Stopping rules and backward error analysis for bound-constrained optimization

Gratton, Serge; Mouffe, Mélodie; Toint, Philippe

Published in:
Numerische Mathematik

DOI:
10.1007/s00211-011-0376-1

Publication date:
2011

Document Version
Early version, also known as pre-print

Stopping rules and backward error analysis for bound-constrained optimization

by S. Gratton[1], M. Mouffe[2] and Ph. L. Toint[3]

14 March 2011

[1] ENSEEIHT-IRIT,
2, rue Camichel, 31000 Toulouse, France
Email: serge.gratton@enseeiht.fr

[2] CERFACS,
42, av. G. Coriolis, 31057 Toulouse, France.
Email: mouffe@cerfacs.fr

[3] Department of Mathematics,
FUNDP-University of Namur,
61, rue de Bruxelles, B-5000 Namur, Belgium.
Email: philippe.toint@fundp.ac.be

Stopping rules and backward error analysis for bound-constrained optimization

Serge Gratton Mélodie Mouffe Philippe L. Toint

14 March 2011

Abstract

Termination criteria for the iterative solution of bound-constrained optimization problems are examined in the light of backward error analysis. It is shown that the problem of determining a suitable perturbation on the problem's data corresponding to the definition of the backward error is analytically solvable under mild assumptions. Moreover, a link between existing termination criteria and this solution is clarified, indicating that some standard measures of criticality may be interpreted in the sense of backward error analysis. The backward error problem is finally considered from the multicriteria optimization point of view and some numerical illustration is provided.

Keywords : Nonlinear optimization, bound constraints, stopping criterion, backward error, multicriteria optimization.

1 Introduction

The definition of many bound-constrained optimization problems contains uncertainties or errors in the associated data, for example when they arise from the discretization error of an underlying continuous problem (Dolan, Moré and Munson [13], Averick and Moré [3]) or because they contain data obtained by actual physical measurements (Fisher [16]). It is then natural to seek a solution of the problem whose accuracy is of the order of (or slightly better than) the level of those uncertainties. If iterative algorithms are used, this translates into the sometimes difficult selection of a suitable termination rule. This is especially problematic when solving industrial applications for which one evaluation of the objective function can be really expensive, which happens typically once per iteration. In general, defining a good stopping criterion corresponds to finding a reasonable balance between robustness and oversolving: one seeks to obtain an accurate solution but also to avoid performing many additional computations for little gain. Moreover, good stopping criteria should have a meaning that is easy to understand for the user.

A wide range of stopping criteria for bound-constrained optimization algorithms is already available in the litterature, if one is ready to ignore the noise in the data caused by the uncertainties and/or errors. They typically consist in requiring a certain optimality (or criticality) measure to fall below a user-specified tolerance. The most commonly used such

measure is the norm of the projection of the negative gradient on the feasible set (see Byrd, Lu, Nocedal and Zhu [7], Hager and Zhang [19] and Xu and Burke [25]). Some trust-region algorithms (Conn, Gould, Sartenaer and Toint [10], Conn, Gould and Toint [11]) use an alternative measure which approximates the maximal linear decrease that can be achieved in the neighbourhood of unit radius. The reduced gradient (that is the gradient where all its components which are pointing in the direction of an already active bound are set to zero) is also used as an optimality measure (for example in Burke ans Moré [5], Calamai and Moré [8], Burke, Moré and Toraldo [6], Burke [4] or Dostal [14]). However, it is usually not entirely obvious how to adapt these approaches to the case where the problem is contaminated by noise. See for instance Moré and Wild [21], for an interesting discussion in a derivative free optimization context.

The purpose of this paper is to present a new approach for defining easily interpretable stopping criteria which take advantage of known uncertainties in the problems' data, with the double objective of ensuring robustness and avoiding unnecessary computations as soon as the solution error becomes smaller than these uncertainties. Our approach is based on the well-known linear-algebraic concept of backward error, a concept which is widely used to define stopping criteria in the solution of linear systems of equations, has been extensively studied in this framework (see Rigal and Gaches [23], Cox and Higham [12], Golub and van Loan [18], Chatelin and Frayssé [9] or Higham [20]) and has already been extended to the solution of nonlinear equations (see Arioli, Duff and Ruiz [1]). The introduction of a backward error estimate in the solution of bound-constrained nonlinear optimization will provide, at each step of the algorithm, a measure of the perturbation of the original problem necessary to define a problem instance of which the incumbent iterate is an exact solution. This then allows a meaningful comparison of this perturbation size with the data uncertainties and suggests an efficient termination of the solution algorithm when the former becomes smaller than the latter.

The paper is organized as follows. In Section 2, we introduce the backward error concept and apply it to our bound-constrained optimization problem. The link between the backward error and several well-known criticality measures is studied in Section 3 and a multicriteria analysis of the backward error problem is presented in Section 4. Finally, the numerical behavior of some interesting criticality measures is illustrated in Section 5 and conclusions discussed in Section 6.

2 Backward error analysis

2.1 Backward error analysis for bound-constrained optimization

We are interested in solving the minimization problem

$$\min_{\mathcal{F}} f(x), \tag{1}$$

where $f(.)$ is a possibly nonlinear objective function and where $\mathcal{F} = \{x \in \Re^n \mid l \leq x \leq u\}$ is a set of bound constraints with $l, u \in \Re^n$. In practice, we are looking for a *first-order critical point* of (1), that is a feasible point x_* where $[\nabla_x f(x_*)]_j = 0$ for all $j \notin \mathcal{A}(x_*)$,

where we denote by $[v]_j$ the j^{th} component of a vector v and where we define the *active set of binding constraints* at $x \in \mathcal{F}$ by $\mathcal{A}(x) = \mathcal{A}^-(x) \cup \mathcal{A}^+(x)$ with

$$\begin{array}{rcl} \mathcal{A}^-(x) = \{j \in \{1,...,n\} \mid [x]_j & = & [l]_j \quad \text{and} \quad [\nabla_x f(x)]_j > 0\}, \\ \mathcal{A}^+(x) = \{j \in \{1,...,n\} \mid [x]_j & = & [u]_j \quad \text{and} \quad [\nabla_x f(x)]_j < 0\}. \end{array}$$

We consider iterative optimization methods which produce a sequence of iterates x_k which converge to a first-order solution x_* of the problem to solve. Our objective is to terminate this sequence as early as possible, especially for large-scale or otherwise expensive problems, in order to achieve a reasonable reliability of the approximate solution while avoiding unnecessary costly iterations. An obvious way of expressing this problem is to stop the iterations when the current iterate x_k is such that

$$||x_k - x_*|| < \epsilon,$$

where ϵ is an acceptable tolerance on the distance between the approximate and the first-order solution and where $||\cdot||$ is a norm making sense for the application considered. But, unless very particular situations are considered such as the testing phase of an optimization algorithm, x_* is not known, and suitable choices for ϵ and $||\cdot||$ are often subjective, making the above test impractical and the exploitation of any knowledge of the uncertainty on the problem data difficult. Our proposal is therefore to adopt the backward error point of view, as has been proposed for linear algebra by Givens [17] and Wilkinson [24]. The idea is to replace the question *How far from the solution is the current approximation x_k?* by *If there exists a minimization problem (P) whose x_k is a first-order solution, how far from the original problem (1) is (P)?* We may then consider terminating the iterative solution algorithm as soon as this latter distance is smaller than the known error (e.g. the discretization error). Notice that we assume the order of magnitude of the error is known a priori. However, in the case where the error is only estimated a posteriori, we have to solve the problem with a very demanding accuracy in order to avoid interferences with the estimation of the error. To make this backward error approach for bound-constrained optimization problem more formal, we consider, for any guess $\tilde{x}$, a perturbed version of the original problem (1) defined by

$$\min_{\mathcal{F}_\Delta} f(x) + \Delta f + \Delta g^T x, \tag{2}$$

with $\mathcal{F}_\Delta = \{x \in \Re^n \mid l + \Delta l \leq x \leq u + \Delta u\}$ and where the perturbations $\Delta f, \Delta g, \Delta l, \Delta u$ are chosen such that $\tilde{x}$ is an exact first-order critical point of (2). The first-order sufficient condition for optimality then implies that $\Delta f, \Delta g, \Delta l$ and Δu satisfy $[\nabla_x f(\tilde{x}) + \Delta g]_j = 0$ for all $j \notin \mathcal{A}_\Delta(\tilde{x})$, where $\mathcal{A}_\Delta(x) = \mathcal{A}^-_\Delta(x) \cup \mathcal{A}^+_\Delta(x)$ is the *perturbed set of binding constraints*, with

$$\begin{array}{rcl} \mathcal{A}^-_\Delta(x) = \{j \in \{1,...,n\} \mid [x]_j & = & [l]_j + [\Delta l]_j \quad \text{and} \quad [\nabla_x f(x) + \Delta g]_j > 0\}, \\ \mathcal{A}^+_\Delta(x) = \{j \in \{1,...,n\} \mid [x]_j & = & [u]_j + [\Delta u]_j \quad \text{and} \quad [\nabla_x f(x) + \Delta g]_j < 0\}. \end{array}$$

Since the value of Δf does not appear in this sufficient condition, we can set $\Delta f = 0$ in (2) without loss of generality, which we do from now on. We now define the backward error

as the minimum of some product norm of the remaining perturbations $\Delta g, \Delta l, \Delta u$. We are then led to define

$$\mathcal{D} \stackrel{\rm def}{=} \{(\Delta g, \Delta l, \Delta u) \in \Re^{3n} \mid \tilde{x} \in \mathcal{F}_\Delta \text{ and } [\nabla_x f(\tilde{x}) + \Delta g]_j = 0 \text{ for all } j \notin \mathcal{A}_\Delta(\tilde{x})\},$$

and to propose terminating the algorithm as soon as

$$\inf_{(\Delta g, \Delta l, \Delta u) \in \mathcal{D}} \| \ (\Delta g, \Delta l, \Delta u) \ \| < \epsilon(\epsilon_g, \epsilon_l, \epsilon_u),$$

where the perturbations Δg, Δl, Δu are respectively measured with the norms $||.||_g, ||.||_l, ||.||_u$ and $\|.\|$ is a product norm of these three norms. The thresholds $\epsilon_g, \epsilon_l, \epsilon_u \in \Re$ represent the known order of magnitude of the error on g, l and u as mesured with the corresponding norm. Notice that $\mathcal{D}$ is always non-empty as it always contains $(-\nabla_x f(\tilde{x}), \tilde{x} - l, \tilde{x} - u)$. Moreover, the infimum may actually be replaced by a minimum, because $\mathcal{D}$ is the union of a finite number of direct products between closed sets (see Mouffe [22]) and is thus itself a closed set and the minimization can be restricted to bounded perturbations $(\Delta g, \Delta l, \Delta u)$ such that $\|(\Delta g, \Delta l, \Delta u)\| \leq \|(-\nabla_x f(\tilde{x}), \tilde{x} - l, \tilde{x} - u)\|$. Thus our proposal is to terminate the algorithm at the first iteration k such that

$$\min_{(\Delta g, \Delta l, \Delta u) \in \mathcal{D}} \| \ (\Delta g, \Delta l, \Delta u) \ \| < \epsilon(\epsilon_g, \epsilon_l, \epsilon_u), \tag{3}$$

where $\min_{(\Delta g, \Delta l, \Delta u) \in \mathcal{D}} \| \ (\Delta g, \Delta l, \Delta u) \ \|$ is the backward error for $\tilde{x} = x_k$, x_k being the current iterate.

