A PROBABILITY-ONE HOMOTOPY ALGORITHM FOR NONSMOOTH EQUATIONS AND MIXED COMPLEMENTARITY PROBLEMS

STEPHEN C. BILLUPS* AND LAYNE T. WATSON†

Abstract. A probability-one homotopy algorithm for solving nonsmooth equations is described. This algorithm is able to solve problems involving highly nonlinear equations, where the norm of the residual has non-global local minima. The algorithm is based on constructing homotopy mappings that are smooth in the interior of their domains. The algorithm is specialized to solve mixed complementarity problems through the use of MCP functions and associated smoothers. This specialized algorithm includes an option to ensure that all iterates remain feasible. Easily satisfiable sufficient conditions are given to ensure that the homotopy zero curve remains feasible, and global convergence properties for the MCP algorithm are developed. Computational results on the MCPLIB test library demonstrate the effectiveness of the algorithm.

Key words. nonsmooth equations, complementarity problems, homotopy methods, smoothing, path following.

AMS subject classifications. 65F10, 65F50, 65H10, 65K10

1. Introduction. The primary attraction of homotopy algorithms is that they are able to reliably solve systems of equations involving highly nonlinear functions, where the norm of the residual may have non-global local minima. This is because, unlike line search or trust region methods, homotopy methods do not rely on descent of a merit function. Instead, they work by following a path, which under certain weak assumptions is known to lead to a solution. Standard probability-one homotopy algorithms require that the system of equations involves only *smooth* (C^2) functions. This paper proposes a probability-one homotopy algorithm for solving *nonsmooth* systems of equations, and specializes this algorithm to solve mixed complementarity problems. The algorithm uses smoothing functions to construct a homotopy mapping that is C^2 in the interior of its domain. This allows the zero curve of the homotopy mapping to be tracked using software from the HOMPACK90 suite of homotopy codes [24]. A preliminary version of this algorithm was presented at the Second International Conference on Complementarity Problems [5]. The algorithm proposed here has two significant improvements: first, a new end game strategy, which makes better use of available information about the behavior of the homotopy zero curve; second, an option for mixed complementarity problems that ensures that all iterates generated by the algorithm are feasible. This is important because many applications involve functions that are not defined outside of the feasible region. For the case of mixed complementarity problems, new convergence results are presented, which establish easily satisfiable sufficient conditions to ensure that the homotopy zero curve always remains strictly feasible.

In order to describe the algorithm, a significant amount of background material is needed. This is given in Section 2, which discusses notation, nonsmooth equations, a generalized Newton method for nonsmooth equations (which will be used in the end

*Department of Mathematics, University of Colorado at Denver, Denver, CO, 802713364 (sbillups@carbon.cudenver.edu), research partially supported through NSF Grant DMS-9973321.

†Departments of Computer Science and Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0106, (ltw@cayuga.cs.vt.edu), research partially supported by AFOSR Grant F496320-99-1-0128 and NSF Grant DMS-9625968.

game), probability-one homotopy methods, complementarity problems, and smoothing functions. Section 3 describes a probability-one homotopy algorithm for nonsmooth equations. This algorithm is then specialized to solve mixed complementarity problems in Section 4. Section 5 addresses implementation details and computational results, and Section 6 concludes.

2. Background.

2.1. Notation. When discussing vectors and vector-valued functions, subscripts are used to indicate components, whereas superscripts are used to indicate the iteration number or some other label. In contrast, for scalars or scalar-valued functions, subscripts refer to labels so that superscripts can be used for exponentiation. The vector of all ones is represented by e.

Unless otherwise specified, $\|\cdot\|$ denotes the Euclidean norm. For a set $C \subset \mathbb{R}^n$, $\pi_C(x)$ represents the orthogonal projection (with respect to the Euclidean norm) of x onto C. The symbol $\mathbb{R}_+$ refers to the nonnegative real numbers. The extended real numbers are denoted by $\overline{\mathbb{R}} := \mathbb{R} \bigcup \{-\infty, +\infty\}$.

Real-valued functions are denoted with lower-case letters like f or ϕ whereas vector-valued functions are represented by upper-case letters like F or Φ. For a function $F : C \subset \mathbb{R}^n \to \mathbb{R}^m$, $\nabla F(x)$ is the $m \times n$ matrix whose i,jth element is $\partial F_i(x)/\partial x_j$. Let $D \subset \mathbb{R}^m$. Then $F^{-1}(D)$ is the set-valued inverse defined by $F^{-1}(D) := \{x \mid F(x) \in D\}$.

Given a function $F : \mathbb{R}^n \to \mathbb{R}^m$, the directional derivative of F at x in the direction d is denoted by $F'(x;d) := \lim_{t \downarrow 0} (F(x+td) - F(x))/t$, provided the limit exists.

2.2. Nonsmooth equations. This paper is concerned with solving equations of the form $F(x) = 0$, where the function $F : \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitzian, but not necessarily continuously differentiable. Such nonsmooth equations provide a unifying framework for the study of many important classes of problems, including constrained optimization, finite-dimensional variational inequalities, complementarity problems, equilibrium problems, generalized equations, partial differential equations, and fixed point problems. The following definitions will be used throughout the paper.

By Rademacher's theorem, since F is locally Lipschitzian, it is differentiable almost everywhere. Let D_F be the set where F is differentiable. Define the *B-subdifferential* by

$$\partial_B F(x) := \left\{ V \;\middle|\; \exists \{x^k\} \to x, x^k \in D_F, \quad \text{with} \quad V = \lim_{k \to \infty} \nabla F(x_k) \right\}.$$

The Clarke subdifferential $\partial F(x)$ is the convex hull of $\partial_B F(x)$.

F is said to be *semismooth* [19] at x if it is directionally differentiable at x and for any $V \in \partial F(x+h)$, $h \to 0$,

$$Vh - F'(x;h) = o(\|h\|).$$

F is said to be *strongly semismooth* [10] if additionally,

$$Vh - F'(x;h) = O(\|h\|^2).$$

A semismooth function $F : \mathbb{R}^n \to \mathbb{R}^n$ is *BD-regular* at x if all elements in $\partial_B F(x)$ are nonsingular, and F is *strongly regular* at x if all elements in $\partial F(x)$ are nonsingular.

2.3. Newton's method for nonsmooth equations. One approach to solving the nonsmooth equation $F(x) = 0$ is a generalization of Newton's method to semismooth equations, which was proposed by Qi [19]. Qi's method is used together with an Armijo line search in the end game of the homotopy algorithm proposed here. Qi's algorithm, which is discussed in detail in [3], is shown in Figure 2.1. θ in this algorithm is the merit function defined by $\theta(x) := \frac{1}{2}F(x)^T F(x)$. Theorem 2.1, which is restated from [19] and [10], shows that this algorithm has the same fast local convergence properties as the standard (smooth) Newton's method under natural generalizations of the standard assumptions.

FIG. 2.1. *Generalized damped newton method*

Step 1 [Initialization] Select line search parameters α, $\sigma \in (0, 1)$, a positive integer m_{max}, a starting point $x^0 \in \mathbb{R}^n$, and a stopping tolerance *tol*. Set $k = 0$.

Step 2 [Direction generation] Choose $V^k \in \partial_B F(x^k)$. If V^k is singular, stop, returning the point x^k along with a failure message. Otherwise choose the direction

$$d^k = -(V^k)^{-1} F(x^k). \tag{2.1}$$

Step 3 [Step length determination] Let m_k be the smallest nonnegative integer $m \leq m_{max}$ such that

$$\theta(x^k + \alpha^m d^k) - \theta(x^k) \leq -\sigma\alpha^m \theta(x^k). \tag{2.2}$$

If no such m_k exists, stop; the algorithm failed. Otherwise set $x^{k+1} = x^k + \alpha^{m_k} d^k$.

Step 4 [Termination check] If $\theta(x^{k+1}) < tol$ stop, returning the point x^{k+1}. Otherwise, return to step 2, with k replaced by $k + 1$.

