Global Convergence of Trust-Region Interior-Point Algorithms for Infinite-Dimensional Nonconvex Minimization Subject to Pointwise Bounds

Michael Ulbrich
Stefan Ulbrich
Matthias Heinkenschloss

CRPC-TR97692
March 1997

Center for Research on Parallel Computation
Rice University
6100 South Main Street
CRPC - MS 41
Houston, TX 77005

submitted April 1997

GLOBAL CONVERGENCE OF TRUST-REGION INTERIOR-POINT ALGORITHMS FOR INFINITE-DIMENSIONAL NONCONVEX MINIMIZATION SUBJECT TO POINTWISE BOUNDS *

MICHAEL ULBRICH†, STEFAN ULBRICH‡, AND MATTHIAS HEINKENSCHLOSS §

Abstract. A class of interior–point trust–region algorithms for infinite–dimensional nonlinear optimization subject to pointwise bounds in L^p-Banach spaces, $2 \leq p \leq \infty$, is formulated and analyzed. The problem formulation is motivated by optimal control problems with L^p-controls and pointwise control constraints. The interior–point trust–region algorithms are generalizations of those recently introduced by Coleman and Li (*SIAM J. Optim.*, 6 (1996), pp. 418–445) for finite–dimensional problems. Many of the generalizations derived in this paper are also important in the finite–dimensional context. They lead to a better understanding of the method and to considerable improvements in their performance. All first– and second–order global convergence results known for trust–region methods in the finite–dimensional setting are extended to the infinite–dimensional framework of this paper.

Key words. Infinite-dimensional optimization, bound constraints, affine scaling, interior-point algorithms, trust-region methods, global convergence, optimal control, nonlinear programming.

AMS subject classifications. 49M37, 65K05, 90C30, 90C48

1. Introduction. This paper is concerned with the development and analysis of a class of interior–point trust–region algorithms for the solution of the following infinite–dimensional nonlinear programming problem:

$$\text{(P)} \qquad \begin{aligned} &\text{minimize} \quad f(u) \\ &\text{subject to} \quad u \in \mathcal{B} \stackrel{\text{def}}{=} \{u \in U \,:\, a(x) \leq u(x) \leq b(x),\ x \in \Omega\}\,. \end{aligned}$$

Here $\Omega \subset \mathbb{R}^n$ is a domain with positive and finite Lebesgue measure $0 < \mu(\Omega) < \infty$. Moreover,

$$U \stackrel{\text{def}}{=} L^p(\Omega), \quad 2 \leq p \leq \infty,$$

denotes the usual Banach space of real-valued measurable functions, and the objective function $f : \mathcal{D} \longrightarrow \mathbb{R}$ is continuous on an open neighborhood $\mathcal{D} \subset U$ of $\mathcal{B}$. All pointwise statements on measurable functions are meant to hold μ-almost everywhere. The lower and upper bound functions $a, b \in V$,

$$V \stackrel{\text{def}}{=} L^\infty(\Omega),$$

are assumed to have a distance of at least $\nu > 0$ from each other. More precisely,

$$b(x) - a(x) \geq \nu \quad \text{ on } \Omega.$$

* This version was generated June 2, 1997.

† Institut für Angewandte Mathematik und Statistik, Technische Universität München, D-80290 München, Germany, E-Mail: mulbrich@statistik.tu-muenchen.de. This author was supported by the DFG under Grant Ul157/1–1 and by the NATO under Grant CRG 960945.

‡ Institut für Angewandte Mathematik und Statistik, Technische Universität München, D-80290 München, Germany, E-Mail: sulbrich@statistik.tu-muenchen.de. This author was supported by the DFG under Grant Ul158/1–1 and by the NATO under Grant CRG 960945.

§ Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005–1892, USA, E-Mail: heinken@rice.edu. This author was supported by the NSF under Grant DMS–9403699, by the DoE under Grant DE-FG03-95ER25257, the AFSOR under Grant F49620–96–1–0329, and the NATO under Grant CRG 960945.

Problems of type (P) arise for instance when the black-box approach is applied to optimal control problems with bound–constrained L^p-control. See, e.g., the problems studied by Burger, Pogu [3], Kelley, Sachs [14], and Tian, Dunn [20].

The algorithms in this paper are extensions of the interior–point trust–region algorithms for bound constrained problems in $\mathbb{R}^N$ introduced by Coleman and Li [6]. Algorithmic enhancements of these methods have been proposed and analyzed in the finite–dimensional context in Branch, Coleman, Li [2], Coleman, Li [5], and Dennis, Vicente [11]. Dennis, Heinkenschloss, Vicente [10], and Heinkenschloss, Vicente [13] extend these methods to solve a class of finite–dimensional constrained optimization problems with bound constraints on parts of the variables. See also Vicente [23]. The interior–point trust–region methods in [6] are based on the reformulation of the Karush–Kuhn–Tucker (KKT) necessary optimality conditions as a system of nonlinear equations using a diagonal matrix D. This affine scaling matrix is computed using the sign of the gradient components and the distance of the variables to the bounds. See § 2. The nonlinear system is then solved by an affine–scaling interior–point method in which the trust–region is scaled by $D^{-\frac{1}{2}}$. These methods enjoy strong theoretical convergence properties as well as a good numerical behavior. The latter is documented in [2], [6], [10], [11] where these algorithms have been applied to various standard finite–dimensional test problems and to some discretized optimal control problems.

The present work is motivated by the application of interior–point trust–region algorithms to optimal control problems with bounds on the controls. Even though the numerical solution of these problems requires a discretization and allows the application of the previously mentioned algorithms to the resulting finite–dimensional problems, it is known that the infinite–dimensional setting dominates the convergence behavior if the discretization becomes sufficiently small. If the algorithm can be applied to the infinite–dimensional problem and convergence can be proven in the infinite–dimensional setting, asymptotically the same convergence behavior can be expected if the algorithm is applied to the finite–dimensional discretized problems. Otherwise, the convergence behavior might – and usually does – deteriorate fast as the discretization is refined.

In the present context, the formulation of the interior–point trust–region algorithms for the solution of the infinite–dimensional problem (P) requires a careful statement of the problem and of the requirements on the function f. This will be done in § 3. The infinite–dimensional problem setting in this paper is similar to the ones in [12], [14], [15], [20]. The general structure of the interior–point trust–region algorithms presented here is closely related to the finite–dimensional algorithms in [6]. However, the statement and analysis of the algorithm in the infinite–dimensional context is more delicate and has motivated generalizations and extensions which are also relevant in the finite–dimensional context. The analysis performed in this paper allows for a greater variety of choices for the affine scaling matrix and the scaling of the trust–region than those presented previously in [6], [11]. Our convergence analysis is more comprehensive than the ones in [5], [6], [11], [23]. In particular, we adapt techniques proposed in Shultz, Schnabel, and Byrd [18] to prove that under mild assumptions every accumulation point satisfies the second–order necessary optimality conditions. Moreover, the convergence results proven in this paper extend all the finite–dimensional ones stated in [17], [18], [19] to our infinite–dimensional context with bound constraints. In the follow up paper [22] we present a local convergence analysis of a superlinearly convergent affine–scaling interior–point Newton method

which is based on equation (13) and prove under appropriate assumptions that in a neighborhood of the solution the generated trial steps are accepted by our trust–region algorithms. There a projection onto the set $\mathcal{B}$ will be used in the computation of trial steps. This extension to the finite–dimensional method, which was originally motivated by the function space framework, has also led to significant improvements of the finite–dimensional algorithm applied to some standard test problems, not obtained from the discretization of optimal control problems. See [22].

Trust–region methods for infinite–dimensional problems like (P) have also been investigated by Kelley, Sachs [15] and Toint [21]. In both papers the constraints are handled by projections. The paper [21] considers trust–region algorithms for minimization on closed convex bounded sets in Hilbert space. They are extensions of the finite–dimensional algorithms by Conn, Gould, Toint [7]. It is proven that the projected gradient converges to zero. A comprehensive finite–dimensional analysis of trust–region methods closely related to those introduced by Toint can be found in Burke, Moré, Toraldo [4]. In contrast to the results in [21], our convergence analysis is also applicable to objective functions that are merely differentiable on a Banach space $L^p(\Omega)$, $p \in (2, \infty]$, which reduces the differentiability requirements substantially compared to the L^2-Hilbert space framework. Furthermore, for the problem class under consideration our convergence results are more comprehensive than the ones in [21]. The infinite–dimensional setting used in [15] fits into the framework of this paper, but is more restrictive. The formulation of their algorithm depends on the presence of a penalty term $\alpha \int_\Omega u^2(x)dx$ in the objective function f and they assume that $\Omega \subset \mathbb{R}$ is an interval. Their algorithm also includes a 'post smoothing' step, which is performed after the trust–region step is computed. The presence of the post smoothing step ensures that existing local convergence results can be applied. Such a 'post smoothing' is not needed in the global analysis of this paper.

We introduce the following notations. $\mathcal{L}(X,Y)$ is the space of linear bounded operators from a Banach space X into a Banach space Y. By $\|\cdot\|_q$ we denote the norm of the Lebesgue space $L^q(\Omega)$, $1 \le q \le \infty$, and we write $(\cdot,\cdot)_2$ for the inner product of the Hilbert space $H \stackrel{\text{def}}{=} L^2(\Omega)$. For $(v,w) \in (L^q(\Omega), L^q(\Omega)^*)$, with $L^q(\Omega)^*$ denoting the dual space of $L^q(\Omega)$, we use the canonical dual pairing $\langle v, w\rangle \stackrel{\text{def}}{=} \int_\Omega v(x)w(x)\,dx$, for which, if $q < \infty$, the dual space $L^q(\Omega)^*$ is given by $L^{q'}(\Omega)$, $1/q + 1/q' = 1$ (in the case $q = 1$ this means $q' = \infty$). Especially, if $q = 2$, we have $L^2(\Omega)^* = L^2(\Omega)$ and $\langle\cdot,\cdot\rangle$ coincides with $(\cdot,\cdot)_2$.

Finally, we set $U' \stackrel{\text{def}}{=} L^{p'}(\Omega)$, $1/p + 1/p' = 1$, which is the same as U^*, if $p < \infty$. Moreover, it is easily seen that $w \longmapsto \langle\cdot, w\rangle$ defines a linear norm-preserving injection from $L^1(\Omega)$ into $L^\infty(\Omega)^*$. Therefore, we may always interpret U' as subspace of U^*. As a consequence of Lemma 5.1 we get the following chain of continuous imbeddings:

$$V \hookrightarrow U \hookrightarrow H = H^* \hookrightarrow U' \hookrightarrow U^* \hookrightarrow V^*.$$

Throughout we will work with differentiability in the Fréchet–sense. We write $g(u) \stackrel{\text{def}}{=} \nabla f(u) \in U^*$ for the gradient and $\nabla^2 f(u) \in \mathcal{L}(U, U^*)$ for the second derivative of f at $u \in \mathcal{B}$ if they exist. The $\|\cdot\|_\infty$-interior of $\mathcal{B}$ is denoted by $\mathcal{B}^\circ$:

$$\mathcal{B}^\circ \stackrel{\text{def}}{=} \bigcup_{\delta>0} \mathcal{B}_\delta\ , \quad \mathcal{B}_\delta \stackrel{\text{def}}{=} \{u \in U\ :\ a(x) + \delta \le u(x) \le b(x) - \delta,\ x \in \Omega\}\,.$$

We often write $f_k, g_k, \ldots$ for $f(u_k), g(u_k), \ldots$

This paper is organized as follows. In the next section we review the basics of the finite–dimensional interior–point trust–region algorithms in [6] and use this to motivate the infinite–dimensional setting applied in this paper. In § 3 we formulate the necessary optimality conditions in the framework needed for the interior–point trust–region algorithms. The interior–point trust–region algorithms are introduced in § 4. Some basic technical results are collected in § 5. The main convergence results are given in § 6, which concerns the global convergence to points satisfying the first–order necessary optimality conditions, and in § 7, which concerns the global convergence to points satisfying the second–order necessary optimality conditions. These convergence results extend all the known convergence results for trust–region methods in finite dimensions to the infinite–dimensional setting of this paper. The local convergence analysis of these algorithms is given in the follow up paper [22], which also contains numerical examples illustrating the theoretical findings of this paper.

2. Review of the finite–dimensional algorithm and infinite–dimensional problem setting. We briefly review the main ingredients of the affine–scaling interior–point trust–region method introduced in [6]. We refer to that paper for more details. The algorithm solves finite–dimensional problems of the form

$$(\mathrm{P_N}) \qquad \begin{aligned} &\text{minimize} \quad f(u) \\ &\text{subject to} \quad u \in \mathcal{B}_N \stackrel{\text{def}}{=} \left\{ u \in \mathbb{R}^N \,:\, a \le u \le b \right\}, \end{aligned}$$

where $f : \mathbb{R}^N \longrightarrow \mathbb{R}$ is a twice continuously differentiable function and $a < b$ are given vectors in $\mathbb{R}^N$. (One can allow components of a and b to be $-\infty$ or ∞, respectively. This is excluded here to simplify the presentation.) Inequalities are understood component wise.

The necessary optimality conditions for $(\mathrm{P_N})$ are given by

$$\begin{aligned} &\nabla f(\bar{u}) - \bar{\mu}^a + \bar{\mu}^b = 0, \\ &a \le \bar{u} \le b, \\ &(\bar{u} - a)^T \bar{\mu}^a + (b - \bar{u})^T \bar{\mu}^b = 0, \\ &\bar{\mu}^a \ge 0,\ \bar{\mu}^b \ge 0. \end{aligned}$$

With the diagonal matrix defined by

$$\left(D(u)\right)_{ii} \stackrel{\text{def}}{=} \begin{cases} (b-u)_i & \text{if} \quad (\nabla f(x))_i < 0, \\ (u-a)_i & \text{if} \quad (\nabla f(x))_i \ge 0, \end{cases} \tag{1}$$

for $i = 1, \ldots, N$, the necessary optimality conditions can be rewritten as

$$\begin{aligned} D(\bar{u})^r \nabla f(\bar{u}) &= 0, \\ a \le \bar{u} &\le b. \end{aligned} \tag{2}$$

where the power $r > 0$ is applied to the diagonal elements. This form of the necessary optimality conditions – we choose $r = 1$ – can now be solved using Newton's method. The i-th component of the function $D(u)$ is differentiable except at points where $(\nabla f(u))_i = 0$. However, this lack of smoothness is benign since $D(u)$ is multiplied by $\nabla f(u)$. One can use

$$D(u)\nabla^2 f(u) + \mathrm{diag}(\nabla f(u)) J(u) \tag{3}$$

where $J(u)$ is the diagonal matrix

$$\Big(J(u)\Big)_{ii} \stackrel{\text{def}}{=} \begin{cases} -1 & \text{if} \quad (\nabla f(x))_i < 0, \\ 1 & \text{if} \quad (\nabla f(x))_i > 0, \\ 0 & \text{else}, \end{cases}$$

as the approximate derivative of $D(u)\nabla f(u)$. After symmetrization, one obtains

$$\hat{M}(u) = D(u)^{1/2}\nabla^2 f(u) D(u)^{1/2} + \operatorname{diag}(\nabla f(u))J(u). \tag{4}$$

One can show that the standard second–order necessary optimality conditions are equivalent to (2) and the positive semi–definiteness of $\hat{M}(\bar{u})$. The standard second–order sufficient optimality conditions are equivalent to (2) and the positive definiteness of $\hat{M}(\bar{u})$.