2.2 Solving the backward error problem

We now wish to investigate how the value of the minimum on the left-hand-side of (3) can be computed in practice for specific choices of the product norm. We start by considering the *weighted sum measure*

$$\chi_{ws} \stackrel{\rm def}{=} \min_{(\Delta g, \Delta l, \Delta u) \in \mathcal{D}} (\alpha_g ||\Delta g||_g + \alpha_l ||\Delta l||_l + \alpha_u ||\Delta u||_u), \tag{4}$$

and the *absolute measure*

$$\chi_{abs} \stackrel{\rm def}{=} \min_{(\Delta g, \Delta l, \Delta u) \in \mathcal{D}} ||\alpha_g |\Delta g| + \alpha_l |\Delta l| + \alpha_u |\Delta u| \ ||_{glu}, \tag{5}$$

where $(\alpha_g, \alpha_l, \alpha_u) > 0$, where $|.|$ denotes the componentwise absolute value, and where $||.||_g, ||.||_l, ||.||_u$ and $||.||_{glu}$ are *monotone* norms, in the sense that each of these norms satisfies the following properties

$$\text{if } \forall j \in \{1, ..., n\} |[u]_j| \geq |[v]_j|, \text{ then } ||u|| \geq ||v|| \quad \forall u, v \in \Re^n.$$

Notice that the product norms defined by (4) and (5) satisfy all the norm properties as long as α_g, α_l and α_u are positive. Notice that, in particular, all the p-norms, $1 \leq p \leq \infty$, are monotone norms. Moreover, the choice left for $||.||_g$, $||.||_l$ and $||.||_u$ in the definition (4) of

χ_{ws} opens the possibility of defining, for instance, $||.||_g$ as the *dual norm* of $||.||_l = ||.||_u$ (on the obvious condition that $||.||_g, ||.||_l, ||.||_u$ are all monotone). Unfortunately, the energy-norm (or A-norm) defined by $||v||_A^2 = v^T Av$, where v is a vector of $\Re^n$ and $A \in \Re^{n\times n}$ is a symmetric positive definite matrix, is not a monotone norm.

This assumption of monotone norms is motivated by the fact that it allows a relatively easy characterization of a set $\mathcal{P} \subseteq \mathcal{D}$ containing the solution set of both problems (4) and (5). Indeed we now show, in a technical theorem, that any optimal solution $(\Delta g^*, \Delta l^*, \Delta u^*)$ of (4), as well as any optimal solution of (5), belongs to a finite set $\mathcal{P} \subseteq \mathcal{D}$ explicitly described as the cartesian product between n subsets of $\Re^3$, each of them containing at most two elements.

Theorem 2.1 *Suppose that $||.||_g$, $||.||_l$, $||.||_u$ and $||.||_{glu}$ are monotone norms and denote by $\mathcal{S}_{ws} \subseteq \mathcal{D}$ the set of solutions of (4) and by $\mathcal{S}_{abs} \subseteq \mathcal{D}$ the set of solutions of (5) for some arbitrary $\tilde{x}$. Let $\mathcal{V}(x) = \mathcal{V}^-(x) \cup \mathcal{V}^+(x)$, where*

$$\begin{array}{lllll} \mathcal{V}^-(x) = \{j \in \{1,...,n\} \mid & [x]_j & < & [l]_j & \text{and} \quad [\nabla_x f(x)]_j > 0\}, \\ \mathcal{V}^+(x) = \{j \in \{1,...,n\} \mid & [x]_j & > & [u]_j & \text{and} \quad [\nabla_x f(x)]_j < 0\}, \end{array}$$

be the set of violated constraints pointed by the negative gradient, and let

$$\mathcal{U} = \{j \in \{1,...,n\} \mid [\nabla_x f(\tilde{x})]_j \neq 0 \text{ and } j \notin \mathcal{A}(\tilde{x}) \text{ and } j \notin \mathcal{V}(\tilde{x})\}$$

be the set of undecided indices. *In addition, denote $\mathcal{F}_j = \{[x]_j \in \Re \mid [l]_j \leq [x]_j \leq [u]_j\}$. Then we have that $\mathcal{S}_{ws} \subseteq \mathcal{P} \subseteq \mathcal{D}$ and $\mathcal{S}_{abs} \subseteq \mathcal{P} \subseteq \mathcal{D}$, where $\mathcal{P}$ is the set of perturbations $(\Delta g, \Delta l, \Delta u) \in \mathcal{D}$ such that, for all $1 \leq j \leq n$,*

$$([\Delta g]_j\,;\,[\Delta l]_j\,;\,[\Delta u]_j) = \left\{ \begin{array}{lll} (a) & (0\,;\,0\,;\,0) & \text{if } [\tilde{x}]_j \in \mathcal{F}_j \text{ and } j \notin \mathcal{U}, \\ (b) & \text{or } ([-\nabla_x f(\tilde{x})]_j\,;\,0\,;\,0) & \text{if } [\tilde{x}]_j \in \mathcal{F}_j \text{ and } j \in \mathcal{U} \\ (c) & (0\,;\,[\tilde{x}-l]_j\,;\,0) & \text{and } [\nabla_x f(\tilde{x})]_j > 0, \\ (d) & \text{or } ([-\nabla_x f(\tilde{x})]_j\,;\,0\,;\,0) & \text{if } [\tilde{x}]_j \in \mathcal{F}_j \text{ and } j \in \mathcal{U} \\ (e) & (0\,;\,0\,;\,[\tilde{x}-u]_j) & \text{and } [\nabla_x f(\tilde{x})]_j < 0, \\ (f) & \text{or } (0\,;\,[\tilde{x}-l]_j\,;\,0) & \text{if } [\tilde{x}]_j \notin \mathcal{F}_j \text{ and } j \notin \mathcal{U}, \\ (g) & (0\,;\,0\,;\,[\tilde{x}-u]_j) & \\ (h) & \text{or } ([-\nabla_x f(\tilde{x})]_j\,;\,0\,;\,[\tilde{x}-u]_j) & \text{if } [\tilde{x}]_j \notin \mathcal{F}_j \text{ and } j \in \mathcal{U} \\ (i) & (0\,;\,[\tilde{x}-l]_j\,;\,[\tilde{x}-u]_j) & \text{and } [\nabla_x f(\tilde{x})]_j > 0, \\ (j) & \text{or } ([-\nabla_x f(\tilde{x})]_j\,;\,[\tilde{x}-l]_j\,;\,0) & \text{if } [\tilde{x}]_j \notin \mathcal{F}_j \text{ and } j \in \mathcal{U} \\ (k) & (0\,;\,[\tilde{x}-l]_j\,;\,[\tilde{x}-u]_j) & \text{and } [\nabla_x f(\tilde{x})]_j < 0. \end{array} \right. \tag{6}$$

Proof. First notice that $\mathcal{P} \subseteq \mathcal{D}$. Indeed, for all undecided indices $j \in \mathcal{U}$ such that $[\tilde{x}]_j \in \mathcal{F}_j$, either we have $[\nabla_x f(\tilde{x}) + \Delta g]_j = 0$ because of (6b) and (6d), or (6c) and (6e) imply that $j \in \mathcal{A}_\Delta(\tilde{x})$. When $[\tilde{x}]_j \notin \mathcal{F}_j$, the violated bound is perturbed in addition to

make $\tilde{x}$ feasible. We now want to prove that $\mathcal{S}_{ws} \subseteq \mathcal{P}$ and $\mathcal{S}_{abs} \subseteq \mathcal{P}$. For this purpose, we consider a perturbation vector $\widehat{\Delta} \stackrel{\rm def}{=} (\widehat{\Delta g}, \widehat{\Delta l}, \widehat{\Delta u}) \in \mathcal{D} \setminus \mathcal{P}$.

In a first step, we prove that there exists at least one $(\Delta g, \Delta l, \Delta u) \in \mathcal{P}$ such that for all j for which (6) does not hold, $([\Delta g]_j\,;\,[\Delta l]_j\,;\,[\Delta u]_j)$ satisfies

$$\begin{array}{lccccccc} & |[\widehat{\Delta g}]_j| \geq |\,[\Delta g]_j\,| & \text{and} & |[\widehat{\Delta l}]_j| \geq |\,[\Delta l]_j\,| & \text{and} & |[\widehat{\Delta u}]_j| \geq |\,[\Delta u]_j\,| & (a) \\ \text{and either} & |[\widehat{\Delta g}]_j| > |\,[\Delta g]_j\,| & \text{or} & |[\widehat{\Delta l}]_j| > |\,[\Delta l]_j\,| & \text{or} & |[\widehat{\Delta u}]_j| > |\,[\Delta u]_j\,| & (b). \end{array} \tag{7}$$

We distinguish three cases.

Suppose first that $j \notin \mathcal{U}$. If $[\tilde{x}]_j \in \mathcal{F}_j$, then equation (6a) implies that

$$[\Delta g]_j = [\Delta l]_j = [\Delta u]_j = 0,$$

and thus (7) obviously holds for any other perturbation $\widehat{\Delta} \in \mathcal{D} \setminus \mathcal{P}$. Otherwise, if $[\tilde{x}]_j < [l]_j$, the optimality condition imposes that $[\widehat{\Delta l}]_j \leq [\tilde{x} - l]_j < 0$ and thus a perturbation satisfying (6f) also ensures (7). The same reasoning applies using (6g) when $[\tilde{x}]_j > [u]_j$.

Suppose now that $j \in \mathcal{U}$ and $[\nabla_x f(\tilde{x})]_j > 0$. Because $\widehat{\Delta} \in \mathcal{D}$, we have both $\tilde{x} \in \mathcal{F}_\Delta$ and either

$$[\widehat{\Delta g}]_j = [-\nabla_x f(\tilde{x})]_j\,, \tag{8}$$

or

$$[\widehat{\Delta l}]_j = [\tilde{x} - l]_j\,, \tag{9}$$

or

$$[\widehat{\Delta u}]_j = [\tilde{x} - u]_j \text{ and } [\widehat{\Delta g}]_j < [-\nabla_x f(\tilde{x})]_j\,. \tag{10}$$

When $[\tilde{x}]_j \in \mathcal{F}_j$, a perturbation satisfying (6b) guarantees (7) for all $\widehat{\Delta}$ satisfying (10). When $[\tilde{x}]_j \notin \mathcal{F}_j$, $j \in \mathcal{U}$ and $[\nabla_x f(\tilde{x})]_j > 0$, it follows that $[\tilde{x}]_j > [u]_j$, and therefore (6h) ensures (7) for all $\widehat{\Delta}$ such that (10) holds. In addition, if $[\tilde{x}]_j \in \mathcal{F}_j$, (6b) and (6c) imply (7) for all $\widehat{\Delta}$ such that (8) holds but $[\widehat{\Delta l}]_j \neq 0$ or $[\widehat{\Delta u}]_j \neq 0$, and for all $\widehat{\Delta}$ such that (9) holds but $[\widehat{\Delta g}]_j \neq 0$ or $[\widehat{\Delta u}]_j \neq 0$, respectively. In the case where $[\tilde{x}]_j \notin \mathcal{F}_j$, we need to impose $[\widehat{\Delta u}]_j \geq [\tilde{x} - u]_j > 0$ to obtain that $\tilde{x} \in \mathcal{F}_\Delta$, and therefore (7) is ensured by a perturbation satisfying (6h) when (8) holds and satisfying (6i) when (9) holds.