THEOREM 2.1. *Suppose that x^* is a solution of $F(x) = 0$ and that F is semismooth and BD-regular at x^*. Then the iteration method defined by $x^{k+1} = x^k + d^k$, where d^k is given by (2.1) is well defined and convergent to x^* Q-superlinearly in a neighborhood of x^*. If F is strongly semismooth at x^*, the iteration sequence converges to x^* Q-quadratically.*

One consequence of this local convergence theorem is that within a neighborhood of a BD-regular solution x^*, the line search criterion (2.2) will be satisfied by $m_k = 0$. Thus, the inner algorithm will take full Newton steps and achieve the fast local convergence rates specified by the theorem.

The damped Newton method described above works very well when started near a solution, or when applied to problems that are nearly linear in the sense that their merit functions do not contain local minima that are not solutions.

For highly nonlinear problems, the damped Newton method tends to fail without a carefully chosen starting point. The reason, of course, is that unless started close to a solution, the iterates may converge only to a local minimum of the merit function. This motivates the consideration of homotopy methods, which are truly globally convergent.

2.4. Homotopy methods. The main theory underlying the present homotopy method is summarized in the following proposition from [5]. This proposition is similar to results presented in [20] and [8, Theorem 2.4]; however, it does not assume F itself to be differentiable. The path γ_a defined in the proposition "reaches a zero of F" in the sense that it contains a sequence $\{(\lambda_k, x^k)\}$ that converges to $(1, \bar{x})$, where $\bar{x}$ is a zero of F.

PROPOSITION 2.2. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a Lipschitz continuous function and suppose there is a C^2 map*

$$\rho : \mathbb{R}^m \times [0,1) \times \mathbb{R}^n \to \mathbb{R}^n$$

such that

1. *$\nabla\rho(a, \lambda, x)$ has rank n on the set $\rho^{-1}(\{0\})$,*
2. *the equation $\rho_a(0, x) = 0$, where $\rho_a(\lambda, x) := \rho(a, \lambda, x)$, has a unique solution $x^a \in \mathbb{R}^n$ for every fixed $a \in \mathbb{R}^m$,*
3. *$\nabla_x \rho_a(0, x^a)$ has rank n for every $a \in \mathbb{R}^m$,*
4. *ρ is continuously extendible (in the sense of Buck [6]) to the domain $\mathbb{R}^m \times [0,1] \times \mathbb{R}^n$, and $\rho_a(1, x) = F(x)$ for all $x \in \mathbb{R}^n$ and $a \in \mathbb{R}^m$, and*
5. *γ_a, the connected component of $\rho_a^{-1}(\{0\})$ containing $(0, x^a)$, is bounded for almost every $a \in \mathbb{R}^m$.*

Then for almost every $a \in \mathbb{R}^m$ there is a zero curve γ_a of ρ_a, along which $\nabla\rho_a$ has rank n, emanating from $(0, x^a)$ and reaching a zero $\bar{x}$ of F at $\lambda = 1$. Further, γ_a does not intersect itself and is disjoint from any other zeros of ρ_a. Also, if γ_a reaches a point $(1, \bar{x})$ and F is strongly regular at $\bar{x}$, then γ_a has finite arc length.

Because γ_a is a smooth curve, it can be parameterized by its arc length away from $(0, x^a)$. This yields a function $(\lambda(s), x(s))$, representing the point on γ_a of arc length s away from $(0, x^a)$.

The construction of a globally convergent probability-one homotopy algorithm entails: (1) constructing a map ρ according to Proposition 2.2, (2) choosing $a \in \mathbb{R}^m$, (3) finding x^a solving $\rho_a(0, x) = 0$, and (4) tracking γ_a starting from $(0, x^a)$ until $\lambda = 1$. Assuming an appropriate ρ exists, the theory guarantees that for almost all a (in the sense of Lebesgue measure), γ_a exists and leads to a solution, hence the term "probability-one".

A simple (and occasionally useful in practice) homotopy mapping is $\rho : \mathbb{R}^n \times [0,1) \times \mathbb{R}^n \to \mathbb{R}^n$ given by

$$\rho(a, \lambda, x) := \lambda F(x) + (1 - \lambda)(x - a). \tag{2.3}$$

If F is C^2 then ρ trivially satisfies properties (1), (2), (3), and (4) but not necessarily (5) of Proposition 2.2. The following theorem gives conditions on F under which the fifth condition is satisfied. This result will be generalized to nonsmooth functions in Theorem 3.2.

THEOREM 2.3. [22] *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a C^2 function such that for some $\tilde{x} \in \mathbb{R}^n$ and $r > 0$,*

$$(x - \tilde{x})^T F(x) \geq 0 \text{ whenever } \|x - \tilde{x}\| = r. \tag{2.4}$$

Then F has a zero in a closed ball of radius r about $\tilde{x}$, and for almost every a in the interior of this ball there is a zero curve γ_a of

$$\rho_a(\lambda, x) := \lambda F(x) + (1 - \lambda)(x - a),$$

along which $\nabla\rho_a(\lambda, x)$ *has full rank, emanating from* $(0, a)$ *and reaching a zero* $\bar{x}$ *of* F *at* $\lambda = 1$. *Further,* γ_a *has finite arc length if* $\nabla F(\bar{x})$ *is nonsingular.*

The actual statement of the theorem in [22] fixes $\tilde{x} = 0$. However, the proof can be modified trivially to yield the more general theorem above. (See the proof of [5, Theorem 2.11] for the necessary modifications). It is interesting to note that in many applications, (2.4) holds for all r sufficiently large (not just for some fixed r). This makes the choice of $\tilde{x}$ irrelevant. Furthermore, in such cases, a can be chosen arbitrarily, (instead of from some neighborhood of $\tilde{x}$), thus making the method truly globally convergent (with probability one).

(2.4) will be referred to as the *global monotonicity* property. If a C^2 function F possesses this property, these theoretical results have some profound implications: the guaranteed existence of a path between almost any starting point and a solution $\bar{x}$ to $F(x) = 0$, which has finite arc length if $\operatorname{rank} \nabla F(\bar{x}) = n$. In theory, to find a solution, one must simply follow the path to a point of γ_a where $\lambda = 1$. In practice, however, the task of constructing a ρ for which γ_a is short and smooth is very difficult, although this has been done for large classes of problems.

Several packages exist to solve root finding problems using homotopy techniques [24]. The implementation here uses the routine STEPNX from the HOMPACK90 suite of software [23] [24, Section 3], which tracks the zero curve of a homotopy mapping specified by the user.

2.5. Complementarity problems. Given a continuously differentiable function $G : \mathbb{R}^n \to \mathbb{R}^n$, the nonlinear complementarity problem NCP(G) is to find some $x \in \mathbb{R}^n$ so that

$$0 \le x \perp G(x) \ge 0, \tag{2.5}$$

where $x \perp G(x)$ means that $x^T G(x) = 0$.

Given a rectangular region $\mathbb{B}_{l,u} := \prod_{i=1}^{n} [l_i, u_i] \subset \overline{\mathbb{R}}^n$ defined by two vectors, l and u in $\mathbb{R}^n$ where $-\infty \le l < u \le \infty$, and a function $G : \mathbb{R}^n \to \mathbb{R}^n$, the mixed complementarity problem MCP($G, \mathbb{B}_{l,u}$) is to find an $x \in \mathbb{B}_{l,u}$ such that for each $i \in \{1, \ldots, n\}$, either 1) $x_i = l_i$ and $G_i(x) \ge 0$, 2) $G_i(x) = 0$, or 3) $x_i = u_i$ and $G_i(x) \le 0$. This is equivalent to the condition that $\operatorname{mid}(x - l, x - u, G(x)) = 0$, where mid represents the componentwise median function. When these conditions are satisfied, write $G(x) \perp x$ and say that x is complementary to $G(x)$. Assume henceforth that G is C^2.

It is well known that NCP(G) can be reformulated as a system of equations. This was first shown by Mangasarian [16]. An excellent review of reformulations of NCP can be found in [18]. To discuss such reformulations requires several definitions, which are equivalent to the NCP function and the BVIP function defined in [18]:

DEFINITION 2.4. *A function* $\phi : \mathbb{R}^2 \to \mathbb{R}$ *is called an NCP function provided* $\phi(a, b) = 0$ *if and only if* $\min(a, b) = 0$.

