Computing the Homology of Real Projective Sets

Felipe Cucker[*]
Dept. of Mathematics
City University of Hong Kong
HONG KONG
e-mail: macucker@cityu.edu.hk

Teresa Krick[†]
Departamento de Matemática & IMAS
Univ. de Buenos Aires & CONICET
ARGENTINA
e-mail: krick@dm.uba.ar

Michael Shub
Department of Mathematics
City College and the Graduate Center of CUNY
New York
USA
e-mail: mshub@ccny.cuny.edu

Abstract. We describe and analyze a numerical algorithm for computing the homology (Betti numbers and torsion coefficients) of real projective varieties. Here numerical means that the algorithm is numerically stable (in a sense to be made precise). Its cost depends on the condition of the input as well as on its size and is singly exponential in the number of variables (the dimension of the ambient space) and polynomial in the condition and the degrees of the defining polynomials. In addition, we show that outside of an exceptional set of measure exponentially small in the size of the data, the algorithm takes exponential time.

Keywords: real projective varieties, homology groups, complexity, condition, exponential time.

AMS classification numbers: 65Y20, 65H10, 55U10.

[*] Partially supported by a GRF grant from the Research Grants Council of the Hong Kong SAR (project number CityU 11310716).

[†] Partially supported by grants BID-PICT-2013-0294, UBACyT-2014-2017-20020130100143BA and PIP-CONICET 2014-2016-11220130100073CO.

1 Introduction

This paper describes and analyzes, both in terms of complexity and numerical stability, an algorithm to compute the topology of a real projective set.

The geometry of the sets of zeros of polynomials equalities, or more generally solutions of polynomial inequalities, is strongly tied to complexity theory. The problem of deciding whether such a set is nonempty is the paramount $\mathsf{NP}_{\mathbb{R}}$-complete problem (i.e., NP-complete over the reals) [7]; deciding whether it is unbounded is $\mathsf{H}\exists$-complete and whether a point is isolated on it is $\mathsf{H}\forall$-complete [9]; computing its Euler characteristic, or counting its points (in the zero dimensional case), $\#\mathsf{P}_{\mathbb{R}}$-complete [8], ...

We do not describe complexity classes in these pages. We content ourselves with the observation that such classes are characterized by restrictions in the use of specific resources (such as computing time or working space) and that complete problems are representatives for them. In this sense, the landscape of classes demanding an increasing amount of resources is paralleled by a collection of problems whose solution appears to be increasingly difficult.

Among the problems whose complexity is poorly understood, the computation of the homology of algebraic or semialgebraic sets —and by this we mean the computation of all their Betti numbers and torsion coefficients— stands out. The use of Cylindrical Algebraic Decomposition [12, 40] allows one to compute a triangulation of the set at hand (and from it, its homology) with a running time doubly exponential in the number of variables (the dimension of the ambient space). On the other hand, the $\#\mathsf{P}_{\mathbb{R}}$-hardness of computing the Euler characteristic (a simpler problem) mentioned above or the PSPACE-hardness of the problem of computing all Betti numbers of a complex algebraic (or projective) set defined over $\mathbb{Z}$, see [31], make clear that the existence of subexponential algorithms for the computation of the homology is unlikely. The obvious question is whether exponential time algorithms for this task exist.

A number of results in recent years have made substantial progress towards an answer to this question. Saugata Basu and collaborators provide algorithms computing the first Betti number of a semialgebraic set in single exponential time (an algorithm to compute the zeroth Betti number within these bounds was already known) [4], as well as an algorithm computing the top ℓ Betti numbers with cost doubly exponential in ℓ (but polynomial for fixed ℓ) [3]. More recently, Peter Scheiblechner [32] considered the class of smooth complex projective varieties and exhibited an algorithm computing all the Betti numbers (but not the torsion coefficients as the paper actually computes the de Rham homology) for sets in this class in single exponential time.

All the algorithms mentioned above are "symbolic", they are direct (as opposed to iterative) and are not meant to work under finite precision. Actually, numerical instability has been observed for many of them and very recent results [26] give some theoretical account for this instability. And partly motivated by this observed instability, an interest in numerical algorithms has developed in tandem with that on symbolic algorithms. An example of the former that bears on this paper is the algorithm in [19] to decide feasibility of semialgebraic sets. The idea was to decide the existence of the desired solution by exploring a grid. While this grid would have exponentially many points, the computation performed at each such point would be fast and accurate, thus ensuring numerical stability in the presence of round-off errors. Both the running time of the algorithm (directly related to the size of the grid) and the machine precision needed to ensure the output's correctness, were shown to depend on a condition number for the system of polynomial inequalities defining the semialgebraic set at hand.

These ideas were extended in [15, 16, 17] to describe and analyze a numerical algorithm for the more difficult question of counting points in zero-dimensional projective sets. Note that in this case the number to be computed coincides with the zeroth Betti number of the set (number of connected components), while higher Betti numbers are all zero.

We now extend them once more to solve the (even more difficult) problem of computing all the homology groups for projective (or spherical) algebraic sets.

In order to state our result, we need to introduce some notation.

Let $m \leq n$, $d_1, \ldots, d_m \in \mathbb{N}$ and $\boldsymbol{d} = (d_1, \ldots, d_m)$. We will denote by $\mathcal{H}_{\boldsymbol{d}}[m]$ the space of polynomial systems $f = (f_1, \ldots, f_m)$ with $f_i \in \mathbb{R}[X_0, \ldots, X_n]$ homogeneous of degree d_i. We may assume here that $d_i \geq 2$ for $1 \leq i \leq m$, since otherwise we could reduce the input to a system with fewer equations and unknowns. We set $D := \max\{d_i,\, 1 \leq i \leq m\}$ and $N := \dim_{\mathbb{R}} \mathcal{H}_{\boldsymbol{d}}[m] = \sum_{i=1}^{m} \binom{n+d_i}{n}$. Note that the last is the *size* of the system f in the sense that it is the number of reals needed to specify this system.

We associate to $f \in \mathcal{H}_{\boldsymbol{d}}[m]$ its zero sets $\mathcal{M}_{\mathbb{S}} := Z_{\mathbb{S}^n}(f)$ on the unit sphere $\mathbb{S}^n \subset \mathbb{R}^{n+1}$ and $\mathcal{M}_{\mathbb{P}} := Z_{\mathbb{P}^n}(f)$ on the projective space $\mathbb{P}^n(\mathbb{R})$. The former is the intersection of the cone of zeros $\mathcal{Z} := Z_{\mathbb{R}^{n+1}}(f)$ of f in $\mathbb{R}^{n+1}$ with $\mathbb{S}^n$ and the latter is the quotient of $\mathcal{M}_{\mathbb{S}}$ by identifying antipodal points. For a generic system f, both $\mathcal{M}_{\mathbb{S}}$ and $\mathcal{M}_{\mathbb{P}}$ are smooth manifolds of dimension $n-m$. We also associate to f a condition number $\kappa(f)$ (whose precise definition will be given in §2.1 below). Finally, we endow the linear space $\mathcal{H}_{\boldsymbol{d}}[m]$ with the Weyl inner product (also defined in §2.1) and consider the unit sphere $\mathbb{S}^{N-1} \subset \mathcal{H}_{\boldsymbol{d}}[m]$ with respect to the norm induced by it.

Theorem 1.1. *We describe an algorithm that, given $f \in \mathcal{H}_{\boldsymbol{d}}[m]$, returns the Betti numbers and torsion coefficients of $\mathcal{M}_{\mathbb{S}}$ (or of $\mathcal{M}_{\mathbb{P}}$), with the following properties.*

(i) *Its cost* $\mathrm{cost}(f)$ *on input f is bounded by $(nD\kappa(f))^{\mathcal{O}(n^2)}$.*

(ii) *Assume $\mathbb{S}^{N-1}$ is endowed with the uniform probability measure. Then, with probability at least $1-(nD)^{-n}$ we have* $\mathrm{cost}(f) \leq (nD)^{\mathcal{O}(n^3)}$.

(iii) *Similarly, with probability at least $1-2^{-N}$ we have* $\mathrm{cost}(f) \leq 2^{\mathcal{O}(N^2)}$.

(iv) *The algorithm is numerically stable.*

We give the proof of Theorem 1.1 in several steps. Part (i) is shown in Propositions 4.3 and 4.4. Parts (ii) and (iii) are in Corollary 5.4. We devote Section 7 to both define what we mean by numerical stability (in a context where we are computing integer numbers) and to sketch why our algorithm is numerically stable.

Remark 1.2. Parts (ii) and (iii) in the statement fit well within the setting of *weak complexity analysis* recently proposed in [2] (but see also [23, Theorem 4.4] for a predecessor of this setting). The idea here is to exclude from the analysis a set of outliers of exponentially small measure (a probability measure in the space of data is assumed). This exclusion may lead to dramatic differences in the quantity to be bounded and provide a better agreement between theoretical analysis and computational experience. A case at hand, studied in [2], is that of the power method to compute dominant eigenpairs. It is an algorithm experienced as efficient in practice (say for symmetric or Hermitian matrices) but whose expected number of iterations (for matrices drawn from the Gaussian orthogonal or unitary ensembles, respectively) is known to be infinite [23]. Theorem 1.4 in [2] shows that the expected number of iterations conditioned to excluding a set of exponentially small measure is polynomially bounded in the dimension n of the input matrix. The authors call this form of analysis *weak average-case*. Parts (ii) and (iii) in the statement can be seen as a form of weak worst-case analysis establishing weak worst-case exponential complexity.

Our algorithm relies on an extension of the ideas in [19] —the use of grids, an exclusion test, and the use of the α-theory of Smale to detect zeros of a polynomial system in the vicinity of a point at hand— to construct a covering of $\mathcal{M}_{\mathbb{S}}$ by open balls in $\mathbb{R}^{n+1}$ of the same radii. This common radius is chosen to ensure that the union of the balls in the covering is homotopically equivalent to $\mathcal{M}_{\mathbb{S}}$. The Nerve Theorem then ensures that this union is homotopically equivalent to the nerve of the covering and we can compute the homology groups of $\mathcal{M}_{\mathbb{S}}$ by computing those of the said nerve. We explain the basic ingredients (condition numbers, Smale's α-theory, the exclusion lemma, ...) in Section 2. Then, in Section 3, we describe and analyze the computation of the covering. Section 4 uses this covering to actually compute the homology groups (part (i) in Theorem 1.1) and Section 5 establishes the probability estimates (parts (ii) and (iii) in Theorem 1.1). Section 6 is devoted to prove a number of results which, to allow for a streamlined exposition, were only stated in Section 2. One of them, Theorem 2.9, links the γ-invariant of Smale with the injectivity radius $\tau(f)$ of the normal bundle of $\mathcal{M}_{\mathbb{S}}$ (in turn related to a number of metric properties of algebraic spherical (or projective) sets). This connection is, to the best of our knowledge, new and is interesting per se. Finally, and as already mentioned, Section 7 deals with issues of finite-precision and numerical stability.

Acknowledgments. We are grateful to Peter Bürgisser who suggested the topic of this paper to us and to the Simons Institute for receiving us in the Fall of 2014, which was where and when the suggestion was made. We also owe an anonymous referee for his very precise and enlightening comments.

2 The basic ingredients

2.1 Condition numbers

We need a condition number as a complexity (and accuracy) parameter. To define one we first fix a norm on the space $\mathcal{H}_{\boldsymbol{d}}[m]$. We follow the (by now well-established) tradition of using the Weyl norm, which is invariant under the action of orthogonal transformations in $\mathbb{R}^{n+1}$: for $f=(f_1,\ldots,f_m)$ with $f_i=\sum_{|\boldsymbol{a}|=d} f_{i,\boldsymbol{a}}X^{\boldsymbol{a}}$, this is $\|f_i\|^2=\sum_{|\boldsymbol{a}|=d} f_{i,\boldsymbol{a}}^2\binom{d}{\boldsymbol{a}}^{-1}$ and then $\|f\|^2:=\sum_{1\le i\le m}\|f_i\|^2$. See e.g. [10, §16.1] for details.

For a point $\xi\in\mathbb{R}^{n+1}$ we denote by $Df(\xi)=\Big(\dfrac{\partial f_i}{\partial x_j}(\xi)\Big)_{1\le i\le m, 0\le j\le n}:\mathbb{R}^{n+1}\to\mathbb{R}^m$ the derivative of f at ξ. We also write

$$\Delta(\xi):=\begin{bmatrix}\|\xi\|^{d_1-1}\sqrt{d_1} & & \\ & \ddots & \\ & & \|\xi\|^{d_m-1}\sqrt{d_m}\end{bmatrix}$$

(or simply Δ, if $\xi\in\mathbb{S}^n$).