Finally, a symmetric reasoning leads to (7) in the case where $j \in \mathcal{U}$ and $[\nabla_x f(\tilde{x})]_j < 0$.

We therefore conclude that, for any $\widehat{\Delta} \in \mathcal{D} \setminus \mathcal{P}$, there always exists $(\Delta g, \Delta l, \Delta u)$ $\in \mathcal{P}$ satisfying (7) for j such that (6) does not hold. In addition, notice that (7a) is actually satisfied for all $j = 1, ..., n$, as it suffices to define

$$([\Delta g]_j\,;\,[\Delta l]_j\,;\,[\Delta u]_j) = ([\widehat{\Delta g}]_j\,;\,[\widehat{\Delta l}]_j\,;\,[\widehat{\Delta u}]_j)$$

for all other j. Moreover, there necessarily exists at least one j such that (7b) holds because we have assumed $\widehat{\Delta} \in \mathcal{D} \setminus \mathcal{P}$.

Using now the monotonicity of $||.||_g$, $||.||_l$ and $||.||_u$, we deduce that

$$||\widehat{\Delta g}||_g \geq ||\Delta g||_g,\ ||\widehat{\Delta l}||_l \geq ||\Delta l||_l \text{ and } ||\widehat{\Delta u}||_u \geq ||\Delta u||_u. \tag{11}$$

Now, as $(\alpha_g, \alpha_l, \alpha_u) \in \Re^3_+$, we have

$$\alpha_g||\widehat{\Delta g}||_g + \alpha_l||\widehat{\Delta l}||_l + \alpha_u||\widehat{\Delta u}||_u \geq \alpha_g||\Delta g||_g + \alpha_l||\Delta l||_l + \alpha_u||\Delta u||_u,$$

which leads to

$$\min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{D}} \alpha_g||\Delta g||_g + \alpha_l||\Delta l||_l + \alpha_u||\Delta u||_u = \min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{P}} \alpha_g||\Delta g||_g + \alpha_l||\Delta l||_l + \alpha_u||\Delta u||_u$$

and, therefore, $\mathcal{S}_{ws} \subseteq \mathcal{P}$. In addition, using the monotonicity of $||.||_{glu}$, we obtain from (7)

$$||\alpha_g|\widehat{\Delta g}| + \alpha_l|\widehat{\Delta l}| + \alpha_u|\widehat{\Delta u}|||_{glu} \geq ||\alpha_g|\Delta g| + \alpha_l|\Delta l| + \alpha_u|\Delta u|||_{glu},$$

and therefore

$$\min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{D}} ||\alpha_g|\Delta g| + \alpha_l|\Delta l| + \alpha_u|\Delta u|||_{glu} = \min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{P}} ||\alpha_g|\Delta g| + \alpha_l|\Delta l| + \alpha_u|\Delta u|||_{glu},$$

which is $\mathcal{S}_{abs} \subseteq \mathcal{P}$. □

We have just proved that the solution of the backward error in each direction corresponds to perturbing the feasible set $\mathcal{F}$ such that the current iterate becomes feasible, and either driving the gradient to zero or perturbing the feasible set further such that the current iterate lies on the boundary pointed by the negative gradient. The required monotonicity of the norms is necessary as shown on the following example. Consider the nonmonotone energy-norm $||v||_A = \sqrt{v^T A v}$ for all vectors $v \in \Re^n$, where

$$A = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

is positive definite. This norm is indeed nonmonotone since we have, for example, $||(-1;1)||^2_A = 1/5$ but $||(1/2;0)||^2_A = 1/4$. Assume, in addition, that $\tilde{x} = (4;3)$, $\nabla_x f(\tilde{x}) = (3;5)$, and that the bound constraints are defined by $l = (0;0)$ and $u = (5;5)$. The set $\mathcal{P}$ defined by Theorem 2.1 is then composed of the vectors

$$\mathcal{P} = \begin{cases} (\Delta g_1, \Delta l_1, \Delta u_1) &=& (\ (-3;-5),\ (0;0),\ (0;0)\ ), \\ (\Delta g_2, \Delta l_2, \Delta u_2) &=& (\ (-3;\ \ 0),\ (0;3),\ (0;0)\ ), \\ (\Delta g_3, \Delta l_3, \Delta u_3) &=& (\ (\ \ 0;-5),\ (4;0),\ (0;0)\ ), \\ (\Delta g_4, \Delta l_4, \Delta u_4) &=& (\ (\ \ 0;\ \ 0),\ (4;3),\ (0;0)\ ). \end{cases}$$

If we now consider the perturbation $(\widehat{\Delta g}, \widehat{\Delta l}, \widehat{\Delta u}) \stackrel{\text{def}}{=} (\ (5;-5),\ (4;-4),\ (0;0)\ )$, it is easy to verify that it belongs to $\mathcal{D} \setminus \mathcal{P}$ and also that

$$||\widehat{\Delta g}||_A + ||\widehat{\Delta l}||_A + ||\widehat{\Delta u}||_A = 4.0249 < 6.0 = \min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{P}} ||\Delta g||_A + ||\Delta l||_A + ||\Delta u||_A.$$

Hence $\mathcal{S}_{ws} \not\subseteq \mathcal{P}$ in this case.

We observe that, if our assumption on norms is strengthened to require *strict monotonicity* of $||.||_g$, $||.||_l$ and $||.||_u$ in the sense that

$$\text{if } \exists j \in \{1, ..., n\} |[u]_j| > |[v]_j|, \text{ then } ||u|| > ||v|| \quad \forall u, v \in \Re^n,$$

then we may deduce in the proof of Theorem 2.1 not only that (11) holds, but also that at least one of the inequalities

$$||\widehat{\Delta g}||_g > ||\Delta g||_g \text{ or } ||\widehat{\Delta l}||_l > ||\Delta l||_l \text{ or } ||\widehat{\Delta u}||_u > ||\Delta u||_u \tag{12}$$

must hold as well. This will be used in Section 4.

3 Practical criticality measures and backward error analysis

If we wish to find an explicit solution of problems (4) and (5), the result of previous section does help, but does not provide a complete solution in that one still has to solve the combinatorial problem of minimizing the perturbation norm over $\mathcal{P}$. We actually have to specify the chosen norms for $||.||_g, ||.||_l, ||.||_u$ and $||.||_{glu}$ to obtain an explicit expression of the solution. The three following Corollaries are simple consequences of the fact that we determine the point $p \in \mathcal{D}$ from Theorem 2.1 with the smallest components and then use the monotonicity of the norms involved. Consequently, we present here the results for three different norm choices, the proof of which are given in the appendix in order to improve the readability of the paper. We start by considering a specific case where an explicit solution is possible, namely the case where χ_{abs} is chosen and $||.||_{glu} = ||.||_p$ for $1 \leq p < \infty$.

Corollary 3.1 *Suppose that* $||.||_{glu} = ||.||_p$, $1 \leq p < \infty$. *Then*

$$\chi^p_{abs} \stackrel{\text{def}}{=} \min_{(\Delta g, \Delta l, \Delta u) \in \mathcal{D}} ||\alpha_g |\Delta g| + \alpha_l |\Delta l| + \alpha_u |\Delta u| \, ||_p = ||\Lambda||_p, \tag{13}$$

where Λ *is defined componentwise and* $[\Lambda]_j$ *is equal to*

$$\begin{cases} 0 & \text{if } [\tilde{x}]_j \in \mathcal{F}_j \text{ and } [\nabla_x f(\tilde{x})]_j = 0, \\ \min\{\alpha_g |[\nabla_x f(\tilde{x})]_j|, \alpha_l |[\tilde{x} - l]_j|\} & \text{if } [\tilde{x}]_j \in \mathcal{F}_j \text{ and } [\nabla_x f(\tilde{x})]_j > 0, \\ \min\{\alpha_g |[\nabla_x f(\tilde{x})]_j|, \alpha_u |[\tilde{x} - u]_j|\} & \text{if } [\tilde{x}]_j \in \mathcal{F}_j \text{ and } [\nabla_x f(\tilde{x})]_j < 0, \\ \alpha_u |[\tilde{x} - u]_j| & \text{if } [\tilde{x}]_j > [u]_j \text{ and } [\nabla_x f(\tilde{x})]_j \leq 0, \\ \min\{\alpha_g |[\nabla_x f(\tilde{x})]_j|, \alpha_l |[\tilde{x} - l]_j|\} + \alpha_u |[\tilde{x} - u]_j| & \text{if } [\tilde{x}]_j > [u]_j \text{ and } [\nabla_x f(\tilde{x})]_j > 0, \\ \alpha_l |[\tilde{x} - l]_j| & \text{if } [\tilde{x}]_j < [l]_j \text{ and } [\nabla_x f(\tilde{x})]_j \geq 0, \\ \min\{\alpha_g |[\nabla_x f(\tilde{x})]_j|, \alpha_u |[\tilde{x} - u]_j|\} + \alpha_l |[\tilde{x} - l]_j| & \text{if } [\tilde{x}]_j < [l]_j \text{ and } [\nabla_x f(\tilde{x})]_j < 0. \end{cases} \tag{14}$$

We now extend this result to the use of the infinity norm in the definition of χ_{abs}.

Corollary 3.2 *Suppose that* $||.||_{glu} = ||.||_\infty$*, then*

$$\chi_{abs}^{\infty} \stackrel{def}{=} \min_{(\Delta g, \Delta l, \Delta u) \in \mathcal{D}} ||\alpha_g |\Delta g| + \alpha_l |\Delta l| + \alpha_u |\Delta u|\ ||_\infty = ||\Lambda||_\infty \tag{15}$$

where the components of the vector Λ *are defined as in (14).*

We finally show that a similar result holds for χ_{ws} when $||.||_g = ||.||_l = ||.||_u = ||.||_1$, because $\chi_{ws} = \chi_{abs}$ with $||.||_{glu} = ||.||_1$ in that specific case.

Corollary 3.3 *Suppose that* $||.||_g = ||.||_l = ||.||_u = ||.||_1$*. Then*

$$\chi_{ws}^{1} \stackrel{\text{def}}{=} \min_{(\Delta g, \Delta l, \Delta u) \in \mathcal{D}} \alpha_g ||\Delta g||_1 + \alpha_l ||\Delta l||_1 + \alpha_u ||\Delta u||_1 = ||\Lambda||_1, \tag{16}$$

where $[\Lambda]_j$ *is defined by (14).*

The above results are particularly interesting because they allows to express a close form termination criterion in the very frequent case where the weights in (4) and (5) are chosen such that $\alpha_l = \alpha_u \stackrel{\text{def}}{=} \alpha_{lu}$. This is often natural since the lower and upper bounds are generally computed similarly. In this case, we define a vector representing an augmented scaled projection of the negative gradient on the feasible set

$$\Gamma(\alpha_g, \alpha_{lu}) \stackrel{\text{def}}{=} \alpha_{lu} \left(\left| \text{Proj}_{\mathcal{F}(\tilde{x})} \left[\tilde{x} - \frac{\alpha_g}{\alpha_{lu}} \nabla_x f(\tilde{x}) \right] - \tilde{x} \right| + |\tilde{x} - \text{Proj}_{\mathcal{F}}(\tilde{x})| \right), \tag{17}$$

where $\text{Proj}_{\mathcal{F}}(x)$ is the orthogonal projection of x onto the (convex) feasible set $\mathcal{F}$ and relate this quantity to the desired backward error, and where $\mathcal{F}(\tilde{x})$ represents the smallest box containing l, u and $\tilde{x}$; for example, $\mathcal{F}(\tilde{x}) = \mathcal{F}$ when $\tilde{x} \in \mathcal{F}$. It is crucial to note that this augmented scaled projection is easily computable (given the weights) and reduces, as we show below, to popular termination rules for specific weight's choices.