DEFINITION 2.5. *A function* $\psi : \mathbb{R} \bigcup \{-\infty\} \times \mathbb{R} \bigcup \{\infty\} \times \mathbb{R}^2 \to \mathbb{R}$ *is called an MCP function provided* $\psi(l, u, a, b) = 0$ *if and only if* $\operatorname{mid}(a - l, a - u, b) = 0$.

It is useful to further distinguish NCP and MCP functions according to their orientations:

DEFINITION 2.6. *An NCP function* ϕ *is called* positively oriented *if for all* $a, b \in \mathbb{R}$,

$$\operatorname{sign}(\phi(a, b)) = \operatorname{sign}(\min(a, b)).$$

An MCP function ψ is called positively oriented *if*

$$\text{sign}(\psi(l,u,a,b)) = \text{sign}(\text{mid}(a-l, a-u, b))$$

for all $l \in \mathbb{R} \bigcup \{-\infty\}, u \in \mathbb{R} \bigcup \{\infty\}$, $l < u$, and $a, b \in \mathbb{R}$.

An NCP function that has been very popular recently is the Fischer-Burmeister function [13] $\phi : \mathbb{R}^2 \to \mathbb{R}$, defined by

$$\phi^{FB}(a,b) := a + b - \sqrt{a^2 + b^2}. \tag{2.6}$$

It is easily seen that $\phi^{FB}(a,b) = 0$ if and only if $0 \le a \perp b \ge 0$. Thus, by defining the function $F : \mathbb{R}^n \to \mathbb{R}^n$ by

$$F_i(x) := \phi^{FB}(x_i, G_i(x)), \tag{2.7}$$

it is clear that $x \in \mathbb{R}^n$ solves NCP(G) if and only if $F(x) = 0$.

While ϕ^{FB} is not differentiable at the origin, $(\phi^{FB})^2$ is continuously differentiable everywhere. This property, together with the fact that ϕ^{FB} is semismooth, makes this reformulation well suited for use in globalization strategies for nonsmooth Newton-based methods (see, for example, [9]).

Given a positively oriented NCP function ϕ, and the convention that $\phi(\infty, b) = \lim_{a\to\infty} \phi(a,b)$ and $\phi(a,\infty) = \lim_{b\to\infty} \phi(a,b)$, an MCP function ψ can be constructed using the following formula, first proposed in [1]:

$$\psi(l,u,a,b) := \phi(a-l, -\phi(u-a, -b)). \tag{2.8}$$

Constructing the function $F : \mathbb{R}^n \to \mathbb{R}^n$ by

$$F_i(x) := \psi(l_i, u_i, x_i, G_i(x)), \tag{2.9}$$

yields a reformulation of the MCP($G, \mathbb{B}_{l,u}$); $F(x) = 0$ if and only if x is a solution to MCP($G, \mathbb{B}_{l,u}$) [2].

Note that for the Fischer-Burmeister function, $\lim_{a\to\infty} \phi^{FB}(a,b) = b$ and $\lim_{b\to\infty} \phi^{FB}(a,b) = a$. Thus, for the MCP case, if l_i is finite, and $u_i = \infty$, then $F_i(x) = \phi^{FB}(x_i - l_i, G_i(x))$; if u_i is finite and $l_i = -\infty$, then $F_i(x) = -\phi^{FB}(u_i - x_i, -G_i(x))$; and if neither bound is finite, $F_i(x) = G_i(x)$.

2.6. Smoothing operators. Consider the system $F(x) = 0$ where F is a nonsmooth function, and suppose there exists a family of functions F^μ parameterized by a *smoothing parameter* μ so that $\lim_{\mu\downarrow 0} F^\mu = F$ in some sense. Under suitable conditions, the solutions to the systems $F^\mu(x) = 0$ converge to a solution to $F(x) = 0$ along a smooth trajectory [7].

DEFINITION 2.7. *Given a nonsmooth continuous function $\phi : \mathbb{R}^p \to \mathbb{R}$, a* smoother *for ϕ is a continuous function $\tilde{\phi} : \mathbb{R}^p \times \mathbb{R}_+ \to \mathbb{R}$ such that*

1. *$\tilde{\phi}(x, 0) = \phi(x)$, and*
2. *$\tilde{\phi}$ is continuously differentiable on the set $\mathbb{R}^p \times \mathbb{R}_{++}$.*

If $\tilde{\phi}$ is C^2 on $\mathbb{R}^p \times \mathbb{R}_{++}$, call $\tilde{\phi}$ a C^2-smoother.

For convenience, define $\phi_\mu(x) := \tilde{\phi}(\cdot, \mu)$. To define smoothers for functions $F : \mathbb{R}^n \to \mathbb{R}^n$, say that $F^\mu : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}^n$ is a smoother for F if for each $i \in \{1 \dots n\}$, F_i^μ is a smoother for F_i.

In the case of complementarity problems, the NCP functions and MCP functions generally have well understood nonsmoothness structure, so C^2-smoothers for these functions can usually be easily constructed. As an example, the following C^2-smoother for the Fischer-Burmeister function was proposed by Kanzow [15]:

$$\tilde{\phi}^K(a,b,\mu) := a + b - \sqrt{a^2 + b^2 + 2\mu}, \tag{2.10}$$

The following smoother is more useful here, since its partial derivative with respect to μ is bounded near the origin.

$$\tilde{\phi}^{BW}(a,b,\mu) := a + b - \sqrt{a^2 + b^2 + \mu^2}, \tag{2.11}$$

Given a smoother $\tilde{\phi}$ for a NCP function ϕ and the convention that $\tilde{\phi}(\infty, b, \mu) = \lim_{a\to\infty} \tilde{\phi}(a,b,\mu)$ and $\tilde{\phi}(a,\infty,\mu) = \lim_{b\to\infty} \tilde{\phi}(a,b,\mu)$, a smoother $\tilde{\psi}$ for the MCP function ψ defined by (2.8) can be constructed according to the formula:

$$\tilde{\psi}(l,u,a,b,\mu) := \phi_\mu(a-l, -\phi_\mu(u-a,-b)). \tag{2.12}$$

Smoothers for (2.7) and (2.9) are then given, respectively, by

$$F_i^\mu(x) := \phi_\mu(x_i, G_i(x)), \quad \text{and} \tag{2.13}$$

$$F_i^\mu(x) := \psi_\mu(l_i, u_i, x_i, G_i(x)). \tag{2.14}$$

Note that for the smoother defined by (2.11), $\lim_{a\to\infty} \tilde{\phi}^{BW}(a,b,\mu) = b$, and $\lim_{b\to\infty} \tilde{\phi}^{BW}(a,b,\mu) = a$. Thus, for the MCP case, if $u_i = \infty$ and l_i is finite, then $F_i^\mu(x) = \tilde{\phi}^{BW}(x_i - l_i, G_i(x), \mu)$; if u_i is finite and $l_i = -\infty$, then $F_i^\mu(x) = -\tilde{\phi}^{BW}(u_i - x_i, -G_i(x), \mu)$; and if neither bound is finite, then $F_i^\mu(x) = G_i(x)$.

3. The algorithm. This section summarizes the probability-one homotopy algorithm for solving nonsmooth equations. It contrasts with an earlier hybrid Newton-homotopy method described in [2]. The earlier method begins by using a nonsmooth version of a damped-Newton's method to solve the root finding problem $F(x) = 0$. If the Newton algorithm stalls, a standard homotopy method is invoked to solve a particular smoothed version of the original problem, $F^\mu(x) = 0$, where μ is fixed. The smoothing parameter μ is chosen based on the level of a merit function on F at the last point $\hat{x}$ generated by the Newton method. Starting from $\hat{x}$, a homotopy method is carried out until it produces a point that yields a better merit value than the previous Newton iterate. The Newton method is then started again and the process repeats until a point is produced that is close enough to a solution or the homotopy method fails. One key feature of that hybrid method is that each time the Newton method stalls, a different homotopy map is constructed. The smoothing parameter μ is chosen based on the level of the merit function when the Newton method stalls, so the homotopy that is then used is

$$\rho_a^\mu(\lambda, x) := \lambda F^\mu(x) + (1-\lambda)(x-a).$$

An alternative approach, described here, is to adopt a pure probability-one homotopy algorithm by fixing the homotopy map and tracking a single homotopy zero curve into the Newton domain of convergence around a solution. Essentially, the idea is to use a standard probability-one homotopy algorithm, but with a specially designed "end game" near a solution. The key to this approach is to define a homotopy mapping that couples the smoothing parameter with the homotopy parameter.