The condition of f at a zero $\xi\in\mathbb{R}^{n+1}\setminus\{0\}$ has been well-studied in the series of papers [33, 34, 35, 37, 36]. It is defined as ∞ when the derivative $Df(\xi)$ of f at ξ is not surjective, and when $Df(\xi)$ is surjective as

$$\mu_{\mathrm{norm}}(f,\xi) := \|f\|\big\|Df(\xi)^{\dagger}\Delta(\xi)\big\|, \tag{1}$$

where $Df(\xi)^{\dagger} : \mathbb{R}^m \to \mathbb{R}^{n+1}$ is the *Moore-Penrose inverse* of the full-rank matrix $Df(\xi)$, i.e. $Df(\xi)^{\dagger} = Df(\xi)^{\mathrm{t}}(Df(\xi)\,Df(\xi)^{\mathrm{t}})^{-1}$, where $Df(\xi)^{\mathrm{t}}$ is the transpose of $Df(\xi)$. This coincides with the inverse of the restricted linear map $Df(\xi)|_{(\ker Df(\xi))^{\perp}}$. Also, the norm in $\|Df(\xi)^{\dagger}\Delta(\xi)\|$ is the spectral norm.

Since the expression in the right of (1) is well-defined for arbitrary points $x \in \mathbb{S}^n$, we can define $\mu_{\mathrm{norm}}(f,x)$ for any such point.

For 0-dimensional homogeneous systems, that is, for systems $f \in \mathcal{H}_{\boldsymbol{d}}[n]$, the quantity $\mu_{\mathrm{norm}}(f,x)$ in (1) is occasionally defined differently, by replacing $Df(x)^{\dagger}$ by $(Df(x)_{|T_x})^{-1}$. Here T_x denotes the orthogonal complement of x in $\mathbb{R}^{n+1}$ and we are inverting the restriction of the derivative $Df(x)$ to this space (see [10, §16.7]). This definition only makes sense when $m = n$ as in this case the restriction $(Df(x)_{|T_x})^{-1} : T_x \to \mathbb{R}^n$ is a linear map between spaces of the same dimension. This is not the case when $m \neq n$. Hence the use here of the Moore-Penrose derivative.

To define the condition of a system f it is not enough to just consider the condition at its zeros. For points $x \in \mathbb{R}^{n+1}$ where $\|f(x)\|$ is non-zero but small, small perturbations of f can turn x into a new zero (and thus change the topology of $\mathcal{Z}$). Following an idea going back to [13] and developed in this context in [17] we define

$$\kappa(f,x) := \frac{\|f\|}{\big\{\|f\|^2\mu_{\mathrm{norm}}^{-2}(f,x) + \|f(x)\|^2\big\}^{1/2}}$$

where $\mu_{\mathrm{norm}}(f,x)$ is defined as in (1) for $x \in \mathbb{S}^n$, with the convention that $\infty^{-1} = 0$ and $0^{-1} = \infty$, and

$$\kappa(f) := \max_{x\in\mathbb{S}^n} \kappa(f,x). \tag{2}$$

Remark 2.1. For any $\lambda \neq 0$ we have $\mu_{\mathrm{norm}}(f,x) = \mu_{\mathrm{norm}}(f,\lambda x)$, since when $Df(x)$ is surjective, $Df(\lambda x)^{\dagger} = \Big(\Lambda Df(x)\Big)^{\dagger} = Df(x)^{\dagger}\Lambda^{-1}$ for $\Lambda = \begin{bmatrix} \lambda^{d_1-1} & & \\ & \ddots & \\ & & \lambda^{d_m-1}\end{bmatrix}$. Similarly, $\mu_{\mathrm{norm}}(f,\xi) = \mu_{\mathrm{norm}}(\lambda f,\xi)$ for all $\lambda \neq 0$, and consequently, $\kappa(\lambda f) = \kappa(f)$.

Note that $\kappa(f) = \infty$ if only if there exists $\xi \in \mathbb{S}^n$ such that $f(\xi) = 0$ (i.e $\xi \in \mathcal{M}_{\mathbb{S}}$) and $Df(\xi)$ is not surjective, i.e., f belongs to the *set of ill-posed systems*

$$\Sigma_{\mathbb{R}} := \big\{f \in \mathcal{H}_{\boldsymbol{d}}[m] \mid \exists\, \xi \in \mathbb{S}^n \text{ such that } f(\xi) = 0 \text{ and } \operatorname{rank}(Df(\xi)) < m\big\}. \tag{3}$$

The following result is proved in Section 6.1. It extends a statement originally shown for square systems in [16] (see also [10, Theorem 19.3]).

Proposition 2.2. *For all $f \in \mathcal{H}_{\boldsymbol{d}}[m]$,*

$$\frac{\|f\|}{\sqrt{2}\,\mathrm{dist}(f,\Sigma_{\mathbb{R}})} \le \kappa(f) \le \frac{\|f\|}{\mathrm{dist}(f,\Sigma_{\mathbb{R}})}.$$

We prove the following in Section 6.2.

Proposition 2.3. *Let $m \leq n+1$. For all $f \in \mathcal{H}_{\boldsymbol{d}}[m]$, $0 \leq \varepsilon \leq \frac{1}{2}$ and $y, z \in \mathbb{S}^n$ such that*

$$\|y - z\| \leq \frac{2\varepsilon}{D^{3/2}\mu_{\text{norm}}(f, y)}$$

we have

$$\frac{1}{1+\frac{5}{2}\varepsilon}\,\mu_{\text{norm}}(f,y) \leq \mu_{\text{norm}}(f,z) \leq \Big(1+\frac{5}{2}\varepsilon\Big)\mu_{\text{norm}}(f,y).$$

2.2 Moore-Penrose Newton and point estimates

Let $f : \mathbb{R}^{n+1} \to \mathbb{R}^m$, $m \leq n+1$, be analytic. The *Moore-Penrose Newton operator* of f at $x \in \mathbb{R}^{n+1}$ is defined (see [1]) as

$$N_f(x) := x - Df(x)^{\dagger} f(x).$$

We say that it is well-defined if $Df(x)$ is surjective.

Definition 2.4. Let $x \in \mathbb{R}^{n+1}$. We say that x *converges to a zero of* f if the sequence $(x_k)_{k\geq 0}$ defined as $x_0 := x$ and $x_{k+1} := N_f(x_k)$ for $k \geq 0$ is well-defined and converges to a zero of f.

Following ideas introduced by Steve Smale in [38], the following three quantities were associated to a point $x \in \mathbb{R}^{n+1}$ in [37],

$$\begin{aligned}
\beta(f,x) &:= \|Df(x)^{\dagger}\, f(x)\|| \\
\gamma(f,x) &:= \max_{k>1} \left\| Df(x)^{\dagger} \frac{D^k f(x)}{k!} \right\|^{\frac{1}{k-1}} \\
\alpha(f,x) &:= \beta(f,x)\gamma(f,x),
\end{aligned}$$

when $Df(x)$ is surjective, and $\alpha(f,x) = \beta(f,x) = \gamma(f,x) = \infty$ when $Df(x)$ is not surjective. The quantity $\beta(f,x) = \|N_f(x) - x\|$ measures the length of the Newton step at x. The value of $\gamma(f,\xi)$, at a zero ξ of f, is related to the radius of the neighborhood of points that converge to the zero ξ of f, and the meaning of $\alpha(f,x)$ is made clear in the main theorem in the theory of point estimates.

Theorem 2.5. *Let $f : \mathbb{R}^{n+1} \to \mathbb{R}^m$, $m \leq n+1$, be analytic. Set $\alpha_0 = 0.125$. Let $x \in \mathbb{R}^{n+1}$ with $\alpha(f,x) < \alpha_0$, then x converges to a zero ξ of f and $\|x - \xi\| < 2\beta(f,x)$. Furthermore, if $n+1 = m$ and $\alpha(f,x) \leq 0.03$, then all points in the ball of center x and radius $\frac{0.05}{\gamma(f,x)}$ converge to the same zero of f.*

Proof. In [37, Th. 1.4] it is shown that under the stated hypothesis, x converges to a zero ξ of f and

$$\|x_{k+1} - x_k\| \leq \left(\frac{1}{2}\right)^{2^k - 1} \|x_1 - x_0\| = \left(\frac{1}{2}\right)^{2^k - 1} \beta(f,x).$$

Therefore

$$\|x_{i+1} - x\| \leq \sum_{0 \leq k \leq i} \left(\frac{1}{2}\right)^{2^k - 1} \beta(f,x) < (2 - \frac{1}{8})\beta(f,x).$$

This implies the first statement. The second is Theorem 4 and Remarks 5, 6 and 7 in [6, Ch. 8]. □

In what follows we will apply the theory of point estimates to the case of polynomial maps $f = (f_1, \dots, f_m)$. In the particular case where the f_i are homogeneous, the invariants α, β and γ are themselves homogeneous in x. We have $\beta(f, \lambda x) = \lambda\beta(f, x)$, $\gamma(f, \lambda x) = \lambda^{-1}\gamma(f, x)$, and $\alpha(f, \lambda x) = \alpha(f, x)$, for all $\lambda \neq 0$. This property motivates the following projective version for them:

$$\begin{aligned}
\beta_{\mathrm{proj}}(f, x) &:= \|x\|^{-1}\|Df(x)^{\dagger}\, f(x)\| \\
\gamma_{\mathrm{proj}}(f, x) &:= \|x\| \max_{k>1} \left\| Df(x)^{\dagger} \frac{D^k f(x)}{k!} \right\|^{\frac{1}{k-1}} \\
\alpha_{\mathrm{proj}}(f, x) &:= \beta_{\mathrm{proj}}(f, x)\gamma_{\mathrm{proj}}(f, x),
\end{aligned}$$

These projective versions coincide with the previous expressions when $x \in \mathbb{S}^n$ and an α-Theorem for them is easily derived from Theorem 2.5 above. Furthermore, β_{proj} still measures the (scaled) length of the Newton step, and γ_{proj} relates to the condition number via the following bound (known as the Higher Derivative Estimate),

$$\gamma_{\mathrm{proj}}(f, x) \leq \frac{1}{2} D^{3/2} \mu_{\mathrm{norm}}(f, x). \tag{4}$$

The proof is exactly the one of [6, Th. 2, p. 267] which still holds for $m \leq n$ and $Df(x)^{\dagger}$ instead of $Df(x)|_{T_x}^{-1}$.

We now move to "easily computable" versions $\overline{\alpha}, \overline{\beta}$ and $\overline{\gamma}$, which we define for $x \in \mathbb{S}^n$:

$$\begin{aligned}
\overline{\beta}(f, x) &:= \mu_{\mathrm{norm}}(f, x)\frac{\|f(x)\|}{\|f\|} \\
\overline{\gamma}(f, x) &:= \frac{1}{2} D^{3/2} \mu_{\mathrm{norm}}(f, x) \\
\overline{\alpha}(f, x) &:= \overline{\beta}(f, x)\overline{\gamma}(f, x) = \frac{1}{2} D^{3/2} \mu^2_{\mathrm{norm}}(f, x)\frac{\|f(x)\|}{\|f\|}.
\end{aligned} \tag{5}$$

For $x \in \mathbb{S}^n$, (4) therefore says that $\gamma(f, x) \leq \overline{\gamma}(f, x)$. We also observe that $\beta(f, x) \leq \overline{\beta}(f, x)$ since

$$\beta(f, x) = \left\|Df(x)^{\dagger} f(x)\right\| \leq \left\|Df(x)^{\dagger}\right\| \|f(x)\| \leq \|f\| \|Df(x)^{\dagger}\Delta\| \frac{\|f(x)\|}{\|f\|} = \overline{\beta}(f, x).$$

Therefore $\alpha(f, x) \leq \overline{\alpha}(f, x)$.

2.3 Curvature and coverings

A crucial ingredient in our development is a result in a paper by Niyogi, Smale and Weinberger [25, Prop.7.1]. The context of that paper (learning on manifolds) is different from ours but this particular result, linking curvature and coverings, is, as we said, central to us.

Consider a compact Riemannian submanifold $\mathcal{M}$ of a Euclidean space $\mathbb{R}^{n+1}$. Consider as well a finite collection of points $\mathcal{X} = \{x_1, \dots, x_K\}$ in $\mathbb{R}^{n+1}$ and also $\varepsilon > 0$. We are interested in conditions guaranteeing that the union of the open balls

$$U_{\varepsilon}(\mathcal{X}) := \bigcup_{x \in \mathcal{X}} B(x, \varepsilon)$$

covers $\mathcal{M}$ and is homotopically equivalent to it. These conditions involve two notions which we next define.