We now verify our claim that (17) is the vector whose norm is the backward error.

Theorem 3.4 *The augmented scaled projection of the negative gradient on the feasible set* $\Gamma(\alpha_g, \alpha_{lu})$ *defined in (17), is such that*

$$\Gamma(\alpha_g, \alpha_{lu}) = \Lambda, \tag{18}$$

where Λ *is defined by (14).*

Proof. In a first step, we show that

$$\left[\left| \text{Proj}_{\mathcal{F}(\tilde{x})} \left[\tilde{x} - \tfrac{\alpha_g}{\alpha_{lu}} \nabla_x f(\tilde{x}) \right] - \tilde{x} \right| \right]_j =$$

$$\begin{cases} 0 & \text{if } [\tilde{x}]_j \in \mathcal{F}_j \text{ and } [\nabla_x f(\tilde{x})]_j = 0, \\ \frac{1}{\alpha_{lu}} \min\{\alpha_g |[\nabla_x f(\tilde{x})]_j|, \alpha_{lu} |[\tilde{x} - l]_j|\} & \text{if } [\tilde{x}]_j \in \mathcal{F}_j \text{ and } [\nabla_x f(\tilde{x})]_j > 0, \\ \frac{1}{\alpha_{lu}} \min\{\alpha_g |[\nabla_x f(\tilde{x})]_j|, \alpha_{lu} |[\tilde{x} - u]_j|\} & \text{if } [\tilde{x}]_j \in \mathcal{F}_j \text{ and } [\nabla_x f(\tilde{x})]_j < 0, \\ 0 & \text{if } [\tilde{x}]_j > [u]_j \text{ and } [\nabla_x f(\tilde{x})]_j \leq 0, \\ \frac{1}{\alpha_{lu}} \min\{\alpha_g |[\nabla_x f(\tilde{x})]_j|, \alpha_{lu} |[\tilde{x} - l]_j|\} & \text{if } [\tilde{x}]_j > [u]_j \text{ and } [\nabla_x f(\tilde{x})]_j > 0, \\ 0 & \text{if } [\tilde{x}]_j < [l]_j \text{ and } [\nabla_x f(\tilde{x})]_j \geq 0, \\ \frac{1}{\alpha_{lu}} \min\{\alpha_g |[\nabla_x f(\tilde{x})]_j|, \alpha_{lu} |[\tilde{x} - u]_j|\} & \text{if } [\tilde{x}]_j < [l]_j \text{ and } [\nabla_x f(\tilde{x})]_j < 0. \end{cases} \tag{19}$$

Assume first that $[\tilde{x}]_j \in \mathcal{F}_j$. The definitions of $\Gamma(\alpha_g, \alpha_{lu})$ and of the orthogonal projection and the fact that $\mathcal{F}(\tilde{x}) = \mathcal{F}$ give that

$$\begin{aligned}\left[\left|\mathrm{Proj}_{\mathcal{F}(\tilde{x})}\left[\tilde{x} - \frac{\alpha_g}{\alpha_{lu}}\nabla_x f(\tilde{x})\right] - \tilde{x}\right|\right]_j \\ &= \begin{cases} \min\left\{\frac{\alpha_g}{\alpha_{lu}}|[\nabla_x f(\tilde{x})]_j|, |[\tilde{x}-l]_j|\right\} & \text{if } \frac{\alpha_g}{\alpha_{lu}}[\nabla_x f(\tilde{x})]_j \geq 0, \\ \min\left\{\frac{\alpha_g}{\alpha_{lu}}|[\nabla_x f(\tilde{x})]_j|, |[\tilde{x}-u]_j|\right\} & \text{if } \frac{\alpha_g}{\alpha_{lu}}[\nabla_x f(\tilde{x})]_j \leq 0, \end{cases} \\ &= \begin{cases} \frac{1}{\alpha_{lu}}\min\{\alpha_g|[\nabla_x f(\tilde{x})]_j|, \alpha_{lu}|[\tilde{x}-l]_j|\} & \text{if } [\nabla_x f(\tilde{x})]_j \geq 0, \\ \frac{1}{\alpha_{lu}}\min\{\alpha_g|[\nabla_x f(\tilde{x})]_j|, \alpha_{lu}|[\tilde{x}-u]_j|\} & \text{if } [\nabla_x f(\tilde{x})]_j \leq 0, \end{cases}\end{aligned}$$

where we used the positiveness of the weigths. Because those two minima are zero when $[\nabla_x f(\tilde{x})]_j = 0$, we have proved (19) for all j such that $[\tilde{x}]_j \in \mathcal{F}_j$. Considering now the infeasible case, we observe that, because of the weights' positiveness, if $[\tilde{x}]_j > [u]_j$ and $[\nabla_x f(\tilde{x})]_j| \leq 0$, the left-hand side of (19) is the projection on $\mathcal{F}(\tilde{x})$ of either the nul vector (if $[\nabla_x f(\tilde{x})]_j| = 0$) or a vector based at one of the bounds and pointing outwards. Therefore this projection is identically zero. The same holds if $[\tilde{x}]_j < [l]_j$ and $[\nabla_x f(\tilde{x})]_j| \geq 0$. On another hand, if $[\tilde{x}]_j > [u]_j$ but $[\nabla_x f(\tilde{x})]_j| > 0$, the projection of the scaled negative gradient on $\mathcal{F}(\tilde{x})$ becomes

$$\min\left\{\frac{\alpha_g}{\alpha_{lu}}|[\nabla_x f(\tilde{x})]_j|, |[\tilde{x}-l]_j|\right\} = \frac{1}{\alpha_{lu}}\min\{\alpha_g|[\nabla_x f(\tilde{x})]_j|, \alpha_{lu}|[\tilde{x}-l]_j|\}$$

and, similarly, the projection of the scaled negative gradient on $\mathcal{F}(\tilde{x})$ is equal to

$$\min\left\{\frac{\alpha_g}{\alpha_{lu}}|[\nabla_x f(\tilde{x})]_j|, |[\tilde{x}-u]_j|\right\} = \frac{1}{\alpha_{lu}}\min\{\alpha_g|[\nabla_x f(\tilde{x})]_j|, \alpha_{lu}|[\tilde{x}-u]_j|\}$$

when $[\tilde{x}]_j < [l]_j$ but $[\nabla_x f(\tilde{x})]_j| < 0$, which concludes the proof of (19). Finally notice that

$$[|x - \mathrm{Proj}_{\mathcal{F}}(x)|]_j = \begin{cases} 0 & \text{if } [\tilde{x}]_j \in \mathcal{F}_j, \\ |[\tilde{x}-u]_j| & \text{if } [\tilde{x}]_j > [u]_j, \\ |[\tilde{x}-l]_j| & \text{if } [\tilde{x}]_j < [l]_j, \end{cases}$$

and the proof is complete. □

Having shown that the augmented scaled projection vector is identical to the solution of the backward error problem in the conditions specified by Corollaries 3.1-3.3, we now restate the explicit forms taken by the associated criticality measures.

Corollary 3.5 *Suppose that $\alpha_l = \alpha_u = \alpha_{lu}$, and that $\|\cdot\|_p$, $1 \leq p \leq \infty$, is used in (5). Then*

$$\chi_{abs}{}^p = ||\Gamma(\alpha_g, \alpha_{lu})||_p . \tag{20}$$

Proof. This is an immediate consequence of Corollary 3.1, Corollary 3.2 and Theorem 3.4. □

Corollary 3.6 *Suppose that $\alpha_l = \alpha_u = \alpha_{lu}$, and that $\|\cdot\|_1$ is used in (4). Then*

$$\chi_{ws}{}^1 = ||\Gamma(\alpha_g, \alpha_{lu})||_1 . \tag{21}$$

Proof. This is an immediate consequence of Corollary 3.3 and Theorem 3.4. □

We now comment on formula (17). In the light of Theorem 3.4, we first see that, if $\tilde{x}$ is feasible, then $||\Gamma(\alpha_g, \alpha_{lu})||$ reduces to the scaled projection of the negative gradient on the feasible space

$$\Gamma(\alpha_g, \alpha_{lu}) \stackrel{\rm def}{=} \alpha_{lu} \left(\mathrm{Proj}_{\mathcal{F}} \left[\tilde{x} - \frac{\alpha_g}{\alpha_{lu}} \nabla_x f(\tilde{x}) \right] - \tilde{x} \right). \tag{22}$$

Moreover, an immediate consequence of Corollary 3.5 and Corollary 3.6 is that, if $\tilde{x}$ is feasible and $\alpha_g = \alpha_{lu} = 1$, then the optimal value for (5) is

$$\chi = \|\Gamma(1,1)\|_p = \|\mathrm{Proj}_{\mathcal{F}} [\tilde{x} - \nabla_x f(\tilde{x})] - \tilde{x}\|_p ,$$

which is a quantity commonly used in actual termination rules for bound-constrained optimization (see Byrd *et al.* [7], Hager and Zhang [19] and Xu and Burke [25]). The choice of the measure (17) however allows acting on the weights α_g and α_{lu}. This feature is useful for instance when the error on the gradient (ϵ_g) is comparatively larger than that on the bounds (ϵ_{lu}), a situation which is not untypical, for instance in discretized contact problems (see Dostal [14]). The formulation (17) then makes the use of a single termination accuracy ϵ reasonable even if these errors are different, by using

$$\alpha_g = 1/\epsilon_g \ \text{ and } \ \alpha_{lu} = 1/\epsilon_{lu}. \tag{23}$$

If the solution process is terminated when $\|\Gamma(\alpha_g, \alpha_{lu})\|_p \leq 0.1$, for instance, this ensures that any accepted solution of the optimization problem (1) has a backward error on the gradient at least an order of magnitude smaller than ϵ_g, and is therefore negligible, the same being true for the backward error on the bounds constraints.

