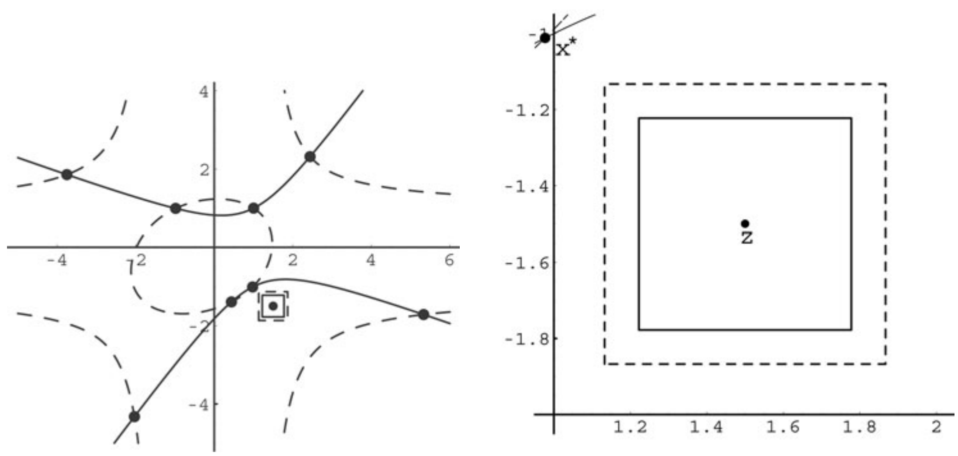

FIG. 7. *Exclusion boxes for Example* 8.3 *at* $z = (1.5, -1.5)$.

$$F'(z) \approx \begin{pmatrix} -2 & 7 \\ -1578.73 & 1761.77 \end{pmatrix}, \qquad F_2[z, z, x] = \begin{pmatrix} 2 & -2 \\ g_2 & g_3 \end{pmatrix},$$

with

$$\begin{aligned} g_1 \approx & -488.75 - 69x_1 - 17x_1^2 + 61.75x_2 + 24x_1x_2 + x_1^2x_2 \\ & + 31.5x_2^2 + 37x_1x_2^2 + 10x_1^2x_2^2 - 12.25x_2^3 - 5x_1x_2^3 + x_1^2x_2^3 \\ & + 24.75x_2^4 + 4x_1x_2^4 + x_1^2x_2^4, \\ g_2 \approx & \quad 73.1563 + 138.063x_2 - 95.875x_2^2 + 68.25x_2^3, \\ g_3 \approx & -536.547 - 12.75x_2 + 7.6875x_2^2. \end{aligned}$$

Performing the necessary computations, we find for $\mathbf{x} = [z - u, z + u]$ with $u = \frac{1}{2}\binom{1}{1}$

$$F'(z)^{-1} \approx \begin{pmatrix} 0.234 & -0.00093 \\ 0.21 & -0.000266 \end{pmatrix}, \qquad \underline{b} = \begin{pmatrix} 0.496 \\ 0.3939 \end{pmatrix},$$

$$\overline{B}_1 = \begin{pmatrix} 1.2895 & 0 \\ 0.5113 & 0 \end{pmatrix}, \quad B_0' = \begin{pmatrix} 1 & 10^{-5} \\ 10^{-5} & 1.00001 \end{pmatrix}, \quad \overline{B}_2 = \begin{pmatrix} 1.5212 & 0.0215 \\ 0.7204 & 0.2919 \end{pmatrix}.$$

Now we use Theorem 7.2 for $v = \binom{1}{1}$ and $C = F'(z)^{-1}$ and get

$$w^\times = \begin{pmatrix} 1.00001 \\ 1.00002 \end{pmatrix}, \quad a^\times = \begin{pmatrix} 2.8322 \\ 1.5236 \end{pmatrix}, \quad D^\times = \begin{pmatrix} 6.6191 \\ 3.4006 \end{pmatrix}, \quad \lambda^\times = 0.277656;$$

