Noname manuscript No.
(will be inserted by the editor)

Exclusion regions for optimization problems

Hermann Schichl · Mihály Csaba Markót · Arnold Neumaier

Received: date / Accepted: date

Abstract Branch and bound methods for finding all solutions of a global optimization problem in a box frequently have the difficulty that subboxes containing no solution cannot be easily eliminated if they are close to the global minimum. This has the effect that near each global minimum, and in the process of solving the problem also near the currently best found local minimum, many small boxes are created by repeated splitting, whose processing often dominates 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 local minimum found, that are guaranteed to contain no other local minimum and hence can safely be discarded. In addition, we will introduce a method for verifying the existence of a feasible point close to an approximate local minimum.

These exclusion regions are constructed using uniqueness tests based on the Krawczyk operator and make use of first, second and third order information on the objective and constraint functions.

Keywords global optimization · validated enclosure · existence test · uniqueness test · inclusion region · exclusion region · branch and bound · cluster effect · Krawczyk operator · Kantorovich theorem · backboxing · affine invariant

Mathematics Subject Classification (2000) primary 65H20 · 65G30

This research was supported by the Austrian Science Found (FWF) Grant Nr. P22239-N13.

Institut für Mathematik, Universität Wien. Nordbergstraße 15, A-1090 Wien, Austria
E-mail: {Hermann.Schichl, Mihaly.Markot, Arnold.Neumaier}@univie.ac.at

1 Introduction

Branch and bound methods for solving global optimization problems frequently have the difficulty that subboxes containing no solution cannot be easily eliminated if one of the best local optima found so far lies nearby the box. This has the effect that near the best local minima found, many small boxes are created by repeated splitting, whose processing often dominates the total work spent on the global search.

This paper discusses in Section 2 the reasons for the occurrence of this so-called cluster effect, and how to reduce the cluster effect by defining exclusion regions around each local minimum found that are guaranteed to contain no other local minimum and hence can safely be discarded. Such exclusion regions in the shape of boxes are the basis for the back-boxing strategy that eliminates the cluster effect in most cases. Back-boxing is a method for identifying a box around a local minimum on which the objective function is strictly convex, and which can therefore contain no other local minimum, see Van Iwaarden (1996). Hence, this box, the exclusion box, can be excluded from the search region during further branch and bound search. There are also other methods for constructing such exclusion boxes like Kearfott (1987, Algorithm 2.6, Step 4) (see also Kearfott (1997, 1996a)), or ε–inflation, a procedure which iteratively tries to expand a uniqueness region around a local minimizer, again leading to an exclusion box, see e.g. Mayer (1995). There is also the possibility to construct pairs of exclusion and inclusion boxes; those have the property that every solution inside the exclusion box necessarily is already in the inclusion box, see e.g. Rump (1998). In this case, uniqueness of the solution within the inclusion box is not necessarily proved. There is no mathematical necessity to construct boxes which can be excluded from the branch and bound search. Sometimes, it is more useful to construct exclusion and inclusion regions, that are spherical, ellipsoidal, or have other shapes.

There are several ways to use a branch and bound

Exclusion regions for systems of equations are traditionally constructed using uniqueness tests based on the Krawczyk operator (see, e.g., Neumaier (1990, Chapter 5)), other interval Newton operators (see, e.g., Kearfott (1996a)), or the Kantorovich theorem (see, e.g., Ortega and Rheinboldt (2000, Theorem 12.6.1)). They can also be constructed by a second order method by Schichl and Neumaier (2005a). All of these methods applied to the Karush-John first order necessary optimality conditions of an optimization problem can be in principal used to construct exclusion regions for local optima.

However, the optimality conditions often lead to a degenerate system, causing this method to fail.

In Section 3 we review known methods for constructing exclusion regions for optimization problems. Then in Section 4 we will revise the second order method presented in Schichl and Neumaier (2005a) and extend it to more generally shaped exclusion regions based on hypernorm balls. The main result of the article will be presented in Sections 5 and 6. Numerical and analytical

examples will be given in Section 7, where we will also show that the exclusion regions are optimally big in a certain sense.

In the following, the notation is as in the book Neumaier (2001). In particular, inequalities are interpreted component-wise, I denotes the identity matrix, intervals and boxes (= interval vectors) are in bold face, 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 $\Box S$.

Throughout the article we consider the global optimization problem

$$\begin{aligned} &\min f(x) \\ &\text{s.t. } F(x) = 0 \\ &\quad x \in \mathbf{x}, \end{aligned} \tag{1}$$

where $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ are three times continuously differentiable. (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.) Observe, that nonlinear inequality constraints are also covered by (1) since they can be converted to bound constraints by the introduction of slack variables. It is a straighforward calculation to eliminate the slack variables from the formulas developed in Sections 5 and 6. Since this increases the complexity of the formulas significantly, we refrained from doing so in this presentation. For a proper implementation, however, that step should be taken.

We will also consider the nonlinear system of equations

$$G(x) = 0, \tag{2}$$

for a twice continuously differentiable function $G : D' \subseteq \mathbb{R}^n \to \mathbb{R}^n$.

For a Lipschitz continuous function G we can always write

$$G(x) - G(z) = G[z, x](x - z) \tag{3}$$

for any two points x and z with a suitable matrix $G[z, x] \in \mathbb{R}^{n\times n}$, called a **slope matrix** for G. While $G[z, x]$ is not uniquely determined, we always have

$$G[z, z] = G'(z). \tag{4}$$

Thus $G[z, x]$ is a slope version of the Jacobian. There are recursive procedures to calculate $G[z, x]$ given x and z, see Krawczyk and Neumaier (1985), Rump (1996), Kolev (1997), and Schichl and Neumaier (2005b); a Matlab implementation is in INTLAB Rump (1999); also the COCONUT environment Schichl and Markót (2012) provides algorithms.

If the slope matrix $G[z, x]$ is Lipschitz-continuous we can further write

$$G[z, x] = G[z, z'] + (x - z')^T G[z, z', x] \tag{5}$$

with the **second order slope tensor** $G[z, z', x] \in \mathbb{R}^{n\times n\times n}$. Here, as throughout this paper, we use the following notation for third order tensors.

For a third order tensor $\mathcal{T} \in \mathbb{R}^{n\times m\times r}$, vectors $u \in \mathbb{R}^n$, $v \in \mathbb{R}^r$, $w \in \mathbb{R}^m$, and matrices $A \in \mathbf{R}^{s\times n}$, $B \in \mathbf{R}^{r\times s}$, and $C \in \mathbf{R}^{s\times m}$ we write

$$
\begin{aligned}
(\mathcal{T}^T)_{ijk} &= \mathcal{T}_{jki} & (\mathcal{T}^{\overline{T}})_{ijk} &= \mathcal{T}_{kij}\\
(u^T\mathcal{T})_{ij} &= \sum_k u_k\mathcal{T}_{kij} & (\mathcal{T}v)_{ij} &= \sum_k \mathcal{T}_{ijk}v_k\\
(w^T\cdot\mathcal{T})_{ij} &= \sum_k w_k\mathcal{T}_{ikj} & (\mathcal{T}\cdot w)_{ij} &= \sum_k \mathcal{T}_{ikj}w_k\\
(A\mathcal{T})_{ijk} &= \sum_\ell A_{i\ell}\mathcal{T}_{\ell jk} & (\mathcal{T}B)_{ijk} &= \sum_\ell \mathcal{T}_{ij\ell}B_{\ell k}\\
(C\cdot\mathcal{T})_{ijk} &= \sum_\ell C_{j\ell}\mathcal{T}_{i\ell k} & (\mathcal{T}\cdot C^T)_{ijk} &= \sum_\ell \mathcal{T}_{i\ell k}C_{\ell j}\\
(u^T\mathcal{T}v)_i &= \sum_{j,k} u_k\mathcal{T}_{kij}v_j & u^Tw^T\cdot\mathcal{T}v &= \sum_{i,j,k} u_iw_j\mathcal{T}_{ijk}v_k\\
(u^TC\cdot\mathcal{T})_{ij} &= \sum_{k,\ell} u_kC_{i\ell}\mathcal{T}_{k\ell j} & (u^TC\cdot\mathcal{T}v)_i &= \sum_{j,k,\ell} u_kC_{i\ell}\mathcal{T}_{k\ell j}v_j.
\end{aligned}
$$

It is important to note that the multiplication $\cdot$ binds more strongly than the "standard implicitly noted multiplication".

If $z = z'$ formula (5) above somewhat simplifies, because of (4), to

$$G[z, x] = G'(z) + (x - z)^T G[z, z, x]. \tag{6}$$

Throughout the article, the notion of a hypernorm will also be important. This is a joint generalization of norms and componentwise absolute values, originally introduced by Fischer (1974); see also Schichl and Neumaier (2011). Here, we will only need real valued hypernorms on $\mathbb{R}^n$.

Definition 1 A mapping $\nu : \mathbb{R}^n \to \mathbb{R}^r$ is called a **hypernorm** on $\mathbb{R}^n$ if for all $v, w \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ we have

(HN1) $\nu(v) \geq 0$ and $\nu(v) = 0$ iff $v = 0$,
(HN2) $\nu(\lambda v) = |\lambda|\nu(v)$,
(HN3) $\nu(v + w) \leq \nu(v) + \nu(w)$.

The hypernorm is called **monotone** if

(HNM) $0 \leq v \leq w$ implies $\nu(v) \leq \nu(w)$.

Let $\nu_n : \mathbb{R}^n \to \mathbb{R}^r$ and $\nu_m : \mathbb{R}^m \to \mathbb{R}^s$ be two hypernorms. A hypernorm $\nu_{m\times n} : \mathbb{R}^{m\times n} \to \mathbb{R}^{s\times r}$ is called **compatible** with ν_n and ν_m if for all $A \in \mathbb{R}^{m\times n}$ and all $v \in \mathbb{R}^n$ we have

$$\nu_m(Av) \leq \nu_{m\times n}(A)\nu_n(v).$$

The concept of compatibility is analogously extended from matrices to higher order tensors.

Let $\nu : \mathbb{R}^n \to \mathbb{R}^r$ be a hypernorm. A compatible hypernorm $\nu^* : \mathbb{R}^{n*} = \mathbb{R}^{1\times n} \to \mathbb{R}^{r*} = \mathbb{R}^{1\times r}$ is called a **dual** hypernorm for ν.

Let $\nu : \mathbb{R}^n \to \mathbb{R}^r$ be a hypernorm, and let $0 \leq u \in \mathbb{R}^r$ and $x \in \mathbb{R}^n$. The set

$$B_{u,\nu}(x) := \{y \in \mathbb{R}^n \mid \nu(x - y) \leq u\}$$

is called the **closed hypernorm ball** with center x and (generalized) radius u. A closed hypernorm ball is a nonempty convex set.