If the current point is feasible, the definition of (22) may also be related to a second case of interest: the criticality measure defined by the norm of the reduced gradient g_{red}, defined by the projection of $\nabla_x f(\tilde{x})$ on the tangent cone of the constraints, or more precisely,

$$[g_{red}]_j \stackrel{\rm def}{=} \begin{cases} 0 & \text{if } j \in \mathcal{A}(\tilde{x}), \\ [\nabla_x f(\tilde{x})]_j & \text{otherwise.} \end{cases}$$

(see Burke and Moré [5], Calamai and Moré [8], Burke *et al.* [6], Burke [4] or Dostal [14], for example). It is interesting to note that in the case where $\alpha_l = \alpha_u$, $\tilde{x} \in \mathcal{P}$ and $\|\cdot\|_{glu} = \|\cdot\|_p$, $1 \leq p \leq \infty$, we have that

$$\|g_{red}\|_p = \lim_{\alpha_{lu} \to \infty} \|\Gamma(1, \alpha_{lu})\|_p. \tag{24}$$

Indeed, since $\lim_{\alpha_{lu}\to\infty} \|\Gamma(1,\alpha_{lu})\|_p = \|\lim_{\alpha_{lu}\to\infty} \Gamma(1,\alpha_{lu})\|_p$, (14), (22) and $\tilde{x} \in \mathcal{F}$ imply that

$$\begin{aligned}\lim_{\alpha_{lu}\to\infty} \Gamma(1,\alpha_{lu}) &= \lim_{\alpha_{lu}\to\infty} \begin{cases} 0 & \text{if } [\nabla_x f(\tilde{x})]_j = 0, \\ \min\{|[\nabla_x f(\tilde{x})]_j|, \alpha_{lu}|[\tilde{x}-l]_j|\} & \text{if } [\nabla_x f(\tilde{x})]_j > 0, \\ \min\{|[\nabla_x f(\tilde{x})]_j|, \alpha_{lu}|[\tilde{x}-u]_j|\} & \text{if } [\nabla_x f(\tilde{x})]_j < 0, \end{cases} \\ &= \lim_{\alpha_{lu}\to\infty} \begin{cases} 0 & \text{if } j \in \mathcal{A}(\tilde{x}) \text{ or } [\nabla_x f(\tilde{x})]_j = 0, \\ \min\{|[\nabla_x f(\tilde{x})]_j|, \alpha_{lu}|[\tilde{x}-l]_j|\} & \text{if } \begin{cases} j \notin \mathcal{A}(\tilde{x}) \text{ and} \\ [\nabla_x f(\tilde{x})]_j > 0, \end{cases} \\ \min\{|[\nabla_x f(\tilde{x})]_j|, \alpha_{lu}|[\tilde{x}-u]_j|\} & \text{if } \begin{cases} j \notin \mathcal{A}(\tilde{x}) \text{ and} \\ [\nabla_x f(\tilde{x})]_j < 0, \end{cases} \end{cases}\end{aligned}$$

because $[\tilde{x}]_j = [l]_j$ when $j \in \mathcal{A}(\tilde{x})$ and $[\nabla_x f(\tilde{x})]_j > 0$ and $[\tilde{x}]_j = [u]_j$ when $j \in \mathcal{A}(\tilde{x})$ and $[\nabla_x f(\tilde{x})]_j < 0$. Finally, taking the limit as α_{lu} goes to infinity gives that

$$\begin{aligned}\lim_{\alpha_{lu}\to\infty} \Gamma(1,\alpha_{lu}) &= \begin{cases} 0 & \text{if } j \in \mathcal{A}(\tilde{x}), \\ \alpha_g |[\nabla_x f(\tilde{x})]_j| & \text{if } j \notin \mathcal{A}(\tilde{x}), \end{cases} \\ &= \|g_{red}\|_p.\end{aligned}$$

When $\tilde{x} \in \mathcal{F}$, we also observe that

$$\lim_{\alpha_g\to\infty} \|\Gamma(\alpha_g,1)\|_p = ||d||_p, \tag{25}$$

where d is a vector joining the current iterate to the corner of the feasible set designated by the negative gradient, and whose components are defined by

$$[d]_j = \begin{cases} [l-\tilde{x}]_j & \text{if } [\nabla_x f(\tilde{x})]_j > 0, \\ [u-\tilde{x}]_j & \text{if } [\nabla_x f(\tilde{x})]_j < 0, \\ 0 & \text{if } [\nabla_x f(\tilde{x})]_j = 0. \end{cases}$$

This is an immediate result of letting α_g tend to infinity in (14) with $\alpha_{lu} = 1$. Equations (24) and (25) are an illustration of the sensitivity of backward error to the weights $\alpha_g, \alpha_l, \alpha_u$. We expect that for large α_g, χ will reflect the distance d from $\tilde{x}$ to the bounds pointed by the negative gradient. For large α_{lu}, χ will behave like the projection of $\nabla_x f(\tilde{x})$ on the tangent cone to the constraints.

After having recovered two well-known criticality measures from our backward analysis approach, we now observe that not every such criticality measure can be viewed under that angle. For example, the measure defined by

$$\mu \stackrel{\text{def}}{=} | \min_{\substack{\tilde{x}+d\in\mathcal{F} \\ \|d\|_\infty \le 1}} \nabla_x f(\tilde{x})^T d| \tag{26}$$

is often used in trust-region algorithms and can be interpreted as giving a first-order approximation of the feasible decrease which can be achieved in a ball of radius one (see Conn

et al. [10]). The use of the infinity norm $\|\cdot\|_\infty$ in this definition is motivated by the observation that the intersection of the feasible set with the unit ball remains a box, which makes the computation of μ straighforward. Unfortunately, μ is in general not a backward error in any norm, as we now show.

Lemma 3.7 *The criticality measure μ is not a backward error in the sense of (3), i.e. there does not exist a product norm $\|\cdot\|_{tr}$ such that, for all problems of the type (1), we have that*

$$\mu = \min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{D}} \|(\Delta g, \Delta l, \Delta u)\|_{tr}. \tag{27}$$

Proof. We only need to find one problem (one specific $\tilde{x}$, f, $\mathcal{F}$) where there is no norm such that (27) holds. We therefore consider the minimization of a linear function subject to some bound constraints $l \leq x \leq u$ and such that its constant gradient is negative, that is $\nabla_x f(\tilde{x}) < 0$ for all feasible iterates $\tilde{x}$, where the inequality is understood componentwise.

If we consider some $\tilde{x} > u - 1$, in that specific case, $d^* = (u - \tilde{x})$ for all k and

$$\mu = |-\nabla_x f(\tilde{x})^T(u-\tilde{x})|.$$

So we suppose that there exists $\|\cdot\|_{tr}$ such that

$$\mu = |-\nabla_x f(\tilde{x})^T(u-\tilde{x})| = \min_{y\in\mathcal{D}} \|(\Delta g, \Delta l, \Delta u)\|_{tr} = \|(\Delta g^*, \Delta l^*, \Delta u^*)\|_{tr}. \tag{28}$$

We then obtain, using the Cauchy-Schwarz inequality,

$$1 \geq \frac{|-\nabla_x f(\tilde{x})^T(u-\tilde{x})|}{\|-\nabla_x f(\tilde{x})\|_2\|u-\tilde{x}\|_2} = \frac{\|(\Delta g^*, \Delta l^*, \Delta u^*)\|_{tr}}{\|-\nabla_x f(\tilde{x})\|_2\|u-\tilde{x}\|_2}. \tag{29}$$

As we consider a feasible $\tilde{x}$, the vectors $\Delta g^*, \Delta l^*, \Delta u^*$ are such that

$$\begin{gathered}
[\nabla_x f(\tilde{x}) + \Delta g^*]_j = 0 \\
\text{or} \\
[l + \Delta l^*]_j = [\tilde{x}]_j \text{ and } [\nabla_x f(\tilde{x}) + \Delta g^*]_j > 0 \\
\text{or} \\
[u + \Delta u^*]_j = [\tilde{x}]_j \text{ and } [\nabla_x f(\tilde{x}) + \Delta g^*]_j < 0.
\end{gathered} \tag{30}$$

Consider now a sequence of iterates $\tilde{x}_k$ and assume first that $[\Delta g^*]_j = [-\nabla_x f(\tilde{x}_k)]_j$ for all j and for k sufficiently large, i.e. for all $k \geq k_1$. In that case, because all norms are equivalent in finite dimension, there exists a constant ν such that

$$\|(\Delta g^*, \Delta l^*, \Delta u^*)\|_{tr} \geq \nu(\|\Delta g^*\|_2 + \|\Delta l^*\|_2 + \|\Delta u^*\|_2) \geq \nu\|\Delta g^*\|_2 = \nu\|-\nabla_x f(\tilde{x}_k)\|_2, \tag{31}$$

where we used the fact that $\|(u,v,w)\| \stackrel{\rm def}{=} \|u\|_2 + \|v\|_2 + \|w\|_2$ is a norm on $\Re^n \times \Re^n \times \Re^n$, where n is the dimension of the problem. Equation (29) therefore gives $1 \geq \nu/\|u-\tilde{x}_k\|_2$. We consider more specifically the sequence of iterates such that $\tilde{x}_k$ is monotonically converging

to the upper bound u such that $[u - \tilde{x}_k]_j = 1/k$ for all j and for all k (implying $\tilde{x}_k > u - 1$ for all k). Then the last equation leads to

$$1 \geq \lim_{k\to\infty} \frac{\nu}{\|u - \tilde{x}_k\|_2} = k\frac{\nu}{\sqrt{n}} = +\infty,$$

which is impossible. We thus conclude that our assumption is false and, because of (30), we deduce that there exists at least one index j and at least one $k \geq k_1$ such that either $[\Delta l^*]_j = [l - \tilde{x}_k]_j$ or $[\Delta u^*]_j = [u - \tilde{x}_k]_j$. This, together with the first inequality of (31), implies that

$$\begin{aligned} \|(\Delta g^*, \Delta l^*, \Delta u^*)\|_{tr} &\geq \nu(\|\Delta g^*\|_2 + \|\Delta l^*\|_2 + \|\Delta u^*\|_2) \\ &\geq \nu \min\{\|\Delta l^*\|_2, \|\Delta u^*\|_2\} \\ &\geq \nu \min\{|[l - \tilde{x}_k]_j|, |[u - \tilde{x}_k]_j|\} \end{aligned}$$

and, therefore, (29) gives that

$$1 \geq \frac{\nu \min\{|[l - \tilde{x}_k]_j|, |[u - \tilde{x}_k]_j|\}}{\| - \nabla_x f(\tilde{x}_k)\|_2 \|u - \tilde{x}_k\|_2}.$$

The assumption $[u - \tilde{x}_k]_j = 1/k$ implies that $\|u - \tilde{x}_k\|_2 = \sqrt{n}/k$ and that there exists k_2 such that for all $k \geq k_2$ we have $|[l - \tilde{x}_k]_j| > |[u - \tilde{x}]_j|$. We obtain

$$1 \geq \frac{\nu|[u - \tilde{x}]_j|}{\| - \nabla_x f(\tilde{x})\|_2 \|u - \tilde{x}\|_2} \geq \frac{\nu}{\sqrt{n}\| - \nabla_x f(\tilde{x})\|_2},$$

which is impossible for all problems where the constant gradient is chosen such that $\| - \nabla_x f(\tilde{x})\|_2 < \nu/\sqrt{n}$. We conclude that our assumption (28) is false, and the proof is complete. □

4 A multicriteria analysis

Solving the backward error problem corresponds to finding the minimal distance between the original problem and the closest problem we have already solved at iteration k. We have so far measured this distance by means of a product norm defined on the space of the perturbations, for instance by constructing a positive linear combination of the individual perturbation norms, as in Section 2.2. This approach is quite natural since one often has information about Δg, Δl and Δu and some choice of norms for $\|\cdot\|_g, \|\cdot\|_l, \|\cdot\|_u$ may be suggested by the underlying application. Aggregating them in a suitable positive linear combination may therefore be reasonable. This is however not the only possibility and we briefly explore, in this section, the use of the multicriteria optimization (MCO) (see Ehrgott [15] for more details on this subject) problem of the form

$$\begin{aligned} &\text{“min”} \ (\|\Delta g\|_g, \|\Delta l\|_l, \|\Delta u\|_u) \\ &\text{s.t.} \ (\Delta g, \Delta l, \Delta u) \in \mathcal{D}. \end{aligned} \tag{32}$$

Notice that the previous definition (4) of the backward error problem can be viewed as a *scalarization* of the more general problem (32), consisting of taking a linear combination of the three objective functions with positive weights. A solution $(\Delta g^*, \Delta l^*, \Delta u^*)$ of the general MCO problem (32) is a *Pareto optimal* solution, if and only if there exists no $(\Delta g, \Delta l, \Delta u) \in \mathcal{D}$ such that

$$\begin{array}{ccccccc} \|\Delta g\|_g \leq \|\Delta g^*\|_g & \text{and} & \|\Delta l\|_l \leq \|\Delta l^*\|_l & \text{and} & \|\Delta u\|_u \leq \|\Delta u^*\|_u \\ \|\Delta g\|_g < \|\Delta g^*\|_g & \text{or} & \|\Delta l\|_l < \|\Delta l^*\|_l & \text{or} & \|\Delta u\|_u < \|\Delta u^*\|_u. \end{array}$$

In that case, we say that the feasible point $(\Delta g^*, \Delta l^*, \Delta u^*)$ is *not dominated* by any other feasible point. The set $\mathcal{D}_E$ of all Pareto optimal solutions is called the *Pareto optimal set*, while $\mathcal{Y}_N$ represents the set of all *nondominated points* $y_n = (\|\Delta g_e\|_g, \|\Delta l_e\|_l, \|\Delta u_e\|_u) \in \Re^3$, where $(\Delta g_e, \Delta l_e, \Delta u_e) \in \mathcal{D}_E$, and is called the *nondominated set*.