A point satisfying the necessary optimality conditions (2) is now computed using the iteration $u_{k+1} = u_k + s_k$, where for a given u_k with $a < u_k < b$, the trial step $s_k = D_k^{1/2}\hat{s}_k$ satisfies $a < u_k + s_k < b$, and $\hat{s}_k$ is an approximate solution of

$$\min \hat{\psi}_k(\hat{s}) \quad \text{subject to} \quad \|\hat{s}\|_2 \le \Delta_k \ , \quad u_k + D_k^{1/2}\hat{s} \in \mathcal{B}_N \tag{5}$$

with $\hat{\psi}_k(\hat{s}) \stackrel{\text{def}}{=} \hat{g}_k^T\hat{s} + \frac{1}{2}\hat{s}^T\hat{M}_k\hat{s}$, $\hat{g}_k \stackrel{\text{def}}{=} D_k^{1/2}\nabla f_k$. The trust–region radius Δ_k is updated from iteration to iteration in the usual fashion. In (5) the Hessian $\nabla^2 f(u_k)$ might be replaced by a symmetric approximation B_k. If the approximate solution $\hat{s}_k$ of (5) satisfies a fraction of Cauchy decrease condition

$$\begin{aligned} &\hat{\psi}_k(\hat{s}_k) < \beta \min\Big\{\hat{\psi}_k(\hat{s}) \ : \ \hat{s} = t\hat{g}_k,\ t \le 0,\ \|\hat{s}\|_2 \le \Delta_k,\ u_k + D_k^{1/2}\hat{s} \in \mathcal{B}_N\Big\}, \\ &\|\hat{s}_k\|_2 \le \beta_0\Delta_k, \end{aligned} \tag{6}$$

then under appropriate, standard conditions one can show the basic trust–region convergence result

$$\liminf_{k\to\infty} \|D(u_k)^{1/2}\nabla f(u_k)\| = 0.$$

Stronger convergence results can be proven if the assumptions on the function f and on the step computation $\hat{s}_k$ are strengthened appropriately. See [6] and [5], [11].

Coleman and Li [6] show that close to nondegenerate KKT–points one obtains trial steps $\hat{s}_k$ which meet these requirements if one first computes an approximate solution of (5) ignoring the bound constraints and then satisfies the interior–point condition $a < u_k + s_k < b$ by a step-size rule. A careful analysis of the proofs in [6] unveils that the same holds true for nearly arbitrary trust–region scalings. It becomes apparent that the crucial role of the affine scaling does *not* consist in the scaling of the trust–region but rather in leading to the additional term $\operatorname{diag}(\nabla f_k)J_k$ in the Hessian $\hat{M}_k$ of $\hat{\psi}_k$. Near nondegenerate KKT–points this positive semi–definite diagonal–matrix shapes the level sets of $\hat{\psi}_k$ in such a way that all 'bad' directions $\hat{s}$ which allow only for small step-sizes to the boundary of the box cannot minimize $\hat{\psi}_k$ on any reasonable trust–region. The trust-region scaling in (5), (6) tends to equilibrate the distance of the origin to the bounding box constraints $\{\hat{s} : u_k + D_k^{1/2}\hat{s} \in \mathcal{B}_N\}$. However, for this feature the equivalence of 2- and ∞-norm is indispensable and thus it does not carry over to our infinite–dimensional framework. In fact, in the infinite–dimensional

setting the affine–scaled trust–region $\{\|\hat{s}\|_p \leq \Delta_k\}$ no longer enjoys the property of reflecting the distance to the bounding box constraints. Therefore we will allow for a very general class of trust–region scalings in our analysis. See also [11]. Since, as mentioned above, the term $\mathrm{diag}(\nabla f_k)J_k$ in the Hessian $\hat{M}_k$ plays the crucial role in this affine–scaling interior–point all convergence results in [6] remain valid. It is also worth mentioning that in our context an approximate solution $\hat{s}_k$ of (5) satisfying (6) can be easily obtained by applying any descent method which starts minimization at $\hat{s} = 0$ along the steepest descent direction $-\hat{g}_k$. Moreover, we show in [22] that near an optimizer satisfying suitable sufficiency conditions admissible trial steps can be obtained from unconstrained minimizers of $\hat{\psi}_k$ by pointwise projection onto $\mathcal{B}$. Here our flexibility in the choice of the trust–region scaling will prove to be valuable.

The finite–dimensional convergence analysis heavily relies on the equivalency of norms in $\mathbb{R}^N$. This is for example used to obtain pointwise ($\|\cdot\|_\infty$) estimates from $\|\cdot\|_2$ estimates. In the infinite–dimensional context the formulation of the algorithm and the proof of its convergence is more delicate.

We will make use of the following Assumptions:

(A1) $f : \mathcal{D} \longrightarrow \mathbb{R}$ is differentiable on $\mathcal{D}$ with g mapping $\mathcal{B} \subset U$ continuously into U'.

(A2) The gradient g satisfies $g(\mathcal{B}) \subset V$.

(A3) There exists $c_1 > 0$ such that $\|g(u)\|_\infty \leq c_1$ for all $u \in \mathcal{B}$.

(A4) f is twice continuously differentiable on $\mathcal{D}$. If $p = \infty$ then $\nabla^2 f(u) \in \mathcal{L}(U, U')$ for all $u \in \mathcal{B}$, and if $(h_k) \subset V$ converges to zero in all spaces $L^q(\Omega)$, $1 \leq q < \infty$, then $\nabla^2 f(u)h_k$ tends to zero in U'.

For $p \in [2, \infty)$ the assumptions (A1) and (A4) simply say that f is continuously Fréchet–differentiable or that f is twice continuously Fréchet–differentiable, respectively. If $p = \infty$, then the requirements that $g(u), \nabla^2 f(u)h \in U' = L^1(\Omega) \neq U^*$ for $u \in \mathcal{B}$, $h \in V$ is a further condition. It allows us to use estimates like $\langle v, g(u)\rangle \leq \|g(u)\|_{p'}\|v\|_p$ for $p \in [2, \infty)$ and $p = \infty$. Moreover, since on $L^1(\Omega)$ the L^1- and $(L^\infty)^*$-norms coincide, assumption (A4) implies that $\nabla^2 f : \mathcal{B} \subset U \longrightarrow \mathcal{L}(U, U')$ is continuous also for $p = \infty$. Finally, (A1) ensures that the gradient $g(u)$ is always at least an L^1-function which will be essential for many reasons, e.g. to allow the definition of a function space analogue for the scaling matrix D.

The assumption (A2) is motivated by the choice of the scaling matrix D and the fraction of Cauchy decrease condition (6) in the finite–dimensional case. The infinite–dimensional analogue $d(u)$ of the diagonal scaling matrix $D(u)$ will be a function. Given the definition (1) of $D(u)$ it is to be expected that $d(u) \in L^\infty(\Omega)$ if $a < u < b$. If $g(u) \in L^{p'}(\Omega)$, then $d(u)^{1/2}g(u) \in L^{p'}(\Omega)$. Hence, the candidate for the Cauchy step satisfies $\hat{s}^c = -td(u)^{1/2}g(u) \in L^{p'}(\Omega)$. Since $p' \neq \infty$, one will in general not be able to find a scaling $\tau > 0$ so that $a < u + \tau d(u)^{1/2}\hat{s}^c < b$. The assumption (A2) assures that $\hat{s}^c = -td(u)g(u) \in V$. The uniform boundedness assumption (A3) is, e.g., used to derive the important estimate (26). We point out that in (A3) the uniform bound on $g(u)$ has to hold only for $u \in \mathcal{B}$ which is a bounded set in $L^\infty(\Omega)$.

The conditions (A1)–(A4) limit the optimal control problems that fit into this framework. However, a large and important class of optimal control problems with L^p-controls satisfy these conditions. For example, the conditions imposed in [12, p. 1270], [20, p. 517] to study the convergence of the gradient projection method

imply our assumptions (A1), (A2), and (A4). The assumption (A3) can be enforced by additional requirements on the functions ϕ and S used in [12], [20]. The boundary control problems for a nonlinear heat equation in [3] and in [14], [15] also satisfy the assumptions (A1)–(A4). See [22].

3. Necessary optimality conditions and affine scaling. The problem under consideration belongs to the class of cone constrained optimization problems in Banach space for which optimality conditions are available (cf. [16]). But we believe that for our particular problem an elementary derivation of the necessary optimality conditions for problem (P) not only is simpler but also more transparent than the application of the general theory. This derivation also helps us to motivate the choice of the affine scaling which is used to reformulate the optimality condition and which is the basis for the interior–point method.

3.1. First–order necessary conditions. The first–order necessary optimality conditions in Theorem 3.1 are completely analogous to those for finite–dimensional problems with simple bounds (cf. § 2, [6]). We only have to replace coordinatewise by pointwise statements and to ensure that the gradient $g(\bar{u})$ is a measurable function.

THEOREM 3.1 (FIRST–ORDER NECESSARY OPTIMALITY CONDITIONS). *Let $\bar{u}$ be a local minimizer of problem* (P) *and assume that f is differentiable at $\bar{u}$ with $g(\bar{u}) \in U'$. Then*

(O1) $\bar{u} \in \mathcal{B}$,

(O2) $g(\bar{u})(x) \begin{cases} = 0 & \text{for } x \in \Omega \text{ with } a(x) < \bar{u}(x) < b(x), \\ \geq 0 & \text{for } x \in \Omega \text{ with } \bar{u}(x) = a(x), \\ \leq 0 & \text{for } x \in \Omega \text{ with } \bar{u}(x) = b(x) \end{cases}$

are satisfied.

Proof. Condition (O1) is trivially satisfied. To verify (O2), define

$$A_- = \{x \in \Omega : \bar{u}(x) = a(x),\ g(\bar{u})(x) < 0\}\ ,\quad A_-^k = \{x \in A_- : g(\bar{u})(x) \leq -1/k\}\,,$$

and assume that A_- has positive measure $\mu(A_-) > \varepsilon > 0$. Since μ is continuous from below and $A_-^k \uparrow A_-$, there exists $l > 0$ with $\mu(A_-^l) \geq \varepsilon$. This yields a contradiction, because $\bar{u} + \tau s \in \mathcal{B}$, $s = \chi_{A_-}(b - a)$, for $0 \leq \tau \leq 1$, and

$$\frac{d}{d\tau} f(\bar{u} + \tau s)|_{\tau=0} = \langle s, g(\bar{u})\rangle \leq -\frac{\varepsilon\nu}{l} < 0.$$

Hence we must have $\mu(A_-) = 0$. In the same way we can show that $\mu(A_+) = 0$ for $A_+ = \{x \in \Omega : \bar{u}(x) = b(x),\ g(\bar{u})(x) > 0\}$. Finally, we look at

$$I = \{x \in \Omega : a(x) < \bar{u}(x) < b(x),\ g(\bar{u})(x) \neq 0\}\,.$$

Assume that $\mu(I) > \varepsilon > 0$. Since $I^k \uparrow I$ with

$$I^k = \{x \in \Omega : a(x) + 1/k \leq \bar{u}(x) \leq b(x) - 1/k,\ |g(\bar{u})(x)| \geq 1/k\}\,,$$

we can find $l > 0$ with $\mu(I^l) \geq \varepsilon$ and obtain for

$$s = -\chi_{I^l} \frac{g(\bar{u})}{|g(\bar{u})|}$$

that $\bar{u} + \tau s \in \mathcal{B}$, $0 \leq \tau \leq 1/l$, and

$$\frac{d}{d\tau} f(\bar{u} + \tau s)|_{\tau=0} = \langle s, g(\bar{u}) \rangle \leq -\frac{\varepsilon}{l} < 0,$$

a contradiction to the local optimality of $\bar{u}$. Hence $\mu(I) = \mu(A_-) = \mu(A_+) = 0$ which means that (O2) holds. □

3.2. Affine scaling. Let assumption (A1) hold. Our algorithm will be based on the following equivalent *affine–scaling* formulation of (O2):

$$d^r(\bar{u}) g(\bar{u}) = 0 \tag{7}$$

where $r > 0$ is arbitrary and $d(u) \in V$, $u \in \mathcal{B}$, is a scaling function which is assumed to satisfy

$$d(u)(x) \begin{cases} = 0 & \text{if } u(x) = a(x) \text{ and } g(u)(x) \geq 0, \\ = 0 & \text{if } u(x) = b(x) \text{ and } g(u)(x) \leq 0, \\ > 0 & \text{else,} \end{cases} \tag{8}$$

for all $x \in \Omega$. The equivalence of (O2) and (7) will be stated and proved in Lemma 3.2. Before we do this, we give two examples of proper choices for d. The first choice $d = d_{\mathrm{I}}$ is motivated by the scaling matrices used in [6] (see (1)). Except for points x with $g(u)(x) = 0$ it equals those used in [6] and [11]:

$$d_{\mathrm{I}}(u)(x) \stackrel{\text{def}}{=} \begin{cases} u(x) - a(x) & \text{if } g(u)(x) > 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } u(x) - a(x) \leq b(x) - u(x), \\ b(x) - u(x) & \text{if } g(u)(x) < 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } b(x) - u(x) < u(x) - a(x). \end{cases} \tag{9}$$

The slight modification in comparison to (1) will enable us to establish the valuable relation (16) without a nondegeneracy assumption.

While the global analysis could be carried out entirely with this choice, the discontinuous response of $d(u)(x)$ to sign changes of $g(u)(x)$ raises difficulties for the design of superlinearly convergent algorithms in infinite dimensions. These can be circumvented by the choice $d = d_{\Pi}$, where

$$d_{\Pi}(u)(x) \stackrel{\text{def}}{=} \begin{cases} \min\{|g(u)(x)|, c(x)\} & \text{if } -g(u)(x) > u(x) - a(x) \\ & \text{and } u(x) - a(x) \leq b(x) - u(x), \\ \min\{|g(u)(x)|, c(x)\} & \text{if } g(u)(x) > b(x) - u(x) \\ & \text{and } b(x) - u(x) \leq u(x) - a(x), \\ \min\{u(x) - a(x), & \\ \quad b(x) - u(x), c(x)\} & \text{else.} \end{cases} \tag{10}$$

Here $c : x \in \Omega \longmapsto \min\{\zeta(b(x) - a(x)), \kappa\}$ with $\zeta \in (0, 1/2]$ and $\kappa \geq 1$.

It is easily seen that $d = d_{\mathrm{I}}$ and $d = d_{\Pi}$ both satisfy (8). An illustrative example for the improved smoothness of the scaling function $d_{\Pi}(u)$ will be given in § 4.1.

LEMMA 3.2. *Let* (A1) *hold and* $\bar{u} \in \mathcal{B}$. *Then* (O2) *is equivalent to* (7) *for all* $r > 0$ *and all* d *satisfying* (8).

Proof. Since d^r, $r > 0$, also satisfies (8), we may restrict ourselves to the case $r = 1$. First assume that (O2) holds. For all $x \in \Omega$ with $g(\bar{u})(x) = 0$ we also have

$d(\bar{u})(x)g(\bar{u})(x) = 0$. If $g(\bar{u})(x) > 0$, then by (O2) $\bar{u}(x) = a(x)$ and if $g(\bar{u})(x) < 0$ then $\bar{u}(x) = b(x)$. In both cases $d(\bar{u})(x) = 0$ and hence $d(\bar{u})(x)g(\bar{u})(x) = 0$. On the other hand, let $d(\bar{u})g(\bar{u}) = 0$ hold. For all $x \in \Omega$ with $a(x) < \bar{u}(x) < b(x)$ we have $d(\bar{u})(x) > 0$ which implies $g(\bar{u})(x) = 0$. For all $x \in \Omega$ with $\bar{u}(x) = a(x)$ we obtain $g(\bar{u})(x) \geq 0$ since $g(\bar{u})(x) < 0$ would yield the contradiction $d(\bar{u})(x) > 0$. Analogously, we see that $g(\bar{u})(x) \leq 0$ for all $x \in \Omega$ with $\bar{u}(x) = b(x)$. Therefore, (O2) holds. □

3.3. Second–order conditions. If assumption (A4) holds, we can derive second–order conditions which are satisfied at all local solutions of (P). These are also analogous to the well known conditions for finite–dimensional problems.

