

Quasi-decidability of a Fragment of the Analytic First-order Theory of Real Numbers*

PETER FRANEK and STEFAN RATSCHAN

Institute of Computer Science, Academy of Sciences of the Czech Republic

PIOTR ZGLICZYNSKI

Jagellonian University in Krakow

January 7, 2012

Abstract

In this paper we consider a fragment of the first-order theory of the real numbers that includes systems of equations of real analytic functions, and for which all functions are computable in the sense that it is possible to compute arbitrarily close piece-wise interval approximations. Even though this fragment is undecidable, we prove that (under the additional assumption that all variables range over closed intervals) there is a (possibly non-terminating) algorithm for checking satisfiability such that (1) whenever it terminates, it computes a correct answer, and (2) it always terminates when the input is robust. A formula is robust, if its satisfiability does not change under small perturbations. As a basic tool for our algorithm we use the notion of degree from the field of (differential) topology.

1 Introduction

It is well known that, while the theory of real numbers with addition and multiplication is decidable [28], any periodic function makes the problem undecidable, since it allows encoding of the integers. Recently, several papers [11, 21, 22, 8] have argued, that in continuous domains (where we have notions of neighborhood, perturbation etc.) such decidability results do not always have much practical relevance. The reason is, that real-world manifestations of abstract mathematical objects in such domains will always be exposed to perturbations (imprecision of production, engineering approximations, unpredictable influences

*This is an extended version of a paper that appeared in the proceedings of the 36th International Symposium on Mathematical Foundations of Computer Science [10]. The work of Stefan Ratschan and Peter Franek was supported by MŠMT project number OC10048 and by the institutional research plan AV0Z100300504.

of the environment etc.). Engineers take these perturbations into account by coming up with *robust* designs, that is, designs that do not change essentially under such perturbations. Hence, in this context, it is sufficient to come up with algorithms that are able to decide such robust problem instances. They are allowed to run forever in non-robust cases, but—since robustness may not be checkable—*must not* return incorrect results, in whatever case. In a recent paper we called problems possessing such an algorithm *quasi-decidable* [23].

In this paper we show quasi-decidability of a certain fragment of the first-order theory of the reals. The basic building blocks are existentially quantified disjunctions of systems of n equalities over n variables and arbitrarily many inequalities. Those blocks may be combined using universal quantifiers, conjunctions, and disjunctions. All variables are assumed to range over closed intervals.

The allowed function symbols include addition, multiplication, exponentiation, and sine. More specifically, they have to be real analytic, and for compact intervals I_1, \dots, I_n , we need to be able to compute an interval $J \supseteq f(I_1, \dots, I_n)$ such that the over-approximation of J over $f(I_1, \dots, I_n)$ can be made arbitrarily small.

The main tool we use is the notion of the *degree of a continuous function* that comes from differential topology [13, 17, 20]. For continuous functions $f : [a, b] \rightarrow \mathbb{R}$, the degree $\deg(f, [a, b], 0)$ is 0 iff $f(a)$ and $f(b)$ have the same sign, otherwise the degree is either -1 or 1 , depending on whether the sign changes from negative to positive or the other way round. Hence, in this case, the degree gives the information given by the intermediate value theorem plus some directional information. For higher dimensional functions, the degree is an integer whose value may be greater than 1, and that generalizes this information to higher dimensions. However, the degree is defined only when the dimensions of the domain and target space of f are equal.

If we can over-approximate the function f arbitrarily precisely on intervals, then the degree is algorithmically computable. Our algorithm for checking the existence of a solution of $f = 0$ consists of over-approximating the connected components of the zero set of f by small neighborhoods and checking the degree of f in those neighborhoods. If for any neighborhood this degree is nonzero, then $f(x) = 0$ has a solution. Otherwise we show that there exists an arbitrarily small perturbation \tilde{f} of f such that $\tilde{f}(x) = 0$ does not have a solution. However, such neighborhoods may not exist for a general continuous or even differentiable function. Therefore, we restrict ourselves to analytic functions. The zero set of an analytic function consists of a finite number of closed connected components and we may take disjoint small neighborhoods around them.

For handling inequalities, universal quantification, conjunction and disjunctions we use interval deductions. We show termination of the resulting algorithm for all robust cases based on robustness properties of the topological degree and convergence properties of the employed interval deductions.

Concerning related work, Collins [7] presents similar result to ours for the special case of systems of n equalities in n variables, formulated in the language of computable analysis [30]. However, the paper unfortunately contains only

very rough sketch proofs, that we were not able to complete into full proofs.

Verification of zeros of systems of equations is a major topic in the interval computation community [19, 24, 15, 12]. However, here people are usually not interested in some form of completeness of their methods, but in usability within numerical solvers for systems of equations or global optimization. Still, our completeness result implies that the algorithm that we use is stronger than the known common algorithms for detecting zeros, like Miranda’s and Borsuk’s theorem.

Since this work applies results from a quite distant field—topology—to automated reasoning, it is not possible to keep the paper self-contained within a reasonable number of pages. Still, we tried to keep the basic material self-contained and to refer to topological results only later in the paper. The necessary topological pre-requisites can be found in standard textbooks [18, e.g.].

In Section 2, we define the notions of a robustness and quasi-decidability, and state the main theorem of the paper. In Section 3, we give the according quasi-decision procedure. Section 4 contains some properties of real analytic functions we need. In Section 5, we present the notion of topological degree, describe its main properties, and give an algorithm for computing it for an \mathbb{R}^n -valued interval-computable function defined on an n -dimensional box. In Section 6, we show the connection between a robust system of n equations in n variables and the topological degree. In particular, we show that such a system $f(x) = 0$ has a robust solution if and only if there exists a neighborhood U of a component of the zero set of f such that $\deg(f, U, 0) \neq 0$. This will be the main ingredient for in Section 7, where we prove the correctness of the algorithm presented in Section 3. In the last section, we compare the algorithm with other known algorithms for deciding satisfiability of a system of equations and we discuss the limitation of our restricting assumption of having not more unknowns than equations.

2 The Main Theorem

Let us assume a class \mathcal{X} of first-order predicate logical formulas and let us fix a certain interpretation for all function and predicate symbols in \mathcal{X} . Moreover, assume a function d that, given two formulas ϕ and ψ in \mathcal{X} , returns a number in $\mathbb{R}^{\geq 0} \cup \{\infty\}$. Intuitively, this function should model the distance between formulas in \mathcal{X} , but—at least for now—we do not require any specific properties.

Definition 1 *Let S be a sentence in \mathcal{X} and $\epsilon > 0$. We say that S is ϵ -robust, if for every sentence S' , $d(S', S) < \epsilon$ implies that S' and S are equi-satisfiable. We say that the sentence S is robust, if there is an $\epsilon > 0$ such that S is ϵ -robust. We say that a sentence S is robustly true, if it is both robust and true (i.e., satisfiable). We say that a sentence S is robustly false, if it is both robust and false (i.e., unsatisfiable).*

Now, consider the following algorithm specification:

Input: A sentence S from \mathcal{X} ,

Output: $a \in \{\mathbf{T}, \mathbf{F}\}$

such that

- if $a = \mathbf{F}$, then the sentence S is false, and
- if $a = \mathbf{T}$, then the sentence S is true,
- if the sentence S is robust then the algorithm terminates.

Definition 2 *Given a class \mathcal{X} of first-order sentences and a function $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\geq 0} \cup \{\infty\}$, \mathcal{X} is quasi-decidable wrt. d iff there exists an algorithm (a quasi-decision procedure) with the above specification.*

We will now concentrate on the real numbers. Our goal is to show quasi-decidability of a non-trivial class of first-order sentences that includes functions such as \sin .

We define a *box* B in \mathbb{R}^n to be the Cartesian product of n closed intervals of finite positive length (i.e., a hyper-rectangle) and a k -dimensional box (or k -box) in \mathbb{R}^n to be a product of k closed intervals of positive finite length and $(n - k)$ constants. The *width* of a box is the maximum of the width of its constituting intervals.

We denote the norm of $x \in \mathbb{R}^n$, by $|x|$ and the norm of a continuous function $f : \Omega \rightarrow \mathbb{R}^n$ by $\|f\| := \sup\{|f(x)|; x \in \Omega\}$ and if Ω is not clear, we use the notation $\|f\|_\Omega$. The solution set $\{f(x) = 0 \mid x \in \Omega\}$ will be denoted simply by $\{f = 0\}$.

For a set $\Omega \subset \mathbb{R}^n$, $\bar{\Omega}$ is its closure, Ω° its interior and $\partial\Omega = \bar{\Omega} \setminus \Omega^\circ$ its boundary with respect to the Euclidean topology. If Ω is a k -box in \mathbb{R}^n , we usually denote $\partial\Omega$ the boundary in the topology of Ω (i.e., union of the $2k$ faces).

Definition 3 *Let $\Omega \subseteq \mathbb{R}^n$. We call a function $f : \Omega \rightarrow \mathbb{R}$ interval-computable iff there exists an algorithm that, for a given box $B \subseteq \Omega$ with rational endpoints, computes a closed (possibly degenerate) interval $I(f)(B)$ such that*

- $I(f)(B) \supseteq \{f(x) \mid x \in B\}$ (i.e., the range of f in B is over-approximated), and
- for every box S , for every $\varepsilon > 0$ there is a $\delta > 0$ such that for every box B with $B \subseteq S$, $\text{diam}(B) < \delta$, for all $y \in I(f)(B)$, there is an $x \in B$ such that $|f(x) - y| \leq \varepsilon$ (i.e., the over-approximation can be made arbitrarily small).

We call a function $f = (f_1, \dots, f_n) : \Omega \rightarrow \mathbb{R}^n$ interval-computable iff each f_i is interval-computable. In this case, the algorithm $I(f)$ returns a tuple of intervals, one for each f_i .

Usually such functions are written in terms of symbolic expressions containing symbols denoting certain basic interval-computable functions such as rational constants, addition, multiplication, exponentiation, and sine. In that case, the first property above follows from the so-called fundamental theorem of interval arithmetic (which can be found in any introductory text on interval computation, for example Formula (5.17) in [25]), and the second property from Lipschitz continuity of interval arithmetic (e.g., Theorem 2.1.1 in Neumaier's book [19]). However, in this paper we do not fix a certain notation and will allow an arbitrary language for denoting interval computable functions. We will use interval computable functions and expressions denoting them interchangeably and assume that for an expression denoting a function f , a corresponding algorithm $I(f)$ is given.