Theorem 2.1 in Section 2.2 has established that the solution of the backward error problem is located in the set $\mathcal{P}$. Looking back at this theorem (relation (11)) and the subsequent comment yielding (12) from the MCO point view, we conclude that all $(\Delta g, \Delta l, \Delta u) \in \mathcal{D} \setminus \mathcal{P}$ are dominated by at least one point of $\mathcal{P}$, and thus cannot be Pareto optimal for the original MCO problem. As a consequence, we deduce that

$$\mathcal{D}_E \subseteq \mathcal{P}.$$

Unfortunately, we cannot say which solution of $\mathcal{P}$ is Pareto optimal without knowing the specific values of $\tilde{x}, l, u$ and $\nabla_x f(\tilde{x})$. In addition, notice that, because the standard definition of the backward error is a scalarization of the MCO problem, a solution of (4) is always also Pareto optimal, that is

$$\mathcal{S}_{ws} \subseteq \mathcal{D}_E \quad \text{and} \quad (\|\Delta g^*\|_g, \|\Delta l^*\|_l, \|\Delta u^*\|_u) \in \mathcal{Y}_N \text{ for all } (\Delta g^*, \Delta l^*, \Delta u^*) \in \mathcal{S}_{ws}.$$

Nevertheless, if $\mathcal{Y}_N$ is not a convex set, we may not access all $y_n \in \mathcal{Y}_N$ by scalarization (see Ehrgott [15], pp 68-73, for a proof of these two properties). We illustrate this observation on a simple example. Consider some iterate $\tilde{x} = (3; 4; 1)$ obtained during the minimization of a problem with the bound constraints $l = (0; 0; 0)$ and $u = (5; 5; 5)$, for which the gradient is equal to $\nabla_x f(\tilde{x}) = (4; 3; 1)$. Assume that we have chosen $\|\cdot\|_g = \|\cdot\|_l = \|\cdot\|_u = \|\cdot\|_\infty$. In that case, $\mathcal{P}$ contains

$$\left\{ \begin{array}{lcl} (\Delta g_1, \Delta l_1, \Delta u_1) & = & (\,(-4; -3; -1), \quad (0; 0; 0), \quad (0; 0; 0)\,) \\ (\Delta g_2, \Delta l_2, \Delta u_2) & = & (\,(-4; -3; \quad 0), \quad (0; 0; 1), \quad (0; 0; 0)\,) \\ (\Delta g_3, \Delta l_3, \Delta u_3) & = & (\,(-4; \quad 0; -1), \quad (0; 4; 0), \quad (0; 0; 0)\,) \\ (\Delta g_4, \Delta l_4, \Delta u_4) & = & (\,(\quad 0; -3; -1), \quad (3; 0; 0), \quad (0; 0; 0)\,) \\ (\Delta g_5, \Delta l_5, \Delta u_5) & = & (\,(-4; \quad 0; \quad 0), \quad (0; 4; 1), \quad (0; 0; 0)\,) \\ (\Delta g_6, \Delta l_6, \Delta u_6) & = & (\,(\quad 0; -3; \quad 0), \quad (3; 0; 1), \quad (0; 0; 0)\,) \\ (\Delta g_7, \Delta l_7, \Delta u_7) & = & (\,(\quad 0; \quad 0; -1), \quad (3; 4; 0), \quad (0; 0; 0)\,) \\ (\Delta g_8, \Delta l_8, \Delta u_8) & = & (\,(\quad 0; \quad 0; \quad 0), \quad (3; 4; 1), \quad (0; 0; 0)\,) \end{array} \right.$$

and we can compute, using the definition of a Pareto optimal solution, the set

$$\mathcal{D}_E = \{(\Delta g_1, \Delta l_1, \Delta u_1), (\Delta g_4, \Delta l_4, \Delta u_4), (\Delta g_6, \Delta l_6, \Delta u_6), (\Delta g_8, \Delta l_8, \Delta u_8)\},$$

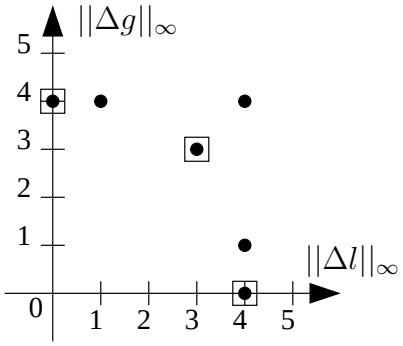

Figure 1: For each $(\Delta g, \Delta l, \Delta u) \in \mathcal{P}$, we represented $(\|\Delta g\|_\infty, \|\Delta l\|_\infty)$ by a big dot and the elements of $\mathcal{Y}_N$ are surrounded by a square. Δu is not represented here because $\|\Delta u\|_\infty \in [0, +\infty)$ for all perturbations in $\mathcal{D}$ and $\|\Delta u\|_\infty = 0$ for all perturbations in $\mathcal{P}$. We see that the Pareto front is not convex so we cannot access $y_2 = (3; 3; 0)$ by means of a scalarization.

leading to $\mathcal{Y}_N = \{y_1 = (4, 0, 0), y_2 = (3, 3, 0), y_3 = (0, 4, 0)\}$. The image of $\mathcal{P}$ and the Pareto front are shown in Figure 1. In this case, the possibly interesting perturbation $y_2 = (3, 3, 0)$ cannot be reached by any scalarization.

The interest of this multicriteria approach to the backward error is that it may lead to terminate the algorithm even sooner than with (4), at a still acceptable approximate solution of the optimization problem. The choice of the interesting point on the Pareto front would be left to a "decision maker" in this approach.

5 Numerical examples

In this section, we illustrate the interest of adapting the stopping criterion of a bound-constrained optimization algorithm according to the error bounds we may know on the data. For this purpose, we consider the well-known minimal surface problem with obstacle

$$\min_{v \in \mathcal{K}} \int_{S_2} \sqrt{1 + \|\nabla_x v\|_2^2}, \tag{33}$$

where $\mathcal{K} = \big\{v \in H^1(S_2) \mid v(x) = v_0(x) \text{ on } \partial S_2\big\}$, S_2 is the unit square $\{(x, y) \in \Re^2 \mid 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1\}$ and where v must satisfy the constraints

$$\begin{array}{ll} v(x) \geq 0.7 & \text{if } \frac{1}{3} \leq x_1, x_2 \leq \frac{2}{3}, \\ v(x) \geq 0 & \text{otherwise.} \end{array}$$

This convex problem is discretized using a finite-element basis defined using a uniform triangulation of S_2, with the same grid spacing, $h = 1/n$, along the two coordinate directions. The basis functions are the standard P1 functions which are linear on each triangle and

take the value 0 or 1 at each vertex. The boundary condition $v_0(x)$ is chosen as

$$v_0(x) = \begin{cases} x_1(1-x_1), & x_2 = 0, \quad 0 \leq x_1 \leq 1, \\ 0, & x_1 = 0, \quad 0 \leq x_2 \leq 1, \\ x_1(1-x_1), & x_2 = 1, \quad 0 \leq x_1 \leq 1, \\ 0, & x_1 = 1, \quad 0 \leq x_2 \leq 1, \end{cases}$$

We then modify the discretized version of this problem, here considered as the original optimization problem, by adding the following linear term : $10^{-2}\sin(1:n)^T x$, where n is the dimension of the discretized problem and $\sin(1:n)$ is a vector of $\Re^n$ whose i^{th} component is equal to $\sin(i)$. This modified problem can be viewed as an approximation of the original discretized problem with an error on the gradient of $\mathcal{O}(10^{-2})$. We now compare the behavior of two different criticality measures during the application of an infinity-norm trust-region algorithm using a projected truncated conjugate gradient algorithm as internal solver applied on this perturbed problem with $n = 3969$ variables. The first measure is the standard 1-norm of the projection of the negative gradient on the feasible set with a stopping threshold set to $\epsilon = 10^{-15}$. The second measure is the scaled version (22) of the previous measure, where the weights are chosen as in (23) with $\epsilon_g = 10^{-2}$ and $\epsilon_{lu} = 10^{-14}$ because the problem has an error of $\mathcal{O}(10^{-2})$ on $\nabla_x f(.)$ but the bounds are computed exactly. In this case, as suggested after (23), the stopping tolerance ϵ is set to 10^{-1} in order to ensure that the final solving error on the gradient will be insignificant in comparison with the error known on its computation. Notice that this choice also ensures that the solving error allowed on the bound constraints will be reduced to the order of 10^{-15} as in the first case.

	Nb Iter	Evals f	Evals $\nabla_x f(x)$	Evals $\nabla_{xx} f(x)$
$\|\Gamma\|_1$	493	147	142	80
$\|\Gamma(1/\epsilon_g, 1/\epsilon_{lu})\|_1$	228	137	132	80

Table 1: Total number of iterations, function, gradient and Hessian evaluations when stopping the algorithm using $\|\Gamma\|_1$ and $\|\Gamma(1/\epsilon_g, 1/\epsilon_{lu})\|_1$.

The total number of iterations, function, gradient and Hessian evaluations at convergence are displayed in Table 1. The number of iterations presented here corresponds to internal iterations or conjugate gradient iterations in the bound-constrained quadratic trust-region subproblem. The number of external trust-region iterations has not been represented because it is equal to the number of function evaluations. As expected, the scaled criticality measure is less restrictive and we can see that stopping the algorithm as soon as the backward error on the gradient is significantly smaller than its intrinsic error implies a huge reduction of the number of conjugate gradient iterations. In addition, the use of an adapted stopping criterion allowed to save function evaluations, which can be crucial when dealing with real industrial problems. Notice that the proportion of internal iterations per function evaluation changes along the algorithm because active bounds are detected progressively, and the conjugate gradient iterations are stopped as soon as two bounds have been activated. Therefore, there are generally few internal iterations per external iteration at the

beginning of the algorithm, but when closer to convergence most of the active bounds have been detected by the algorithm, and it is allowed to perform many more conjugate gradient steps before finishing one external iteration. In addition, the equal number of Hessian evaluations comes from the fact that the Hessian matrix is not computed at every iteration. More precisely, the Hessian matrix is only computed whenever the preceding iteration is not successful enough or when $\|g_k - g_{k-1} - H_{k-1}s_{k-1}\|_2 > \epsilon_H \|g_k\|_2$, where ϵ_H has been set to 10^{-3}. This condition may not hold at the last iterations because the minimized function behaves like a quadratic when approaching convergence, which explains the same number of Hessian evaluations. Moreover, no significant improvement has been obtained on the objective function value with the more stringent stopping criteria (the relative difference between the two values is actually 2.662e-9). Of course the scope of illustration remains limited, but it definitely suggests that the use of termination rule based on backward error analysis can be beneficial.