THEOREM 3.3 (SECOND–ORDER NECESSARY OPTIMALITY CONDITIONS). *Let* (A4) *be satisfied and* $g(\bar{u}) \in U'$ *hold at the local minimizer* $\bar{u}$ *of problem* (P). *Then* (O1), (O2) *and*

(O3) $\langle s, \nabla^2 f(\bar{u})s\rangle \geq 0$ *for all* $s \in T(\mathcal{B}, \bar{u})$

are satisfied, where

$$T(\mathcal{B}, \bar{u}) \stackrel{\text{def}}{=} \{s \in V \,:\, s(x) = 0 \text{ for all } x \in \Omega \text{ with } \bar{u}(x) \in \{a(x), b(x)\}\}$$

denotes the tangent space of the active constraints.

Proof. Let the assumptions hold. As shown in Theorem 3.1, (O1) and (O2) are satisfied. In particular, we have that $sg(\bar{u}) = 0$ for all $s \in T(\mathcal{B}, \bar{u})$. Now assume the existence of $s \in T(\mathcal{B}, \bar{u})$ and $\varepsilon > 0$ with $\langle s, \nabla^2 f(\bar{u})s\rangle < -\varepsilon$. Let $I = \{x \in \Omega \,:\, a(x) < \bar{u}(x) < b(x)\}$,

$$I_k = \{x \in \Omega \,:\, a(x) + 1/k \leq \bar{u}(x) \leq b(x) - 1/k\} \tag{11}$$

and define restrictions $s^k = \chi_{I_k} s \in V$. Since $I_k \uparrow I$ and $s = \chi_I s$, we get $\|s^k - s\|_q^q \leq \mu(I \setminus I_k)\|s\|_\infty^q$. Hence, the restrictions s^k converge to s in all spaces $L^q(\Omega)$, $1 \leq q < \infty$. Therefore, $\nabla^2 f(\bar{u})(s - s^k)$ tends to zero in U' by (A4) and, using the symmetry of $\nabla^2 f(\bar{u})$,

$$\begin{aligned}\langle s^k, \nabla^2 f(\bar{u})s^k\rangle &= \langle s, \nabla^2 f(\bar{u})s\rangle - \langle s + s^k, \nabla^2 f(\bar{u})(s - s^k)\rangle \\ &< 2\,\|\nabla^2 f(\bar{u})(s - s^k)\|_{p'}\|s\|_p - \varepsilon \\ &\leq -\varepsilon/2\end{aligned}$$

for all sufficiently large k. Let $l > 0$ be such that $\langle s^l, \nabla^2 f(\bar{u})s^l\rangle \leq -\varepsilon/2$. The observations that $s^l \in T(\mathcal{B}, \bar{u})$ and $\bar{u} + \tau s^l \in \mathcal{B}$ for $0 \leq \tau \leq 1/(l\|s\|_\infty)$ now yield the desired contradiction:

$$\begin{aligned}\frac{d}{d\tau} f(\bar{u} + \tau s^l)|_{\tau=0} &= \langle s, g(\bar{u})\rangle = 0\ , \\ \frac{d^2}{d\tau^2} f(\bar{u} + \tau s^l)|_{\tau=0} &= \langle s^l, \nabla^2 f(\bar{u})s^l\rangle \leq -\varepsilon/2 < 0.\end{aligned}$$

This readily shows that (O3) holds. □

4. The algorithm.

4.1. A Newton-like iteration. The key idea of the method to be developed consists in solving the equation $d(u)g(u) = 0$ by means of a Newton–like method augmented by a trust–region globalization. The bound constraints on u are enforced by, e.g., a scaling of the Newton-like step. In particular, all iterates will be strictly feasible with respect to the bounds: $u_k \in \mathcal{B}^\circ$.

In general it is not possible to find a function d satisfying (8) that depends smoothly on u. For an efficient method, however, we need a suitable substitute for the derivative of dg. Formal application of the product rule suggests to choose an approximate derivative of the form

$$D(u)\nabla^2 f(u) + d_u(u)g(u)\ ,\ u \in \mathcal{B}^\circ,$$

with $d_u(u)w \in \mathcal{L}(U, U')$, $w \in U'$, replacing the in general non-existing derivative of $u \in \mathcal{B} \longmapsto d(u)w \in U'$ at u. Here and in the sequel the linear operator $D^r(u)$, $r \geq 0$, denotes the pointwise multiplication operator associated with $d^r(u)$, i.e.

$$D^r(u) : v \longmapsto d(u)^r v.$$

Since $d^r(u) \in V$, $D^r(u)$ maps $L^q(\Omega)$, $1 \leq q \leq \infty$, continuously into itself. Moreover, if the assumption (D2) below is satisfied and $u \in \mathcal{B}^\circ$, then $D^r(u)$ defines an automorphism of $L^q(\Omega)$, $1 \leq q \leq \infty$, with inverse $D^{-r}(u)$. In fact, for all $u \in \mathcal{B}^\circ$ there exists $0 < \delta \leq \delta_d$ such that $u \in \mathcal{B}_\delta$, and thus $d(u)(x) \geq \varepsilon_d(\delta)$ on Ω by (D2). If we look at the special case $d = d_{\mathrm{I}}$, the choice $d_u(u)w = d'_{\mathrm{I}}(u)w$ with

$$d'_{\mathrm{I}}(u)(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } g(u)(x) > 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } u(x) - a(x) \leq b(x) - u(x) \\ -1 & \text{if } g(u)(x) < 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } b(x) - u(x) < u(x) - a(x) \end{cases} \tag{12}$$

for $u \in \mathcal{B}$, $x \in \Omega$ seems to be the most natural.

For the general case this suggests the choice

$$D(u)\nabla^2 f(u) + E(u)\ ,$$

where $E(u) : v \longmapsto e(u)v$ is a multiplication operator, $e(u) \in V$, which approximates $d_u(u)g(u)$. Properties of E will be specified below.

We are now able to formulate the following Newton-like iteration for the solution of $d(u)g(u) = 0$:

Given $u_k \in \mathcal{B}^\circ$, compute the new iterate $u_{k+1} := u_k + s_k \in \mathcal{B}^\circ$ where $s_k \in U$ solves

$$(D_k B_k + E_k)s_k = -d_k g_k, \tag{13}$$

and B_k denotes a symmetric approximation of (or replacement for) $\nabla^2 f(u_k)$, i.e. $\langle v, B_k w\rangle = \langle w, B_k v\rangle$ for all $v, w \in U$.

We assume that B_k satisfies the following condition:

(A5) The norms $\|B_k\|_{U,U'}$ are uniformly bounded by a constant $c_2 > 0$.

In the following, we will not restrict our investigations to special choices of d and e. Rather, we will develop an algorithm that is globally convergent for all affine scalings d and corresponding e satisfying the assumptions (D1)–(D5):

(D1) The scaling d satisfies (8) for all $u \in \mathcal{B}$.

(D2) There exists $\delta_d > 0$ such that for all $\delta \in (0, \delta_d]$ there is $\varepsilon_d = \varepsilon_d(\delta) > 0$ such that $d(u)(x) \geq \varepsilon_d$ for all $u \in \mathcal{B}$ and all $x \in \Omega$ with $a(x) + \delta \leq u(x) \leq b(x) - \delta$.

(D3) The scaling satisfies $d(u)(x) \leq d_{\mathrm{I}}(u)(x)$ for all $u \in \mathcal{B}$, $x \in \Omega$ and d_{I} given by (9). In particular, $d(u)(x) \leq c_d$ for some $c_d > 0$.

(D4) For all $u \in \mathcal{B}$ the function $e(u)$ satisfies $0 \leq e(u)(x) \leq c_e$ for all $x \in \Omega$ and $g(u)(x) = 0$ implies $e(u)(x) = 0$.

(D5) The function $e(u)$ is given by $e(u) = d'(u)g(u)$, where $d'(u)$ satisfies $|d'(u)(x)| \leq c_{d'}$ for all $u \in \mathcal{B}$ and $x \in \Omega$.

We have seen that assumption (D1) is essential for the reformulation of the first–order necessary optimality conditions and that (D2) ensures the continuous invertability of the scaling operator $D(u)$ for $u \in \mathcal{B}^\circ$. Furthermore, assumption (D2) will be used in the second–order convergence analysis. The assumption (D4) together with (A5) is needed to ensure uniform boundedness of the Hessian approximations $\hat{M}_k$ to be defined later (see Remark 4.2). The assumption (D5) is needed to prove second–order convergence results.

Obviously, (D1)–(D3) hold for either $d = d_{\mathrm{I}}$ and $d = d_{\mathrm{II}}$. The assumption (D4) is satisfied for $e(u) = d'_{\mathrm{I}}(u)g(u)$, where $d'_{\mathrm{I}}(u)$ is given by (12), provided that $\|g(u)\|_\infty$ is uniformly bounded on $\mathcal{B}$, i.e. provided that (A3) holds. The following example illustrates that the relaxed requirements on the scaling function d can be used to improve the smoothness of d and the scaled gradient dg substantially[1]:

EXAMPLE 4.1. The quadratic function

$$f(u) = \frac{1}{2}\|u\|_2^2 - \frac{1}{4}\left(\int_0^1 u(x)\,dx\right)^2$$

is smooth on $L^2([0,1])$. The gradient and the (strictly positive) Hessian are given by

$$g(u) = u - \frac{1}{2}\int_0^1 u(x)\,dx\ , \quad \nabla^2 f(u) : v \longmapsto v - \frac{1}{2}\int_0^1 v(x)\,dx.$$

f assumes its strict global minimum on the box $\mathcal{B} = \{x + \frac{1}{2} \leq u(x) \leq 2\}$ at the lower bound $\bar{u}$, $\bar{u}(x) = x + \frac{1}{2}$. At $u_\varepsilon = \bar{u} + \varepsilon(x + \frac{1}{100})$, $\varepsilon > 0$, the gradient $g(u_\varepsilon) = x + \varepsilon(x - \frac{49}{200})$ becomes negative for small x. Plot (a) in Figure 1 shows $d_{\mathrm{I}}(u_\varepsilon)$ (dashed), $d_{\mathrm{II}}(u_\varepsilon)$ (solid) and $|g(u_\varepsilon)|$ (dotted) for $\varepsilon = 0.001$. Note that d_{II} is continuous at the sign–change of $g(u_\varepsilon)$ and retains its order of magnitude in contrast to d_{I}. Plot (b) depicts the remainder terms

$$\varepsilon^{-1}\left|d_i(\bar{u})g(\bar{u}) - d_i(u_\varepsilon)g(u_\varepsilon) - (d'_i(u_\varepsilon)g(u_\varepsilon) + d'_i(u_\varepsilon)\nabla^2 f(u_\varepsilon))(\bar{u} - u_\varepsilon)\right|$$

for $i = \mathrm{I}$ (dashed) and $i = \mathrm{II}$ (solid) while $|g(u_\varepsilon)|$ is again dotted. Here d'_i is as in (12). Note that the remainder term for $d_{\mathrm{I}}g$ does *not* tend to zero near the sign–change of $g(u_\varepsilon)$ in contrast to $d_{\mathrm{II}}g$. In fact, it follows from our investigations in [22] that

$$\|d_{\mathrm{II}}(\bar{u})g(\bar{u}) - d_{\mathrm{II}}(u)g(u) - (d'_{\mathrm{II}}(u)g(u) + d'_{\mathrm{II}}(u)\nabla^2 f(u))(\bar{u} - u)\|_q = o(\|u - \bar{u}\|_\infty)$$

for $u \in \mathcal{B}$, $1 \leq q \leq \infty$, and $\bar{u}$ satisfying (O1), (O2), if $g : \mathcal{B} \subset V \longrightarrow V$ is locally Lipschitz at $\bar{u}$ and $g : \mathcal{B} \subset V \longrightarrow L^q(\Omega)$ is continuously differentiable in a neighborhood of $\bar{u}$. Our example admits the choice $q = \infty$.

[1] The advantages of the improved smoothness will be seen in the local convergence analysis [22].

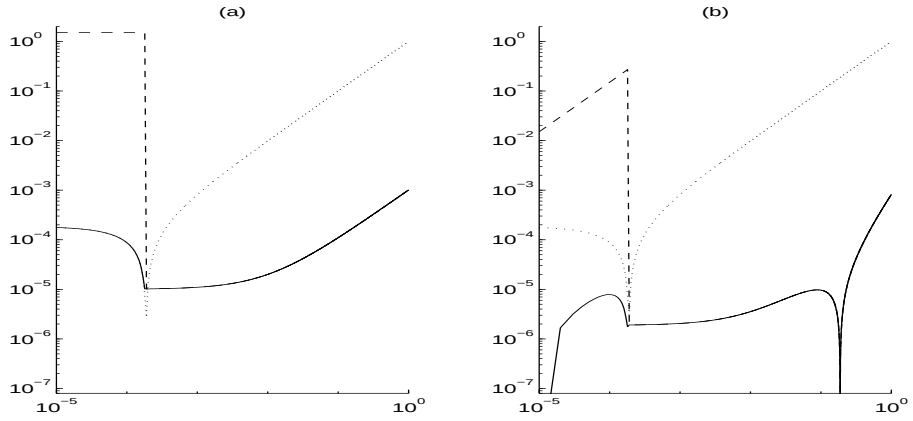

FIG. 1. *Smoothness properties of the scaling functions* $d_{\mathrm{I}}(u)$ *and* $d_{\mathrm{II}}(u)$

4.2. New coordinates and symmetrization. Since neither the well-definedness nor the global convergence of the Newton-like iteration (13) can be ensured, we intend to safeguard and globalize it by means of a closely related trust–region method. To this end we have to transform (13) into an equivalent quadratic programming problem. While the iterates are required to stay strictly feasible with respect to the bound constraints, we want to use an affine–scaling interior–point approach to reduce the effect of the interfering bound constraints in the quadratic subproblem as far as possible. The affine scaling can be expressed by a change of coordinates $s \rightsquigarrow \hat{s}$ and has to be performed in such a way that we get enough distance from the boundary of the box $\mathcal{B}$ to be able to impose a useful fraction of Cauchy decrease condition on the trial step. An appropriate change of coordinates $s \rightsquigarrow \hat{s}$ is given by $\hat{s} \stackrel{\text{def}}{=} d_k^{-r} s$. Here $r \geq 1/2$ is arbitrary but fixed throughout the iteration. Performing this transformation and applying D_k^{r-1}, the multiplication operator associated with d_k^{r-1}, from the left to (13) leads to the equivalent equation

$$\hat{M}_k \hat{s}_k = -\hat{g}_k \tag{14}$$

with $\hat{g}(u) \stackrel{\text{def}}{=} d^r(u) g(u)$, $\hat{M}_k \stackrel{\text{def}}{=} \hat{B}_k + \hat{C}_k$, where $\hat{B}_k \stackrel{\text{def}}{=} D_k^r B_k D_k^r$, and $\hat{C}_k \stackrel{\text{def}}{=} E_k D_k^{2r-1}$.