We denote by $\tau(\mathcal{M})$ the *injectivity radius of the normal bundle of* $\mathcal{M}$, i.e., the largest t such that the open normal bundle around $\mathcal{M}$ of radius t

$$N_t(\mathcal{M}) := \big\{(x,v) \in \mathcal{M} \times \mathbb{R}^{n+1} \mid v \in N_x\mathcal{M}, \|v\| < t\big\}$$

is embedded in $\mathbb{R}^{n+1}$. That is, the largest t for which $\phi_t : N_t(\mathcal{M}) \to \mathbb{R}^{n+1}$, $(x,v) \mapsto x+v$, is injective. Therefore, its image $\mathrm{Tub}_{\tau(\mathcal{M})}$ is an open tubular neighborhood of $\mathcal{M}$ with its canonical orthogonal projection map $\pi_0 : \mathrm{Tub}_{\tau(\mathcal{M})} \to \mathcal{M}$ mapping every point $x \in \mathrm{Tub}_{\tau(\mathcal{M})}$ to the (unique) point in $\mathcal{M}$ closest to x. In particular, $\mathcal{M}$ is a deformation retract of $\mathrm{Tub}_{\tau(\mathcal{M})}$.

Also, we recall that the Hausdorff distance between two subsets $A, B \subset \mathbb{R}^{n+1}$ is defined as

$$d_H(A,B) := \max\Big\{\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|a-b\|\Big\}.$$

If both A and B are compact, we have that $d_H(A,B) \le r$ if and only if for all $a \in A$ there exists $b \in B$ such that $\|a-b\| \le r$ and for all $b \in B$ there exists $a \in A$ such that $\|a-b\| \le r$.

The following is a slight variation of [25, Prop.7.1].

Proposition 2.6. *Let* $\overline{\tau} \le \tau(\mathcal{M})$ *and* $0 < r < (3-\sqrt{8})\overline{\tau}$. *If* $d_H(\mathcal{X},\mathcal{M}) \le r$ *then* $\mathcal{M}$ *is a deformation retract of* $U_\varepsilon(\mathcal{X})$ *for every* ε *satisfying*

$$\varepsilon \in \left(\frac{(r+\overline{\tau}) - \sqrt{r^2+\overline{\tau}^2-6r\overline{\tau}}}{2}, \frac{(r+\overline{\tau}) + \sqrt{r^2+\overline{\tau}^2-6r\overline{\tau}}}{2}\right). \qquad \square$$

Remark 2.7. If we start with $r > 0$ for which $6r < \tau(\mathcal{M})$ we can take $\overline{\tau} := 6r$. In this case the interval we obtain for the admissible values of ε is $[3r, 4r]$.

The quantity $\tau(\mathcal{M})$ is strongly related to the curvature of $\mathcal{M}$ as shown in Propositions 6.1, 6.2, and 6.3 in [25]. Even though we won't make use of these results, we summarize them in the following statement.

Theorem 2.8. *Let* $\tau := \tau(\mathcal{M})$.

(i) *The norm of the second fundamental form of* $\mathcal{M}$ *is bounded by* $\frac{1}{\tau}$ *in all directions.*

(ii) *For* $p, q \in \mathcal{M}$ *let* $\phi(p,q)$ *be the angle between their tangent spaces* T_p *and* T_q, *and* $d_{\mathcal{M}}(p,q)$ *their geodesic distance. Then* $\cos(\phi(p,q)) \ge 1 - \frac{1}{\tau}d_{\mathcal{M}}(p,q)$.

(iii) *For* $p, q \in \mathcal{M}$, $d_{\mathcal{M}}(p,q) \le \tau - \tau\sqrt{1 - \dfrac{2\|p-q\|}{\tau}}$. $\square$

2.4 Curvature and condition

Theorem 2.8 shows a deep relationship between the curvature of a submanifold $\mathcal{M}$ of Euclidean space and the value of $\tau(\mathcal{M})$. One of the main results in this paper is a further connnection, for the particular case where $\mathcal{M} = \mathcal{M}_{\mathbb{S}}$, the set of zeros of $f \in \mathcal{H}_{\boldsymbol{d}}[m]$ in $\mathbb{S}^n$, between $\tau(\mathcal{M}_{\mathbb{S}})$ and the values of γ on $\mathcal{M}_{\mathbb{S}}$. Define

$$\tau(f) := \tau(\mathcal{M}_{\mathbb{S}}) \quad \text{and} \quad \Gamma(f) := \max_{x\in\mathcal{M}_{\mathbb{S}}} \max\{1, \gamma(f,x)\}.$$

In Section 6.3 we prove the following.

Theorem 2.9. *We have*

$$\tau(f) \geq \frac{1}{87\,\Gamma(f)}.$$

Note that as $\max\{1, \gamma(f,x)\} \leq \overline{\gamma}(f,x)$ we obtain

Corollary 2.10.

$$\tau(f) \geq \frac{1}{87\,\overline{\Gamma}(f)}.$$

where $\overline{\Gamma}(f) := \max_{x\in\mathcal{M}_{\mathbb{S}}} \overline{\gamma}(f,x)$.

2.5 Grids and exclusion results

Our algorithm works on a grid $\mathcal{G}_\eta$ on $\mathbb{S}^n$, which we construct by projecting onto $\mathbb{S}^n$ a grid on the cube $\mathsf{C}^n = \{y \in \mathbb{R}^{n+1} \mid \|y\|_\infty = 1\}$. We make use of the (easy to compute) bijections $\phi : \mathsf{C}^n \to \mathbb{S}^n$ and $\phi^{-1} : \mathbb{S}^n \to \mathsf{C}^n$ given by $\phi(y) = \frac{y}{\|y\|}$ and $\phi^{-1}(x) = \frac{x}{\|x\|_\infty}$.

Given $\eta := 2^{-k}$ for some $k \geq 1$, we consider the uniform grid $\mathcal{U}_\eta$ of mesh η on C^n. This is the set of points in C^n whose coordinates are of the form $i2^{-k}$ for $i \in \{-2^k, -2^k+1, \ldots, 2^k\}$, with at least one coordinate equal to 1 or -1. We denote by $\mathcal{G}_\eta$ its image by ϕ in $\mathbb{S}^n$. An argument in elementary geometry shows that for $y_1, y_2 \in \mathsf{C}^n$,

$$\|\phi(y_1) - \phi(y_2)\| \leq d_{\mathbb{S}}(\phi(y_1), \phi(y_2)) \leq \frac{\pi}{2}\|y_1 - y_2\| \leq \frac{\pi}{2}\sqrt{n+1}\,\|y_1 - y_2\|_\infty, \tag{6}$$

where $d_{\mathbb{S}}(x_1, x_2) := \arccos(\langle x_1, x_2\rangle) \in [0, \pi]$ denotes the angular distance, for $x_1, x_2 \in \mathbb{S}^n$.

Given $\varepsilon > 0$, we denote by $B(x, \varepsilon) := \{y \in \mathbb{R}^{n+1} \mid \|y - x\| < \varepsilon\}$, for $x \in \mathbb{R}^{n+1}$, the open ball with respect to the Euclidean distance, and by $B_{\mathbb{S}}(x, \varepsilon) = \{y \in \mathbb{S}^n \mid d_{\mathbb{S}}(y, x) < \varepsilon\}$, for $x \in \mathbb{S}^n$, the open ball with respect to the angular distance. We also set from now on

$$\mathsf{sep}(\eta) := \eta\,\sqrt{n+1} \quad \text{and} \quad \delta(f, \eta) := 1.1\sqrt{D(n+1)}\|f\|\eta. \tag{7}$$

Lemma 2.11. *The union $\cup_{x\in\mathcal{G}_\eta} B(x, \mathsf{sep}(\eta))$ covers the sphere $\mathbb{S}^n$.*

Proof. Let $z \in \mathbb{S}^n$ and $y = \phi^{-1}(z) \in \mathsf{C}^n$. There exists $y' \in \mathcal{U}_\eta$ such that $\|y' - y\|_\infty \leq \frac{\eta}{2}$. Let $x = \phi(y') \in \mathcal{G}_\eta$. Then, equation (6) shows that $\|x - z\| \leq \frac{\eta}{2}\,\frac{\pi}{2}\,\sqrt{n+1} < \eta\,\sqrt{n+1}$. □

In [15, Lem. 3.1] and [10, Lem. 19.22], the following Exclusion Lemma is proved (the statement there is for $n = m$ but the proof holds for general m).

Lemma 2.12. (Exclusion lemma.) *Let $f \in \mathcal{H}_{\boldsymbol{d}}[m]$ and $x, y \in \mathbb{S}^n$ be such that $0 < d_{\mathbb{S}}(x, y) \leq \sqrt{2}$. Then,*

$$\|f(x) - f(y)\| < \|f\|\sqrt{D}\; d_{\mathbb{S}}(x, y).$$

In particular, if $f(x) \neq 0$, there is no zero of f in the ball $B_{\mathbb{S}}\big(x, \frac{\|f(x)\|}{\|f\|\sqrt{D}}\big)$. □

Corollary 2.13. *Let η be such that $\mathsf{sep}(\eta) \leq \frac{1}{2}$, and let $x \in \mathbb{S}^n$ satisfy $\|f(x)\| > \delta(f, \eta)$. Then $f(y) \neq 0$ on the ball $B(x, \mathsf{sep}(\eta))$.*

Proof. Let $y \in \mathbb{R}^{n+1}$ such that $\|y-x\| < \mathsf{sep}(\eta) \le \frac{1}{2}$. Define $h(\varepsilon) = \sqrt{2-2\sqrt{1-\varepsilon^2}}$. We have $\|\phi(y)-x\| \le h(\|y-x\|)$. Since $h(\varepsilon)/\varepsilon$ is monotonically increasing on $[0,1]$,

$$\|\phi(y)-x\| \le 2h(1/2)\|y-x\| < 1.035\|y-x\| < 0.5175 \text{ for } \ \|y-x\| < \frac{1}{2}.$$

Then,

$$d_{\mathbb{S}}(\phi(y),x) = 2\arcsin\Big(\frac{\|\phi(y)-x\|}{2}\Big) \le 1.012\|\phi(y)-x\| < 1.1\|x-y\| < 1.1\,\mathsf{sep}(\eta)$$

since arcsin is a convex function on the interval $[0, 0.5175]$. Therefore the hypothesis on $\|f(x)\|$ implies that

$$\|f(x)\| > 1.1\,\|f\|\sqrt{D}\,\mathsf{sep}(\eta) > \|f\|\sqrt{D}\,d_{\mathbb{S}}(\phi(y),x)$$

i.e., that $d_{\mathbb{S}}(\phi(y),x) < \frac{\|f(x)\|}{\|f\|\sqrt{D}}$. Lemma 2.12 then shows, since $f(x) \neq 0$, that $f(\phi(y)) \neq 0$ and we conclude that $f(y) \neq 0$ as f is homogeneous. $\square$

3 Computing a homotopically equivalent covering

Set $k := \lceil \log_2 4\sqrt{n+1} \rceil$ so that $\mathsf{sep}(\eta) \le \frac{1}{4}$ for $\eta = 2^{-k}$, where $\mathsf{sep}(\eta)$ is defined in (7). Our algorithm works on the grid $\mathcal{G}_\eta$ on $\mathbb{S}^n$ constructed in the previous section, and makes use of the quantities $\overline{\beta}, \overline{\gamma}$ and $\overline{\alpha}$ introduced in (5) and $\delta(f,\eta)$ defined in (7). We recall $\alpha_0 := 0.125$.

Algorithm 1. Covering

Input: $f \in \mathcal{H}_{\boldsymbol{d}}[m]$

Preconditions: $f \neq 0$

```
let $\eta := 2^{-k}$
repeat
    $\mathcal{X} := \emptyset$
    $r := \sqrt{\mathsf{sep}(\eta)}$
    $\varepsilon := 3.5\,r$
    for all $x \in \mathcal{G}_\eta$
        if $\overline{\alpha}(f,x) \le \alpha_0$ and $\frac{1}{531\,\overline{\gamma}(f,x)} \ge r$ and $2.2\,\overline{\beta}(f,x) < r$ then
            $\mathcal{X} := \mathcal{X} \cup \{x\}$
        elsif $\|f(x)\| \ge \delta(f,\eta)$ then do nothing
        elsif go to (*)
    return the pair $\{\mathcal{X},\varepsilon\}$ and halt
    end for
    (*) $\eta := \eta/2$
```

Output: $\{\mathcal{X},\varepsilon\}$

Postconditions: The algorithm halts if $f \notin \Sigma_{\mathbb{R}}$. If $\mathcal{X} = \emptyset$ then $\mathcal{M}_{\mathbb{S}}$ is empty. Otherwise, the set $\mathcal{X}$ is closed by the involution $x \mapsto -x$, and the union of the balls $\{B(x,\varepsilon) \mid x \in \mathcal{X}\}$ covers $\mathcal{M}_{\mathbb{S}}$ and is homotopically equivalent to it.