1. Every norm on $\mathbb{R}^n$ is a hypernorm, its dual norm is a dual hypernorm, and a compatible operator norm on $\mathbb{R}^{n\times n}$ is a compatible hypernorm on $\mathbb{R}^{n\times n}$. The closed hypernorm balls of a norm are the closed norm balls. If the norm is monotone, it is also monotone as a hypernorm.
2. The componentwise absolute value $|\ \ | : \mathbb{R}^n \to \mathbb{R}^n$ is a monotone hypernorm. It is dual to itself, and the componentwise absolute value on matrices is a compatible hypernorm. Its closed hypernorm balls are boxes.
3. For two hypernorms $\nu_1 : \mathbb{R}^n \to \mathbb{R}^r$ and $\nu_2 : \mathbb{R}^m \to \mathbb{R}^s$ the mapping $(\nu_1, \nu_2) : \mathbb{R}^{n+m} \to \mathbb{R}^{r+s}$ is again a hypernorm. Compatible and dual hypernorms can then be stacked accordingly. If both hypernorms are monotone, the stacked hypernorm is monotone, too.

2 The cluster effect

Branch and bound methods using constraint propagation methods (see, e.g., Van Hentenryck et al (1997)) for solving (1) in a verified way suffer from the so called **cluster effect**, see Du and Kearfott (1994). 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 reviewed in this section.

Lets consider the simplest case of (1), where $m = 0$ and $\mathbf{x} = \mathbf{R}^n$. In this unrestricted case, the problem reduces to finding the best local minimum of the function f. 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 $O(\varepsilon^k)$

$$f(\mathbf{x}) \in (1 + C\varepsilon^k)\square(\{f(x) \mid x \in \mathbf{x}\}) \tag{7}$$

for $k \geq 1$ and ε sufficiently small. The exponent k depends on the method used for the computation of $f(\mathbf{x})$.

Let x^* be a local minimum of f (so that $\nabla^2 f(x^*)$ is positive definite, i.e. it satisfies the sufficient second order optimality conditions), and assume (7). Then no box of diameter ε can be eliminated that contains a point x with

$$\|\nabla^2 f(x^*)(x - x^*)\|_\infty \leq \Delta = C\varepsilon^k. \tag{8}$$

This inequality describes a parallelepiped of volume

$$V = \frac{\Delta^n}{\det \nabla^2 f(x^*)}.$$

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

$$\begin{aligned} &\frac{C^n}{\det \nabla^2 f(x^*)} && \text{if } k = 2, \\ &\frac{(C\varepsilon)^n}{\det \nabla^2 f(x^*)} && \text{if } k = 3. \end{aligned}$$

For $k = 2$ 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 $k = 3$ 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 minimum. However, for ill-conditioned minima (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 cubic approximation property $k = 3$. Hence, Hessian 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 (using inclusion functions constructed by natural extension) 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 also not sufficient to avoid the cluster effect. Interval Newton-methods (see, e.g., Hansen (1978), Kearfott (1996b)) and second order centered forms (see, e.g., Schichl and Markót (2012)) provide information with $k = 3$, except near ill-conditioned or singular zeros.

For singular (and hence for sufficiently ill-conditioned) zeros, the argument above 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 (1996b).

3 Prerequisites

We consider the system of equations (2). We need the following theorem by Kahan (1968).

Theorem 1 *(Kahan) Let $z \in S$ for a compact convex set S. If there is a regular matrix $C \in \mathbb{R}^{n\times n}$ such that the Krawczyk operator*

$$K(z, x) := z - CG(z) - (CG[z, x] - I)(x - z) \tag{9}$$

satisfies $K(z,x) \in S$ for all $x \in S$ then S contains a zero of G.

Proof By Brouwer's fixed point theorem we can conclude that $K(z, .)$ has a fixed point $x^* \in S$. Then

$$K(z,x^*) = z - CG(z) - (CG[z,x^*] - I)(x^* - z) = x^*,$$

and so

$$0 = CG(z) + CG[z,x^*](x^* - z) = C(G(z) + G[z,x^*](x^* - z)) = CG(x^*).$$

Since C is regular, we conclude that $G(x^*) = 0$. □

4 Exclusion regions close to a zero of a system of equations

This section is devoted to proving a generalization of Schichl and Neumaier (2005a, Theorem 4.3) to allow more generally shaped exclusion regions.

Suppose that z is an approximate solution of the nonlinear system of equations (2). We will construct a pair of regions around z, an **inclusion region** R_i and an **exclusion region** R_e with the property that every solution x^* of (2) which lies in the interior of R_e must lie within R_i.

Take a fixed preconditioning matrix $C \in \mathbb{R}^{n\times n}$, and let $\nu_0 : \mathbb{R}^n \to \mathbb{R}^r$ be a hypernorm, and $\nu_1 : \mathbb{R}^{n\times n} \to \mathbb{R}^{r\times r}$ and $\nu_2 : \mathbb{R}^{n\times n\times n} \to \mathbb{R}^{r\times r\times r}$ be compatible hypernorms. In addition, assume that the hypernorm bounds

$$\begin{aligned} \overline{h} &\geq \nu_0(CG(z)) \geq \underline{h}, \\ H_0 &\geq \nu_1(CG'(z) - I), \\ \mathcal{H}(x) &\geq \nu_2(C \cdot G[z,z,x]) \end{aligned} \tag{10}$$

are satisfied for all $x \in X$, where X is a closed convex set.

We take an approximate zero z of G, and we choose C to be an approximation of $G'(z)^{-1}$. Now we prove a hypernorm version of Schichl and Neumaier (2005a, Proposition 4.1).

Proposition 1 *For every solution $x \in X$ of (2), the deviation*

$$s := \nu_0(x - z)$$

satisfies

$$0 \leq s \leq \Big(H_0 + s^T\mathcal{H}(x)\Big)s + \overline{h}. \tag{11}$$

Proof By (3) we have $G[z,x](x - z) = G(x) - G(z) = -G(z)$, because x is a zero. Hence, using (6), we compute

$$\begin{aligned} -(x-z) &= -(x-z) + C(G[z,x](x-z) + G(z) + G'(z)(x-z) \\ &\qquad - G'(z)(x-z)) \\ &= C(G[z,x] - G'(z))(x-z) + (CG'(z) - I)(x-z) + CG(z) \\ &= \Big(CG'(z) - I + (x-z)^T C \cdot G[z,z,x]\Big)(x-z) + CG(z). \end{aligned}$$

Now we apply the hypernorms, use (10), and get

$$\begin{aligned} s = \nu_0(x-z) &\le \Big(\nu_1(CG'(z)-I) + \nu_0(x-z)^T\nu_2(C\cdot G[z,z,x])\Big)\nu_0(x-z) \\ &\quad + \nu_0(CG(z)) \\ &\le \Big(H_0 + s^T\mathcal{H}(x)\Big)\, s + \overline{h}. \end{aligned}$$

□

Our next step will be to closely inspect the proof of Schichl and Neumaier (2005a, Theorem 4.2) and to dissect the results in such a way that we can make optimal use of them in the optimization setting.

Proposition 2 *Let $0 \le u \in \mathbb{R}^r$ be such that*

$$\Big(H_0 + u^T\overline{\mathcal{H}}\Big)u + \overline{h} \le u \tag{12}$$

with $\mathcal{H}(x) \le \overline{\mathcal{H}}$ for all $x \in M_u$, where

$$M_u := B_{u,\nu_0}(z) \cap X. \tag{13}$$

Then for $K(x) := x - CG(x)$ we find

$$K(x) \in B_{u,\nu_0}(z) \tag{14}$$

for all $x \in M_u$.

Proof Take any $x \in M_u$. We get

$$\begin{aligned} K(x) = x - CG(x) &= z - CG(z) - (CG[z,x] - I)(x-z) \\ &= z - CG(z) - \Big(C\Big(G'(z) + (x-z)^TG[z,z,x]\Big) - I\Big)\,(x-z), \end{aligned}$$

hence

$$K(x) = z - CG(z) - \Big(CG'(z) - I + (x-z)^TC\cdot G[z,z,x]\Big)\,(x-z). \tag{15}$$

Applying hypernorms we find

$$\begin{aligned} \nu_0(K(x)-z) &= \nu_0\left(-CG(z) - \Big(CG'(z) - I + (x-z)^TC\cdot G[z,z,x]\Big)\,(x-z)\right) \\ &\le \nu_0(CG(z)) + \Big(\nu_1(CG'(z)-I) + \nu_0(x-z)^T\nu_2(C\cdot G[z,z,x])\Big) \\ &\qquad \cdot \nu_0(x-z) \\ &\le \overline{h} + \Big(H_0 + u^T\overline{\mathcal{H}}\Big)\,u. \end{aligned} \tag{16}$$

Now assume (12). Then (16) gives $\nu_0(K(x)-z) \le u$, hence (14). □

Theorem 2 *In the situation of Proposition 2 let $B_{u,\nu_0}(z) \subseteq X$. Then there exists a solution x^* of (2) in M_u.*

Proof By Proposition 2 we have (14) for arbitrary $x \in M_u$. Since $B_{u,\nu_0}(z) \subseteq X$ this implies $K(x) \in M_u$. Because M_u is compact and convex, by Theorem 1 there exists a solution of (2) which lies in M_u. □

Note that (12) implies $H_0 u \le u$. If $J = \{j \mid u_j = 0\}$ and $V = \langle\{e_\ell \mid \ell \in J\}\rangle \subseteq \mathbf{R}^r$ is the subspace of all vectors v with $v_J = 0$, then $H_0(V) \subseteq V$ and the spectral radius $\rho(H_0|_V) \le 1$. In the applications, we can make $\overline{h}$ very small by choosing z as an approximate zero. For C we can choose an approximate inverse of $G'(z)$ such that $CG'(z) \approx I$.

Now the only thing that remains is the hypernorm version of Schichl and Neumaier (2005a, Theorem 4.3).

Theorem 3 *Let $S \subseteq X$ be any set containing z, and take*

$$\overline{\mathcal{H}} \ge \mathcal{H}(x) \quad \text{for all } x \in S. \tag{17}$$

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

$$w := (I - H_0)v, \quad a := v^T \overline{\mathcal{H}} v. \tag{18}$$

We suppose that

$$D_j = w_j^2 - 4a_j \overline{h}_j > 0 \tag{19}$$

for all $j = 1, \ldots, r$, and define

$$\lambda_j^e := \frac{w_j + \sqrt{D_j}}{2a_j}, \quad \lambda_j^i := \frac{\overline{h}_j}{a_j \lambda_j^e}, \tag{20}$$

$$\lambda^e := \min_{j=1,\ldots,r} \lambda_j^e, \quad \lambda^i := \max_{j=1,\ldots,r} \lambda_j^i. \tag{21}$$

If $\lambda^e > \lambda^i$ then there is at least one zero x^* of (2) in the (inclusion) region

$$R^i := B_{\lambda^i v,\nu_0}(z) \cap S. \tag{22}$$

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

$$R^e := B_{\lambda^e v,\nu_0}(z) \cap S. \tag{23}$$

Proof Let $0 < v \in \mathbb{R}^r$ be arbitrary, and set $u = \lambda v$. We check for which λ the vector u satisfies property (12) of Proposition 2. The requirement

$$\begin{aligned} \lambda v = u &\ge \left(H_0 + u^T \overline{\mathcal{H}}\right) u + \overline{h} = \left(H_0 + \lambda v^T \overline{\mathcal{H}}\right) \lambda v + \overline{h} \\ &= \overline{h} + \lambda H_0 v + \lambda^2 v^T \overline{\mathcal{H}} v, \end{aligned}$$

considered component-wise, gives a system of quadratic inequalities for λ. Hence, for every $\lambda \in [\lambda^i, \lambda^e]$ (this interval is nonempty by assumption), the vector u satisfies (12).