REMARK 4.2. Assumptions (D4) and (A5) imply that $\|\hat{M}_k\|_{U,U'}$ are uniformly bounded by a constant $c_3 > 0$.

Since $\hat{M}_k$ is symmetric, $\hat{s}_k$ is a solution of (14) if and only if it is a stationary point of the quadratic function

$$\hat{\psi}_k(\hat{s}) \stackrel{\text{def}}{=} \langle \hat{s}, \hat{g}_k \rangle + \frac{1}{2} \langle \hat{s}, \hat{M}_k \hat{s} \rangle .$$

We will return to this issue later.

4.3. Second–order necessary conditions revisited. If $B_k = \nabla^2 f(u_k)$, then the operator

$$\hat{M}(u) \stackrel{\text{def}}{=} D(u)^r \nabla^2 f(u) D(u)^r + E(u) D(u)^{2r-1} \tag{15}$$

also plays an important role in the second–order necessary optimality conditions. In fact, we will show that if conditions (O1), (O2) hold at $\bar{u}$, then (O3) can be equivalently replaced by

(O3′) $\langle s, \hat{M}(\bar{u})s\rangle \geq 0$ for all $s \in T(\mathcal{B}, \bar{u})$

or even

(O3″) $\langle s, \hat{M}(\bar{u})s\rangle \geq 0$ for all $s \in V$.

The proof requires the following two lemmas.

LEMMA 4.3. *Let* (D1) *be satisfied, let* $g(\bar{u}) \in U'$, *and suppose that* (O1), (O2) *hold at* $\bar{u}$. *Then*

$$I^* \stackrel{\text{def}}{=} \{x \in \Omega : d(\bar{u})(x) > 0\} = \{x \in \Omega : a(x) < \bar{u}(x) < b(x)\} \stackrel{\text{def}}{=} I. \tag{16}$$

Proof. The inclusion $I \subset I^*$ is obvious from (8). Now let $x \in I^*$ be given. Then $g(\bar{u})(x) = 0$ by (O2) and Lemma 3.2. From (8) we conclude $\bar{u}(x) \notin \{a(x), b(x)\}$, i.e. $x \in I$. □

LEMMA 4.4. *Let* (D1) *and* (D4) *be satisfied, let* $g(\bar{u}) \in U'$, *and suppose that* (O1), (O2) *hold at* $\bar{u}$. *Moreover, assume that* f *is twice continuously differentiable at* $\bar{u}$ *with* $\nabla^2 f(\bar{u}) \in \mathcal{L}(U, U')$. *Then the statements* (O3′) *and* (O3″) *are equivalent.*

Proof. Obviously ii) implies i). To show the opposite direction, assume that i) holds. Set $A = \Omega \setminus I$, were I is the set defined in (16). For arbitrary $s \in V$ we perform the splitting $s = s_I + s_A$, $s_I = \chi_I s \in T(\mathcal{B}, \bar{u})$, $s_A = \chi_A s$. Lemma 4.3 implies that $d^r(\bar{u})s_A = 0$ and we obtain

$$\begin{aligned}\langle s, \hat{M}(\bar{u})s\rangle &= \langle s_I, \hat{M}(\bar{u})s_I\rangle + 2\langle s_A, e(\bar{u})d^{2r-1}(\bar{u})s_I\rangle \\ &\quad + 2\langle d^r(\bar{u})s_A, \nabla^2 f(\bar{u})d^r(\bar{u})s_I\rangle + \langle d^r(\bar{u})s_A, \nabla^2 f(\bar{u})d^r(\bar{u})s_A\rangle \\ &\quad + \langle s_A, e(\bar{u})d^{2r-1}(\bar{u})s_A\rangle \\ &= \langle s_I, \hat{M}(\bar{u})s_I\rangle + \langle s_A, e(\bar{u})d^{2r-1}(\bar{u})s_A\rangle \geq \langle s_I, \hat{M}(\bar{u})s_I\rangle \geq 0.\end{aligned}$$

This completes the proof. □

THEOREM 4.5. *Let* (D1), (D2), *and* (D4) *be satisfied. Then in Theorem* 3.3 *condition* (O3) *can be equivalently replaced by* (O3′) *or* (O3″).

Proof. Since the conditions of Theorem 3.3 and Lemma 4.4 guarantee that (O3′) and (O3″) are equivalent, we only need to show that (O3) can be replaced by (O3′).

Let (O1), (O2) be satisfied. Then for all $s \in T(\mathcal{B}, \bar{u})$ we have $sg(\bar{u}) = 0$. To show that (O3) implies (O3′), let $s \in T(\mathcal{B}, \bar{u})$ be arbitrary. Then $h = d^r(\bar{u})s$ is also contained in $T(\mathcal{B}, \bar{u})$. Therefore, $\langle s, \hat{M}(\bar{u})s\rangle \geq \langle h, \nabla^2 f(\bar{u})h\rangle \geq 0$.

To prove the opposite direction, assume that there exist $s \in T(\mathcal{B}, \bar{u})$ and $\varepsilon > 0$ with $\langle s, \nabla^2 f(\bar{u})s\rangle < -\varepsilon$. As carried out in the proof of Theorem 3.3, we can find $l > 0$ such that $s^l = \chi_{I_l} s \in T(\mathcal{B}, \bar{u})$, I_l as defined in (11), satisfies $\langle s^l, \nabla^2 f(\bar{u})s^l\rangle \leq -\varepsilon/2$. Since $d(\bar{u})$ is bounded away from zero on I_l by assumption (D2), we obtain that $h = \chi_{I_l} d^{-r}(\bar{u})s$ is an element of $T(\mathcal{B}, \bar{u})$ that satisfies $\langle h, \hat{M}(\bar{u})h\rangle = \langle s^l, \nabla^2 f(\bar{u})s^l\rangle \leq -\varepsilon/2$ (note that $e(\bar{u})h = 0$ by (D4)). This contradicts (O3′). □

Define

$$\hat{\psi}[u](\hat{s}) \stackrel{\text{def}}{=} \langle \hat{s}, \hat{g}(u)\rangle + \frac{1}{2}\langle \hat{s}, \hat{M}(u)\hat{s}\rangle$$

with $\hat{M}(u)$ given by (15) and $\hat{g}(u) \stackrel{\text{def}}{=} d^r(u)g(u)$. Note that $\hat{\psi}[u_k] = \hat{\psi}_k|_{B_k = \nabla^2 f(u_k)}$. The previous results show that $\hat{\psi}[\bar{u}](\hat{s})$ is convex and admits a global minimum at $\hat{s} = 0$ if $\bar{u}$ is a local solution of (P).

4.4. Trust–region globalization. The results on the second–order conditions in the previous section indicate that the Newton–like iteration (14) can be used locally under appropriate conditions on B_k. To globalize the iteration, we minimize $\hat{\psi}_k(\hat{s})$ over the intersection of the ball $\|\hat{w}_k\hat{s}\|_p \leq \Delta_k$ and the box $\mathcal{B}$ which leads to the following trust–region subproblem:

Compute an approximate solution $\hat{s}_k$ with $u_k + d_k^r\hat{s}_k \in \mathcal{B}^\circ$ of

$$\min \hat{\psi}_k(\hat{s}) \quad \text{subject to} \quad \|\hat{w}_k\hat{s}\|_p \leq \Delta_k\,,\ u_k + d_k^r\hat{s} \in \mathcal{B} \tag{17}$$

Here $\hat{w}_k \in V$ is a positive scaling function for the trust–region, see assumption (W) below. As noted in § 2, the crucial contributions of the affine scaling are the term $E(u)D(u)^{2r-1}$ in the Hessian $\hat{M}(u)$ and the scaling $\hat{g}$ of the gradient. The trust–region serves as a tool for globalization. Therefore, more general trust–region scalings can be admitted, as long as they satisfy (W) below. This freedom in the scaling of the trust–region will be important for the infinite–dimensional local convergence analysis of this method. See [22].

We will work with the original variables in terms of which the above problem reads

Compute s_k with $u_k + s_k \in \mathcal{B}^\circ$ as an approximate solution of

$$\min \psi_k(s) \quad \text{subject to} \quad \|w_k s\|_p \leq \Delta_k\,,\ u_k + s \in \mathcal{B} \tag{18}$$

with $\psi_k(s) = \langle s, g_k\rangle + \frac{1}{2}\langle s, M_k s\rangle$, $M_k = B_k + C_k$, $C_k = E_k D_k^{-1}$, and $w_k = d_k^{-r}\hat{w}_k$.

The only restriction on the trust–region scaling is that w_k^{-1} as well as $\hat{w}_k = d_k^r w_k$ are pointwise bounded uniformly in k:

(W) There exist $c_w > 0$ and $c_{w'} > 0$ such that $\|d_k^r w_k\|_\infty \leq c_w$ and $\|w_k^{-1}\|_\infty \leq c_{w'}$ for all k.

Examples for w_k are $w_k = d_k^{-r}$ which yields a ball in the $\hat{s}$–variables, and $w_k = 1$ which leads to a ball in the s–variables. Both choices satisfy (W) if (D3) holds. See also [11].

The functions d_k^{-r} and d_k^{-1} are only well defined if $u_k \in \mathcal{B}^\circ$. Therefore, the condition $u_k + d_k^r\hat{s}_k \in \mathcal{B}^\circ$ on the trial iterate is essential. However, it is important to remark that the bound constraints do not need to be strictly enforced when computing $\hat{s}_k$. For example, in the finite–dimensional algorithms in [6], [11], an approximate solution of

$$\min \hat{\psi}_k(\hat{s}) \quad \text{subject to} \quad \|\hat{w}_k\hat{s}\|_p \leq \Delta_k$$

is computed and then scaled by $\tau_k > 0$ so that $u_k + \tau_k d_k^r\hat{s}_k \in \mathcal{B}^\circ$. Similar techniques also apply in the infinite–dimensional framework. Practical choices for the infinite–dimensional algorithm will be discussed in [22].

4.5. Cauchy decrease for the trial steps. An algorithm which is based on the iterative approximate solution of subproblem (18) can be expected to converge to a local solution of (P) only if the trial steps s_k produce a sufficiently large decrease of ψ_k. A well established way to impose such a condition is the requirement that the decrease provided by s_k should be at least a fraction of the *Cauchy decrease*.

Here the Cauchy decrease denotes the maximum possible decrease along the steepest descent direction of ψ_k at $s = 0$ with respect to an appropriate norm (or, equivalently, appropriate coordinates) inside the feasible region of the subproblem. We will see in Lemma 6.1 that the new coordinates $\hat{s} = d_k^{-r} s$ indeed provide enough distance to the boundary of $\mathcal{B}$ to allow the implementation of a useful Cauchy decrease strategy.

Unless in the Hilbert space case $p = 2$, the steepest descent direction of $\hat{\psi}_k$ at $\hat{s} = 0$ is *not* given by the negative gradient $-\hat{g}_k$ but rather by any $\hat{s}^d \neq 0$ satisfying $\langle \hat{s}^d, \hat{g}_k \rangle = \|\hat{s}^d\|_p \|\hat{g}_k\|_{p'}$. On the other hand, if $\hat{g}_k \in H$ then $-\nabla\hat{\psi}_k(0) = -\hat{g}_k$ is the $\|\cdot\|_2$-steepest descent direction of $\hat{\psi}_k$ at $\hat{s} = 0$. This is a strong argument for choosing this direction as basis for the Cauchy decrease condition. Of course this approach is only useful if we ensure that $u_k - \tau d_k^r \hat{g}_k \in \mathcal{B}^\circ$ for all $\tau > 0$ sufficiently small which can be done by imposing condition (A2) on g which is not very restrictive. Assuming this, we may take $-d_k^r \hat{g}_k = -d_k^{2r} g_k$ as Cauchy decrease direction of ψ_k, and therefore define the following *fraction of Cauchy decrease condition*:

There exist $\beta, \beta_0 > 0$ (fixed for all k) such that s_k is an approximate solution of (18) in the following sense:

$$\|w_k s_k\|_p \leq \beta_0 \Delta_k\,,\ u_k + s_k \in \mathcal{B}^\circ\,,\ \text{and}\ \psi_k(s_k) < \beta \psi_k(s_k^c), \tag{19a}$$

where s_k^c is a solution of the one-dimensional problem

$$\min \psi_k(s) \quad \text{subject to} \quad s = -t d_k^{2r} g_k\,,\ t \geq 0\,,\ u_k + s \in \mathcal{B}\,,\ \|w_k s\|_p \leq \Delta_k. \tag{19b}$$

4.6. Formulation of the algorithm. For the update of the trust–region radius Δ_k and the acceptance of the step we use a very common strategy. It is based on the demand that the actual decrease

$$ared_k(s_k) \stackrel{\text{def}}{=} f_k - f(u_k + s_k) \tag{20}$$

should be a sufficiently large fraction of the predicted decrease

$$pred_k(s_k) \stackrel{\text{def}}{=} -\langle s_k, g_k \rangle - \frac{1}{2}\langle s_k, B_k s_k \rangle = -\psi_k(s_k) + \frac{1}{2}\langle s_k, C_k s_k \rangle \tag{21}$$

promised by the quadratic model. Since the model error is at most $O(\|s_k\|_p^2)$, the decrease ratio

$$\rho_k \stackrel{\text{def}}{=} \frac{ared_k(s_k)}{pred_k(s_k)} \tag{22}$$

will tend to one for $s_k \to 0$. This suggests the following strategy for the update of the trust–region radius:

ALGORITHM 4.6 (UPDATE OF THE TRUST–REGION RADIUS Δ_k).

Let $0 < \eta_1 < \eta_2 < \eta_3 < 1$, and $0 < \gamma_1 < 1 < \gamma_2 < \gamma_3$.

1. If $\rho_k \leq \eta_1$ then choose $\Delta_{k+1} \in (0, \gamma_1 \Delta_k]$.
2. If $\rho_k \in (\eta_1, \eta_2)$ then choose $\Delta_{k+1} \in [\gamma_1 \Delta_k, \Delta_k]$.
3. If $\rho_k \in [\eta_2, \eta_3)$ then choose $\Delta_{k+1} \in [\Delta_k, \gamma_2 \Delta_k]$.
4. If $\rho_k \geq \eta_3$ then choose $\Delta_{k+1} \in [\gamma_2 \Delta_k, \gamma_3 \Delta_k]$.

REMARK 4.7. The forms of predicted and actual decrease follow the choices used in [11], [23] (and [10] for the constrained case). In [6] the decreases and the ratio are computed as follows:

$$pred_k^1(s_k) \stackrel{\text{def}}{=} -\psi_k(s_k) \ , \quad ared_k^1(s_k) \stackrel{\text{def}}{=} ared_k(s_k) - \frac{1}{2}\langle s_k, C_k s_k\rangle \ , \quad \rho_k^1 \stackrel{\text{def}}{=} \frac{ared_k^1(s_k)}{pred_k^1(s_k)}.$$

Since the crucial estimates (25) and (38) also are true for $pred_k^1(s_k)$, and under (D4) the relations

$$pred_k^1(s_k)(\rho_k^1 - 1) = pred_k(s_k)(\rho_k - 1) \ , \quad ared_k^1(s_k) \le ared_k(s_k)$$

hold, all convergence results presented in this paper remain valid if ρ_k is replaced by ρ_k^1 . We restrict the presentation to the choice (20), (21).

The algorithm iteratively computes a trial step s_k satisfying the fraction of Cauchy decrease condition. Depending on the decrease ratio ρ_k the trial step is accepted or rejected, and the trust–region radius is adjusted.