Definition 4 *We will denote by \mathcal{A} the class of all first-order predicate logical formulas such that*

- *all variables are bounded by closed intervals with rational endpoints (for such an interval I , we will introduce those bounds with corresponding quantifiers in the form $\exists x \in I, \forall x \in I$),*
- *all terms denote functions that are continuous, analytic, and interval-computable, and*
- *the allowed predicate symbols are $=, \leq, <$ with their usual interpretation over the real numbers.*

Throughout we will use the convention that logical connectives bind stronger than quantifiers. Moreover, we use brackets to denote Boolean structure of formulas. Sometimes we will use line breaks instead of brackets for this purpose. We will use the symbol \equiv to denote equality of first-order formulas.

Now let us define some notion of distance on \mathcal{A} . First we define the case where it is finite:

Definition 5 *Let F, G be two sentences in the class \mathcal{A} . We say that F and G have the same structure, if they one can be obtained from the other by only exchanging terms (i.e., they have the same Boolean and quantification structure).*

For example the sentences

$$\exists x \in [0, 1] \forall y \in [0, 1] . x^2 - y = xy \wedge x = y$$

and

$$\exists x \in [0, 1] \forall y \in [0, 1] . x^2 - y = xy + 1 \wedge x = y^2$$

have the same structure, because the only difference is in the terms involved.

Definition 6 *Assume two sentences F and G . If they do not have the same structure, then $d(F, G) := \infty$. In the case where they do have the same structure, assume that the sentence F contains terms denoting functions f_1, \dots, f_p and the*

sentence G contains on the corresponding places terms denoting the functions g_1, \dots, g_p . We define the distance

$$d(F, G) := \max_{i \in \{1, \dots, p\}} \|f_i - g_i\|_{\Omega_i},$$

where Ω_i denotes the respective domain of those functions.

For example the sentences defined above have distance equal to 1, because $d(x^2 - y, x^2 - y) = 0$, $d(xy, xy + 1) = 1$, $d(x, x) = 0$ and $d(y, y^2) = 1$ for $x, y \in [0, 1]$.

Further, we will be interested in a subclass \mathcal{B} of sentences in \mathcal{A} . Even though this sub-class is undecidable for unbounded quantifiers [5, 29], and its decidability is unknown for bounded quantifiers, we will give an algorithm that determines, for a given robust sentences in this subclass, whether it is true or not.

Definition 7 Let \mathcal{B} be the smallest subset of \mathcal{A} such that

(a) \mathcal{B} contains all formulas of the form

$$\exists x \in B . [f_1 = 0 \wedge f_2 = 0 \wedge \dots \wedge f_n = 0 \wedge g_1 \geq 0 \wedge g_2 \geq 0 \wedge \dots \wedge g_k \geq 0]$$

where f_i, g_j are functions, B is an m -box (the expression $\exists x \in B$ denoting a block of m existential quantifiers) and either $n \geq m$ or $n = 0$. The integer k may be arbitrary and we also admit $k = 0$ (i.e., the case without inequalities).

(b) Let $I \subset \mathbb{R}$ be a closed bounded interval. If U is in \mathcal{B} , then

$$\forall x \in I . U$$

is also in \mathcal{B} .

(c) If U, V are in \mathcal{B} , then

$$U \wedge V \quad U \vee V$$

are also in \mathcal{B} .

The formulas corresponding to (a) represent systems of equations and inequalities. However, we assume that there are no more existential quantifiers than equations in (a), corresponding to the condition $n \geq m$.

The following sentence is an example of a formula in class \mathcal{B} :

$$\begin{aligned} &\forall x \in [-1, 1] \\ &\quad \exists y \in [-1, 1] \exists z \in [-1, 1] \\ &\quad \quad [x^2 - y^2 - z^2 = 0 \wedge x^3 - y^3 - z^3 = 0]. \end{aligned}$$

The following sentence is an example of a sentence not in \mathcal{B}

$$\exists x \in [0, 1] \exists y \in [0, 1] . x - y = 0$$

because the domain of the particular function is a 2-dimensional box and there is only one equation, so the assumptions in (a) are violated.

Note that if $\exists x F_1$ and $\exists x F_2$ are in the class \mathcal{B} , then $\exists x \in B . (F_1 \vee F_2)$ is equi-satisfiable with the formula $(\exists x \in B . F_1) \vee (\exists x \in B . F_2)$ in \mathcal{B} , and hence every quasi-decision procedure for \mathcal{B} can trivially handle disjunctions within existential quantification, too.

Also, note that the case of strict equalities $\exists x \in B . f = 0 \wedge g > 0$ is robust if and only if $\exists x \in B . f = 0 \wedge g \geq 0$ is robust and they are equi-satisfiable in such case. So, we will restrict ourselves to the case of non-strict inequalities.

Now we state the main theorem of this paper:

Theorem 1 *The class \mathcal{B} is quasi-decidable wrt. the function d .*

3 The Quasi-decision Procedure

In this paper, we construct an algorithm that decides, whether or not a robust sentence in \mathcal{B} is true.

For any formula $U \in \mathcal{B}$, variable x and $x_0 \in \mathbb{R}$ we denote by $U[x \leftarrow x_0]$ the formula derived from U by substituting x_0 for x in every free occurrence of x in U . We also allow x to be an n -tuple of variables, and $x_0 \in \mathbb{R}^n$, in which case $U[x \leftarrow x_0]$ denotes the parallel substitution of entries of x_0 with their corresponding entries of x .

In our algorithms, we use an alternative form of the Cartesian product that concatenates tuples from the argument sets, instead of forming pairs. That is, for sets $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ it produces the set $\{(x_1, \dots, x_n, y_1, \dots, y_m) \mid (x_1, \dots, x_n) \in X, (y_1, \dots, y_m) \in Y\}$. Especially, for the set $\{()\}$ containing the 0-tuple, $\{()\} \times X$ will be X .

For technical reasons, we construct an algorithm for the following, more general, specification:

Input:

- a formula S from \mathcal{B} in l free variables p ,
- a $|p|$ -box P ,
- $r \in \mathbb{R}^{>0}$,

such that the width of P is at most r .

Output: a subset of $\{\mathbf{T}, \mathbf{F}\}$.

with the following two properties:

Correctness: If the algorithm terminates with $\{\mathbf{T}\}$ ($\{\mathbf{F}\}$), then for all $p_0 \in P$, $S[p \leftarrow p_0]$ is robustly true (robustly false).

Definiteness: If for a given box B , either for all $p_0 \in B$ the sentence $S[p \leftarrow p_0]$ is robustly true or for all $p \in B$ the sentence S robustly false, then there

exists an $\epsilon > 0$ such that for every $r \leq \epsilon$ and every sub-box $P \subset B$ with width smaller than r , the algorithm terminates with $\{\mathbf{T}\}$ or $\{\mathbf{F}\}$ (as opposed to $\{\mathbf{T}, \mathbf{F}\}$).

We will denote this algorithm by $\text{CheckSat}(S, P, r)$. The following theorem is clear:

Theorem 2 *Assume that CheckSat fulfills its specification described above. Then the algorithm below is a quasi-decision procedure for \mathcal{B} .*

```

 $\epsilon \leftarrow 1$ 
loop
   $R \leftarrow \text{CheckSat}(S, \{\}, \epsilon)$ 
  if  $|R| = 1$  then
    return  $s$  s.t.  $s \in R$ 
  else
     $\epsilon \leftarrow \epsilon/2$ 

```

Note however, that that the above specification does not only result in a quasi-decision procedure. Moreover, it checks robustness of the input:

Theorem 3 *Assume that CheckSat fulfills its specification described above. Then the quasi-decision procedure of Theorem 2 terminates if and only if the input is robust.*

We will now define the algorithm $\text{CheckSat}(S, P, r)$ in details. We will leave the correctness proof to Section 7.

The algorithm is recursive, following the definition of class \mathcal{B} . We will describe the parts corresponding to the individual cases of this definition. We will prove in Section 7 that these algorithms have the required properties.

The following technical lemma, which holds by induction, using Items (a)–(c) of Definition 7, ensures that we remain in the class \mathcal{B} :

Lemma 1 *Let U be a formula in \mathcal{B} where x occurs only freely and $x_0 \in \mathbb{R}$. Then $U[x \leftarrow x_0]$ is also in \mathcal{B} .*

In all algorithms we work with intervals with rational endpoints. In particular, for any $\epsilon > 0$, we are able to split a given box B into a grid of sub-boxes of side-width smaller than ϵ .

3.1 System of Equations and Inequalities

We first consider the case (a) of Definition 7, that is, a formula S of the form

$$\exists x \in B . [f_1 = 0 \wedge \dots \wedge f_n = 0 \wedge g_1 \geq 0 \wedge \dots \wedge g_k \geq 0]$$

where B is an m -box. In an abuse of notion we also use f_1, \dots, f_n and g_1, \dots, g_k for the functions denoted by those expressions. In the first case they are functions in $\mathbb{R}^{l+m} \rightarrow \mathbb{R}$ and in the second case functions in $\mathbb{R}^{l+m} \rightarrow \mathbb{R}$, where l is the number of free variables of the formula S . We assume that the order of the arguments of those functions is the same as the order in which the respective variables are quantified in the overall formula. Finally, we denote by $f : \mathbb{R}^{l+m} \rightarrow \mathbb{R}^n$ the function defined by the components (f_1, \dots, f_n) and by $g : \mathbb{R}^{l+m} \rightarrow \mathbb{R}^k$ the function defined by the components (g_1, \dots, g_k) .

In the algorithm, we use the notion of the degree from the field of differential topology [13, 17, 20]. For a smooth function $f : \Omega \rightarrow \mathbb{R}^n$ where Ω is an n -dimensional manifold or an n -box and $p \notin f(\partial\Omega)$, the degree of f with respect to Ω and a point $p \in \mathbb{R}^n$ is an integer denoted by $\deg(f, \Omega, p)$. More details on the degree are given in Section 5 below, where we show that the degree is algorithmically computable for interval-computable analytic functions $f : B \rightarrow \mathbb{R}^n$, with B being an n -dimensional box.

The algorithm looks as follows:

```

Algorithm SoEI( $S, P, r$ )           // System of equations and inequalities
Let  $S_r$  by a grid of boxes in  $B$  of width at most  $r$ .
if for every box  $A \in S_r$ 
    there is no  $y_f \in I(f)(P \times A)$ ,  $y_g \in I(g)(P \times A)$  s.t.  $y_f = 0$  and  $y_g \geq 0$  then
        return {F} //  $f = 0 \wedge g \geq 0$  has no solution
if m=n then
    Merge all boxes in  $S_r$  containing a common face  $C$  s.t.  $0 \in I(f)(P \times C)$ .
    Remove all boxes in  $S_r$  containing a face  $C$  s.t.  $C \subset \partial B$  and  $0 \in I(f)(P \times C)$ .
    Let  $p_0$  be an arbitrary element of  $P$ 
    for each grid element  $A$ 
        if  $\deg(f(p_0), A, 0) \neq 0$  then // equations hold, so check inequalities
            let  $S_r(A)$  be a grid of boxes in  $A$  of width at most  $r$ 
            if for all  $E \in S_r(A)$ , for all  $y_g \in I(g)(P \times E)$ ,  $y_g > 0$  then
                return {T}
    return {T, F}
return {T, F}

```

Here we suppose that f is present in the formula (i.e., $n > 0$). The algorithm can be easily adapted to the case, where it is not. The checks for positivity/negativity etc. of all box elements can be easily implemented by checking the end-points of the intervals defining the box.