Another interesting property of the scaled criticality measure is that the choice of the weigths in the scaling may have a significant influence on the shape of the acceptable solution. For the purpose of illustration, consider now the following quadratic problem

$$\begin{aligned} -\Delta u(x)/10 &= f(x) \text{ in } S_2 \\ u(x) &= 0 \text{ on } \partial S_2, \end{aligned}$$

where $f(x)$ is such that the analytical solution to this problem is $u(x) = 2x_2(x_2 - 1) + 2x_1(x_1 - 1)$. The problem is submitted to the following bound constraint

$$\begin{array}{ll} u(x) \geq 7.5 & \text{if } \frac{4}{9} \leq x_1, x_2 \leq \frac{5}{9} \\ u(x) \geq 5 & \text{if } \frac{1}{9} \leq x_1, x_2 \leq \frac{2}{9}, \text{ or } \frac{1}{9} \leq x_1 \leq \frac{2}{9} \text{ and } \frac{7}{9} \leq x_2 \leq \frac{8}{9}, \\ & \text{or } \frac{7}{9} \leq x_1 \leq \frac{8}{9} \text{ and } \frac{1}{9} \leq x_2 \leq \frac{2}{9}, \text{ or } \frac{7}{9} \leq x_1, x_2 \leq \frac{8}{9} \\ u(x) \geq 0 & \text{otherwise,} \end{array}$$

and is discretized using a 5-point finite-difference scheme with $h = 1/3969$. We consider four approximate solutions of this problem, acceptable for the scaled criticality measure with weights chosen as in (23) and where the tolerances are arbitrarily chosen as $\epsilon_g = 10^{-8}$ and $\epsilon_{lu} = 10^{-8}$, $\epsilon_g = 10^{-8}$ and $\epsilon_{lu} = 10^{-2}$, $\epsilon_g = 10^{-2}$ and $\epsilon_{lu} = 10^{-8}$, and finally $\epsilon_g = 10^{-2}$ and $\epsilon_{lu} = 10^{-2}$. Notice that $\|\Gamma(1/10^{-8}, 1/10^{-8})\|_1$ is a representative of standard stopping criteria. Figure 2 first shows the distance between the approximate solution and the bound constraint for all active components at the exact solution (this restriction is denoted by the subscript a), while Figure 3 illustrates the gradient of the approximate solutions for all inactive components at the exact solution (this restriction is denoted by the subscript i). Table 2 contains the ℓ_1-norm of the same quantities in the three situations considered, together with the value of the corresponding criticality measure.

	$\|\nabla_x f(\tilde{x}_i)\|_1$	$\|\tilde{x}_a - l_a\|_1$	$\|\Gamma(1/\epsilon_g, 1/\epsilon_{lu})\|_1$
$\epsilon_g = 10^{-8}, \epsilon_{lu} = 10^{-8}$	9.1601e-11	0	0.0092
$\epsilon_g = 10^{-8}, \epsilon_{lu} = 10^{-2}$	9.1586e-11	5.5315e-05	0.0645
$\epsilon_g = 10^{-2}, \epsilon_{lu} = 10^{-8}$	6.2852e-04	0	0.0629
$\epsilon_g = 10^{-2}, \epsilon_{lu} = 10^{-2}$	6.2852e-04	5.5226e-05	0.0684

Table 2: The 1-norm of the gradient of approximate solutions on all inactive components at the exact solution, the 1-norm of the distance between the approximate solution and the bound constraint on all active components at the exact solution and the value of the scaled criticality measure (22) are presented with regard to different values of ϵ_g and ϵ_{lu}.

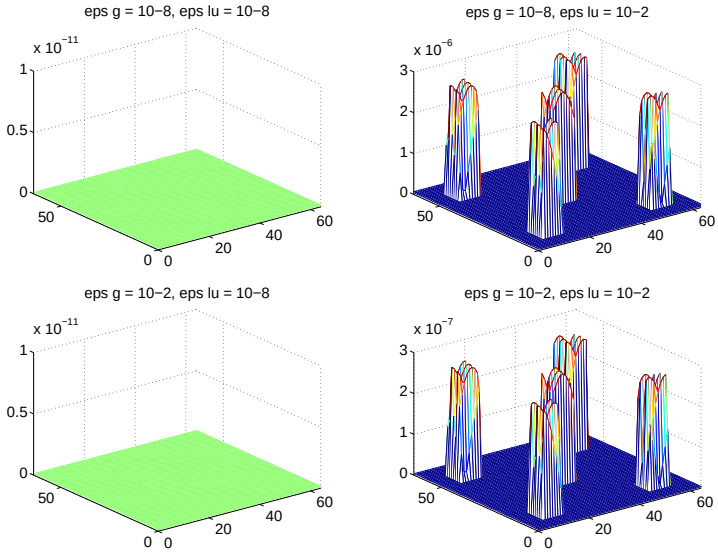

Figure 2: The distance between acceptable solutions for the scaled criticality measure with different values of ϵ_g and ϵ_{lu} and the bound constraint for all active components at the exact solution.

We see on this example that the gradient and the distance to the bound constraints is handled differently when the weights of the scaled criticality measure are changed. For example, the flexibility left on the accuracy required on the gradient has been used in the second and the third cases, without negatively affecting the accuracy on the distance to the bounds when ϵ_{lu} is set to 10^{-8}. In practice, of course, the shape of the approximate solution obtained with a specific criticality measure will also depends on the choice of the algorithm producing the iterates. For instance, if the algorithm is designed to identify quickly the correct active set, it is possible that the backward error on the bound constraints remains insignificant for all reasonable values of ϵ_{lu} when using the scaled criticality measure.

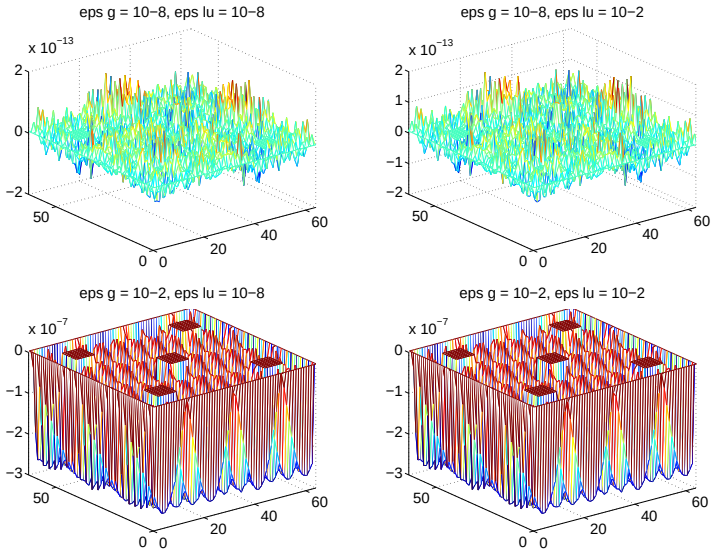

Figure 3: The gradient of acceptable solutions for the scaled criticality measure with different values of ϵ_g and ϵ_{lu} at all inactive components at the exact solution.

6 Conclusion

We have applied the concept of backward error analysis on the problem of finding meaningful stopping criteria for nonlinear bound-constrained optimization algorithms. We have first shown that known criticality measures for this problem based on the projected and reduced gradient can be viewed as backward error measures. Variants of the first of these measures have been suggested for the case where the error on the gradient and on the bounds are known and of different orders of magnitude. We have also indicated why a measure constructed on the feasible linear decrease in a unit ball can not be interpreted in this way, and have defined a multicriteria backward error that opens the way to the use of new stopping criteria. A numerical example has finally been presented to illustrate the potential benefits of our approach.

The authors believe that backward-error-based termination criteria have a real potential for avoiding oversolving optimization problems, both at the nonlinear level and at the level of the subproblem solution, where approximate formulations are typically considered. For instance the present results already cover the solution of the ℓ_∞ trust-region subproblem, but the case of the Euclidean norm is also of interest. These ideas of course need further analysis and more extensive numerical confirmation, but the initial results are encouraging. Moreover, the extention of the theory to the use of possibly nonmonotone norms, such as energy norms (see Arioli, Loghin and Wathen [2]) should be considered in the future in order to apply backward error stopping criteria to the solution of optimization problems resulting from the discretization of partial differential equations.

Acknowledgement

This paper was finalized with the support of the "Assimilation de Données pour la Terre, l'Atmosphère et l'Océan (ADTAO)" project, funded by the Fondation "Sciences et Technologies pour l'Aéoronautique et l'Espace (STAE)", Toulouse, France, within the "Réseau Thématique de Recherche Avancée (RTRA)".

References

[1] M. Arioli, I. S. Duff, and D. Ruiz. Stopping criteria for iterative solvers. *SIAM Journal on Matrix Analysis and Applications*, 13(1):138–144, 1992.

[2] M. Arioli, D. Loghin, and A. Wathen. Stopping criteria for iterations in finite element methods. *Numerische Mathematik*, 99(3):381–410, 2004.

[3] B. M. Averick and J. J. Moré. The Minpack-2 test problem collection. Technical Report ANL/MCS-P153-0694, Mathematics and Computer Science, Argonne National Laboratory, Argonne, Illinois, USA, 1992.

[4] J. V. Burke. On the identification of active constraints II: the nonconvex case. *SIAM Journal on Numerical Analysis*, 27(4):1081–1102, 1990.

[5] J. V. Burke and J. J. Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

[6] J. V. Burke, J. J. Moré, and G. Toraldo. Convergence properties of trust region methods for linear and convex constraints. *Mathematical Programming, Series A*, 47(3):305–336, 1990.

[7] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[8] P. H. Calamai and J. J. Moré. Projected gradient methods for linearly constrained problems. *Mathematical Programming*, 39:93–116, 1987.

[9] F. Chaitin-Chatelin and V. Frayssé. *Lectures on Finite Precision Computations*. SIAM, Philadelphia, USA, 1996.

[10] A. R. Conn, N. I. M. Gould, A. Sartenaer, and Ph. L. Toint. Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints. *SIAM Journal on Optimization*, 3(1):164–221, 1993.

[11] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. Number 01 in MPS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000.

[12] A. J. Cox and N. J. Higham. Backward error bounds for constrained least squares problems. *BIT*, 39(2):210–227, 1999.

[13] E. D. Dolan, J. J. Moré, and T. S. Munson. Benchmarking optimization software with COPS 3.0. Technical Report ANL/MCS-TM-237, Mathematics and Computer Science, Argonne National Laboratory, Argonne, Illinois, USA, 2004.