ALGORITHM 4.8 (TRUST–REGION INTERIOR–POINT ALGORITHM).

Let $\eta_1 > 0$ as in Algorithm 4.6.

1. Choose $u_0 \in \mathcal{B}^\circ$ and $\Delta_0 > 0$.
2. For $k = 0, 1, \ldots$
   - 2.1. If $\hat{g}_k = 0$ then STOP with result u_k.
   - 2.2. Compute s_k satisfying (19).
   - 2.3. Compute ρ_k as defined in (22).
   - 2.4. If $\rho_k > \eta_1$ then set $u_{k+1} = u_k + s_k$, else set $u_{k+1} = u_k$.
   - 2.5. Compute Δ_{k+1} using Algorithm 4.6.

5. Norm estimates. In this section we collect several useful norm estimates for L^q-spaces. The first lemma states that $\|\cdot\|_{q_1}$ is majorizable by a multiple of $\|\cdot\|_{q_2}$ if $q_2 \ge q_1$.

LEMMA 5.1. *For all $1 \le q_1 \le q_2 \le \infty$ and $v \in L^{q_2}(\Omega)$ we have*

$$\|v\|_{q_1} \le m_{q_1,q_2}\|v\|_{q_2}$$

with $m_{q_1,q_2} = \mu(\Omega)^{\frac{1}{q_1}-\frac{1}{q_2}}$. Here $1/\infty$ is to be interpreted as zero.

Proof. See e.g. [1, Thm. 2.8]. □

As a consequence of Hölder's inequality we obtain the following result, which allows us to apply the principle of boundedness in the high- and convergence in the low-norm.

LEMMA 5.2. (Interpolation inequality) *Given $1 \le q_1 \le q_2 \le \infty$ and $0 \le \theta \le 1$, let $1 \le q \le \infty$ satisfy $1/q = \theta/q_1 + (1-\theta)/q_2$. Then for all $v \in L^{q_2}(\Omega)$ the following is true:*

$$\|v\|_q \le \|v\|_{q_1}^{\theta}\|v\|_{q_2}^{1-\theta} \tag{23}$$

Proof. In the nontrivial cases $0 < \theta < 1$ and $q < \infty$ observe that $[q_1/(\theta q)]^{-1} + [q_2/((1-\theta)q)]^{-1} = 1$ and apply Hölder's inequality:

$$\|v\|_q^q = \left\||v|^{\theta q}|v|^{(1-\theta)q}\right\|_1 \le \left\||v|^{\theta q}\right\|_{\frac{q_1}{\theta q}}\left\||v|^{(1-\theta)q}\right\|_{\frac{q_2}{(1-\theta)q}} = \|v\|_{q_1}^{(\theta q)}\|v\|_{q_2}^{(1-\theta)q}.$$

□

The next lemma will be used in the proof of Lemma 7.1.

LEMMA 5.3. *For $v \in L^q(\Omega)$, $1 \le q < \infty$ and all $\delta > 0$ holds*

$$\mu\left(\{x \in \Omega : |v(x)| \ge \delta\}\right) \le \delta^{-q}\|v\|_q^q.$$

Proof.

$$\|v\|_q^q = \||v|^q\|_1 \ge \|\chi_{\{|v|\ge\delta\}}|v|^q\|_1 \ge \mu(\{|v| \ge \delta\})\delta^q.$$

□

6. Convergence to first–order optimal points. The convergence of the algorithm is mainly achieved by two ingredients: A lower bound for the predicted decrease for trial steps satisfying the fraction of Cauchy decrease condition, and the relation $ared_k(s_k) > \eta_1 pred_k(s_k)$ which is always satisfied for successful steps s_k. The lower bound on the predicted decrease is established in the following lemma:

LEMMA 6.1. *Let the assumptions* (A1), (A2), (D1)–(D4), *and* (W) *hold. Then there exists $c_4 > 0$ such that for all $u_k \in \mathcal{B}^\circ$ with $\hat{g}_k \ne 0$ and all s_k satisfying* (19) *the following holds:*

$$pred_k(s_k) \ge -\psi_k(s_k) \ge \frac{1}{2}\beta\|\hat{g}_k\|_2^2 \min\left\{\frac{\Delta_k}{c_w\|\hat{g}_k\|_p}, \frac{\|\hat{g}_k\|_2^2}{\|\hat{M}_k\|_{U,U'}\|\hat{g}_k\|_p^2}, \frac{c_d^{1-2r}}{\|g_k\|_\infty}\right\} \tag{24}$$

$$\ge c_4\|\hat{g}_k\|_{p'}^2 \min\left\{\frac{\Delta_k}{c_w\|\hat{g}_k\|_p}, \frac{\|\hat{g}_k\|_{p'}^2}{\|\hat{M}_k\|_{U,U'}\|\hat{g}_k\|_p^2}, \frac{c_d^{1-2r}}{\|g_k\|_\infty}\right\}. \tag{25}$$

Proof. Since C_k is obviously positive by (D4), we have

$$pred_k(s_k) = -\psi_k(s_k) + \frac{1}{2}\langle s_k, C_k s_k\rangle \ge -\psi_k(s_k).$$

Now we will derive an upper bound for the minimum of $\phi(\tau) = \psi_k(-\tau d_k^{2r} g_k)$ on $[0, \tau^+]$ with $\tau^+ = \min\{\tau_{\mathcal{B}}, \tau_\Delta\}$, where

$$\tau_{\mathcal{B}} = \max\Big\{\tau : b(x) - u_k(x) + \tau d_k^{2r}(x)g_k(x) \ge 0\,, \text{ and}$$
$$u_k(x) - a(x) - \tau d_k^{2r}(x)g_k(x) \ge 0 \quad \forall\, x \in \Omega\Big\}$$

and

$$\tau_\Delta = \frac{\Delta_k}{\|w_k d_k^{2r} g_k\|_p} = \frac{\Delta_k}{\|\hat{w}_k\hat{g}_k\|_p} \ge \frac{\Delta_k}{c_w\|\hat{g}_k\|_p}.$$

Therefore, using (D3),

$$\tau_{\mathcal{B}} = \min\left\{\inf_{\{g_k(x)<0\}} \frac{b(x) - u_k(x)}{-d_k^{2r}(x)g_k(x)}, \inf_{\{g_k(x)>0\}} \frac{u_k(x) - a(x)}{d_k^{2r}(x)g_k(x)}\right\} = \inf_{\{g_k(x)\ne 0\}} \frac{d_{\mathrm{I}}(u_k)(x)}{d_k^{2r}(x)|g_k(x)|}$$

$$\ge \inf_{\{g_k(x)\ne 0\}} \frac{d_k^{1-2r}(x)}{|g_k(x)|} \ge \inf_{\{g_k(x)\ne 0\}} \frac{c_d^{1-2r}}{|g_k(x)|} \ge \frac{c_d^{1-2r}}{\|g_k\|_\infty}.$$

We have $\phi(\tau) = -\kappa_1\tau + \frac{1}{2}\kappa_2\tau^2$ with

$$\kappa_1 = \langle d_k^{2r} g_k, g_k\rangle = \|\hat{g}_k\|_2^2 \quad , \quad \kappa_2 = \langle d_k^{2r} g_k, M_k d_k^{2r} g_k\rangle = \langle \hat{g}_k, \hat{M}_k\hat{g}_k\rangle,$$

and observe $|\kappa_2| \le \|\hat{M}_k\|_{U,U'}\|\hat{g}_k\|_p^2$. Let τ^* be a minimizer for ϕ on $[0,\tau^+]$. If $\tau^* < \tau^+$ then $\kappa_2 > 0$, $\tau^* = \kappa_1/\kappa_2$, and

$$\phi(\tau^*) = -\frac{1}{2}\frac{\kappa_1^2}{\kappa_2} \le -\frac{1}{2}\frac{\|\hat{g}_k\|_2^4}{\|\hat{M}_k\|_{U,U'}\|\hat{g}_k\|_p^2}.$$

If $\tau^* = \tau_\Delta$ and $\kappa_2 > 0$ then $\kappa_1/\kappa_2 \ge \tau_\Delta$ and

$$\phi(\tau^*) = -\kappa_1\tau_\Delta + \frac{\kappa_2}{2}\tau_\Delta^2 \le -\frac{\kappa_1}{2}\tau_\Delta \le -\frac{1}{2}\frac{\|\hat{g}_k\|_2^2}{c_w\|\hat{g}_k\|_p}\Delta_k.$$

If $\tau^* = \tau_\Delta$ and $\kappa_2 \le 0$ then even $\phi(\tau^*) \le -\kappa_1\tau_\Delta$. For $\tau^* = \tau_\mathcal{B}$ the same arguments show

$$\phi(\tau^*) \le -\frac{\kappa_1}{2}\tau_\mathcal{B} \le -\frac{1}{2}c_d^{1-2r}\frac{\|\hat{g}_k\|_2^2}{\|g_k\|_\infty}.$$

The first inequality (24) now follows from these estimates and (19). The second inequality (25) follows from (24) and the application

$$\|\hat{g}_k\|_{p'} \le m_{p',2}\|\hat{g}_k\|_2$$

of Lemma 5.1. Note that $p \ge 2$ and $1/p + 1/p' = 1$ yield $p' \le 2$. □

REMARK 6.2. The sequence of inequalities for the estimation of $\tau_\mathcal{B}$ uses the inequality $d_k^{2r-1} \le c_d^{2r-1}$. This is where we need the restriction to $r \ge 1/2$.

Let the assumptions of Lemma 6.1 hold. If the kth iteration of Algorithm 4.8 is successful, i.e. $\rho_k > \eta_1$ (or equivalently $u_{k+1} \ne u_k$), then Lemma 6.1 provides an estimate for the actual decrease:

$$f_k - f_{k+1} > \eta_1 c_4 \|\hat{g}_k\|_{p'}^2 \min\left\{\frac{\Delta_k}{c_w\|\hat{g}_k\|_p}, \frac{\|\hat{g}_k\|_{p'}^2}{\|\hat{M}_k\|_{U,U'}\|\hat{g}_k\|_p^2}, \frac{c_d^{1-2r}}{\|g_k\|_\infty}\right\}.$$

If in addition the assumptions (A3) and (A5) hold, Remark 4.2 and the previous inequality imply the existence of $c_5 > 0$ with

$$f_k - f_{k+1} > c_5\|\hat{g}_k\|_{p'}^2 \min\left\{\Delta_k, \|\hat{g}_k\|_{p'}^2, c_d^{1-2r}\right\}. \tag{26}$$

The next statement is trivial:

LEMMA 6.3. *Let (Δ_k) and (ρ_k) be generated by Algorithm 4.8. If $\rho_k \ge \eta_2$ for sufficiently large k then (Δ_k) is bounded away from zero.*

Now we can prove a first global convergence result.

THEOREM 6.4. *Let assumptions* (A1)–(A3), (A5), (D1)–(D4), *and* (W) *hold. Let the sequence (u_k) be generated by Algorithm 4.8. Then*

$$\liminf_{k\to\infty} \|d_k^r g_k\|_{p'} = 0.$$

Even more:

$$\liminf_{k\to\infty} \|d_k^r g_k\|_q = 0 \quad \textit{for all } 1 \le q < \infty.$$

Proof. Assume that there are $K > 0$ and $\varepsilon > 0$ with $\|\hat{g}_k\|_{p'} \geq \varepsilon$ for all $k \geq K$. First we will show that this implies $\sum_{k=0}^{\infty} \Delta_k < \infty$. If there is only a finite number of successful steps then $\Delta_{k+1} \leq \gamma_1 \Delta_k$ for large k and we are done. Otherwise, if the sequence (k_i) of successful steps does not terminate, we conclude from $f_k \downarrow$ and the boundedness of f that $\sum_{k=0}^{\infty}(f_k - f_{k+1}) < \infty$.

For all $k = k_i$ we may use (26) and obtain, since $\|\hat{g}_{k_i}\|_{p'} \geq \varepsilon$ for $k_i \geq K$, that Δ_{k_i} tends to zero and, moreover, obeys the inequality

$$\Delta_{k_i} < \frac{1}{c_5 \varepsilon^2}(f_{k_i} - f_{k_i+1})$$

for all k_i sufficiently large. This shows $\sum_{i=0}^{\infty} \Delta_{k_i} < \infty$. Since for all successful steps $k \in \{k_i\}$ we have $\Delta_{k+1} \leq \gamma_2 \Delta_k$ and for all others $\Delta_{k+1} \leq \gamma_1 \Delta_k$, we conclude

$$\sum_{k=0}^{\infty} \Delta_k \leq \sum_{i=0}^{\infty} \Delta_{k_i} \left(1 + \frac{\gamma_2}{1 - \gamma_1}\right) < \infty. \tag{27}$$

In a second step we will show that $|\rho_k - 1| \to 0$. Due to

$$\|u_{k+1} - u_k\|_p \leq \|s_k\|_p \leq \beta_0 \|w_k^{-1}\|_{\infty} \Delta_k \leq \beta_0 c_{w'} \Delta_k \tag{28}$$

and (27), (u_k) is a Cauchy sequence in U. Furthermore,

$$\begin{aligned} 2\left|\psi_k(s_k) - \langle s_k, g_k\rangle - \frac{1}{2}\langle s_k, C_k s_k\rangle\right| &= |\langle s_k, B_k s_k\rangle| \leq \|B_k\|_{U,U'} \|s_k\|_p^2 \\ &\leq c_2 \beta_0^2 c_{w'}^2 \Delta_k^2. \end{aligned}$$

The mean value theorem yields $f(u_k + s_k) - f_k = \langle s_k, \bar{g}_k\rangle$ for some $\tau_k \in [0,1]$ and $\bar{g}_k = g(u_k + \tau_k s_k)$, and hence

$$\begin{aligned} |pred_k(s_k)||\rho_k - 1| &= \left|f(u_k + s_k) - f_k + \frac{1}{2}\langle s_k, C_k s_k\rangle - \psi_k(s_k)\right| \\ &\leq \left|\langle s_k, g_k\rangle + \frac{1}{2}\langle s_k, C_k s_k\rangle - \psi_k(s_k)\right| + |\langle s_k, \bar{g}_k - g_k\rangle| \\ &\leq \left(\frac{c_2}{2}\beta_0^2 c_{w'}^2 \Delta_k + \beta_0 c_{w'} \|\bar{g}_k - g_k\|_{p'}\right) \Delta_k. \end{aligned}$$

Since (u_k) converges in the closed set $\mathcal{B}$, g is continuous, and (Δ_k) as well as $(\|s_k\|_p)$ (see (28)) tend to zero, the first factor in the last expression converges to zero, too. Lemma 6.1 garantees that $|pred_k(s_k)|/\Delta_k$ is uniformly bounded away from zero for $k \geq K$, since by assumption $\|\hat{g}_k\|_{p'} \geq \varepsilon$. This shows $|\rho_k - 1| \to 0$. But now Lemma 6.3 yields a contradiction to $\Delta_k \to 0$. Therefore, the assumption is wrong and the first part of the assertion holds.

The second part follows from Lemma 5.1 for $1 \leq q \leq p'$ and from (A3) and the interpolation inequality (23) for $p' < q < \infty$. □

Now we will show that if $\hat{g}$ is uniformly continuous the limites inferiores in Theorem 6.4 can be replaced by limites.

We introduce the following assumptions:

(A6) The scaled gradient $\hat{g} = d^r g : \mathcal{B} \subset U \longrightarrow U'$ is uniformly continuous.

(A6′) The gradient $g : \mathcal{B} \subset U \longrightarrow U'$ is uniformly continuous and $d = d_{\mathrm{I}}$ or $d = d_{\mathrm{II}}$.