3.2 Universal Quantifiers

The recursive call corresponding to Case (b) of Definition 7 looks as follows:

Algorithm Univ($\forall x \in I . S, P, r$):

Let I_r be a grid of sub-intervals I of width at most r
Let $(R_1, \dots, R_{|I_r|})$ be the $|I_r|$ -tuple $(\text{CheckSat}(S, P \times I', r) \mid I' \in I_r)$
return $\{t_1 \wedge \dots \wedge t_{|I_r|} \mid t_1 \in R_1 \dots t_{|I_r|} \in R_{|I_r|}\}$

Here, in the return statement, the symbol \wedge denotes the Boolean function corresponding to conjunction. See Section 8.1 on a discussion of why the dual approach to this algorithm does not work for existential quantification.

3.3 Conjunctions and disjunctions

Finally, the recursive call corresponding to Case (c) of Definition 7 looks as follows:

Algorithm $\text{Conj}(S \wedge T, P, r)$

return $\{u \wedge v \mid u \in \text{CheckSat}(S, P_1, r), v \in \text{CheckSat}(T, P_2, r)\}$
where P_1 (P_2) is the projection of P
to the free variables of S (T , respectively).

Here, in the return statement, the symbol \wedge denotes the Boolean function corresponding to conjunction. The algorithm for disjunction is completely analogous, replacing \wedge with \vee .

4 Zeros of Analytic Functions

A function $f : \Omega \rightarrow \mathbb{R}^n$ defined on an open set Ω is analytic, iff each of its components f_1, \dots, f_n is an analytic function, that is, iff for each $i \in \{1, \dots, n\}$, for each point $x_0 \in \Omega$, there exists a neighborhood U of x_0 and a sequence of numbers $\{c_j\}_{j \in \mathbb{N}_0}$ such that $f_i = \sum_{j \in \mathbb{N}_0} c_j (x - x_0)^j$ on U .

We will call a function with closed domain $f : \bar{\Omega} \rightarrow \mathbb{R}^n$ analytic, if it is a restriction of an analytic function $\tilde{f} : U \rightarrow \mathbb{R}^n$ for some open set U such that $\Omega \subseteq U$.

The set of analytic functions is closed with respect to addition, multiplication, division by nonzero function, composition and differentiation.

We will need the following statement later:

Theorem 4 *For an open, connected and bounded set Ω , an analytic function $f : \bar{\Omega} \rightarrow \mathbb{R}^n$, the set $\{f = 0\}$ consists of a finite number of connected components.*

Proof. It follows from Lojasiewicz's theorem [16] that the zero set of a real valued analytic functions is locally a union of a finite number of manifolds of various dimensions. So, the zero set of a real valued analytic function defined on a compact set $\bar{\Omega}$ has a finite number of connected components and the set $\{f = 0\}$ coincides with the zero set of the real valued analytic function $\sum_i f_i^2$ on $\bar{\Omega}$. ■

An analogous statement for smooth (but not analytic) functions f does not hold. One can easily construct a smooth function $f : [0, 1] \rightarrow \mathbb{R}$ such that $\{f = 0\}$ is the Cantor set which is totally disconnected.

5 Degree of a Continuous Function

In this section we describe some basic properties of the degree, and show, how it can be computed. We have already mentioned in the introduction that in the one-dimensional case, the degree captures the information provided by the intermediate value theorem.

Let $\Omega \subset \mathbb{R}^n$ be open and bounded, $f : \bar{\Omega} \rightarrow \mathbb{R}^n$ continuous and smooth (i.e., infinitely often differentiable) in Ω , $p \notin f(\partial\Omega)$. For regular values $p \in \mathbb{R}^n$ (i.e., values p such that for all y with $f(y) = p$, $\det f'(y) \neq 0$), a generalization of the directional information used in the one-dimensional case, is the sign of the determinant $\det f'(y)$. Adding up those signs results in the explicit definition [17] of $\deg(f, \Omega, p)$ by

$$\deg(f, \Omega, p) := \sum_{y \in f^{-1}(p)} \text{sign } \det f'(y).$$

See standard textbooks for a generalization to non-regular values [17]. Here we give an alternative, axiomatic definition, that can be shown to be unique. In this approach $\deg(f, \Omega, p)$ is the unique integer satisfying the following properties [9, 20, e.g.]:

1. For the identity function I , $\deg(I, \Omega, p) = 1$ iff $p \in \Omega$
2. If $\deg(f, \Omega, p) \neq 0$ then $f(x) = p$ has a solution in Ω
3. If there is a continuous function (a ‘‘homotopy’’) $h : [0, 1] \times \bar{\Omega} \rightarrow \mathbb{R}^n$ such that $h(0) = f$, $h(1) = g$ and $p \notin h(t, \partial\Omega)$ for all t , then $\deg(f, \Omega, p) = \deg(g, \Omega, p)$
4. If $\Omega_1 \cap \Omega_2 = \emptyset$ and $p \notin f(\partial\Omega_1 \cup \partial\Omega_2)$, then $\deg(f, \Omega_1 \cup \Omega_2, p) = \deg(f, \Omega_1, p) + \deg(f, \Omega_2, p)$
5. $\deg(f, \Omega, p)$, as a function of p , is constant on any connected component of $\mathbb{R}^n \setminus f(\partial\Omega)$.

This can be extended to the case where Ω has dimension n but is embedded into some higher-dimensional space (in geometric terms, f is a differentiable function between two compact oriented manifolds of the same dimensions). For example, if f is a function from a segment c of a curve (i.e., a set of dimension 1) in \mathbb{R}^k to another segment of a curve in \mathbb{R}^k , and if $f \neq 0$ on the endpoints of c , then $\deg(f, c, 0)$ is well-defined.

The literature provides several articles [27, 3, 1, e.g.] that claim to provide an algorithm that automatically computes the topological degree. However, they mostly just contain general recipes for which the precise input/output

specifications are unavailable, that contain real-number operations for which it is not clear how to implement them on computers, or whose correctness relies on unknown Lipschitz constants. In order to clarify the situation, we give an algorithm here that is based on ideas readily available in the literature, but that does not have those deficiencies.

The algorithm is based on a theorem that recursively reduces the computation of the degree wrt. to a k -dimensional box to the computation of the degree wrt. to a $(k - 1)$ -dimensional box. The theorem uses the notion of orientation that has a specific meaning in differential topology [13, 17, 20]. In order to make the material digestible to a more general audience, and in order to demonstrate algorithmic implementability, we describe here an equivalent, but simpler formalization for the special case of boxes (instead of general manifolds):

We define the orientation of a box in \mathbb{R}^n to be a sign $s \in \{1, -1\}$. Let us consider a k -dimensional box B with orientation s . Observe that we can obtain faces of B by replacing one of the intervals constituting B by either its lower or upper bound (the resulting face is a $(k - 1)$ -dimensional box). Assume that this interval is the r -th (non-constant) interval of B (so $r \in \{1, \dots, k\}$). Then, if the face results from taking a lower bound, we define the *induced* orientation¹ of the face to be $(-1)^r s$, if it results from taking an upper bound, the orientation is $(-1)^{r+1} s$.

Let D be a finite union of oriented k -boxes. The orientation of a union of oriented boxes is, for our purposes, just the information about the orientation of each box in the union. The induced orientation of ∂D is the set ∂D , consisting of $(k - 1)$ -dimensional boxes with orientations induced from the boxes in D .

The algorithm for computing the degree is based on the following theorem that recursively reduces the degree on a box to the degree on boxes of one dimension less.

Theorem 5 *Let B be an oriented finite union of n -dimensional boxes with connected interior. Let $f : B \rightarrow \mathbb{R}^n$ be continuous such that $f \neq 0$ on the boundary of B . Let D_1, \dots, D_k be disjoint subsets of the boundary of B such that D_i is a finite union of $(n - 1)$ -dimensional boxes and the interior of D_i in ∂B is connected. We denote the boundary of D_i in the topology of ∂B by ∂D_i . The orientation of B induces an orientation of each D_i , $i \in \{1, \dots, k\}$.*

For $r \in \{1, \dots, k\}$ we denote by f_r the r -th component of f and $f_{-r} := (f_1, \dots, f_{r-1}, f_{r+1}, \dots, f_n)$.

Now assume that there exists an $r \in \{1, \dots, k\}$, and $s \in \{-1, 1\}$ such that

- *for all $i \in \{1, \dots, k\}$, f_r has constant sign s in D_i ,*
- *$\bigcup_{i \in \{1, \dots, k\}} D_i$ contains all zeros from ∂D of f_{-r} for which f_r has sign s , and*

¹The attentive reader might find it arbitrary that the induced orientation depends on whether the index r is even or odd. In fact, permutations of coordinates will change the definition, but as long as the permutation is even, this results in a definition that is equivalent for our purposes.

- for all $i \in \{1, \dots, k\}$, $0 \notin f_{-r}(\partial D_i)$

Then

$$\deg(f, B, 0) = (-1)^{r-1} s \sum_{i \in \{1, \dots, k\}} \deg(f_{-r}, D_i, 0).$$

Finally, for a one dimensional closed, connected union of oriented boxes D^1 of B with left face l and right face r (according to orientation)

$$\deg(f, D^1, 0) = \begin{cases} 1 & \text{if } f(l) < 0 < f(r) \\ -1 & \text{if } f(r) < 0 < f(l) \\ 0 & \text{if } f(l)f(r) > 0 \end{cases}$$

Here, the notion of left/right face is defined based on the observation that one-dimensional boxes have two faces (that correspond to points) with opposite induced orientation. We define the *left* face to be the face with the opposite induced orientation as the original box, and the *right* face to be the face with the same induced orientation.

When starting the recursion in the above theorem with an n -dimensional box in \mathbb{R}^n of orientation 1, if in the base case D^1 consists of more than one box, then every left face of a box in D^1 is either a boundary point, or the right face of another box. This makes the notion of a left face l and right face r of D^1 well-defined.