Now assume that x is a solution of (2) in $\operatorname{int}(R^e) \setminus R^i$. Let λ be minimal with $\nu_0(x-z) \le \lambda v$. By construction, $\lambda^i < \lambda < \lambda^e$. By the properties of the Krawczyk operator, we know that $x = K(x) = K(z,x)$, hence

$$\begin{aligned}\nu_0(x-z) &\le \nu_0(CG(z)) \\ &\quad + \Big(\nu_1(CG'(z)-I) + \nu_0(x-z)^T \nu_2(C\cdot G[z,z,x])\Big)\nu_0(x-z) \\ &\le \overline{h} + \lambda H_0 v + \lambda^2 v^T \overline{\mathcal{H}} v < \lambda v,\end{aligned} \tag{24}$$

since $\lambda > \lambda^i$. But this contradicts the minimality of λ. So there are indeed no solutions of (2) in $\operatorname{int}(R^e) \setminus R^i$. □

We will show in Example 1 that this hypernorm generalization is also the best possible in some cases.

We observe that the inclusion region from Theorem 3 can usually be further improved by noting that $x^* = K(z,x^*)$ and (15) imply

$$\begin{aligned}x^* \in K(z,\mathbf{x}^i) &= z - CG(z) - \Big(CG'(z) - I + (\mathbf{x}^i - z)^T C \cdot F[z,z,\mathbf{x}^i]\Big)(\mathbf{x}^i - z) \\ &\subset \operatorname{int}(\mathbf{x}^i).\end{aligned}$$

So after computing $\mathbf{x}^i$ by Theorem 3, performing the iteration $\mathbf{x}^i_{n+1} = K(z,\mathbf{x}^i_n)$ with $\mathbf{x}^i_0 = x^i$ will further shrink the inclusion region. A few iterations will be sufficient, since usually $\mathbf{x}^i$ is already quite small and the iteration converges quadratically.

An important special case is when $G(x)$ is quadratic in x. For such a function $G[z,x]$ is linear in x, and therefore all $G[z,z,x]$ are constant in x. This, in turn, means that $\mathcal{H}(x) = \mathcal{H}$ is constant as well. So we can set $\overline{\mathcal{H}} = \mathcal{H}$, and the estimate (17) becomes valid everywhere.

If the hypernorms ν_i are norms $\|\ \ \|$, then Theorem 3 simplifies, and no vector v needs to be chosen.

Corollary 1 *Let $\|\ \ \|$ denote a norm on $\mathbb{R}^n$ and corresponding compatible norms on $\mathbf{R}^{n\times n}$ and $\mathbf{R}^{n\times n\times n}$. Let furthermore for a fixed preconditioning matrix $C \in \mathbb{R}^{n\times n}$ the norm bounds*

$$\begin{aligned}\alpha &\ge \|CG(z)\|, \\ \beta &\ge \|CG'(z) - I\|, \\ \gamma &\ge \|C\cdot G[z,z,x]\|\end{aligned} \tag{25}$$

be satisfied for all $x \in S$, where $S \subseteq X$ is a set containing z. We set

$$\delta := (1-\beta)^2 - 4\alpha\gamma, \qquad \lambda^e := \frac{1-\beta+\sqrt{\delta}}{2\gamma}, \qquad \lambda^i := \frac{\alpha}{\gamma\lambda^e}. \tag{26}$$

If $\delta > 0$ and $\lambda^e > \lambda^i$ then there is at least one zero x^* of (2) in the (inclusion) region

$$\mathbf{x}^i := B_{\lambda^i}(z) \cap S. \tag{27}$$

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

$$\mathbf{x}^e := B_{\lambda^e}(z) \cap S. \tag{28}$$

Note that in general not only the choice of v matters in Theorem 3. The choice of S is also very important. If S is chosen too big, then the overestimation for $\overline{\mathcal{H}}$ might be large. In this case, the positivity requirement for the D_j will force the λ^e to be very small, much smaller than the size of S. Thus, it is necessary to balance the size of S and the expected size of the exclusion region.

One possible way to come up with a starting set S is to choose a box $z + v[-r, r]$ where the radius r is computed by Algorithm 1. There, we use the fact that $|u_1^T C \cdot G[z, z, x]u_2|$ can be computed in $O(p)$, where p is the effort for one point evaluation of G, but computing $\overline{\mathcal{H}}$ needs an effort of $O(n^2 p)$. We choose $u_1 = e_i$ and $u_2 = \hat{v}$ with $\nu_0(\hat{v}) = v$ to get a rough estimate of the size of $v^T \overline{\mathcal{H}} v$.

```
1  FindTrialBox(z, x, v);
   input  : z – the approximate zero
   input  : x – the search box
   input  : v – the vector v > 0
   output: S – trial box S
2  r_max := max{r | z + [−r, r]v ⊆ x};
3  C ≈ G'(z)^{-1};
4  r := r_max;
5  Compute v̂ with ν_0(v̂) = v;
6  for i := 1, ..., n do
7      while true do
8          q := e_i^T C · G[z, z, z + [−r, r]v̂];
9          ρ := 1/(‖v‖_1 ‖ν_0(q)‖_1);
10         if r = r_max ∧ ρ ≥ r then break;
11         if max(ρ, r)/ min(ρ, r) < 2 then
12             r := min(r_max, max(ρ, r));
13             break;
14         end
15         if ρ = 0 then
16             r := (1/2)√r;
17         else
18             r := √(ρr);
19         end
20         if r > r_max then r := r/2;
21     end
22     r_max := r;
23 end
24 return S := z + (1/2)[−r_max, r_max];
```

The tensor $\overline{\mathcal{H}}$ can be constructed using interval arithmetic, for a given reference box $\mathbf{x}$ around z. Using backward evaluation schemes (see e.g. Schichl and Markót (2010)) the effort for computing this third order tensor is $O(n^2 p)$.

5 Exclusion regions for optimization problems

In this section we consider problem (1). We want to find an analogous result to Theorem 3 in the situation of an optimization problem. We could formulate the Karush-John first order necessary optimality conditions as a system of equations and apply Theorem 3 directly. However, in many cases this system is degenerate and hence cannot be verified. Therefore, we take a more direct approach.

Throughout the section we will need the Karush-John first order optimality conditions for (1), see Karush (1939); John (1948).

Theorem 4 (General first order optimality conditions) *Let $f : U \to \mathbb{R}$ and $F : U \to \mathbb{R}^m$ be functions continuously differentiable on a neighborhood U of $x^* \in \mathbb{R}^n$. If x^* is a locally optimal point of the nonlinear program* (1)*, then there exist a scalar $\kappa^* \geq 0 \in \mathbb{R}$, a vector $w^* \in \mathbb{R}^m$, and a vector $q^* \in \mathbb{R}^n$ such that*

$$q^* = \kappa^* \nabla f(x^*) + \nabla F(x^*) w^*, \tag{29}$$

$$q_k^* \begin{cases} \geq 0 & \text{if } \underline{x}_k = x_k^* < \overline{x}_k \\ \leq 0 & \text{if } \underline{x}_k < x_k^* = \overline{x}_k \\ = 0 & \text{if } \underline{x}_k < x_k^* < \overline{x}_k \end{cases} \tag{30}$$

and

$$\kappa^*, w^* \quad \text{are not both zero.} \tag{31}$$

Let $z \in \mathbb{R}^n$ be an approximate local solution of (1), $y \in \mathbb{R}^m$ an approximation for the corresponding multiplier vector w^*, and $\sigma \geq 0$ an approximation for κ^*. As usual, we define the Lagrange function

$$L(x, w, \kappa) := \kappa f(x) + w^T F(x).$$

We set $s := \nabla_x L(z, y, \sigma)^T \approx q^*$. Then we consider the complementarity conditions for q^*, which are approximately satisfied for s. For $0 < \alpha \in \mathbb{R}^n$ we define the α–**active set** $I_\alpha := \{i \mid |s_i| \geq \alpha_i\}$ and the α–**inactive set** $J_\alpha := \{i \mid |s_i| < \alpha_i\}$. We choose $\alpha > 0$ such that $|J_\alpha| \geq m$. This is needed, since we need enough free variables to satisfy the equality constraints. If no such α exists, this method does not work due to degeneracy. One common source of such a degeneracy is the replacement of equality constraints by pairs of inequality constraints in certain automatic reformulations. This is the reason, why the COCONUT Environment (Schichl and Markót (2013)) automatically detects and undoes such reformulations. There are, however, classes of problems, e.g. those described in Kearfott et al (2012), that are naturally degenerate. In an LP context some degeneracy can be removed by automatic preprocessing. For nonlinear global optimization, symbolic preprocessing steps or rational arithmetic are necessary for coping with degeneracy. On the other hand, such methods are difficult to implement and computationally expensive in general.

Furthermore, we require for $i \in I_\alpha$ that

$$z_i = x_i^b := \begin{cases} \underline{x}_i & \text{if } s_i \geq \alpha_i \\ \overline{x}_i & \text{if } s_i \leq -\alpha_i. \end{cases}$$

If this is not satisfied from the beginning, we change z accordingly.

In the following we fix $0 < \alpha \in \mathbb{R}^n$ and set $J := J_\alpha$ and $I := I_\alpha$. We also introduce the following short-hand notation for the larger formulas later:

$$t := \begin{pmatrix} x_J \\ w \\ \kappa \\ x_I \end{pmatrix}, \quad \bar{t} := \begin{pmatrix} x_J \\ w \\ \kappa \end{pmatrix}, \quad u := \begin{pmatrix} z_J \\ y \\ \sigma \\ z_I \end{pmatrix}, \quad \bar{u} := \begin{pmatrix} z_J \\ y \\ \sigma \end{pmatrix}.$$

For the further estimates, especially on the third order tensors, we define W to be the index set for the multipliers w and introduce the combined index set $M := J \cup W \cup \{\kappa\}$, so $\bar{t}_W = t_W = w$, $\bar{t}_J = t_J = x_J$, $u_M = \bar{u}$, etc.

Now we construct a function G for which we can apply Theorem 3. We use the inactive part of equation (29), the original equality constraints, a smooth version of (31), and the active boundary constraints

$$G(t) := \begin{pmatrix} \nabla_J L(t) \\ F(x) \\ \kappa^2 + w^T w - 1 \\ x_I - x_I^b \end{pmatrix}. \tag{32}$$

However, for the formulation of the final result we get rid of the simple components corresponding to the bound constraints. So to properly apply Theorem 3 we need to compute estimates analogous to (24) and before that we must find a proper preconditioning matrix C. We set

$$C' := \begin{pmatrix} \nabla^2_{JJ} L(u) & \nabla_J F(z) & \nabla_J f(z) \\ \nabla_J F(z)^T & 0 & 0 \\ 0 & 2y^T & 2\sigma \end{pmatrix} \tag{33}$$

and $C \approx {C'}^{-1}$.