[14] Z. Dostál. *Optimal Quadratic Programming Algorithms*. Springer Verlag, Heidelberg, Berlin, New York, 2009.

[15] M. Ehrgott. *Multicriteria Optimization*. Lecture Notes in Economics and Mathematical Systems. Springer Verlag, Heidelberg, Berlin, New York, second edition, 2005.

[16] M. Fisher. Minimization algorithms for variational data assimilation. In *Recent Developments in Numerical Methods for Atmospheric Modelling*, pages 364–385, Reading, UK, 1998. European Center for Medium-Range Weather Forecasts.

[17] W. Givens. Numerical computation of the characteristic values of a real matrix. Technical Report 1574, Oak Ridge National Laboratory, Oak Ridge, USA, 1954.

[18] G. H. Golub and C. F. Van Loan. *Matrix computations*. North Oxford Academic, Oxford, UK, 1983.

[19] W. W. Hager and H. Zhang. A new active set algorithm for box constrained optimization. *SIAM Journal on Optimization*, 17(2):526–557, 2006.

[20] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, USA, 1996.

[21] J. J. Moré and S. M. Wild. Estimating computational noise. Technical Report ANL-MCS-P1721-0210, Mathematics and Computer Science, Argonne National Laboratory, Argonne, Illinois, USA, 2010.

[22] M. Mouffe. *Multilevel optimization in infinity norm and associated stopping criteria*. PhD thesis, Department of Mathematics, FUNDP - University of Namur, Namur, Belgium, 2009.

[23] J. L. Rigal and F. Gaches. On the compatibility of a given solution with the data of a linear system. *Journal of the ACM*, 14(3):543–548, 1967.

[24] J. H. Wilkinson. Error analysis of direct methods of matrix inversion. *Journal of the ACM*, 8(3):281–330, 1961.

[25] L. Xu and J. Burke. ASTRAL: An active set ℓ_∞-trust-region algorithm for box-constrained optimization. Technical Report preprint, Department of Mathematics, University of Washington, Seattle, USA, 2007.

A Proofs of Corollaries 3.1-3.3

Proof of Corollary 3.1: The application of Theorem 2.1 and the definition of the p-norm first give

$$\chi^p_{abs} = \min_{(\Delta g, \Delta l, \Delta u) \in \mathcal{P}} \sqrt[p]{\sum_{j=1}^{n} ([\alpha_g |\Delta g| + \alpha_l |\Delta l| + \alpha_u |\Delta u|]_j)^p}.$$

Then, considering the positiveness of the terms and the definition of $\mathcal{P}$, together with the fact that all the components j of the elements of $\mathcal{P}$ are chosen independently between at most two possibilities, we have that

$$\chi^p_{abs} = \sqrt[p]{\sum_{j=1}^{n} \left(\min_{\substack{([\Delta g]_j \,;\, [\Delta l]_j \,;\, [\Delta u]_j), \\ (\Delta g, \Delta l, \Delta u) \in \mathcal{P}}} (\alpha_g |[\Delta g]_j| + \alpha_l |[\Delta l]_j| + \alpha_u |[\Delta u]_j|) \right)^p}.$$

The measure χ^p_{abs} is thus equal to the p-norm of the vector Λ defined by

$$[\Lambda]_j = \min_{\substack{([\Delta g]_j \,;\, [\Delta l]_j \,;\, [\Delta u]_j), \\ (\Delta g, \Delta l, \Delta u) \in \mathcal{P}}} (\alpha_g |[\Delta g]_j| + \alpha_l |[\Delta l]_j| + \alpha_u |[\Delta u]_j|),$$

the value of which will be determined in the second part of the proof. Consider first the case where $[\tilde{x}]_j \in \mathcal{F}_j$. The definition of $\mathcal{P}$ yields $[\Delta g]_j = [\Delta l]_j = [\Delta u]_j = 0$ that for all $j \notin \mathcal{U}$. Otherwise, that is if $j \in \mathcal{U}$, the definition of $\mathcal{P}$ leaves the choice between two solutions (depending on the sign of the gradient) for the minimization corresponding to the j-th component. The first solution is

$$([\Delta g]_j \,;\, [\Delta l]_j \,;\, [\Delta u]_j) = ([-\nabla_x f(\tilde{x})]_j \,;\, 0 \,;\, 0)$$

and $[\Lambda]_j$ is equal to $\alpha_g |[\Delta g]_j| = \alpha_g |[\nabla_x f(\tilde{x})]_j|$ in this case. If the second solution

$$([\Delta g]_j \,;\, [\Delta l]_j \,;\, [\Delta u]_j) = \begin{cases} (0 \,;\, [\tilde{x} - l]_j \,;\, 0) & \text{if } [\nabla_x f(\tilde{x})]_j > 0, \\ (0 \,;\, 0 \,;\, [\tilde{x} - u]_j) & \text{if } [\nabla_x f(\tilde{x})]_j < 0, \end{cases}$$

is preferred, then $[\Lambda]_j$ is equal to either $\alpha_l |[\Delta l]_j| = \alpha_l |[\tilde{x} - l]_j|$ if $[\nabla_x f(\tilde{x})]_j > 0$, or to $\alpha_u |[\Delta u]_j| = \alpha_u |[\tilde{x} - u]_j|$ if $[\nabla_x f(\tilde{x})]_j < 0$. Thus $[\Lambda]_j$ is defined as in (14) when $[\tilde{x}]_j \in \mathcal{F}_j$. Now consider the infeasible case. If $j \notin \mathcal{U}$, the definition of $\mathcal{P}$ implies that $[\Lambda]_j$ is equal to either $(0 \,;\, [\tilde{x} - l]_j \,;\, 0)$ when $[\tilde{x}]_j < [l]_j$, or to $(0 \,;\, 0 \,;\, [\tilde{x} - u]_j)$ when $[\tilde{x}]_j > [u]_j$. Therefore we have that

$$[\Lambda]_j = \begin{cases} \alpha_u |[\tilde{x} - u]_j| & \text{if } [\tilde{x}]_j > [u]_j \text{ and } [\nabla_x f(\tilde{x})]_j \leq 0, \\ \alpha_l |[\tilde{x} - l]_j| & \text{if } [\tilde{x}]_j < [l]_j \text{ and } [\nabla_x f(\tilde{x})]_j \geq 0. \end{cases}$$

If $j \in \mathcal{U}$ and $[\nabla_x f(\tilde{x})]_j > 0$, notice that the infeasibility automatically implies $[\tilde{x}]_j > [u]_j$. In that case, the definition of $\mathcal{P}$ lets the choice between two solutions: $([-\nabla_x f(\tilde{x})]_j \,;\, 0 \,;\, [\tilde{x} - u]_j)$ and $(0 \,;\, [\tilde{x} - l]_j \,;\, [\tilde{x} - u]_j)$, leading to a value of the objective function equal to

$$[\Lambda]_j = \min\{\alpha_g |[\nabla_x f(\tilde{x})]_j|, \alpha_l |[\tilde{x} - l]_j|\} + \alpha_u |[\tilde{x} - u]_j|.$$

Similarly, the two solutions in $\mathcal{P}$ when $j \in \mathcal{U}$ and $[\nabla_x f(\tilde{x})]_j < 0$ correspond to a situation where $[\tilde{x}]_j < [l]_j$, ane give that

$$[\Lambda]_j = \min\{\alpha_g|[\nabla_x f(\tilde{x})]_j|, \alpha_u|[\tilde{x}-u]_j|\} + \alpha_l|[\tilde{x}-l]_j|.$$

Gathering the values obtained in the different cases, we finally obtain that $\chi^p_{abs} = ||\Lambda||_p$, with $[\Lambda]_j$ defined by (14). □

Proof of Corollary 3.2: First notice that the definition of $\mathcal{P}$ implies that $\#\mathcal{P}$, the cardinal of $\mathcal{P}$, is a finite number since it is smaller than 2^n (because we have the choice between at most two solutions for each $j = 1, ..., n$). As a consequence, Theorem 2.1 implies that

$$\begin{aligned}
\chi^\infty_{abs} &= \min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{P}} ||\alpha_g|\Delta g| + \alpha_l|\Delta l| + \alpha_u|\Delta u|\ ||_\infty \\
&= \min\left\{||(\Delta g_1, \Delta l_1, \Delta u_1)||_\infty, ..., ||(\Delta g_{\#\mathcal{P}}, \Delta l_{\#\mathcal{P}}, \Delta u_{\#\mathcal{P}})||_\infty\right\} \\
&= \min\left\{\lim_{p\to\infty} ||(\Delta g_1, \Delta l_1, \Delta u_1)||_p, ..., \lim_{p\to\infty} ||(\Delta g_{\#\mathcal{P}}, \Delta l_{\#\mathcal{P}}, \Delta u_{\#\mathcal{P}})||_p\right\}
\end{aligned}$$

where we used the identity $\lim_{p\to\infty} ||.||_p = ||.||_\infty$. The fact that $\#\mathcal{P}$ is finite now allows us to write that

$$\begin{aligned}
\chi^\infty_{abs} &= \lim_{p\to\infty} \min\{||(\Delta g_1, \Delta l_1, \Delta u_1)||_p, ..., ||(\Delta g_{\#\mathcal{P}}, \Delta l_{\#\mathcal{P}}, \Delta u_{\#\mathcal{P}})||_p\} \\
&= \lim_{p\to\infty} \min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{D}} ||\alpha_g|\Delta g| + \alpha_l|\Delta l| + \alpha_u|\Delta u|\ ||_p,
\end{aligned}$$

where we used Theorem 2.1 to derive the last equality. Finally, Corollary 3.1 then gives that

$$\chi^\infty_{abs} = \lim_{p\to\infty} ||\Lambda||_p = ||\Lambda||_\infty,$$

where Λ is defined by (14). □

Proof of Corollary 3.3: We prove this result by showing that $\chi^1_{ws} = \chi^p_{abs}$ where $p = 1$. Applying the definitions of χ^1_{ws} and of the 1-norm, we first obtain that

$$\begin{aligned}
\chi^1_{ws} &= \min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{D}} \alpha_g||\Delta g||_1 + \alpha_l||\Delta l||_1 + \alpha_u||\Delta u||_1 \\
&= \min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{D}} \alpha_g \sum_{j=1}^n |[\Delta g]_j| + \alpha_l \sum_{j=1}^n |[\Delta l]_j| + \alpha_u \sum_{j=1}^n |[\Delta u]_j| \\
&= \min_{(\Delta g, \Delta l, \Delta u)\in\mathcal{D}} \sum_{j=1}^n [\alpha_g|\Delta g| + \alpha_l|\Delta l| + \alpha_u|\Delta u|]_j.
\end{aligned}$$

Then, the positiveness of the terms and the definitions of χ^p_{abs} and of the 1-norm give that

$$\begin{aligned}\chi_{ws} &= \min_{(\Delta g,\Delta l,\Delta u)\in\mathcal{D}} \sum_{j=1}^{n} |[\alpha_g|\Delta g| + \alpha_l|\Delta l| + \alpha_u|\Delta u|]_j| \\ &= \min_{(\Delta g,\Delta l,\Delta u)\in\mathcal{D}} ||\alpha_g|\Delta g| + \alpha_l|\Delta l| + \alpha_u|\Delta u|\,||_1 \\ &= \chi^p_{abs}\end{aligned}$$

with $p = 1$. We conclude the proof by applying Corollary 3.1. □