In the sequel we use the quantity

$$\mathbf{C} := \max\left\{12\,(n+1)D,\ \frac{531^2}{2}\sqrt{n+1}\,D^3\right\}. \tag{8}$$

Note that we have $\mathbf{C} = \mathcal{O}(n\,D^3)$.

Proposition 3.1. *Algorithm* Covering *is correct (it computes a list $\{\mathcal{X}, \varepsilon\}$ satisfying its postconditions). Furthermore, its cost is bounded by*

$$\mathcal{O}\left(\log_2(\mathbf{C}\kappa(f))\,nN(2\mathbf{C}\kappa^2(f))^n\right) = (nD\kappa(f))^{\mathcal{O}(n)}$$

and the number K of points in the returned $\mathcal{X}$ is bounded by $(nD\kappa(f))^{\mathcal{O}(n)}$.

The rest of this section is devoted to prove Proposition 3.1.

Lemma 3.2. *Let $x \in \mathbb{S}^n$ and $y \in \mathcal{Z}(f)$ be such that $\|x-y\| \leq 0.7$. Then the point $\phi(y) := \frac{y}{\|y\|} \in \mathcal{M}_{\mathbb{S}}$ satisfies $\|x - \phi(y)\| \leq 1.1\|x-y\|$.*

PROOF. The proof goes exactly as the proof of Corollary 2.13. □

The following two lemmas deal with the correctness of the algorithm.

Assume the algorithm halts for a certain value η. Let $\mathcal{X}$ be the set constructed by the execution at this stage and set $r = \sqrt{\mathsf{sep}(\eta)}$.

Lemma 3.3. *The sets $\mathcal{X}$ and $\mathcal{M}_{\mathbb{S}}$ satisfy $d_H(\mathcal{X}, \mathcal{M}_{\mathbb{S}}) \leq r$. Furthermore, for all $y \in \mathcal{M}_{\mathbb{S}}$, there exists $x \in \mathcal{X}$ such that $\|y - x\| \leq r^2$.*

PROOF. The points in $\mathcal{G}_\eta$ divide into two groups that satisfy, respectively:

$\boxed{x \in \mathcal{G}_\eta \setminus \mathcal{X}}$ This happens when $\|f(x)\| \geq \delta(f,\eta)$, and therefore, by Corollary 2.13, there are no zeros of f in the ball $B(x, \mathsf{sep}(\eta)) = B(x, r^2)$.

$\boxed{x \in \mathcal{X}}$ This happens when in particular $\overline{\alpha}(f,x) < \alpha_0$, and therefore, by Theorem 2.5, there exist zeros of f in the ball $B(x, 2\beta(f,x)) \subset B(x, r/1.1)$ since $2.2\overline{\beta}(f,x) < r$. This implies, because of Lemma 3.2, that $\mathcal{M}_{\mathbb{S}} \cap B(x,r) \neq \emptyset$.

This last sentence shows that for $x \in \mathcal{X}$, there exists $y \in \mathcal{M}_{\mathbb{S}}$ with $\|y - x\| < r$. In addition, since by Lemma 2.11, $\cup_{x \in \mathcal{G}_\eta} B(x, r^2)$ covers the sphere $\mathbb{S}^n$ and there are no points of $\mathcal{M}_{\mathbb{S}}$ in $\cup_{x \in \mathcal{G}_\eta \setminus \mathcal{X}} B(x, r^2)$, it follows that $\mathcal{M}_{\mathbb{S}} \subset \cup_{x \in \mathcal{X}} B(x, r^2)$ and therefore for all $y \in \mathcal{M}_{\mathbb{S}}$, there exists $x \in \mathcal{X}$ such that $\|y - x\| \leq r^2 < r$. This shows that $d_H(\mathcal{X}, \mathcal{M}_{\mathbb{S}}) \leq r$. □

Lemma 3.4. *Let $\overline{\tau} := 6r$. Then $\overline{\tau} < \tau(f)$.*

PROOF. Let $y \in \mathcal{M}_{\mathbb{S}}$ be such that $\overline{\Gamma}(f) = \overline{\gamma}(f,y)$, for $\overline{\Gamma}(f)$ defined in Identity (2.10). By Lemma 3.3 there exists $x \in \mathcal{X}$ such that $\|x - y\| < r$. Hence,

$$\|x-y\| \,<\, r \,\leq\, \frac{1}{531\,\overline{\gamma}(f,x)} \,=\, \frac{2}{531\,D^{3/2}\mu_{\text{norm}}(f,x)}.$$

By Proposition 2.3 (with $\varepsilon = \frac{1}{531}$) we have $\mu_{\text{norm}}(f,y) \leq (1 + \frac{5}{1062})\mu_{\text{norm}}(f,x) \leq 1.005\,\mu_{\text{norm}}(f,x)$. Consequently, $\overline{\gamma}(f,y) \leq 1.005\overline{\gamma}(f,x)$ and therefore,

$$\overline{\tau} = 6r \leq \frac{6}{531\,\overline{\gamma}(f,x)} \leq \frac{6.03}{531\,\overline{\gamma}(f,y)} < \frac{1}{87\,\overline{\gamma}(f,y)} = \frac{1}{87\,\overline{\Gamma}(f)} \leq \tau(f),$$

the last by Theorem 2.9. □

To bound the complexity we rely on the following.

Lemma 3.5. *Let* $\mathbf{C}$ *be defined in* (8)*. Suppose* $\eta \leq \frac{1}{\mathbf{C}\kappa^2(f)}$ *and let* $\mathcal{X}$ *be the set constructed by the algorithm for this* η*. Then, for all* $x \in \mathcal{G}_\eta$ *either* $x \in \mathcal{X}$ *or* $\|f(x)\| > \delta(f,\eta)$*.*

PROOF. Let $x \in \mathcal{G}_\eta$. By the definition of $\kappa(f)$ in (2),

$$\frac{1}{\kappa^2(f)} \leq 2\max\left\{\mu_{\mathrm{norm}}^{-2}(f,x), \frac{\|f(x)\|^2}{\|f\|^2}\right\}.$$

We accordingly divide the proof into two cases.

Assume first that $\max\left\{\mu_{\mathrm{norm}}^{-2}(f,x), \frac{\|f(x)\|^2}{\|f\|^2}\right\} = \frac{\|f(x)\|^2}{\|f\|^2}$.
In this case

$$\eta \leq \frac{1}{\mathbf{C}\kappa^2(f)} \leq \frac{2\|f(x)\|^2}{\mathbf{C}\|f\|^2},$$

which implies

$$\|f(x)\| \geq \frac{\sqrt{\eta\,\mathbf{C}}\,\|f\|}{\sqrt{2}} > \frac{\eta\sqrt{\mathbf{C}}\|f\|}{\sqrt{2}} \geq 1.1\sqrt{(n+1)D}\,\|f\|\,\eta = \delta(f,\eta),$$

the second inequality since $\eta < 1$ and the third since $\mathbf{C} \geq 12(n+1)D$.

Now assume instead that $\max\left\{\mu_{\mathrm{norm}}^{-2}(f,x), \frac{\|f(x)\|^2}{\|f\|^2}\right\} = \mu_{\mathrm{norm}}^{-2}(f,x)$.
In this case

$$\eta \leq \frac{1}{\mathbf{C}\kappa^2(f)} \leq \frac{2}{\mathbf{C}\mu_{\mathrm{norm}}^2(f,x)}. \tag{9}$$

We will show that the condition $\frac{1}{531\,\overline{\gamma}(f,x)} \geq \sqrt{\mathsf{sep}(\eta)}$ of the algorithm holds true, and that when any of the other two conditions doesn't hold, then $\|f(x)\| > \delta(f,\eta)$.

Indeed,

$$\overline{\gamma}(f,x) = \frac{1}{2}D^{3/2}\mu_{\mathrm{norm}}(f,x) \underset{(9)}{\leq} \frac{\sqrt{2}}{2}D^{3/2}\frac{1}{\sqrt{\mathbf{C}\eta}} \leq \frac{1}{531\,\sqrt{\eta}\,(n+1)^{1/4}} = \frac{1}{531\,\sqrt{\mathsf{sep}(\eta)}},$$

the second inequality since $\sqrt{\mathbf{C}} \geq \frac{\sqrt{2}}{2}531(n+1)^{1/4}D^{3/2}$.

Assume now that $\overline{\alpha}(f,x) > \alpha_0$. Then

$$\alpha_0 < \frac{1}{2}D^{3/2}\mu_{\mathrm{norm}}^2(f,x)\frac{\|f(x)\|}{\|f\|}$$

which implies

$$\|f(x)\| > \|f\|\frac{2\alpha_0}{D^{3/2}\mu_{\mathrm{norm}}^2(f,x)} \underset{(9)}{\geq} \|f\|\mathbf{C}\eta\frac{\alpha_0}{D^{3/2}} \geq 1.1\,\sqrt{D(n+1)}\,\|f\|\,\eta = \delta(f,\eta),$$

the last inequality since $\mathbf{C} \geq \frac{531^2}{2}\sqrt{n+1}\,D^3 \geq \frac{1.1\sqrt{n+1}D^2}{\alpha_0}$.

Assume finally that $2.2\,\overline{\beta}(f,x) \geq \sqrt{\mathsf{sep}(\eta)}$, i.e.

$$2.2\frac{\|f(x)\|}{\|f\|}\mu_{\mathrm{norm}}(f,x) \geq \sqrt{\eta}(n+1)^{1/4}.$$

This implies

$$\|f(x)\| \geq \|f\|\sqrt{\eta}\frac{(n+1)^{1/4}}{2.2\,\mu_{\mathrm{norm}}(f,x)} \underset{(9)}{\geq} \|f\|\eta\frac{\sqrt{\mathbf{C}}(n+1)^{1/4}}{2.2\sqrt{2}} \geq 1.1\sqrt{(n+1)D}\,\|f\|\,\eta = \delta(f,\eta),$$

since $\mathbf{C} \geq 12(n+1)D$. □

Proof of Proposition 3.1. Lemmas 3.3, 3.4 and Remark 2.7 show that if the algorithm halts, then the current value of r when halting and that of $\overline{\tau} := 6r$ satisfy the hypothesis of Proposition 2.6. The fact that $\overline{\tau} = 6r$ shows that with the choice $\varepsilon := 3.5r$ the manifold $\mathcal{M}_{\mathbb{S}}$ is a deformation retract of $U_\varepsilon(\mathcal{X})$ and, hence, the two are homotopically equivalent. Finally, the fact that $\mathcal{X}$ is closed under the involution $x \mapsto -x$ is straightforward. This shows correctness.

To evaluate the complexity, note that Lemma 3.5 shows that the algorithm halts as soon as

$$\eta \leq \eta_0 := \frac{1}{\mathbf{C}\,\kappa^2(f)}.$$

This gives a bound of $\mathcal{O}(\log_2(\mathbf{C}\kappa(f)))$ for the number of iterations.

At each such iteration there are at most $R_\eta := 2(n+1)\left(\frac{2}{\eta}\right)^n$ points in the grid $\mathcal{G}_\eta$. For each such point x we can evaluate $\mu_{\mathrm{norm}}(f,x)$ and $\|f(x)\|$, both with cost $\mathcal{O}(N)$ (cf. [10, Prop. 16.45 and Lem. 16.31]). It follows that the cost of each iteration is $\mathcal{O}(R_\eta N)$.

Since at these iterations $\eta \geq \eta_0$, we have $R_\eta \leq 2(n+1)\big(2\mathbf{C}\kappa^2(f)\big)^n$. Using this estimate in the $\mathcal{O}(R_\eta N)$ cost of each iteration and multiplying by the bound $\mathcal{O}(\log_2(\mathbf{C}\kappa(f)))$ for the number of iterations, we obtain a bound of $N(nD\kappa(f))^{\mathcal{O}(n)}$ for the total cost. The claimed bound follows by noting that $N = (nD)^{\mathcal{O}(n)}$.

Finally, the number of points K of the returned $\mathcal{X}$ satisfies

$$K = R_{\eta_0} \leq 2(n+1)\big(2\mathbf{C}\,\kappa^2(f)\big)^n = (nD\kappa(f))^{\mathcal{O}(n)}. \qquad \square$$

4 Computing the Betti numbers and torsion coefficients of spherical and projective algebraic sets

Let X be a topological space and $\{U_i\}_{i\in I}$ a collection of open subsets covering X. We recall that the *nerve* of this covering is the abstract simplicial complex $\mathcal{N}(U_i)$ defined on I so that a finite set $J \subset I$ belongs to $\mathcal{N}(U_i)$ if and only if the intersection $\cap_{j\in J}U_j$ is nonempty. In general the complex does not reflect the topology of X, except when intersections are contractible, in which case there is the Nerve Theorem, that we quote here from [5, Theorem 10.7].