Using this preconditioning matrix we compute the hypernorm bounds (24) for the special case we consider. We fix monotone hypernorms $\nu_1 : \mathbb{R}^{\widetilde{N}} \to \mathbb{R}^r$, $\nu_0 : \mathbb{R}^{\tilde{I}} \to \mathbb{R}^{\tilde{r}}$, and compatible hypernorms $\nu_{2,0} : \mathbb{R}^{\widetilde{N} \times \tilde{I}} \to \mathbb{R}^{r \times \tilde{r}}$, $\nu_{2,1} : \mathbb{R}^{\widetilde{N} \times \widetilde{N}} \to \mathbb{R}^{r \times r}$, $\nu_{3,1} : \mathbb{R}^{\widetilde{N} \times \widetilde{N} \times N} \to \mathbb{R}^{r \times r \times (r+\tilde{r})}$, $\nu_{3,2} : \mathbb{R}^{\widetilde{N} \times \tilde{I} \times N} \to \mathbb{R}^{r \times \tilde{r} \times R}$, where $N = n + m + 1$, $\widetilde{N} = |J| + m + 1$, $\tilde{I} = |I|$, and $R = r + \tilde{r}$. Furthermore, we let

$$\begin{aligned} \bar{b} &\geq \nu_1 \left(C \begin{pmatrix} \nabla_J L(u) \\ F(z) \\ z^2 + y^T y - 1 \end{pmatrix} \right) \\ B_0 &\geq \nu_{2,1}(CC' - I). \end{aligned} \tag{34}$$

For the estimation we define the tensor valued functions $\widetilde{\mathcal{B}} : \mathbb{R}^N \to \mathbb{R}^{\widetilde{N}\times\widetilde{N}\times N}$ and $\widetilde{\mathcal{T}} : \mathbb{R}^N \to \mathbb{R}^{\widetilde{N}\times\tilde{I}\times N}$ in block form as

$$\begin{aligned}
\widetilde{\mathcal{B}}_{J::}(t,u) &:= C \cdot \begin{pmatrix} (\nabla_J L)[u,u,t]_{J:J} & 0 & 0 \\ F[z,z,x]_{J:J} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \widetilde{\mathcal{B}}_{I::}(t,u) := C \cdot \begin{pmatrix} (\nabla_J L)[u,u,t]_{I:J} & 0 & 0 \\ F[z,z,x]_{I:J} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
\widetilde{\mathcal{B}}_{W::}(t,u) &:= C \cdot \begin{pmatrix} \nabla^2_{JJ} F(z)^{\overline{T}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & e_W^T & 0 \end{pmatrix} \qquad \widetilde{\mathcal{B}}_{\kappa::}(t,u) := C \cdot \begin{pmatrix} \nabla^2_{JJ} f(z) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\
\widetilde{\mathcal{T}}_{J::}(t,u) &:= C \cdot \begin{pmatrix} (\nabla_J L)[u,u,t]_{J:I} \\ F[z,z,x]_{J:I} \\ 0 \end{pmatrix} \qquad \widetilde{\mathcal{T}}_{I::}(t,u) := C \cdot \begin{pmatrix} (\nabla_J L)[u,u,t]_{I:I} \\ F[z,z,x]_{I:I} \\ 0 \end{pmatrix} \\
\widetilde{\mathcal{T}}_{W::}(t,u) &:= C \cdot \begin{pmatrix} \nabla^2_{JI} F(z)^{\overline{T}} \\ 0 \\ 0 \end{pmatrix} \qquad \widetilde{\mathcal{T}}_{::\kappa}(t,u) := C \cdot \begin{pmatrix} \nabla^2_{JI} f(z) \\ 0 \\ 0 \end{pmatrix}.
\end{aligned} \tag{35}$$

For the next step we consider $S \subseteq X$ and boxes $\mathbf{w} \subseteq \mathbb{R}^m$, $\mathbf{k} \subseteq \mathbf{R}+$ and choose a vector $0 < v^T = (v_J^T, v_W^T, v_\kappa, v_I) \in \mathbb{R}^R$ and analogously to (17)–(21) proceed by requiring the estimates

$$\begin{aligned}
\mathcal{B} &\geq \nu_{3,1}(\widetilde{\mathcal{B}}(t,u)) \\
\mathcal{T} &\geq \nu_{3,2}(\widetilde{\mathcal{T}}(t,u))
\end{aligned} \tag{36}$$

for all $x \in S \subseteq X$, $w \in \mathbf{w}$, and $\kappa \in \mathbf{k}$, and define

$$w := (I - B_0)v_M, \quad a := v^T(\mathcal{B}v_M + \mathcal{T}v_I). \tag{37}$$

For all $j \in \{1,\ldots,r\}$ we set

$$D_j = w_j^2 - 4a_j\bar{b}_j, \tag{38}$$

and for all $j \in M_0 := \{i \in \{1,\ldots,r\} \mid D_j > 0\}$

$$\lambda_j^e := \frac{w_j + \sqrt{D_j}}{2a_j}, \quad \lambda_j^i := \frac{\bar{b}_j}{a_j\lambda_j^e}, \quad \lambda^e := \min_{j\in M_0} \lambda_j^e, \quad \lambda^i := \max_{j\in M_0} \lambda_j^i. \tag{39}$$

In case that $J = \emptyset$, we set $\lambda^e = \infty$ and $\lambda^i = 0$.

Next we have to take care of the boundary. Until now we have fixed the solution to the boundary in all components x_i for $i \in I$. To extend the exclusion region into those components we calculate the following estimates. We set $\delta := \nabla_I L(u)$,

$$C_B(t,u) := \big((\nabla_I L)[u,t]_{:J},\ \nabla_I F(z),\ \nabla_I f(z)\big), \tag{40}$$

and compute for all $i \in I$ the estimates

$$\begin{aligned}
Y_i &\geq \nu_1^*(C_B(t,u)_{i:}), \quad \text{and} \\
Z_i &\geq \nu_0^*(Z_{:i}^L(t,u)),
\end{aligned} \tag{41}$$

for all $t \in B_{\lambda^e v,\xi_0}(u)$ with $x \in \mathbf{x}$, where $\xi_0 := (\nu_1, \nu_0)^T : \mathbb{R}^N \to \mathbb{R}^R$ is the stacked hypernorm and ν_0^* and ν_1^* are dual hypernorms of ν_0 and ν_1, respectively. Here we set for $j, k \in I$

$$Z^L_{jk}(t,u) = \begin{cases} (\nabla_k L)[u,t]_j & j \in I, (\nabla_k L)[u,t]_j < 0, z_j \text{ active} \\ 0 & \text{otherwise.} \end{cases} \tag{42}$$

Now we have assembled everything to formulate the exclusion box theorem.

Theorem 5 *Choose a vector $0 < v \in \mathbb{R}^R$ such that all estimates in (34)–(41) are valid in the set S. Assume further $D_i > 0$ for all $i = \{1, \ldots, r\}$. We define*

$$\mu_i := \frac{|\delta_i|}{Y^T_{i:} v_M + Z^T_{i:} v_I}, \quad \text{for } i \in I, \text{ and } \quad \mu^e := \min(\min_{i \in I} \mu_i, \lambda^e). \tag{43}$$

If now $\mu^e > \lambda^i$ then there exists a KJ–point (x^*, w^*, κ^*) for (1) in the inclusion region $R^i := B_{\lambda^i v_M, \nu_1}(z_J, y, \sigma) \times \{x^b_I\} \cap S$. All KJ–points of (1) in the interior of the exclusion region $R^e := B_{\mu^e v, \xi_0}(z, y, \sigma) \cap S$ are in R^i.

Proof We start by considering the function G as defined in (32). The point $u = (z, y, \sigma)$ is an approximate solution of $G(t) = 0$. To apply Theorem 3 we must calculate the hypernorm bounds (10). We find

$$G[u,t] = \begin{pmatrix} (\nabla_J L)[u,t]_{:J} & \nabla_J F(z) & \nabla_J f(z) & (\nabla_J L)(u,t)_{:I} \\ F[z,x]_{:J} & 0 & 0 & F[z,x]_{:I} \\ 0 & w^T + y^T & \kappa + \sigma & 0 \\ 0 & 0 & 0 & I \end{pmatrix}, \tag{44}$$

and hence

$$G'(u) = \begin{pmatrix} \nabla^2_{JJ} L(u) & \nabla_J F(z) & \nabla_J f(z) & \nabla^2_{JI} L(u) \\ \nabla_J F(z)^T & 0 & 0 & \nabla_I F(z)^T \\ 0 & 2y^T & 2\sigma & 0 \\ 0 & 0 & 0 & I \end{pmatrix} = \begin{pmatrix} C' & C'' \\ 0 & I \end{pmatrix}. \tag{45}$$

The second order slopes can also be calculated

$$G[u,u,t]_{J::} = \begin{pmatrix} (\nabla_J L)[u,u,t]_{J:J} & 0\,0 & (\nabla_J L)[u,u,t]_{J:I} \\ F[z,z,x]_{J:J} & 0\,0 & F[z,z,x]_{J:I} \\ 0 & 0\,0 & 0 \\ 0 & 0\,0 & 0 \end{pmatrix}, \tag{46}$$

$$G[u,u,t]_{W::} = \begin{pmatrix} \nabla^2_{JJ} F(z)^{\overline{T}} & 0 & 0 & \nabla_{JI} F(z)^{\overline{T}} \\ 0 & 0 & 0 & 0 \\ 0 & e_W^T & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{47}$$

$$G[u,u,t]_{\kappa::} = \begin{pmatrix} \nabla^2_{JJ} f(z) & 0\,0 & \nabla^2_{JI} f(z) \\ 0 & 0\,0 & 0 \\ 0 & 0\,1 & 0 \\ 0 & 0\,0 & I \end{pmatrix}, \tag{48}$$

$$G[u,u,t]_{I::} = \begin{pmatrix} (\nabla_J L)[u,u,t]_{I:J} & 0\,0 & (\nabla_J L)[u,u,t]_{I:I} \\ F[z,z,x]_{I:J} & 0\,0 & F[z,z,x]_{I:I} \\ 0 & 0\,0 & 0 \\ 0 & 0\,0 & 0 \end{pmatrix}. \tag{49}$$

We see that $G'(u)$ is regular iff C' is, and that the matrix

$$C_G := \begin{pmatrix} C & -CC'' \\ 0 & I \end{pmatrix}$$

is an approximate inverse.

By comparing (46)–(49) and (35), we further note that $\mathcal{B}(t,u) = \big(C_G \cdot G[u,u,t]\big)_{:MM}$, $\mathcal{T}(t,u) = \big(C_G \cdot G[u,u,t]\big)_{:MI}$, and that all the remaining components of $C_G \cdot G[u,u,t]$ vanish.

