Superlinear Convergence of Affine-Scaling Interior-Point Newton Methods for Infinite-Dimensional Nonlinear Problems with Pointwise Bounds

Michael Ulbrich and Stefan Ulbrich

CRPC-TR97697
Revised October 1997

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

Submitted in September 1997

SUPERLINEAR CONVERGENCE OF AFFINE-SCALING INTERIOR-POINT NEWTON METHODS FOR INFINITE-DIMENSIONAL NONLINEAR PROBLEMS WITH POINTWISE BOUNDS

MICHAEL ULBRICH * AND STEFAN ULBRICH †

Abstract. We develop and analyze a superlinearly convergent affine-scaling interior-point Newton method for infinite-dimensional problems with pointwise bounds in L^p-space. The problem formulation is motivated by optimal control problems with L^p-controls and pointwise control constraints. The finite-dimensional convergence theory by Coleman and Li (*SIAM J. Optim.*, 6 (1996), pp. 418–445) makes essential use of the equivalence of norms and the exact identifiability of the active constraints close to an optimizer with strict complementarity. Since these features are not available in our infinite-dimensional framework, algorithmic changes are necessary to ensure fast local convergence. The main building block is a Newton-like iteration for an affine-scaling formulation of the KKT-condition. We demonstrate in an example that a stepsize rule to obtain an interior iterate may require very small stepsizes even arbitrarily close to a nondegenerate solution. Using a pointwise projection instead we prove superlinear convergence under a weak strict complementarity condition and convergence with Q-rate >1 under a slightly stronger condition if a smoothing step is available. We discuss how the algorithm can be embedded in the class of globally convergent trust-region interior-point methods recently developed by M. Heinkenschloss and the authors. Numerical results for the control of a heating process confirm our theoretical findings.

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

AMS subject classifications. 49K27, 49M15, 49M37, 65K05, 90C30, 90C48

1. Introduction. We introduce an affine-scaling interior-point Newton method for the solution of the infinite-dimensional nonlinear optimization problem

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

and study its local convergence behavior in detail. Here $\Omega \subset \mathbb{R}^n$ is a domain with positive and finite Lebesgue measure $0 < \mu(\Omega) < \infty$, and

$$L^q = L^q(\Omega) \quad , \quad 1 \le q \le \infty,$$

denotes the usual Banach space of (equivalence classes of) real-valued measurable functions for which the norm

$$\|u\|_q \stackrel{\text{def}}{=} \left(\int_\Omega |u(x)|^q \, dx \right)^{1/q} \quad (q < \infty) \quad , \quad \|u\|_\infty \stackrel{\text{def}}{=} \operatorname*{ess\,sup}_{x \in \Omega} |u(x)|$$

* Lehrstuhl für Angewandte Mathematik und Mathematische Statistik, Zentrum Mathematik, Technische Universität München, 80290 München, Germany (mulbrich@statistik.tu-muenchen.de). This author was supported by Deutsche Forschungsgemeinschaft under Grant Ul157/1-1 and by the North Atlantic Treaty Organization under Grant CRG 960945.

† Lehrstuhl für Angewandte Mathematik und Mathematische Statistik, Zentrum Mathematik, Technische Universität München, 80290 München, Germany (sulbrich@statistik.tu-muenchen.de). This author was supported by Deutsche Forschungsgemeinschaft under Grant Ul158/1-1 and by the North Atlantic Treaty Organization under Grant CRG 960945.

is bounded. Let $2 \leq p \leq \infty$ and assume that the objective function $f : \mathcal{D} \longrightarrow \mathbb{R}$ is continuous on an open neighborhood $\mathcal{D} \subset L^p$ of $\mathcal{B}$. Additional requirements on f will be given below. The lower and upper bound functions $a, b \in L^\infty$ are assumed to have positive distance from each other, i.e.

$$\operatorname*{ess\,inf}_{x \in \Omega} (b(x) - a(x)) > 0.$$

Then $\mathcal{B}$ has a nonempty L^∞-interior

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

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 [5], Kelley, Sachs [15], Sachs [22], and Tian, Dunn [23].

The algorithm presented in this paper is based on the application of a Newton-like iteration to an affine-scaling formulation of the first-order necessary optimality conditions. For finite-dimensional problems this class of algorithms has been introduced and analyzed by Coleman and Li [6], [7]. Extensions to problems with additional equality constraints were studied in Dennis, Heinkenschloss, Vicente [8], Heinkenschloss, Vicente [13], and Vicente [25], [26]. In all of the above papers except for [26] the affine-scaling Newton iteration is embedded in a trust-region interior-point algorithm to achieve global convergence. In a recent paper (Ulbrich, Ulbrich, Heinkenschloss [24]) we extended the finite-dimensional global convergence theory of Coleman and Li [7] for trust-region interior-point algorithms to the infinite-dimensional problem class (P). The present paper continues these investigations and focuses on the local superlinear convergence of a closely related affine-scaling interior-point Newton method which plays the same important role in our setting as the ordinary Newton method does in the local analysis of trust-region algorithms for unconstrained optimization. Problem (P) is a special type of cone constrained optimization problems in Banach space. For this very general class of problems Alt [2] developed a Lagrange-Newton-SQP method and proved quadratic convergence. A drawback of SQP-type methods consists in the fact that in each step a linearly cone-constrained quadratic problem or, equivalently, a linear generalized equation has to be solved. In our setting each SQP-subproblem would have the form (P) with the objective f replaced by a quadratic approximation. The solution of these problems is by no means trivial and requires a multiple of the effort needed to perform a Newton-like step. Therefore, although SQP-methods are quadratically convergent, their efficiency crucially depends on the availability of fast solvers for the subproblems.

During the last fifteen years several attempts have been undertaken to develop algorithms for which each iteration is not much more expensive than an ordinary Newton step. One of these is the projected Newton method which was introduced by Bertsekas [3] for finite-dimensional bound-constrained problems. Kelley and Sachs [15] extended this method to problems of type (P) with special structure and proved local convergence with Q-rate $1 + \beta$, $0 < \beta < 1$. The class of problems addressed in [15] is essentially the same as the one discussed in §8 of this work. Although it is possible in the finite-dimensional case to prove quadratic convergence, see [3], Kelley and Sachs could not establish this result in their infinite-dimensional setting. In this paper we develop local convergence results for infinite-dimensional affine-scaling interior-point Newton methods which are similar to those by Kelley and Sachs [15] for projected

Newton methods. Like Kelley and Sachs, we observe a gap between the achievable convergence rate in the finite- and infinite-dimensional setting. Our theory covers a more comprehensive problem class and requires weaker assumptions than that for projected Newton methods in [15]. The cost for one iteration of our algorithm is dominated by the solution of a linear equation and is therefore comparable to that of a projected Newton step.

The development of a local convergence theory for our infinite-dimensional setting turns out to be much more delicate than in the finite-dimensional case. First of all, strict complementarity, i.e. $g(\bar{u})(x) \neq 0$ for a.a. $x \in \Omega$ with $\bar{u}(x) \in \{a(x), b(x)\}$, at a local solution $\bar{u} \in \mathcal{B}$ of (P) does not guarantee that the absolute value of the gradient $g(\bar{u})$ is uniformly bounded away from zero on the active set. As a consequence, even for $u \in \mathcal{B}$ arbitrarily close to $\bar{u}$ the active set at $\bar{u}$ cannot be identified exactly by means of the information available at u. And, finally, since the L^t- and L^∞-norm, $1 \leq t < \infty$, are not equivalent, an iterate u_k may be very close to the solution $\bar{u}$ in L^t but still deviate substantially from $\bar{u}$ on a small set of nonzero measure. These are the main reasons why – in contrast to the finite-dimensional case – it seems not to be possible to achieve quadratic convergence in our general setting. This has an important effect on the expressiveness of the finite-dimensional quadratic convergence rate: Let (PD) be a finite-dimensional bound-constrained problem obtained by discretizing a problem of type (P). To compute an approximate solution of (P) we apply a finite-dimensional analogue of our affine-scaling interior-point Newton method to the discretized problem (PD). Then, under appropriate assumptions, the finite-dimensional convergence theory promises quadratic convergence, whereas on account of the close relationship to (P) and the infinite-dimensional convergence results we expect only superlinear instead of quadratic convergence. In fact, the convergence behavior is dominated by the infinite-dimensional theory until the iterates enter a neighborhood of the local solution $\bar{u}$ where the requirements for quadratic convergence are satisfied. Especially for fine discretizations this set is typically very small, because it is closely related to the neighborhood of $\bar{u}$ where the active set at $\bar{u}$ can be identified exactly. Hence, for increasingly accurate discretizations the domain of quadratic convergence will shrink whereas the domain of superlinear convergence will be stable. It is important to note that our convergence results require modifications of the finite-dimensional algorithm investigated in [6] and [7], especially the enforcement of strict feasibility by a modified projection instead of a stepsize rule. This argumentation shows that the development of efficient algorithms for the solution of infinite-dimensional optimization problems also leads to improved finite-dimensional methods.

In the following we give a rough outline of the theory developed in this paper. As mentioned above, the heart of our algorithm is a Newton-like step applied to the affine-scaling formulation $d(u)g(u) = 0$ of the Karush-Kuhn-Tucker (KKT) conditions. Here $d(u) \in L^\infty$ denotes a suitably chosen weighting function, the affine-scaling function. In the Newton equation the in general non-existing derivative of $u \longmapsto d(u)g(u)$ is replaced by an appropriate operator $G(u)$. If $u_k^s \in \mathcal{B}^\circ$ denotes the current (actually smoothed, see below) iterate then the affine-scaling Newton step reads

$$G(u_k^s)(u_{k+1}^n - u_k^s) = -d(u_k^s)g(u_k^s).$$

Under a regularity assumption on $G(u)$ we establish for suitable $q < s$ the estimate $\|u_{k+1}^n - \bar{u}\|_q = o(\|u_k^s - \bar{u}\|_s)$ if strict complementarity holds at the local solution $\bar{u}$ of

(P) and $\|u_{k+1}^n - \bar{u}\|_q \leq C\|u_k^s - \bar{u}\|_s^{1+\beta}$, $0 < \beta < 1$, if a slightly stronger strict complementarity condition is satisfied. This discrepancy of the norms is, among other things, caused by the fact that the complementarity can be arbitrarily weak on small sets. To overcome this difficulty we follow [15] and assume the availability of a smoothing step $S_k^\circ : \mathcal{B}^\circ \subset L^q \longrightarrow \mathcal{B}^\circ \subset L^s$, $u_k \longmapsto u_k^s = S_k^\circ(u_k)$ with $\|u_k^s - \bar{u}\|_s \leq C_S\|u_k - \bar{u}\|_q$. Moreover, since u_{k+1}^n may lie outside of $\mathcal{B}^\circ$, we define a back-transport $u \longmapsto P[u_k^s](u) \in \mathcal{B}^\circ$ by an interior-point modification of the pointwise projection onto $\mathcal{B}$. We will see that a stepsize rule is inappropriate in our framework, although it yields quadratic convergence in the finite-dimensional case. We prove that the combination

$$u_k \rightsquigarrow u_k^s = S_k^\circ(u^k) \rightsquigarrow u_{k+1}^n \rightsquigarrow u_{k+1} = P[u_k^s](u_{k+1}^n)$$

generates sequences (u_k) and (u_k^s) that converge superlinearly to $\bar{u}$ in L^q and L^s, respectively. If the stronger strict complementarity condition holds we prove convergence with Q-rate $1+\beta$. We apply our results to a class of problems with L^2-regularization for which a projected Newton method was analyzed in [15] and show that the assumptions therein imply ours. For this problem class a smoothing step can be derived from a fixed point formulation of the KKT-conditions. Moreover, we show that the second-order sufficiency condition of Dunn and Tian [9] implies our regularity assumption on G. Finally, we discuss how our algorithm can be embedded in the globally convergent class of trust-region interior-point methods recently introduced in [24]. The resulting method is applied to the boundary control of a heating process which was already considered in [5], [18].

This paper is organized as follows. In §2 we introduce some notation and put together several important estimates for L^p-spaces. Moreover, we resume the first-order necessary optimality conditions for problem (P) in standard- and affine-scaling formulation. Our particular choice of the affine-scaling function and the basic affine-scaling Newton step are introduced in §3. Here we also discuss why an iteration based on this step alone is in general neither well-defined nor convergent and sketch the idea of a smoothing step and a back-transport that take care of these problems. An outline of our algorithm and its convergence properties in a clearly arranged abstract setting is given in §4. In §5 we carry out a thorough analysis of the Newton-like step. In §6 the affine-scaling interior-point Newton algorithm is formulated. Moreover, we introduce a back-transport based on a pointwise projection onto $\mathcal{B}$, explain why in our infinite-dimensional setting a stepsize-rule is not suitable for a back-transport, and address the smoothing step. Our convergence results are presented in §7. In §8 we apply our results to a class of L^2-regularized problems and show that our assumptions are weaker than those used in [15]. In §9 we discuss the relationship between sufficient second-order conditions developed in [9] and the regularity assumptions we impose on the approximate derivative operator G. §10 addresses the question how our algorithm can be used to accelerate the globally convergent class of trust-region interior-point algorithms recently proposed in [24]. Finally, we present numerical results for the boundary control of a heating process in §11.

2. Preliminaries.

2.1. Notation. We write $B^c = \Omega \setminus B$ for the complement of a measurable set $B \subset \Omega$ and denote the characteristic function of B by χ_B, i.e. $\chi_B(x) = 1$ for $x \in B$, and $\chi_B(x) = 0$, otherwise. If $v : \Omega \longrightarrow \mathbb{R}$ is measurable then we set $v_B \stackrel{\text{def}}{=} \chi_B v$. Moreover, we write $\|\cdot\|_{t,B}$ for $\|\chi_B \cdot\|_t$, $1 \leq t \leq \infty$.

$\mathcal{L}(Y,Z)$ is the space of bounded linear operators from the Banach space Y into the Banach space Z. The operator norm on $\mathcal{L}(L^{q_1},L^{q_2})$ is denoted by $\|\cdot\|_{q_1,q_2}$. We write I for the identity operator $y \longmapsto y$. As representation of the dual space of L^t, $1 \le t < \infty$, we choose $L^{t'}$, $1/t + 1/t' = 1$, with the corresponding dual pairing $\langle v, w\rangle = \int_\Omega v(x)w(x)dx$, $v \in L^t$, $w \in L^{t'}$.

Our minimum assumptions on the objective function f are

ASSUMPTION.

(A1) $f : \mathcal{D} \subset L^p \longrightarrow \mathbb{R}$ is twice continuously Fréchet differentiable with derivatives

$$\begin{aligned} g \stackrel{\text{def}}{=} \nabla f &: \mathcal{D} \longrightarrow L^{p'} \\ \nabla^2 f &: \mathcal{D} \longrightarrow \mathcal{L}(L^p, L^{p'}) \end{aligned} \quad , \quad \frac{1}{p} + \frac{1}{p'} = 1.$$

Moreover, there is $C_g > 0$ such that $\|g(u)\|_\infty < C_g$ for all $u \in \mathcal{B}$.

2.2. Some inequalities. For convenience, we recall a couple of well known norm estimates for L^p-spaces.

LEMMA 2.1. *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 $1/\infty$ has to be interpreted as zero.

Proof. See e.g. [1, Thm. 2.8]. □

LEMMA 2.2 (INTERPOLATION INEQUALITY). *Given $1 \le q_1 \le q_2 \le \infty$ and $0 \le \theta \le 1$, let $1 \le q_0 \le \infty$ satisfy $1/q_0 = \theta/q_1 + (1-\theta)/q_2$. Then for all $v \in L^{q_2}$:*

$$\|v\|_{q_0} \le \|v\|_{q_1}^{\theta}\|v\|_{q_2}^{1-\theta} \tag{1}$$

Proof. See [24, Lem. 5.2]. □

LEMMA 2.3. *Let $q_0 \in [1,\infty]$ and $q_1, q_1' \in [1,\infty]$ with $1/q_1 + 1/q_1' = 1$ be given. Then for all $u \in L^{q_0 q_1}$ and $v \in L^{q_0 q_1'}$ we have*

$$\|uv\|_{q_0} \le \|u\|_{q_0 q_1}\|v\|_{q_0 q_1'}.$$

Proof. In the nontrivial case $q_0 < \infty$, apply Hölder's inequality:

$$\|u\|_{q_0 q_1}^{q_0}\|v\|_{q_0 q_1'}^{q_0} = \||u|^{q_0}\|_{q_1}\||v|^{q_0}\|_{q_1'} \ge \||u|^{q_0}|v|^{q_0}\|_1 = \|uv\|_{q_0}^{q_0}.$$

□

LEMMA 2.4. *For $v \in L^q$, $1 \le q < \infty$, and all $\delta > 0$ holds*

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

Proof.

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

□

2.3. Necessary optimality conditions. The method is based on an affine-scaling formulation of the first-order necessary optimality conditions. A detailed derivation of these conditions can be found in [24]. Therein we also prove second-order necessary conditions which are not needed in our context.

THEOREM 2.5 (FIRST-ORDER NECESSARY OPTIMALITY CONDITIONS, KARUSH-KUHN-TUCKER (KKT) CONDITIONS).
Let $\bar{u}$ be a local minimizer of problem (P) *and assume that f is differentiable at $\bar{u}$. In the case $p = \infty$ assume in addition that the gradient satisfies $g(\bar{u}) \in L^1$. Then*

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

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

are satisfied.

Proof. See [24, Thm. 3.1]. □

The inequality (O2) can be converted into an equation by pointwise multiplication with an *affine-scaling* function $d(\bar{u})$, where $d : \mathcal{B} \longrightarrow L^\infty$ satisfies

$$d(u)(x) \begin{cases} = 0 & \text{if } u(x) = a(x) \text{ and } g(u)(x) \geq 0, \\ = 0 & \text{if } u(x) = b(x) \text{ and } g(u)(x) \leq 0, \\ > 0 & \text{else} \end{cases} \tag{2}$$

for a.a. $x \in \Omega$. For details we refer to [24]. The idea was first introduced by Coleman and Li in [7] for the finite-dimensional case.

LEMMA 2.6. *Let $f : \mathcal{D} \subset L^p \longrightarrow \mathbb{R}$ be differentiable and $\bar{u} \in \mathcal{B}$. In the case $p = \infty$ assume in addition that the gradient satisfies $g(u) \in L^1$, $u \in \mathcal{B}$. Then* (O2) *is equivalent to*

$$d(\bar{u})g(\bar{u}) = 0 \tag{3}$$

for all d satisfying (2).

Proof. See [24, Lem. 3.2]. □

3. A Newton-like step. As for all efficient methods, we aim to apply Newton's method to a suitable formulation of the optimality system. In our approach we take equation (3) which, according to Lemma 2.6, is equivalent to the first-order necessary condition (O2). We use the freedom provided by (2) to choose the affine-scaling function d in such a way that dg is as smooth as possible in a neighborhood of a KKT-point $\bar{u}$ of (P). Since the function space analogue of the affine-scaling matrix of Coleman and Li [7],

$$d_{\mathrm{I}}(u)(x) \stackrel{\text{def}}{=} \begin{cases} u(x) - a(x) & \text{if } g(u)(x) > 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } u(x) - a(x) \leq b(x) - u(x), \\ b(x) - u(x) & \text{if } g(u)(x) < 0 \text{ or} \\ & g(u)(x) = 0 \text{ and } b(x) - u(x) < u(x) - a(x) \end{cases} \tag{4}$$

is not even continuous in u at a KKT-point $\bar{u}$, we work with a different choice for d. The discontinuity of d_{I} results from the fact that $|d_{\mathrm{I}}(u) - d_{\mathrm{I}}(\bar{u})| = b - a - |u - \bar{u}|$ on $\{x \in \Omega : g(u)(x)g(\bar{u})(x) < 0\}$. Nevertheless, our theory can be extended to this choice of d. One has to exploit that the above mentioned subset of Ω is small and

that $d(u)g(u)$ is small on this set as well. We have included a few remarks on this issue. We introduce the following affine-scaling function: Choose $\zeta \in (0, 1/2]$, $\kappa > 0$, and define

$$c : x \in \Omega \longmapsto \min\{\zeta(b(x) - a(x)), \kappa\}\ , \quad \nu \stackrel{\text{def}}{=} \operatorname*{ess\,inf}_{x\in\Omega}\ c(x).$$

Then our affine-scaling function is given by $d : \mathcal{B} \longrightarrow L^\infty$,

$$d(u)(x) = \begin{cases} \min\{|g(u)(x)|, c(x)\} & \text{if } -g(u)(x) > u(x) - a(x) \\ & \text{and } u(x) - a(x) \le b(x) - u(x), \\ \min\{|g(u)(x)|, c(x)\} & \text{if } g(u)(x) > b(x) - u(x) \\ & \text{and } b(x) - u(x) \le u(x) - a(x), \\ \min\{u(x) - a(x), b(x) - u(x), c(x)\} & \text{else.} \end{cases}$$

As we will see in Lemma 5.1, a suitable approximate derivative $G(u) \in \mathcal{L}(L^p, L^{p'})$ of $d(u)g(u)$ can be obtained by formally applying the product rule which yields

$$G(u) = d(u)\nabla^2 f(u) + d'(u)g(u)I \tag{5}$$

with $d' : \mathcal{B} \longrightarrow L^\infty$ suitably chosen. We recall that the only requirements on d' needed for the global convergence analysis in [24] are the conditions $d'(u)g(u) \ge 0$ and $\|d'(u)\|_\infty \le c_{d'}$ for all $u \in \mathcal{B}$. Our choice

$$d'(u) \stackrel{\text{def}}{=} \chi_{\{d(u)<c\}}\operatorname{sgn}(g(u))\ , \quad u \in \mathcal{B},$$

can be motivated as follows: Let $\bar{u}$ be a KKT-point and u tend to $\bar{u}$ in L^p. Then the sets $\{g(u) > 0 \wedge d(u) = u - a \wedge g(\bar{u}) > 0\}$ tend to $\{g(\bar{u}) > 0\}$ in measure. Analogously, the sets $\{g(u) < 0 \wedge d(u) = b - u \wedge g(\bar{u}) < 0\}$ tend to $\{g(\bar{u}) < 0\}$. On these sets, the choice $d'(u) = \operatorname{sgn}(g(u))$ is obtained by formal differentiation w.r.t. $u(x)$. Furthermore, $d'(u)g(u)$ tends to zero on the set $\{g(\bar{u}) = 0\}$ in $L^{p'}$ since $\|d'(u)\|_\infty$ is bounded. It turns out that the contribution of $d'(u)g(u)$ on this set is small enough for any uniformly bounded choice of $d'(u)$ to get a sufficiently good approximation $G(u)(u - \bar{u})$ of $d(u)g(u) - d(\bar{u})g(\bar{u})$ (cf. Lemma 5.1).