Theorem 4.1. *Let X be a triangulable topological space and $\{U_i\}_{i\in I}$ a locally finite family of open subsets (or a finite family of closed subsets) such that $X = \cup_{i\in I}U_i$. If every nonempty finite intersection $\cap_{j\in J}U_j$ is contractible, then X and the nerve $\mathcal{N}(U_i)$ are homotopically equivalent.* □

Here we use the Nerve Theorem in the case where the sets U_i in the statement of the theorem are the open balls $B(x_i,\varepsilon)$ for $x_i \in \mathcal{X}$ where $\{\mathcal{X},\varepsilon\}$ is the output of Algorithm 1 and X is their union. Note that as balls are convex, so is their intersection. Hence, these intersections, if nonempty, are contractible, and we can apply the Nerve Theorem. That is, given $\{\mathcal{X},\varepsilon\}$ we want to compute first its nerve $\mathscr{N} := \mathscr{N}(U_i)$ and then, the Betti numbers and torsion coefficients of $\mathscr{N}$. Proposition 3.1 and Theorem 4.1 ensure that these quantities coincide for $\mathscr{N}$ and $\mathcal{M}_{\mathbb{S}}$.

In what follows, we assume that we have ordered the set $\mathcal{X}$ so that $\mathcal{X} = \{x_1 < x_2 < \ldots < x_K\}$ where $K = |\mathcal{X}|$ is the cardinality of $\mathcal{X}$. Then, for $k \geq 0$, the abelian group C_k of k-chains of $\mathscr{N}$ is free, generated by the set of k-faces

$$\big\{J \subset \{x_1,\ldots,x_K\} \mid |J| = k \text{ and } \bigcap_{x_j \in J} B(x_j,\varepsilon) \neq \emptyset\big\}. \tag{10}$$

To determine the faces of C_k from $\{\mathcal{X},\varepsilon\}$ we need to be able to decide whether, given a subset $\{x_{i_1},\ldots,x_{i_k}\}$ of $\mathcal{X}$, the intersection of the balls $B(x_{i_j},\varepsilon)$, $j = 1,\ldots,k$, is nonempty. This is equivalent to say that the smallest ball containing all the points $\{x_{i_1},\ldots,x_{i_k}\}$ has radius smaller than ε, and we can do so if we have at hand an algorithm computing this smallest ball. Since we are looking here for a deterministic algorithm, we do not apply the efficient but randomized algorithm of [22, pp. 60–61], whose (expected) cost is bounded by $\mathcal{O}((n+2)!k)$, but we apply a deterministic quantifier elimination algorithm to the following problem: given $x_{i_1},\ldots,x_{i_k} \in \mathbb{R}^{n+1}$ and $\varepsilon > 0$, decide whether

$$\exists\, z \in \mathbb{R}^{n+1} \text{ s.t. } \|x_{i_j} - z\| < \varepsilon \text{ for } 1 \leq j \leq k.$$

This can be solved using for instance [27] in time linear in $k^{\mathcal{O}(n)}$. As there are $\binom{K}{k} \leq K^k$ subsets of k elements in I, the following result is clear.

Lemma 4.2. *The cost of constructing C_k is bounded by $K^k \cdot k^{\mathcal{O}(n)}$.* □

For $k \geq 1$ the boundary map $\partial_k : C_k \to C_{k-1}$ is defined, for a simplex $J \in C_k$, $J = \{x_{i_1},\ldots,x_{i_k}\}$, with $i_1 < i_2 < \ldots < i_k$, by

$$\partial_k(J) = \sum_{j=1}^{k} (-1)^j \{x_{i_1},\ldots,\widehat{x_{i_j}},\ldots,x_{i_k}\}$$

where the $(k-1)$-face $\{x_{i_1},\ldots,\widehat{x_{i_j}},\ldots,x_{i_k}\}$ is obtained by deleting the jth element in J. This map is therefore represented by a matrix M_k with O_{k-1} rows and O_k columns with entries in $\{-1,0,1\}$, where O_k denotes the number of faces in (10).

Proposition 4.3. *We can compute the Betti numbers $b_0(\mathcal{M}_{\mathbb{S}}),\ldots,b_{n-m}(\mathcal{M}_{\mathbb{S}})$ as well as the torsion coefficients of $\mathcal{M}_{\mathbb{S}}$ with cost*

$$(nD\kappa(f))^{\mathcal{O}(n^2)}.$$

Proof. Algorithm Covering produces, as shown in Proposition 3.1, a pair $\{\mathcal{X},\varepsilon\}$ such that the union $U_\varepsilon(\mathcal{X})$ of the balls $B(x,\varepsilon)$, for $x \in \mathcal{X}$, covers $\mathcal{M}_{\mathbb{S}}$ and is homotopically equivalent to it. Theorem 4.1 then ensures that the nerve $\mathscr{N}$ of this covering is homotopically equivalent to $U_\varepsilon(\mathcal{X})$ (and hence to $\mathcal{M}_{\mathbb{S}}$). It is therefore enough to compute the Betti numbers and torsion coefficients of $\mathscr{N}$. To do so, we construct, for $k = 0,\ldots,n-m+1$, the group C_k (i.e., we determine its faces). This has cost

$$\sum_{k=0}^{n-m+1} K^k \cdot k^{\mathcal{O}(n)} = \sum_{k=0}^{n-m+1} (nD\kappa(f))^{\mathcal{O}(nk)} k^{\mathcal{O}(n)} = (nD\kappa(f))^{\mathcal{O}(n^2)}$$

by Lemma 4.2 and the bound for K in Proposition 3.1.

With the groups C_k at hand we write down the matrices M_k corresponding to the boundary maps ∂_k, for $k = 1, \dots, n-m+1$. Next we compute their Smith normal forms D_k,

$$D_k = \begin{bmatrix} b_{k,1} & & & & & \\ & \ddots & & & & \\ & & b_{k,t_k} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}.$$

Then, $\dim \mathrm{Im}\partial_k = \mathrm{rank}(D_k) = t_k$, and consequently $\dim\ker\partial_k = O_k - \mathrm{rank}(D_k) = O_k - t_k$. For $k = 1, \dots, n-m$ we thus obtain the Betti numbers

$$b_k(\mathcal{M}_{\mathbb{S}}) = \dim\big(\ker\partial_k/\mathrm{Im}\partial_{k+1}\big) = O_k - t_k - t_{k+1}$$

and the same formula yields $b_0(\mathcal{M}_{\mathbb{S}})$ and $b_{n-m}(\mathcal{M}_{\mathbb{S}})$ by taking $t_0 = 0$. Furthermore, it is well-known that the kth homology group of $\mathcal{N}$ (and hence that of $\mathcal{M}_{\mathbb{S}}$ as well) has the structure

$$H_k(\mathcal{M}_{\mathbb{S}}) \simeq \mathbb{Z}^{b_k(\mathcal{M}_{\mathbb{S}})} \oplus \mathbb{Z}_{b_{k+1,1}} \oplus \mathbb{Z}_{b_{k+1,2}} \oplus \dots \oplus \mathbb{Z}_{b_{k+1,t_{k+1}}},$$

that is, its torsion coefficients are $b_{k+1,1}, b_{k+1,2}, \dots, b_{k+1,t_{k+1}}$.

The cost of this last computations is that of computing the Smith normal forms $D_1, \dots, D_{n-m}$. The one for D_k can be done (see [39]) with cost

$$\mathcal{O}^{\sim}\big((\min\{O_k, O_{k-1}\})^5 \max\{O_k, O_{k-1}\}\big) = \mathcal{O}^{\sim}\big(K^{6n}\big) = (nD\kappa(f))^{\mathcal{O}(n^2)}$$

(here $\mathcal{O}^{\sim}(g)$ denotes $\mathcal{O}(g\log^c g)$ for some constant c) and hence the same bound holds for the cost of computing all of them. □

The reasoning above extends in a simple manner to compute the homology of $\mathcal{M}_{\mathbb{P}}$. Indeed, projective space $\mathbb{P}^n$ is homeomorphic to the quotient $\mathbb{S}^n/\sim$ where $\sim$ is the equivalence relation that identifies antipodal points. Now consider the map

$$\mathbb{S}^n \xrightarrow{[\ ]} \mathbb{P}^n$$

associating to x its class $[x] = \{x, -x\}$. Because the set $\mathcal{X}$ is closed by taking antipodal points, its image $\overline{\mathcal{X}}$ under [ ] is well-defined and so is the ball in projective space $B_{\mathbb{P}}([x], \varepsilon) := \{B(x,\varepsilon), B(-x,\varepsilon)\}$. Then, the retraction from the union of the balls $B(x,\varepsilon)$ onto $\mathcal{M}_{\mathbb{S}}$ induces a retraction in projective space from the union of the balls $B_{\mathbb{P}}([x],\varepsilon)$ onto $\mathcal{M}_{\mathbb{P}}$.

Also, given $x_{i_1}, \dots, x_{i_k}$ in $\mathcal{X}$, the intersection of $B([x_{i_j}],\varepsilon)$ is nonempty if and only if there exist representatives of $[x_{i_1}], \dots, [x_{i_k}]$ such that the Euclidean balls centered at these representatives have nonempty intersection. That is, if and only if there exist $e_1, \dots, e_k \in \{-1, 1\}$ such that the balls $B(e_1x_{i_1},\varepsilon), B(e_2x_{i_2},\varepsilon), \dots, B(e_kx_{i_k},\varepsilon)$ have nonempty intersection. This can be checked by brute force, by checking each of the 2^k possibilities. Furthermore, if this is the case we get, since $\varepsilon < 1$,

$$\begin{aligned}\bigcap_{1\le j\le k} B_{\mathbb{P}}([x_{i_j}],\varepsilon) &= [B(e_1x_{i_1},\varepsilon) \cap \dots \cap B(e_kx_{i_k},\varepsilon)] \\ &= \big\{B(e_1x_{i_1},\varepsilon) \cap \dots \cap B(e_kx_{i_k},\varepsilon), B(-e_1x_{i_1},\varepsilon) \cap \dots \cap B(-e_kx_{i_k},\varepsilon)\big\}.\end{aligned}$$

Since if $B(e_1x_{i_1},\varepsilon)\cap\ldots\cap B(e_kx_{i_k},\varepsilon)$ contracts to $y\in\mathbb{R}^{n+1}$ then $B(-e_1x_{i_1},\varepsilon)\cap\ldots\cap B(-e_kx_{i_k},\varepsilon)$ contracts to $-y$, then the intersection of $B([x],\varepsilon)$ contracts to $\{y,-y\}=[y]\in\mathbb{P}^n$ and the Nerve Theorem applies: it implies that the nerve $\overline{\mathcal{N}}$ of the family $\{B([x],\varepsilon)\mid[x]\in\overline{\mathcal{X}}\}$ is homotopically equivalent to the union of this family. The reasoning of Proposition 4.3 straightforwardly applies to prove the following result.

Proposition 4.4. *We can compute the Betti numbers $b_0(\mathcal{M}_{\mathbb{P}}),\ldots,b_{n-m}(\mathcal{M}_{\mathbb{P}})$ as well as the torsion coefficients of $\mathcal{M}_{\mathbb{P}}$ with cost*

$$(nD\kappa(f))^{\mathcal{O}(n^2)}. \qquad \square$$

5 On the cost of computing coverings for random systems

The following result is a part of Theorem 21.1 in [10].

Theorem 5.1. *Let $\Sigma\subset\mathbb{R}^{p+1}$ be contained in a real algebraic hypersurface, given as the zero set of a homogeneous polynomial of degree d and, for $a\in\mathbb{R}^{p+1}$, $a\neq 0$,*

$$\mathscr{C}(a):=\frac{\|a\|}{\operatorname{dist}(a,\Sigma)}.$$

Then, for all $t\geq(2d+1)p$,

$$\operatorname*{Prob}_{a\in\mathbb{S}^p}\{\mathscr{C}(a)\geq t\}\ \leq\ 4e\,dp\,\frac{1}{t}$$

and

$$\mathop{\mathbb{E}}_{a\in\mathbb{S}^p}\big(\log_2\mathscr{C}(a)\big)\leq\log_2 p+\log_2 d+\log_2(4e^2). \qquad \square$$

Remark 5.2. For condition numbers over the complex numbers, one can improve the tail estimate in Theorem 5.1 to show a rate of decay of the order of $t^{-2(p+1-\ell)}$ where ℓ is the (complex) dimension of $\Sigma\subset\mathbb{C}^{p+1}$ (see [21, Theorem 4.1]). Over the reals, such an estimate (with the 2 in the exponent removed) has only been proved in the case where Σ is complete intersection [24]. We suspect that a similar estimate holds for $\kappa(f)$.