The theorem follows directly from Kearfott [14, Theorem 2.2], which again is a direct consequence of some results of Stenger [27].

Now the algorithm just recursively reduces the computation of the topological degree in dimension n to lower dimension using the above theorem. If f is analytic, $\{f_r = 0\}$ and $\{f_{-r} = 0\}$ have a finite number of connected components (see Section 4) and the sets D_i can be found by computing an increasingly fine decomposition of the boundary of B into boxes, and checking the necessary conditions using interval arithmetic. Due to the second property in Definition 3 this procedure will eventually approximate f on the boundary of B arbitrarily closely, and hence it will eventually find such a decomposition.

Note that if we would suppose that f is only continuous (not analytic), the zero set of f_r could have an infinite number of connected components and it may be impossible to find a finite number of D_i such that $\cup D_i$ contain all zeros of f_{-r} and f_r has constant sign on each D_i .

6 Degree and Robustness

In this section, we will clarify the connection between solutions of a function f from the closure of an open set $\Omega \subset \mathbb{R}^n$ to \mathbb{R}^n , robustness, and the degree of f .

In the case of a formula $S \equiv (\exists x \in B f(x) = 0)$, where f is a function from an m -box to \mathbb{R}^n , we say that f has a *robust zero in B* if the sentence S is robustly true. First we prove that a function from an m -dimensional box to a higher-dimensional space \mathbb{R}^n ($n > m$) never has a robust zero. Then we prove

that if $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $\deg(f, \Omega, 0) \neq 0$, then f has a robust zero in Ω . In the rest of the section we prove a partial converse of this. We will show that if the degree is zero and the set of solution $f = 0$ is connected, then f does *not* have a robust zero in Ω . Finally, we will show that if $f = 0$ has a robust solution in Ω , then there exists an open set $U \subset \Omega$ such that $0 \notin f(\partial U)$ and $\deg(f, U, 0) \neq 0$.

In this section, we will always assume that Ω is an open connected bounded set and that it is the interior of $\bar{\Omega}$ (to avoid degenerate cases like $\Omega = [-1, 1] \setminus \{0\}$). We will call the closure $\bar{\Omega}$ of an open connected bounded set a *closed region*.

Theorem 6 *Let $\bar{\Omega}$ be a closed region in \mathbb{R}^m , $n > m$ and $f : \Omega \rightarrow \mathbb{R}^n$ be continuous. Then the sentence $S \equiv (\exists x \in \Omega f = 0)$ is not robustly true.*

Proof. Assume that S is ϵ -robust and true. It follows from the Stone-Weierstrass theorem that the continuous function f may be approximated arbitrarily close with a smooth function \tilde{f} (even with a polynomial). If \tilde{f} has a robust zero, then $\tilde{f}(\Omega)$ contains an open neighborhood of $0 \in \mathbb{R}^n$. However, all values in $\tilde{f}(\Omega)$ are critical values and it follows from Sard's theorem that the set of critical values has zero measure in \mathbb{R}^n , so it cannot contain a neighborhood of $0 \in \mathbb{R}^n$.

The same argument is valid if Ω is an m -dimensional manifold and $f : \Omega \rightarrow \mathbb{R}^n$.

Further, we will consider the case of equal dimension of the domain and range of f .

Theorem 7 *Let $\bar{\Omega} \subset \mathbb{R}^n$ be a closed region, $f : \bar{\Omega} \rightarrow \mathbb{R}^n$ be continuous and smooth on Ω , $0 \notin f(\partial\Omega)$ and let $\deg(f, \Omega, 0) \neq 0$. Then f has a robust zero in Ω .*

Proof. Let $\epsilon < \min_{x \in \partial\Omega} |f|$. For any g such that $\|g - f\|_{\bar{\Omega}} < \epsilon$, we define a homotopy $h(t, x) = tf(x) + (1-t)g(x)$ between f and g . We see that for $x \in \partial\Omega$ and $t \in [0, 1]$,

$$|h(t, x)| = |tf(x) + (1-t)g(x)| = |f(x) + (1-t)(g(x) - f(x))| \geq |f(x)| - \epsilon > 0$$

so that $h(t, x) \neq 0$ for $x \in \partial\Omega$. From Section 5, Properties 2 and 3, we see that $g(x) = 0$ has a solution. ■

We will now consider the case when the degree is zero.

Lemma 2 *Let B be homeomorphic to an n -dimensional ball, $f : B \rightarrow \mathbb{R}^n$ be continuous, nowhere zero on ∂B and let $\deg(f, B, 0) = 0$. Then there exists a continuous nowhere zero function $g : B \rightarrow \mathbb{R}^n$ such that $g = f$ on ∂B and $\|g\| \leq \|f\|$.*

Proof. We may assume, without loss of generality, that $B = \{x \in \mathbb{R}^n; |x| \leq 1\}$ and $\partial B = S^{n-1}$ is the $(n-1)$ -sphere. Let $R : \mathbb{R}^n \setminus \{0\} \rightarrow S^{n-1}$ be defined

by $R(x) := x/|x|$. The degree $\deg(f, B, 0) = 0$ is equal to the degree of the function $R \circ f|_{S^{n-1}} : S^{n-1} \rightarrow S^{n-1}$. The Hopf theorem [17, pp. 51]) states that the degree classifies continuous self-functions of a sphere up to homotopy. So, $R \circ f|_{S^{n-1}}$ is homotopy equivalent to a constant map. So, there exists a homotopy $F : [0, 1] \times S^{n-1} \rightarrow S^{n-1}$ such that $F(1, x) = R \circ f(x)$ and $F(0, x) = c \in S^{n-1}$. Let $r \in [0, 1]$ and $x \in S^{n-1}$. Define the function $g : B \rightarrow \mathbb{R}^n$ by

$$g(rx) = F(r, x)(r|f(x)| + (1-r)\|f\|).$$

This function is continuous, nowhere zero and well defined because $g(0x) = g(0) = c\|f\|$ is independent of $x \in S^{n-1}$. Clearly, $g(x) = f(x)$ on S^{n-1} and for any $x \in S^{n-1}$ and $r \in [0, 1]$, $|g(rx)| \leq r|f(x)| + (1-r)\|f\| \leq \|f\|$. ■

Further, we will need the following technical lemma:

Lemma 3 *Let $\bar{\Omega} \subset \mathbb{R}^n$ be a closed region, $f : \bar{\Omega} \rightarrow \mathbb{R}^n$ continuous and smooth on Ω , $0 \notin f(\partial\Omega)$. Then there exists a continuous function $\tilde{f} : \bar{\Omega} \rightarrow \mathbb{R}^n$, smooth on Ω , with the following properties:*

1. $\tilde{f} = f$ on $\partial\Omega$,
2. 0 is a regular value of \tilde{f} ,
3. $\|\tilde{f}\| \leq 2\|f\|$,
4. \tilde{f} is homotopy equivalent to f through a homotopy $h(t)$ such that $0 \notin h(t)(\partial\Omega)$.

Proof. Let U be an open neighborhood of $\partial\Omega$ in $\bar{\Omega}$ such that $f \neq 0$ on U and $\min\{|f(x)|; x \in \bar{U}\} = \epsilon > 0$. From Sard's theorem [17], there exists a regular value x_0 of f with $|x_0| < \epsilon/2$. It follows that 0 is a regular value of the function $f(x) - x_0$. Let ϕ be a smooth function supported in \bar{U} such that $\phi = 1$ on $\partial\Omega$ and $\phi = 0$ on $\bar{\Omega} \setminus U$. Define $\tilde{f}(x) = f(x) - (1 - \phi(x))x_0$. Clearly, $\tilde{f} = f$ on $\partial\Omega$. The function \tilde{f} is nowhere zero on \bar{U} , because $|\tilde{f}(x)| \geq |f(x)| - |x_0| \geq \epsilon/2$ on \bar{U} . On $\bar{\Omega} \setminus \bar{U}$, $\tilde{f}(x) = f(x) - x_0$. In particular, 0 is a regular value of \tilde{f} and $\|\tilde{f}\| \leq \|f\| + |x_0| \leq 2\|f\|$. Finally, a homotopy between f and \tilde{f} may be given by $h(t, x) = f(x) + (1-t)\tilde{f}(x)$. ■

We have seen in Lemma 2 that if the degree of f on a ball is zero, then we may define a nowhere zero function on the ball that coincides with f on the boundary. We will now see that this is true not only for a ball, but for any closed region $\bar{\Omega} \subset \mathbb{R}^n$.

Lemma 4 *Let $\bar{\Omega}$ be a closed region in \mathbb{R}^n , $f : \bar{\Omega} \rightarrow \mathbb{R}^n$ continuous and smooth on Ω , $0 \notin f(\partial\Omega)$ and $\deg(f, \Omega, 0) = 0$. Then there exists a continuous nowhere zero function $g : \bar{\Omega} \rightarrow \mathbb{R}^n$ such that $g = f$ on $\partial\Omega$ and $\|g\|_{\bar{\Omega}} \leq 2\|f\|_{\bar{\Omega}}$.*

Proof. If the dimension $n = 1$, the function f is defined on an interval $[a, b]$ and the degree assumption implies that $f(a)$ and $f(b)$ have equal signs. So, we

may define g to be a linear function coinciding with f on a and b and the lemma is proved.

Now consider the case $n \geq 2$. From Lemma 3, we construct a continuous function $\tilde{f} : \bar{\Omega} \rightarrow \mathbb{R}^n$ smooth on Ω , $\tilde{f} = f$ on $\partial\Omega$, $\|\tilde{f}\| \leq 2\|f\|$, having 0 as a regular value, homotopy equivalent to f . In particular, $\deg(f, \Omega, 0) = 0$.

The preimage $\tilde{f}^{-1}(0)$ is discrete, because 0 is a regular value and the compactness of $\bar{\Omega}$ implies that $\tilde{f}^{-1}(0)$ is finite. Because $\deg(\tilde{f}, \Omega, 0) = 0$, we may enumerate the points in $\tilde{f}^{-1}(0)$ as $\{x_1, \dots, x_{2m}\}$ so that \tilde{f} is orientation-preserving in the neighborhoods of x_1, x_2, \dots, x_m and orientation-reversing in the neighborhoods of x_{m+1}, \dots, x_{2m} .