If u is an interior point of $\mathcal{B}$ w.r.t. the L^∞-norm, more precisely $u \in \mathcal{B}^\circ$, then the multiplication operator $d(u)I$ is an automorphism of L^t for all $1 \le t \le \infty$. Since our algorithm will rely on the bijectivity of $d(u_k)I$ at each iterate u_k we require $u_k \in \mathcal{B}^\circ$ for all k. Given a current iterate $u^c \in \mathcal{B}^\circ$, we define a Newton-like step for the solution of the affine-scaling equation (3):

$$G(u^c)(u^n - u^c) = -d(u^c)g(u^c) \tag{6}$$

Let $\bar{u} \in \mathcal{B}$ be a KKT-point, i.e. $d(\bar{u})g(\bar{u}) = 0$. Then subtracting the trivial identity $G(u^c)(\bar{u} - \bar{u}) = -d(\bar{u})g(\bar{u})$ from (6) yields the equivalent equation

$$G(u^c)(u^n - \bar{u}) = R(u^c) \tag{7}$$

with

$$R(u) \stackrel{\text{def}}{=} d(\bar{u})g(\bar{u}) - d(u)g(u) - G(u)(\bar{u} - u). \tag{8}$$

For a classical analysis of the Newton-like iteration induced by (6) we would typically need that for suitable q_1, q_2 and $u \in \mathcal{B}^\circ \subset L^{q_1}$ close to $\bar{u}$ the operator $G(u)$ admits an

inverse $G(u)^{-1} \in \mathcal{L}(L^{q_2}, L^{q_1})$ and that $\|G(u)^{-1}R(u)\|_{q_1} = o(\|u - \bar{u}\|_{q_1})$. Moreover, for $u^c \in \mathcal{B}^\circ$ close to $\bar{u}$ the solution u^n of (6) is required to lie again in $\mathcal{B}^\circ$ to keep the iteration alive. The presence of the multiplication operator $d'(u)g(u)I$ in $G(u)$ implicates that only the choice $q_1 = q_2 = q$ makes sense. Furthermore, we will show in Lemma 5.6 that it is untenable to assume the uniform boundedness of $\|G(u)^{-1}\|_{q,q}$ in a neighborhood of $\bar{u}$. Hence, we will introduce a multiplication operator $W(u) \in \mathcal{L}(L^q, L^q)$ such that the uniform boundedness of $(W(u)G(u))^{-1}$ in $\mathcal{L}(L^q, L^q)$ is a relatively weak requirement which is, e.g., implied by assumptions used for the analysis of a projected Newton method in [15]. In Lemma 5.11 we will show that under suitable assumptions $\|W(u)R(u)\|_q = o(\|u - \bar{u}\|_s)$ for some $s > q$. There seems to be no way to prove the more favorable estimate $\|W(u)R(u)\|_q = o(\|u - \bar{u}\|_q)$. Even the weaker estimate $\|R(u)\|_q = o(\|u - \bar{u}\|_q)$ requires at least the continuity of the gradient $g(u)$ from L^q to L^∞. For details see Lemma 5.1 and the proof of Lemma 5.7. Kelley and Sachs [15] overcame similar difficulties by introducing a smoothing step $u \in L^q \longmapsto u^s \in L^s$ with the property $\|u^s - \bar{u}\|_s \leq \text{const}\|u - \bar{u}\|_q$. We take the same approach. Finally, it is very likely that the iteration eventually breaks down with an $u^n \notin \mathcal{B}^\circ$. Therefore, we must include a back-transport that takes u^n back into the interior of $\mathcal{B}$. This back-transport can be implemented as an interior-point modification of the pointwise projection $P(u) = \max\{a, \min\{b, u\}\}$ which satisfies $|P(u) - \bar{u}| \leq |u - \bar{u}|$.

4. Outline of the algorithm in an abstract setting. The fundamental building blocks and convergence properties of the algorithm can be described most conveniently in the following abstract framework. Let X_0, X_1 and X_2 be Banach spaces, $\mathcal{K}^\circ \subset X_1$ be a convex nonempty set, and $X_1 \subset X_0$ continuously embedded. Denote by $\mathcal{K}$ the closure of $\mathcal{K}^\circ$ in X_1. Given the mapping $E : \mathcal{K} \longrightarrow X_2$, we want to solve the equation

$$E(u) = 0 \quad , \quad u \in \mathcal{K}. \tag{9}$$

To this end, we define a Newton-like iteration based on the linear approximation $E(u + s) - E(u) \approx G(u)s$, $G : \mathcal{K}^\circ \longrightarrow \mathcal{L}(X_0, X_2)$. The iteration is augmented by a *smoothing step* $u_k \longmapsto u_k^s = S_k^\circ(u_k)$ with operator $S_k^\circ : \mathcal{K}^\circ \subset X_0 \longrightarrow \mathcal{K}^\circ \subset X_1$, and a *back-transport* $P[v] : X_0 \longrightarrow \mathcal{K}^\circ$, $v \in \mathcal{K}^\circ$, see below:

ALGORITHM 4.1 (ABSTRACT NEWTON ITERATION).

1. Choose $u_0 \in \mathcal{K}^\circ$.
2. For $k = 0, 1, 2, \ldots$
   - 2.1 If $E(u_k) = 0$, STOP.
   - 2.2 Perform a smoothing step: $u_k^s = S_k^\circ(u_k)$.
   - 2.3 Compute $u_{k+1}^n \in X_0$ from
   $$G(u_k^s)(u_{k+1}^n - u_k^s) = -E(u_k^s) \quad \text{(Newton-like step)}$$
   - 2.4 Transport u_{k+1}^n back to $\mathcal{K}^\circ$: $u_{k+1} = P[u_k^s](u_{k+1}^n)$.

Let $\bar{u} \in \mathcal{K}$ be a solution to (9). Then we can rewrite the equation in step 2.3 as follows:

$$G(u_k^s)(u_{k+1}^n - \bar{u}) = E(\bar{u}) - E(u_k^s) - G(u_k^s)(\bar{u} - u_k^s) \stackrel{\text{def}}{=} R(u_k^s)$$

Algorithm 4.1 is locally superlinear convergent under the following general assumptions:

ABSTRACT ASSUMPTIONS.

There are constants $\rho, C_S, C_P > 0$ and monotone increasing functions

$$\delta_P, \delta_R, \gamma : [0, C_S\rho) \longrightarrow [0, \infty)$$

such that

1. For all k with $\|u_k - \bar{u}\|_{X_0} < \rho$ holds

$$\|S_k^\circ(u_k) - \bar{u}\|_{X_1} \leq C_S \|u_k - \bar{u}\|_{X_0}.$$

2. For all $u \in X_0$, $v \in \mathcal{K}^\circ$, $\|v - \bar{u}\|_{X_1} < C_S\rho$,

$$\|P[v](u) - \bar{u}\|_{X_0} \leq C_P \|u - \bar{u}\|_{X_0} + \delta_P(\|v - \bar{u}\|_{X_1})\|v - \bar{u}\|_{X_1}.$$

3. There are a Banach space X_3 and an operator $W : \mathcal{K}^\circ \longrightarrow \mathcal{L}(X_2, X_3)$ such that
   a) for all $u \in \mathcal{K}^\circ$, $\|u - \bar{u}\|_{X_1} < C_S\rho$, and $r \in X_3$ there exists $s \in X_1$ with

$$W(u)G(u)s = r \ , \quad \|s\|_{X_0} \leq \gamma(\|u - \bar{u}\|_{X_1})\|r\|_{X_3}.$$

   b) for all $u \in \mathcal{K}^\circ$, $\|u - \bar{u}\|_{X_1} < C_S\rho$,

$$\|W(u)R(u)\|_{X_3} \leq \delta_R(\|u - \bar{u}\|_{X_1})\|u - \bar{u}\|_{X_1}.$$

   c) $\lim\limits_{t \to 0^+} \gamma(t)\delta_R(t) = 0$ and $\lim\limits_{t \to 0^+} \delta_P(t) = 0$.

THEOREM 4.2. *Let $\bar{u} \in \mathcal{K}$ be a solution to (9). Assume that the above assumptions hold. Then there is $0 < \rho_0 \leq \rho$ such that for all $u_0 \in \mathcal{K}^\circ$, $\|u_0 - \bar{u}\|_{X_0} < \rho_0$, Algorithm 4.1 is well-defined and either terminates with $u_k \in \mathcal{K}^\circ$ solving (9) or generates sequences $(u_k) \subset \mathcal{K}^\circ$ and $(u_k^s) \subset \mathcal{K}^\circ$ that converge superlinearly to $\bar{u}$ in X_0 and X_1, respectively.*

Proof. We introduce the abbreviations $\varepsilon_k \stackrel{\text{def}}{=} \|u_k - \bar{u}\|_{X_0}$ and $\varepsilon_k^s \stackrel{\text{def}}{=} \|u_k^s - \bar{u}\|_{X_1}$. Let $u_k \in \mathcal{K}^\circ$ satisfy $\varepsilon_k < \rho$. Then $\varepsilon_k^s < C_S\rho$ by 1., and thus, using the assumptions,

$$\begin{aligned} \varepsilon_{k+1} &= \|P[u_k^s](u_{k+1}^n) - \bar{u}\|_{X_0} \leq C_P \|u_{k+1}^n - \bar{u}\|_{X_0} + \delta_P(\varepsilon_k^s)\varepsilon_k^s \\ &\leq C_P\gamma(\varepsilon_k^s)\|W(u_k^s)R(u_k^s)\|_{X_3} + \delta_P(\varepsilon_k^s)\varepsilon_k^s \leq (C_P\gamma(\varepsilon_k^s)\delta_R(\varepsilon_k^s) + \delta_P(\varepsilon_k^s))\varepsilon_k^s \\ &\leq C_S(C_P\gamma(C_S\varepsilon_k)\delta_R(C_S\varepsilon_k) + \delta_P(C_S\varepsilon_k))\varepsilon_k. \end{aligned} \tag{10}$$

Moreover,

$$\varepsilon_{k+1}^s \leq C_S\varepsilon_{k+1} \leq C_S(C_P\gamma(\varepsilon_k^s)\delta_R(\varepsilon_k^s) + \delta_P(\varepsilon_k^s))\varepsilon_k^s \tag{11}$$

By Assumption 3c), there is $0 < \rho_0 \leq \rho$ such that

$$C_S(C_P\gamma(C_Sz)\delta_R(C_Sz) + \delta_P(C_Sz)) < 1 \ \text{ for all } 0 \leq z < \rho_0.$$

Therefore, if $\varepsilon_0 < \rho_0 \leq \rho$, we have $\varepsilon_k < \rho_0 \leq \rho$ for all k. In particular, the algorithm is well-defined. Now (10) yields superlinear convergence of (u_k) to $\bar{u}$ in X_0, and (11) superlinear convergence of (u_k^s) to $\bar{u}$ in X_1. □

REMARK 4.3. It is easier to find a smoothing operator S_k that satisfies all requirements in 1. except for the condition $S_k(\mathcal{K}^\circ) \subset \mathcal{K}^\circ$. If the operator $P[v]$ can be

defined in such a way that, in addition to 2., for all $u \in X_1$, $\|u-\bar{u}\|_{X_1} < C_S\rho$, and $v \in \mathcal{K}^\circ$, $\|v-\bar{u}\|_{X_0} < \rho$,

$$\|P[v](u)-\bar{u}\|_{X_1} \le \bar{C}_P\|u-\bar{u}\|_{X_1} + \bar{C}'_P\|v-\bar{u}\|_{X_0}, \tag{12}$$

then obviously $S_k^\circ : u \in \mathcal{K}^\circ \longmapsto P[u](S_k(u))$ defines a smoothing step satisfying 1. with C_S replaced by $\bar{C}_P C_S + \bar{C}'_P$. For our problem (P) we will be able to define $P[v]$ in such a way that (12) holds, see Lemma 6.4. □

In our setting we have $\mathcal{K}^\circ = \mathcal{B}^\circ$ and, consequently, $\mathcal{K} = \mathcal{B}$. The mapping E is given by $u \longmapsto d(u)g(u)$. The crucial topics of our analysis consist in the proper choice of the spaces X_i, the weighting operator W, and the proof that under appropriate conditions the above abstract assumptions hold. A few remarks on the 'nonstandard' building blocks of Algorithm 4.1 are in order. If there exists a projection $P : X_0 \longrightarrow \mathcal{K} \subset X_0$ onto $\mathcal{K}$ that is Lipschitz at $\bar{u}$, e.g. $P(u) = \min\{b, \max\{a, u\}\}$ for $X_0 = L^q$ and $\mathcal{K} = \mathcal{B}$, then the back-transport operator $P[v]$ can (and will) be implemented by an interior-point modification of P. More specifically, $P[v](u)$ will consist in the projection $P(u)$ of u onto $\mathcal{K}$ followed by a tiny step towards the point $v \in \mathcal{K}^\circ$ to achieve $P[v](u) \in \mathcal{K}^\circ$. The idea of a smoothing step was already used by Kelley and Sachs [15]. It is a tool to compensate the discrepancy of the X_0-norm on the left side and the stronger X_1-norm on the right side of the inequality

$$\|u_{k+1}^n - \bar{u}\|_{X_0} \le \gamma(\|u_k^s-\bar{u}\|_{X_1})\delta_R(\|u_k^s-\bar{u}\|_{X_1})\|u_k^s-\bar{u}\|_{X_1}$$

which is obtained by combining assumptions 3a) and b).

5. Analysis of the Newton-like iteration. We return to the affine-scaling Newton iteration (7) and begin to verify the abstract assumptions of §4. The following Lemma states a pointwise estimate for the remainder term $R(u)$.

LEMMA 5.1. *Let* (A1) *hold. In addition, let* (O1) *and* (O2) *be satisfied at* $\bar{u}$. *Then for all* $u \in \mathcal{B}$ *the inequality*

$$|R(u)| \le d(u)|g(\bar{u}) - g(u) - \nabla^2 f(u)(\bar{u}-u)| + (|g(\bar{u})|d(u) + |g(u)||\bar{u}-u|) \tag{13}$$

holds on Ω *and, moreover,*

$$\begin{aligned} |R(u)| \le\ & d(u)|g(\bar{u}) - g(u) - \nabla^2 f(u)(\bar{u}-u)| \\ & + \min\{\max\{d(u), |g(u)|\}, |g(\bar{u})-g(u)|\}\max\{|\bar{u}-u|, |g(\bar{u})-g(u)|\} \end{aligned} \tag{14}$$

is satisfied on $J \overset{\text{def}}{=} \{x \in \Omega :\ |u(x)-\bar{u}(x)| < c(x)\}$.

Proof. Let $u \in \mathcal{B}$ be given and set $I = \{x \in \Omega :\ d(u)(x) < c(x)\}$. Then we get

$$\begin{aligned} & d(\bar{u})g(\bar{u}) - d(u)g(u) - G(u)(\bar{u}-u) = \\ & \quad = d(\bar{u})g(\bar{u}) - d(u)g(u) - d(u)\nabla^2 f(u)(\bar{u}-u) - \chi_I|g(u)|(\bar{u}-u) \\ & \quad = d(u)\Big(g(\bar{u}) - g(u) - \nabla^2 f(u)(\bar{u}-u)\Big) + g(\bar{u})(d(\bar{u})-d(u)) - \chi_I|g(u)|(\bar{u}-u) \end{aligned}$$

Since (O1) and (O2) are satisfied at $\bar{u}$ we have $d(\bar{u})g(\bar{u}) = 0$ by Lemma 2.6 and hence the first estimate is obvious. We complete the proof by verifying that for a.a. $x \in J$

$$\begin{aligned} R_1(u)(x) \le\ & \min\{\max\{d(u)(x), |g(u)(x)|\}, |g(\bar{u})(x) - g(u)(x)|\} \\ & \cdot \max\{|\bar{u}(x)-u(x)|, |g(\bar{u})(x)-g(u)(x)|\}, \end{aligned} \tag{15}$$

$$R_1(u)(x) \stackrel{\text{def}}{=} \Big| g(\bar{u})(x)\Big(d(\bar{u})(x) - d(u)(x)\Big) - \chi_I(x)|g(u)(x)|\Big(\bar{u}(x) - u(x)\Big)\Big|.$$

We use again $d(\bar{u})g(\bar{u}) = 0$. On the subset of all $x \in J$ with $g(\bar{u})(x) = 0$ we get

$$R_1(u) = \chi_I|g(u)||\bar{u} - u| \le |g(u) - g(\bar{u})||\bar{u} - u|,$$

and (15) is obvious.

For all $x \in J$ with $g(\bar{u})(x) \neq 0$ we have $d(\bar{u})(x) = 0$. (O2) implies that only the cases $\bar{u}(x) = a(x)$ and $g(\bar{u})(x) > 0$ or $\bar{u}(x) = b(x)$ and $g(\bar{u})(x) < 0$ can occur.

We first look at $x \in J$ with $\bar{u}(x) = a(x)$ and $g(\bar{u})(x) > 0$. Since $\zeta \le 1/2$ and $x \in J$, we get $u(x) - a(x) < b(x) - u(x)$. Hence, we obtain (mind that $\bar{u}(x) = a(x)$)

$$d(u)(x) = \begin{cases} \min\{|g(u)(x)|, c(x)\} & \text{if } -g(u)(x) > u(x) - \bar{u}(x) \ge 0, \\ \min\{u(x) - \bar{u}(x), c(x)\} & \text{else.} \end{cases}$$

If $d(u)(x) = u(x) - \bar{u}(x) < c(x)$, then $x \in I$ and using $d(\bar{u})(x) = 0$, $g(\bar{u})(x) \ge 0$ we get for all these x

$$R_1(u) = \Big||g(\bar{u})| - |g(u)|\Big||\bar{u} - u| \le |g(\bar{u}) - g(u)||\bar{u} - u|.$$

If, in addition, $|g(\bar{u})(x) - g(u)(x)| \le d(u)(x)$ then (15) holds, for

$$R_1(u)(x) \le \min\{d(u)(x), |g(\bar{u})(x) - g(u)(x)|\}|\bar{u}(x) - u(x)|.$$

Otherwise, we have $|g(\bar{u})(x) - g(u)(x)| > d(u)(x) = u(x) - \bar{u}(x) = |\bar{u}(x) - u(x)|$, and therefore

$$R_1(u)(x) \le \min\{d(u)(x), |g(\bar{u})(x) - g(u)(x)|\}|g(\bar{u})(x) - g(u)(x)|$$

which implies (15). If $d(u)(x) = |g(u)(x)| < c(x)$ then $x \in I$, $g(u)(x) \le 0 \le g(\bar{u})(x)$. Thus, we have for all such x that $\max\{|g(u)|, |g(\bar{u})|\} \le |g(\bar{u}) - g(u)|$ and

$$\begin{aligned} R_1(u) &= \Big|-g(\bar{u})|g(u)| - |g(u)|(\bar{u} - u)\Big| \\ &= |g(u)|\Big||\bar{u} - u| - |g(\bar{u})|\Big| \le |g(u)| \max\{|\bar{u} - u|, |g(\bar{u})|\} \\ &\le \min\{|g(u)|, |g(\bar{u}) - g(u)|\} \max\{|\bar{u} - u|, |g(\bar{u}) - g(u)|\}. \end{aligned}$$

It remains the case $x \in J \cap I^c$, i.e. $x \in J$ and $d(u)(x) = c(x)$. Here $u(x) - \bar{u}(x) \ge c(x) = d(u)(x)$ or $-g(u)(x) \ge c(x) = d(u)(x)$. Since $x \in J$, the first case cannot occur. Therefore, we have $g(u)(x) \le -d(u)(x) \le 0 \le g(\bar{u})(x)$ for all $x \in J \cap I^c$, and hence

$$R_1(u) = |g(\bar{u})d(u)| \le |g(\bar{u})||g(u)| \le \min\{|g(u)|, |g(\bar{u}) - g(u)|\}|g(\bar{u}) - g(u)|.$$

For $\bar{u}(x) = b(x)$ and $g(\bar{u})(x) < 0$ the same arguments can be used and the proof is complete. □

REMARK 5.2. For the Coleman-Li affine-scaling function d_{I} the estimate (13) holds as well. A simplified version of (14) can be established on $J = \{g(u)g(\bar{u}) \ge 0\}$. In contrast to the set J defined in Lemma 5.1, J^c is not a set of measure zero for $\|u - \bar{u}\|_\infty$ sufficiently small. This additional technical difficulty can be overcome by

using the fact that the measure of J^c tends to zero under the strict complementarity condition (C) below. □

Let (O1) and (O2) hold for $\bar{u}$. We define the active set $\bar{A}$ and the inactive set $\bar{I}$,

$$\bar{A} = \{x \in \Omega : \ \bar{u}(x) \in \{a(x), b(x)\}\} \ , \quad \bar{I} = \bar{A}^c.$$

Furthermore, the usual strict complementarity condition shall hold at $\bar{u}$ (note that $|g(\bar{u})|$ is a Lagrange multiplier):

ASSUMPTION (STRICT COMPLEMENTARITY CONDITION).

(C) $g(\bar{u})(x) \neq 0$ for a.a. $x \in \bar{A}$.

In contrast to the finite-dimensional case the active set can in general not be identified after a finite number of iterations under the strict complementarity condition (C), since the gradient may be arbitrarily small on the active set, especially near its boundary. But we shall use (C) to show that the residual set of 'uncertainty' is small. We need the following continuity property of d.

LEMMA 5.3. *Let the assumptions of Lemma* 2.6 *hold. In addition, let* (O1) *and* (O2) *be satisfied at* $\bar{u}$. *Then for all* $u \in \mathcal{B}$ *the inequality*

$$|d(u) - d(\bar{u})| \leq \max\{|u - \bar{u}|, |g(u) - g(\bar{u})|\}$$

holds on $J \stackrel{\text{def}}{=} \{x \in \Omega : \ |u(x) - \bar{u}(x)| < (b(x) - a(x))/2\}$.

Proof. Let $x \in J$ be arbitrary. Since (O1) and (O2) hold at $\bar{u}$, the identity $d(\bar{u})g(\bar{u}) = 0$ is valid by Lemma 2.6. In addition, (O2) assures that

$$d(\bar{u})(x) = \min\{\bar{u}(x) - a(x), b(x) - \bar{u}(x), c(x)\}.$$

By definition, we have

$$d(u)(x) = \min\{u(x) - a(x), b(x) - u(x), c(x)\} \quad \text{or} \tag{16}$$

$$d(u)(x) \leq \min\{|g(u)(x)|, c(x)\} \geq \min\{u(x) - a(x), b(x) - u(x), c(x)\}. \tag{17}$$

For all x from case (16) as well as all x with $d(u)(x) = \min\{|g(u)(x)|, c(x)\} \leq d(\bar{u})(x)$ we get

$$|d(\bar{u}) - d(u)| \leq |\min\{\bar{u} - a, b - \bar{u}, c\} - \min\{u - a, b - u, c\}| \leq |u - \bar{u}|,$$

where we have used the inequality (cf. [24, Lem. 9.3])

$$|\min\{a_1, \ldots, a_n\} - \min\{b_1, \ldots, b_n\}| \leq \max\{|a_1 - b_1|, \ldots, |a_n - b_n|\}.$$

For all x with $d(u)(x) = \min\{|g(u)(x)|, c(x)\} > d(\bar{u})(x)$ we have

$$|d(u) - d(\bar{u})| \leq d(u) \leq |g(u)| \leq |g(u) - g(\bar{u})|.$$

The last inequality is obvious if $g(\bar{u})(x) = 0$. For $g(\bar{u})(x) \neq 0$ it follows from the observation that $g(u)(x)$ and $g(\bar{u})(x)$ have different signs. In fact, by (O2) only the cases $\bar{u}(x) = a(x)$, $g(\bar{u})(x) > 0$ or $\bar{u}(x) = b(x)$, $g(\bar{u})(x) < 0$ can occur. If $\bar{u}(x) = a(x)$ and $g(\bar{u})(x) > 0$ then $u(x) - a(x) < b(x) - u(x)$ since $x \in J$. Hence, by the definition of $d(u)$, $d(u)(x) = \min\{|g(u)(x)|, c(x)\}$ is only possible for $g(u)(x) < 0$. Finally, if $\bar{u}(x) =$

$b(x), g(\bar{u})(x) < 0$ then $b(x) - u(x) < u(x) - a(x)$ and $d(u)(x) = \min\{|g(u)(x)|, c(x)\}$ requires $g(u)(x) > 0$. □

REMARK 5.4. An analogue of Lemma 5.3 can be established for $d_I(u)$ on the set $J = \{g(u)g(\bar{u}) \geq 0\}$ (cf. Remark 5.2): $|d_I(u) - d_I(\bar{u})| \leq |u - \bar{u}|$ on J. □

The pointwise estimates in Lemma 5.1 and 5.3 can be converted into norm estimates by making the following assumption:

ASSUMPTION.