3.1. The homotopy map. Given a function F and an associated C^2-smoother F^μ, construct a homotopy mapping with F^μ where the smoothing parameter μ is a function of the homotopy parameter λ so that $\mu \downarrow 0$ as $\lambda \uparrow 1$. If this homotopy satisfies the conditions in Proposition 2.2, a well behaved path exists from almost any starting point to a solution, and standard curve tracking techniques can reliably solve the equation $F(x) = 0$.

Throughout this section, assume that F is a Lipschitz continuous function on $\mathbb{R}^n$ and that F^μ is a C^2-smoother for F. Take $\mu : [0,1] \to \mathbb{R}_+$ to be a decreasing C^2 function such that $\mu(\lambda) > 0$ for $\lambda < 1$ and $\mu(1) = 0$. For example,

$$\mu(\lambda) := \alpha(1-\lambda) \tag{3.1}$$

for some parameter $\alpha > 0$. Define the homotopy map $\rho_a : [0,1) \times \mathbb{R}^n \to \mathbb{R}^n$, nonlinear in λ, by

$$\rho_a(\lambda, x) := \lambda\, F^{\mu(\lambda)}(x) + (1-\lambda)(x-a) \tag{3.2}$$

and let γ_a be the connected component of the set $\rho_a^{-1}(\{0\})$ that contains $(0, a)$. Notice that this mapping is a generalization of (2.3), since if F is C^2, then $F^\mu := F$ suffices.

In order to ensure that a well behaved zero curve exists, conditions on F and its smoother are required so that Proposition 2.2 can be invoked. The following weak assumption on the smoother will be useful in the theory that follows.

ASSUMPTION 3.1. *There is a nondecreasing function $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\lim_{\nu \downarrow 0} \eta(\nu) = 0$ such that for all x in $\mathbb{R}^n$ and all ν in $\mathbb{R}_+$*

$$\|F^\nu(x) - F(x)\|_\infty \le \eta(\nu).$$

Note (by [2, Proposition 2.14]) that if F^ν is constructed by (2.14), with ϕ_μ defined either by (2.10) or (2.11), Assumption 3.1 is satisfied with $\eta(\nu) := 3\sqrt{2\nu}$ or $\eta(\nu) := 3\nu$, respectively.

The following theorem [5, Theorem 2.11] is a generalization of Theorem 2.3.

THEOREM 3.2. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a Lipschitz continuous function such that for some fixed $r > 0$ and $\tilde{x} \in \mathbb{R}^n$,*

$$(x - \tilde{x})^T F(x) \ge 0 \text{ whenever } \|x - \tilde{x}\| = r,$$

and let F^μ be a smoother for F satisfying Assumption 3.1. Further, suppose that the smoothing parameter $\mu(\lambda)$ is such that

$$\eta(\mu(\lambda)) < \frac{1-\lambda}{\lambda} M \text{ for } 0 < \lambda \le 1 \tag{3.3}$$

for some $M \in (0, r)$. Then γ_a is bounded for almost every $a \in \mathbb{R}^n$ such that $\|a - \tilde{x}\| < \tilde{r} := r - M$.

A direct application of Proposition 2.2 gives the main convergence theorem.

THEOREM 3.3. *Under the assumptions of Theorem 3.2, F has a zero in a closed ball of radius r about $\tilde{x}$, and for almost every a in the interior of a ball of radius $\tilde{r}$ about $\tilde{x}$, there is a zero curve γ_a of*

$$\rho(a, \lambda, x) := \rho_a(\lambda, x) := \lambda\, F^{\mu(\lambda)}(x) + (1-\lambda)(x-a),$$

along which $\nabla \rho_a(\lambda, x)$ has full rank, emanating from $(0, a)$ and reaching a zero $\bar{x}$ of F at $\lambda = 1$. Further, γ_a has finite arc length if F is strongly regular at $\bar{x}$.

Observe that in applications, the r in Theorem 3.2 can be arbitrarily large, hence so can $\tilde{r} = r - M$, and thus $\|a - \tilde{x}\| < \tilde{r}$ is really no restriction at all.

3.2. Tracking the zero curve. As discussed in Section 2.4, the zero curve can, with probability one, be parameterized by arc length: let $(\lambda(s), x(s))$ be the point on γ_a of arc length s away from $(0, x^a)$. Tracking the zero curve involves generating a sequence of points $\{y^k\} \subset \mathbb{R}^{n+1}$, with $y^0 = (0, x^a)$ that lie approximately on the curve in order of increasing arc length. That is, $y^k \approx (\lambda(s_k), x(s_k))$, where $\{s_k\}$ is some increasing sequence of arc lengths.

The subroutine STEPNX from HOMPACK90 [24] is used to handle the curve tracking. At each iteration, STEPNX uses a predictor-corrector algorithm to generate the next point on the curve. The prediction phase requires for each iterate y^k the corresponding unit tangent vector to the curve, $(y')^k \approx (\lambda'(s_k), x'(s_k))$. This is accomplished by finding an element η of the null space of $\nabla\rho_a(y^k)$, and setting $(y')^k := \pm\eta/\|\eta\|$, where the sign is chosen so that $(y')^k$ makes an acute angle with $(y')^{k-1}$, for $k > 0$. On the first iterate, the sign is chosen so that the first component (corresponding to λ) of $(y')^0$ is positive.

At each iteration after the first, STEPNX approximates the zero curve with a Hermite cubic polynomial $c^k(s)$, which is constructed using the last two points y^{k-1} and y^k, along with the associated unit tangent vectors $(y')^{k-1}$ and $(y')^k$. A step of length h along this cubic yields the predicted point $w^{k,0} := c(s_k + h)$. The first iteration uses a linear predictor instead, which is constructed using the starting point y^0 and its associated unit tangent vector.

Once the predicted point is calculated, a normal flow corrector algorithm [24] is used to return to the zero curve. Starting with the initial point $w^{k,0}$, the corrector iterates $w^{k,j}, j = 1, \ldots$ are calculated via the formula $w^{k,j+1} := w^{k,j} + z^{k,j}$, $j = 0, 1, \ldots$, where the step $z^{k,j}$ is the unique minimum-norm solution to the equation

$$\nabla\rho_a(w^{k,j})z^{k,j} = -\rho_a(w^{k,j}). \tag{3.4}$$

The corrector algorithm terminates when one of the following conditions is satisfied: the normalized correction step $z^{k,j} / \left(1 + \left\|w^{k,j}\right\|\right)$ is sufficiently small; some maximum number of iterations (usually 4) is exceeded; or a rank-deficient Jacobian matrix is encountered in (3.4). In the first case, set $y^{k+1} := w^{k,j}$, calculate an optimal step size h for the next iteration, and proceed to the next prediction step. In the second case, discard the point and return to the prediction phase using a smaller step size if possible; otherwise, terminate curve tracking with an error return. In the third case, terminate the curve tracking, since $\operatorname{rank}\nabla\rho_a < n$ should theoretically not happen, and indicates serious difficulty. The step size in h is also never reduced beyond relative machine precision.

3.2.1. Step size control. At each iteration, STEPNX estimates an "optimal" step size to be used in computing the predicted point. This calculation is governed by several user-defined parameters. Successful termination of the corrector phase occurs when the norm of the residual $\left\|\rho(w^{k,j})\right\|$ is sufficiently small. In some cases, this can happen even when the converged point is not close to the true zero curve. As the tracking progresses, the computed points may slowly drift farther and farther from the zero curve, while continuing to meet the criterion on the norm of the residual. Eventually, the iterates may leave the Newton domain of attraction, and the corrector phase may fail to converge, no matter how small the predictor step is. To avoid such difficulties, STEPNX calculates several quantities that measure the "quality" of the step.