Condition (A6) is not so easy to verify for most choices of d. With Lemma 6.5, however, we provide a very helpful tool to check the validity of (A6). Moreover, we show in Lemma 6.6 that (A6′) implies (A6). The proofs of both lemmas can be found in the appendix. As a by-product of our investigations we get the valuable result that $\hat{g}$ inherits the continuity of g if we choose $d = d_{\mathrm{I}}$ or $d = d_{\mathrm{II}}$. We will derive the results concerning continuity and uniform continuity of $\hat{g}$ simultaneously. Additional requirements for the uniform continuity are written in parentheses.

LEMMA 6.5. *Let* (A1)–(A3), (D3) *hold and* $g : \mathcal{B} \subset U \longrightarrow U'$ *be* (*uniformly*) *continuous. Assume that* $\|\chi_{\{g(u)g(\tilde{u})>0\}}(d(u) - d(\tilde{u}))\|_{p'}$ *tends to zero* (*uniformly in* $u \in \mathcal{B} \subset U$) *for* $\tilde{u} \to u$ *in* $\mathcal{B} \subset U$. *Then* $\hat{g} = d^r g : \mathcal{B} \subset U \longrightarrow U'$ *is* (*uniformly*) *continuous.*

Proof. See appendix. □

The previous lemma is now applicable to the choices $d = d_{\mathrm{I}}$ and $d = d_{\mathrm{II}}$:

LEMMA 6.6. *Let* (A1)–(A3) *hold and* $d = d_{\mathrm{I}}$ *or* $d = d_{\mathrm{II}}$. *Then* $\hat{g} = d^r g : \mathcal{B} \subset U \longrightarrow U'$ *is continuous. If, in addition,* g *is uniformly continuous, then the same is true for* $\hat{g}$.

Proof. See appendix. □

Now we state the promised variant of Theorem 6.4.

THEOREM 6.7. *Let assumptions* (A1)–(A3), (A5), (D1)–(D4), (W), *and* (A6) *or* (A6′) *hold. Then the sequence* (u_k) *generated by Algorithm* 4.8 *satisfies*

$$\lim_{k\to\infty} \|d_k^r g_k\|_{p'} = 0. \tag{29}$$

Even more:

$$\lim_{k\to\infty} \|d_k^r g_k\|_q = 0 \quad \textit{for all } 1 \le q < \infty. \tag{30}$$

Proof. Since, due to Lemma 6.6, $\hat{g} = d^r g$ is uniformly continuous, it suffices to show that under the assumption $\|\hat{g}_k\|_{p'} \ge \varepsilon_1 > 0$ for an infinite number of iterations k there exists a sequence of index pairs (m_i, l_i) with $\|\hat{g}_{m_i} - \hat{g}_{l_i}\|_{p'} \ge \delta > 0$ but $\|u_{m_i} - u_{l_i}\|_p \to 0$, which is a contradiction to the uniform continuity of $\hat{g}$.
So let us assume that (29) does not hold. Then there is $\varepsilon_1 > 0$ and a sequence (m_i) with $\|\hat{g}_{m_i}\|_{p'} \ge \varepsilon_1$. Theorem 6.4 yields a sequence (k_i) with $\|\hat{g}_{k_i}\|_{p'} \to 0$. For arbitrary $0 < \varepsilon_2 < \varepsilon_1$ we can thus find a sequence (l_i) such that

$$\|\hat{g}_k\|_{p'} \ge \varepsilon_2, \quad m_i \le k < l_i, \quad \|\hat{g}_{l_i}\|_{p'} < \varepsilon_2.$$

Since $\hat{g}_{l_i} \ne \hat{g}_{l_i-1}$, iteration $l_i - 1$ is successful and one has for all successful iterations k, $m_i \le k < l_i$, by Lemma 6.1 and (26)

$$f_k - f_{k+1} > c_5 \varepsilon_2^2 \min\left\{\Delta_k, \varepsilon_2^2, c_d^{1-2r}\right\}. \tag{31}$$

The left hand side converges to zero, because (f_k) is nonincreasing and bounded from below, i.e. is a Cauchy sequence. We conclude that Δ_k tends to zero for successful steps $m_i \le k < l_i$ and get with (28) that

$$f_k - f_{k+1} \ge c_5 \varepsilon_2^2 \Delta_k \ge \frac{c_5 \varepsilon_2^2}{\beta_0 c_{w'}} \|u_{k+1} - u_k\|_p \stackrel{\text{def}}{=} c_6 \|u_{k+1} - u_k\|_p,$$

which is clearly true also for unsuccessful iterations. Summing and using the triangle inequality yields

$$f_{m_i} - f_{l_i} \geq c_6 \|u_{m_i} - u_{l_i}\|_p.$$

Since (f_k) is a Cauchy sequence, the left hand side converges to zero for $i \to \infty$. Hence, $\|u_{m_i} - u_{l_i}\|_p \longrightarrow 0$ but

$$\|\hat{g}_{m_i} - \hat{g}_{l_i}\|_{p'} \geq \|\hat{g}_{m_i}\|_{p'} - \|\hat{g}_{l_i}\|_{p'} \geq \varepsilon_1 - \varepsilon_2 > 0.$$

This is a contradiction to the uniform continuity of $\hat{g}$. The second assertion follows as in the proof of Theorem 6.4. □

7. Convergence to second–order optimal points. The first–order convergence results in the previous section could be shown under rather weak conditions on the trust–region step s_k and for arbitrary symmetric and bounded Hessian 'approximations'. If stronger assumptions are imposed on B_k and on s_k, then it can be shown that every accumulation point of (u_k) satisfies the second–order necessary optimality conditions. This will be done in this section. We need the following assumption on the Hessian approximation:

(A7) For all accumulation points $\bar{u} \in U$ of (u_k) and all $\varepsilon > 0$ there is $\delta = \delta(\bar{u}, \varepsilon) > 0$ such that $\|u_k - \bar{u}\|_p \leq \delta$ implies $\|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \leq \varepsilon$.

Obviously (A7) is satisfied if $B_k = \nabla^2 f(u_k)$ and if (A4) holds. However, (A7) also applies in other important situations. For example, (A7) applies if f is a least squares functional, $f(\bar{u}) = 0$, and B_k is the Gauss–Newton approximation of the Hessian.

The fraction of Cauchy decrease condition does not take into account any properties of the quadratic part of ψ_k. Apparently, this condition is too weak to guarantee the positivity of $\hat{M}(\bar{u})$ at accumulation points of (u_k). The decrease condition has to be strengthened in such a way that for $\bar{u}$ satisfying (O1) and (O2) but not (O3″) there are $\alpha, \varepsilon, c > 0$ such that $\psi_k(s_k) \leq -c \min\{\Delta_k^2, \alpha^2\}$ for all iterates u_k with $\|u_k - \bar{u}\|_p \leq \varepsilon$. For the finite–dimensional problem one can establish such an inequality near nondegenerate points $\bar{u}$ by using techniques similar to those of Coleman and Li [6] if the s_k satisfy a finite–dimensional fraction of optimal decrease condition of the form

$$\|w_k s_k\|_2 \leq \beta_0 \Delta_k\,,\ u_k + s_k \in \mathcal{B}^\circ\,,\ \text{and}\ \psi_k(s_k) < \beta \psi_k(\tau_k s_k^\circ),$$

where $\tau_k = \max\{\tau \geq 0 : u_k + \tau s_k^\circ \in \mathcal{B}\}$ and s_k° solves

$$\min \psi_k(s) \quad \text{subject to} \quad \|w_k s\|_2 \leq \Delta_k.$$

This approach is not directly transferable to our setting because the example

$$\min \quad -\int_0^1 t s^2(t)\, dt \quad \text{subject to} \quad \|s\|_2 \leq \Delta \tag{32}$$

shows that even in a Hilbert space s_k° may not exist. Moreover, the proofs in [6] use extensively a convenient characterization of s_k° derived from the Karush–Kuhn–Tucker conditions (cf. [19]) and the equivalence of 2- and ∞-norm in $\mathbb{R}^N$. Since, as shown by (32), in Banach space the quadratic subproblem may not have a solution, this is not applicable in our framework. Our convergence proof requires that the trial steps yield

a fraction of the Cauchy decrease, and, moreover, a fraction of the decrease achievable along directions of negative curvature of ψ_k at $s = 0$. For convenience and simplicity of notation, however, we favor a more intuitive but stronger *fraction of optimal decrease condition*:

There exist $\beta, \beta_0 > 0$ (fixed for all k) such that

$$\|w_k s_k\|_p \le \beta_0 \Delta_k\,,\ u_k + s_k \in \mathcal{B}^\circ\,,\ \text{and}\ \psi_k(s_k) < \beta \psi_k^o, \tag{33a}$$

where

$$\psi_k^o \stackrel{\text{def}}{=} \inf \psi_k(s) \quad \text{subject to}\ \ u_k + s \in \mathcal{B}\,,\ \|w_k s\|_p \le \Delta_k \tag{33b}$$

In the next lemma we show that in a neighborhood of an accumulation point $\bar{u}$ of (u_k) at which (O1), (O2), but not (O3″) hold, one can find a direction of negative curvature h^n of ψ_k such that $u_k \pm h^n \in \mathcal{B}$.

LEMMA 7.1. *Let assumptions* (A1), (A2), (A5), (A7), (D1)–(D5) *hold and let the sequence* (u_k) *be generated by Algorithm* 4.8. *Assume that* $\bar{u} \in \mathcal{B}$ *is an accumulation point of* (u_k) *with* $\hat{g}(\bar{u}) = 0$ *and that there are* $\bar{h} \in V$, $\bar{h} \neq 0$, *and* $\lambda > 0$ *with*

$$\langle \bar{h}, \hat{M}(\bar{u})\bar{h}\rangle \le -\lambda \|\bar{h}\|_p^2. \tag{34}$$

Then there exist $\varepsilon, \alpha, \hat{\lambda} > 0$ *such that for all* u_k *with* $\|u_k - \bar{u}\|_p \le \varepsilon$ *one can find* $h \in V$, $\|h\|_p = 1$, *with* $u_k + \tau\alpha d_k^r h \in \mathcal{B}$ *for all* $\tau \in [-1, 1]$ *and*

$$\langle h, \hat{M}_k h\rangle \le -\hat{\lambda}\|h\|_p^2.$$

Proof. Since $\bar{u} \in \mathcal{B}$ and $\hat{g}(\bar{u}) = 0$, (O1) and (O2) are satisfied due to Lemma 3.2. Lemma 4.3 yields $I \stackrel{\text{def}}{=} \{x \in \Omega :\ a(x) < \bar{u}(x) < b(x)\} = \{x \in \Omega :\ d(\bar{u})(x) > 0\} \stackrel{\text{def}}{=} I^*$. Define $I_\delta = \{x \in \Omega :\ a(x) + \delta \le \bar{u}(x) \le b(x) - \delta\}$ for arbitrary $0 < \delta < 4\delta_d$ with δ_d as in (D2). We first show that (34) implies the existence of $\tilde{h} \in V$ with $\|\tilde{h}\|_p = 1$, $\{\tilde{h} \neq 0\} \subset I_\delta$, and

$$\langle \tilde{h}, \hat{M}(\bar{u})\tilde{h}\rangle \le -\frac{\lambda}{2}. \tag{35}$$

From $\hat{g}(\bar{u}) = d^r(\bar{u})g(\bar{u}) = 0$ we see that $g(\bar{u})(x) = 0$ on $I^* = I$. We write $v_A = \chi_A v$ for measurable functions v and measurable sets $A \subset \Omega$. Then

$$\begin{aligned} 0 > -\lambda\|\bar{h}\|_p^2 &\ge \langle \bar{h}, \hat{M}(\bar{u})\bar{h}\rangle = \langle \bar{h}, |d'(\bar{u})g(\bar{u})| d^{2r-1}(\bar{u})\bar{h}\rangle + \langle d^r(\bar{u})\bar{h}, \nabla^2 f(\bar{u}) d^r(\bar{u})\bar{h}\rangle \\ &\ge \langle d^r(\bar{u})\bar{h}_I, \nabla^2 f(\bar{u}) d^r(\bar{u})\bar{h}_I\rangle = \langle \bar{h}_I, \hat{M}(\bar{u})\bar{h}_I\rangle. \end{aligned}$$

So, $\bar{h}_I \in V \setminus \{0\}$ and (34) holds for $\bar{h}_I$ instead of $\bar{h}$. Furthermore, using the symmetry of $\hat{M}(\bar{u})$ and the identity $\bar{h}_I = \bar{h}_{I_\delta} + \bar{h}_{I \setminus I_\delta}$,

$$\begin{aligned} \langle \bar{h}_{I_\delta}, \hat{M}(\bar{u})\bar{h}_{I_\delta}\rangle &= \langle \bar{h}_I, \hat{M}(\bar{u})\bar{h}_I\rangle - \langle \bar{h}_{I_\delta} + \bar{h}_I, \hat{M}(\bar{u})\bar{h}_{I\setminus I_\delta}\rangle \\ &\le -\lambda\|\bar{h}_I\|_p^2 + 2\|\hat{M}(\bar{u})\bar{h}_{I\setminus I_\delta}\|_{p'}\|\bar{h}_I\|_p. \end{aligned}$$

Since the measure of $I \setminus I_\delta$ can be made arbitrarily small by reducing $\delta > 0$ and thus $\bar{h}_{I\setminus I_\delta}$ tends to zero for $\delta \to 0$ in all spaces $L^q(\Omega)$, $1 \le q < \infty$, we find $\delta > 0$ with

$$2\|\hat{M}(\bar{u})\bar{h}_{I\setminus I_\delta}\|_{p'} \le \frac{\lambda}{2}\|\bar{h}_I\|_p.$$

Then (35) holds with $\tilde{h} = \dfrac{\bar{h}_{I_\delta}}{\|\bar{h}_{I_\delta}\|_p}$. Obviously, $\{\tilde{h} \neq 0\} \subset I_\delta$. For $\varepsilon > 0$ and u_k with $\|u_k - \bar{u}\|_p \leq \varepsilon$, define $h \in V$ by

$$h(x) = \begin{cases} \tilde{h}(x)\dfrac{d^r(\bar{u})(x)}{d_k^r(x)} & \text{if} \quad \min\{u_k(x) - a(x), b(x) - u_k(x)\} > \dfrac{\delta}{4}, \\ 0 & \text{else.} \end{cases}$$

We have $I_h \overset{\text{def}}{=} \{h \neq 0\} \subset I_\delta$ and conclude from assumptions (D2) and (D3) that $\varepsilon_d(\delta/4) \leq d_k(x) \leq c_d$ on I_h and $\varepsilon_d(\delta/4) \leq d(\bar{u})(x) \leq c_d$ on I_δ, which implies that

$$\|h\|_p \leq \gamma \ , \ \|h\|_\infty \leq \gamma\|\tilde{h}\|_\infty \ , \ \|h\|_p \geq \frac{1}{\gamma}\|\tilde{h}_{I_h}\|_p \ \text{ with } \ \gamma = \frac{c_d^r}{\varepsilon_d^r(\delta/4)}. \tag{36}$$

From $\bar{u}(x) - a(x) \geq \delta \leq b(x) - \bar{u}(x)$ on I_δ follows

$$I_\delta \setminus I_h \subset \{x \in \Omega : \ |u_k(x) - \bar{u}(x)| \geq 3\delta/4\}.$$