(A2) There are $2 \leq q < r \leq s \leq \infty$, $s \geq p$, such that $g : \mathcal{B} \subset L^s \longrightarrow L^r$ is Lipschitz continuous with constant L_g and $g : \mathcal{B} \subset L^s \longrightarrow L^q$ is Lipschitz continuously Fréchet differentiable. We denote the Lipschitz constant of $\nabla g = \nabla^2 f$ by $L_{g'}$.

As a consequence of Lemma 5.3 we get the Lipschitz continuity of d at $\bar{u}$:

LEMMA 5.5. *If* (O1), (O2) *hold at* $\bar{u}$ *and the assumptions* (A1) *and* (A2) *are satisfied then for all* $u \in \mathcal{B}$

$$\|d(u) - d(\bar{u})\|_r \leq \left(m_{r,s} + L_g + \frac{2\kappa m_{r,s}}{\nu}\right) \|u - \bar{u}\|_s \stackrel{\text{def}}{=} L_d \|u - \bar{u}\|_s$$

with $m_{r,s}$ *defined as in Lemma* 2.1.

Proof. On $B \stackrel{\text{def}}{=} \{x \in \Omega : |u(x) - \bar{u}(x)| < \nu/2\}$ Lemma 5.3 is applicable and yields with (A2) and Lemma 2.1

$$\|d(u) - d(\bar{u})\|_{r,B} \leq \|\max\{|u - \bar{u}|, |g(u) - g(\bar{u})|\}\|_r \leq (m_{r,s} + L_g)\|u - \bar{u}\|_s.$$

Since $|d(u)(x) - d(\bar{u})(x)| \leq \kappa$, we get on B^c

$$\|d(u) - d(\bar{u})\|_{r,B^c} \leq \|\kappa\|_{r,B^c} \leq \left\|\kappa \frac{2|u - \bar{u}|}{\nu}\right\|_{r,B^c} \leq \frac{2\kappa}{\nu}\|u - \bar{u}\|_r \leq \frac{2\kappa m_{r,s}}{\nu}\|u - \bar{u}\|_s.$$

The triangle inequality completes the proof. □

In the finite-dimensional case the existence and uniform boundedness of $G(u)^{-1}$ in a neighborhood of $\bar{u}$ can be ensured if $\bar{u}$ satisfies sufficient second-order conditions with strict complementarity, see [7]. The following considerations show that the requirement of uniform boundedness of $G(u)^{-1}$ close to $\bar{u}$ is unacceptably strong in the infinite-dimensional setting. Since

$$g(\bar{u})(x) = 0 \quad \text{a.e. on } \bar{I}\,, \quad d(\bar{u})(x) = 0 \quad \text{a.e. on } \bar{A}, \tag{18}$$

and by (A2) and Lemma 5.3

$$\|d(u) - d(\bar{u})\|_r + \|g(u) - g(\bar{u})\|_r \leq (L_d + L_g)\|u - \bar{u}\|_s,$$

the set

$$N_\varepsilon(u) \stackrel{\text{def}}{=} \{x \in \Omega : |g(u)(x)| + d(u)(x) \leq \varepsilon\} \tag{19}$$

may have nonzero measure for arbitrarily small $\varepsilon > 0$ if $\|u - \bar{u}\|_s$ is small enough. Typically, an open neighborhood of a part of $\partial\bar{A}$ is contained in $N_\varepsilon(u)$.

Let $1 \leq q_2 \leq q_1 \leq \infty$ and assume that $\|\nabla^2 f(u)\|_{q_1,q_2}$ is uniformly bounded on an L^s-neighborhood of $\bar{u}$. The following lemma shows that in the above scenario

$\|G(u)\|_{q_1,q_2}$ is uniformly bounded, but $\|G(u)^{-1}\|_{q_2,q_1}$ is not. This is caused by the fact that the operator

$$H(u) = W(u)G(u) \quad \text{with} \quad W(u) = \frac{1}{|g(u)| + d(u)} I\,, \quad u \in \mathcal{B}, \tag{20}$$

is still uniformly bounded in $\mathcal{L}(L^{q_1}, L^{q_2})$ although $\|W(u)\|_{q_2,q_2} \to \infty$ as $\|u - \bar{u}\|_s$ tends to zero. More precisely, we have

LEMMA 5.6. *Let $u \in \mathcal{B}^\circ$, $1 \le q_2 \le q_1 \le \infty$, and $\nabla^2 f(u) \in \mathcal{L}(L^{q_1}, L^{q_2})$. Then*

i) $G(u) \in \mathcal{L}(L^{q_1}, L^{q_2})$, $\quad \|G(u)\|_{q_1,q_2} \le m_{q_2,q_1}\|g(u)\|_\infty + \kappa\|\nabla^2 f(u)\|_{q_1,q_2}$,

ii) $H(u) \in \mathcal{L}(L^{q_1}, L^{q_2})$, $\quad \|H(u)\|_{q_1,q_2} \le m_{q_2,q_1} + \|\nabla^2 f(u)\|_{q_1,q_2}$,

iii) *If $G(u)$ is invertible in $\mathcal{L}(L^{q_1}, L^{q_2})$ then*

$$\|G(u)^{-1}\|_{q_2,q_1} \ge \varepsilon^{-1}\|H(u)\|^{-1}_{q_1,q_2}.$$

for all $\varepsilon > 0$ with $\mu(N_\varepsilon(u)) > 0$.

Here m_{q_2,q_1} is as in Lemma 2.1.

Proof. Assertion i) follows immediately from the definition of $G(u)$. The estimate

$$\begin{aligned}\|H(u)v\|_{q_2} &\le \left\| \frac{\chi_{\{d(u)<c\}}|g(u)|}{|g(u)| + d(u)} v \right\|_{q_2} + \left\| \frac{d(u)}{|g(u)| + d(u)} \nabla^2 f(u) v \right\|_{q_2} \\ &\le m_{q_2,q_1}\|v\|_{q_1} + \|\nabla^2 f(u)\|_{q_1,q_2}\|v\|_{q_1}\end{aligned}$$

yields ii). To prove iii) let $G(u) \in \mathcal{L}(L^{q_1}, L^{q_2})$ be invertible and $\varepsilon > 0$ such that $\mu(N_\varepsilon(u)) > 0$. Then $\|w_\varepsilon\|_{q_2} > 0$ for $w_\varepsilon \stackrel{\text{def}}{=} \chi_{N_\varepsilon(u)}$, and, setting $v_\varepsilon = G(u)^{-1}w_\varepsilon$,

$$\|H(u)v_\varepsilon\|_{q_2} \le \|H(u)\|_{q_1,q_2}\|G(u)^{-1}\|_{q_2,q_1}\|w_\varepsilon\|_{q_2}.$$

On the other hand, the definition of $N_\varepsilon(u)$ yields

$$\|H(u)v_\varepsilon\|_{q_2} = \left\| \frac{w_\varepsilon}{|g(u)| + d(u)} \right\|_{q_2} \ge \frac{\|w_\varepsilon\|_{q_2}}{\varepsilon}.$$

Combining both estimates gives iii). □

The identity

$$H(u) = \frac{\chi_{\{d(u)<c\}}|g(u)|}{|g(u)| + d(u)} I + \frac{d(u)}{|g(u)| + d(u)} \nabla^2 f(u).$$

shows that the operator $H(u)$ is 'almost' a pointwise convex combination of the identity and the Hessian $\nabla^2 f(u)$. If (A2) and the strict complementarity condition (C) hold then, using (18) and Lemma 5.5, one can show with the same techniques as in the proof of Lemma 8.3 that

$$\frac{\chi_{\{d(u)<c\}}|g(u)|}{|g(u)| + d(u)} \xrightarrow{L^q} \chi_{\bar{A}} \quad \text{and} \quad \frac{d(u)}{|g(u)| + d(u)} \xrightarrow{L^q} \chi_{\bar{I}} \quad \text{a.e. as} \quad u \in \mathcal{B}^\circ \xrightarrow{L^s} \bar{u}.$$

Thus, they converge in all spaces L^t, $1 \le t < \infty$, by (A1) and the interpolation inequality of Lemma 2.2.

Hence, we impose the following assumption on $G(u)$ which, as we will see, is implied by the assumptions in the paper of Kelley and Sachs [15] on the projected Newton method (cf. Lemma 8.3) and in important cases by a sufficient second-order condition of Dunn and Tian [9] (see Theorem 9.4).

ASSUMPTION.

(A3) There is $0 < \rho_H \leq 1$ such that the operator $H(u)$ defined in (20) satisfies $H(u) \in \mathcal{L}(L^q, L^q)$ and is invertible for all $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_s < \rho_H$, with uniformly bounded inverse, more precisely, $\|H(u)^{-1}\|_{q,q} < C_H$.

We now return to the analysis of (7). Since for $u \in \mathcal{B}^\circ$ there is $\delta > 0$ with $d(u) > \delta$, the multiplication operator $W(u)$ defined in (20) is a linear continuous isomorphism of L^t, $1 \leq t \leq \infty$. Applying $W(u^c)$ from the left to (7) yields the equivalent equation

$$H(u^c)(u^n - \bar{u}) = W(u^c)R(u^c). \tag{21}$$

Since $H(u^c) \in \mathcal{L}(L^q, L^q)$ is invertible by (A3) if $\|u^c - \bar{u}\|_s < \rho_H$, we derive an upper bound for the L^q-norm of the right hand side:

LEMMA 5.7. *Let* (O1), (O2) *hold at* $\bar{u}$. *Moreover, let* (A1) *and* (A2) *be satisfied. Then for all* $u \in \mathcal{B}^\circ$ *holds:*

$$\begin{aligned} \|W(u)R(u)\|_q \leq\ & L_{g'}\|u - \bar{u}\|_s^2 + (m_{r,s} + L_g)\|Q(u)\|_{\tilde{q}}\|u - \bar{u}\|_s \\ & + \max\{\|g(\bar{u})\|_\infty, \|b - a\|_\infty\}\, \nu^{-\frac{s}{q}}\|u - \bar{u}\|_s^{\frac{s}{q}} \end{aligned} \tag{22}$$

where $\tilde{q} \stackrel{\text{def}}{=} \dfrac{qr}{r-q}$ $(= q$ *if* $r = \infty)$,

$$Q(u) = \frac{\min\{\max\{d(u), |g(u)|\}, |g(u) - g(\bar{u})|\}}{|g(u)| + d(u)}, \tag{23}$$

and the last term has to be interpreted as zero in the case $s = \infty$ *for* $\|u - \bar{u}\|_\infty < \nu$.

Proof. For $J \stackrel{\text{def}}{=} \{x \in \Omega : |u(x) - \bar{u}(x)| < \nu\}$ we may apply Lemma 5.1 and get with (23), (A2) and the mean value theorem

$$\begin{aligned} \|W(u)R(u)\|_q \leq\ & \left\| \frac{d(u)}{|g(u)| + d(u)} \right\|_\infty \|g(\bar{u}) - g(u) - \nabla^2 f(u)(\bar{u} - u)\|_q \\ & + \|Q(u)\max\{|u - \bar{u}|, |g(u) - g(\bar{u})|\}\|_{q,J} + \left\| \frac{|g(\bar{u})|d(u) + |g(u)||u - \bar{u}|}{|g(u)| + d(u)} \right\|_{q,J^c} \\ \leq\ & \sup_{\tau \in [0,1]} \|\nabla^2 f(u + \tau(\bar{u} - u)) - \nabla^2 f(u)\|_{s,q}\|u - \bar{u}\|_s \\ & + \|Q(u)\|_{\frac{qr}{r-q}}\|\max\{|u - \bar{u}|, |g(u) - g(\bar{u})|\}\|_r \\ & + \max\{\|g(\bar{u})\|_\infty, \|b - a\|_\infty\}\, \mu(J^c)^{1/q}, \end{aligned}$$

where we have applied Lemma 2.3 with $q_0 = q$ and $q_1 = r/q$ in the last step. Now (A2) immediately yields the first two terms on the right hand side of (22). To finish the proof, we first observe that $\mu(J^c) = 0$ for $\|u - \bar{u}\|_\infty < \nu$. Hence, we have (22) with the mentioned interpretation for $s = \infty$. If finally $s < \infty$, we have

$$\mu(J^c) = \|1\|_{s,J^c}^s \leq \|(u - \bar{u})/\nu\|_{s,J^c}^s \leq \nu^{-s}\|u - \bar{u}\|_s^s.$$

Using this in the last term of the above inequality, we get (22). ☐

It is important to notice that the term $Q(u)$ is crucial for our analysis since

$$|Q(u)(x)| = \left| \frac{\min\{\max\{d(u)(x), |g(u)(x)|\}, |g(u)(x) - g(\bar{u})(x)|\}}{|g(u)(x)| + d(u)(x)} \right| = O(1)$$

on $\{\max\{d(u), |g(u)|\} \leq \text{const}|g(u) - g(\bar{u})|\}$. In contrast to the finite-dimensional case, these sets may have nonzero measure under any reasonable strict complementarity condition even if $\|u - \bar{u}\|_\infty$ is arbitrarily small. On the other hand, we get under assumption (A2) on the complement of the set $N_\varepsilon(u)$ defined in (19) the estimate

$$|Q(u)(x)| \leq \frac{|g(u)(x) - g(\bar{u})(x)|}{\varepsilon} \quad \forall\ x \in N_\varepsilon(u)^c.$$

REMARK 5.8. Since the estimate (22) is sharp and usually $\|Q(u)\|_\infty = O(1)$ for $u \in \mathcal{B}^\circ \xrightarrow{L^\infty} \bar{u}$, an estimate of the form

$$\|u^n - \bar{u}\|_\infty = o(\|u^c - \bar{u}\|_\infty) \quad (u^c \in \mathcal{B}^\circ \xrightarrow{L^\infty} \bar{u})$$

for the solution u^n of the affine-scaling Newton equation (6) does in general not hold even if (A2),(A3) are satisfied with $q = \infty$. ☐

The following Lemma estimates the Lebesgue measure of the residual sets $N_\varepsilon(u)$.

LEMMA 5.9. *Let* (A1), (A2) *hold. If* $\bar{u}$ *satisfies* (O1), (O2), *and* (C), *then the following is true:*

i) $\omega : [0, \infty) \longrightarrow [0, \infty)$, $\omega(\varepsilon) \stackrel{\text{def}}{=} \mu(N_\varepsilon(\bar{u}))$ *is monotone increasing and satisfies*

$$\lim_{\varepsilon \to 0+} \omega(\varepsilon) = \omega(0) = 0. \tag{24}$$

ii) *For all* $u \in \mathcal{B}$ *holds*

$$\mu(N_\varepsilon(u)) \leq \omega(2\varepsilon) + \varepsilon^{-r}(L_g + L_d)^r \|u - \bar{u}\|_s^r.$$

with the obvious interpretation for $r = s = \infty$ *by setting* $\alpha^\infty = 0$ *for* $\alpha \in [0, 1)$.

Proof. ω is nonnegative and increasing, since $N_{\tilde{\varepsilon}}(\bar{u}) \subset N_\varepsilon(\bar{u})$ for $0 < \tilde{\varepsilon} \leq \varepsilon$. Hence, $\lim_{\varepsilon \to 0+} \omega(\varepsilon)$ exists and

$$\lim_{\varepsilon \to 0+} \omega(\varepsilon) = \mu\left(\bigcap_{\varepsilon > 0} N_\varepsilon(\bar{u}) \right).$$

By (C) and the definition of d there is a set N of measure zero with

$$|g(\bar{u})(x)| + d(\bar{u})(x) > 0 \quad \forall\ x \in N^c.$$

Hence, $N_0(\bar{u}) \subset N$ and thus $\omega(0) = \mu(N_0(\bar{u})) = 0$. Moreover, for all $x \in N^c$ there is $\varepsilon_0 > 0$ with $x \notin N_\varepsilon(\bar{u})$ for all $0 < \varepsilon < \varepsilon_0$. This shows

$$\bigcap_{\varepsilon > 0} N_\varepsilon(\bar{u}) \subset N$$

which implies (24). To prove ii), we use the triangle inequality and get

$$\begin{aligned} N_\varepsilon(u) &= \{|g(u)| + d(u) \le \varepsilon\} \\ &\subset \{|g(\bar u)| + d(\bar u) \le \varepsilon + |g(u) - g(\bar u)| + |d(u) - d(\bar u)|\} \\ &\subset N_{2\varepsilon}(\bar u) \cup \{|g(u) - g(\bar u)| + |d(u) - d(\bar u)| \ge \varepsilon\}. \end{aligned}$$

In the case $r = \infty$, we have by (A2) and Lemma 5.5

$$\left\| |g(u) - g(\bar u)| + |d(u) - d(\bar u)| \right\|_\infty \le (L_g + L_d)\|u - \bar u\|_\infty$$

Hence, $N_\varepsilon(u) \subset N_{2\varepsilon}(\bar u)$ for $(L_g + L_d)\|u - \bar u\|_\infty < \varepsilon$, which is the obvious interpretation of ii) for $r = s = \infty$. For $r < \infty$ we have by (A2), Lemma 2.4, and Lemma 5.5

$$\begin{aligned} \mu\left(\{|g(u) - g(\bar u)| + |d(u) - d(\bar u)| \ge \varepsilon\}\right) &\le \varepsilon^{-r} \left\| |g(u) - g(\bar u)| + |d(u) - d(\bar u)| \right\|_r^r \\ &\le \varepsilon^{-r} (L_g + L_d)^r \|u - \bar u\|_s^r. \end{aligned}$$

This proves ii). □

The following stronger strict complementarity condition will enable us to prove convergence with Q-rate > 1, since we get additional control on the growth of $\omega(\varepsilon)$:

Assumption (Strong strict complementarity condition).

(CS) There are $\bar q > 0$, $C_C > 0$, and $\varepsilon_0 > 0$ such that

$$\omega(\varepsilon) = \mu(\{|g(\bar u)| + d(\bar u) \le \varepsilon\}) \le C_C \varepsilon^{\bar q} \quad \text{for all} \quad 0 < \varepsilon < \varepsilon_0.$$

Remark 5.10. It is easy to see that condition (CS) is satisfied if the following regularity assumptions on $\bar u$ and the active set $\bar A$ hold. They are relaxations of Assumption 2.4 in [15]:

There is $c_0 > 0$ such that for all sufficiently small $\delta > 0$

$$\mu\left(\{x \in \Omega : \operatorname{dist}(x, \partial \bar A) \le \delta\}\right) \le c_0 \delta$$

and for suitable $c_1 > 0$ the following growth estimates hold true:

$$\begin{aligned} |g(\bar u)(x)| &\ge c_1 (\operatorname{dist}(x, \partial \bar A))^{1/\bar q} \quad \forall\ x \in \bar A \\ \min\{\bar u(x) - a(x), b(x) - \bar u(x)\} &\ge c_1 (\operatorname{dist}(x, \partial \bar A))^{1/\bar q} \quad \forall\ x \in \bar I = \bar A^c. \end{aligned}$$

□

The previous lemma enables us to estimate the norm of $Q(u)$. Together with Lemma 5.7 we get

Lemma 5.11. *Let* (O1), (O2), *and* (C) *hold at* $\bar u$. *Assume that* (A1) *and* (A2) *are satisfied. Let* $\bar p \in (0,1)$ *and* $\rho \in (0,1]$ *be arbitrary such that* $(L_g + L_d)\rho^{1-\bar p} \le 1$. *Then there is* $C_{WR} > 0$ *only depending on* $\mu(\Omega), \|b - a\|_\infty, \|g(\bar u)\|_\infty, L_g$, *and* $L_{g'}$, *but not on* q, r, s *such that for all* $u \in \mathcal{B}^\circ$, $\|u - \bar u\|_s < \rho$,

$$\|W(u)R(u)\|_q \le C_{WR} \Phi_{\bar p}(\|u - \bar u\|_s) \|u - \bar u\|_s \ , \tag{25}$$

$$\Phi_{\bar p}(z) = \omega(2z^{\bar p})^{1/\tilde q} + z^{(1-\bar p)\min\{1, r/\tilde q\}} + \left(\frac{z}{\nu}\right)^{\frac{s-q}{q}} \tag{26}$$

where $\tilde{q} = qr/(r-q)$ *and* ω *is as in Lemma* 5.9.

Proof. Let $u \in \mathcal{B}^\circ$ be arbitrary with $\|u - \bar{u}\|_s < \rho$. According to Lemma 5.7 we have to estimate $\|Q(u)\|_{\tilde{q}}$ with $\tilde{q} = qr/(r-q)$ and Q given by (23). Let $\bar{p} \in (0,1)$ be arbitrary. We decompose Ω into the set

$$N(u) \stackrel{\text{def}}{=} N_{\|u-\bar{u}\|_s^{\bar{p}}}(u) = \left\{ x \in \Omega \,:\, |g(u)(x)| + d(u)(x) \le \|u - \bar{u}\|_s^{\bar{p}} \right\}$$

and its complement $N(u)^c$. Assumption (A2) yields with the definition of $N(u)^c$

$$\|Q(u)\|_{r,N(u)^c} \le \left\| \frac{|g(u) - g(\bar{u})|}{|g(u)| + d(u)} \right\|_{r,N(u)^c} \le \frac{\|g(u) - g(\bar{u})\|_r}{\|u - \bar{u}\|_s^{\bar{p}}} \le L_g \|u - \bar{u}\|_s^{1-\bar{p}} \le 1. \tag{27}$$

If $\tilde{q} \le r$, i.e. $r \ge 2q$, one has

$$\|Q(u)\|_{\tilde{q},N(u)^c} \le m_{\tilde{q},r} \|Q(u)\|_{r,N(u)^c},$$

and for $\tilde{q} > r$, i.e. $q < r < 2q$, application of Lemma 2.2 with $q_0 = \tilde{q}$, $q_1 = r$, $q_2 = \infty$ yields by using $\|Q(u)\|_\infty \le 1$

$$\|Q(u)\|_{\tilde{q},N(u)^c} \le \|Q(u)\|_{r,N(u)^c}^{r/\tilde{q}}.$$

Combining this and (27) gives

$$\|Q(u)\|_{\tilde{q},N(u)^c} \le C_1 \|u - \bar{u}\|_s^{(1-\bar{p})\min\{1,r/\tilde{q}\}} \tag{28}$$

with $C_1 = L_g^{\min\{1,r/\tilde{q}\}} \max\{m_{\tilde{q},r}, 1\}$. Since $\|Q(u)\|_\infty \le 1$, we get on the other hand from Lemma 5.9 and Minkowski's inequality

$$\begin{aligned} \|Q(u)\|_{\tilde{q},N(u)} \le \mu(N(u))^{1/\tilde{q}} &\le \left(\omega(2\|u - \bar{u}\|_s^{\bar{p}}) + \left(\frac{(L_g + L_d)\|u - \bar{u}\|_s}{\|u - \bar{u}\|_s^{\bar{p}}} \right)^r \right)^{1/\tilde{q}} \\ &\le \omega(2\|u - \bar{u}\|_s^{\bar{p}})^{1/\tilde{q}} + (L_g + L_d)^{r/\tilde{q}} \|u - \bar{u}\|_s^{(1-\bar{p})r/\tilde{q}}. \end{aligned} \tag{29}$$

Combining (22), (28), (29), and $\|Q(u)\|_{\tilde{q}} \le \|Q(u)\|_{\tilde{q},N(u)^c} + \|Q(u)\|_{\tilde{q},N(u)}$ gives (25). Since $m_{q_1,q_2} \le \max\{1, \mu(\Omega)\}$, it is easy to see that C_{WR} only depends on the quantities listed above. □

Our first main result is the following:

THEOREM 5.12. *Let* (O1), (O2) *and* (C) *hold at* $\bar{u}$. *If the assumptions* (A1), (A2) *and* (A3) *are satisfied then for all* $u^c \in \mathcal{B}^\circ$ *with* $\|u^c - \bar{u}\|_s < \rho_H$ *the equation* (6) *has a unique solution* $u^n \in L^q$. *In addition, for every* $\bar{p} \in (0,1)$ *and* $0 < \rho \le \rho_H$ *satisfying* $(L_g + L_d)\rho^{1-\bar{p}} \le 1$ *there is* $C > 0$ *only depending on* $\mu(\Omega), \|b - a\|_\infty, \|g(\bar{u})\|_\infty, L_g, L_{g'}$, *and* C_H, *but not on* q, r, s *such that for all* $u^c \in \mathcal{B}^\circ$ *with* $\|u^c - \bar{u}\|_s < \rho$

$$\|u^n - \bar{u}\|_q \le C \Phi_{\bar{p}}(\|u^c - \bar{u}\|_s) \|u^c - \bar{u}\|_s \tag{30}$$

with $\Phi_{\bar{p}}$ *given by* (26).