The first quantity is the contraction factor

$$\left\|z^{k,1}\right\| / \left\|z^{k,0}\right\| ,$$

which measures how much the Newton step shrinks from the first corrector iteration to the second. The second quantity is the residual factor

$$\left\|\rho_a(w^{k,1})\right\| / \left\|\rho_a(w^{k,0})\right\| .$$

The third quantity is the distance factor

$$\left\|w^{k,1} - y^{k+1}\right\| / \left\|w^{k,0} - y^{k+1}\right\| ,$$

which approximates how much the distance from the zero curve shrinks from the first iteration to the second. Since Newton's method has quadratic local convergence, each of these quantities should be small when the predicted point is close to the zero curve. Through the use of input parameters, the user is able to specify ideal values (`lideal, rideal, dideal`, respectively) for each of these quantities. If the quantities are smaller than the ideal, the step size will be increased; if the quantities are larger than ideal, the step size will be decreased. The amount of increase or decrease is also controlled by user-defined parameters. Generally, default values for all of these parameters work very well. However, occasionally, it is necessary to choose more conservative parameter values in order to avoid losing the zero curve.

As a final consideration, the default limit on the number of Newton iterations in the corrector phase is 4 (a HOMPACK90 parameter). In some cases, increasing this limit to 6 or 8 improved performance.

3.3. The end game. The standard homotopy method used by HOMPACK90 concludes the curve tracking with an end game strategy that zeros in on a point (λ, x) on the zero curve with $\lambda = 1$. This end game strategy, which is a robust blend of secant iterations with Newton corrections, is begun when a point (λ, x) is found on the zero curve with $\lambda > 1$. However, this approach requires that $\rho(\lambda, x)$ be defined for $\lambda > 1$—a requirement that is not desirable here since the smoother $F^{\mu(\lambda)}$ may not be defined for $\lambda > 1$. Therefore the standard end game is replaced with the generalized Newton method given in Figure 2.1, which is begun while $\lambda < 1$ still.

The Newton end game is invoked when one of the following criteria is satisfied:

1. The point generated by the cubic predictor (with step length h) has $\lambda > 1$.
2. A linear predictor with the same step length has $\lambda > 1$.
3. The corrector phase of the algorithm generates a point with $\lambda > 1$.

In all cases, a starting point for the Newton end game is the prediction of where the zero curve crosses the hyperplane $\lambda = 1$. The precise details follow.

1. First, try to find a point (λ^c, x^c) for which the cubic approximation has $\lambda^c = 1$. If this point occurs within a step length shorter than $2h$, then x^c will be the starting point.
2. Otherwise, find a point (λ^l, x^l) for which the linear approximation has $\lambda^l = 1$. Then x^l will be the starting point.

If the curve tracking fails for any reason before the end game criteria are met, then attempt the nonsmooth Newton's method with the starting point x, where (λ, x) is the last point found on the zero curve.

The starting point generated by the above procedure is usually quite good. However, in some cases, the Newton end game may fail to converge. In that event, simply return to tracking the zero curve, picking up from the last point y^k on γ_a, but with the step size (computed by STEPNX) cut in half, and with the STEPNX tracking tolerances `abserr` and `relerr` also reduced.

Note that this approach differs from the end game strategy described in [5], which simply invoked the Newton end game with a starting point x whenever a point (λ, x)

was found on the zero curve with λ sufficiently close to 1. The new end game strategy has two main advantages over this earlier approach. First, using the cubic predictor to estimate where the zero curve crosses $\lambda = 1$ results is a significantly more accurate approximation for the solution as a starting point for Newton's method. Second, the new method takes better advantage of available information in determining when to enter the end game. Specifically, on difficult problems, the Newton domain of convergence near the final solution will be small, so it is desirable to track the zero curve very close to $\lambda = 1$ before trying Newton's method. This is exactly what happens since, in this case, the step size will likely be very small. In contrast, for easier problems, larger step sizes will be used, and the end game will be started earlier. Again this is acceptable because the Newton domain of convergence around the solution will likely be large.

In order to solve the system $F(x) = 0$, the nonsmooth Newton's method requires that F be semismooth. If, in addition, F is BD-regular at a solution x^*, Newton's method will converge superlinearly in some neighborhood about x^*. Theoretically, to use the homotopy approach and guarantee the end game's success, F should satisfy the global monotonicity property and be strongly regular at every solution. This guarantees that the homotopy's zero curve crosses the hyperplane $\lambda = 1$ transversally rather than tangentially, and ensures that the zero curve will have finite arc length. For most homotopies used in practice in other contexts, even if the zero curve γ_a is tangent to the hyperplane $\lambda = 1$, a point with $\lambda > 1$ near $\rho_a^{-1}(\{0\})$ will be generated, and the usual end game provided in HOMPACK90 will succeed (to modest accuracy, since $\nabla F(\bar{x})$ is singular).

4. Solving mixed complementarity problems. This section specializes the algorithm described above in order to solve mixed complementarity problems. The approach taken here is to reformulate the MCP by defining the function $F : \mathbb{R}^n \to \mathbb{R}^n$ according to (2.8) and (2.9), where ϕ is a positively oriented NCP function, and defining a smoother for F according to (2.12) and (2.14), where ϕ_μ is a smoother for ϕ. Once these functions are defined, the homotopy algorithm described in the previous section can be used to find a zero of F, which corresponds to a solution of MCP. Because of the special structure of these functions, stronger convergence results are possible than for the general nonsmooth equations problem. The first results presented in this section are tailored to particular choices of ϕ and ϕ_μ, namely the Fischer-Burmeister NCP function (2.6), and the smoother (2.11). More general results are given in Theorem 4.3 and Corollary 4.4. In describing these results it will be useful to refer to the following index set:

$$I_{l,u} = \{ i \mid -\infty < l_i < u_i < \infty \}.$$

That is, $I_{l,u}$ is the set of indices for which both the lower and upper bounds are finite.

THEOREM 4.1. *Let ϕ be the positively oriented NCP function in* (2.6)*, and let $\tilde{\phi}$ be the smoother for ϕ in* (2.11)*. Let ψ be defined by* (2.8) *with associated smoother $\tilde{\psi}$ defined by* (2.12)*. Choose $a \in \operatorname{int} \mathbb{B}_{l,u}$. Let F^μ be defined by* (2.14)*, where $\mu : [0,1] \to \mathbb{R}_+$ is a decreasing C^2 function satisfying $\mu(1) = 0$ and*

$$\mu(\lambda)^2 \leq 2\frac{1-\lambda}{\lambda}(u_i - a_i)(u_i - l_i) \text{ for all } i \in I_{l,u},\ \lambda \in (0,1]. \tag{4.1}$$

Define $\rho_a : [0,1) \times \mathbb{R}^n \to \mathbb{R}^n$ by (3.2)*, and let γ_a be the connected component of $\rho_a^{-1}(\{0\})$ containing $(0,a)$. Then γ_a is contained in $[0,1) \times (\operatorname{int} \mathbb{B}_{l,u})$.*

Proof. Let $(\hat{\lambda}, \hat{x})$ be an arbitrary point on γ_a. If $\hat{\lambda} = 0$, then $\hat{x} = a \in \operatorname{int} \mathbb{B}_{l,u}$; so assume $0 < \hat{\lambda} < 1$. First suppose that $\hat{x}_i \le l_i$ for some i. Then

$$0 = \rho_i(\hat{\lambda}, \hat{x}) = \hat{\lambda} F_i^{\mu(\hat{\lambda})}(\hat{x}) + (1 - \hat{\lambda})(\hat{x}_i - a_i)$$

or

$$F_i^{\mu(\hat{\lambda})}(\hat{x}) = -\frac{1 - \hat{\lambda}}{\hat{\lambda}}(\hat{x}_i - a_i) > 0, \tag{4.2}$$

where the last inequality follows from $\hat{x}_i \le l_i < a_i$, since a is interior to $\mathbb{B}_{l,u}$. Also $F_i^{\mu(\hat{\lambda})}(\hat{x}) = \tilde{\phi}(\hat{x}_i - l_i, \zeta, \mu)$, where $\zeta := -\tilde{\phi}(u_i - \hat{x}_i, -G_i(\hat{x}), \mu)$. Thus,

$$F_i^{\mu(\hat{\lambda})}(\hat{x}) = \tilde{\phi}(\hat{x}_i - l_i, \zeta, \mu) \le \phi(\hat{x}_i - l_i, \zeta) \le 0,$$

contradicting (4.2). It follows that every point (λ, x) on γ_a satisfies $l < x$.