Choose m smooth, pairwise disjoint, non-self-intersecting curves c_i in Ω connecting x_i and x_{m+i} . This is possible, because the dimension $n \geq 2$ and the complement of a smooth non-self-intersecting curve in an open connected set $\Omega \subset \mathbb{R}^n$ is still open and connected. For these smooth curves, there exist disjoint neighborhoods homeomorphic to balls B_1, \dots, B_m (see e.g. [17, Product neighborhood theorem]). Because $\deg(\tilde{f}, B_i, 0) = 1 - 1 = 0$, we may apply Lemma 2 to construct nowhere zero functions $g_i : \bar{B}_i \rightarrow \mathbb{R}^n$ such that $g_i = \tilde{f}$ on ∂B_i and $|g_i(x)| \leq \|\tilde{f}\|_{\bar{B}_i} \leq 2\|f\|_{\bar{\Omega}}$. The resulting function $g(x)$ defined by $g = g_i$ on B_i and \tilde{f} elsewhere is continuous and has the properties required. ■

We now prove a partial converse of Theorem 7.

Theorem 8 *Let $\bar{\Omega} \subset \mathbb{R}^n$ be a closed region, $f : \bar{\Omega} \rightarrow \mathbb{R}^n$ continuous and smooth on Ω , $0 \notin f(\partial\Omega)$. Let f have a robust zero in Ω and assume that the set $\{f = 0\} \subset \Omega$ is connected. Then $\deg(f, \Omega, 0) \neq 0$.*

Proof. Let $\epsilon > 0$. Since $\{f = 0\}$ is connected, it is contained in a single connected component Ω' of the open set $\{x; |f(x)| < \epsilon\}$. Let $\deg(f, \Omega, 0) = 0$. Applying Lemma 3 to the set Ω' , we construct a continuous function $\tilde{f} : \bar{\Omega}' \rightarrow \mathbb{R}^n$ smooth on Ω' , homotopy equivalent to $f : \bar{\Omega}' \rightarrow \mathbb{R}^n$, $\tilde{f} = f$ on $\partial\Omega'$, having 0 as a regular value and $\|\tilde{f}\| \leq 2\|f\|_{\bar{\Omega}'} \leq 2\epsilon$. Because the set $\{f = 0\}$ is connected and contained in Ω' , we obtain that $\deg(f, \Omega, 0) = \deg(f, \Omega', 0) = \deg(\tilde{f}, \Omega', 0) = 0$. Using Lemma 4, we obtain a continuous function $g : \bar{\Omega}' \rightarrow \mathbb{R}^n$ such that $g = \tilde{f}$ on $\partial\Omega'$, $g \neq 0$ on $\bar{\Omega}'$ and $|g| \leq 2\|\tilde{f}\| \leq 4\epsilon$ on $\bar{\Omega}'$. Extending g to all $\bar{\Omega}$ by $g = f$ on $\bar{\Omega} \setminus \Omega'$, we obtain an everywhere nonzero continuous function g such that $\|g - f\| \leq 5\epsilon$. This can be done for any ϵ and it follows that f has no robust zero in Ω . ■

Using the last theorem, we see that even if 0 is in the interior of f , 0 does not have to be a robust zero of f . Let Ω be an open unit ball in \mathbb{R}^2 . Let (r, φ) be polar coordinates on Ω , i.e. $0 \leq r < 1$ and $\varphi \in \mathbb{R}$. Define the function $f : \Omega \rightarrow \mathbb{R}^2$ in polar coordinates by $f(r, \varphi) = (r, 4 \sin \varphi)$. This map preserves the diameter, and restricted to any circle of diameter r , the image is the whole circle, because the angle φ ranges from -4 to 4 , covering an interval larger than 2π . So, the image of f is the whole unit ball Ω and 0 is in the interior of $f(\Omega)$. However, the degree $\deg(f, \Omega, 0)$ is clearly zero (the boundary map from the unit circle to itself has zero winding number), so 0 is not a robust zero of f .

In the last part of this section, we will show that if C is a component of the zero set $\{f = 0\}$ that has a non-empty intersection with the boundary $\partial\Omega$, an arbitrary small perturbation of f exists such that this component vanishes. First, we need a technical lemma.

Lemma 5 *Let U be an open set in \mathbb{R}^n , $f : U \rightarrow \mathbb{R}^n$ be continuous. Let N be a neighborhood of $x^0 \in U$ such that $0 \notin f(N)$ and let $k \in \mathbb{N}$. Then there exists a function f_1 such that the following conditions are satisfied:*

- (1) $f_1 = f$ on $U \setminus N$
- (2) $\|f_1\| \leq \|f\|$
- (3) 0 is a regular value of $f_1|_N$
- (4) N contains $2k$ points $x_1, \dots, x_k, y_1, \dots, y_k$ such that $f_1(x_i) = f_1(y_i) = 0$, f_1 is orientation-preserving in the neighborhood of x_i and orientation-reversing in the neighborhood of y_i .

Proof. Choose $\delta > 0$ such that $x^0 + [-2\delta, 2\delta]^n \subseteq N$. We construct f_1 such that $f_1(x) = f(x)$ for $x \notin (x^0 + [-2\delta, 2\delta]^n)$. For $x \in (x^0 + [-\delta, \delta]^n)$ we set

$$(f_1)_i(x) = \left(\frac{|x_i - x_i^0|}{\delta} - \frac{1}{2} \right) f_i(x^0).$$

It is easy to see that $f_1^{-1}(0)$ contains in $(x^0 + [\pm\delta])^{2^n}$ points of the form $(x_1^0 \pm \delta/2, x_2^0 \pm \delta/2, \dots, x_n^0 \pm \delta/2)$, half of them preserve orientation and half reverse orientation. Clearly, $|f_1(x)| \leq |f(x^0)| \leq \|f\|$ on $x_0 + [\pm\delta]$. Because $\deg(f_1, x^0 + [\pm\delta], 0) = \deg(f, x^0 + [\pm 2\delta]) = 0$, it is easy to see that f_1 may be extended to $x^0 + [\pm 2\delta]$ so that $f_1 = f$ on $\partial(x^0 + [\pm 2\delta])$, f_1 is nonzero in $x_0 + [\pm 2\delta] \setminus (x_0 + [\pm\delta])$ and the norm $\|f_1\| \leq \|f\|$. The only zero points of f_1 in N are $(x_1^0 \pm \delta/2, \dots, x_n^0 \pm \delta/2)$, so 0 is a regular value of $f_1|_N$. The details are left to the reader.

To produce more zeros we can choose any point $x_1 \in N$ s.t. $f_1(x_1) \neq 0$ and a small neighborhood of x_1 in N where f_1 is nonzero and continue in the same way. ■

Lemma 6 *Let $\bar{\Omega}$ be a closed region with interior Ω , $f : \bar{\Omega} \rightarrow \mathbb{R}^n$ be an analytic function and C be a connected component of $\{f = 0\}$ s.t. $C \cap \partial\Omega \neq \emptyset$. Then for every $\epsilon > 0$, there exists a neighborhood U of C and a continuous function $\tilde{f} : \bar{\Omega} \rightarrow \mathbb{R}^n$ such that*

- (1) $U \cap \{f = 0\} = C$
- (2) $|\tilde{f} - f| < \epsilon$
- (3) $\tilde{f} \neq 0$ on U
- (4) $\tilde{f} = f$ on $\bar{\Omega} \setminus U$.

In other words, if a component C of the zero set of f crosses the boundary $\partial\Omega$, then an arbitrary small perturbation of f exists such that this zero component vanishes.

Proof. Due to our definition of analyticity of a function with closed domain, we know that we can extend f to an analytic function whose domain is open

and contains $\bar{\Omega}$. In the proof we will possibly modify f outside of $\bar{\Omega}$ close to the boundary. Let D be a bounded open set in \mathbb{R}^n such that $\bar{\Omega} \subset D$.

We know from the analyticity of f that $\{f = 0\}$ has a finite number of connected components. Let ϵ_1 be so small that the component U of $\{|f| < \epsilon_1\} \cap \bar{\Omega}$ containing C satisfies $U \cap \{f = 0\} = C$. Clearly, $\partial U \cap \{f = 0\} \subset \partial\Omega$. Consider a map $f_1(x) = f(x) - x_0$ for small x_0 . By Sard's theorem we can find $x_0 < \epsilon_1/2$ such that 0 is a regular value for $f_1 : D \rightarrow \mathbb{R}^n$. Let U_2 be an open neighborhood of U in D such that $|f(x)| < \epsilon_1$ for $x \in U_2$. In particular, $U_2 \cap \bar{\Omega} = U$. We can modify f_1 so that $f = f_1$ on $D \setminus U_2$, $|f_1| < 2\epsilon_1$ on U_2 , $|f_1 - f| < \epsilon_1$ and 0 is still a regular value of f_1 , when restricted to U_2 . Let us split $f_1^{-1}(0) \cap U$ into A_+ and A_0 , where $A_+(f_1, U) = \{x \in f_1^{-1}(0) \cap U; \det(f_1'(x)) > 0\}$ and $A_-(f_1, U) = \{x \in f_1^{-1}(0) \cap U; \det(f_1'(x)) < 0\}$. In the proof of Lemma 4 it was shown how to remove a pair of zeros $(x_+, x_-) \in A_+(f_1, U) \times A_-(f_1, U)$ with small perturbation of f_1 . Hence if $|A_+| = |A_-|$ then we can remove all zeros and the resulting function \tilde{f} will be nonzero on U , satisfy $|\tilde{f} - f_1| < 2\epsilon_1$ and equal to f on $\bar{\Omega} \setminus U$.

When $|A_+| \neq |A_-|$, we can use Lemma 5 to modify f_1 in $U_2 \setminus U$ (hence outside of $\bar{\Omega}$), to create a map f_2 such that $f_2 = f_1$ on U , $|f_2(x)| < \epsilon_1$ for $x \in U_2$ and hence $\|f_2 - f_1\| < 2\epsilon$, 0 is a regular value of f_2 and f_2 has in $U_2 \setminus U$ at least $||A_+(f_1, U)| - |A_-(f_1, U)||$ zeros of both orientations. Now we can pair all points from $A_{\pm}(f_2, U)$ with points from $A_{\mp}(f_2, U_2)$, to remove all zeros from U (there may remain some zeros in $U_2 \setminus U$). So, we may choose \tilde{f} to be the restriction of f_2 to $\bar{\Omega}$. ■

Finally, we prove the following theorem summarizing the results in this chapter:

Theorem 9 *Let $\bar{\Omega}$ be a closed region with interior $\Omega \subset \mathbb{R}^n$ and $f : \bar{\Omega} \rightarrow \mathbb{R}^n$ be analytic. Then f has a robust zero in $\bar{\Omega}$ if and only if there exists an open set $U \subseteq \Omega$ such that $0 \notin f(\partial U)$, and $\deg(f, U, 0) \neq 0$.*

Proof. If the degree $\deg(f, U, 0) \neq 0$, it follows from Theorem 7 that f has a robust zero in U , so it has a robust zero in $\bar{\Omega}$.