Proof. For $u^c \in \mathcal{B}^\circ$, $\|u^c - \bar{u}\|_s < \rho_H$, the unique solvability of (6) is obvious by the assumptions. Now let in addition $\|u^c - \bar{u}\|_\infty < \rho$ hold. Since $\bar{u}$ satisfies (O1), (O2), the equations (6) and (21) are equivalent. By the choice of $\rho > 0$ we may apply (A3) to obtain

$$\|u^n - \bar{u}\|_q \le \|H(u^c)^{-1}\|_{q,q} \|W(u^c)R(u^c)\|_q \le C_H \|W(u^c)R(u^c)\|_q.$$

Lemma 5.11 completes the proof with $C = C_H C_{WR}$. □

For the important case $r = s = \infty$ the proof of Theorem 5.12 can be obtained without the careful analysis of residual sets in Lemmata 5.5, 5.9 and 5.7 since these sets have measure zero for $\|u - \bar{u}\|_\infty$ small. We have the

COROLLARY 5.13. *Under the additional assumptions $r = s = \infty$ and $\rho \le \nu$, Theorem* 5.12 *holds with* (30) *replaced by*

$$\|u^n - \bar{u}\|_q \le C \left(\omega(2\|u^c - \bar{u}\|_\infty^{\bar{p}})^{1/q} + \|u^c - \bar{u}\|_\infty^{1-\bar{p}}\right) \|u^c - \bar{u}\|_\infty. \tag{31}$$

In the very likely case that in addition the strong strict complementarity condition (CS) holds we get even more:

COROLLARY 5.14. *Let in addition to the assumptions of Theorem* 5.12 *condition* (CS) *hold. Then with the choice $\bar{p} = \min\{r/(r+\tilde{q}), \tilde{q}/(\tilde{q}+\bar{q})\}$ the estimate* (30) *implies that for all $u^c \in \mathcal{B}^\circ$ with $\|u^c - \bar{u}\|_s < \rho$*

$$\|u^n - \bar{u}\|_q \le \bar{C} \left(\|u^c - \bar{u}\|_s^{\frac{\bar{q}}{\tilde{q}+\max\{1,\tilde{q}/r\}\bar{q}}} + \left(\frac{\|u^c - \bar{u}\|_s}{\nu}\right)^{\frac{s-q}{q}} \right) \|u^c - \bar{u}\|_s,$$

where $\tilde{q} = qr/(r-q)$ and $\bar{C}$ depends on $\mu(\Omega), \|b-a\|_\infty, \|g(\bar{u})\|_\infty, L_g, L_{g'}, C_H, C_C, \bar{q}$, and ε_0, but not on q, r, s. For $\rho \le (\varepsilon_0/2)^{1/\bar{p}}$ the constant $\bar{C}$ can be chosen independently of C_C, ε_0, and $\bar{q}$.

Proof. From (CS) we have $\omega(\varepsilon) \le C_C \varepsilon^{\bar{q}}$ for a fixed $\bar{q} > 0$ and all $\varepsilon \in]0, \varepsilon_0[$. Obviously, if we choose $C_C \ge \mu(\Omega)\varepsilon_0^{-\bar{q}}$ and remember $\omega(0) = 0$, the bound for $\omega(\varepsilon)$ holds for all $\varepsilon \ge 0$. We determine the optimal choice of $\bar{p}$ in (30) from

$$\frac{\bar{p}\bar{q}}{\tilde{q}} = (1-\bar{p}) \min\{1, r/\tilde{q}\}.$$

If $r \le \tilde{q}$ this gives $\bar{p} = \dfrac{r}{r+\bar{q}} \le \dfrac{\tilde{q}}{\tilde{q}+\bar{q}}$ and the common exponent $\dfrac{\bar{p}\bar{q}}{\tilde{q}} = \dfrac{\bar{q}}{\tilde{q}+\bar{q}(\tilde{q}/r)}$.
If $r > \tilde{q}$ we get $\bar{p} = \dfrac{\tilde{q}}{\tilde{q}+\bar{q}} < \dfrac{r}{r+\bar{q}}$ and $\dfrac{\bar{p}\bar{q}}{\tilde{q}} = \dfrac{\bar{q}}{\tilde{q}+\bar{q}}$. □

REMARK 5.15. It is possible to prove an even higher convergence speed by splitting Ω in the proof of Lemma 5.11 not only in $N_{\|u-\bar{u}\|_s^{\bar{p}}}(u)$ and its complement, but in $N^0(u), N^1(u) \setminus N^0(u), \ldots, N^l(u) \setminus N^{l-1}(u), N^l(u)^c$, where $N^k(u) \stackrel{\text{def}}{=} N_{\|u-\bar{u}\|_s^{\bar{p}_k}}(u)$ and $1 > \bar{p}_0 > \bar{p}_1 > \cdots > \bar{p}_l > 0$. Now the $\bar{p}_k$ can be chosen in such a way that the smallest exponent is maximized. In favor of the clarity of the presentation we have not applied this more sophisticated technique. □

For $r = s = \infty$ we state the more handy result

COROLLARY 5.16. *If in Corollary* 5.13 *the condition* (C) *is replaced by the strong strict complementarity condition* (CS), *then for all $u^c \in \mathcal{B}^\circ$ with $\|u^c - \bar{u}\|_\infty < \rho$*

$$\|u^n - \bar{u}\|_q \le \bar{C} \|u^c - \bar{u}\|_\infty^{1+\bar{q}/(q+\bar{q})}.$$

Assume for a moment that the iteration $u_0 \in \mathcal{B}^\circ$, $G(u_k)(u_{k+1} - u_k) = -d(u_k)g(u_k)$, is well-defined, i.e. in particular $(u_k) \subset \mathcal{B}^\circ$. We have already observed that the sequence (u_k) may fail to converge superlinearly in L^∞ even if (A2), (A3) hold with $q = \infty$.

As pointed out in [15] and [16] the same is true for directly applied projected Newton methods because the active set cannot be identified on a residual set of nonzero measure. In these papers a smoothing step is used to achieve fast L^∞-convergence. Theorem 5.12 will enable us to add such a modification if an appropriate smoothing operator is available. The same problems arise also in the case $s < \infty$, since a result of the form (30) requires $s > q$.

Moreover, as we will see in Example 6.3, the case $u_{k+1} \notin \mathcal{B}^\circ$ occurs very likely for some k. Hence, a back-transport into $\mathcal{B}^\circ$ is necessary. Therefore, we will use the following ingredients to design a superlinearly convergent algorithm (cf. the outline in §4):

$$\begin{aligned}
&\text{SMOOTHING STEP :} && u_k \in \mathcal{B}^\circ \longmapsto u_k^s = S_k^\circ(u_k) \in \mathcal{B}^\circ \text{ with } \|u_k^s - \bar{u}\|_s \le C_S \|u_k - \bar{u}\|_q.\\
&\text{NEWTON STEP} \quad : && u_{k+1}^n \in L^q \text{ solves } G(u_k^s)(u_{k+1}^n - u_k^s) = -d(u_k^s) g(u_k^s).\\
&\text{BACK-TRANSPORT:} && u_{k+1}^n \in L^q \longmapsto u_{k+1} = P[u_k^s](u_{k+1}^n) \in \mathcal{B}^\circ\\
& && \text{with } \|u_{k+1} - \bar{u}\|_q \le C_P \|u_{k+1}^n - \bar{u}\|_q + C_P' \|u_k^s - \bar{u}\|_s^2.
\end{aligned}$$

Here C_S, C_P, and C_P' are positive constants.

5.1. An affine-scaling Newton algorithm. Provided that smoothing step and back-transport with the above properties are available, the previous considerations and the abstract convergence theory in §4 suggest the following algorithm:

ALGORITHM 5.17 (AFFINE-SCALING INTERIOR-POINT NEWTON ALGORITHM).

1. Choose $u_0 \in \mathcal{B}^\circ$.
2. For $k = 0, 1, 2, \ldots$
   - 2.1 If $d(u_k) g(u_k) = 0$, STOP.
   - 2.2 Perform a smoothing step: $u_k^s = S_k^\circ(u_k)$.
   - 2.3 Compute $u_{k+1}^n \in L^q$ from
   $$G(u_k^s)(u_{k+1}^n - u_k^s) = -d(u_k^s) g(u_k^s) \quad \text{(Affine-scaling Newton step)}$$
   - 2.4 Transport u_{k+1}^n back to $\mathcal{B}^\circ$: $u_{k+1} = P[u_k^s](u_{k+1}^n)$.

6. Back-transport and smoothing-step.

6.1. The back-transport. Since the solution u_{k+1}^n of the affine-scaling Newton equation in step 2.3 is not necessarily an interior point of $\mathcal{B}$, a back-transport into $\mathcal{B}^\circ$ is needed. In [7] a stepsize rule is used for this purpose. A reflection technique was proposed in [4] and [6]. We will see that in our function space setting very small stepsizes σ_k may be necessary to achieve $u_k^s + \sigma_k(u_{k+1}^n - u_k^s) \in \mathcal{B}^\circ$. Thus, a stepsize rule fails to provide superlinear convergence, cf. Example 6.3. Therefore, we will propose and analyze a projection technique which is also an attractive alternative to reflection techniques in the finite-dimensional case.

6.1.1. Back-transport by projection. Since $\bar{u} \in \mathcal{B}$, the pointwise projection $P(u)$ of u onto $\mathcal{B}$ with $P : L^1 \longrightarrow \mathcal{B}$ defined by

$$P(u) = \max\{a, \min\{b, u\}\} \tag{32}$$

obviously satisfies $|P(u) - v| \le |u - v|$ on Ω for all $v \in \mathcal{B}$. Hence,

$$\|P(u) - v\|_t \le \|u - v\|_t \tag{33}$$

for all $t \in [1,\infty]$, $v \in \mathcal{B}$, and $u \in L^t$.

As mentioned earlier, an interior-point modification $P[v]$, $v \in \mathcal{B}^\circ$, of P can be used to obtain a back-transport satisfying the required property

$$\|P[v](u) - \bar{u}\|_q \le C_P \|u - \bar{u}\|_q + C_P' \|v - \bar{u}\|_s^2. \tag{34}$$

In fact, for $\xi \in (0,1)$, typically $\xi > 0.9$, and $v \in \mathcal{B}^\circ$ choose

$$P[v] : L^q \longrightarrow \mathcal{B}^\circ, \quad P[v](u) = v + \max\left\{\xi, 1 - \|P(u) - v\|_q\right\}(P(u) - v). \tag{35}$$

Then obviously

$$P[v](u) - P(u) = \min\left\{1 - \xi, \|P(u) - v\|_q\right\}(v - P(u)),$$

and hence

$$\begin{aligned} &\|P[v](u) - P(u)\|_q \le \|P(u) - v\|_q^2, \\ &\|P[v](u) - P(u)\|_t \le \|b - a\|_t \|P(u) - v\|_q, \quad 1 \le t \le \infty. \end{aligned} \tag{36}$$

Using this, we can derive (34):

LEMMA 6.1. *Let P and $P[v]$, $v \in \mathcal{B}^\circ$, be defined according to* (32) *and* (35). *Then condition* (34) *holds with $C_P = (2\|b - a\|_q + 1)$, $C_P' = 2m_{q,s}^2$.*

Proof. Let $v \in \mathcal{B}^\circ$ and $u \in L^q$. Using the properties of $P[v]$ yields

$$\begin{aligned} \|P[v](u) - \bar{u}\|_q &\le \|P[v](u) - P(u)\|_q + \|P(u) - \bar{u}\|_q \le \|P(u) - v\|_q^2 + \|u - \bar{u}\|_q \\ &\le 2\left(\|P(u) - \bar{u}\|_q^2 + \|v - \bar{u}\|_q^2\right) + \|u - \bar{u}\|_q \\ &\le (2\|b - a\|_q + 1)\|u - \bar{u}\|_q + 2\|v - \bar{u}\|_q^2 \\ &\le (2\|b - a\|_q + 1)\|u - \bar{u}\|_q + 2m_{q,s}^2 \|v - \bar{u}\|_s^2. \end{aligned}$$

□

6.1.2. Projection vs. stepsize rule for back-transport. The following arguments and Example 6.3 below show that even if (A2), (A3) hold for $q = \infty$ and $\|u_k^s - \bar{u}\|_\infty$ is arbitrarily small, stepsizes $\sigma_k \le \varepsilon \ll 1$ may be necessary to ensure $u_k^s + \sigma_k(u_{k+1}^n - u_k^s) \in \mathcal{B}^\circ$: Let $u_k^s \in \mathcal{B}^\circ$ be arbitrary. From step 2.3 we deduce for x with $d(u_k^s)(x) < c(x)$ and $g(u_k^s)(x) \ne 0$

$$u_k^s(x) - u_{k+1}^n(x) = \left(\operatorname{sgn}\left(g(u_k^s)(x)\right) + \frac{(\nabla^2 f(u_k^s)(u_{k+1}^n - u_k^s))(x)}{|g(u_k^s)(x)|}\right) d(u_k^s)(x) \tag{37}$$

If we look at those $x \in \Omega$ where in addition $\bar{u}(x) = a(x)$ and $|g(u_k^s)(x)|$ is small, say $|g(u_k^s)(x)| \le u_k^s(x) - a(x)$, then

$$d(u_k^s)(x) = u_k^s(x) - a(x)$$

and we need stepsize $\sigma_k \le \varepsilon$ if

$$\left(\nabla^2 f(u_k^s)(u_{k+1}^n - u_k^s)\right)(x) \ge \left(1 + \varepsilon^{-1}\right)|g(u_k^s)(x)|.$$

But even for $\|u_k^s - \bar{u}\|_\infty$ arbitrarily small the set

$$\left\{x \in \Omega : \ \bar{u}(x) = a(x), \ \left(\nabla^2 f(u_k^s)(u_{k+1}^n - u_k^s)\right)(x) \geq \left(1 + \varepsilon^{-1}\right) |g(u_k^s)(x)|\right\}$$

may have nonzero measure, because $|g(u_k^s)|$ is very small on a neighborhood of $\partial \bar{A}$.

Since superlinear convergence can only be guaranteed if the sequence of stepsizes converges to one, a stepsize rule for the Newton-like step is unsuitable for the infinite-dimensional case although it was proven to give quadratic convergence in the finite-dimensional case (see [7]).

REMARK 6.2. In the finite-dimensional case one can easily show by using a componentwise version of (37) that $\sigma_k = 1 - O(\|u_{k+1} - u_k\|)$ if second-order sufficiency conditions with strict complementarity hold at $\bar{u}$. See [6], [7]. □

The following example illustrates that the above scenario can really occur. Moreover, we will see that the use of a stepsize rule may lead to almost a stagnation of the iteration whereas the proposed projection technique yields fast convergence.

EXAMPLE 6.3. We consider problem (P) with quadratic objective function

$$f : u \in L^2([0,1]) \longmapsto \frac{1}{2}\|u\|_2^2 - \frac{1}{4}\left(\int_0^1 u(x)\, dx\right)^2$$

and feasible set $\mathcal{B} \stackrel{\text{def}}{=} \{u \in L^2([0,1]) : \ a(x) \stackrel{\text{def}}{=} x - \frac{1}{2} \leq u(x) \leq 10 \stackrel{\text{def}}{=} b(x) \text{ a.e.}\}$. f is smooth with

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

and strictly convex, since by Jensen's inequality $(v, \nabla^2 f(u)v)_2 \geq \frac{1}{2}\|v\|_2^2$ for all $v \in L^2$. The unique global minimum of f on $\mathcal{B}$ is given by $\bar{u}(x) = \max\{y, a(x)\}$ with $y = 3/2 - \sqrt{2}$, because f is strictly convex, $g(\bar{u}) = \bar{u} - y = 0$ on the inactive set $\bar{I} = [0, \hat{x})$, $\hat{x} = y + 1/2$, and $g(\bar{u}) = \bar{u} - y \geq 0$ on the active set $\bar{A} = [\hat{x}, 1] = \{\bar{u} = a\}$. It is easy to check that (A1)–(A3), (C), and (CS) hold for $p = q = 2$, $r = s = \infty$. For $0 < \varepsilon < 1$ the function $u_\varepsilon \in \mathcal{B}^\circ$, $u_\varepsilon(x) \stackrel{\text{def}}{=} \bar{u}(x) + \varepsilon|x - \hat{x}| + \varepsilon^2/10$, is strictly feasible with $\|u_\varepsilon - \bar{u}\|_\infty = \hat{x}\varepsilon + \varepsilon^2/10 < \varepsilon$. Moreover, the gradient $g(u_\varepsilon)$ is negative in a neighborhood of the boundary point $\hat{x}$ of $\bar{A}$ which leads to the above scenario of small stepsizes:

$$g(u_\varepsilon)(\hat{x}) = \frac{\varepsilon}{20}\left(-45 + 30\sqrt{2} + \varepsilon\right) < -\frac{\varepsilon}{20} \quad , \quad 0 < \varepsilon < 1.$$

Now we analyze what happens if we take u_ε as starting point for an affine-scaling Newton step s_ε, i.e.

$$G(u_\varepsilon)s_\varepsilon = -d(u_\varepsilon)g(u_\varepsilon),$$

or, in detail,

$$s_\varepsilon - \frac{1}{2}\frac{d(u_\varepsilon)}{d'(u_\varepsilon)g(u_\varepsilon) + d(u_\varepsilon)}\int_0^1 1 \cdot s_\varepsilon(x)\, dx = -\frac{d(u_\varepsilon)g(u_\varepsilon)}{d'(u_\varepsilon)g(u_\varepsilon) + d(u_\varepsilon)}. \tag{38}$$

Since the operator $(d'(u_\varepsilon)g(u_\varepsilon) + d(u_\varepsilon))^{-1}G(u_\varepsilon)$ on the left (which coincides with $H(u_\varepsilon)$ for ε small enough) is a 'rank-one modification' of the identity, its inverse

can be explicitly determined by applying the Sherman-Morrison-Woodbury-Lemma in $L^2([0,1])$. It is possible to derive a closed formula for s_ε. Table 1 shows the maximum stepsize $\sigma_{\max} \stackrel{\text{def}}{=} \max\{\sigma \in [0,1] : u_\varepsilon + \sigma s_\varepsilon \in \mathcal{B}\}$ for $c \stackrel{\text{def}}{=} 2$, and the relative L^2-norm of the part of s_ε that would be cut off by a pointwise projection. Fig. 1 depicts for $\varepsilon = 1/100$ a plot of $-s_\varepsilon$ and $u_\varepsilon - a$ (dashed) close to the sign change of the gradient $g(u_\varepsilon)$ at $x_0 = 0.58706$.

ε	$\sigma_{\max}$	$\frac{\|(u_\varepsilon+s_\varepsilon)-P(u_\varepsilon+s_\varepsilon)\|_2}{\|s_\varepsilon\|_2}$
1.0E−2	1.77E−2	4.88E−3
1.0E−3	1.78E−3	1.41E−3
1.0E−4	1.78E−4	4.43E−4
1.0E−5	1.78E−5	1.40E−4

TAB. 1

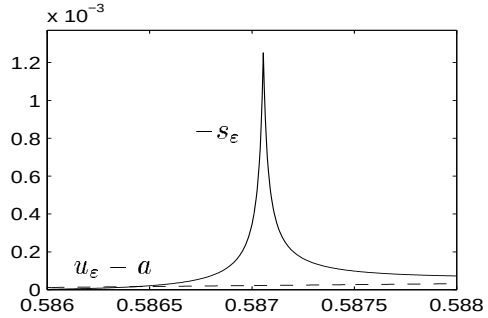

FIG. 1

Apparently, a stepsize rule yields very small stepsizes whereas the pointwise projection leads only to a tiny change of the step with respect to the L^2-norm.

Now we compare the performance of two variants of the affine-scaling interior-point Newton method:

(I) Algorithm 5.17.

(II) Algorithm 5.17 with 2.4 replaced by a stepsize rule:

2.4′ Transport u^n_{k+1} back to $\mathcal{B}^\circ$ by the following stepsize rule:

$$s^n_k = u^n_{k+1} - u^s_k \quad , \quad \sigma_{k,\max} = \max\{\sigma \in [0,1] : u^s_k + \sigma s^n_k \in \mathcal{B}\} \quad ,$$
$$u_{k+1} = u^s_k + \max\{\xi, 1 - \|s^n_k\|_2\}\, \sigma_{k,\max} s^n_k \quad , \quad \xi \text{ as in (35)}.$$

In both variants we apply smoothing to u_k only if the L^2- and L^∞-norm of $u_k - u^s_{k-1}$ differ too much:

$$S^\circ_k(u) \stackrel{\text{def}}{=} \begin{cases} P[u](u - g(u)) & \text{if } k \geq 1,\ u = u_k, \text{ and } \|u_k - u^s_{k-1}\|_\infty \geq 3\|u_k - u^s_{k-1}\|_2, \\ u & \text{else.} \end{cases}$$

The interior-point modification of the projected gradient step is in deed a smoothing step. This follows from Remark 4.3 and the discussion in §8. For the numerical realization of both methods we have discretized the problem by approximating $L^2([0,1])$ with piecewise linear functions on a uniform grid with 200 points. To check the decrease properties of the new iterates, we use the fact that u^n_{k+1} solves the affine-scaling Newton equation in step 2.3 if and only if u^n_{k+1} is a stationary point of the quadratic function $\psi[u^s_k](u)$ defined by (57), cf. §10. This function is used as quadratic model in the interior-point trust-region methods recently analyzed in [24]. Since $\psi[u^s_k]$ is strictly convex in our context, it attains its global minimum at u^n_{k+1}. We start both iterations with $u_0 = u_\varepsilon$, $\varepsilon = 0.5$, and $\xi = 0.999995$. The distances $\|u_{k+1} - \bar{u}\|_2$, $\|u^s_{k+1} - \bar{u}\|_\infty$ from the solution $\bar{u}$ of the discrete problem and the decrease ratio $\psi[u^s_k](u_{k+1})/\psi[u^s_k](u^n_{k+1})$ are shown in Table 2. For method (II) we have also added $\sigma_{k,\max}$. Fig. 2.1 depicts $-s^n_0$ and the distance to the lower bound $u_0 - a$ (dashed). Fig. 2.2 and 2.3 show the same quantities, i.e. $-s^n_1$ and $u_1 - a$, after one step

of algorithm (I) and (II), respectively. We see that the stepsize rule in (II) leads to an iterate u_1 and a new search direction s_1^n that requires a very small stepsize of 0.0128 yielding almost no progress. The reason is depicted in Fig. 2.3: s_1^n has a small peak on the set where the distance to the lower bound is small. On the other hand, the part of $s_k^n = u_{k+1}^n - u_k^s$ that is cut off by a projection is very small. Hence, the projection leads to a nearly optimal decrease of $\psi[u_k^s]$ in every step. Fig. 3.1 and 3.2 show the first iterates for both iterations. While our algorithm (I) converges in 5 steps to high accuracy, method (II) needs 28 iterations to enter the region of quadratic convergence which exists according to the finite-dimensional theory. Then it converges in two more steps to high accuracy.