Now suppose $\hat{x}_i \ge u_i$ for some i. Note that this implies that u_i is finite. In this case (analogous to (4.2)),

$$F_i^{\mu(\hat{\lambda})}(\hat{x}) = -\frac{1 - \hat{\lambda}}{\hat{\lambda}}(\hat{x}_i - a_i) \le -\frac{1 - \hat{\lambda}}{\hat{\lambda}}(u_i - a_i) \tag{4.3}$$

and $\zeta = -\tilde{\phi}(u_i - \hat{x}_i, -G_i(\hat{x}), \mu) > 0$ since $\mu(\lambda) > 0$ for $\lambda < 1$. If $l_i = -\infty$, then $F_i^{\mu(\hat{\lambda})}(\hat{x}) = \zeta > 0$, contradicting (4.3). If l_i is finite, then from (6) and (11), for any $\alpha, \beta \in \mathbb{R}$,

$$\tilde{\phi}(\alpha, \beta, \mu) - \phi(\alpha, \beta) > -\frac{\mu^2}{2\sqrt{\alpha^2 + \beta^2}}. \tag{4.4}$$

Then, using $\zeta > 0$, $\hat{x}_i \ge u_i$, the monotonicity of $\tilde{\phi}$, (4.4) and (4.1) gives

$$\begin{aligned} F_i^{\mu(\hat{\lambda})}(\hat{x}) &= \tilde{\phi}(\hat{x}_i - l_i, \zeta, \mu) \\ &\ge \tilde{\phi}(u_i - l_i, 0, \mu) \\ &> \phi(u_i - l_i, 0) - \frac{\mu^2}{2\sqrt{(u_i - l_i)^2}} \\ &= -\frac{\mu^2}{2(u_i - l_i)} \\ &\ge -\frac{1 - \hat{\lambda}}{\hat{\lambda}}(u_i - a_i), \end{aligned}$$

contradicting (4.3). Therefore every point $(\hat{\lambda}, \hat{x})$ on γ_a satisfies $l < \hat{x} < u$. □

Note that if $I_{l,u}$ is empty, then the condition on $\mu(\lambda)$ in the above theorem is achieved by any decreasing C^2 function satisfying $\mu(1) = 0$. If $I_{l,u}$ is not empty, the condition is easily achieved by choosing a deep in the interior of the feasible region $\mathbb{B}_{l,u}$. For example, if $u_i - a_i \ge \frac{1}{2}(u_i - l_i)$ for all $i \in I_{l,u}$, then

$$\mu(\lambda) = \left[\min_{i \in I_{l,u}} (u_i - l_i)\right](1 - \lambda)$$

suffices, since for $0 < \lambda \leq 1$,

$$\begin{aligned}\mu(\lambda)^2 &= \left[\min_{i \in I_{l,u}} (u_i - l_i)\right]^2 (1-\lambda)^2 \\ &\leq 2\left[\min_{i \in I_{l,u}} (u_i - a_i)(u_i - l_i)\right](1-\lambda)^2 \\ &\leq 2\left[\min_{i \in I_{l,u}} (u_i - a_i)(u_i - l_i)\right]\frac{(1-\lambda)}{\lambda}.\end{aligned}$$

The above theorem has two important consequences. First, because γ_a always stays in the feasible region, it is possible to implement the algorithm without ever having to evaluate functions outside of the feasible region. This is important because many applications involve functions that are not defined outside the feasible region. The second consequence is the guarantee that when all bounds are finite, the zero curve γ_a is bounded. The implications of this are stated in the following corollary.

COROLLARY 4.2. *Let ϕ and ϕ_μ be defined by* (2.6) *and* (2.11)*, respectively. Assume that all the bounds of the MCP are finite, choose $\kappa \in (0, \sqrt{2})$ and take*

$$\mu(\lambda) = \kappa\left[\min_i (u_i - l_i)\right](1-\lambda), \tag{4.5}$$

Then for almost all $a \in \operatorname{int} \mathbb{B}_{l,u}$ satisfying $u_i - a_i \geq \kappa^2(u_i - l_i)/2$ for $1 \leq i \leq n$ and ρ_a defined as in Theorem 4.1*, there is a zero curve γ_a of ρ_a emanating from $(0, a)$, along which $\nabla\rho_a(\lambda, x)$ has full rank, that remains in $[0,1) \times (\operatorname{int} \mathbb{B}_{l,u})$ and reaches a point $(1, \bar{x})$, where $\bar{x}$ solves the MCP. γ_a does not intersect itself, is disjoint from any other zeros of ρ_a, and has finite arc length if F is strongly regular at $\bar{x}$.*

Proof. The first four hypotheses of Proposition 2.2 are satisfied trivially. The choice of ϕ_μ, $\mu(\lambda)$, and the restrictions on a suffice to carry out the proof of Theorem 4.1. Hence γ_a remains in $[0,1) \times (\operatorname{int} \mathbb{B}_{l,u})$, and is bounded since $\mathbb{B}_{l,u}$ is bounded. □

The remainder of this section generalizes the above results to other choices of ϕ and ϕ_μ.

THEOREM 4.3. *Let ϕ be a positively oriented NCP function, and let $\tilde{\phi}$ be a C^2-smoother for ϕ, monotone in its first two variables, satisfying*

$$\phi(\alpha, \beta) \geq \tilde{\phi}(\alpha, \beta, \mu), \quad \text{for all } \alpha, \beta \in \overline{\mathbb{R}},\ \mu > 0, \text{ and} \tag{4.6}$$

$$\tilde{\phi}(\alpha, 0, \mu) > -\frac{c\mu^p}{\alpha}, \quad \text{for } \mu > 0,\ 0 < \alpha < \infty, \tag{4.7}$$

where c and p are positive constants. Define ψ by (2.8) *and the smoother $\tilde{\psi}$ by* (2.12)*. Choose $a \in \operatorname{int} \mathbb{B}_{l,u}$, and let $\mu : [0,1] \to \mathbb{R}_+$ be a decreasing C^2 function with $\mu(1) = 0$ satisfying*

$$\mu(\lambda)^p \leq \frac{1-\lambda}{c\lambda}(u_i - a_i)(u_i - l_i) \quad \text{for } i \in I_{l,u},\ \lambda \in (0,1]. \tag{4.8}$$

Define F^μ by (2.14)*, define $\rho_a : [0,1) \times \mathbb{R}^n \to \mathbb{R}^n$ by* (3.2)*, and let γ_a be the connected component of $\rho_a^{-1}(\{0\})$ containing $(0, a)$. Then γ_a is contained in $[0,1) \times (\operatorname{int} \mathbb{B}_{l,u})$.*

Proof. The proof is identical to the proof of Theorem 4.1 except that in place of (4.4), the inequality (4.7) is used. Then by similar arguments using (4.8),

$$\begin{aligned} F_i^{\mu(\hat{\lambda})}(\hat{x}) &> -\frac{c\mu^p}{u_i - l_i} \\ &\geq -\frac{1-\hat{\lambda}}{\hat{\lambda}}(u_i - a_i), \end{aligned}$$

contradicting (4.3). □

COROLLARY 4.4. *Let ϕ, $\tilde{\phi}$, ψ, $\tilde{\psi}$, and F^μ be defined as in Theorem* 4.3. *Assume that all the bounds of the MCP are finite, choose $\kappa \in (0,1)$, and take*

$$\mu(\lambda) = \kappa \left(\frac{1-\lambda}{c}\right)^{1/p} \left[\min_i (u_i - l_i)\right]^{2/p}$$

Then for almost all $a \in \operatorname{int} \mathbb{B}_{l,u}$ satisfying $u_i - a_i \geq \kappa^p (u_i - l_i)$ for $1 \leq i \leq n$ and ρ_a defined as in Theorem 4.1, *there is a zero curve γ_a of ρ_a emanating from $(0, a)$, along which $\nabla \rho_a(\lambda, x)$ has full rank, that remains in $[0,1) \times (\operatorname{int} \mathbb{B}_{l,u})$ and reaches a point $(1, \bar{x})$, where $\bar{x}$ solves the MCP. γ_a does not intersect itself, is disjoint from any other zeros of ρ_a, and has finite arc length if F is strongly regular at $\bar{x}$.*

4.1. Ensuring feasibility. Since some MCP applications involve functions that are not defined outside the feasible region, the algorithm includes an option to ensure that all iterates are feasible. The following discussion assumes that the MCP algorithm is based on the particular choices of ϕ and ϕ_μ given by (2.6) and (2.11).