If f has a robust zero in $\bar{\Omega}$, then there exists a component Z of the zero set and a neighborhood $U \subset \bar{\Omega}$ of $Z \subset \bar{\Omega}$ such that $U \cap \{f = 0\} = Z$ and f has a robust zero in U . It follows from Lemma 6 that $Z \cap \partial\Omega = \emptyset$ and we may assume that $0 \notin f(\partial U)$. Theorem 8 states that $\deg(f, U, 0) \neq 0$. ■

In particular, the last theorem is true if $\bar{\Omega}$ is a box, or a finite union of boxes with nonempty interior.

7 Algorithm Correctness Proof

We will prove here that the algorithm CheckSat proposed in Section 3 fulfills its specification, which proves the main theorem of this paper. The correctness

proof will again be divided into the cases constituting the definition of class \mathcal{B} , from which correctness of the overall, recursive algorithm follows by induction.

Before that, we prove some technical results on the relationship between the class \mathcal{B} and robustness.

7.1 Robustness and the Class \mathcal{B}

Lemma 7 *Let S be a formula from \mathcal{B} . If $S[p \leftarrow p_0]$ is a robust sentence, then there exists an open neighborhood U of p_0 , such that for all $u \in U$, $S[p \leftarrow u]$ is robust, and equi-satisfiable to $S[p \leftarrow p_0]$.*

Proof. Assume that $S[p \leftarrow p_0]$ is robust. Then there is an $\epsilon > 0$ such for all T with $d(S[p \leftarrow p_0], T) < \epsilon$, T and $S[p \leftarrow p_0]$ are equi-satisfiable. Since $d(S[p \leftarrow u], S[p \leftarrow p_0]) = d(p_0, u)$, for $d(p_0, u) < \epsilon$, also $S[p \leftarrow u]$ is equi-satisfiable and robust. ■

Lemma 8 *Let S be a sentence from \mathcal{B} . If S is false, then it is robustly false.*

Proof. We proceed by induction, following the cases of Definition 7. Let S be the formula $\exists x \in B . f = 0 \wedge g \geq 0$, where $f = 0$, and $g \geq 0$ is the usual short-cut for conjunctions of equalities, and inequalities, respectively. Let S be false. The box B is a compact set, so $f(B)$ and $g(B)$ are compact. The compact set $f(B) \subset \mathbb{R}^n$ and the closed set $\{0\} \subset \mathbb{R}^n$ are disjoint. Similarly, the compact set $g(B) \subset \mathbb{R}^k$ and the closed set $[0, \infty)^k$ are disjoint. A compact and a disjoint closed set have a positive distance, so there exists a $d > 0$ such that the distance of $f(B)$ from $\{0\}$ is larger than d and the distance of $g(B)$ from $(0, \infty)$ is larger than d . So, S is d -robust.

Furthermore, assume that $I \subset \mathbb{R}$ is a compact interval and $\forall x \in I . S$ is a false sentence. Then there exists an $x_0 \in I$ such that $S[x \leftarrow x_0]$ is false. From the induction hypothesis, it is robustly false. Let $\epsilon > 0$ be such that $S[x \leftarrow x_0]$ is ϵ -robust and let S' be a formula such that $d(\forall x S', \forall x S) \leq \epsilon$. Then $d(S'[x \leftarrow x_0], S[x \leftarrow x_0]) \leq \epsilon$ and $S'[x \leftarrow x_0]$ is false. So, $\forall x \in I' . S'$ is false and it follows that $\forall x \in I . S$ is robustly false.

Finally, let U and V be sentences in \mathcal{B} and $U \wedge V$ be false. Then either U or V is false and the induction hypothesis says that it is robustly false. So, $U \wedge V$ is robustly false. Similarly, if $U \vee V$ is false, then both U and V are robustly false and $U \vee V$ is robustly false. ■

Lemma 9 *Let S be a formula containing a free variable x from a bounded closed interval I . Then the sentence $S[x \leftarrow x_0]$ is robustly true for all x_0 in I , if and only if the sentence $\forall x \in I . S$ is robustly true.*

Proof. Let $\forall x \in I . S$ be ϵ -robust and true, $x_0 \in I$ and let X be a sentence such that $d(X, S[x \leftarrow x_0]) < \epsilon$. Consider the formula $U := (S + X - S[x \leftarrow x_0])$ where the subtraction is applied on each function involved in the formulas. Clearly,

$d(\forall x \in I . S, \forall x \in I . U) = d(S, U) = d(X, S[x \leftarrow x_0]) < \epsilon$ and $\forall x \in I . U$ is true. In particular, $U[x \leftarrow x_0] \equiv X$ is true and it follows that $S[x \leftarrow x_0]$ is ϵ -robust and true.

For the converse, assume that for all $x_0 \in I$, $S[x \leftarrow x_0]$ is robustly true. Let

$$\mu(x_0) := \sup\{\mu > 0; S[x \leftarrow x_0] \text{ is } \mu\text{-robust}\}.$$

Clearly, $\mu(x_0)$ is a continuous function and has a minimum m on the compact interval I . So, for each $x_0 \in I$, $S[x \leftarrow x_0]$ is m -robust. If $d(\forall x \in I . S, \forall x \in I . U) < m$, then for each $x_0 \in I$, $d(S[x \leftarrow x_0], U[x \leftarrow x_0]) < m$ and $U[x \leftarrow x_0]$ is true. So, $\forall x \in I . U$ is true and $\forall x \in I . S$ is robustly true. ■

7.2 System of equations and inequalities

We again start with the case (a) of Definition 7 without disjunctions, that is, a formula S of the form

$$\exists x \in B . [f_1 = 0 \wedge f_2 = 0 \wedge \dots \wedge f_n = 0 \wedge g_1 \geq 0 \wedge g_2 \geq 0 \wedge \dots \wedge g_k \geq 0]$$

where B is an m -box. Again, we denote by f the function defined by the components (f_1, \dots, f_n) and g the function defined by the components (g_1, \dots, g_k) .

Theorem 10 *The algorithm $\text{SoEI}(S, P, r)$ fulfills the specification for $\text{CheckSat}(S, P, r)$.*

Proof. We divide the proof into several parts:

Correctness:

Assume first that the algorithm terminates with a negative result $\{\mathbf{F}\}$. It follows directly from Definition 3, that the input sentence S is false. Lemma 8 implies robustness.

Now assume that it terminates with a positive result $\{\mathbf{T}\}$. Then there exists a grid element A such that $\deg(f(p_0), A, 0) \neq 0$ and it follows from Theorem 7 that $f(p_0) = 0$ has a robust solution in A . For any $p \in P$, p and p_0 can be connected by a curve $\phi(t) \subset P$ and $f \circ \phi(t)$ is a homotopy between $f(p_0)$ and $f(p)$ nonzero on ∂A . So, $\deg(f(p), A, 0) \neq 0$ and it follows from Theorem 7 that $f(p) = 0$ has a robust solution on B . Moreover, the successful check whether for all $E \in S_r(A)$, for all $y_g \in I(g)(P \times E)$, $y_g > 0$ implies that for some small enough $d > 0$, for all $p \in P, x \in A$, $g(p, x) > 0$. It follows that for all $p \in P, x \in A$, $g(p, x) > d$ and the input sentence is robustly true for all parameter values in P .

Definiteness—Negative case:

Assume that $\exists x \in B . f(p_0) = 0 \wedge g(p_0) \geq 0$ is robustly false for each $p_0 \in P$. The sets $X = \{(p, x) \in P \times B \mid f(p, x) = 0\}$ and $Y = \{(p, x) \in P \times B \mid g(p, x) \geq 0\}$ are compact and disjoint, so they have a positive distance. For a small enough $\alpha > 0$, the sets $X' = \{(p, x) \in P \times B \mid |f(p, x)| \geq \alpha\}$ and $Y' = \{(p, x) \in P \times B \mid g(p, x) \geq (-\alpha, \dots, -\alpha)\}$ are still disjoint and have a positive distance $d > 0$. If ϵ_0 is fine enough, any ϵ_0 -box that has a nonempty intersection with Y' has an empty intersection with X' .

The second property of interval computability implies that if $\epsilon < \epsilon_0$ is small enough, then any ϵ -box $A \subset B$ and ϵ -box $P' \subset P$ have the following properties:

- If $P' \times A$ has empty intersection with Y' , then there is no $y_g \in I(g)(P' \times A)$ such that $y_g \geq 0$.
- If $P' \times A$ has empty intersection with X' , then there is no $y_f \in I(f)(P' \times A)$ such that $y_f = 0$.

So, for every $A \subset B$ in the S_ϵ -grid, and every $P' \subset P$ of width smaller than ϵ , either $P' \times A$ has empty intersection with X' or it has empty intersection with Y' and due to the above property, each box A satisfies that there is no $y_f \in I(f)(P' \times A), y_g \in I(g)(P' \times A)$ s.t. $y_f = 0$ and $y_g \geq 0$. So the algorithm terminates with $\{\mathbf{F}\}$.

Definiteness—Positive Case:

Assume now that $\exists x \in B . f(p_0) = 0 \wedge g(p_0) \geq 0$ is robustly true for each $p_0 \in P$. Then for each $p_0 \in P$, $f(p_0) = 0$ has a robust solution on the set $\{x \in B \mid g(p_0) \geq 0\}$ and even on $\{x \in B \mid g(p_0) \geq \alpha\}$ for a small enough $\alpha > 0$. It follows from Theorem 6 that $m = n$. Let $\Omega \subseteq B$ be an open neighborhood of $\{g(p_0) \geq \alpha\}$ such that $\Omega \subseteq \{g(p_0) \geq \alpha/2\}$. The equation $f(p_0) = 0$ has a robust solution on Ω and it follows from Theorem 9 that there exists an open set $U \subseteq \Omega$ such that $0 \notin f(p_0)(\partial U)$ and $\deg(f, U, 0) \neq 0$. Let $U(p_0) \subseteq P$ be a neighborhood of $\{p_0\}$ such that $(U(p_0) \times \Omega) \subseteq \{(p, x) \mid g(p, x) \geq \alpha/4\}$ and let $\epsilon_0(p_0)$ be so small that for every sub-box $K \subseteq U(p_0) \times \Omega$ of width less than $\epsilon_0(p_0)$,

$$I(g)(K) \subset (0, \infty)^k. \quad (1)$$

Possibly making $U(p_0)$ smaller, we may assume that $0 \notin f(\overline{U(p_0)} \times \partial U)$. Let $V \subset \Omega$ be a neighborhood of ∂U open in \mathbb{R}^n such that $0 \notin f(\overline{U(p_0)} \times \bar{V})$. The compactness of $\overline{U(p_0)} \times \bar{V}$ implies that $|f|$ has a positive minimum on this set and the second property of the definition of interval-computability implies that there exists a $\epsilon_2(p_0) < \epsilon_1(p_0)$ such that for every sub-box $K \subset U(p_0) \times V$ of width smaller than $\epsilon_2(p_0)$,

$$0 \notin I(f)(K). \quad (2)$$

Let $\epsilon_3(p_0) < \epsilon_2(p_0)$ be so that each box of width less than $\epsilon_3(p_0)$ that has a nonempty intersection with ∂U lies in V .