	Alg. (I) (Projection)			Alg. (II) (Stepsize rule)			
k	$\|u_{k+1} - \bar{u}\|_2$	$\|u_{k+1}^s - \bar{u}\|_\infty$	r_k †	$\|u_{k+1} - \bar{u}\|_2$	$\|u_{k+1}^s - \bar{u}\|_\infty$	r_k †	$\sigma_{k,\max}$
0	4.2275E−2	8.1290E−2	0.9999	7.6898E−2	1.4443E−1	0.9288	0.7332
1	4.2356E−3	8.5289E−3	0.9998	7.6053E−2	1.4287E−1	0.0255	0.0128
2	1.8431E−4	1.5081E−3	1.0000	7.4395E−2	1.3981E−1	0.0502	0.0254
3	4.3224E−5	3.1222E−6*	1.0000	7.1203E−2	1.3391E−1	0.0973	0.0499
4	6.3244E−10	8.7489E−9	1.0000	6.5276E−2	1.2295E−1	0.1833	0.0963

† $r_k = \psi[u_k^s](u_{k+1})/\psi[u_k^s](u_{k+1}^n)$ * Smoothing occurred, i.e. $u_{k+1}^s \neq u_{k+1}$

TAB. 2

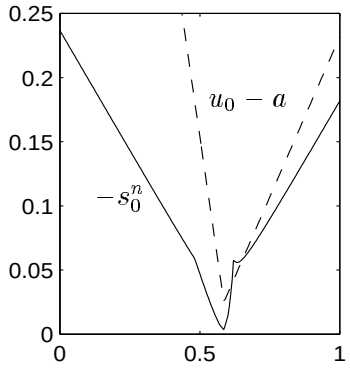

FIG. 2.1

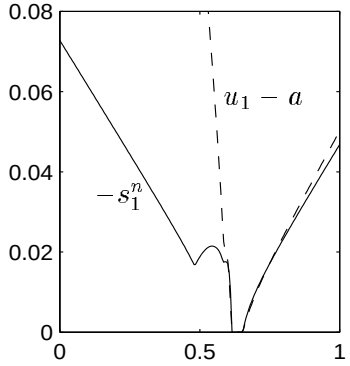

FIG. 2.2

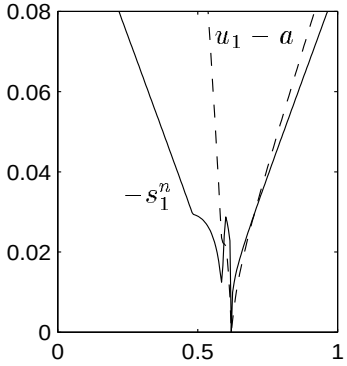

FIG. 2.3

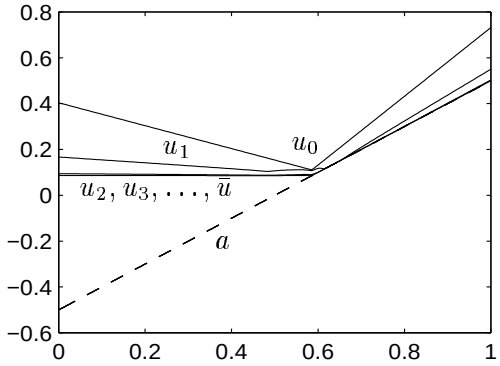

FIG. 3.1

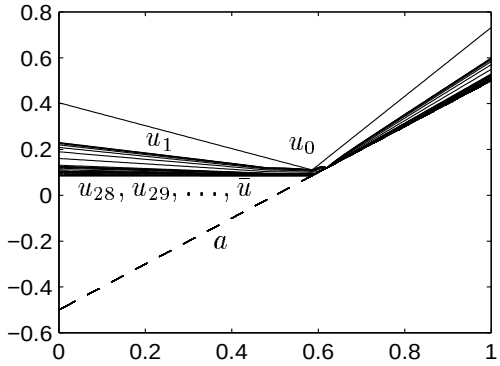

FIG. 3.2

□

These considerations make it evident that the projection technique should be used

instead of a stepsize rule to obtain an interior point $u_{k+1} \in \mathcal{B}^\circ$ from u_{k+1}^n.

6.2. The smoothing step. We have already observed that a smoothing step is necessary because the strongest available estimate after one iteration of (6) is (30) with $q < s$. In the further analysis we assume that a smoothing operator

$$S_k : \mathcal{B} \subset L^q \longrightarrow L^s \quad , \quad k \geq 0 \tag{39}$$

is available. Let $\bar{u} \in \mathcal{B}$ satisfy (O1) and (O2). In order to require smoothing only if it is really necessary we make the following

ASSUMPTION (SMOOTHING PROPERTY).

(S) There are $\rho_S > 0$ and $L_S > 0$ such that the operators S_k defined in (39) possess the following smoothing property:

$$\|S_k(u_k) - \bar{u}\|_s \leq L_S \|u_k - \bar{u}\|_q \quad \text{for all } k \text{ with } \|u_k - \bar{u}\|_q < \rho_S.$$

This assumption allows to choose $S_k(u) = u$ on the set $\{u : \|u - \bar{u}\|_s \leq L_S \|u - \bar{u}\|_q\}$ where smoothing is not necessary. As already outlined in Remark 4.3, the operator $S_k^\circ : u \longmapsto P[u](S_k(u))$ is an interior-point modification of S_k:

LEMMA 6.4. *Let $\bar{u}$ satisfy* (O1), (O2) *and let* (S) *hold. If $P[v]$ is defined by* (35) *then $S_k^\circ : u \in \mathcal{B}^\circ \longmapsto P[u](S(u))$ satisfies $S_k^\circ(\mathcal{B}^\circ) \subset \mathcal{B}^\circ$ and*

$$\|S_k^\circ(u_k) - \bar{u}\|_s \leq C_S \|u_k - \bar{u}\|_q$$

for all k with $\|u_k - \bar{u}\|_q < \rho_S$, where $C_S = (m_{q,s}\|b - a\|_s + 1)L_S + \|b - a\|_s$.

Proof. $P[u](S_k(u)) \in \mathcal{B}^\circ$ does obviously hold for all $u \in \mathcal{B}^\circ$ and $k \geq 0$. Now let $k \geq 0$ be arbitrary with $\|u_k - \bar{u}\|_q < \rho_S$. Using the properties (33), (36) of P and $P[u_k]$, we get

$$\begin{aligned}
\|P[u_k](S_k(u_k)) - \bar{u}\|_s &\leq \|P[u_k](S_k(u_k)) - P(S_k(u_k))\|_s + \|P(S_k(u_k)) - \bar{u}\|_s \\
&\leq \|b - a\|_s \|P(S_k(u_k)) - u_k\|_q + \|S_k(u_k) - \bar{u}\|_s \\
&\leq \|b - a\|_s \left(\|P(S_k(u_k)) - \bar{u}\|_q + \|u_k - \bar{u}\|_q \right) + \|S_k(u_k) - \bar{u}\|_s \\
&\leq (m_{q,s}\|b - a\|_s + 1)\|S_k(u_k) - \bar{u}\|_s + \|b - a\|_s \|u_k - \bar{u}\|_q \\
&\stackrel{\text{def}}{=} \bar{C}_P \|S_k(u_k) - \bar{u}\|_s + \bar{C}'_P \|u_k - \bar{u}\|_q \leq (\bar{C}_P L_S + \bar{C}'_P)\|u_k - \bar{u}\|_q,
\end{aligned}$$

where we have used (S) in the last step. □

We will show in §8 how a smoothing operator can be constructed for a class of regularized problems by using a fixed point formulation of the KKT-conditions (O1),(O2).

7. The convergence result. In the following we will always work with the smoothing operator

$$S_k^\circ : u \in \mathcal{B}^\circ \longmapsto S_k^\circ(u) \stackrel{\text{def}}{=} P[u](S_k(u)) \in \mathcal{B}^\circ \quad , \quad S_k \text{ as in (39)}.$$

We will now prove that Algorithm 5.17 converges superlinearly (resp. with Q-order $1 + \bar{q}/(\tilde{q} + \max\{1, \tilde{q}/r\}\,\bar{q})$) in L^s to $\bar{u}$ if $\bar{u}$ satisfies the first-order necessary conditions

with strict complementarity (C) (resp. (CS)) as well as (A3), a smoothing step exists, and $\|u_0 - \bar{u}\|_q$ is small enough. More precisely, we have the

THEOREM 7.1. *Let $\bar{u}$ satisfy* (O1), (O2) *and* (C). *If* (A1)–(A3) *and* (S) *hold then for $\bar{p} \in (0,1)$ there is $\rho > 0$ such that for all $u_0 \in \mathcal{B}^\circ$ with $\|u_0 - \bar{u}\|_q < \rho$ Algorithm* 5.17 *is well-defined and produces iterates with*

$$\|u_{k+1} - \bar{u}\|_q \le \bar{C}_1 \Phi_{\bar{p}}(C_S \|u_{k+1} - \bar{u}\|_q) \|u_{k+1} - \bar{u}\|_q \tag{40}$$

$$\|u^s_{k+1} - \bar{u}\|_s \le \bar{C}_2 \Phi_{\bar{p}}(\|u^s_{k+1} - \bar{u}\|_s) \|u^s_{k+1} - \bar{u}\|_s \tag{41}$$

where $\bar{C}_1, \bar{C}_2 > 0$ depend on $\mu(\Omega), \|b-a\|_\infty, \|g(\bar{u})\|_\infty, L_g, L_{g'}, C_H, L_S$, but not on q, r, s and $\Phi_{\bar{p}}$ is given by (26), *i.e.*

$$\Phi_{\bar{p}}(z) = \omega(2z^{\bar{p}})^{1/\tilde{q}} + z^{(1-\bar{p})\min\left\{1, \frac{r}{\tilde{q}}\right\}} + \left(\frac{z}{\nu}\right)^{\frac{s-q}{q}}.$$

In the case $r = s = \infty$ the function $\Phi_{\bar{p}}$ simplifies to $\Phi_{\bar{p}}(z) = \omega(2z^{\bar{p}})^{1/q} + z^{1-\bar{p}}$.

Proof. Choose $0 < \rho \le \rho_S$. We will reduce ρ as the proof proceeds. From Lemma 6.4 we deduce

$$\|u^s_k - \bar{u}\|_s \le C_S \|u_k - \bar{u}\|_q \tag{42}$$

By choosing $\rho > 0$ appropriately we can apply Theorem 5.12 with ρ replaced by $C_S\rho$ and $u^c = u^s_k$. We obtain that for all $u_k \in \mathcal{B}^\circ$, $\|u_k - \bar{u}\|_q < \rho$,

$$\|u^n_{k+1} - \bar{u}\|_q \le C\Phi_{\bar{p}}(\|u^s_k - \bar{u}\|_s)\|u^s_k - \bar{u}\|_s \le CC_S\Phi_{\bar{p}}(C_S\|u_k - \bar{u}\|_q)\|u_k - \bar{u}\|_q \tag{43}$$

where $\Phi_{\bar{p}}$ is given by (26). Lemma 6.1 yields

$$\begin{aligned} \|u_{k+1} - \bar{u}\|_q &= \|P[u^s_k](u^n_{k+1}) - \bar{u}\|_q \le C_P\|u^n_{k+1} - \bar{u}\|_q + C'_P\|u^s_k - \bar{u}\|^2_s \\ &\le C_S\left(CC_P\Phi_{\bar{p}}(C_S\|u_k - \bar{u}\|_q) + C_SC'_P\|u_k - \bar{u}\|_q\right)\|u_k - \bar{u}\|_q. \end{aligned} \tag{44}$$

This proves (40), since the $\Phi_{\bar{p}}$-term is of lowest order. By the properties of ω, see Lemma 5.9, $\Phi_{\bar{p}}(z)$ tends to zero as $z \to 0$. Hence, possibly after a further reduction of ρ, the algorithm is well-defined, since $u_0 \in \mathcal{B}^\circ$, $\|u_0 - \bar{u}\|_q < \rho$ implies $u_k \in \mathcal{B}^\circ$, $\|u_k - \bar{u}\|_q < \rho$ for all k. Now (41) is obtained by combining (42) with k replaced by $k+1$, and the first inequalities in (43) and (44):

$$\begin{aligned} \|u^s_{k+1} - \bar{u}\|_s &\le C_S\|u_{k+1} - \bar{u}\|_q \\ &\le C_S\left(C_P\|u^n_{k+1} - \bar{u}\|_q + C'_P\|u^s_k - \bar{u}\|^2_s\right) \\ &\le C_S\left(CC_P\Phi_{\bar{p}}(\|u^s_k - \bar{u}\|_s) + C'_P\|u^s_k - \bar{u}\|_s\right)\|u^s_k - \bar{u}\|_s. \end{aligned}$$

□

If in addition (CS) holds we get convergence with Q-order > 1:

COROLLARY 7.2. *Let in addition to the assumptions of Theorem* 7.1 *condition* (CS) *hold at $\bar{u}$. Then with the choice $\bar{p} = \min\{r/(r+\tilde{q}), \tilde{q}/(\tilde{q}+\tilde{q})\}$ Theorem* 7.1 *yields*

$$\|u_{k+1} - \bar{u}\|_q \le \bar{C}_1\Phi_{CS}(C_S\|u_k - \bar{u}\|_q)\|u_k - \bar{u}\|_q$$

$$\|u^s_{k+1} - \bar{u}\|_s \le \bar{C}_2\Phi_{CS}(\|u^s_k - \bar{u}\|_s)\|u^s_k - \bar{u}\|_s$$

with $\bar{C}_1, \bar{C}_2 > 0$ *as in Theorem* 7.1, $\tilde{q} = \dfrac{qr}{r-q}$, *and*

$$\Phi_{CS}(z) = z^{\frac{\bar{q}}{\bar{q}+\max\{1,\tilde{q}/r\}\bar{q}}} + \left(\frac{z}{\nu}\right)^{\frac{s-q}{q}}.$$

In the case $r = s = \infty$ *the function* Φ_{CS} *assumes the simple form* $\Phi_{CS}(z) = z^{\frac{\bar{q}}{q+\bar{q}}}$.

Proof. This follows immediately from Theorem 7.1 and Corollary 5.14. □

8. Application to a class of regularized problems. In this section we apply our convergence theory to the following class of regularized problems which contains the one considered in the analysis of projected Newton methods by Kelley and Sachs [15]: We investigate problem (P) with the L^2-regularized objective function

$$f : u \in \mathcal{D} \subset L^p \longmapsto k(u) + \frac{1}{2}\|\sqrt{\alpha}(u - u^0)\|_2^2,$$

where $\alpha, u^0 \in L^\infty$ and $k : \mathcal{D} \mapsto \mathbb{R}$ such that (A1) holds. The gradient is given by

$$g(u) = \alpha u - \alpha u^0 + \nabla k(u) \stackrel{\text{def}}{=} \alpha u + K(u).$$

We make the following

ASSUMPTION.

(A2′) $g(u) = \alpha u + K(u)$ with $\alpha \in L^\infty$, $\alpha(x) \geq \alpha_0 > 0$ for a.a. $x \in \Omega$. Furthermore, there are $2 \leq q < s \leq \infty$ such that $g : \mathcal{B} \subset L^s \longrightarrow L^q$ is Lipschitz continuously Fréchet differentiable and K has the following smoothing property:

$$K : \mathcal{B} \subset L^q \longrightarrow L^s$$

is Lipschitz continuous with Lipschitz constant L_K.

Obviously, (A2′) implies the Lipschitz continuity of $g : \mathcal{B} \subset L^s \longrightarrow L^s$ with Lipschitz constant $L_g = \|\alpha\|_\infty + m_{q,s}L_K$. Hence, (A2′) implies (A2) with $r = s$.

To perform a smoothing step we use a technique proposed in [15]. The following fixed point formulation of the optimality conditions (O1),(O2) is essential:

LEMMA 8.1. *Let* (A1) *hold. Then* (O1),(O2) *are satisfied at* $\bar{u}$ *if and only if*

$$\bar{u} = P(\bar{u} - \sigma g(\bar{u})) \tag{45}$$

where $\sigma \in L^\infty$, $\sigma > 0$ *a.e., is arbitrary.*

If in addition (A2′) *holds then* $\bar{u}$ *satisfies* (O1),(O2) *if and only if*

$$\bar{u} = P(-\alpha^{-1}K(\bar{u})) \ (= P(\bar{u} - \alpha^{-1}g(\bar{u}))).$$

Furthermore, for all $u, v \in \mathcal{B}$ *the following holds true:*

$$\left\|P(-\alpha^{-1}K(u)) - P(-\alpha^{-1}K(v))\right\|_s \leq \left\|\alpha^{-1}(K(u) - K(v))\right\|_s \leq \frac{L_K}{\alpha_0}\|u - v\|_q, \tag{46}$$

i.e. $S_k : u \in \mathcal{B} \longmapsto P(-\alpha^{-1}K(u))$ *has the smoothing property* (S).

Proof. Let $\bar{u} \in \mathcal{D}$ be arbitrary. Then (45) is satisfied if and only if (O1) holds (since the right hand side of (45) is in $\mathcal{B}$) and

$$\sigma(x)g(\bar{u})(x) \begin{cases} \geq 0, & \text{if } \bar{u}(x) = a(x), \\ \leq 0, & \text{if } \bar{u}(x) = b(x), \\ = 0, & \text{else} \end{cases} \qquad \text{a.e. on } \Omega.$$

Since $\sigma(x) > 0$ for a.a. $x \in \Omega$, this is nothing else but (O2). If in addition (A2′) holds then $\bar{u} - \sigma g(\bar{u}) = (1 - \sigma\alpha)\bar{u} - \sigma K(\bar{u})$ and the choice $\sigma = \alpha^{-1}$ establishes the second assertion. (46) is easily obtained by using the smoothing property in (A2′) and the fact that $\|P(v) - P(w)\|_s \leq \|v - w\|_s$ for all $v, w \in L^s$, since P is a pointwise projection. □

REMARK 8.2. The smoothing step is a scaled projected gradient step obtained by making a scaled gradient step $-\alpha^{-1}g(u)$ and projecting the result pointwise onto $\mathcal{B}$. Moreover, $P(-\alpha^{-1}K(u)) - u$ is a descent direction for f at $u \in \mathcal{B}$ (cf. [11]): since P is also the projection onto $\mathcal{B}$ in the scaled Hilbert space $(L^2, (\alpha\cdot\,.\,,\,.\,)_2)$ we get by using well known properties of projections on closed convex sets in Hilbert space

$$0 \geq \Big(\alpha(u - P(u - \alpha^{-1}g(u))), u - \alpha^{-1}g(u) - P(u - \alpha^{-1}g(u))\Big)_2$$

and hence (note that we use the L^2 inner product as dual pairing)

$$\Big\langle P(u - \alpha^{-1}g(u)) - u, g(u)\Big\rangle \leq -\Big\|\sqrt{\alpha}(P(u - \alpha^{-1}g(u)) - u)\Big\|_2^2.$$

□

The preceding Lemma shows that the convergence results of the previous section hold for the considered class of regularized problems if $S_k(u) = P(-\alpha^{-1}K(u))$ is used as smoothing operator.

We have already mentioned that (CS) is weaker than the corresponding assumption in [15] for the analysis of the projected Newton method. To allow a further comparison with the results in [15] we will show that assumption (A3) is implied by Assumptions 2.1 and 2.3 in [15] which are stronger than (A2′) and the requirement that

$$\tilde{H}(u) \stackrel{\text{def}}{=} I + \alpha^{-1}\chi_{\bar{I}}K'(u)\chi_{\bar{I}}, \quad \bar{I} = \Omega \setminus \bar{A}, \tag{47}$$

has an inverse for all $u \in \mathcal{B}$, $\|u - \bar{u}\|_s < \tilde{\rho}$ with $\|\tilde{H}(u)^{-1}\|_{q,q} \leq C_{\tilde{H}}$ (in [15] only $s = \infty$ is considered).

We use the following analogue of Assumption 2.2 in [15] which implies the local Lipschitz continuity of $K : \mathcal{B} \subset L^q \longrightarrow L^s$ in $\bar{u}$:

ASSUMPTION.

(A4) There is $\rho_K > 0$ such that for all $u \in \mathcal{B}$ with $\|u - \bar{u}\|_s < \rho_K$

$$\|K'(u)\|_{q,s} \leq C_{K'}.$$

Here $K'(u) \in \mathcal{L}(L^p, L^{p'})$ denotes the Fréchet derivative of K at u.

The following Lemma shows that (A3) is satisfied if (A3) holds for $\tilde{H}(u)$ defined in (47) instead of $H(u)$. Hence, (A3) is implied by Assumptions 2.2 and 2.3. in [15] for the choice $s = \infty$.