We consider the hypernorm $\xi_0 := (\nu_1, \nu_0)^T : \mathbb{R}^N \to \mathbb{R}^R$, the compatible hypernorm

$$\xi_1 := \begin{pmatrix} \nu_{2,1}\ \nu_{2,0} \\ \hat{\nu}_{2,3}\ \hat{\nu}_{2,2} \end{pmatrix} : \mathbb{R}^{N\times N} \to \mathbb{R}^{R\times R},$$

where $\hat{\nu}_{2,3} : \mathbb{R}^{\tilde{I}\times\tilde{N}} \to \mathbb{R}^{\tilde{r}\times r}$, and $\hat{\nu}_{2,2} : \mathbb{R}^{\tilde{I}\times\tilde{I}} \to \mathbb{R}^{\tilde{r}\times\tilde{r}}$ are hypernorms compatible to ν_0 and ν_1, and the compatible hypernorm

$$\xi_2 := \begin{pmatrix} \nu_{3,1}\ \nu_{3,2} \\ \hat{\nu}_{3,3}\ \hat{\nu}_{3,4} \end{pmatrix} : \mathbb{R}^{N\times N\times N} \to \mathbb{R}^{R\times R\times R},$$

where $\hat{\nu}_{3,3} : \mathbb{R}^{\tilde{I}\times\tilde{N}\times N} \to \mathbb{R}^{\tilde{r}\times r\times R}$ and $\hat{\nu}_{3,4} : \mathbb{R}^{\tilde{I}\times\tilde{I}\times N} \to \mathbb{R}^{\tilde{r}\times\tilde{r}\times R}$ are hypernorms compatible to ν_0, ν_1, $\nu_{2,0}$, $\nu_{2,1}$, $\hat{\nu}_{2,2}$, and $\hat{\nu}_{2,3}$.

Then

$$\xi_0(C_G G(z,w,\sigma)) \le \begin{pmatrix} \overline{b} \\ 0 \end{pmatrix} =: \overline{h}$$

$$\xi_1(C_G G'(z) - I) \le \begin{pmatrix} B_0\ 0 \\ 0\ \ 0 \end{pmatrix} =: H_0 \quad \text{and} \quad \xi_2(C_G \cdot G[u,u,t]) \le \begin{pmatrix} \mathcal{B}\ \mathcal{T} \\ 0\ \ 0 \end{pmatrix} =: \overline{\mathcal{H}}.$$

Now we calculate (17) and (18)

$$\tilde{w} := (I - H_0)v = \begin{pmatrix} (I - B_0)v_M \\ v_I \end{pmatrix} = \begin{pmatrix} w \\ v_I \end{pmatrix},$$
$$\tilde{a} := v^T \mathcal{H} v = \begin{pmatrix} v^T(\mathcal{B}v_M + \mathcal{T}v_I) \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ 0 \end{pmatrix}.$$

For (19)–(20) we split the indices in two parts. For those j that belong to the last $\tilde{r}$ indices of $\mathbb{R}^{r+\tilde{r}}$ we find $\widetilde{D}_j = v_j$, $\tilde{\lambda}^e_j = +\infty$, and $\tilde{\lambda}^i_j = 0$, so they do not play a role for calculating the sizes of the exclusion and inclusion regions, respectively.

For the first r indices we find the expressions (37)–(39). The λ^e and λ^i calculated in (39) define the sizes of exclusion and inclusion regions, provided that the solution does not leave the boundary in the active indices I. So this is the maximal size of an exclusion region we can expect.

Let $i \in I$ be an index of an active component. By the definition of I we have $|\delta_i| > 0$. We do not reach another KJ–point of (1) if we make sure that $\nabla_i L(x, w, \kappa)$ stays away from 0 when the point moves from the boundary of $x_i = x_i^b$ into the relative interior of the feasible set.

Fix $t \in \mathrm{int} R^e$. Then

$$\begin{aligned}\nabla_i L(t) &= \delta_i + (\nabla_i L)[u,t](t-u)\\ &= \delta_i + (\nabla_i L)[u,t]_{:I}(x_I - z_I) + (\nabla_i L)[u,t]_{:J}(x_J - z_J) + \nabla_i F(z)(w-y)\\ &\quad + \nabla_i f(z)(\kappa - \sigma)\\ &= \delta_i + C_B(t,u)_{i:}(\bar{t} - \bar{u}) + (\nabla_i L)[u,t]_I(x_I - z_I).\end{aligned}$$

Thus

$$\begin{aligned}\left|\nabla_i L(t)\right| &\geq \left|\delta_i + (\nabla_i L)[u,t]_I(x_I - z_I)\right| - \left|C_B(t,u)_{i:}(\bar{t} - \bar{u})\right|\\ &\geq \left|\delta_i + Z_{Ii}^L(t,u)^T(x_I - z_I)\right| - \left|C_B(t,u)_{i:}(\bar{t} - \bar{u})\right|\\ &\geq |\delta_i| - \nu_0^*(Z_{Ii}^L(t,u))^T\nu_0(x_I - z_I) - \nu_1^*(C_B(t,u)_{i:})^T\nu_1(\bar{t} - \bar{u})\\ &> |\delta_i| - \mu^e(Z_i^T v_I + Y_i^T v_M)\\ &\geq |\delta_i| - \mu_i(Z_i^T v_I + Y_i^T v_M) = 0,\end{aligned}$$

because of (52), (43), and since the hypernorms are monotone. Hence, every component of $\nabla_I L$ is nonzero in the interior of R^e, so there cannot be another KJ–point there. This proves, that R^e is indeed an exclusion region. □

A small disadvantage of Theorem 5 is that the inclusion/exclusion regions also involve the multipliers. However, often it is possible to give good bounds on the multipliers using the KJ–System. If this is the case, the exclusion regions can often be projected to the x component.

Corollary 2 *Let the situation be as in Theorem 5. If* $x \in X^e = \mathrm{int}\,\mathrm{Pr}_x(R^e)$ *and* (29)–(31) *imply* $(x, w, \kappa) \in \mathrm{int} R^e$*, then all solutions* $x^* \in X^e$ *are in the inclusion region* $X^i = \mathrm{Pr}_x(R^i)$*. (Here* Pr_x *specifies the projection to the* x *components.)*

Proof Take a solution $x^* \in \mathrm{int}\,\mathrm{Pr}_x(R^e)$. Then by Theorem 4 there exist w^* and κ^* which satisfy the KJ–conditions (29)–(31). By assumption, we get $(x^*, w^*, \kappa^*) \in \mathrm{int} R^e$. Then Theorem 5 implies $(x^*, w^*, \kappa^*) \in R^i$, and thus $x^* \in X^i$. □

Remark 1
- If $DF(\hat{\mathbf{x}})$ has full rank or F is linear, then $\kappa \neq 0$. In that case, κ can be omitted from all equations, and the system simplifies.
- The KJ system provides a linear interval equation for w, which can be used to get an estimate for w and prove the condition of Corollary 2.
- If z is an approximate strict local minimum, then usually it can be proved that there exists a strict local minimum in R^i.

A very important special case is the bound constrained case, where $m = 0$:

$$\begin{aligned} &\min f(x) \\ &\text{s.t. } x \in \mathbf{x}. \end{aligned} \tag{50}$$

Then, w does not appear in the KJ–conditions (29) and (31), and $\kappa = 1$ can always be assumed. Hence, all estimates and Theorem 5 simplify significantly.

In this case $s = \nabla_x f(z)$ and I and J are defined analogously as before without any restrictions on J. The preconditioning matrix is $C \approx (\nabla^2_{JJ} f(z))^{-1}$. We choose again monotone hypernorms $\nu_1 : \mathbb{R}^{|J|} \to \mathbb{R}^r$, $\nu_0 : \mathbb{R}^{|I|} \to \mathbb{R}^{\tilde{r}}$ and compatible operator hypernorms ν_i. We compute the following estimates:

$$\begin{aligned} \overline{b} &\geq \nu_1(C\nabla_J f(z)) \\ B_0 &\geq \nu_2(C\nabla_{JJ} f(z) - I) \\ \widetilde{\mathcal{B}}(x,z) &:= C \cdot (\nabla_J f)[z,z,x] \\ \mathcal{B} &\geq \nu_3(\widetilde{\mathcal{B}}(x,z)) \quad \text{for all } x \in S \subseteq X. \end{aligned} \tag{51}$$

Choose $0 < v \in \mathbb{R}^R$ and set $w := (I - B_0)v$, $a := v^T \mathcal{B} v$. Define D_j for $j \in J$, λ^e, and λ^i as in (38) and (39).

We set $\delta := \nabla_I f(z)$, and compute for all $i \in I$ the estimates

$$\begin{aligned} &Y_i \geq \nu_1^*((\nabla_i f)[z,x]_J), \quad \text{and} \\ &Z_i \geq \nu_0^*(Z^L_{:i}(x,z)) \end{aligned} \tag{52}$$

for all $x \in B_{\lambda^e v, \xi_0}(z)$ with $x \in \mathbf{x}$, where $\xi_0 := (\nu_1, \nu_0)^T : \mathbb{R}^N \to \mathbb{R}^R$ is again the stacked hypernorm and ν_0^* and ν_1^* are dual hypernorms of ν_0 and ν_1, respectively. Here we set for $j, k \in I$

$$Z^L_{jk}(x,z) = \begin{cases} (\nabla_k f)[z,x]_j & j \in I, (\nabla_k f)[z,x]_j < 0, z_j \text{ active} \\ 0 & \text{otherwise.} \end{cases} \tag{53}$$

Corollary 3 *Let the estimates* (51) *and* (52) *be valid in the set S and define*

$$\begin{aligned} \mu_i &:= \frac{|\nabla_i f(z)|}{Y_i^T v_J + Z_i^T v_I}, \quad \textit{for } i \in I \\ \mu^e &:= \min(\min_{i \in I} \mu_i, \lambda^e). \end{aligned}$$

If $D_j > 0$ for all $j = 1, \ldots, r$ and $\mu^e > \lambda^i$ then there exists a critical point x^* for (50) *in the inclusion region $R^i := B_{\lambda^i v_J, \nu_1}(z_J) \times \{x^b_I\} \cap S$. These are the only critical points of* (50) *in the interior of the exclusion region $R^e := B_{\mu^e v, \xi_0}(z) \cap S$.*

Proof This follows directly from Theorem 5. □

An even more special case is unconstrained optimization

$$\begin{aligned} &\min f(x), \\ &\text{s.t. } x \in \mathbb{R}^n \end{aligned} \tag{54}$$

for which the exclusion regions can be calculated with even less effort. We fix a single monotone hypernorm $\nu_1 : \mathbb{R}^n \to \mathbb{R}^r$, and get the following result. Compute $C \approx (\nabla^2 f(z))^{-1}$.

Corollary 4 *Let the estimates*

$$\begin{aligned} \bar{b} &\geq \nu_1(C\nabla f(z)) \\ B_0 &\geq \nu_2(C\nabla^2 f(z) - I) \\ \mathcal{B} &\geq \nu_3(C \cdot (\nabla f)[z, z, x]) \quad \textit{for all } x \in S \subseteq X \end{aligned}$$

be valid. Fix $0 < v \in \mathbb{R}^r$ *and set* $w := (I - B_0)v$, $a := v^T \mathcal{B} v$. *Define* D_j *for* $j \in J$, λ^e, *and* λ^i *as in* (38) *and* (39).

If now $\lambda^e > \lambda^i$ *then there exists a critical point* x^* *for* (54) *in the inclusion region* $R^i := B_{\lambda^i v, \nu_1}(z) \cap S$. *All critical points of* (54) *in the interior of the exclusion region* $R^e := B_{\lambda^e v, \nu_1}(z) \cap S$ *are in* R^i.