If $p = \infty$ we achieve $\tilde{h}_{I_\delta \setminus I_h} = 0$ for $\varepsilon < 3\delta/4$. Otherwise, due to Lemma 5.3, we can make $\|\tilde{h}_{I_\delta \setminus I_h}\|_p \leq \mu(I_\delta \setminus I_h)^{\frac{1}{p}}\|\tilde{h}\|_\infty$ arbitrarily small by making $\varepsilon > 0$ small. Hence, in all cases we can reduce ε such that

$$\|h\|_p \geq \frac{1}{\gamma}\|\tilde{h}_{I_h}\|_p \geq \frac{1}{\gamma}\left(\|\tilde{h}\|_p - \|\tilde{h}_{I_\delta \setminus I_h}\|_p\right) \geq \frac{1}{2\gamma}. \tag{37}$$

By the definition of h and the fact that $g(\bar{u})(x) = 0$ on $I_\delta \supset I_h$ we get

$$\begin{aligned}
\langle h, \hat{M}_k h\rangle &= \langle h, d_k' g_k d_k^{2r-1} h\rangle + \langle d_k^r h, B_k d_k^r h\rangle \\
&\leq \|d_k'\|_\infty \|d_k\|_\infty^{2r-1} \|\chi_{I_h} g_k\|_{p'} \|h\|_\infty \|h\|_p + \langle d^r(\bar{u})\tilde{h}_{I_h}, B_k d^r(\bar{u})\tilde{h}_{I_h}\rangle \\
&\leq c_{d'} c_d^{2r-1} \|\chi_{I_h} g_k\|_{p'} \|h\|_\infty \|h\|_p + \langle \tilde{h}_{I_h}, \hat{M}(\bar{u})\tilde{h}_{I_h}\rangle \\
&\quad + \|d(\bar{u})\|_\infty^{2r} \|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \|\tilde{h}_{I_h}\|_p^2 \\
&\leq c_{d'} c_d^{2r-1} \|g_k - g(\bar{u})\|_{p'} \|h\|_\infty \|h\|_p + \langle \tilde{h}, \hat{M}(\bar{u})\tilde{h}\rangle - \langle \tilde{h} + \tilde{h}_{I_h}, \hat{M}(\bar{u})\tilde{h}_{I_\delta \setminus I_h}\rangle \\
&\quad + c_d^{2r} \|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \|\tilde{h}_{I_h}\|_p^2 \\
&\leq c_{d'} c_d^{2r-1} \|g_k - g(\bar{u})\|_{p'} \|h\|_\infty \|h\|_p + \langle \tilde{h}, \hat{M}(\bar{u})\tilde{h}\rangle \\
&\quad + 2\|\tilde{h}\|_p \|\hat{M}(\bar{u})\|_{U,U'} \|\tilde{h}_{I_\delta \setminus I_h}\|_p + c_d^{2r} \|B_k - \nabla^2 f(\bar{u})\|_{U,U'} \|\tilde{h}_{I_h}\|_p^2 \\
&\leq \Big(2c_{d'} c_d^{2r-1} \gamma^2 \|\tilde{h}\|_\infty \|g_k - g(\bar{u})\|_{p'} - \frac{\lambda}{2\gamma^2} + 8\gamma^2 \|\hat{M}(\bar{u})\|_{U,U'} \|\tilde{h}_{I_\delta \setminus I_h}\|_p \\
&\quad + c_d^{2r} \gamma^2 \|B_k - \nabla^2 f(\bar{u})\|_{U,U'}\Big) \|h\|_p^2.
\end{aligned}$$

In the derivation of the last inequality we have used (35), (36), (37), and $\|\tilde{h}\|_p = 1$. We have already shown that $\|\tilde{h}_{I_\delta \setminus I_h}\|_p$ can be made arbitrarily small by making $\varepsilon > 0$ small. By continuity the same is true for $\|g_k - g(\bar{u})\|_{p'}$ and by (A7) for $\|B_k - \nabla^2 f(\bar{u})\|_{U,U'}$ (since $\|u_k - \bar{u}\|_p \leq \varepsilon$). Hence, there exist $\varepsilon > 0$ and $\hat{\lambda} > 0$ such that for all u_k with $\|u_k - \bar{u}\|_p \leq \varepsilon$ we can carry out the above construction to obtain $h \in V \setminus \{0\}$ with

$$\langle h, \hat{M}_k h\rangle \leq -\hat{\lambda}\|h\|_p^2.$$

Since $h \neq 0$, $\|h\|_\infty \leq \gamma \|\tilde{h}\|_\infty$ and $\|h\|_p \geq \dfrac{1}{2\gamma}$, where γ only depends on δ, we get

$$\frac{\|h\|_\infty}{\|h\|_p} \leq 2\gamma^2 \|\tilde{h}\|_\infty \stackrel{\text{def}}{=} C.$$

In addition, we have by construction $I_h \subset \{x \in \Omega : \ a(x) + \delta/4 \leq u_k(x) \leq b(x) - \delta/4\}$ and consequently

$$u_k + \tau \frac{\delta}{4Cc_d^r} d_k^r \frac{h}{\|h\|_p} \in \mathcal{B} \ \text{ for all } \ \tau \in [-1,1].$$

Setting $\alpha = \dfrac{\delta}{4Cc_d^r}$ and renorming h to unity completes the proof. □

Now we establish the required decrease estimate.

LEMMA 7.2. *Let assumption* (W) *hold and s_k satisfy* (33). *If for u_k there exist $\hat{\lambda}, \alpha > 0$, $h_k \in V$, $\|h_k\|_p = 1$, with $u_k + \tau \alpha d_k^r h_k \in \mathcal{B}$ for all $\tau \in [-1,1]$ and*

$$\langle h_k, \hat{M}_k h_k \rangle \leq -\hat{\lambda} \|h_k\|_p^2,$$

then

$$pred_k(s_k) \geq -\psi_k(s_k) \geq \frac{\beta \hat{\lambda}}{2} \min \left\{ \frac{\Delta_k^2}{c_w^2}, \alpha^2 \right\}. \tag{38}$$

Proof. The first inequality is obvious. Now let $\hat{\lambda}, \alpha > 0$ be given. For all u_k which admit $h_k \in V$, $\|h_k\|_p = 1$, with $u_k \pm \alpha d_k^r h_k \in \mathcal{B}$ and $\langle h_k, \hat{M}_k h_k \rangle \leq -\hat{\lambda} \|h_k\|_p^2$, set

$$\hat{s}_k^n = \pm \min \{\Delta_k / c_w, \alpha\} h_k \ \text{ and } \ s_k^n = d_k^r \hat{s}_k^n,$$

and choose the sign such that $\langle \hat{s}_k^n, \hat{g}_k \rangle \leq 0$. Then $\|w_k s_k^n\|_p \leq \Delta_k$ by assumption (W) and $u_k + s_k^n \in \mathcal{B}$. Hence s_k^n is admissible for (33b) and can be used to get an upper bound for $\psi_k(s_k)$: The fraction of optimal decrease condition (33) gives

$$\begin{aligned} \psi_k(s_k) &\leq \beta \psi_k(s_k^n) = \beta \hat{\psi}_k(\hat{s}_k^n) = \beta \langle \hat{s}_k^n, \hat{g}_k \rangle + \frac{\beta}{2} \langle \hat{s}_k^n, \hat{M}_k \hat{s}_k^n \rangle \leq \frac{\beta}{2} \langle \hat{s}_k^n, \hat{M}_k \hat{s}_k^n \rangle \\ &\leq -\frac{\beta \hat{\lambda}}{2} \|\hat{s}_k^n\|_p^2 = -\frac{\beta \hat{\lambda}}{2} \min \left\{ \frac{\Delta_k^2}{c_w^2}, \alpha^2 \right\}. \end{aligned}$$

□

For a large class of trust–region algorithms for unconstrained finite–dimensional problems Shultz, Schnabel, and Byrd [18] proposed a very elegant way to prove that all accumulation points of the iterates satisfy the second–order necessary optimality conditions. The key idea is to increase the trust–region radius after exceedingly successful steps (case 4. in Algorithm 4.6). The following convergence theorem is an analogue to [18, Thm 3.2].

THEOREM 7.3. *Let assumptions* (A1)–(A7), (D1)–(D5), *and* (W) *hold. Moreover, let the sequence (u_k) be generated by the Algorithm* 4.8 *and let all s_k satisfy* (33). *Then every accumulation point $\bar{u} \in U$ of (u_k) satisfies the second–order necessary conditions* (O1)–(O3).

Proof. Let $\bar{u} \in U$ be an accumulation point of u_k. Then $\bar{u} \in \mathcal{B}$ and, since $\hat{g} : \mathcal{B} \subset U \longrightarrow U$ is continuous, $\hat{g}(\bar{u}) = 0$ by Theorem 6.7. Using Lemma 3.2, this implies (O1) and (O2).

Now assume that (O3) does not hold at $\bar{u}$. Then due to Theorem 4.5 there are $\bar{h} \in V$, $\bar{h} \neq 0$, and $\lambda > 0$ with $\langle \bar{h}, \hat{M}(\bar{u})\bar{h}\rangle \leq -\lambda \|\bar{h}\|_p^2$. Lemmas 7.1 and 7.2 yield $\alpha, c_7, \varepsilon > 0$ with $pred_k(s_k) \geq c_7 \min\{\Delta_k^2, \alpha^2\}$ for all u_k satisfying $\|u_k - \bar{u}\|_p \leq \varepsilon$. By choosing $0 < \Delta \leq \alpha$ we achieve that for all k with $\Delta_k \leq \Delta$ and $\|u_k - \bar{u}\|_p \leq \varepsilon$

$$pred_k(s_k) \geq c_7 \Delta_k^2.$$

Using this estimate, (A4), (A7), and $\|s_k\|_p \leq \beta_0 c_{w'} \Delta_k$ (see (28)) we find – possibly after reducing ε – with appropriate $\tau_k \in [0,1]$

$$\begin{aligned} pred_k(s_k)\,|\rho_k - 1| &= \left| f(u_k + s_k) - f_k + \frac{1}{2}\langle s_k, C_k s_k\rangle - \psi_k(s_k) \right| \\ &= \frac{1}{2}\left| \langle s_k, (\nabla^2 f(u_k + \tau_k s_k) - B_k) s_k\rangle \right| \\ &\leq \frac{1}{2}\left(\|\nabla^2 f(u_k + \tau_k s_k) - \nabla^2 f(\bar{u})\|_{U,U'} + \|\nabla^2 f(\bar{u}) - B_k\|_{U,U'} \right) \|s_k\|_p^2 \\ &\leq (1 - \eta_3) c_7 \Delta_k^2 \leq (1 - \eta_3) pred_k(s_k). \end{aligned}$$

This shows $\rho_k \geq \eta_3$ for all k with $\Delta_k \leq \Delta$ and $\|u_k - \bar{u}\|_p \leq \varepsilon$ and hence $\Delta_{k+1} \in [\gamma_2 \Delta_k, \gamma_3 \Delta_k]$.

For all $K > 0$ there is $l > K$ with $\|u_l - \bar{u}\|_p \leq \varepsilon/2$ and $\rho_l > \eta_1$. In fact, since $\bar{u}$ is an accumulation point of (u_k), we can find $l' > K$ with $\|u_{l'} - \bar{u}\|_p \leq \varepsilon/2$. Now $\rho_k \leq \eta_1$ for all $k \geq l'$ cannot occur, because then $\Delta_k \leq \gamma_1^{k-l'} \Delta_{l'}$ eventually satisfies $\Delta_k \leq \Delta$ and consequently $\rho_k \geq \eta_3 > \eta_1$. Hence, there is $l \geq l' > K$ with $u_l = u_{l'}$ and $\rho_l > \eta_1$.

Since $\Delta_{k+1} \geq \gamma_2 \Delta_k$ for all k with $\|u_k - \bar{u}\|_p \leq \varepsilon$ and $\Delta_k \leq \Delta$, it is easily seen that

1. $\Delta_l > \Delta$ or
2. $\Delta_l \leq \Delta$ and there is $m > l$ such that $\|u_k - \bar{u}\|_p \leq \varepsilon$ and $\Delta_k \leq \Delta$ for $l \leq k < m$, and
   2.1 $\Delta_m > \Delta$ or
   2.2 $\Delta_m \leq \Delta$ and $\|u_m - \bar{u}\|_p > \varepsilon$.

In case 1. we get

$$f_l - f_{l+1} > \eta_1 c_7 \min\{\Delta_l^2, \alpha^2\} \geq \eta_1 c_7 \Delta^2.$$

For 2.1. we have $\Delta \geq \Delta_{m-1} \geq \Delta_m/\gamma_3 > \Delta/\gamma_3$, and $\rho_{m-1} \geq \eta_3$, hence

$$f_{m-1} - f_m \geq \eta_3 c_7 \Delta_{m-1}^2 \geq \eta_3 c_7 \frac{\Delta^2}{\gamma_3^2}.$$

In case 2.2. we get $\Delta_{k+1} \geq \gamma_2 \Delta_k$, $k = l, \ldots, m-1$. This implies $\Delta_k \leq \gamma_2^{k-m+1} \Delta_{m-1}$ and

$$\begin{aligned}\frac{\varepsilon}{2} &\leq \|u_m - \bar{u}\|_p - \|u_l - \bar{u}\|_p \leq \|um - u_l\|_p = \left\| \sum_{k=l}^{m-1} s_k \right\|_p \\ &\leq \sum_{k=l}^{m-1} \|s_k\|_p \leq \beta_0 c_{w'} \sum_{k=l}^{m-1} \Delta_k \leq \beta_0 c_{w'} \Delta_{m-1} \sum_{k=l}^{m-1} \gamma_2^{k-m+1} \leq \beta_0 c_{w'} \Delta_{m-1} \frac{\gamma_2}{\gamma_2 - 1}.\end{aligned}$$

This yields

$$f_{m-1} - f_m \geq \eta_3 c_7 \Delta_{m-1}^2 \geq \eta_3 c_7 \left(\frac{\varepsilon(\gamma_2 - 1)}{2\beta_0 c_{w'} \gamma_2} \right)^2.$$

Therefore, we get for infinitely many steps k a decrease $f_k - f_{k+1}$ of at least a constant value which yields $f_k \to -\infty$. This contradicts the boundedness of f on $\mathcal{B}$, which follows from (A1)–(A3). Thus, (O3) must hold at $\bar{u}$. □

8. Conclusions and future work. We have introduced and analyzed a globally convergent class of interior–point trust–region algorithms for infinite–dimensional nonlinear optimization subject to pointwise bounds in function space. The methods are generalizations of those presented by Coleman and Li [6] for finite–dimensional problems. We have extended all first– and second–order global convergence results that are available for the finite–dimensional setting to our infinite–dimensional L^p-Banach space framework. The analysis was carried out in a unified way for $2 \leq p \leq \infty$. The lack of the equivalence of norms required the development of new proof techniques. This is also a valuable contribution to the finite–dimensional theory because our results are derived completely without using norm equivalences and hence are almost independent of the problem dimension. In this sense our convergence theory can be considered to be mesh–independent. Moreover, we have carried out our analysis for a very general class of affine scaling operators, and almost arbitrary scaling of the trust–region. This is new also from the finite–dimensional viewpoint. Numerical results for optimal control problems governed by a nonlinear parabolic PDE which prove the efficiency of our algorithms can be found in the forthcoming paper [22]. Furthermore, we present therein results for finite–dimensional standard test–examples compiled in [8] which verify that a combination of the findings in this work and [22] yield improvements also for finite–dimensional problems. Our investigations suggest to incorporate a projection onto the box in the computation of approximate solutions of the trust–region subproblems. This new technique was tested in an implementation of the methods described in [10], [13], and [23], and proved to be superior to other choices.