LEMMA 8.3. *Let* (O1), (O2) *and* (C) *hold at* $\bar{u}$, $\bar{A}$ *denote the active set, and* $\bar{I} = \bar{A}^c$. *If the assumptions* (A1), (A2′) *and* (A4) *are satisfied, then the following is true: If there is* $\bar{\rho} > 0$ *such that for all* $u \in \mathcal{B}$, $\|u - \bar{u}\|_s < \bar{\rho}$,

$$\bar{H}(u) \stackrel{\text{def}}{=} I + \alpha^{-1}\chi_{\bar{I}}K'(u) : L^q \longrightarrow L^q$$

is invertible with $\|\bar{H}(u)^{-1}\|_{q,q} \le C_{\bar{H}}$ *then* (A3) *holds for* $\rho_H > 0$ *sufficiently small.*

The Lemma remains true if $\bar{H}(u)$ *is replaced by* $\tilde{H}(u)$ *defined in* (47) *since the uniformly bounded invertibility of* $\tilde{H}(u)$ *implies the one of* $\bar{H}(u)$:

$$\bar{H}(u)^{-1} = \tilde{H}(u)^{-1} \cdot \left(I - \alpha^{-1}\chi_{\bar{I}}K'(u)\chi_{\bar{A}}I\right) \tag{48}$$

Proof. We note that $H(u)$ in assumption (A3) can be equivalently replaced by

$$\hat{H}(u) \stackrel{\text{def}}{=} \frac{\chi_{\{d(u)<c\}}|g(u)|}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}I + \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\nabla^2 f(u)$$

if (A1) and (A2′) are satisfied. This follows from the identity

$$\hat{H}(u) = \frac{|g(u)| + d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}H(u)$$

and the fact that the first factor is continuously invertible, since

$$\frac{|g(u)| + d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)} \in \begin{cases} \left[\min\left\{\|\alpha\|_\infty^{-1}, 1\right\}, \max\left\{\alpha_0^{-1}, 1\right\}\right] & \text{on } \{d(u) < c\}, \\ \left[\|\alpha\|_\infty^{-1}, \alpha_0^{-1}(1 + \nu^{-1}C_g)\right] & \text{on } \{d(u) \ge c\}. \end{cases}$$

In particular, there exists a constant $C_{\hat{H}H}$ with

$$\|H(u)^{-1}\|_{q,q} \le C_{\hat{H}H}\|\hat{H}(u)^{-1}\|_{q,q}.$$

According to a standard result of operator theory we can establish (A3) with $C_H = 2C_{\hat{H}H}C_{\bar{H}}$ by finding $\rho_H > 0$ such that $\|\bar{H}(u) - \hat{H}(u)\|_{q,q} \le 1/(2C_{\bar{H}})$ for all $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_s < \rho_H$. To this end, let $\rho \le \min\{1, \bar{\rho}, \rho_K\}$ and $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_s < \rho$, be arbitrary. We will adjust ρ as the proof proceeds. We observe that

$$\bar{H}(u) - \bar{H}(u) = \left(\frac{\chi_{\bar{I}}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right)K'(u), \tag{49}$$

apply Lemma 2.3 with $q_0 = q$, $q_1 = s/q$, $q_1' = s/(s-q)$, use (A4), and obtain

$$\begin{aligned} \|\bar{H}(u) - \hat{H}(u)\|_{q,q} &\le \left\|\frac{\chi_{\bar{I}}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{\frac{qs}{s-q}} \|K'(u)\|_{q,s} \\ &\le \left\|\frac{\chi_{\bar{I}}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)}\right\|_{\frac{qs}{s-q}} C_{K'}. \end{aligned}$$

We notice that $q \le \hat{q} \stackrel{\text{def}}{=} qs/(s-q) < \infty$ ($\hat{q} = q$ if $s = \infty$) and split Ω to estimate the first factor in the last expression. For

$$B(u) \stackrel{\text{def}}{=} \left\{x \in \Omega : \chi_{\{d(u)<c\}}(x)|g(u)(x)| + \alpha(x)d(u)(x) \le \sqrt{\rho}\right\}$$

and $\rho < \nu^2\alpha_0^2$ we have $B(u) \subset \{d(u) < c\}$, and thus with $\alpha_1 = \min\{\alpha_0, 1\}$

$$B(u) \subset \{x \in \Omega : \ \alpha_1|g(u)(x)| + \alpha_1 d(u)(x) \leq \sqrt{\rho}\} = N_{\sqrt{\rho}/\alpha_1}(u).$$

Denote the complement of $B(u)$ by $B^c(u)$. The key observation is that the parenthesis in (49) is small on $B^c(u)$ and the measure of the residual set $B(u)$ is small as well. We get by Lemma 5.9 and Minkowski's inequality

$$\begin{aligned}\left\| \frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)} \right\|_{\hat q, B(u)} &\leq \frac{1}{\alpha_0}\mu(B(u))^{1/\hat q} \leq \frac{1}{\alpha_0}\mu(N_{\sqrt{\rho}/\alpha_1}(u))^{1/\hat q} \\ &\leq \frac{1}{\alpha_0}\left(\omega\left(2\frac{\sqrt{\rho}}{\alpha_1}\right) + ((L_g + L_d)\alpha_1\sqrt{\rho})^s\right)^{1/\hat q} \\ &\leq \frac{1}{\alpha_0}\left(\omega\left(2\frac{\sqrt{\rho}}{\alpha_1}\right)^{1/\hat q} + ((L_g + L_d)\alpha_1\sqrt{\rho})^{s/\hat q}\right).\end{aligned}$$

Moreover, we obtain as in the proof of Theorem 5.12

$$\left\| \frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)} \right\|_{\hat q, B^c(u)} \leq C_1 \left\| \frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)} \right\|_{s, B^c(u)}^{\min\{1, s/\hat q\}}$$

by applying Lemma 2.2 in the case $s < \hat q$ (note that the function under the norm is nonnegative and pointwise bounded by $1/\alpha_0$). We can choose $C_1 = m_{\hat q, s}$ if $\hat q \leq s$ and $C_1 = \alpha_0^{-1+s/\hat q}$ if $s < \hat q$. Since $g(\bar u) = 0$ a.e. on $\bar I$ by (O2) we get with (A2)

$$\begin{aligned}\left\| \frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)} \right\|_{s, B^c(u)\cap \bar I} &= \left\| \frac{\chi_{\{d(u)<c\}}|g(u) - g(\bar u)|}{\alpha\left(\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)\right)} \right\|_{s, B^c(u)\cap \bar I} \\ &\leq \left\| \frac{g(u) - g(\bar u)}{\alpha_0\sqrt{\rho}} \right\|_s \leq \frac{L_g}{\alpha_0}\sqrt{\rho}.\end{aligned}$$

The fact that $d(\bar u) = 0$ a.e. on $\bar A$ yields together with Lemma 5.5

$$\begin{aligned}\left\| \frac{\chi_{\bar I}}{\alpha} - \frac{d(u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)} \right\|_{s, B^c(u)\setminus \bar I} &= \left\| \frac{d(u) - d(\bar u)}{\chi_{\{d(u)<c\}}|g(u)| + \alpha d(u)} \right\|_{s, B^c(u)\setminus \bar I} \\ &\leq \left\| \frac{d(u) - d(\bar u)}{\sqrt{\rho}} \right\|_s \leq L_d\sqrt{\rho}.\end{aligned}$$

Hence, there are constants $C_1, C_2 > 0$ such that

$$\|\bar H(u) - \hat H(u)\|_{q,q} \leq \left(C_1\omega\left(2\frac{\sqrt{\rho}}{\alpha_1}\right)^{1/\hat q} + C_2\sqrt{\rho}^{\min\left\{1, \frac{s}{\hat q}\right\}}\right) C_{K'}.$$

Due to Lemma 5.9, after a possible reduction of $\rho > 0$ the right hand side is $\leq 1/(2C_{\bar H})$ and the choice $\rho_H = \rho$ completes the first part of the proof.

Now assume that the assumptions hold for $\tilde H(u)$ instead of $\bar H(u)$. We only have to verify the explicit formula (48) for $\bar H(u)^{-1}$. For $v \in L^q$ we look at the equation

$$v = \bar H(u)h = \tilde H(u)h + \alpha^{-1}\chi_{\bar I}K'(u)\chi_{\bar A}h. \tag{50}$$

Premultiplication by $\chi_{\bar{A}}$ shows $h_{\bar{A}} = v_{\bar{A}}$, and hence

$$h = \tilde{H}(u)^{-1}\left(v - \alpha^{-1}\chi_{\bar{I}}K'(u)v_{\bar{A}}\right).$$

Therefore, the operator given in (48) is a left inverse of $\bar{H}(u)$. It is also a right inverse; to see this we note that $\chi_{\bar{A}}\tilde{H}(u) = \chi_{\bar{A}}I$, hence $\chi_{\bar{A}}\tilde{H}(u)^{-1} = \chi_{\bar{A}}I$, and, consequently,

$$(\tilde{H}(u) + \alpha^{-1}\chi_{\bar{I}}K'(u)\chi_{\bar{A}}I)\tilde{H}(u)^{-1}(I - \alpha^{-1}\chi_{\bar{I}}K'(u)\chi_{\bar{A}}I) = I,$$

where we have used that $\chi_{\bar{A}}\alpha^{-1}\chi_{\bar{I}} = 0$. □

REMARK 8.4. The results of Lemma 8.3 remain true if α also depends on u. □

9. Second-order sufficient conditions. We will now study how Algorithm 5.17 behaves in the neighborhood of a point $\bar{u}$ satisfying the second-order sufficient condition given by Dunn and Tian in [9]. We will show that it implies (A3) in the case $q = 2$ under the additional assumptions of the previous section and also for $q > 2$ if the range of $\tilde{H}(u)$ is dense in L^q. In §10 we will use this sufficiency condition to show that the developed affine-scaling Newton method produces acceptable steps for the trust-region globalization considered in [24] if the iterates u_k are close enough to $\bar{u}$. In our notation, the formal second-order sufficiency conditions by Dunn and Tian [9] read

ASSUMPTION (SECOND-ORDER SUFFICIENT CONDITIONS BY DUNN AND TIAN).

(OS) Condition (A1) holds and there are $t \in [1,\infty]$, $c_r > 0$ such that

$$|\langle v, \nabla^2 f(u)w\rangle| \le c_r\|v\|_2\|w\|_2 \quad \forall\ u \in \mathcal{B}, v, w \in L^\infty \tag{51}$$

$$\lim_{\substack{u\in\mathcal{B}\\ \|u-\bar{u}\|_t\to 0}} \sup_{\substack{w\in L^\infty\\ \|w\|_2=1}} \langle w, (\nabla^2 f(u) - \nabla^2 f(\bar{u}))w\rangle = 0. \tag{52}$$

Moreover, (O1),(O2) are satisfied at $\bar{u}$ and there are sets $A \subset \bar{A}$, $I = A^c$, and constants $c_1, c_2 > 0$ with

$$g(\bar{u}) \ge c_1 \ \text{ on } A \quad , \quad \langle \chi_I w, \nabla^2 f(\bar{u})\chi_I w\rangle \ge c_2\|\chi_I w\|_2^2 \ \ \forall\ w \in L^\infty. \tag{53}$$

REMARK 9.1. Condition (OS) is weaker (stronger) than the sufficient second-order condition of Maurer in [21] if $\|\cdot\|_2$ (resp. $\|\cdot\|_1$) is chosen as the weak norm. Since we prefer a result of the form $f(u) - f(\bar{u}) \ge C\|u - \bar{u}\|_l^2$ for $l = 2$ rather than for $l = 1$, condition (OS) meets our requirements better. Moreover it is obvious that in view of Lemma 2.2 the requirement $t \in [1,\infty)$ could be equivalently replaced by $t \in \{2,\infty\}$ since the relative topology of L^t on $\mathcal{B}$ is the same for all $t \in [1,\infty)$. □

9.1. L^∞-optimality. The following Theorem shows that (OS) implies the L^∞-optimality of $\bar{u}$ for (P) (cf. [9]).

THEOREM 9.2. *Let the formal second-order sufficiency condition* (OS) *hold. Then $\bar{u}$ is a strict L^∞-optimizer for* (P)*, more precisely: there are $\rho > 0$ and $C > 0$ such that*

$$u \in \mathcal{B},\ \|u - \bar{u}\|_\infty < \rho \Longrightarrow f(u) - f(\bar{u}) \ge C\|u - \bar{u}\|_2^2.$$

Proof. The proof is a variant of the one given for Lemma 1 in [9]. Let $u \in \mathcal{B}$. With $v = u - \bar{u}$ we get from (OS)

$$\begin{aligned} f(u) - f(\bar{u}) &= \langle v, g(\bar{u})\rangle + \frac{1}{2}\left(\langle v_I, \nabla^2 f(u) v_I\rangle + \langle v_A, \nabla^2 f(u)(v_A + 2v_I)\rangle\right) + o(\|v\|_2^2) \\ &\geq c_1\|v_A\|_1 + \frac{c_2}{2}\|v_I\|_2^2 - \frac{c_r}{2}\left(\|v_A\|_2^2 + 2\|v_A\|_2\|v_I\|_2\right) + o(\|v\|_2^2) \\ &\geq c_1\frac{\|v_A\|_2^2}{\|v_A\|_\infty} + \frac{c_2}{2}\|v_I\|_2^2 - \frac{c_r}{2}\left(1 + \frac{2c_r}{c_2}\right)\|v_A\|_2^2 - \frac{c_2}{4}\|v_I\|_2^2 + o(\|v\|_2^2) \\ &\geq \left(\frac{c_1}{\|v_A\|_\infty} - \frac{c_r(2c_r + c_2)}{2c_2}\right)\|v_A\|_2^2 + \frac{c_2}{4}\|v_I\|_2^2 + o(\|v\|_2^2), \end{aligned}$$

where we have used $2\alpha\beta \leq c\alpha^2 + \beta^2/c$ with $c = c_2/(2c_r)$. Note that $o(\|v\|_2^2)$ is meant for $\|v\|_\infty \to 0$. We see that the assertion follows for all $u \in \mathcal{B}$ with $\|u - \bar{u}\|_\infty < \rho$ if $\rho > 0$ is small enough. □

9.2. L^2-optimality. We make now additional assumptions on the structure of the second derivative which are met by the class of regularized problems considered in the previous section and are similar to those in [23]:

ASSUMPTION.

(A5) $\nabla^2 f(u) = \beta(u)I + K'(u)$ where $\beta : \mathcal{B} \subset L^2 \longrightarrow L^\infty$ is continuous and K' satisfies (A4) for suitable $2 \leq q < s \leq \infty$.

We have the following variant of Theorem 4 in [9]:

THEOREM 9.3. *Let the formal second-order sufficiency condition* (OS) *with* $t < \infty$ (*i.e. also for* $t = 2$) *and* (A5) *hold. If in addition* $\beta(\bar{u})(x) \geq \beta_0 > 0$ *a.e. on* Ω *then* $\bar{u}$ *is a strict* L^2*-optimizer* (*and hence* L^t*-optimizer,* $t \in [1, \infty]$) *for* (P) *in the following sense: there are* $\rho > 0$ *and* $C > 0$ *such that*

$$u \in \mathcal{B},\ \|u - \bar{u}\|_2 < \rho \Longrightarrow f(u) - f(\bar{u}) \geq C\|u - \bar{u}\|_2^2.$$

Proof. We compute as in the proof of Theorem 9.2

$$\begin{aligned} f(u) - f(\bar{u}) &\geq c_1\|v_A\|_1 + \frac{c_2}{2}\|v_I\|_2^2 + \frac{\beta_0}{2}\|v_A\|_2^2 + \frac{1}{2}\langle v_A, K'(u)(v_A + 2v_I)\rangle + o(\|v\|_2^2) \\ &\geq c_1\|v_A\|_1 + \frac{c_2}{2}\|v_I\|_2^2 + \frac{\beta_0}{2}\|v_A\|_2^2 - \|K'(u)\|_{q,s}\|v_A\|_{s'}\|v\|_q + o(\|v\|_2^2) \end{aligned}$$

where $v = u - \bar{u}$ and $1/s + 1/s' = 1$. Then $1 \leq s' < 2 \leq q < s$ and $1/q + 1/s' \stackrel{\text{def}}{=} 1 + \delta > 1$ because of $s > q$. Now by Lemma 2.2

$$\|v_A\|_{s'} \leq \|v_A\|_1^{1/s'}\|v_A\|_\infty^{1-1/s'} \quad , \quad \|v\|_q \leq \|v\|_2^{2/q}\|v\|_\infty^{1-2/q}.$$

Since $\|v_A\|_\infty \leq \|v\|_\infty \leq \|b - a\|_\infty$, we find $C_1 > 0$ with

$$\|v_A\|_{s'}\|v\|_q \leq C_1\|v_A\|_1^{1/s'}\|v\|_2^{2/q} \leq C_1\left(\frac{1}{p_1'}\|v_A\|_1^{p_1'/s'} + \frac{1}{p_1}\|v\|_2^{2p_1/q}\right)$$

for all $p_1, p_1' \in (1, \infty)$, $1/p_1 + 1/p_1' = 1$, according to Young's inequality. We choose $p_1 = q(1 + \varepsilon)$ and get

$$\frac{1}{p_1'} = 1 - \frac{1}{q(1+\varepsilon)} = \frac{1}{s'} + \frac{1}{q} - \frac{1}{q(1+\varepsilon)} - \delta < \frac{1}{s'(1+\varepsilon')}$$

for $\varepsilon, \varepsilon' > 0$ small enough. Hence, for $\|v\|_2$ small

$$\|K'(u)\|_{q,s}\|v_A\|_{s'}\|v\|_q \leq C_K C_1(\|v_A\|_1^{1+\varepsilon'} + \|v_A\|_2^{2+2\varepsilon}),$$

which completes the proof. □

We shall now study in which cases condition (A3) is implied by the formal second-order sufficiency condition (OS). We will thereby restrict ourselves to problems which satisfy the structural assumptions of §8.

THEOREM 9.4. *Let* (OS), (A2′) *with* $t < \infty$, *and* (A4) *hold. Then the following is true:*

1) *If* $q = 2$ *then* (A3) *is satisfied.*
2) *For* $q > 2$ *there are* $\rho > 0$ *and* $C_{\tilde{H}} > 0$ *such that for all* $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_s < \rho$, *the operator* $\tilde{H}(u)$ *in* (47) *has the properties:*
   i) $\tilde{H}(u) \in \mathcal{L}(L^q, L^q)$ *and* $\|\tilde{H}(u)v\|_q \geq C_{\tilde{H}}\|v\|_q$ *for all* $v \in L^q$.
   ii) *The range of* $\tilde{H}(u) : L^q \longrightarrow L^q$ *is closed in* L^q.

   Hence, if $K' : \mathcal{B} \subset L^s \longrightarrow \mathcal{L}(L^q, L^q)$ *is continuous at* $\bar{u}$ *and if the range of* $\tilde{H}(\bar{u})$ *is dense in* L^q *then* (A3) *is satisfied.*

Proof. Let $0 < \rho \leq \rho_K$ and $u \in \mathcal{B}$, $\|u - \bar{u}\|_s < \rho$. We will adjust ρ in the sequel. By (A4) and the definition of $\tilde{H}(u)$ we have $\tilde{H}(u) \in \mathcal{L}(L^q, L^q)$ and

$$\langle \alpha v, \tilde{H}(u)w\rangle \leq (m_{2,q}^2\|\alpha\|_\infty + m_{q',s}C_{K'})\|v\|_q\|w\|_q \quad \forall\ v, w \in L^q. \tag{54}$$

From $t < \infty$, Lemma 2.1, and Lemma 2.2 we deduce that $u \in \mathcal{B}$, $\|u - \bar{u}\|_s \to 0$ implies $\|u - \bar{u}\|_t \to 0$. Using $\alpha(\tilde{H}(u) - \tilde{H}(\bar{u})) = \chi_{\bar{I}}(\nabla^2 f(u) - \nabla^2 f(\bar{u}))\chi_{\bar{I}}$ we obtain from (A2′) and (52) by a density argument in L^q

$$\lim_{\substack{u\in\mathcal{B}\\ \|u-\bar{u}\|_s\to 0}} \sup_{\substack{w\in L^q\\ \|w\|_2=1}} \langle w, (\tilde{H}(u) - \tilde{H}(\bar{u}))w\rangle = 0. \tag{55}$$

For arbitrary $w \in L^q$ we have with (OS)

$$\begin{aligned}
\langle \alpha w, \tilde{H}(\bar{u})w\rangle &= \langle \alpha w, w\rangle + \langle \alpha w_{\bar{I}}, \alpha^{-1}K'(\bar{u})w_{\bar{I}}\rangle \\
&= \langle \alpha w_{\bar{A}}, w_{\bar{A}}\rangle + \langle \alpha w_{\bar{I}}, w_{\bar{I}}\rangle + \langle w_{\bar{I}}, K'(\bar{u})w_{\bar{I}}\rangle \\
&= \langle \alpha w_{\bar{A}}, w_{\bar{A}}\rangle + \langle w_{\bar{I}}, \nabla^2 f(\bar{u})w_{\bar{I}}\rangle \\
&\geq \alpha_0\|w_{\bar{A}}\|_2^2 + c_2\|w_{\bar{I}}\|_2^2 \geq \min\{\alpha_0, c_2\}\,\|w\|_2^2 \stackrel{\text{def}}{=} C_1\|w\|_2^2.
\end{aligned}$$

Together with (55) this shows that for sufficiently small $\rho > 0$ we have

$$\langle \alpha w, \tilde{H}(u)w\rangle \geq \frac{C_1}{2}\|w\|_2^2 \quad \forall\ w \in L^q. \tag{56}$$

Hence, in the case $q = 2$ the symmetric operator $\alpha\tilde{H}(u) \in \mathcal{L}(L^2, L^2)$ is bounded by (54) and positive by (56). We therefore may apply the Lax-Milgram theorem (which in our symmetric case is an immediate consequence of Riesz's repesentation theorem), yielding that $\tilde{H}(u)$ is continuously invertible in $\mathcal{L}(L^2, L^2)$ with $\|\tilde{H}(u)^{-1}\|_{2,2} \leq 2\|\alpha\|_\infty/C_1$. By Lemma 8.3 this implies (A3) for sufficiently small $\rho_H > 0$.

Next we assume $q > 2$ and establish the second part of 2i). Let $w \in L^q$ be arbitrary. For $c > 0$ which will be adjusted later we consider two cases: If $\|w\|_2 \geq c\|w\|_q$ then by (56)

$$\frac{cC_1}{2}\|w\|_2\|w\|_q \leq \frac{C_1}{2}\|w\|_2^2 \leq \langle \alpha w, \tilde{H}(u)w\rangle \leq \|\alpha\|_\infty m_{q',2}\|w\|_2\|\tilde{H}(u)w\|_q.$$

Whence,

$$\|\tilde{H}(u)w\|_q \geq \frac{cC_1}{2\|\alpha\|_\infty m_{q',2}}\|w\|_q.$$

In the case $\|w\|_2 < c\|w\|_q$ we compute

$$\|\tilde{H}(u)w\|_q \geq \|w\|_q - \frac{1}{\alpha_0}\|\chi_{\bar{I}}K'(u)w_{\bar{I}}\|_q \geq \|w\|_q - \frac{1}{\alpha_0}\|K'(u)w_{\bar{I}}\|_q.$$

Applying Lemma 2.2 we obtain with (A4) and suitable $\theta \in (0,1)$

$$\|K'(u)w_{\bar{I}}\|_q \leq \|K'(u)w_{\bar{I}}\|_2^\theta \|K'(u)w_{\bar{I}}\|_s^{1-\theta} \leq \|K'(u)w_{\bar{I}}\|_2^\theta (C_{K'}\|w\|_q)^{1-\theta}.$$

Obviously, (54) also holds with the left side replaced by $|\langle v, \nabla^2 f(u)w\rangle|$. Therefore, (51) implies by a density argument together with (A2′) and (A4) that

$$\|K'(u)v\|_2 = \|\nabla^2 f(u)v - \alpha v\|_2 \leq (c_r + \|\alpha\|_\infty)\|v\|_2 \stackrel{\text{def}}{=} C_r\|v\|_2 \quad \forall\ v \in L^q.$$

This gives

$$\|K'(u)w_{\bar{I}}\|_q \leq (C_r\|w\|_2)^\theta (C_{K'}\|w\|_q)^{1-\theta} \leq (C_r c)^\theta C_{K'}^{1-\theta}\|w\|_q,$$

and choosing $c > 0$ small enough we achieve

$$\|\tilde{H}(u)w\|_q \geq \frac{1}{2}\|w\|_q$$

as long as $\|w\|_2 < c\|w\|_q$. Since $c > 0$ can be adjusted independently of w, i) is shown.