Let $P' \subset U(p_0)$ be a box of width at most $\epsilon_3(p_0)$. We will show that $\text{SoEI}(S, P', \epsilon_3(p_0))$ terminates with $\{\mathbf{T}\}$. The algorithm creates a grid of boxes S_ϵ , where ϵ is at most $\epsilon_3(p_0)$. It merges boxes containing a face C such that $0 \in I(f)(P' \times C)$ and removes grid elements containing a face $C \subset \partial B$ such that $0 \in I(f)(P' \times C)$. Let M be the smallest union of grid element in S_ϵ containing U . Clearly, $\partial M \subset V$, $0 \notin I(f)(P' \times C)$ for any boundary box $C \subset \partial M$ (due to (2)) and $\deg(f, M, 0) \neq 0$. So, there exists a subset $M' \subseteq M$, merging of grid elements in S_ϵ where the algorithm finds that $\deg(f, M', 0) \neq 0$ (otherwise, M would be a union of subsets on which f has zero degree, contradicting $\deg(f, M, 0) \neq 0$). Then it checks the condition whether for all $E \in S_r(M')$,

for all $y_g \in I(g)(P \times E)$, $y_g > 0$. This is satisfied due to (1) and the algorithm terminates with $\{\mathbf{T}\}$.

All this can be done for any $p_0 \in P$. So, we have a covering $\{U(p_0); p_0 \in P\}$ of the compact set P and can choose a finite sub-covering $\{U(p_1), \dots, U(p_s)\}$. There exists an ϵ' such that each box $P' \subset P$ of width smaller than ϵ' is contained in some $U(p_j)$. Let ϵ be the minimum of ϵ' and all the $\epsilon_3(p_j)$. For any $P' \subset P$ of width at most ϵ , $\text{SoEI}(S, P', \epsilon)$ terminates with a positive result $\{\mathbf{T}\}$. ■

7.3 Universal quantifiers

Theorem 11 *Let S be a formula containing free variables p . Let P be an $|p|$ -box and I a closed interval. Assume that an algorithm CheckSat satisfying the specification is given. Then also the algorithm $\text{Univ}(\forall x \in I . S, P, r)$ satisfies the specification.*

Proof. Concerning correctness, if $\text{Univ}(\forall x \in I . S, P, r)$ returns $\{\mathbf{F}\}$, then $\text{CheckSat}(S, P \times I', r')$ returned $\{\mathbf{F}\}$ for some $I' \in I_r$ and it follows that for all $p_0 \in P$ and $x_0 \in I'$, $S[p \leftarrow p_0][x \leftarrow x_0]$ is robustly false. Then $\forall x \in I . S[p \leftarrow p_0]$ is false for each $p_0 \in P$ and it follows from Lemma 8 that it is robustly false.

If the algorithm returns $\{\mathbf{T}\}$, then $\text{CheckSat}(S, P \times I', r')$ returned $\{\mathbf{T}\}$ for all $I' \in I_r$ and the sentence $S[p \leftarrow p_0][x \leftarrow x_0]$ is robustly true for all $x_0 \in I$ and $p_0 \in P$. It follows from Lemma 9 that for each $p_0 \in P$, $\forall x \in I . S[p \leftarrow p_0]$ is robustly true, so the result is correct.

Concerning definiteness, assume that for all $p_0 \in B$, the sentence $\forall x \in I . S[p \leftarrow p_0]$ is robustly true. Then, by Lemma 9, for all $p_0 \in B$ and all $x_0 \in I$, $S[p \leftarrow p_0][x \leftarrow x_0]$ is robustly true and the property follows directly from the assumption on CheckSat .

Assume now that for all $p_0 \in B$, $\forall x \in I . S[p \leftarrow p_0]$ is robustly false. Then for all p_0 , there exists a $x_0 \in I$ such that $S[p \leftarrow p_0][x \leftarrow x_0]$ is false, and hence, due to Lemma 8, it is also robustly false. From this, Lemma 7 implies that there is a neighborhood $P_0(p_0)$ of p_0 and I_0 of x_0 such that for all $p'_0 \in P_0(p_0)$ and $x'_0 \in I_0$, $S[p \leftarrow p'_0][x \leftarrow x'_0]$ is false. It follows from the assumption on CheckSat that there exists an ϵ_{p_0} such that for all $P' \subset P_0(p_0)$, $I' \subseteq I$ of width at most ϵ_{p_0} , $\text{CheckSat}(P' \times I', S, \epsilon_{p_0})$ terminates with $\{\mathbf{F}\}$.

Because P is compact, we can cover it by $\{P_0(p_0); p_0 \in \Lambda\}$ for a finite set Λ . It is easy to see that there exists a ϵ' such that any P' -box of side-length smaller than ϵ' is in at least one of these $P_0(p_0)$. Now, choose ϵ to be smaller than ϵ' and smaller than ϵ_{p_0} for all $p_0 \in \Lambda$. For any box P' of side-length at most ϵ , the algorithm $\text{Univ}(\forall x \in I . S, P', \epsilon)$ terminates with $\{\mathbf{F}\}$. ■

If we would use a dual algorithm to Univ for handling additional existential quantifiers (in addition to those from the base case of our class \mathcal{B}), the definiteness part of the proof would not go through in the case of a robustly true input (corresponding to the robustly false case for universal quantification): If

$\exists x \in I$. S is robustly true, then this does *not* imply that there exists a $x_0 \in I$ such that $S[x \leftarrow x_0]$ is robustly true. The topological degree was a way to get around this problem in the case where the number of variables is the same as the number of equations. Note however, that even in the case where we have more variables than equations, where intuition suggests higher robustness, the problem persists (see Section 8.1 for a detailed discussion).

7.4 Conjunction and Disjunction

Theorem 12 *Let S and T be two formulas in \mathcal{B} and assume that CheckSat fulfills the specification when applied to S , or T . Then $\text{Conj}(S \wedge T, P, r)$ also fulfills the specification.*

Proof. Let p_S and p_T be the projection of P to the free variables contained in S , resp. T .

Concerning correctness, if Conj returned $\{\mathbf{T}\}$ then the recursive calls for both S and T returned $\{\mathbf{T}\}$. Hence, by correctness of the result of the recursive calls, for all $p_0 \in P$, $S[p_S(p) \leftarrow p_S(p_0)]$ and $T[p_T(p) \leftarrow p_T(p_0)]$ are robustly true, and hence also $(S \wedge T)[p \leftarrow p_0]$.

If Conj returned $\{\mathbf{F}\}$ then the recursive calls for either S or T returned $\{\mathbf{F}\}$. Hence, by correctness of the result of the recursive calls, either for all $p_0 \in P$, $S[p_S(p) \leftarrow p_S(p_0)]$ is robustly false, or for all $p_0 \in P$, $T[p_T(p) \leftarrow p_T(p_0)]$ is robustly false. Hence, also for all $p_0 \in P$, $(S \wedge T)[p \leftarrow p_0]$ is robustly false.

Concerning definiteness, we first assume that for all $p_0 \in P$ the sentence $(S \wedge T)[p \leftarrow p_0]$ is robustly true. Then for all $p_0 \in P$, $S[p_S(p) \leftarrow p_S(p_0)]$ is robustly true, and also for all $p_0 \in P$, $T[p_T(p) \leftarrow p_T(p_0)]$ is robustly true. So, by definiteness of the recursive call, there exists an $\epsilon_1 > 0$ such that if $r < \epsilon_1$ and the width of $P_1 \subset P$ is less than ϵ_1 , then $\text{CheckSat}(S, p_S(P_1), r)$ terminates with $\{\mathbf{T}\}$. Moreover, an analogous ϵ_2 exists for T . For $\epsilon < \min\{\epsilon_1, \epsilon_2\}$, $P' \subset P$ of width less than ϵ and $r < \epsilon$, $\text{Conj}(S \wedge T, P', r)$ terminates with $\{\mathbf{T}\}$.

Now let us consider definiteness in the case where for all $p_0 \in P$, $(S \wedge T)[p \leftarrow p_0]$ is robustly false. Then, for any $p_0 \in P$, either $S[p_S(p) \leftarrow p_S(p_0)]$ or $T[p_T(p) \leftarrow p_T(p_0)]$ is robustly false. Assume, without loss of generality that $S[p_S(p) \leftarrow p_S(p_0)]$ is robustly false. By Lemma 7 there exists a neighborhood U of p_0 such that for every $u \in U$, $S[p_S(p) \leftarrow p_S(u)]$ is robustly false. Since P is compact, we can cover it with finitely many open sets within which either S or T is robustly false. For every element U from this cover there exists an $\epsilon_U > 0$ such that if $U' \subset U$ has width less than ϵ_U and $r < \epsilon_U$, then either $\text{CheckSat}(S, p_S(U'), r)$, or the respective call $\text{CheckSat}(T, p_T(U'), r)$, terminates with $\{\mathbf{F}\}$. Now we can select ϵ smaller than all ϵ_U and so small that every ϵ -box is contained in some U . ■

For disjunctions the situation is analogous.

Together with Theorem 2 this concludes the proof of the main theorem of the paper.

8 Possible Generalizations

We needed analyticity of the functions f involved in the formulas in \mathcal{B} only to be sure that $\{f = 0\}$ decomposes into a finite number of connected components. If f has this property, (in particular if $\{f = 0\}$ is discrete), then the degree $\deg(f, B, 0)$ can be calculated using the algorithm described in Section 5 and the algorithm SoEI terminates for a robust input.