so we conclude that there are no zeros of F in the interior of the exclusion box

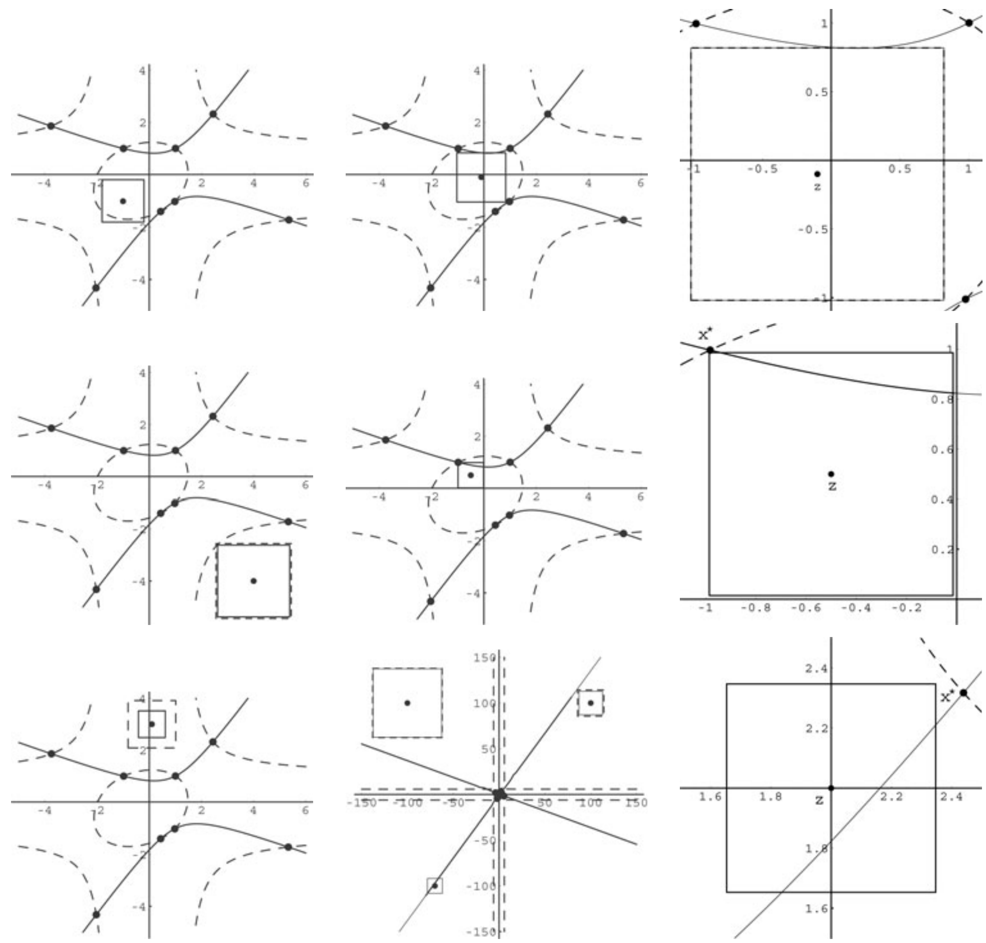

FIG. 8. *Exclusion boxes for Example* 8.3 *in various regions of* $\mathbb{R}^2$.

$$\mathbf{x}^{\times} = \begin{pmatrix} [1.22234, 1.77766] \\ [-1.77766, -1.22234] \end{pmatrix}.$$

However, the choice $C = F'(z)^{-1}$ is not best possible in this situation. If we take

$$C = \begin{pmatrix} 1 & 0.002937 \end{pmatrix},$$

we compute $\lambda^{\times} = 0.367223$ and find the considerably larger exclusion box

$$\mathbf{x}^{\times} = \begin{pmatrix} [1.13278, 1.86722] \\ [-1.86722, -1.13278] \end{pmatrix}.$$

Figure 7 shows both boxes, the bigger one in dashed lines.

Finally, Figure 8 shows various exclusion boxes for nonzeros, and Figure 9 contains exclusion boxes and some inclusion boxes for all of the zeros of F.

While the previous examples were low dimensional, our final example shows that the improvements over traditional results may even be more pronounced for higher dimensional problems with poorly conditioned zeros.