We define

$$\Sigma_{\mathbb{C}}:=\Big\{f\in\mathcal{H}_{\boldsymbol{d}}[m]\mid\exists\,x\in\mathbb{C}^{n+1}\text{ such that }\sum_{0\leq j\leq n}x_j^2=1,\ f(x)=0\text{ and }\operatorname{rank}(Df(x))<m\Big\}.$$

The discriminant variety $\Sigma_{\mathbb{R}}$ defined in (3) is contained in $\Sigma_{\mathbb{C}}$.

Proposition 5.3. *Let U be a set of $N=\dim_{\mathbb{R}}\mathcal{H}_{\boldsymbol{d}}[m]$ variables. Then there exists a polynomial $G\in\mathbb{Q}[U]\setminus\{0\}$ such that $G|_{\Sigma_{\mathbb{C}}}=0$ and $\deg(G)\leq m^{n+2}(n+1)D^{n+1}$. (Here $G(f)$ for $f\in\Sigma_{\mathbb{C}}$ means specializing G at the coefficients of the polynomials in f.)*

Proof. Observe that for generic $f=(f_1,\ldots,f_m)\in\mathcal{H}_{\boldsymbol{d}}[m]$ the map $x\mapsto Df(x)$, $x\in\mathbb{C}^{n+1}$, is surjective, that is $\operatorname{rank}(Df(x))=m$, and that the condition $\operatorname{rank}(Df(x))<m$ is equivalent to the vanishing of all maximal minors of the matrix $Df(x)\in\mathbb{C}^{m\times(n+1)}$.

For convenience, we write $U = \{u_{i,\alpha} \mid i = 1,\dots,m, |\alpha| = d_i\}$. We consider the general $(n+1)$-variate polynomials of degree d_i,

$$F_i = \sum_{|\alpha|=d_i} u_{i,\alpha} X^\alpha \ \in \mathbb{Q}[U][X], \quad 1 \le i \le m.$$

Let $DF(U,X) \in \mathbb{Q}[U][X]^{m\times(n+1)}$ be the Jacobian matrix of $F = (F_1,\dots,F_m)$ w.r.t. X, and denote by $M_k(U,X)$, $1 \le k \le t$, all its maximal minors. We consider the polynomials

$$\sum_{0\le j\le m} X_j^2 - 1,\, F_i(u_i,X),\, M_k(U,X), \quad 1 \le i \le m,\, 1 \le k \le t. \tag{11}$$

These polynomials have no common zeros in $\overline{\mathbb{Q}(U)}^{n+1}$ because they have no common zeros for a generic specialization of U as mentioned at the beginning of the proof, and we can apply [20, Cor.4.20]. We have

$$\deg_X(F_i) = d_i \le D,\, \deg_X\left(\sum X_j^2 - 1\right) = 2,\, \deg_X(M_k) \le m(D-1),$$
$$\deg_U(F_i) = 1,\, \deg_U\left(\sum X_j^2 - 1\right) = 0,\, \deg_U(M_k) \le m,$$

and therefore there exists $G \in \mathbb{Q}[U] \setminus \{0\}$ such that G belongs to the ideal in $\mathbb{Q}[U,X]$ generated by the polynomials in (11) with

$$\deg_U(G) \le (mD)^{n+1} \sum_{0\le\ell\le n} m \le m^{n+2}(n+1)D^{n+1}.$$

Clearly this polynomial G vanishes on all $f \in \Sigma_{\mathbb{C}}$. □

Corollary 5.4. *Let* $\mathrm{cost}_{\mathbb{S}}(f)$ *and* $\mathrm{cost}_{\mathbb{P}}(f)$ *denote the costs of computing the Betti numbers and torsion coefficients of* $\mathcal{M}_{\mathbb{S}}$ *and* $\mathcal{M}_{\mathbb{P}}$*, respectively. For* f *drawn from the uniform distribution on* $\mathbb{S}(\mathcal{H}_{\boldsymbol{d}}[m]) = \mathbb{S}^{N-1}$ *we have the following:*

(i) *With probability at least* $1 - (nD)^{-n}$ *we have* $\mathrm{cost}_{\mathbb{S}}(f) \le (nD)^{\mathcal{O}(n^3)}$*. Similarly for* $\mathrm{cost}_{\mathbb{P}}(f)$*.*

(ii) *With probability at least* $1 - 2^{-N}$ *we have* $\mathrm{cost}_{\mathbb{S}}(f) \le 2^{\mathcal{O}(N^2)}$*. Similarly for* $\mathrm{cost}_{\mathbb{P}}(f)$*.*

Proof. For all $t \ge \big(2(n+1)m^{n+2}D^{n+1}+1\big)N$, it follows from Theorem 5.1 and Propositions 2.2 and 5.3, that we have

$$\mathop{\mathrm{Prob}}_{f\in\mathbb{S}^{N-1}}\{\kappa(f) \ge t\} \ \le\ 4e\, m^{n+2}(n+1)D^{n+1}N\,\frac{1}{t}.$$

By taking $t = (nD)^{cn}$ for a constant c large enough, we have

$$\mathop{\mathrm{Prob}}_{f\in\mathbb{S}^{N-1}}\{\kappa(f) \ge (nD)^{cn}\} \ \le\ 4e\, m^{n+2}(n+1)D^{n+1}N\,(nD)^{-cn} \le (nD)^{-n}.$$

By Propositions 4.3 and 4.4, for f with $\kappa(f) \le (nD)^{cn}$ we have $\mathrm{cost}_{\mathbb{S}}(f), \mathrm{cost}_{\mathbb{P}}(f) \le (nD)^{\mathcal{O}(n^3)}$. This proves (i).

To prove part (ii) we take $t = 2^{cN}$ for c large enough. Then,

$$\mathop{\mathrm{Prob}}_{f\in\mathbb{S}^{N-1}}\{\kappa(f) \ge 2^{cN}\} \ \le\ 4e\,(n+1)m^{n+2}D^{n+1}N2^{-cN} \le 2^{-N}.$$

Using Propositions 4.3 and 4.4 again, we have that for f such that $\kappa(f) \le 2^{cN}$, $\mathrm{cost}_{\mathbb{S}}(f), \mathrm{cost}_{\mathbb{P}}(f) \le (nD)^{\mathcal{O}(n^2)}2^{\mathcal{O}(n^2N)} \le 2^{\mathcal{O}(N^2)}$, the last as $N \ge \frac{n^2}{2}$. □

6 Remaining proofs

6.1 Proof of Proposition 2.2

We start by defining a fiber version of $\Sigma_{\mathbb{R}}$. For $x \in \mathbb{S}^n$ we let

$$\Sigma_{\mathbb{R}}(x) := \big\{g \in \mathcal{H}_{\boldsymbol{d}}[m] : g(x) = 0 \text{ and } \operatorname{rank}(Dg(x)) < m\big\}.$$

Note that, for all $x \in \mathbb{S}^n$, $\Sigma_{\mathbb{R}}(x)$ is a cone in $\mathbb{R}^N$. In particular, $0 \in \Sigma_{\mathbb{R}}(x)$. The following result is the heart of our proof.

Proposition 6.1. *For all $f \in \mathcal{H}_{\boldsymbol{d}}[m]$ and $x \in \mathbb{S}^n$,*

$$\frac{\|f\|}{\sqrt{2}\,\operatorname{dist}(f, \Sigma_{\mathbb{R}}(x))} \le \kappa(f, x) \le \frac{\|f\|}{\operatorname{dist}(f, \Sigma_{\mathbb{R}}(x))}.$$

Proof. We only need to prove the statement for $f \notin \Sigma_{\mathbb{R}}(x)$. As we saw in Remark 2.1, $\kappa(\lambda f, x) = \kappa(f, x)$ for all $\lambda \neq 0$, and also $\operatorname{dist}(\lambda f, \Sigma_{\mathbb{R}}(x)) = |\lambda| \operatorname{dist}(f, \Sigma_{\mathbb{R}}(x))$. We can therefore assume, without loss of generality, that $\|f\| = 1$.

Because the orthogonal group $\mathscr{O}(n+1)$ in $n+1$ variables acts on $\mathcal{H}_{\boldsymbol{d}}[m] \times \mathbb{S}^n$ and leaves $\mu_{\text{norm}}, \kappa$ and the distance to $\Sigma_{\mathbb{R}}$ invariant, we may assume without loss of generality that $x = e_0 := (1, 0, \ldots, 0)$.

For $1 \le i \le m$ write

$$\begin{aligned} f_i(X) &= \sum_{q=0}^{d_i} X_0^{d_i - q} f_{i,q}(X_1, \ldots, X_n) = X_0^{d_i} f_{i,0} + \sum_{q=1}^{d_i} X_0^{d_i - q} f_{i,q}(X_1, \ldots, X_n) \\ &= X_0^{d_i} f_i(e_0) + X_0^{d_i - 1} \sum_{1 \le j \le n} \frac{\partial f_i}{\partial X_j}(e_0) X_j + Q_i(X) \end{aligned} \tag{12}$$

where in the first line $f_{i,q}$ is a homogeneous polynomial of degree q, and in the second, $\deg_{X_0}(Q_i) \le d_i - 2$. In particular $f_{i,1} = \sum_{1 \le j \le n} \frac{\partial f_i}{\partial X_j}(e_0) X_j$.

We first prove that $\kappa(f, e_0) \le 1/\operatorname{dist}(f, \Sigma_{\mathbb{R}}(e_0))$, or equivalently,

$$\operatorname{dist}(f, \Sigma_{\mathbb{R}}(e_0))^2 \le \kappa(f, e_0)^{-2} = \mu_{\text{norm}}^{-2}(f, e_0) + \|f(e_0)\|^2.$$

Write $f_{i,1}(X_1, \ldots, X_n) = \sqrt{d_i} a_{i1} X_1 + \cdots + \sqrt{d_i} a_{in} X_n$ for suitable a_{ij}. Therefore

$$\frac{\partial f_i}{\partial X_j}(e_0) = \begin{cases} d_i f_i(e_0) & \text{if } j = 0 \\ \sqrt{d_i} a_{ij} & \text{if } j \ge 1. \end{cases}$$

Define

$$h_i := f_i - X_0^{d_i} f_{i,0} = \sum_{q=1}^{d_i} X_0^{d_i - q} f_{i,q}(X_1, \ldots, X_n)$$

for $1 \le i \le m$. Then

$$\|f - h\|^2 = \sum_{i \le m} f_{i,0}^2 = \sum_{i \le m} f_i(e_0)^2 = \|f(e_0)\|^2. \tag{13}$$

In addition $h_i(e_0)=0$ and for $0\le j\le n$,

$$\frac{\partial h_i}{\partial X_j}(e_0)=\begin{cases}\frac{\partial f_i}{\partial X_j}(e_0)-d_if_i(e_0)\;=\;0 & \text{if } j=0\\ \frac{\partial f_i}{\partial X_j}(e_0)=\sqrt{d_i}a_{ij} & \text{if } j\ge 1.\end{cases}$$

Therefore, we have (recall the definition of Δ from §2.1)

$$\Delta^{-1}Df(e_0)=\begin{bmatrix}\sqrt{d_1}\,f_1(e_0) & a_{11} & \dots & a_{1n}\\ \sqrt{d_2}\,f_2(e_0) & a_{21} & \dots & a_{2n}\\ \vdots & & & \vdots\\ \sqrt{d_m}f_m(e_0) & a_{m1} & \dots & a_{mn}\end{bmatrix}$$

and

$$\Delta^{-1}Dh(e_0)=\begin{bmatrix}0 & a_{11} & \dots & a_{1n}\\ 0 & a_{21} & \dots & a_{2n}\\ \vdots & & & \vdots\\ 0 & a_{m1} & \dots & a_{mn}\end{bmatrix}.$$

Let $A=(a_{ij})\in\mathbb{R}^{m\times n}$ so that $\Delta^{-1}Dh(e_0)=[0\,A]$. We know that $\operatorname{rank}(A)\le m$. If $\operatorname{rank}(A)\le m-1$, then $h\in\Sigma_{\mathbb{R}}(e_0)$ and hence, by (13)

$$\operatorname{dist}(f,\Sigma_{\mathbb{R}}(e_0))^2\le\|f-h\|^2=\|f(e_0)\|^2\le\mu_{\text{norm}}^{-2}(f,e_0)+\|f(e_0)\|^2.$$

If $\operatorname{rank}(A)=m$, then (the inequality by [14, Lemma 3]),

$$\mu_{\text{norm}}(f,e_0)=\|(\Delta^{-1}Df(e_0))^{\dagger}\|\le\|(\Delta^{-1}Dh(e_0))^{\dagger}\|=\mu_{\text{norm}}(h,e_0). \tag{14}$$