Proof This follows directly from Corollary 3. □

Theorem 5 and Corollaries 3 and 4 rely heavily on third order information. Great care has to be taken in the implementation that the effort for computing this information is not too high. If implemented correctly the second order slopes of a first derivative can be computed in two efficient ways. Either they can be computed directly as second order slopes from the first derivative expressions from the KJ conditions. Alternatively, they can be directly calculated in backward mode like third derivatives as $w^T \nabla^3 L v$, as described in Schichl and Markót (2010). Both methods require an effort of $O(n^2 f)$, where f denotes the computation time for one function evaluation. The direct method has one big advantage, though: Preconditioning the third order tensors in the form $C \cdot \nabla^3 L$ does not cause additional effort, since the multiplication can be evaluated by choosing the w as the rows of C. That keeps the computational complexity at $O(n^2 f)$ and avoids the additional $O(n^4)$ operations required for explicitly evaluating the matrix-tensor product. All the remaining linear algebra needed is $O(n^3)$, so that the overall effort for computing an inclusion/exclusion region is $O(n^2(n + f))$.

If computation algorithms for first and second order slopes of first derivatives are not available, then all expressions of the form $(\nabla L)[z, \mathbf{x}]$ can be replaced by $\nabla^2 L(\mathbf{x})$ and the second order slopes $(\nabla L)[z, z, \mathbf{x}]$ and $(\nabla L)[z, \mathbf{x}, \mathbf{x}]$ can be estimated by $\frac{1}{2}\nabla^3 L(\mathbf{x})$. In that case, Theorem 5 already proves uniqueness of the local optimizer, and Theorem 7 below is not needed.

The question will be, whether it is possible to avoid the third order information altogether. This is not easy. As discussed in Section 2, at least second order information is needed to avoid the cluster effect.

- We could apply first order methods to the KJ system. This does not work very well, except in a backboxing setting, and does not provide large exclusion regions in a single calculation.
- It might be possible to utilize the second order necessary optimality conditions:
$$s_J = 0, F'(x)s = 0 \Rightarrow s^T \hat{G}(x)s \geq 0,$$
where $\hat{G}$ is the reduced Hessian of the Lagrange function. We could use zero-order information on this system. But this does not work well either, because it is difficult to represent the set of s satisfying the left hand side of the implication properly by zero order information ($F'(\mathbf{x})s = 0$ describes a fairly complicated set if $F'(\mathbf{x})$ is enclosed in an interval matrix, see Rohn (1989); Neumaier (1990)). Using the equation $F'(x)s = 0$ efficiently directly in the right hand side of the implication is only possible symbolically, and only if the function F has a special structure (e.g. being affine, etc.).
- Interval Newton type methods could be tried, but those are even in the unconstrained case significantly weaker than the presented approach using third order information.

The regularity of C' in (33) implies that z is an approximate strict local minimum. If this is not the case the problem is too degenerate, and Theorem 5 is not applicable. The prerequisite $|J_\alpha| \geq m$ is absolutely necessary to ensure regularity of the matrix C'. This means that there are enough free variables to make $F(x) = 0$ possible.

The exclusion regions are constructed not to contain another KJ–points, even if they are just maxima or saddle-points. So incorporating the additional constraint $g(x) \leq \overline{f}$ for $\overline{f} \geq f(R^i)$ in some form for a second iteration, where $g(x) \leq f(x)$, might possibly increase the size of the exclusion region.

Of course, as for systems of equations, the size of the initial set S has a significant influence on the size of the computed exclusion box. An analogue to Algorithm 1 needs to be used to compute this initial box.

6 Uniqueness regions

An important aspect of exclusion regions is the notion of uniqueness regions, where we can prove that strict local minima are unique.

Before we can approach that goal, we need to generalize the uniqueness regions of Schichl and Neumaier (2005a, Theorem 6.1) to hypernorm variants.

Theorem 6 *Take an approximate solution $z \in \mathbf{x}^e$ of (2), and let $B \in \mathbb{R}^{r\times r}$ be a matrix such that*

$$\nu_1(CG[z,x] - I) + \nu_0(x - z)^T \nu_2(C \cdot G[z,x,x]) \leq B \tag{55}$$

for all $x \in \mathbf{x}^i$. If $\|B\| < 1$ for some monotone norm then $\mathbf{x}^e$ contains at most one solution x^* of (2).

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

$$0 = G(x') - G(x) = G[x, x'](x' - x) = \Big(G[x, z] + (x' - z)^T G[z, x, x']\Big)(x' - x). \tag{56}$$

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

$$x - x' = \Big((CG[z, x] - I) + (x' - z)^T C \cdot G[z, x, x']\Big)(x' - x). \tag{57}$$

Applying hypernorms, and using (55), we find

$$\begin{aligned}\nu_0(x' - x) &\le \Big(\nu_1(CG[z, x] - I) + \nu_0(x' - z)^T \nu_2(C \cdot G[z, x, x'])\Big)\nu_0(x' - x)\\ &\le B\,\nu_0(x' - x).\end{aligned}$$

This, in turn, implies $\|\nu_0(x' - x)\| \le \|B\|\,\|\nu_0(x' - x)\|$. If $\|B\| < 1$ we immediately conclude $\nu_0(x' - x) = 0$, hence $x = x'$. □

Using this result we can construct regions in which there is a unique strict local minimum, in the following way. First one verifies as in Section 5 an exclusion region R^e which contains no local minimum except in a much smaller inclusion region R^i. Then we try to further refine the inclusion region by some iterations similar to Krawczyk's method, which generally converges quickly if the initial inclusion box is already verified, as follows:

We define the function $\widehat{G} : \mathbb{R}^{\widetilde{N}} \to \mathbb{R}^{\widetilde{N}}$ by

$$\widehat{G}(\bar{t}) := \begin{pmatrix} (\nabla_J L|_{x_I = x_I^b})(\bar{t}) \\ (F|_{x_I = x_I^b})(x_J) \\ \kappa^2 + w^T w - 1 \end{pmatrix},$$

and the third order Krawczyk-type operator

$$K(S, \bar{u}) := \big\{\bar{u} - C\widehat{G}(\bar{u}) - \big(CC' - I - (\bar{t} - \bar{u})^T \widetilde{\mathcal{B}}_{J::}(\bar{t})\big)(\bar{t} - \bar{u}) \;\big|\; \bar{t} \in S\big\},$$

where C and C' were defined in (33) and $\widetilde{\mathcal{B}}$ was defined in (35). Then we compute a smaller inclusion region by the iteration

$$R_1^i := R^i|_{x_I = x_I^b}, \quad \text{and} \quad R_{k+1}^i := K(R_k^i, \bar{u}_k) \cap R_k^i, \tag{58}$$

where $\bar{u}_k \in R^i$. Let R_0^i be the approximate limit set of this iteration. It is usually a really tiny set, whose width is determined by rounding errors only.

Clearly, $\mathrm{int}(R^e)$ contains a unique minimum iff $R_0^i \times \{x_I^b\}$ contains at most one minimum. Thus, it suffices to have a condition under which a tiny region contains at most one local minimum. This can be done even in fairly ill-conditioned cases by the following test.

Theorem 7 *Take an approximate solution $u \in R_0^i \times \{x_I^b\}$ of (1), where R_0^i is computed by the iteration (58) from an inclusion region as provided by Theorem 5. Let the hypernorms ν_1, $\nu_{2,1}$, and $\nu_{3,1}$ be as in Section 5 and define*

$$C_U(\bar{t},\bar{u}) := \begin{pmatrix} (\nabla_J L|_{x_I=x_I^b})[\bar{u},\bar{t}] & \nabla_J F(z)^T & \nabla_J f(z)^T \\ (F|_{x_I=x_I^b})[z_J,x_J] & 0 & 0 \\ 0 & 2y^T & 2\sigma \end{pmatrix},$$

$$\widehat{\mathcal{B}}(\bar{t},\bar{u}) := \begin{pmatrix} (\nabla_J L|_{x_I=x_I^b})[\bar{u},\bar{t},\bar{t}] & (\nabla_J F|_{x_I=x_I^b})[z_J,x_J]^T & (\nabla_J f|_{x_I=x_I^b})[z_J,x_J]^T \\ (F|_{x_I=x_I^b})[z_J,x_J,x_J] & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

If there exists a matrix $B \in \mathbb{R}^{r\times r}$ with $\|B\| < 1$ for some monotone matrix norm such that

$$\nu_{2,1}\big(CC_U(\bar{t},\bar{u}) - I\big) + \nu_1(\bar{t}-\bar{u})^T \nu_{3,1}\big(C\cdot\widehat{\mathcal{B}}(\bar{t},\bar{u})\big) \le B, \tag{59}$$

for all $\bar{t} \in R_0^i$ then $R_0^i \times \{x_I^b\}$ contains at most one solution (x^*, w^*, κ^*) of (1).

Proof We compute $\widehat{G}[t,s] = C_U(t,s)$ and $\widehat{G}[t,s,s] = \widehat{\mathcal{B}}(t,s)$; thus the result follows from Theorem 6. □

Since B is nonnegative, $\|B\| < 1$ holds for some norm iff the spectral radius of B is less than one (see, e.g., (Neumaier, 1990, 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 is larger. (Essentially, this amounts to testing $I - B$ for being an H-matrix; cf. (Neumaier, 1990, Proposition 3.2.3).) A candidate matrix B can be efficiently calculated by interval analysis.

How much ill-condition can be tolerated by the test in Theorem 7 depends primarily on the size of the inclusion region R_0^i. If this set is small enough then $C_U(\overline{t},\overline{u})$ is a very thin interval matrix, and C is an approximate midpoint inverse of C_U, making the first matrix in (59) close to zero. The second matrix is multiplied by $\nu_1(\overline{t}-\overline{u})$ which is also close to zero. The size of the components of $\widehat{B}$ is mostly determined by the curvature of the Lagrangian. Their size has to be significantly smaller than $1/\,(\mathit{width})(R_0^i)$ to make the test work.

7 Examples

We illustrate the theory with a few low-dimensional examples. Some preliminary test calculations on the COCONUT test set Shcherbina et al (2003) have shown evidence that the difference of the radii of the exclusion box computed by Theorem 5 and the one computed by a backboxing scheme based on the

interval Newton method increases with increasing dimension. In a branch and bound context this leads to a higher portion of the search space that can be removed by a number of branching steps which is linear in the dimension. Since the cluster effect is one reason why a branch and bound method needs exponential effort for some problems, a large enough exclusion box that eliminates the cluster effect can cause hitherto intractable problems to become tractable, see Example 4.

Example 1 In this example we will do all calculations symbolically, hence free of rounding errors, assuming a known zero. (This idealizes the practically relevant case where a good approximation of a local minimum is available from a standard optimization algorithm.)

We consider the bound constrained optimization problem

$$\begin{aligned} &\min\ \tfrac{1}{3}x_1^3 + x_1x_2^2 - 25x_1 - 24x_2 \\ &\text{s.t. } x_i \in [-10, 10], \end{aligned}$$

which has 3 local solutions $(-10,-10)$, $(-10,10)$, and $(4,3)$.