To prove ii), let $(w_k) \subset L^q$ be arbitrary. Then

$$\begin{aligned}
&\tilde{H}(u)w_k \stackrel{L^q}{\to} v \in L^q \quad (k \to \infty)\\
&\stackrel{\text{i)}}{\Longrightarrow} \quad \|w_k - w_l\|_q \leq C_{\tilde{H}}^{-1}\|\tilde{H}(u)w_k - \tilde{H}(u)w_l\|_q \to 0 \quad (k,l \to \infty)\\
&\Longrightarrow \quad w_k \stackrel{L^q}{\to} w \in L^q \quad (k \to \infty) \stackrel{\text{i)}}{\Longrightarrow} \tilde{H}(u)w_k \stackrel{L^q}{\to} \tilde{H}(u)w \quad (k \to \infty).
\end{aligned}$$

If the range of $\tilde{H}(\bar{u}) : L^q \longrightarrow L^q$ is dense then $\tilde{H}(\bar{u})$ is injective by i) and surjective by ii). Thus, it has a continuous inverse by the open mapping theorem and i) shows $\|\tilde{H}(\bar{u})^{-1}\|_{q,q} \leq C_{\tilde{H}}^{-1}$. If in addition $K' : \mathcal{B} \subset L^s \longrightarrow \mathcal{L}(L^q, L^q)$ is continuous at $\bar{u}$ then for $\|u - \bar{u}\|_s$ small enough $\tilde{H}(u)^{-1} \in \mathcal{L}(L^q, L^q)$ exists and (A3) is satisfied for sufficiently small $\rho_H > 0$. □

The previous result shows that – at least in the case $q = 2$ – the application of Algorithm 5.17 to the class of problems considered in §8 leads to superlinear convergence in a neighborhood of a point $\bar{u}$ satisfying (OS). This is especially important since formal sufficiency conditions of type (OS) are the usual starting point for proving that a rapidly convergent local method meets the trial step requirements of a globally convergent algorithm in a neighborhood of a local optimizer. Hence, it is important that the local convergence theory can be established under a sufficiency condition that is as weak as possible.

10. Trust region globalization. The aim of this section is to show that near a local optimizer satisfying (OS) Algorithm 5.17 produces admissible trial steps for the globally convergent affine-scaling interior-point trust-region algorithm that we proposed and analyzed in [24]. The trust-region globalization extends ideas of Coleman and Li in [7] and uses the fact that $u^n \in L^q$ solves (6) for given $u^c \in \mathcal{B}^\circ$ if and only if u^n is a stationary point of the quadratic function $\psi[u^c] : L^q \longrightarrow \mathbb{R}$,

$$\psi[u^c](u) \stackrel{\text{def}}{=} \langle u - u^c, g(u^c)\rangle + \frac{1}{2}\langle u - u^c, M(u^c)(u - u^c)\rangle, \tag{57}$$

$$\text{where } M(u) \stackrel{\text{def}}{=} d(u)^{-1}G(u) = \chi_{\{d(u)<c\}}|g(u)|d(u)^{-1}I + \nabla^2 f(u).$$

Here and in the following we use the standard notation s for the trial steps although it collides with the norm index occurring in (A2). There is no danger of ambiguity. As shown in [24], a globally convergent algorithm can be obtained as follows: Denote by $u_k \in \mathcal{B}^\circ$ the current iterate. We compute a trial step s_k as approximate solution to the trust-region subproblem

$$\text{minimize } \psi[u_k](u_k + s) \;\text{ subject to }\; u_k + s \in \mathcal{B},\ \|s\|_q \le \Delta_k. \tag{58}$$

This trial step is required to satisfy the

Fraction of Cauchy decrease condition:

(D) $u_k + s_k \in \mathcal{B}^\circ$, $\|s_k\|_q \le \beta_0\Delta_k$, and $\psi[u_k](u_k + s_k) \le \beta\psi[u_k]^c$, where

$$\psi[u_k]^c \stackrel{\text{def}}{=} \beta \min\left\{\psi[u_k](u_k + s) \;:\; s = -\tau d_k^\vartheta g_k,\ \tau \ge 0,\ u_k + s \in \mathcal{B},\ \|s\|_q \le \Delta_k\right\}$$

with fixed constants $\beta_0 > 0$, $0 < \beta < 1$, $\vartheta \ge 1$. The step s_k is accepted, i.e. $u_{k+1} = u_k + s_k$, if $r_k > \eta_1$, where $0 < \eta_1 < 1$ is fixed and the decrease ratio $r_k = r(u_k, s_k)$ is given by

$$r(u_k, s_k) \stackrel{\text{def}}{=} \frac{f(u_k + s_k) - f(u_k)}{\langle s_k, g(u_k)\rangle + \langle s_k, \nabla^2 f(u_k)s_k\rangle/2}.$$

Otherwise, i.e. if $r_k \le \eta_1$, the step is rejected: $u_{k+1} = u_k$. For our presentation it is convenient to use an update rule for the trust-region radius Δ_k that is slightly different from the one given in [24]. However, it is not hard to verify that all the convergence results stated therein remain valid. In our update rule we fix $0 < \eta_1 < \eta_2 < \eta_3 < 1$, $0 < \beta_0\gamma_0 \le \gamma_1 < 1 < \gamma_2 \le \gamma_3$, $\Delta_{\min} > 0$, and choose

$$\Delta^+ \in \begin{cases} [\gamma_0\|s_k\|_q, \gamma_1\Delta_k] & \text{if } r_k \le \eta_1, \\ [\gamma_1\Delta_k, \Delta_k] & \text{if } \eta_1 < r_k < \eta_2, \\ [\Delta_k, \gamma_2\Delta_k] & \text{if } \eta_2 < r_k < \eta_3, \\ [\gamma_2\Delta_k, \gamma_3\Delta_k] & \text{else.} \end{cases}, \quad \Delta_{k+1} := \begin{cases} \Delta^+ & \text{if } r_k \le \eta_1, \\ \max\{\Delta_{\min}, \Delta^+\} & \text{else.} \end{cases}$$

For a detailed formulation of the algorithm and its convergence properties we refer to [24]. The theory developed therein (adapted to our update rule) states that under assumption (A1) each accumulation point of the sequence (u_k) satisfies the first-order necessary optimality conditions (O1),(O2), and, moreover, the second-order necessary condition [24, Thm. 3.3, (O3)] if (D) is replaced by a fraction of optimal decrease condition.

Having in mind that trust-region methods for unconstrained problems inherit their local convergence behavior from Newton's method, it is natural to try to accelerate the

above trust-region method by means of Algorithm 5.17. We combine both methods as follows:

ALGORITHM 10.1 (TRUST-REGION INTERIOR-POINT NEWTON METHOD).

1. Choose $\Delta_k \geq \Delta_{\min}$, $u_0 \in \mathcal{B}^\circ$.
2. For $k = 0, 1, 2, \ldots$
   - 2.1 If $d(u_k)g(u_k) = 0$, STOP.
   - 2.2 Perform a smoothing step (if necessary): $u_k^s = S_k^\circ(u_k)$.
   - 2.3 Compute a trial step by Algorithm 5.17: $s_k = \min\{1, \Delta_k/\|s_k^N\|_q\}s_k^N$, where
     $$s_k^N = u_{k+1}^N - u_k^s\ ,\quad u_{k+1}^N = P[u_k^s](u_{k+1}^n)\ ,\quad u_{k+1}^n = u_k^s - G(u_k^s)^{-1}d(u_k^s)g(u_k^s).$$
     If s_k satisfies (D) for $\psi[u_k^s]$ then goto step 2.5.
   - 2.4 Compute a trial step that satisfies (D) for $\psi[u_k^s]$, e.g. by a descent method that starts with a line search along $-d(u_k^s)^\vartheta g(u_k^s)$.
   - 2.5 Compute the decrease ratio $r_k = r(u_k^s, s_k)$ and the new trust-region radius Δ_{k+1}. If $r_k > \eta_1$ then set $u_{k+1} = u_k^s + s_k$. Otherwise set $u_{k+1} = u_k^s$ and go to step 2.2.

Now suppose that one of the accumulation points $\bar{u} \in \mathcal{B}$ of (u_k) satisfies the second-order sufficiency condition (OS). The question is: Does this globally convergent method eventually turn into Algorithm 5.17 and thus inherit its superlinear convergence?

It is beyond the scope of this paper to answer this question in full generality, since this would require to analyze the effect of the smoothing step on the global convergence behavior of the trust-region algorithm. We try to find a reasonable compromise by developing results that are rigorously applicable whenever the smoothing steps do not affect the global convergence. This is certainly the case if the smoothing steps decrease the objective function f; see Remark 8.2 in this context. Moreover, we require

ASSUMPTION.

(A6) $\bar{u} \in \mathcal{B}$ is an accumulation point of (u_k) at which (OS) holds. Moreover, condition (A2) is satisfied with $r = s = \infty$.

As a first result we show that the quadratic model $\psi[u_k^s]$ has a unique minimizer if $\|u_k^s - \bar{u}\|_\infty$ is sufficiently small. To show this, we first prove

LEMMA 10.2. *Let* (A6) *hold. Then there are* $\rho > 0$, $C_M > 0$ *such that*

$$\langle v, M(u)v\rangle \geq C_M\|v\|_2^2 \quad \forall\ v \in L^q$$

for all $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_\infty < \rho$.

Proof. We know that (O1),(O2) hold at $\bar{u}$. Let I and A be defined as in (OS). Since $|g(\bar{u})| \geq c_1$ a.e. on A, we get by (O2) that $d(\bar{u}) = 0$ a.e. on A and hence for sufficiently small $\rho > 0$ and all $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_\infty < \rho$

$$|g(u)| \geq c_1/2 \quad \text{and} \quad c > d(u) \leq L_d\rho \quad \text{a.e. on } A$$

by (A2) and Lemma 5.5, respectively. (A2) yields with a density argument in L^q that (51)–(53) also hold for L^∞ replaced by L^q and thus, possibly after reducing ρ, we have

$$\langle v_I, \nabla^2 f(u)v_I\rangle \geq \frac{c_2}{2}\|v_I\|_2^2 \quad \forall\ v \in L^q$$

as long as $u \in \mathcal{B}^\circ$, $\|u - \bar{u}\|_\infty < \rho$. Hence, for all $v \in L^q$

$$\begin{aligned}\langle v, M(u)v\rangle &= \left\langle v_A, \frac{|g(u)|}{d(u)} v_A \right\rangle + \left\langle v_I, \frac{\chi_{\{d(u)<c\}}|g(u)|}{d(u)} v_I \right\rangle \\ &\quad + \langle v_I, \nabla^2 f(u) v_I\rangle + \langle v_A, \nabla^2 f(u)(2v_I + v_A)\rangle \\ &\geq \frac{c_1}{2L_d\rho}\|v_A\|_2^2 + \frac{c_2}{2}\|v_I\|_2^2 - c_r\|v_A\|_2^2 - 2c_r\|v_A\|_2\|v_I\|_2.\end{aligned}$$

With the standard estimate

$$2c_r\|v_A\|_2\|v_I\|_2 \leq \frac{c_2}{4}\|v_I\|_2^2 + \frac{4c_r^2}{c_2}\|v_A\|_2^2$$

we arrive at

$$\langle v, M(u)v\rangle \geq \left(\frac{c_1}{2L_d\rho} - c_r - \frac{4c_r^2}{c_2}\right)\|v_A\|_2^2 + \frac{c_2}{4}\|v_I\|_2^2.$$

Now for $\rho > 0$ sufficiently small the assertion follows. □

THEOREM 10.3. *Let* (A3) *and* (A6) *hold and* (u_k), (u_k^s), (u_k^n), (u_k^N) *be generated by Algorithm* 10.1. *If* ρ *is sufficiently small and* $\|u_k^s - \bar{u}\|_\infty < \rho$ *then* $u_{k+1}^n \in L^q$ *is a global minimizer of* $\psi[u_k^s]$ *and*

$$\psi[u_k^s](u_{k+1}^n) = -\frac{1}{2}\langle u_{k+1}^n - u_k^s, M(u_k^s)(u_{k+1}^n - u_k^s)\rangle \leq -\frac{C_M}{2}(\|s^e\|_2^2 + \|s^i\|_2^2) \tag{59}$$

$$\psi[u_k^s](P(u_{k+1}^n)) \leq \psi[u_k^s](u_{k+1}^n) - \frac{C_M}{2}\|s^e\|_2^2 + (c_r + \nu^{-1}L_g\rho)\|s^e\|_2\|s^i\|_2 \tag{60}$$

with $s^e = u_{k+1}^n - P(u_{k+1}^n)$, $s^i = P(u_{k+1}^n) - u_k^s$. *Moreover,*

$$\psi[u_k^s](u_{k+1}^N) \leq \max\{\xi, 1 - \|s^i\|_q\}\psi[u_k^s](P(u_{k+1}^n)), \tag{61}$$

and hence

$$\frac{\psi[u_k^s](u_{k+1}^N)}{\psi[u_k^s](u_{k+1}^n)} \geq \max\{\xi, 1 - \|s^i\|_q\}\left(1 - O\left(\frac{\|s^e\|_2}{\|s^i\|_2}\right)\right). \tag{62}$$

Proof. Setting $s = u_{k+1}^n - u_k^s$ we have $s = s^e + s^i$ and $s^e s^i \geq 0$ a.e. on Ω. We use the abbreviations $d_k = d(u_k^s)$, $g_k = g(u_k^s)$, $M_k = M(u_k^s)$. According to step 3) in Algorithm 5.17, we have $M_k s = -g_k$. Hence, u_{k+1}^n is a stationary point of $\psi[u_k^s]$ and, therefore, its global minimum by Lemma 10.2. Moreover,

$$\psi[u_k^s](u_{k+1}^n) = -\frac{1}{2}\langle s, M_k s\rangle \leq -\frac{C_M}{2}\|s^e + s^i\|_2^2 \leq -\frac{C_M}{2}(\|s^e\|_2^2 + \|s^i\|_2^2).$$

To prove the second inequality we observe that for all $x \in \Omega$ with $s(x)g_k(x) < 0$ and $d_k(x) < c(x)$ we have

$$s^e(x) = 0 \quad \text{or} \quad |s^i(x)| \geq d_k(x).$$

In fact, if $s^e(x) \neq 0$ then either $s^i(x) = b(x) - u_k^s(x) > 0$ or $s^i(x) = a(x) - u_k^s(x) < 0$. In the first case we have $g_k(x) < 0$ and thus $d_k(x) = b(x) - u_k^s(x) = s^i(x)$ or $d_k(x) < c(x) < b(x) - u_k^s(x) = s^i(x)$. The second case enforces $g_k(x) > 0$ which

implies $d_k(x) = u_k^s(x) - a(x) = |s^i(x)|$ or $d_k(x) < c(x) < u_k^s(x) - a(x) = |s^i(x)|$. Hence, we get with $N = \{x \in \Omega : s(x)g_k(x) < 0\}$ and $J = \{x \in \Omega : d_k(x) < c(x)\}$

$$\left\langle s_N^e, \frac{|g_k|_J}{d_k} s^i \right\rangle \geq -\langle s_{N\cap J}^e, g_k \rangle$$

Using this, we obtain

$$\begin{aligned}
\psi[u_k^s](u_{k+1}^n) &= \langle s^e + s^i, g_k \rangle + \frac{1}{2}\left\langle s^e + s^i, M_k(s^e + s^i)\right\rangle \\
&= \psi[u_k^s](P(u_{k+1}^n)) + \langle s_{N^c}^e, g_k\rangle + \langle s_{N\setminus J}^e, g_k\rangle + \langle s_{N\cap J}^e, g_k\rangle + \left\langle s_N^e, \frac{|g_k|_J}{d_k} s^i\right\rangle \\
&\quad + \left\langle s_{N^c}^e, \frac{|g_k|_J}{d_k} s^i\right\rangle + \frac{1}{2}\langle s^e, M_k s^e\rangle + \langle s^e, \nabla^2 f(u_k^s) s^i\rangle \\
&\geq \psi[u_k^s](P(u_{k+1}^n)) + \langle s_{N\setminus J}^e, g_k\rangle + \frac{1}{2}\langle s^e, M_k s^e\rangle + \langle s^e, \nabla^2 f(u_k^s) s^i\rangle
\end{aligned}$$

We still need an estimate for $\langle s_{N\setminus J}^e, g_k\rangle$. To this end, we use the inclusion $N \setminus J \subset J^c$. Since $|d_k(x) - d(\bar u)(x)| \leq L_d\rho$ and $d_k(x) \geq c(x) \geq \nu$ on J^c we have $d(\bar u)(x) > 0$ a.e. on J^c for $\rho > 0$ small enough. (O2) yields $g(\bar u)(x) = 0$ on J^c and thus

$$|\langle s_{N\setminus J}^e, g_k\rangle| \leq \|s_e\|_{1,J^c}\|g_k - g(\bar u)\|_\infty \leq L_g\rho\mu(J^c)^{1/2}\|s_e\|_2.$$

Furthermore, $|g_k(x)| \leq L_g\rho < \nu$ on J_c for small ρ, which, since $d_k(x) \geq \nu$, requires

$$\nu \leq d_k(x) \leq \min\{b(x) - u_k^s(x), u_k^s(x) - a(x)\} \leq |s^i(x)|.$$

Hence, by Lemma 2.4

$$\mu(J^c) \leq \mu\{x \in \Omega : |s^i(x)| \geq \nu\} \leq \nu^{-2}\|s^i\|_2^2.$$

We conclude $|\langle s_{N\setminus J}^e, g_k\rangle| \leq \nu^{-1}L_g\rho\|s_e\|_2\|s^i\|_2$. Therefore, (60) holds. Now

$$u_{k+1}^N - u_k^s = P[u_k^s](u_{k+1}^n) - u_k^s = \max\{\xi, 1 - \|s^i\|_q\}(P(u_{k+1}^n) - u_k^s) \stackrel{\text{def}}{=} \tau(P(u_{k+1}^n) - u_k^s).$$

This implies (61), for

$$\begin{aligned}
\psi[u_k^s](P[u_k^s](u_{k+1}^n)) &= \tau\langle s^i, g_k\rangle + \frac{\tau^2}{2}\langle s^i, M_k s^i\rangle \\
&\leq \tau\left(\langle s^i, g_k\rangle + \frac{1}{2}\langle s^i, M_k s^i\rangle\right) = \tau\psi[u_k^s](P(u_{k+1}^n))
\end{aligned}$$

where the inequality follows from $0 \leq \tau < 1$ and $\langle s^i, M_k s^i\rangle \geq 0$, see Lemma 10.2. Now (59)–(61) and a straightforward calculation give (62). □

Let the assumptions of Theorem 5.12 hold. Using (30) we have for $\|u_k^s - \bar u\|_\infty$ small enough

$$\begin{aligned}
\|u_{k+1}^n - u_k^s\|_q &\leq \|u_{k+1}^n - \bar u\|_q + \|u_k^s - \bar u\|_q \leq (C\Phi_{\bar p}(\|u_k^s - \bar u\|_\infty) + m_{q,\infty})\|u_k^s - \bar u\|_\infty \\
&\leq C_\Delta\|u_k^s - \bar u\|_\infty
\end{aligned}$$

with C_Δ appropriately chosen. Now

$$\|u_{k+1}^N - u_k^s\|_q = \|P[u_k^s](u_{k+1}^n) - u_k^s\|_q \leq \|P(u_{k+1}^n) - u_k^s\|_q \leq C_\Delta\|u_k^s - \bar u\|_\infty. \tag{63}$$

Hence, for $\|u_k^s - \bar{u}\|_\infty$ small enough we have (cf. (D))

$$u_{k+1}^N \in \mathcal{B}^\circ, \quad \|u_{k+1}^N - u_k^s\|_q \le \beta_0 \Delta_{\min}.$$

Using this in Theorem 10.3 we can show that if $\|s^e\|_2/\|s^i\|_2$ eventually remains small enough then Algorithm 10.1 turns into the superlinearly convergent Algorithm 5.17. In particular, this happens if no smoothing steps are required:

THEOREM 10.4. *Let the assumptions of Theorem* 10.3 *as well as* (C) *and* (S) *hold. Then there are $\rho > 0$, $\varepsilon > 0$ such that if step $k-1$ was accepted and $\|u_k - \bar{u}\|_q < \rho$, then $s_k = s_k^N$ and step k is accepted whenever*

$$\frac{\|u_{k+1}^n - P(u_{k+1}^n)\|_2}{\|P(u_{k+1}^n) - u_k^s\|_2} < \varepsilon \tag{64}$$

holds. If there is $C_1 > 0$ with $\|u_k^s - \bar{u}\|_\infty \le C_1 \|u_k^s - \bar{u}\|_q$ then (64) *is automatically satisfied for $\|u_k - \bar{u}\|_q$ small enough.*

Proof. Assume that step $k-1$ was accepted and $\|u_k - \bar{u}\|_q < \rho$ with $\rho > 0$ sufficiently small. We use s^e and s^i as defined in Theorem 10.3. Since $\|u_k^s - \bar{u}\|_\infty \le C_S \|u_k - \bar{u}\|_q$ by (S), we get with (63)

$$\|u_{k+1}^N - u_k^s\|_q \le \|s^i\|_q \le C_\Delta C_S \|u_k - \bar{u}\|_q \le C_\Delta C_S \rho.$$

Hence, $u_{k+1}^N = u_k^s + s_{k+1}^N \in \mathcal{B}^\circ$ with $\|s_k^N\|_q \le \Delta_{\min} \le \Delta_k$ for ρ small enough. Choose $0 < \tilde{\varepsilon} < 1$ such that $(1-\tilde{\varepsilon})^2 > \beta$ with β given in (D). Possibly after reducing ρ we achieve $\|s^i\|_q \le \tilde{\varepsilon}$. For $0 < \varepsilon < 1$ sufficiently small we have by (62) and (64)

$$\frac{\psi[u_k^s](u_{k+1})}{\psi[u_k^s](u_{k+1}^n)} \ge (1-\tilde{\varepsilon})^2 > \beta.$$

Since u_{k+1}^n is the global minimizer of $\psi[u_k^s]$ by Theorem 10.3, $s_k = s_k^N$ obviously satisfies (D) for $\psi[u_k^s]$.