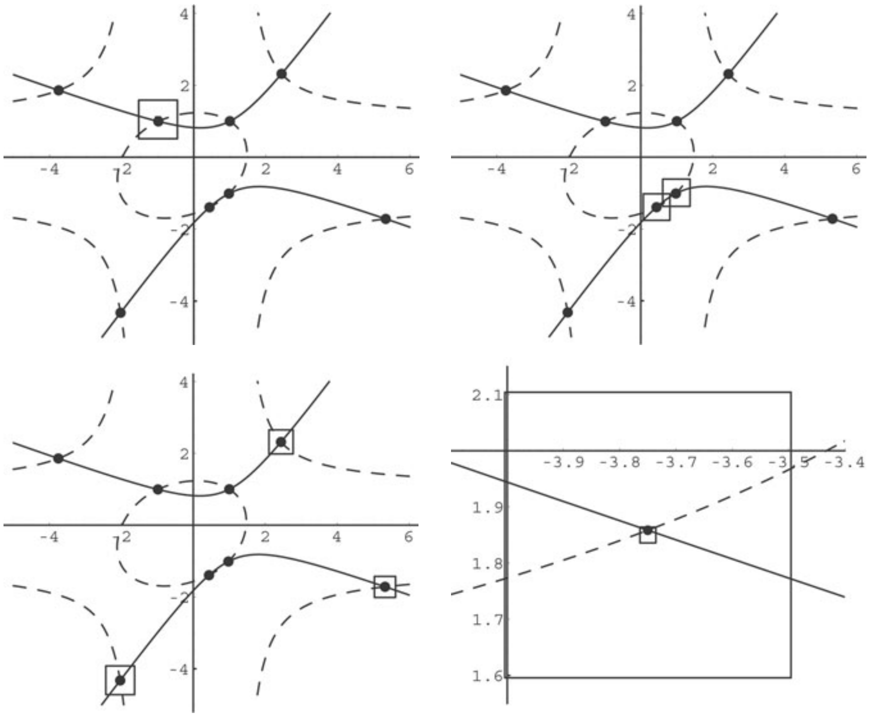

FIG. 9. *Exclusion boxes for all zeros of F in Example* 8.3.

Example 8.4. We consider the set of equations

$$\sum_{k=1}^{n} x_k^i = H(n, -i) \quad \text{for } i = 1, \ldots, n,$$

where the *harmonic numbers* $H(n,m)$ are defined as

$$H(n,m) := \sum_{k=1}^{n} k^{-m}.$$

Clearly, $x_k^* = k$ is a solution, and the complete set of solutions is given by all permutations of this vector.

We compare the results provided by Theorem 4.3 with the exclusion box obtained by strong regularity of the slope $F[z, \mathbf{x}]$ (which in the previous examples was the best among the traditional choices). The vector v needed in Theorem 4.3 was chosen as the all-one vector e. All numerical calculations were performed in double precision arithmetic.

The results are collected in Table 1; R denotes the radius of the exclusion box computed by Theorem 4.3, r the radius of the exclusion box implied by strong regularity of $F[z, \mathbf{x}]$, and κ the condition number of $F'(x^*)$. All numbers are approximate.

From the logarithmic plot in Figure 10, we see that the radii of the exclusion boxes decrease in both cases exponentially with n. However, the quotient of the two radii increases exponentially with n. This shows that our new method suffers much less from the double deterioration due to the increase of both dimension and the Jacobian condition number at the zero.

TABLE 1

n	R	r	R/r	κ
2	1	1	1.000	10.91
3	0.41316	0.127017	3.253	153.155
4	0.197355	0.0206925	9.538	3021.56
5	0.082	0.00359092	22.835	76819.8
6	0.034	0.00063524	53.523	$2.38489 \cdot 10^{6}$
7	0.013	0.00011303	115.007	$8.7331 \cdot 10^{7}$
8	0.005	0.000020137	248.296	$3.68207 \cdot 10^{9}$
9	0.00185847	$3.58494 \cdot 10^{-6}$	518.408	$1.75585 \cdot 10^{11}$
10	0.00068	$6.3732199 \cdot 10^{-7}$	1066.960	$9.34062 \cdot 10^{12}$
11	0.00025	$1.1311565 \cdot 10^{-7}$	2210.130	$5.48274 \cdot 10^{14}$
12	0.000092	$2.00428 \cdot 10^{-8}$	4590.190	$3.52073 \cdot 10^{16}$
13	0.000034	$3.5455649 \cdot 10^{-9}$	9589.450	$2.46174 \cdot 10^{18}$
14	0.0000125	$6.26252 \cdot 10^{-10}$	19960.000	$5.6081 \cdot 10^{19}$
15	$4.5043 \cdot 10^{-6}$	$1.1045 \cdot 10^{-10}$	40781.400	$2.64518 \cdot 10^{20}$
16	$1.6527 \cdot 10^{-6}$	$1.94493 \cdot 10^{-11}$	84975.400	$9.40669 \cdot 10^{21}$

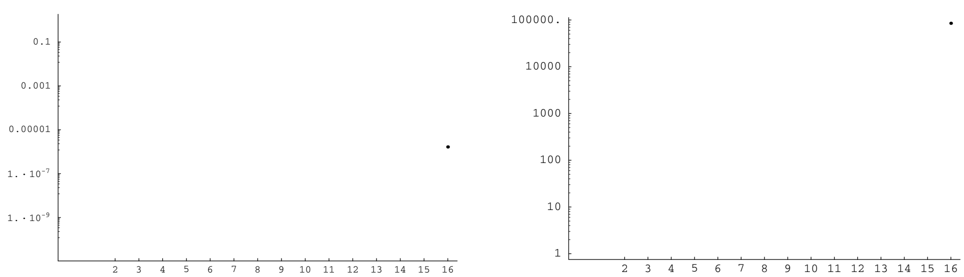

FIG. 10. *Radii of the exclusion boxes and quotient of the radii for Example* 8.4.