Feasibility of the path γ_a can be assured by Theorem 4.1, provided that the initial point a and the function $\mu(\lambda)$ are chosen appropriately. The following procedure achieves this while choosing the initial point a near the starting point x^0 provided by the user: define a by

$$a_i := \begin{cases} \operatorname{mid}(l_i + \nu_i, x_i^0, u_i - \nu_i), & \text{if } i \in I_{l,u}, \\ \max(l_i + \nu, x_i^0) & \text{for } u_i = \infty, l_i \text{ finite}, \\ \min(u_i - \nu, x_i^0) & \text{for } l_i = -\infty, u_i \text{ finite}, \\ x_i^0 & \text{if } l_i = -\infty, u_i = \infty, \end{cases} \tag{4.9}$$

where $\nu_i := \kappa_{\min}^2 (u_i - l_i)/2$ for $i \in I_{l,u}$, and $\kappa_{\min} \in (0,1)$ and $\nu > 0$ are constants that ensure the strict feasibility of a. Next, define $\mu(\lambda)$ by (3.1), with α given by

$$\alpha = \begin{cases} \min\left(c, \kappa \left[\min_{i \in I_{l,u}} (u_i - l_i)\right]\right), & \text{if } I_{l,u} \neq \emptyset, \\ c & \text{otherwise}, \end{cases} \tag{4.10}$$

where c is some positive constant, and κ is defined as follows if $I_{l,u} \neq \emptyset$:

$$\kappa := \min_{i \in I_{l,u}} \sqrt{\frac{2(u_i - a_i)}{u_i - l_i}}. \tag{4.11}$$

Note that if $I_{l,u}$ is not empty, this choice of a and κ ensures that $\kappa \geq \kappa_{\min}$ and also that (4.1) is satisfied. Thus, the assumptions of Theorem 4.1 are satisfied, so γ_a remains strictly within the feasible region. Feasibility is maintained by exploiting STEPNX's built-in logic for handling domain violations. Precisely, whenever a STEPNX call to evaluate $F(x)$ produces an infeasible point (either in the prediction

phase or the correction phase), that domain violation is reported to STEPNX. The result is that STEPNX cuts the step size in half (after sanity checks to prevent an infinite loop) and calculates a new predicted point. Since the zero curve is strictly feasible for $\lambda < 1$, eventually (assuming adequate machine precision) a feasible step will be taken.

Finally, to ensure feasibility of the iterates generated in the end game, the generalized damped Newton method in Figure 2.1 is modified according to the general descent framework described in [12]. Specifically, the Newton direction d^k is projected back onto the feasible region to produce the modified direction

$$\tilde{d}^k := \pi_{\mathbb{B}_{l,u}}(x^k + d^k) - x^k.$$

Note that $x^k + \tilde{d}^k$ is feasible. Step 3 in Figure 2.1 is then replaced with the following:

Step 3' If $\theta(x^k + \tilde{d}^k) \leq (1-\sigma)\theta(x^k)$, set $x^{k+1} = x^k + \tilde{d}^k$. Otherwise, take a projected gradient step as follows: let m_k be the smallest nonnegative integer $m \leq m_{max}$ such that

$$\theta(x^k(\alpha^m)) \leq \theta(x^k) - \sigma\nabla\theta(x^k)(x^k - x^k(\alpha^m)), \tag{4.12}$$

where $x^k(t) := \pi_{\mathbb{B}_{l,u}}(x^k - t\nabla\theta(x^k))$. If no such m_k exists, stop; the algorithm failed. Otherwise, set $x^{k+1} = x^k(\alpha^k)$.

Note that for any feasible x^*, $\left\|x^k + \tilde{d}^k - x^*\right\| \leq \left\|x^k + d^k - x^*\right\|$. This ensures, by [12, Theorem 4.5] and Theorem 2.1, that in a neighborhood of a strongly regular solution $\bar{x}$, the iterates generated by the feasible end game strategy described above converge Q-superlinearly to $\bar{x}$.

The projected gradient step in the above algorithm requires that θ be differentiable. This is true when ϕ is the Fischer-Burmeister function (2.6), but is not true in general.

5. Solver implementation and testing. The MCP algorithm described in the previous section was implemented using the Fischer-Burmeister NCP function for ϕ and the smoother defined by (2.11). The nonsmooth Newton's method described in Figure 2.1 was used for the Newton end game. To construct the homotopy mapping defined in (3.2), the parameter a was constructed according to (4.9), with $\kappa_{\min} := 0.1$, and $\nu = 0.0001$. The function $\mu(\lambda)$ was defined by (4.10), with $c = 1.0$, and κ defined by (4.11).

The algorithm was implemented in C with a link to the Fortran 90 subroutine STEPNX from HOMPACK90. The code is interfaced with the GAMS modeling language, enabling it to be tested using the MCPLIB suite of GAMS test problems [11, 4]. All linear algebra was performed using the LUSOL sparse factorization routine [14] from MINOS [17].

Computational results on the MCPLIB problems are shown in Table 5.1. Many of the problems in this test library include multiple runs, which vary the starting point x^0 or other parameters defining the problem. All of the problems were run using default parameter settings, and the number of successes and failures over all runs are reported in the third column of Table 5.1. The notation $m(n)$ means that the problem included $m+n$ runs, and for those, there were m successes and n failures. The default parameters were chosen as follows:

- Curve tracking parameters: `abserr` = `relerr` = 10^{-4}. Maximum step size $h_{\max} = 100,000$. The normal default for this parameter used by HOMPACK90 is $h_{\max} = 1$. However, many problems in the MCPLIB test library

were poorly scaled so had very long zero curves. The large value of $h_{\max}$ was therefore used to allow these curves to be tracked in a reasonable number of iterations. All other curve tracking parameters were the defaults chosen by STEPNX.

- Newton parameters (See Figure 1): $\alpha = \sigma = 0.5$. $m_{\max} = 20$. Maximum number of Newton iterations = 30.
- Stopping criteria: An iterate x^k was considered to solve the problem when $\|F(x^k)\|_\infty \,/\, \left(1 + \|x^k\|_\infty\right) < 10^{-6}$.

In cases where the problem was not solved by the default parameters, the algorithm was restarted using more conservative parameters: `abserr` = `relerr` = 10^{-6}, `dideal` = 0.01, `lideal` = 0.01, `rideal` = 0.005, and $h_{\max} = \max(.1, \texttt{arclen}/100)$, where `arclen` is the arc length of the zero curve calculated using the default parameters. Results from these runs are shown in the fourth column of Table 5.1.

For the problems that were not solved by the conservative settings, the last column of Table 5.1 describes the reason for failure. The notation "∞" indicates that the zero curve appeared to go off to infinity. This behavior is common for problems that do not satisfy the global monotonicity assumption. The notation "lost" indicates that STEPNX was unable to continue tracking the zero curve. This is generally due to a poorly conditioned Jacobian matrix. The notation "r" indicates failure due to exceeding resource limits–either the limit of 5000 homotopy steps, or 1000 CPU seconds. Finally, the notation "v" indicates failure due to domain violations.

While the algorithm failed to solve a number of problems that have been solved by other algorithms, it is encouraging to note that it performed very well on some problems that are generally regarded as very hard. Notable among these are the billups, pgvon105, pgvon106, and simple-ex problems. Thus, the homotopy algorithm should be viewed as an important supplement to other approaches.

It should also be noted that the algorithm solved several problems for which it was not able to track the zero curve all the way to $\lambda = 1$. This occurred for the bert_oc, obstacle, opt_cont* problems. However, for these problems the Newton end-game was able to find the solution.

Except for the cases "v" and "r", the failures are of two types: numerical instability or unbounded homotopy zero curve γ_a. No attempt was made to scale, reformulate, or precondition the test problems, or to tune the tracking parameters for a particular problem. There is little doubt that a concerted pursuit of all these options would have removed all the failures due to numerical instability. The unbounded zero curves are a more fundamental problem, indicating that the default homotopy map (3.2) is inadequate (which is no surprise, since in engineering practice the default map is virtually never used). It is likely that replacing (3.2) by $\lambda F^{\mu(\lambda)}(x) + (1-\lambda)G(a, \lambda, x)$, where G is carefully crafted for each problem, could remove the other failures. This remains the topic of future work.