The results of this paper and [22] represent a first important step towards a rigorous justification why trust–region interior–point and trust–region interior–point SQP methods perform so well on discretized control problems. See [10], [13], and [22] for applications. The extension of our theory to methods with additional equality constraints is in progress.

Acknowledgements. This work was done while the first and second author were visiting the Department of Computational and Applied Mathematics and the Center for Research on Parallel Computation, Rice University. They are greatly indebted

to John Dennis for giving them the opportunity to work in this excellent research environment.

We would like to thank Richard Byrd, Colorado State University, and Philippe Toint, Facultés Universitaires Notre-Dame de la Paix, for pointing us to the second–order convergence result in [18] which led to an improvement of generality and elegance in our presentation. We also are grateful to John Dennis, Rice University, and Luís Vicente, Universidade de Coimbra, for their helpful suggestions.

9. Appendix. In this section we present proofs of Lemma 6.5 and 6.6. These proofs require the following three technical results:

LEMMA 9.1. *For $0 < r \leq 1$, $1 \leq q \leq \infty$, $v_1, v_2 \in L^q(\Omega)$, $v_1, v_2 \geq 0$, the following holds:*

$$\|v_1^r - v_2^r\|_q \leq m_{q,q/r} \|v_1 - v_2\|_q^r. \tag{39}$$

Proof. For $r = 1$ the assertion is trivial. For $\alpha, \beta \geq 0$, $0 < r < 1$, we use the estimate

$$|\alpha^r - \beta^r| \leq |\alpha - \beta|^r. \tag{40}$$

This estimate can be seen as follows. Due to symmetry we may assume that $\alpha \geq \beta \geq 0$. The function $h(\alpha) = |\alpha - \beta|^r - |\alpha^r - \beta^r|$ satisfies $h(\beta) = 0$,

$$h'(\alpha) = r\left((\alpha - \beta)^{r-1} - \alpha^{r-1}\right) \geq 0 \quad (\alpha > \beta)$$

and, thus, $h(\alpha) \geq 0$ for all $\alpha \geq \beta$.

In the case $q = \infty$ the assertion follows immediately from (40). For $1 \leq q < \infty$ we use Lemma 5.1 to get

$$\begin{aligned} \|v_1^r - v_2^r\|_q &\leq m_{q,q/r} \|v_1^r - v_2^r\|_{q/r} = m_{q,q/r} \left(\int_\Omega |v_1(x)^r - v_2(x)^r|^{\frac{q}{r}} \, dx \right)^{\frac{r}{q}} \\ &\leq m_{q,q/r} \left(\int_\Omega |v_1(x) - v_2(x)|^q \, dx \right)^{\frac{r}{q}} = m_{q,q/r} \|v_1 - v_2\|_q^r. \end{aligned}$$

This completes the proof. □

LEMMA 9.2. *For $r \geq 1$, $1 \leq q \leq \infty$, $v_1, v_2 \in V$, $v_1, v_2 \geq 0$, the following inequality holds:*

$$\|v_1^r - v_2^r\|_q \leq r \max\{\|v_1\|_\infty, \|v_2\|_\infty\}^{r-1} \|v_1 - v_2\|_q. \tag{41}$$

Proof. In the case $r = 1$ there is nothing to show. First we prove that for all $r > 1$, $\alpha, \beta \in [0, \gamma]$, $\gamma > 0$, we have $h(\alpha) \stackrel{\text{def}}{=} r\gamma^{r-1}|\alpha - \beta| - |\alpha^r - \beta^r| \geq 0$. In fact, we may assume $\alpha \geq \beta$ and compute $h(\beta) = 0$,

$$h'(\alpha) = r(\gamma^{r-1} - \alpha^{r-1}) \geq 0 \quad (\beta \leq \alpha \leq \gamma).$$

Therefore,

$$|v_1^r(x) - v_2^r(x)| \leq r \max\{\|v_1\|_\infty, \|v_2\|_\infty\}^{r-1} |v_1(x) - v_2(x)| \quad \text{for all } x \in \Omega$$

which immediately implies (41). □

LEMMA 9.3. *Let $\alpha_1, \ldots, \alpha_n$, and $\beta_1, \ldots, \beta_n$ be arbitrary real numbers. Then*

$$|\min\{\alpha_1, \ldots, \alpha_n\} - \min\{\beta_1, \ldots, \beta_n\}| \leq \max\{|\alpha_1 - \beta_1|, \ldots, |\alpha_n - \beta_n|\}$$

Proof. Without restriction, let $\beta_k = \min\{\beta_1, \ldots, \beta_n\} \leq \min\{\alpha_1, \ldots, \alpha_n\}$. Then the assertion follows from

$$|\min\{\alpha_1, \ldots, \alpha_n\} - \min\{\beta_1, \ldots, \beta_n\}| = \min\{\alpha_1, \ldots, \alpha_n\} - \beta_k \leq \alpha_k - \beta_k.$$

□

9.1. Proof of Lemma 6.5. We write $\|\cdot\|_{q,A}$ for $\|\chi_A \cdot\|_q$, $A \subset \Omega$ measurable. For arbitrary $u, \tilde{u} \in \mathcal{B}$ set $N = \{x \in \Omega : g(u)(x)g(\tilde{u})(x) > 0\}$. The triangle inequality gives the following estimate

$$\begin{aligned}
\|\hat{g}(u) - \hat{g}(\tilde{u})\|_{p'} &= \|d^r(u)g(u) - d^r(\tilde{u})g(\tilde{u})\|_{p'} \\
&\leq \|d(u)\|_\infty^r \|g(u) - g(\tilde{u})\|_{p'} + \|(d^r(u) - d^r(\tilde{u}))g(\tilde{u})\|_{p'} \\
&\leq \|d(u)\|_\infty^r \|g(u) - g(\tilde{u})\|_{p'} + \|d^r(u) - d^r(\tilde{u})\|_\infty \|g(\tilde{u})\|_{p',\Omega\setminus N} \\
&\quad + \|g(\tilde{u})\|_\infty \|d^r(u) - d^r(\tilde{u})\|_{p',N}.
\end{aligned}$$

We use the fact that $|g(u) - g(\tilde{u})| \geq |g(\tilde{u})|$ on $\Omega \setminus N$ and obtain

$$\begin{aligned}
\|\hat{g}(u) - \hat{g}(\tilde{u})\|_{p'} &\leq (\|d(u)\|_\infty^r + \|d^r(u) - d^r(\tilde{u})\|_\infty)\|g(u) - g(\tilde{u})\|_{p'} \\
&\quad + \|g(\tilde{u})\|_\infty \|d^r(u) - d^r(\tilde{u})\|_{p',N} \\
&\leq 3c_d^r \|g(u) - g(\tilde{u})\|_{p'} + c_1 \|d^r(u) - d^r(\tilde{u})\|_{p',N}.
\end{aligned}$$

Now the (uniform) continuity of $\hat{g}$ follows from Lemma 9.1, Lemma 9.2, the (uniform) continuity of g, and the assumption $\|\chi_N(d(u) - d(\tilde{u}))\|_{p'} \longrightarrow 0$ (uniformly in u) on the scaling. □

9.2. Proof of Lemma 6.6. We restrict ourselves to the more complicated case $d = d_\Pi$. The result follows from Lemma 6.5 if we verify that

$$\|\chi_{\{g(u)g(\tilde{u})>0\}}(d(u) - d(\tilde{u}))\|_{p'} \longrightarrow 0 \quad \text{as } \tilde{u} \longrightarrow u \text{ (uniformly in } u).$$

Let $u, \tilde{u} \in \mathcal{B}$ be arbitrary. Using symmetries, it is easily seen that we are done if we are able to establish appropriate upper bounds for $|d(u)(x) - d(\tilde{u})(x)|$ for the three cases that $g(u)(x) > 0$, $g(\tilde{u})(x) > 0$ and

a) $d_\Pi(u)(x)$ and $d_\Pi(\tilde{u})(x)$ are both determined by the second case in (10),
b) $d_\Pi(u)(x)$ and $d_\Pi(\tilde{u})(x)$ are both determined by the else-case in (10),
c) $d_\Pi(u)(x)$ is determined by the second and $d_\Pi(\tilde{u})(x)$ by the else-case in (10).

Set $\rho(x) = |d_\Pi(u)(x) - d_\Pi(\tilde{u})(x)|$. We will use Lemma 9.3 several times.

Case a):

$$\rho(x) = |\min\{g(u)(x), c(x)\} - \min\{g(\tilde{u})(x), c(x)\}| \leq |g(u)(x) - g(\tilde{u})(x)|.$$

Case b):

$$\begin{aligned}\rho(x) &= |\min\{u(x)-a(x), b(x)-u(x), c(x)\} \\ &\quad - \min\{\tilde{u}(x)-a(x), b(x)-\tilde{u}(x), c(x)\}| \\ &\leq |u(x)-\tilde{u}(x)|.\end{aligned}$$

Case c): From $b(x)-u(x) \leq u(x)-a(x)$ follows $u(x)-a(x) \geq c(x)$ and therefore

$$\begin{aligned}d_\Pi(u)(x) &= \min\{u(x)-a(x), g(u)(x), c(x)\} \\ &\geq \min\{u(x)-a(x), b(x)-u(x), c(x)\}.\end{aligned}$$

If $g(\tilde{u})(x) > b(x)-\tilde{u}(x)$ then $b(x)-\tilde{u}(x) > \tilde{u}(x)-a(x)$ and hence

$$d_\Pi(\tilde{u})(x) = \min\{\tilde{u}(x)-a(x), g(\tilde{u})(x), c(x)\}.$$

Therefore, we obtain

$$\rho(x) \leq \max\{|u(x)-\tilde{u}(x)|, |g(u)(x)-g(\tilde{u})(x)|\}.$$

Otherwise, if $g(\tilde{u})(x) \leq b(x)-\tilde{u}(x)$, we have in the case $d_\Pi(u)(x) \geq d_\Pi(\tilde{u})(x)$ that

$$\begin{aligned}\rho(x) &\leq \min\{u(x)-a(x), g(u)(x), c(x)\} - \min\{\tilde{u}(x)-a(x), g(\tilde{u})(x), c(x)\} \\ &\leq \max\{|u(x)-\tilde{u}(x)|, |g(u)(x)-g(\tilde{u})(x)|\},\end{aligned}$$

and for $d_\Pi(u)(x) < d_\Pi(\tilde{u})(x)$ we get

$$\begin{aligned}\rho(x) &\leq \min\{\tilde{u}(x)-a(x), b(x)-\tilde{u}(x), c(x)\} \\ &\quad - \min\{u(x)-a(x), b(x)-u(x), c(x)\} \\ &\leq |u(x)-\tilde{u}(x)|.\end{aligned}$$

Taking all cases together, this shows that

$$\begin{aligned}\|\chi_{\{g(u)g(\tilde{u})>0\}}\rho\|_{p'} &\leq \|u-\tilde{u}\|_{p'} + \|g(u)-g(\tilde{u})\|_{p'} \\ &\leq m_{p',p}\|u-\tilde{u}\|_p + \|g(u)-g(\tilde{u})\|_{p'}.\end{aligned}$$

Now, the application of Lemma 6.5 shows that $\hat{g}$ inherits the (uniform) continuity of g. □

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] M. A. Branch, T. F. Coleman, and Y. Li, *A subspace, interior, and conjugate gradient method for large–scale bound–constrained minimization problems*, CTC95TR217, Center for Theory and Simulation in Science and Engineering, Cornell University, Ithaca, NY 14853–3801, 1995. Available via the URL `http://www.tc.cornell.edu/Research/Tech.Reports/index.html`.

[3] J. Burger and M. Pogu, *Functional and numerical solution of a control problem originating from heat transfer*, J. Optim. Theory Appl., 68 (1991), pp. 49–73.

[4] J. V. Burke, J. J. Moré, and G. Toraldo, *Convergence properties of trust region methods for linear and convex constraints*, Math. Programming, 47 (1990), pp. 305–336.

[5] T. F. Coleman and Y. Li, *On the convergence of interior–reflective Newton methods for nonlinear minimization subject to bounds*, Math. Programming, 67 (1994), pp. 189–224.

[6] ———, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM J. Optimization, 6 (1996), pp. 418–445.

[7] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Global convergence of a class of trust region algorithms for optimization with simple bounds*, SIAM J. Numer. Anal., 25 (1988), pp. 433–460. See [9].

[8] ———, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Math. Comp., 50 (1988), pp. 399–430.

[9] ———, *Correction to the paper on global convergence of a class of trust region algorithms for optimization with simple bounds*, SIAM J. Numer. Anal., 26 (1989), pp. 764–767.

[10] J. E. Dennis, M. Heinkenschloss, and L. N. Vicente, *Trust–region interior–point algorithms for a class of nonlinear programming problems*, Tech. Rep. TR94–45, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005–1892, 1994. Available via the URL `http://www.caam.rice.edu/~trice/trice_soft.html`.

[11] J. E. Dennis and L. N. Vicente, *Trust–region interior–point algorithms for minimization methods with simple bounds*, in Applied Mathematics and Parallel Computing, Festschrift for Klaus Ritter, H. Fischer, B. Riedmüller, and S. Schäffler, eds., Heidelberg, 1996, Physica–Verlag, pp. 97–107.

[12] J. C. Dunn, *On l^2 sufficient conditions and the gradient projection method for optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1270–1290.

[13] M. Heinkenschloss and L. N. Vicente, *Analysis of inexact trust–region interior–point SQP algorithms*, Tech. Rep. TR95–18, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005–1892, 1995. Available via the URL `http://www.caam.rice.edu/~trice/trice_soft.html`.

[14] C. T. Kelley and E. W. Sachs, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Scientific Computing, 15 (1994), pp. 645–667.

[15] ———, *A trust region method for parabolic boundary control problems*, Tech. Rep. CRSC–TR96–28, Center for Research in Scientific Computing, North Carolina State University, Raleigh, NC, 1996. Available via the URL `http://www4.ncsu.edu/eos/users/ctkelley/www/pubs.html`.

[16] H. Maurer and J. Zowe, *First and second–order necessary and sufficient optimality conditions for infinite–dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[17] J. J. Moré, *Recent developments in algorithms and software for trust region methods*, in Mathematical Programming, The State of The Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer Verlag, Berlin, Heidelberg, New York, 1983, pp. 258–287.

[18] G. A. Shultz, R. B. Schnabel, and R. H. Byrd, *A family of trust region based algorithms for unconstrained minimization with strong global convergence properties*, SIAM J. Numer. Anal., 22 (1985), pp. 47–67.

[19] D. C. Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.

[20] T. Tian and J. C. Dunn, *On the gradient projection method for optimal control problems with nonnegative L^2 inputs*, SIAM J. Control Optim., 32 (1994), pp. 516–552.

[21] P. L. Toint, *Global convergence of a class of trust–region methods for nonconvex minimization in Hilbert space*, IMA Journal of Numerical Analysis, 8 (1988), pp. 231–252.

[22] M. Ulbrich and S. Ulbrich, *Superlinear convergence of affine–scaling interior–point Newton methods for infinite–dimensional nonlinear problems with pointwise bounds*, TR97–05, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005–1892, 1997. Available via the URL `http://www.statistik.tu-muenchen.de/LstAMS/sulbrich/papers/papers.html`.

[23] L. N. Vicente, *Trust–region interior–point algorithms for a class of nonlinear programming problems*, PhD thesis, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005–1892, 1996. Available via the URL `http://www.mat.uc.pt/~lvicente/papers/papers.html`.