We start with the solution $x^* = (4,3)$. We have no active constraints, so we choose $I = \emptyset$. Hence, we can use Corollary 3. We find

$$\nabla f(x) = \begin{pmatrix} x_1^2 + x_2^2 - 25 \\ 2x_1x_2 - 24 \end{pmatrix}.$$

With respect to the solution $x^* = \binom{4}{3}$, we have

$$\begin{aligned} \nabla f(x) - \nabla f(x^*) &= \begin{pmatrix} x_1^2 - 4^2 + x_2^2 - 3^2 \\ 2x_1x_2 - 2\cdot 4\cdot 3 \end{pmatrix} \\ &= \begin{pmatrix} (x_1+4)(x_1-4) + (x_2+3)(x_2-3) \\ 2x_2(x_1-4) + 2\cdot 4(x_2-3) \end{pmatrix}, \end{aligned}$$

so that we can take

$$(\nabla f)[x^*, x] = \begin{pmatrix} x_1 + 4 & x_2 + 3 \\ 2x_2 & 8 \end{pmatrix}.$$

This has the form (6) with

$$\nabla^2 f(x^*) = \begin{pmatrix} 8 & 6 \\ 6 & 8 \end{pmatrix}, \qquad (\nabla f)[x^*, x^*, x] = \left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \middle| \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix} \right),$$

and we obtain

$$\mathcal{B} = \frac{1}{28} \left(\begin{pmatrix} 8 & 0 \\ 6 & 0 \end{pmatrix} \middle| \begin{pmatrix} 12 & 8 \\ 16 & 6 \end{pmatrix} \right).$$

Since we calculate without rounding errors and we have a true zero, both B_0 and $\overline{b}$ vanish. From (39) we get

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

$$a_1 = \tfrac{1}{28}(8v_1^2 + 12v_1v_2 + 8v_2^2), \quad a_2 = \tfrac{1}{28}(6v_1^2 + 16v_1v_2 + 6v_2^2),$$

and for the particular choice $v = \binom{1}{1}$, we get from (39)

$$\lambda^i = 0, \qquad \lambda^e = 1. \tag{60}$$

Thus, Corollary 1 implies that the interior of the box

$$[x^* - v, x^* + v] = B_{1,|\,|}(4,3) = \begin{pmatrix}[3,5]\\[2,4]\end{pmatrix}$$

contains no solution apart form $\binom{4}{3}$. This is best possible, since there is another Kuhn-Tucker point $\binom{3}{4}$ 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.

Next we consider the optimizer $z = (-10, -10)$. In this situation, we need to apply Corollary 3. We have $J = \emptyset$, and so by definition $\lambda^e = \infty$ while $\lambda^i = 0$. We calculate

$$\delta = \nabla f(z) = \begin{pmatrix}175\\176\end{pmatrix}, \quad (\nabla f)[z,x] = \begin{pmatrix}x_1 - 10 & x_2 - 10\\ 2x_2 & -20\end{pmatrix},$$

since $J = \emptyset$ also Y is empty, and

$$Z^L(x,z) = \begin{pmatrix}x_1 - 10 & 2x_2\\ x_2 - 10 & -20\end{pmatrix},$$

if $x_2 \le 0$ and $x_1 \le 10$. Hence,

$$Z = \begin{pmatrix}20 & 20\\ 20 & 20\end{pmatrix}.$$

Choosing $v = (1,1)$ we get

$$\mu_1 = \tfrac{175}{40}, \quad \mu_2 = \tfrac{176}{40}, \quad \mu^e = \mu_1 = \tfrac{35}{8},$$

and therefore the exclusion region is

$$\mathrm{int}B_{\frac{35}{8},|\,|}(-10,10) \cap \mathbf{x} = \begin{pmatrix}[-10, -5.625\,[\\ [-10, -5.625\,[\end{pmatrix},$$

not of perfect size but still reasonably big.

Example 2 Consider the optimization problem

$$\begin{aligned} &\min\ (x_1+2)^2 + (x_2-2)^2\\ &\text{s.t. } 4x_1 + x_1^2 - x_2^3 + 2x_2^2 = -1\\ &\qquad x_1 \in [-5,5], \quad x_2 \in [-2,1] \end{aligned} \tag{61}$$

which has 3 local solutions $(-2,-1)$, $(-2-\sqrt{2},1)$, $(-2+\sqrt{2},1)$, see Figure 1.

If we calculate without rounding errors and start with the solution $z = x^* = (-2-\sqrt{2}, 1)$, we get $y = -\frac{\sqrt{2}}{2}$, $\sigma = \frac{\sqrt{2}}{2}$, and $\nabla L(z, y, \sigma) = (0, -\frac{3\sqrt{2}}{2})$, so $J = \{1\}$ and $I = \{2\}$. Then we compute

$$C' = \begin{pmatrix} 0 & -2\sqrt{2} & -2\sqrt{2} \\ -2\sqrt{2} & 0 & 0 \\ 0 & -\sqrt{2} & \sqrt{2} \end{pmatrix}, \quad C = \begin{pmatrix} 0 & -\frac{\sqrt{2}}{4} & 0 \\ -\frac{\sqrt{2}}{8} & 0 & -\frac{\sqrt{2}}{4} \\ -\frac{\sqrt{2}}{8} & 0 & \frac{\sqrt{2}}{4} \end{pmatrix}.$$

Since we calculate without rounding errors, the terms $\overline{b}$ and B_0 both vanish. For the third order tensors we compute

$$\begin{aligned}
\widetilde{\mathcal{B}}_{J::}(t,u) &:= C \cdot \begin{pmatrix} 0\,0\,0 \\ 1\,0\,0 \\ 0\,0\,0 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\sqrt{2}} & 0\,0 \\ 0 & 0\,0 \\ 0 & 0\,0 \end{pmatrix}, & \widetilde{\mathcal{T}}_{J::}(t,u) &:= C \cdot 0 = 0, \\
\widetilde{\mathcal{B}}_{W::}(t,u) &:= C \cdot \begin{pmatrix} 2\,0\,0 \\ 0\,0\,0 \\ 0\,1\,0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 \\ -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & 0 \end{pmatrix}, & \widetilde{\mathcal{T}}_{W::}(t,u) &:= C \cdot 0 = 0, \\
\widetilde{\mathcal{B}}_{\kappa::}(t,u) &:= C \cdot \begin{pmatrix} 2\,0\,0 \\ 0\,0\,0 \\ 0\,0\,1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ -\frac{1}{2\sqrt{2}} & 0 & -\frac{1}{2\sqrt{2}} \\ -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} \end{pmatrix}, & \widetilde{\mathcal{T}}_{::\kappa}(t,u) &:= C \cdot 0 = 0, \\
\widetilde{\mathcal{T}}_{I::}(t,u) &:= C \cdot \begin{pmatrix} 0 \\ -x_2 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{x_2}{2\sqrt{2}} \\ 0 \\ 0 \end{pmatrix}, & \widetilde{\mathcal{B}}_{I::}(t,u) &:= C \cdot 0 = 0.
\end{aligned} \tag{62}$$

As hypernorms we choose the componentwise absolute value. Thus, the only term which is not straightforward is in $\widetilde{\mathcal{T}}$ where x_2 will be estimated as $|\mathbf{x}_2| = 2$. We choose $v = e$ and calculate

$$w = v = e, \quad a = \begin{pmatrix} \frac{3}{2\sqrt{2}} \\ \sqrt{2} \\ \sqrt{2} \end{pmatrix}, \quad D = e, \quad \lambda^e_{\cdot} = \begin{pmatrix} \frac{2\sqrt{2}}{3} \\ \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}, \quad \lambda^e = \frac{\sqrt{2}}{2}, \quad \lambda^i = 0.$$

For the border terms we find

$$\delta = -\frac{3\sqrt{2}}{2}, \quad C_B(t,u) = (0,1,-1), \quad Z^L_{22}(t,u) = \begin{cases} \frac{1+3x_2}{\sqrt{2}} & x_2 \le -\frac{1}{3} \\ 0 & \text{otherwise} \end{cases}$$

$$Y = (0,1,1), \quad Z = 0, \quad x_2 \ge -\tfrac{1}{3}$$

$$\mu_2 = -\frac{3\sqrt{2}}{4}, \quad \mu^e = \tfrac{\sqrt{2}}{2}.$$

Theorem 5 tells us therefore that there is no solution in the interior of the box

$$R^e = [-2 - \tfrac{3\sqrt{2}}{2}, -2 - \tfrac{\sqrt{2}}{2}] \times [1 - \tfrac{\sqrt{2}}{2}, 1]$$

except for x^*. The box again is of considerable size, although it is not optimal.

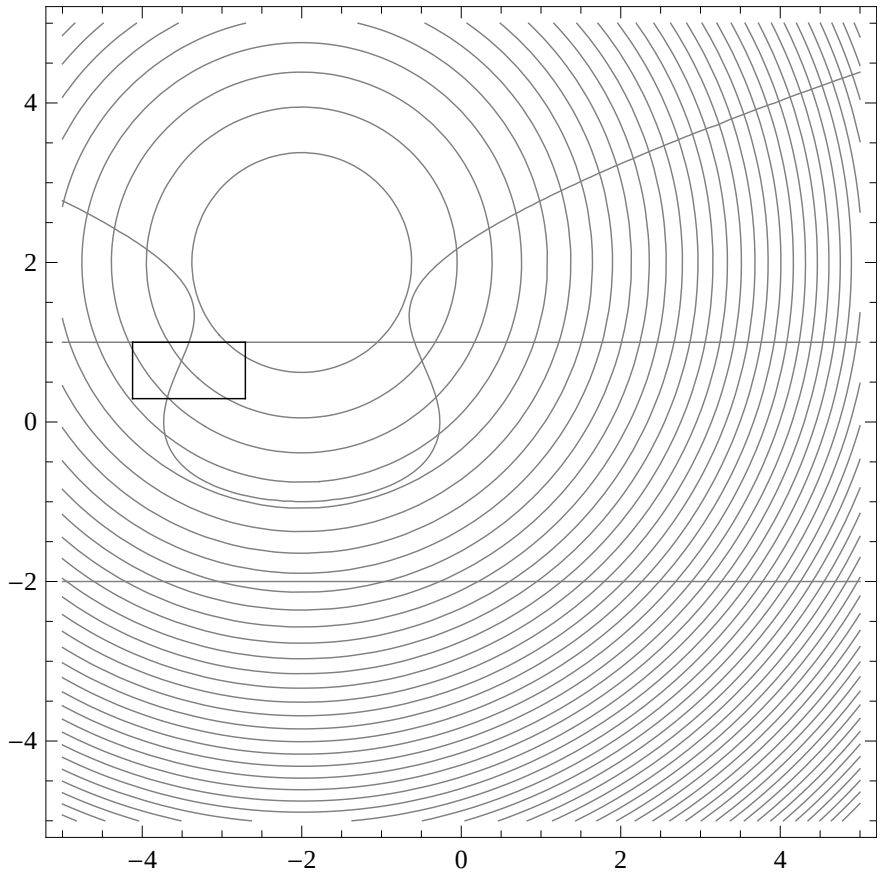

Fig. 1 Exclusion Box for Problem (61).