8.1 Generalization of System of Equations

The main restriction in case of the equation $\exists x \in B . f = 0$ is the condition $m \leq n$, where B is an m -box and f is \mathbb{R}^n -valued. Assume that $m > n$. In some cases, we could fix some $m - n$ input variables in f to be constants $a \in \mathbb{R}^{m-n}$ and try to solve the equation $\exists x \in B^a . f(a) = 0$. This is a function from an n -box in \mathbb{R}^m to \mathbb{R}^n and if it has a robust zero in $B^a = \{x \in \mathbb{R}^m; (a, x) \in B\}$, so, clearly, f has a robust zero in B . The converse, however, is not true. If $f(a)$ does not have a robust zero in B^a for any fixed choice of $a \in \mathbb{R}^{m-n}$ (a ranging from any subset of $m - n$ from the total number of n variables), f still may have a robust zero in B . For an example, consider the Hopf fibration $H : S^3 \rightarrow S^2$ and define a map $\tilde{h} : \mathbb{R}^4 \rightarrow \mathbb{R}^3$ in polar coordinates by $\tilde{h}(rx) = rH(x)$ for $r \geq 0$ and $x \in S^3$. Let h be the restriction of \tilde{h} to the box $B = [-1, 1]^4$. Clearly, 0 is the only zero of h . For any $a \in [-1, 1]$, $h^a(x) = h(a, x)$ has not a robust zero in $[-1, 1]^3$ (for $a = 0$, a small perturbation of h^a is nowhere zero). However, $\exists x \in B . h = 0$ is robustly true. To show this, assume it is not and let h_1 is a nowhere zero small perturbation of h . Then $F(t) = th + (1 - t)h_1$ is a homotopy between h and h_1 , $0 \notin F(t)(\partial B)$ for all t . So, the map $H_1 := h_1/|h_1|$ from S^3 to S^2 is homotopic to the Hopf map $H = h/|h| : S^3 \rightarrow S^2$. Further, H_1 is homotopic to the trivial map via $G(t) : x \mapsto h_1(tx)/|h_1(tx)|$, but H is not homotopically trivial. This is a contradiction and therefore, h contains a robust zero in B . This makes a straightforward generalization of our result difficult.

The ideas presented in Section 6 may be easily generalized to any dimensions, if the condition of a nonzero degree is replaced by the condition "the map $f/|f|$ from $\partial\Omega$ to the sphere S^{m-1} can be continuously extended to a map from Ω to S^{m-1} ". This is the extension problem in computational homotopy theory. If Ω is the unit ball in \mathbb{R}^2 , $\partial\Omega = S^1$ and X is an arbitrary space, the question whether or not $f : \partial\Omega \rightarrow X$ presented algorithmically can be extended to $f : \Omega \rightarrow X$ is equivalent to the word problem and there is no algorithm to solve it [26]. The question whether or not such an algorithm exists for X being the sphere and Ω arbitrary, is—up to our knowledge—an open problem. Recent work of Krcal, Cadek and Sergeaert shows that such an algorithm is possible in the cases where the dimension of m of B is not larger then $2n$, but the algorithm is very complicated and involves the computations of homotopy groups of spheres, which is possible but highly nontrivial [4].

On the other hand, an algorithm that would decide, whether or not the sentence $\exists x \in B . f = 0$ is robustly true, may be used for proving that a map $f : S^k \rightarrow S^l$ is homotopically nontrivial. Assume that CheckSat is an algorithm

that decides, whether a robust sentence $S \equiv (\exists x \in B . f = 0)$ is true or false. Let $h : S^k \rightarrow S^l$ be interval-computable and define $H : \mathbb{R}^{k+1} \rightarrow \mathbb{R}^{l+1}$ by $H(rx) = rh(x)$ for $r \geq 0$ and $x \in S^k$. Then h is homotopically nontrivial if and only if S is robustly true.

8.2 Alternative Criteria for Detecting Zeros

Basic existence theorems that are commonly used for proving that an equation $f = 0$ has a solution in B are Kantorovich, Miranda's and Borsuk's theorem. Among these Borsuk's theorem is the strongest [2, 12], that is, if the assumptions of the other theorems are fulfilled, then the assumptions of Borsuk's theorem are fulfilled as well.

We will now remind the Borsuk's theorem and the compare its power for proving existence of a zero with the degree:

Theorem 13 (Borsuk's theorem) *If B is convex and symmetric with respect to an interior point x , $f : B \rightarrow \mathbb{R}^n$ is continuous such that $f(x) \neq 0$ on ∂B and on for any $x + y \in \partial B$ and $\lambda > 0$,*

$$f(x + y) \neq \lambda f(x - y),$$

then $f = 0$ has a solution in B .

If the assumption of this theorem is fulfilled, we can easily see that the degree $\deg(f, B, 0)$ is ± 1 , and the degree provides a test that is at least as strong as Borsuk's theorem.

On the other hand, if f has an isolated zero of degree $\pm 2, \pm 3, \dots$, the assumption of Borsuk's theorem would fail, and hence it is not strong enough for proving existence of arbitrary robust zeros.

The simplest example of such a case is the function

$$f : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x^2 - y^2 \\ 2xy \end{pmatrix}$$

defined in a neighborhood of $(0, 0)$. This function has clearly a robust zero in any neighborhood Ω of the origin, because it is the square map $f(z) = z^2$ under the identification $\mathbb{R}^2 \simeq \mathbb{C}$ and $\deg(f, \Omega, 0) = 2$. But the assumption of Borsuk's theorem are not fulfilled in any neighborhood of the origin.

8.3 Problem with Existential Quantifiers

The generalization of the algorithm SoEI to universal quantifiers is quite easy. The main ingredient is the fact that $\forall x \in I . S$ is robustly true if and only if for each $x_0 \in I$, the sentence $S[x \leftarrow x_0]$ is robustly true (Lemma 9). However, nothing like that holds for existence quantifiers. The sentence $\exists x \in [-1, 1] . x = 0$ is robustly true but for any $x_0 \in [-1, 1]$, the sentence $x_0 = 0$ (x_0 is considered to be a constant function here) is not robustly true. A topological reformulation of adding an existence quantifier to the beginning of a formula would be desirable and could be a subject of future research.

9 Conclusion

In the paper, we have proved that the problem of checking satisfiability of systems of equations of real analytic functions in a box is quasi-decidable in the sense that there exists an algorithm that successfully can do this check in all robust cases. Hence, problems that correspond to application domains where perturbations in the form of modeling errors, manufacturing imprecision etc. occur, are solvable in practice (provided enough computing power is available).

The generalization to the full first-order case is an open problem.

References

- [1] O. Aberth. Computation of topological degree using interval arithmetic, and applications. *Mathematics of Computation*, 62(205):171–178, 1994.
- [2] G. Alefeld, A. Frommer, G. Heindl, and J. Mayer. On the existence theorems of Kantorovich, Miranda and Borsuk. *Electronic Transactions on Numerical Analysis*, 17:102–111, 2004.
- [3] T. E. Boulton and K. Sikorski. Complexity of computing topological degree of Lipschitz functions in n dimensions. *J. Complexity*, 2:44–59, 1986.
- [4] M. Čádek, M. Krčál, J. Matoušek, F. Sergeraert, L. Vokřínek, and U. Wagner. Computing all maps into a sphere. <http://arxiv.org/abs/1105.6257>, 2011.
- [5] B. F. Caviness. On canonical forms and simplification. *J. ACM*, 17(2):385–396, 1970.
- [6] B. F. Caviness and J. R. Johnson, editors. *Quantifier Elimination and Cylindrical Algebraic Decomposition*. Springer, Wien, 1998.
- [7] P. Collins. Computability and representations of the zero set. *Electron. Notes Theor. Comput. Sci.*, 221:37–43, December 2008.
- [8] W. Damm, G. Pinto, and S. Ratschan. Guaranteed termination in the verification of LTL properties of non-linear robust discrete time hybrid systems. *International Journal of Foundations of Computer Science (IJFCS)*, 18(1):63–86, 2007.
- [9] I. Fonseca and W. Gangbo. *Degree Theory in Analysis and Applications*. Clarendon Press, Oxford, 1995.
- [10] P. Franek, S. Ratschan, and P. Zgliczynski. Satisfiability of systems of equations of real analytic functions is quasi-decidable. In *MFCS 2011: 36th International Symposium on Mathematical Foundations of Computer Science*, volume 6907 of *LNCS*. Springer, 2011.

- [11] M. Fränzle. Analysis of hybrid systems: An ounce of realism can save an infinity of states. In J. Flum and M. Rodriguez-Artalejo, editors, *Computer Science Logic (CSL'99)*, number 1683 in LNCS. Springer, 1999.
- [12] A. Frommer and B. Lang. Existence tests for solutions of nonlinear equations using Borsuk's theorem. *SIAM Journal on Numerical Analysis*, 43(3):1348–1361, 2005.
- [13] M. Hirsch. *Differential topology*. Springer, 1976.
- [14] R. Kearfott, J. Dian, and A. Neumaier. Existence verification for singular zeros of complex nonlinear systems. *SIAM J. Numer. Anal.*, 38(2):360–379, 2000.
- [15] R. B. Kearfott. On existence and uniqueness verification for non-smooth functions. *Reliable Computing*, 8(4):267–282, 2002.
- [16] S. Krantz and H.R.Parks. *A Primer of Real Analytic Functions*. Birkhäuser, 2002.
- [17] J. W. Milnor. *Topology from the differential viewpoint*. Princeton University Press, 1997.
- [18] J. Munkres. *Topology*. Prentice Hall, 1999.
- [19] A. Neumaier. *Interval Methods for Systems of Equations*. Cambridge Univ. Press, Cambridge, 1990.
- [20] D. O'Regan, Y. Cho, and Y.Q.Chen. *Topological Degree Theory and Applications*. Chapman & Hall, 2006.
- [21] S. Ratschan. Quantified constraints under perturbations. *Journal of Symbolic Computation*, 33(4):493–505, 2002.
- [22] S. Ratschan. Efficient solving of quantified inequality constraints over the real numbers. *ACM Transactions on Computational Logic*, 7(4):723–748, 2006.
- [23] S. Ratschan. Safety verification of non-linear hybrid systems is quasi-semidecidable. In *TAMC 2010: 7th Annual Conference on Theory and Applications of Models of Computation*, volume 6108 of LNCS, pages 397–408. Springer, 2010.
- [24] S. M. Rump. A note on epsilon-inflation. *Reliable Computing*, 4:371–375, 1998.
- [25] S. M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, pages 287–449, 2010.

- [26] F. Sergeraert. Introduction to combinatorial homotopy theory. Summer School and Conference Mathematics, Algorithms and Proofs, Lecture notes, <http://www-fourier.ujf-grenoble.fr/~sergerar/Papers/Trieste-Lecture-Notes.pdf>, 2008.
- [27] F. Stenger. Computing the topological degree of a mapping in R^n . *Numerische Mathematik*, 25:23–38, 1975.
- [28] A. Tarski. *A Decision Method for Elementary Algebra and Geometry*. Univ. of California Press, Berkeley, 1951. Also in [6].
- [29] P. S. Wang. The undecidability of the existence of zeros of real elementary functions. *J. ACM*, 21(4):586–589, 1974.
- [30] K. Weihrauch. *Introduction to Computable Analysis*. Texts in Theoretical Computer Science. Springer, Heidelberg, 2000.