Now assume $\|u_k^s - \bar{u}\|_\infty \le C_1 \|u_k^s - \bar{u}\|_q$. Then Lemma 2.2 yields with $\theta = 2/q$

$$\|u_k^s - \bar{u}\|_q \le \|u_k^s - \bar{u}\|_2^\theta \|u_k^s - \bar{u}\|_\infty^{1-\theta} \le C_1^{1-\theta} \|u_k^s - \bar{u}\|_2^\theta \|u_k^s - \bar{u}\|_q^{1-\theta}$$

and thus

$$\|u_k^s - \bar{u}\|_2 \ge C_1^{1-\frac{1}{\theta}} \|u_k^s - \bar{u}\|_q \stackrel{\text{def}}{=} C_3 \|u_k^s - \bar{u}\|_q.$$

To show (64) for ρ small enough we use $\|u_{k+1}^n - P(u_{k+1}^n)\|_2 \le \|u_{k+1}^n - \bar{u}\|_2$ and get

$$\begin{aligned} \|P(u_{k+1}^n) - u_k^s\|_2 &\ge \|u_k^s - \bar{u}\|_2 - \|P(u_{k+1}^n) - u_{k+1}^n\|_2 - \|u_{k+1}^n - \bar{u}\|_2 \\ &\ge \|u_k^s - \bar{u}\|_2 - 2\|u_{k+1}^n - \bar{u}\|_2 \ge \frac{C_3}{C_1} \|u_k^s - \bar{u}\|_\infty - 2\|u_{k+1}^n - \bar{u}\|_2. \end{aligned}$$

Moreover, Theorem 5.12 yields

$$\|u_{k+1}^n - \bar{u}\|_2 \le m_{2,q} \|u_{k+1}^n - \bar{u}\|_q \le m_{2,q} C \Phi_{\bar{p}}(\|u_k^s - \bar{u}\|_\infty) \|u_k^s - \bar{u}\|_\infty.$$

Hence, for $\|u_k^s - \bar{u}\|_\infty$ sufficiently small we get

$$\frac{\|u_{k+1}^n - P(u_{k+1}^n)\|_2}{\|P(u_{k+1}^n) - u_k^s\|_2} \le \left(\frac{C_3 \|u_k^s - \bar{u}\|_\infty}{C_1 \|u_{k+1}^n - \bar{u}\|_2} - 2 \right)^{-1} \le \left(\frac{C_3}{C_1 m_{2,q} C \Phi_{\bar{p}}(\|u_k^s - \bar{u}\|_\infty)} - 2 \right)^{-1}$$

and the last term is $< \varepsilon$ for small ρ, since $\Phi_{\bar{p}}(\|u_k^s - \bar{u}\|_\infty) \le \Phi_{\bar{p}}(C_s \rho)$ tends to zero as $\rho \to 0$. ∎

11. Application to a control problem. In this section we present numerical results for the application of Algorithm 5.17 to a boundary control problem governed by a nonlinear heat equation which is a simplified model for the heating of a probe in a kiln. Let $Q \stackrel{\text{def}}{=} (0,1)$ denote the spatial domain with $x = 0$ at the boundary and $x = 1$ at the inside of the probe. The temperature $y(x,t)$, $(x,t) \in Q \times (0,T) \stackrel{\text{def}}{=} Q_T$ of the probe satisfies the nonlinear heat equation

$$\begin{aligned} \tau(y)y_t - (\kappa(y)y_x)_x &= h && \text{on } Q_T, \\ \kappa(y(0,t))y_x(0,t) &= \zeta(y(0,t) - u(t)) && t \in (0,T), \\ \kappa(y(1,t))y_x(1,t) &= 0 && t \in (0,T), \\ y(x,0) &= y_0(x) && x \in Q. \end{aligned} \tag{65}$$

where $y_0 : Q \longrightarrow \mathbb{R}$ is the initial temperature, $\tau, \kappa : \mathbb{R} \longrightarrow \mathbb{R}$ denote the specific heat capacity and the heat conduction, respectively, $h : Q_T \longrightarrow \mathbb{R}$ is a source term, $\zeta \in \mathbb{R}$ a given scalar and $u : (0,T) \longrightarrow \mathbb{R}$ the control. For consistency with our notations let $\Omega \stackrel{\text{def}}{=} (0,T)$.

The control u shall be determined in such a way that the temperature $y(1,t)$ inside the probe follows a given temperature profile $y_d(t)$. Since it is well known that this nonlinear inverse heat conduction problem is ill posed, we add a regularization in the control space and choose as objective function

$$J(y,u) = \frac{1}{2}\int_0^T \Big((y(1,t) - y_d(t))^2 + \alpha\, u(t)^2 \Big)\, dt$$

with $y_d \in L^\infty((0,T))$. The problem was considered in [5]. Let $V = H^1(Q)$, V' its dual, $H = L^2(Q)$ and $W(0,T) \stackrel{\text{def}}{=} \{y \in L^2(0,T;V) : y_t \in L^2(0,T;V')\}$, equipped with the norm $\|y\|_{W(0,T)} \stackrel{\text{def}}{=} \|y\|_{L^2(0,T;V)} + \|y_t\|_{L^2(0,T;V')}$. It is well known that $W(0,T)$ is a Hilbert space and that the embedding $W(0,T) \hookrightarrow C(0,T;H)$ is continuous. Under the assumption that $\kappa, \tau \in C(\mathbb{R})$ with

$$0 < \kappa_1 \le \kappa(s) \le \kappa_2, \quad 0 < \tau_1 \le \tau(s) \le \tau_2 \quad \forall\, s \in \mathbb{R}$$

it is shown in [5] that for all $h \in L^2(0,T;H)$, $y_0 \in H$, and $u \in L^2((0,T))$ there exists a solution $y \in W(0,T)$ of the state equation (65) which satisfies the stability estimate

$$\|y\|_{W(0,T)} \le C \left(\|h\|_{L^2(0,T;H)} + \|u\|_2 + \|y_0\|_H \right). \tag{66}$$

Uniqueness is proven under the additional assumption $y_x \in L^\infty(0,T;L^r(Q))$, $r > 2$, and $\kappa, \tau \in C^1(\mathbb{R})$. Furthermore, it was shown that for $\alpha > 0$ there exists an optimal solution $\bar{u} \in L^2(\Omega)$ of the control problem

$$\text{minimize } J(y,u) \quad \text{subject to} \quad y \in W(0,T),\ u \in L^2(\Omega) \text{ satisfy (65).} \tag{67}$$

With the lower and upper bounds $a, b \in L^\infty(\Omega)$, $b - a \ge \nu > 0$, we introduce the additional box constraints

$$u \in \mathcal{B} \stackrel{\text{def}}{=} \{a \le u \le b\}. \tag{68}$$

Since $\mathcal{B}$ is a closed bounded convex subset of $L^2(\Omega)$, exactly the same arguments as in [5] can be used to prove the existence of an optimal control $\bar{u} \in \mathcal{B}$ for $\alpha \ge 0$. Assuming

that for $u \in \mathcal{B}$ the solution $y = y(u)$ to (65) is unique, we can define the objective function $f(u) \stackrel{\text{def}}{=} J(y(u), u)$ for which (67), (68) is equivalent to (P).

For the rest of this paragraph assume that κ and τ are constant. Then the above existence and stability result is well known and it implies that the operator $u \in L^2(\Omega) \longmapsto y(1,.) \in L^q(\Omega)$ is completely continuous for $2 \le q < 4$: We use the symbol '$\hookrightarrow$' for continuous and '$\hookrightarrow\hookrightarrow$' for compact embeddings. (66) shows the continuity of $u \in L^2(\Omega) \longmapsto y \in W(0,T)$. To complete the argument we show that $y \in W(0,T) \longmapsto y(1,.) \in L^q(\Omega)$ is compact for $1 \le q < 4$. Let $1/2 < \theta < \Theta < 1$. Since $V \hookrightarrow\hookrightarrow H^\Theta(Q) \hookrightarrow V'$, we have by a Lions lemma that $W(0,T) \hookrightarrow\hookrightarrow L^2(0,T;H^\Theta(Q))$ (see [19, Thm. 5.1]). Moreover, from the interpolation result $H^\theta(Q) = [H, H^\Theta(Q)]_{\theta/\Theta}$ it can be deduced that

$$\|\cdot\|_{L^{2\Theta/\theta}(0,T;H^\theta(Q))} \le C \|\cdot\|^{1-\theta/\Theta}_{L^\infty(0,T;H)} \|\cdot\|^{\theta/\Theta}_{L^2(0,T;H^\Theta(Q)}.$$

Hence, $W(0,T) \hookrightarrow L^\infty(0,T;H)$ and $W(0,T) \hookrightarrow\hookrightarrow L^2(0,T;H^\Theta(Q))$ yield the compact embedding $W(0,T) \hookrightarrow\hookrightarrow L^{2\Theta/\theta}(0,T;H^\theta(Q))$. Finally, since $H^\theta(Q) \hookrightarrow C([0,1])$, we conclude that $y \in W(0,T) \longmapsto y(1,.) \in L^{2\Theta/\theta}(\Omega)$ is compact. Now (66) shows the complete continuity of $u \in L^2(\Omega) \longmapsto y(1,.) \in L^q(\Omega)$ for $1 \le q < 4$. Hence, $u \in L^2(\Omega) \to J(y(u), u)$ is Fréchet differentiable and (67) is ill posed for $\alpha = 0$.

By standard results (see [20]) the gradient representation g of $u \longmapsto J(y(u), u)$ w.r.t. the inner product on $L^2(\Omega)$ is given by $g(u) = \alpha u + K(u)$, where $K(u) = p(1,.)$ and the *adjoint state* p satisfies

$$\begin{array}{rll} p_t + p_{xx} = 0 & & \text{on } Q_T, \\ p_x(0,t) = 0 & & t \in (0,T), \\ p_x(1,t) = \alpha(y(1,t) - y_d(t) - p(1,t)) & & t \in (0,T), \\ p(x,T) = 0 & & x \in Q \end{array} \tag{69}$$

in the weak sense. Using Green's function, p is given by an integral equation of Volterra type with weakly singular kernel from which one can deduce that (69) defines a completely continuous affine linear mapping $y(1,.) \in L^q(\Omega) \longmapsto C(Q_T)$ for all $q > 2$ (see e.g. [22]). Combining this with the previous considerations we obtain the complete continuity of the affine linear mapping $u \in L^2(\Omega) \longmapsto K(u) = p(1,.) \in C(\Omega)$. Obviously, the Fréchet-derivative K' of $K : L^2(\Omega) \longrightarrow C(\Omega)$ exists and is given by the compact linear operator $K'(u) : v \in L^2(\Omega) \longmapsto K(v) - K(0) \in C(\Omega)$. We conclude that the assumptions (A1), (A2$'$) and (A4) are satisfied for $q = 2$, $s = r = \infty$ and the results of §8 can be applied.

Since the same regularity properties can also be shown for the state equation in the case $h \equiv 0$, $y_0 \in C(Q)$, we get as a byproduct that under these assumptions on h and y_0 the mapping $u \in L^q(\Omega) \longmapsto y \in C(Q_T)$ is completely continuous for $q > 2$.

While similar results can be shown for nonlinear boundary conditions (cf. [22], [15], [17]), a differentiability result for the nonlinear problem (67) seems not to be available. Since (67) is of importance in applications, e.g. the sterilization of canned food, we nevertheless present numerical results for the nonlinear problems and content ourselves with the complete justification of our assumptions for the case of constant κ and τ.

11.1. Discretization. As in [12], [18] we use the discretization of (67) proposed in [5]. For the space discretization we approximate V in the variational formulation

of (65) by the space $V_{\Delta x}$ of continuous functions that are piecewise linear on the intervals $[i\Delta x, (i+1)\Delta x]$, $\Delta x \stackrel{\text{def}}{=} 1/N_x$, $i = 0, \ldots, N_x - 1$. Since the time differentiation in the variational form of (65) is linear with respect to the transformed state $\phi(y) \stackrel{\text{def}}{=} \int_0^y \tau(\xi)\, d\xi$, a discontinuous Galerkin method w.r.t. ϕ is used where $L^2(0,T;V)$ is approximated by the space Y_Δ of $V_{\Delta x}$-valued functions that are piecewise constant on $(k\Delta t, (k+1)\Delta t]$, $\Delta t \stackrel{\text{def}}{=} 1/N_t$, $k = 0, \ldots, N_t - 1$ (the same discretization is obtained by applying backward Euler). This leads in a natural way to the approximation of h and y_0 by their L^2-projection onto Y_Δ and $V_{\Delta x}$, respectively. The discrete control space $U_{\Delta t}$ consists of piecewise constant functions on the same partition of $(0,T]$ and y_d is approximated by its L^2 projection onto the same space. For details we refer to [5], [12], [18].

It was shown in [5] that the resulting implicit scheme admits a unique solution for $\Delta t/\Delta x^2 \le \lambda < (\tau_2/\kappa_1 - \tau_1/\kappa_2)^{-1}/6$ that converges to a solution of (65) as $\Delta t, \Delta x$ tend to zero.

11.2. Numerical Tests. For the application of Algorithm 5.17 we use Example 1 of [18], see also [12] : $T = 0.5$, $\zeta = 1$ and

$$\tau(y) = 4 + y,\ \kappa(y) = 4 - y,\ y_d(t) = 2 - e^{-t},\ y_0(x) = 2 + \cos \pi x$$
$$h(x,t) = (-6 + 2\pi^2)\, e^{-t} \cos \pi x + \pi^2\, e^{-2t} - (1 + 2\pi^2)\, e^{-2t} \cos^2 \pi x$$

Then the optimal control for $\alpha = 0$ without bound constraints is $u^*(t) = 2 + e^{-t}$ with associated state $y^*(x,t) = 2 + e^{-t} \cos \pi x$. The regularization parameter was set to $\alpha = 10^{-4}$. The L^2 gradient representation of $f(u) \stackrel{\text{def}}{=} J(y(u), u)$ in U_Δ was computed via the discrete adjoint equation, cf. [5]. Since – at least in the case of constant κ and τ – $\nabla^2 f(u)$ is a compact perturbation of αI, a quasi-Newton approximation of $\nabla^2 f(u)$ like BFGS or PSB is efficient also in the L^2-Hilbert space setting, see [10], [14]. Thus, we may expect that a BFGS- or PSB-approximation of the Hessian in the discrete model performs nearly independent of the discretization, cf. [14] for the mesh-independence of BFGS. For the numerical tests Algorithm 5.17 was embedded in the trust-region framework of [24] as described in Algorithm 10.1. We took a L^2-trust-region and used an extension of the Steihaug CG-iteration in the scaled variables $\hat{s} \stackrel{\text{def}}{=} d_k^{-1} s$ to compute an approximate solution to (58) satisfying the decrease condition (D): Let u_k^s be the current iterate and B_k the approximation of $\nabla^2 f(u_k^s)$. A CG-iteration in the scaled variables $\hat{s}$ is started. If the process leaves the trust-region or $\mathcal{B}$ or if negative curvature[1] is detected, Steihaug's method yields s_k^{SH} and $s_k^1 = \sigma s_k^{SH}$ is a candidate for step 2.4 in Algorithm 10.1. Here $\sigma \in (0,1]$ is chosen maximal such that $u_k^s + 1.0005\, s_k^1 \in \mathcal{B}$. In contrast to Steihaug's algorithm we continue the CG-iteration as long as no negative curvature is detected even if it leaves the trust-region or $\mathcal{B}$ until an inexact unconstrained minimizer $u_k^s + s_k^n$ of $\psi[u_k^s]$ with $\|d_k \nabla \psi[u_k^s](u_k^s + s_k^n)\|_2 \le 10^{-4} \|d_k g_k\|_2$ is found. Then $u_k^s + s_k^n$ is an approximation for u_{k+1}^n in Algorithm 10.1 with $\nabla^2 f(u_k^s)$ replaced by B_k. If the CG-iteration left the trust-region or $\mathcal{B}$ we take $s_k^2 = \min\left(\Delta_k / \|s_k^N\|_2, 1\right) s_k^N$ with the projected step $s_k^N = P[u_k^s](u_k^s + s_k^n) - u_k^s$ according to Algorithm 10.1 and $s_k^3 = \min(\Delta_k / \|s_k^S\|_2, 1) s_k^S$ with s_k^S obtained from s_k^n by the stepsize rule 2.4′ as further candidates. In (35) and 2.4′ we took $\xi = 0.99995$. Now we

[1] This does not apply to BFGS-approximations

set $u_{k+1} = u_k^s + s_k$ where $s_k = s_k^i$, $i \in \{1,2,3\}$ is the trial step that provides the best reduction of $\psi[u_k^s]$. As smoothing step in 2.2 of Algorithm 10.1 we use (cf. §8)

$$S_k^\circ(u) \stackrel{\text{def}}{=} \begin{cases} P[u](u - \alpha^{-1} g(u)) & \text{if } k \geq 1,\ u = u_k, \text{ and } \|u_k - u_{k-1}^s\|_\infty \geq 3\|u_k - u_{k-1}^s\|_2, \\ u & \text{else.} \end{cases}$$

For our numerical results we used a BFGS-approximation of the Hessian with $B_0 = \alpha I$. In Algorithm 10.1 we set $\vartheta = 2$, $\Delta_0 = 1$, $\eta_1 = 0.1$, $\eta_2 = 0.75$, $\eta_3 = 0.9$ and $\gamma_1 = 0.5$, $\gamma_2 = \gamma_3 = 2$. The stopping criterion was $\|d(u_k^s) g(u_k^s)\|_2 \leq 10^{-10}$. The upper and lower bounds for the control were $a \equiv -1000$, $b \equiv 0.8$ and we used $c \stackrel{\text{def}}{=} 0.075 \min\{b - a, 0.8\}$ in the definiton of the discrete scaling function d. The optimization was started with $u_0 \equiv 0.05$.

	Proj. & Smooth.		Projection		Stepsize rule	
k	$\|s_k^s\|_\infty^\dagger$	$\|d_{k+1}^s g_{k+1}^s\|_2^\dagger$	$\|s_k^s\|_\infty^\dagger$	$\|d_{k+1}^s g_{k+1}^s\|_2^\dagger$	$\|s_k^s\|_\infty^\dagger$	$\|d_{k+1}^s g_{k+1}^s\|_2^\dagger$
	$N_t = 100 \quad N_x = 20$					
	grad-evals: 8		grad-evals: 8		grad-evals: 13	
0	2.041E−01	3.128E−06	2.041E−01	3.128E−06	2.041E−01	3.128E−06
1	4.093E−01	1.752E−06	4.093E−01	1.752E−06	4.093E−01	1.752E−06
2	3.381E−01	1.304E−07	3.381E−01	1.304E−07	1.366E−01	1.146E−06
3	7.382E−02	4.110E−10*	7.381E−02	9.494E−09	6.285E−02	8.221E−07
4	2.821E−04	1.984E−12*	1.590E−02	1.931E−09	7.082E−02	5.951E−07
5			6.501E−03	3.038E−10	7.028E−02	4.029E−07
6			1.634E−03	1.908E−11	6.683E−02	2.379E−07
7					7.704E−02	6.209E−08
8					1.849E−02	1.211E−08
9					9.845E−03	1.640E−09
10					3.251E−03	2.489E−10
11					1.790E−03	4.387E−11
	$N_t = 400 \quad N_x = 80$					
	grad-evals: 8		grad-evals: 9		grad-evals: 14	

† $s_k^s = u_{k+1}^s - u_k^s$, $d_k^s = d(u_k^s)$, $g_k^s = g(u_k^s)$ * Smoothing occurred, i.e. $u_{k+1}^s \neq u_{k+1}$

TAB. 3: Results for $N_t = 100$, $N_x = 20$

Table 3 shows $\|u_{k+1}^s - u_k^s\|_\infty$ and the norm $\|d(u_{k+1}^s) g(u_{k+1}^s)\|_2$ of the scaled gradient ($u_k^s = u_k$ if no smoothing step occurs) for three different algorithms. The first algorithm is as described above. It uses also the projection step s_k^2 as candidate for the trial step and performs a smoothing step if necessary (see above). The second algorithm is the same but without smoothing. The third algorithm is the same as the second but uses only the trial steps s_k^1, s_k^3 and not the projected step s_k^2 that is suggested by our investigations. There were no rejected trial steps in all three algorithms. Except for the first two iterations the projected step s_k^2 was chosen by the first two algorithms. Obviously the first algorithm provides the fastest convergence. But also if the smoothing steps are omitted the usage of the projected step s_k^2 leads to a significant acceleration of the local convergence in comparison to a stepsize-based algorithm. To demonstrate that the iteration numbers are nearly independent of the mesh-size, we list also the number of gradient evaluations for $N_t = 400$, $N_x = 80$.

Conclusions. We have developed an affine-scaling interior-point Newton algorithm for bound-constrained minimization subject to pointwise bounds in L^p-space. The method is an extension of the algorithms by Coleman and Li [6], [7] for finite-dimensional problems. Our infinite-dimensional framework raised a couple of difficulties which are not present in the finite-dimensional case. A careful analysis led to several modifications of the original algorithm which enabled us to prove superlinear convergence for the resulting method. Under a slightly stronger strict complementarity condition we proved convergence with Q-rate >1. Our main modifications are the introduction of a smoothing step and the implementation of the back-transport by a projection instead of the usual stepsize rule. The smoothing step takes care of the fact that, in general, we only can show that for suitable $q < s$ the affine-scaling Newton step produces a point which is much closer to the solution in L^q (but not necessarily in L^s) than the current iterate was in L^s. The necessity of a smoothing step was also observed by Kelley and Sachs [15] in their study on projected Newton methods. The back-transport is required because the solution of the affine-scaling Newton equation may lie outside of the feasible set $\mathcal{B}$. In the finite-dimensional case one can prove that a stepsize rule to enforce strict feasibility generates stepsizes that converge to one. In our infinite-dimensional setting, however, this is no longer true as we have demonstrated in Example 6.3. Therefore, we have defined a back-transport on the basis of the pointwise projection onto $\mathcal{B}$. We have discussed how smoothing steps can be obtained for a class of regularized problems. Moreover, we have shown that our theory is applicable under the assumptions used by Kelley and Sachs [15] as well as those by Dunn and Tian [9]. We have demonstrated that our algorithm can be used as accelerator for the class of globally convergent trust-region interior-point methods introduced in [24]. The good performance of this algorithm is documented by our numerical results for the boundary control of a heating process.

Acknowledgments. The major part of this work was done while the authors were visiting the Department of Computational and Applied Mathematics and the Center for Research on Parallel Computation, Rice University. They would like to thank John Dennis, Rice University, and Klaus Ritter, Technische Universität München, for making this pleasant and fruitful visit possible.

This work was initiated and influenced by many discussions with Matthias Heinkenschloss, Rice University. His support is greatly acknowledged. We also are grateful to John Dennis, Rice University, and Luís Vicente, Universidade de Coimbra, for several stimulating discussions.

REFERENCES

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

[2] W. Alt, *The Lagrange–Newton method for infinite dimensional optimization problems*, Numer. Funct. Anal. Optim. 11 (1990), pp. 201–224.

[3] D. P. Bertsekas, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.

[4] M. A. Branch, T. F. Coleman, and Y. Li, *A subspace, interior, and conjugate gradient method for large–scale bound–constrained minimization problems*, Tech. Rep. 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`.

[5] 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.

[6] 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.

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

[8] 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`.

[9] J. C. Dunn and T. Tian, *Variants of the Kuhn-Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1384.

[10] A. Griewank, *The local convergence of Broyden-like methods on Lipschitzian problems in Hilbert spaces*, SIAM J. Numer. Anal., 24 (1987), pp. 684–705.

[11] M. Heinkenschloss, *A trust region method for norm constrained problems*, ICAM Report 94–08–01, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, 1994. Available via the URL `http://www.caam.rice.edu/~heinken/papers/Papers.html`.

[12] ———, *Projected sequential quadratic programming methods*, SIAM J. Optimization, 6 (1996), pp. 373–417.

[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, *Quasi-Newton methods and unconstrained optimal control problems*, SIAM J. Control Optim., 25 (1987), pp. 1503–1516.

[15] ———, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Scientific Computing, 15 (1994), pp. 645–667.

[16] ———, *Solution of optimal control problems by a pointwise projected Newton method*, SIAM J. Control Optim., 33 (1995), pp. 1731–1757.

[17] ———, *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`.

[18] F.–S. Kupfer and E. W. Sachs, *Numerical solution of a nonlinear parabolic control problem by a reduced SQP method*, Comput. Optim. Appl., (1992), pp. 113–135.

[19] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[20] ———, *Optimal control of systems governed by partial differential equations*, Springer, New York, 1971.

[21] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.

[22] E. Sachs, *A parabolic control problem with a boundary condition of the Stefan-Boltzmann type*, Z. Angew. Math. Mech., 58 (1978), pp. 443–449.

[23] 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.

[24] M. Ulbrich, S. Ulbrich, and M. Heinkenschloss, *Global convergence of trust–region interior–point algorithms for infinite–dimensional nonconvex minimization subject to pointwise bounds*, Tech. Rep. TR97–04, 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`.

[25] 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`.

[26] ———, *On interior–point Newton algorithms for discretized optimal control problems with state constraints*, Tech. Rep. 96–18, Departamento de Matemática, Universidade de Coimbra, 3000 Coimbra, Portugal, 1996. Available via the URL `http://www.mat.uc.pt/~lvicente/papers/papers.html`.