REFERENCES

[1] M. BERZ AND J. HOEFKENS, *Verified high-order inversion of functional dependencies and interval Newton methods*, Reliab. Comput., 7 (2001), pp. 379–398.

[2] P. DEUFLHARD AND G. HEINDL, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal., 16 (1979), pp. 1–10.

[3] E. HANSEN, *Preconditioning linearized equations*, Computing, 58 (1997), pp. 187–196.

[4] C. JANSSON, *On self-validating methods for optimization problems*, in Topics in Validated Computations, J. Herzberger, ed., Elsevier, Amsterdam, 1994, pp. 381–438.

[5] W. M. KAHAN, *A More Complete Interval Arithmetic*, Lecture notes for an engineering summer course in numerical analysis, University of Michigan, Ann Arbor, MI, 1968.

[6] F. KALOVICS, *Creating and handling box valued functions used in numerical methods*, J. Comput. Appl. Math., 147 (2002), pp. 333–348.

[7] L. B. KANTOROVICH, *Functional analysis and applied mathematics*, Uspekhi Mat. Nauk, 3 (1948), pp. 89–185 (in Russian). Translated by C. D. Benster, Nat. Bur. Stand. Rep. 1509, Washington, DC, 1952.

[8] R. B. KEARFOTT, *Empirical evaluation of innovations in interval branch and bound algorithms for nonlinear systems*, SIAM J. Sci. Comput., 18 (1997), pp. 574–594.

[9] R. B. KEARFOTT, *A review of techniques in the verified solution of constrained global optimization problems*, in Applications of Interval Computations, R. B. Kearfott and V. Kreinovich, eds., Kluwer, Dordrecht, The Netherlands, 1996, pp. 23–60.

[10] R. B. KEARFOTT, *Rigorous Global Search: Continuous Problems*, Kluwer, Dordrecht, The Netherlands, 1996.

[11] R. B. Kearfott and K. Du, *The cluster problem in multivariate global optimization*, J. Global Optim., 5 (1994), pp. 253–265.

[12] R. B. Kearfott, C. Hu, and M. Novoa, III, *A review of preconditioners for the interval Gauss-Seidel method*, Interval Comput., 1 (1991), pp. 59–85.

[13] L. V. Kolev, *Use of interval slopes for the irrational part of factorable functions*, Reliab. Comput., 3 (1997), pp. 83–93.

[14] R. Krawczyk and A. Neumaier, *Interval slopes for rational functions and associated centered forms*, SIAM J. Numer. Anal., 22 (1985), pp. 604–616.

[15] A. Kuntsevich and F. Kappel, *SolvOpt*, http://www.kfunigraz.ac.at/imawww/kuntsevich/solvopt/(1997).

[16] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.

[17] A. Neumaier, *Introduction to Numerical Analysis*, Cambridge University Press, Cambridge, UK, 2001.

[18] A. Neumaier, *Taylor forms—use and limits*, Reliab. Comput., 9 (2002), pp. 43–79.

[19] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Classics Appl. Math. 30, SIAM, Philadelphia, 2000.

[20] S. M. Rump, *Expansion and estimation of the range of nonlinear functions*, Math. Comp., 65 (1996), pp. 1503–1512.

[21] S. M. Rump, *INTLAB—INTerval LABoratory*, in Developments in Reliable Computing, T. Csendes, ed., Kluwer, Dordrecht, The Netherlands, 1999, pp. 77–104.

[22] Z. Shen and A. Neumaier, *The Krawczyk operator and Kantorovich's theorem*, J. Math. Anal. Appl., 149 (1990), pp. 437–443.

[23] P. Van Hentenryck, L. Michel, and Y. Deville, *Numerica. A Modeling Language for Global Optimization*, MIT Press, Cambridge, MA, 1997.

[24] R. J. Van Iwaarden, *An Improved Unconstrained Global Optimization Algorithm*, Ph.D. thesis, University of Colorado, Denver, CO, http://www-math.cudenver.edu/graduate/thesis/rvan.ps.gz (1996).