Because of the Condition Number Theorem [10, Corollaries 1.19 and 1.25] there exists a matrix $P\in\mathbb{R}^{m\times n}$ such that $A+P$ is a non-zero matrix of rank less than m and

$$\|P\|_F=\|A^{\dagger}\|^{-1}=\|[0\;A]^{\dagger}\|^{-1}=\|(\Delta^{-1}Dh(e_0))^{\dagger}\|^{-1}=\mu_{\text{norm}}^{-1}(h,e_0).$$

Let $E=(e_{ij})=\Delta P\in\mathbb{R}^{m\times n}$ and consider the polynomials

$$g_i(X):=h_i(X)+X_0^{d_i-1}\sum_{j=1}^{n}e_{ij}X_j,\quad 1\le i\le m.$$

Then g_i are not all zero, $g_i(e_0)=h_i(e_0)=0$, $\frac{\partial g_i}{\partial X_0}(e_0)=\frac{\partial h_i}{\partial X_0}(e_0)=0$, and $\frac{\partial g_i}{\partial X_j}(e_0)=\frac{\partial h_i}{\partial X_j}(e_0)+e_{ij}=c_{ij}+e_{ij}=\sqrt{d_i}a_{ij}+e_{ij}$ for $1\le j\le n$. It follows that

$$Dg(e_0)=[0\;\Delta A+E]=[0\;\Delta(A+P)]$$

and therefore $\operatorname{rank}(Dg(e_0))<m$. Hence, $g\in\Sigma_{\mathbb{R}}(e_0)$. In addition,

$$\|g-h\|^2=\sum_{\substack{1\le i\le m\\ 1\le j\le n}}\binom{d_i}{d_i-1,1}^{-1}e_{ij}^2=\sum_{\substack{1\le i\le m\\ 1\le j\le n}}d_i^{-1}e_{ij}^2=\sum_{\substack{1\le i\le m\\ 1\le j\le n}}p_{ij}^2=\|P\|_F^2=\mu_{\text{norm}}^2(h,e_0).$$

We conclude as

$$\|g-f\|^2=\|g-h\|^2+\|h-f\|^2\underset{(13)}{=}\mu_{\text{norm}}^{-2}(h,e_0)+\|f(e_0)\|^2\underset{(14)}{\le}\mu_{\text{norm}}^{-2}(f,e_0)+\|f(e_0)\|^2,$$

and hence, $\mathrm{dist}(f,\Sigma_{\mathbb{R}}(e_0))^2 \le \|f-g\|^2 \le \mu_{\mathrm{norm}}^{-2}(f,e_0)+\|f(e_0)\|^2$.

We now prove that $\kappa(f,e_0) \ge \frac{1}{\sqrt{2}\,\mathrm{dist}(f,\Sigma_{\mathbb{R}}(e_0))}$, or equivalently, that

$$2\,\mathrm{dist}(f,\Sigma_{\mathbb{R}}(e_0))^2 \ge \mu_{\mathrm{norm}}^{-2}(f,e_0)+\|f(e_0)\|^2.$$

Let $g \in \Sigma_{\mathbb{R}}(e_0)$ be such that $\mathrm{dist}(f,\Sigma_{\mathbb{R}}(e_0))^2 = \|f-g\|^2$. As in Identity (12), write

$$g_i(X) = X_0^{d_i-1}\sum_{1\le j\le n}\frac{\partial g_i}{\partial X_j}(e_0)X_j + \widetilde{Q}_i(X),$$

where we used that $g(e_0)=0$. From this equality and (12) it follows that

$$f_i - g_i = X_0^{d_i}f_i(e_0) + \left[X_0^{d_i-1}\sum_{1\le j\le n}\left(\frac{\partial f_i}{\partial X_j}(e_0)X_j - \frac{\partial g_i}{\partial X_j}(e_0)X_j\right)\right] + \left[Q_i(X)-\widetilde{Q}_i(X)\right].$$

As the three terms in this sum do not share monomials,

$$\begin{aligned}\|f_i-g_i\|^2 &\ge f_i(e_0)^2 + \sum_{1\le j\le n}\left(\frac{\partial f_i}{\partial X_j}(e_0) - \frac{\partial g_i}{\partial X_j}(e_0)\right)^2\\ &\ge \frac{1}{2}f_i(e_0)^2 + \frac{1}{2}\left[\frac{1}{d_i}f_i(e_0)^2 + \frac{1}{d_i}\sum_{1\le j\le n}\left(\frac{\partial f_i}{\partial X_j}(e_0) - \frac{\partial g_i}{\partial X_j}(e_0)\right)^2\right]\end{aligned}$$

and hence,

$$\|f-g\|^2 \ge \frac{1}{2}\left(\|f(e_0)\|^2 + \left\|\mathrm{diag}\Big(\frac{1}{\sqrt{d_i}}\Big)Df(e_0) - \mathrm{diag}\Big(\frac{1}{\sqrt{d_i}}\Big)Dg(e_0)\right\|_F^2\right).$$

But $\mathrm{rank}\big(\mathrm{diag}(\frac{1}{\sqrt{d_i}})Dg(e_0)\big) < m$, and therefore, by the Eckart-Young theorem,

$$\left\|\mathrm{diag}\Big(\frac{1}{\sqrt{d_i}}\Big)Df(e_0) - \mathrm{diag}\Big(\frac{1}{\sqrt{d_i}}\Big)Dg(e_0)\right\|_F \ge \sigma_m,$$

the smallest singular value of $\mathrm{diag}(\frac{1}{\sqrt{d_i}})Df(e_0)$. On the other hand,

$$\mu_{\mathrm{norm}}(f,e_0)^{-2} = \big\|Df(e_0)^{\dagger}\mathrm{diag}(\sqrt{d_i})\big\|^{-2} = \left\|\Big(\mathrm{diag}\Big(\frac{1}{\sqrt{d_i}}\Big)Df(e_0)\Big)^{\dagger}\right\|^{-2} = \Big(\frac{1}{\sigma_m}\Big)^{-2} = \sigma_m^2.$$

This concludes the proof since

$$\|f-g\|^2 \ge \frac{1}{2}\big(\|f(e_0)\|^2+\sigma_m^2\big) = \frac{1}{2}\big(\|f(e_0)\|^2 + \mu_{\mathrm{norm}}(f,e_0)^{-2}\big)$$

as desired. □

Proof of Proposition 2.2. We can assume again $\|f\|=1$. We note that

$$\mathrm{dist}(f,\Sigma_{\mathbb{R}}) = \min\{\mathrm{dist}(f,g) : g\in\Sigma_{\mathbb{R}}\} = \min\{\mathrm{dist}(f,\Sigma_{\mathbb{R}}(x)) : x\in\mathbb{S}^n\},$$

since $\Sigma_{\mathbb{R}} = \bigcup_{x\in\mathbb{S}^n} \Sigma_{\mathbb{R}}(x)$. Then, using Proposition 6.1,

$$\kappa(f) = \max_{x\in\mathbb{S}^n} \kappa(f,x) \leq \max_{x\in\mathbb{S}^n} \frac{1}{\operatorname{dist}(f,\Sigma_{\mathbb{R}}(x))} = \frac{1}{\min\limits_{x\in\mathbb{S}^n} \operatorname{dist}(f,\Sigma_{\mathbb{R}}(x))} = \frac{1}{\operatorname{dist}(f,\Sigma_{\mathbb{R}})}.$$

Analogously,

$$\kappa(f) = \max_{x\in\mathbb{S}^n} \kappa(f,x) \geq \max_{x\in\mathbb{S}^n} \frac{1}{\sqrt{2}\operatorname{dist}(f,\Sigma_{\mathbb{R}}(x))} = \frac{1}{\sqrt{2}\min\limits_{x\in\mathbb{S}^n} \operatorname{dist}(f,\Sigma_{\mathbb{R}}(x))} = \frac{1}{\sqrt{2}\operatorname{dist}(f,\Sigma_{\mathbb{R}})}.$$

□

6.2 Proof of Proposition 2.3

The following simple quadratic map, which was introduced by S. Smale in [38], is useful in several places in our development,

$$\psi: \ [0,\infty) \to \mathbb{R}, \quad u \mapsto 1-4u+2u^2. \tag{15}$$

It is monotonically decreasing and nonnegative in $[0, 1-\frac{\sqrt{2}}{2}]$.

Lemma 6.2. *Let $u := \|z-y\|\gamma(f,y)$. For all $\varepsilon \in (0,1/2]$, if $u \leq \varepsilon$ then*

$$\mu_{\rm norm}(f,z) \leq \Big(1+\frac{5}{2}\varepsilon\Big)\mu_{\rm norm}(f,y).$$

Proof. As $Df(y)Df(y)^\dagger = \mathrm{Id}_{\mathbb{R}^m}$ we have

$$\begin{aligned}\mu_{\rm norm}(f,z) &= \|f\|\|Df(z)^\dagger\Delta\| \ = \ \|f\|\|Df(z)^\dagger Df(y)Df(y)^\dagger\Delta\| \\ &\leq \|f\|\|Df(z)^\dagger Df(y)\|\|Df(y)^\dagger\Delta\| \ \leq \ \frac{(1-u)^2}{\psi(u)}\mu_{\rm norm}(f,y)\end{aligned}$$

the last inequality by [37, Lemma 4.1(11)]. We now use that

$$\frac{(1-u)^2}{\psi(u)} = 1 + u\Big(\frac{2-u}{1-4u+2u^2}\Big) \leq 1+\frac{5}{2}\varepsilon$$

the last as $u \leq \varepsilon \leq \frac{1}{2}$ and the fact that $\frac{2-u}{1-4u+2u^2} \leq \frac{5}{2}$ in the interval $[0,\frac{1}{2}]$. □

Proof of Proposition 2.3. Because of (4) we have

$$\|y-z\| \leq \frac{2\varepsilon}{D^{3/2}\mu_{\rm norm}(f,y)} = \frac{\varepsilon}{\overline{\gamma}(f,y)} \leq \frac{\varepsilon}{\gamma(f,y)}.$$

Hence, we can apply Lemma 6.2 to deduce the inequality on the right. For the inequality on the left, assume it does not hold. That is,

$$\mu_{\rm norm}(f,z) < \frac{1}{1+\frac{5}{2}\varepsilon}\,\mu_{\rm norm}(f,y) < \mu_{\rm norm}(f,y).$$

Then, $\|y-z\| \leq \frac{\varepsilon}{\mu_{\rm norm}(f,y)} \leq \frac{\varepsilon}{\mu_{\rm norm}(f,z)}$ and we can use Lemma 6.2 with the roles of y and z exchanged to deduce that

$$\mu_{\rm norm}(f,y) \leq \Big(1+\frac{5}{2}\varepsilon\Big)\mu_{\rm norm}(f,z)$$

which contradicts our assumption. □

6.3 Proof of Theorem 2.9

Recall that $\mathcal{Z}$ denotes $f^{-1}(0) \subset \mathbb{R}^{n+1}$ and $\mathcal{M}_{\mathbb{S}} = \mathcal{Z} \cap \mathbb{S}^n$. The idea of the proof is to show that if $p, q \in \mathcal{M}_{\mathbb{S}}$, $p \neq q$, then there are fixed radius balls around p and q such that the normals at p and q to $\mathcal{M}_{\mathbb{S}}$, i.e., the normal spaces of their tangent spaces at $\mathcal{M}_{\mathbb{S}}$, do not intersect in the intersection of the two balls. Either the two points are so far that there will be no intersection between the two balls, or there are close and in that case, $\mathcal{M}_{\mathbb{S}}$ around p can be described as an analytic map by the implicit function theorem. This enables us to analyze the normals at p and q and their possible intersection.

For the rest of this section we fix an arbitrary point $p \in \mathcal{M}_{\mathbb{S}}$, i.e., such that $f(p) = 0$ and $\|p\| = 1$, with a full-rank derivative $Df(p)$ and we set $\gamma_p := \max\{\gamma(f,p), 1\}$.

For any $\varepsilon > 0$ and any linear subspace $H \subset \mathbb{R}^{n+1}$ we denote by $B_{\varepsilon,H}(0)$ the open ε-ball in H centered at 0 and by $B_{\varepsilon,H}(p) := p + B_{\varepsilon,H}(0)$ the same but centered at p. In the special case that $H = \mathbb{R}^{n+1}$ we simply write $B_\varepsilon(0)$ and $B_\varepsilon(p)$.