6. Conclusions. This paper described a probability-one homotopy algorithm for solving nonsmooth systems of equations and complementarity problems. These methods are an extension to nonsmooth equations of the probability-one homotopy methods described in [8, 21, 23, 24] and they are attractive because they are able to solve a qualitatively different class of problems than methods relying on merit functions. This claim is justified both theoretically and computationally. The key to success of the method is the global monotonicity assumption. When this is satisfied, the zero curve is known to lead to a solution. This result is formalized in Theorem 3.2. In the case of complementarity problems, an easily satisfiable condition

TABLE 5.1
MCPLIB Test Problems

Problem Name	size	Default Settings Success(Failure)	Conservative Settings Success(Failure)	Notes
badfree	5	1(0)		
bert_oc	5000	3(1)	3(1)	r
bertsekas	15	5(1)	6(0)	
billups	1	3(0)		
bratu	5625	1(0)		
choi	13	1(0)		
colvdual	20	4(0)		
colvnlp	15	6(0)		
colvtemp	20	4(0)		
cycle	1	1(0)		
degen	2	1(0)		
duopoly	63	0(1)	0(1)	∞
ehl_k40	41	2(1)	3(0)	
ehl_k60	61	2(1)	3(0)	
ehl_k80	81	2(1)	3(0)	
ehl_kost	101	1(2)	1(2)	lost
electric	158	0(1)	0(1)	∞
eta2100	296	0(1)	1(0)	
explcp	16	1(0)		
forcebsm	184	0(1)	0(1)	∞
forcedsa	186	0(1)	0(1)	∞
freebert	15	7(0)		
gafni	5	3(0)		
games	16	25(0)		
hanskoop	14	10(0)		
hydroc06	29	0(1)	0(1)	∞
hydroc20	99	0(1)	0(1)	∞
jel	6	2(0)		
josephy	4	8(0)		
kojshin	4	8(0)		
lincont	419	0(1)	0(1)	∞
mathinum	3	6(0)		
mathisum	4	7(0)		
methan08	31	0(1)	0(1)	∞
multi-v	48	0(3)	0(3)	lost
nash	10	4(0)		
ne-hard	3	1(0)		
obstacle	2500	7(1)	8(0)	

was established, which ensures that the homotopy zero curve always remains strictly feasible. This condition can always be enforced in the algorithm by choosing the initial point a properly. A simple consequence of this result is that for finitely bounded mixed complementarity problems, the zero curve is bounded, and by Proposition 2.2, is guaranteed to lead to a solution.

TABLE 5.1
MCPLIB Test Problems (cont.)

Problem Name	size	Default Settings Success(Failure)	Conservative Settings Success(Failure)	Notes
olg	249	0(1)	0(1)	lost
opt_cont127	4096	1(0)		
opt_cont	288	1(0)		
opt_cont255	8192	1(0)		
opt_cont31	1024	1(0)		
opt_cont511	16384	1(0)		
pgvon105	105	4(0)		
pgvon106	106	5(1)	6(0)	
pies	42	0(1)	1(0)	
powell	16	5(1)	5(1)	∞
powell_mcp	8	6(0)		
qp	4	1(0)		
romer	214	0(2)	0(2)	lost
scarbsum	40	1(1)	2(0)	
scarfanum	13	4(0)		
scarfasum	14	1(3)	1(3)	v
scarfbnum	39	0(2)	2(0)	
scarfbsum	40	1(1)	2(0)	
shubik	30	7(41)	13(35)	r
simple-ex	17	1(0)		
simple-red	13	1(0)		
sppe	27	3(0)		
tinloi	146	10(54)	64(0)	
tobin	42	4(0)		
trade12	600	1(1)	1(1)	lost
trafelas	2376	0(2)	0(2)	r

REFERENCES

[1] S. C. BILLUPS, *Algorithms for Complementarity Problems and Generalized Equations*, PhD thesis, University of Wisconsin–Madison, Madison, Wisconsin, Aug. 1995.

[2] ———, *A homotopy based algorithm for mixed complementarity problems*, UCD/CCM Report No. 124, Department of Mathematics, University of Colorado at Denver, Denver, Colorado, 1998.

[3] ———, *Improving the robustness of descent-based mehtods for semi-smooth equations using proximal perturbations*, Mathematical Programming, 87 (2000), pp. 153–176.

[4] S. C. BILLUPS, S. P. DIRKSE, AND M. C. FERRIS, *A comparison of large scale mixed complementarity problem solvers*, Computational Optimization and Applications, 7 (1997), pp. 3–25.

[5] S. C. BILLUPS, A. L. SPEIGHT, AND L. T. WATSON, *Nonmonotone path following methods for nonsmooth equations and complementarity problems*, in Applications and Algorithms of Complementarity, M. C. Ferris, O. L. Mangasarian, and J.-S. Pang, eds., Kluwer Academic Publishers, forthcoming.

[6] C. BUCK, *Advanced Calculus*, McGraw–Hill, New York, NY, 3rd ed., 1978.

[7] B. CHEN AND X. CHEN, *A global and local superlinear continuation-smoothing method for P_0 + R_0 NCP or monotone NCP*, SIAM Journal on Optimization, 9 (1999), pp. 624–645.

[8] S.-N. CHOW, J. MALLET-PARET, AND J. A. YORKE, *Finding zeros of maps: homotopy methods that are constructive with probability one*, Mathematics of Computation, 32 (1978), pp. 887–899.

[9] T. De Luca, F. Facchinei, and C. Kanzow, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Mathematical Programming, 75 (1996), pp. 407–439.

[10] ———, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems*, Computational Optimization and Applications, forthcoming, (1999).

[11] S. P. Dirkse and M. C. Ferris, *MCPLIB: A collection of nonlinear mixed complementarity problems*, Optimization Methods and Software, 5 (1995), pp. 319–345.

[12] M. C. Ferris, C. Kanzow, and T. S. Munson, *Feasible descent algorithms for mixed complementarity problems*, Mathematical Programming, 86 (1999), pp. 475–497.

[13] A. Fischer, *A special Newton–type optimization method*, Optimization, 24 (1992), pp. 269–284.

[14] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *Maintaining LU factors of a general sparse matrix*, Linear Algebra and Its Applications, 88/89 (1987), pp. 239–270.

[15] C. Kanzow, *Some equation–based methods for the nonlinear complementarity problem*, Optimization Methods and Software, 3 (1994), pp. 327–340.

[16] O. L. Mangasarian, *Equivalence of the complementarity problem to a system of nonlinear equations*, SIAM Journal on Applied Mathematics, 31 (1976), pp. 89–92.

[17] B. A. Murtagh and M. A. Saunders, *MINOS 5.0 user's guide*, Technical Report SOL 83.20, Stanford University, Stanford, California, 1983.

[18] L. Qi, *Regular pseudo-smooth NCP and BVIP functions and globally and quadratically convergent generalized Newton methods for complementarity and variational inequality problems*, Mathematics of Operations Research, 24 (1999), pp. 440–471.

[19] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Mathematical Programming, 58 (1993), pp. 353–368.

[20] L. T. Watson, *An algorithm that is globally convergent with probability one for a class of nonlinear two-point boundary value problems*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 394–401.

[21] ———, *A globally convergent algorithm for computing fixed points of C^2 maps*, Applied Mathematics and Computation, 5 (1979), pp. 297–311.

[22] ———, *Solving the nonlinear complementarity problem by a homotopy method*, SIAM Journal on Control and Optimization, 17 (1979), pp. 36–46.

[23] L. T. Watson, S. C. Billups, and A. P. Morgan, *Algorithm 652: HOMPACK: A suite of codes for globally convergent homotopy algorithms*, ACM Transactions on Mathematical Software, 13 (1987), pp. 281–310.

[24] L. T. Watson, R. C. Melville, A. P. Morgan, and H. F. Walker, *Algorithm 777: HOMPACK90: A suite of FORTRAN 90 codes for globally convergent homotopy algorithms*, ACM Transactions on Mathematical Software, 23 (1997), pp. 514–549.