Example 3 The system of equations (2) with

$$G(x) = \begin{pmatrix} x_1^2 + x_1x_2 + 2x_2^2 - x_1 - x_2 - 2 \\ 2x_1^2 + x_1x_2 + 3x_2^2 - x_1 - x_2 - 4 \end{pmatrix} \tag{63}$$

has the solutions $\binom{1}{1}$, $\binom{1}{-1}$, $\binom{-1}{1}$, cf. Figure 2. We consider the first solution $x^* = \binom{1}{1}$. We can easily compute that

$$\begin{aligned} G[z,x] &= \begin{pmatrix} x_1 + x_2 + z_1 - 1 & 2x_2 + z_1 + 2z_2 - 1 \\ 2x_1 + x_2 + 2z_1 - 1 & 3x_2 + z_1 + 3z_2 - 1 \end{pmatrix}, \\ G'(z) &= \begin{pmatrix} 2z_1 + z_2 - 1 & z_1 + 4z_2 - 1 \\ 4z_1 + z_2 - 1 & z_1 + 6z_2 - 1 \end{pmatrix}, \\ G[z,z,x] &= \left(\begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix} \middle| \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} \right). \end{aligned}$$

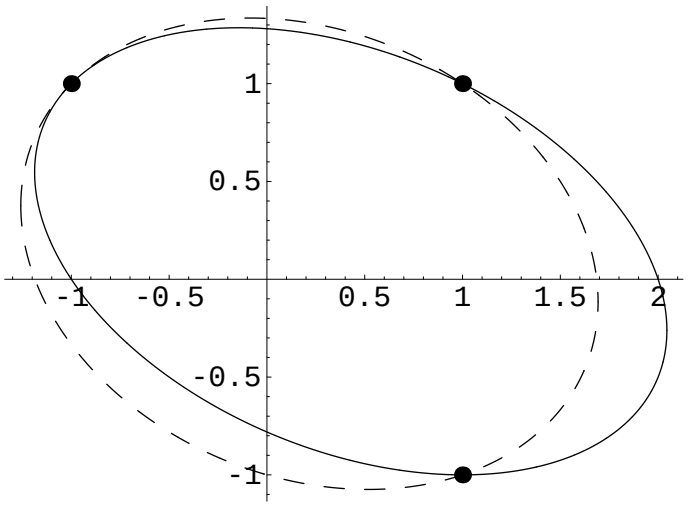

Fig. 2 Two quadratic equations in two variables.

For

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

we get $CG(x^*) = 0$ and $CG'(x^*) - I = 0$, since we calculate without rounding errors. Then

$$C \cdot G[x^*, x^*, x] = \left(\begin{pmatrix} 0.5 & 0 \\ 0 & 0 \end{pmatrix} \middle| \begin{pmatrix} -0.5 & 0 \\ 0.5 & 0.5 \end{pmatrix} \right)$$

We choose the 2–norm as hypernorms ν_0 and ν_1, then compatible hypernorms are the matrix 2–norm, i.e. the maximal singular value, and the 3–tensor 2–norm for the second index, i.e. the maximal mode-2 singular value. Computing the HOSVD of $C \cdot G[x^*, x^*, x]$ (see De Lathauwer et al (2000)) we get the singular value tensor

$$\mathcal{S} = \left(\begin{pmatrix} -0.8536 & 0 \\ 0 & 0.3536 \end{pmatrix} \middle| \begin{pmatrix} 0 & -0.3536 \\ 0.1464 & 0 \end{pmatrix} \right)$$

and the norm $\|C \cdot G[x^*, x^*, x]\|_{2,2} = \|\mathcal{S}\|_{2,2} = 0.866$. We use Corollary 1 to compute $\delta = 1$, $\lambda^i = 0$, and $\lambda^e = 1/0.866 = 1.1547$. The exclusion region R^e is therefore the circle with center $(1, 1)$ and radius 1.1547.

The box exclusion region for this example was already calculated in Schichl and Neumaier (2005a) as $[0, 2] \times [0, 2]$. Its radius and its volume as compared to the exclusion circle are slightly smaller.

There it was already shown that the other known methods for computing uniqueness areas perform worse by at least an order of magnitude.

Example 4 This example is from Kieffer et al (2011), where the authors considered the nonlinear parameter estimation problem

$$\begin{aligned}\min f(x) &= \sum_{i=1}^{N} (y_m(x, t_i) - y_i)^2 \\ \text{s.t. } & x \in [0.01, 1]^3,\end{aligned}$$

where the y_i are constants values, y_m is specified as

$$y_m(x, t_i) = \left(x_1 e^{-x_1 t_i/2}\right)\left(e^{-x_2 t_i/2}\right)\left(e^{-x_3 t_i/2}\right)\left(-\frac{2}{\sqrt{a(x)}}\sinh(\sqrt{a(x)}t_i/2)\right),$$

with

$$a(x) = (x_3 - x_2 + x_1)^2 + 4x_1x_2,$$

for all $i = 1, \ldots, N$, and we set $N := 15$ and $t_i := i$. We solved the above bound constrained problem with an interval branch and bound method using Corollary 3 for the exclusion/inclusion box techniques.

Backboxing techniques using interval Newton operator were not able to find an inclusion/exclusion pair. They could be used to generate some exclusion boxes which did not contain a solution.

The best approximate solution to the problem found by the local optimizer IPOPT Waechter et al (2009) was

$$\begin{aligned}z = &[0.6049537177358995, 0.6049537177358995], \\ &[0.1444752644582018, 0.1444752644582018], \\ &[0.3660122310991678, 0.3660122310991678]\end{aligned}$$

with a function value enclosure of

$$f(z) \in [6.721779280230305, 6.721779280254950] \cdot 10^{-5}.$$

The exclusion and inclusion boxes computed using the result of Corollary 3 were

$$\begin{aligned}R^e = ([&0.6048367258681456, 0.6050707096036533], \\ &[0.1443582725904481, 0.1445922563259555], \\ &[0.3658952392314140, 0.3661292229669215]),\end{aligned}$$

and

$$\begin{aligned}R^i = ([&0.6049444542702035, 0.6049629812015954], \\ &[0.1444660009925059, 0.1444845279238977], \\ &[0.3660029676334718, 0.3660214945648637]).\end{aligned}$$

The inclusion box could be further improved by a third order Krawczyk–like iteration to

$$\begin{aligned}([&0.604961728242, 0.604961728246], \\ &[0.144474180373, 0.144474180376], \\ &[0.366021184203, 0.366021184210])\end{aligned}$$

with an enclosure of the global optimum of

$$[6.72177710824, 6.72177710827] \cdot 10^{-5},$$

close to the maximal precision possible with standard double-precision floating-point computations.

References

De Lathauwer L, De Moor B, Vandewalle J, et al (2000) A multilinear singular value decomposition. SIAM Journal on Matrix Analysis and Applications 21(4):1253–1278

Du K, Kearfott R (1994) The cluster problem in multivariate global optimization. Journal of Global Optimization 5(3):253–265

Fischer H (1974) Hypernormbälle als abstrakte Schrankenzahlen. Computing 12(1):67–73

Hansen E (1978) Interval forms of Newtons method. Computing 20:153–163, URL `http://dx.doi.org/10.1007/BF02252344`, 10.1007/BF02252344

John F (1948) Extremum problems with inequalities as subsidiary conditions. In: Studies and Essays presented to R. Courant on his 60th Birthday, pp 187–204

Kahan W (1968) A more complete interval arithmetic. Lecture notes for an engineering summer course in numerical analysis

Karush W (1939) Minima of functions of several variables with inequalities as side constraints. Master's thesis, Dept. of Mathematics, Univ. of Chicago

Kearfott R (1996a) A review of techniques in the verified solution of constrained global optimization problems, Kluwer, Dordrecht, pp 23–60

Kearfott R (1996b) Rigorous global search: continuous problems. Kluwer, Dordrecht

Kearfott R (1997) Empirical evaluation of innovations in interval branch and bound algorithms for nonlinear systems. SIAM Journal on Scientific Computing 18:574–594

Kearfott R, Muniswamy S, Wang Y, Li X, Wang Q (2012) On smooth reformulations and direct non-smooth computations in global optimization for minimax problems. Journal of Global Optimization To appear

Kearfott RB (1987) Abstract generalized bisection and a cost bound. Mathematics of Computation 49(179):187–202

Kieffer M, Markót MC, Schichl H, Walter E (2011) Verified global optimization for estimating the parameters of nonlinear models. In: Modeling, Design, and Simulation of Systems with Uncertainties, Springer, Berlin, pp 129–151

Kolev L (1997) Use of interval slopes for the irrational part of factorable functions. Reliable Computing 3(1):83–93

Krawczyk R, Neumaier A (1985) Interval slopes for rational functions and associated centered forms. SIAM Journal on Numerical Analysis 22(3):604–616

Mayer G (1995) Epsilon-inflation in verification algorithms. Journal of Computational and Applied Mathematics 60(1):147–169

Neumaier A (1990) Interval methods for systems of equations. Cambridge University Press, Cambridge

Neumaier A (2001) Introduction to numerical analysis. Cambridge University Press, Cambridge

Ortega J, Rheinboldt W (2000) Iterative solution of nonlinear equations in several variables. Society for Industrial Mathematics (SIAM)

Rohn J (1989) Systems of linear interval equations. Linear algebra and its applications 126:39–78

Rump S (1996) Expansion and estimation of the range of nonlinear functions. Mathematics of Computation 65(216):1503–1512

Rump S (1999) INTLAB – INTerval LABoratory, Kluwer, Dordrecht, pp 77–104. URL `http://www.ti3.tu-harburg.de/rump/intlab/index.html`

Rump SM (1998) A note on epsilon-inflation. Reliable Computing 4(4):371–375

Schichl H, Markót MC (2010) Interval analysis on directed acyclic graphs for global optimization. Higher order methods. URL `http://www.mat.univie.ac.at/~herman/papers/dag2.pdf`, manuscript

Schichl H, Markót MC (2012) Algorithmic differentiation techniques for global optimization in the COCONUT environment. Optimization Methods and Software 27(2):359–372

Schichl H, Markót MCea (2013) The COCONUT Environment. software. URL `www.mat.univie.ac.at/coconut-environment`

Schichl H, Neumaier A (2005a) Exclusion regions for systems of equations. SIAM Journal on Numerical Analysis 42(1):383–408

Schichl H, Neumaier A (2005b) Interval analysis on directed acyclic graphs for global optimization. Journal of Global Optimization 33(4):541–562

Schichl H, Neumaier A (2011) Transposition theorems and hypernorms. URL `http://www.mat.univie.ac.at/~herman/trans2.pdf`, manuscript

Shcherbina O, Neumaier A, Sam-Haroud D, Vu XH, Nguyen TV (2003) Benchmarking global optimization and constraint satisfaction codes. In: et al CB (ed) Global Optimization and Constraint Satisfaction, Springer, Berlin, pp 211–222, URL `http://www.mat.univie.ac.at/~neum/papers.html\#bench`

Van Hentenryck P, Michel L, Deville Y (1997) Numerica: a modeling language for global optimization. MIT press

Van Iwaarden R (1996) An improved unconstrained global optimization algorithm. PhD thesis, University of Colorado at Denver

Waechter A, Laird C, Margot F, Kawajir Y (2009) Introduction to ipopt: A tutorial for downloading, installing, and using ipopt