We recall that, because of Euler's formula, $p \in \ker Df(p)$. We define

$$T := \langle p \rangle^\perp, \qquad H_1 := \ker Df(p) \cap T, \qquad H_2 := \ker Df(p)^\perp \subset T, \qquad H_3 := H_2 + \langle p \rangle,$$

and consider the orthogonal projections $\pi_i : \mathbb{R}^{n+1} \to H_i$ for $i = 1, 2, 3$. Note that H_1, H_2, H_3 are linear spaces of dimension $n - m$, m, and $m + 1$ respectively. In addition, $T = H_1 \perp H_2$ and $\mathbb{R}^{n+1} = H_1 \perp H_3 = \ker Df(p) \perp H_2$, where the symbol $\perp$ denotes orthogonal direct sum.

Proposition 6.3. *Define $c_1 = 0.024$. Then $\mathcal{Z} \cap B_{\frac{c_1}{\gamma_p},T}(p)$ is contained in the graph of a real analytic map $\omega : B_{\frac{c_1}{\gamma_p},H_1}(p) \to H_2$ satisfying $\omega(p) = 0$, $\|D\omega(p+x)\| \leq 2.3\,\|x\|\gamma_p$ and $\|\omega(p+x)\| \leq 1.15\,\|x\|^2\gamma_p$, for all $x \in B_{\frac{c_1}{\gamma_p},H_1}(0)$.*

Figure 1 below attempts to summarize the situation described in Proposition 6.3.

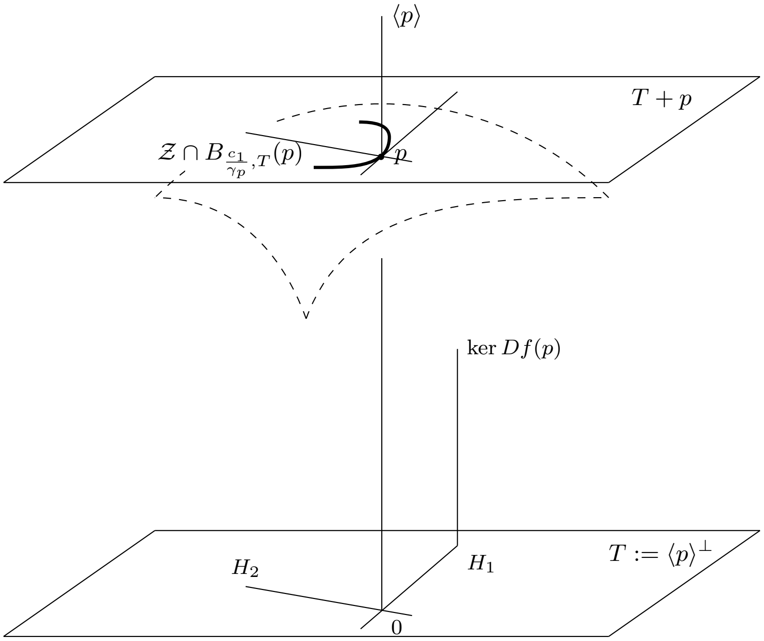
$\langle p\rangle$
$T+p$
$\mathcal{Z}\cap B_{\frac{c_1}{\gamma_p},T}(p)$
p
$\ker Df(p)$
H_2
H_1
$T:=\langle p\rangle^\perp$
0

Figure 1

Proof. The general idea is to first apply (and get explicit bounds for) the Implicit Function Theorem to get a real analytic map $\omega_0 : B_{\frac{c_1}{\gamma_p},\ker Df(p)}(p)\to H_2$ satisfying that $\mathcal{Z}\cap B_{\frac{c_1}{\gamma_p}}(p)$ is contained in the graph of ω_0 with $\omega_0(p)=0$, $\|D\omega_0(p+x)\|\le 2.3\|x\|\gamma_p$ and $\|\omega_0(p+x)\|\le 1.15\gamma_p\|x\|^2$ for all $x\in B_{\frac{c_1}{\gamma_p},\ker Df(p)}(0)$. We then restrict $B_{\frac{c_1}{\gamma_p}}(p)$ to $B_{\frac{c_1}{\gamma_p},T}(p)$ and ω_0 to $H_1\subset\ker Df(p)$ to obtain ω satisfying all the stated conditions.

The process is involved and we describe it as a sequence of claims.

Claim 1. *For all $z\in\mathbb{R}^{n+1}$ such that $u=u(z):=\|z\|\gamma(f,p)<1-\frac{\sqrt{2}}{2}$ the derivative $Df(p+z)|_{H_2}$ of f with respect to H_2 at $p+z$ is invertible.*

Indeed,

$$\|Df(p)|_{H_2}^{-1}Df(p+z)|_{H_2}-\mathrm{Id}_{H_2}\|\le\|Df(p)^\dagger Df(p+z)-\pi_2\|<\frac{1}{(1-u)^2}-1<1$$

the first inequality by properties of Moore-Penrose inverse and the second by [37, Lem. 4.1(9)]. Therefore, by [10, Lem. 15.7], $Df(p)|_{H_2}^{-1}Df(p+z)|_{H_2}$ is invertible, which implies $Df(p+z)|_{H_2}$ invertible as desired. This proves Claim 1.

From now on, since $\mathbb{R}^{n+1}=\ker Df(p)\oplus H_2$, we write indistinctly $f(p)$ or $f(p,0)$ as $p\in\ker Df(p)$, and for $z=(x,y)\in\ker Df(p)\oplus H_2$, $f(p+z)$ or $f(p+x,y)$.

Let

$$\Omega:=\Big\{z=(x,y)\in\ker Df(p)\oplus H_2\mid\|z\|\le\Big(1-\frac{\sqrt{2}}{2}\Big)\frac{1}{\gamma(f,p)}\Big\}.$$

For all $z=(x_0,y_0)\in\Omega$, Claim 1 ensures that $Df(p+x_0,y_0)|_{H_2}$ is invertible. If $f(p+z)=0$, the Analytic Implicit Function Theorem ensures the existence of an open set $U\subset\ker Df(p)$

around x_0, an open set $V \subset H_2$ around y_0 and a real analytic map $\omega_z : p + U \to V$ such that

$$\{(p+x, \omega_z(p+x)) \mid x \in U\} = \{(p+x, y) \in (p+U) \times V \mid f(p+x, y) = 0\}. \qquad (16)$$

Recall the decreasing map ψ defined in (15) and consider also the function

$$\phi(u) := \frac{(1-u)^2}{\psi(u)}\Big(\frac{1}{(1-u)^2} - 1\Big) = \frac{2u(1-\frac{u}{2})}{\psi(u)}.$$

We observe that $\phi(u) < 2.2\,u$ for $u < 0.024 =: c_1$.

Claim 2. *Let $z = (x_0, y_0) \in \Omega$ and $u = u(z) = \|z\|\gamma(f,p)$. If $f(p+z) = 0$ then $\|D\omega_z(p+x_0)\| \le \phi(u)$.*

This is Lemma 5.1 in [37] (with x, y and σ there corresponding to $p, p+z$ and $D\omega_z$ in our context, and in the particular case where $f(p+z) = 0$).

Let now ω_0 be ω_z for $z = (0,0)$, and denote by $0 \in U_0 \subset \ker Df(p)$ and $0 \in V_0 \subset H_2$ the open sets given by the Implicit Function Theorem in last paragraph. We observe that by Claim 2, we have $D\omega_0(p) = 0$ since $\phi(0) = 0$.

Claim 3. *We have*

$$\|D^2\omega_0(p)\| \le 2\gamma(f,p).$$

First note that by the Implicit Function Theorem, $D\omega_0(p) = -(Df(p,0)|_{H_2})^{-1} \circ Df(p,0)|_{\ker Df(p)} = 0$ and $f \circ (\mathrm{Id}, \omega_0) = 0$ in $(p+U_0) \times V_0$, so

$$\begin{aligned} 0 &= D^2(f \circ (\mathrm{Id}, \omega_0))(p) \\ &= D^2 f((\mathrm{Id}, 0), (\mathrm{Id}, 0))(p) + (Df(\mathrm{Id}, \omega_0)(0, D^2\omega_0))(p) \\ &= D^2 f((\mathrm{Id}, 0), (\mathrm{Id}, 0))(p) + Df(p)|_{H_2} D^2\omega_0(p). \end{aligned}$$

Note we have removed the symbol $\circ$ in the compositions from the second line above. We have done, and keep doing, this to make the notation lighter.

So, $D^2\omega_0(p) = -Df(p)^\dagger D^2 f((\mathrm{Id}, 0), (\mathrm{Id}, 0))(p)$ and we obtain the inequality

$$\|D^2\omega_0(p)\| = \|Df(p)^\dagger D^2 f((\mathrm{Id}, 0), (\mathrm{Id}, 0))(p)\| \le 2 \max_{k>1} \left\| Df(p)^\dagger \frac{D^k f(p)}{k!} \right\|^{1/k-1} = 2\gamma(f,p)$$

from the definition of $\gamma(f,p)$. Claim 3 is proved.

Claim 4. *Recall $c_1 = 0.024$. The analytic map $\omega_0 : p + U_0 \to V_0$ can be analytically extended on the open ball $B_{\frac{c_1}{\gamma_p}, \ker Df(p)}(p)$, and for all $p + x \in B_{\frac{c_1}{\gamma_p}, \ker Df(p)}(p)$, its extension –also denoted by ω_0– satisfies the following:*

(i) *$\|D\omega_0(p+x)\| < 2.3\,\|x\|\gamma_p$, and*

(ii) *$\|\omega_0(p+x)\| < 1.15\,\|x\|^2\gamma_p < \frac{0.0007}{\gamma_p}$.*

Since ω_0 is defined in $p + U_0$, there exists r, $0 < r \le \frac{c_1}{\gamma_p}$, such that ω_0 is defined on $B_{r, \ker Df(p)}(p)$ and satisfies Conditions (i) and (ii). To see (i) we note that the equality

$\|D\omega_0(p)\| = 0$ along with Claim 3, the Mean Value Theorem and the fact that ω_0 is defined and C^2 on $p + U_0$ imply that

$$\|D\omega_0(p+x)\| < 2.3\|x\|\gamma_p \tag{17}$$

for x sufficiently close to 0. For (ii), from (17) and the Fundamental Theorem of Calculus, we have

$$\begin{aligned} \|\omega_0(p+x)\| &= \|\omega_0(p+x) - \omega_0(p)\| \ \leq \int_0^1 \|D\omega_0(p+tx)x\| dt \\ &\leq \int_0^1 \|D\omega_0(p+tx)\|\|x\| dt \ \underset{\text{(i)}}{<} \int_0^1 2.3\, t\, \|x\|^2 \gamma_p \, dt \\ &= 1.15\, \|x\|^2 \gamma_p \ \leq \ 1.15\, \frac{c_1^2}{\gamma_p} \ < \ \frac{0.0007}{\gamma_p}. \end{aligned} \tag{18}$$

Let us show that the supremum r_0 of all $0 < r \leq \frac{c_1}{\gamma_p}$ such that $\omega_0(p+x)$ can be analytically extended to $B_{r,\ker Df(p)}(p)$ and satisfies Conditions (i) and (ii) is exactly $r_0 = \frac{c_1}{\gamma_p}$. We assume the contrary, that $r_0 < \frac{c_1}{\gamma_p}$, and show that in that case ω_0 can be extended a little further.

Let x_0 be any point in $\ker Df(p)$ with $\|x_0\| = r_0$. We note that the continuous map ω_0 is bounded on the ball $B_{r_0,\ker Df(p)}(p)$ because Condition (i) holds there. Thus we can consider the limit $y_0 := \lim_{t\to 1^-} \omega_0(p + tx_0)$. Then, reasoning as in (18)

$$\|y_0\| \leq \int_0^1 \|D\omega_0(p+tx_0)x_0\| dt \ < \ 1.15\, \|x_0\|^2 \gamma_p \ \leq \ 1.15\, \frac{c_1^2}{\gamma_p} \ < \ \frac{0.0007}{\gamma_p}.$$

Using the inequality above and the triangle inequality we obtain

$$\begin{aligned} \|(x_0, y_0)\|\gamma_p \quad &< \quad \Big(\|x_0\| + 1.15\, \|x_0\|^2 \gamma_p\Big)\gamma_p = \|x_0\|\gamma_p\Big(1 + 1.15\, \|x_0\|\gamma_p\Big) \\ &\underset{r_0 < \frac{c_1}{\gamma_p}}{\leq} \quad c_1\Big(1 + 1.15\, c_1\Big) < \Big(1 - \frac{\sqrt{2}}{2}\Big). \end{aligned} \tag{19}$$

Hence, $z := (x_0, y_0) \in \Omega$ and $f(x_0, y_0) = 0$. This implies that there exist an open ball $U \subset \ker Df(p)$ and an open set $V \subset H_2$ around x_0 and y_0 respectively, and a real analytic $\omega_z : p + U \to V$ such that (16) holds.

Since $\|y_0\| < 1.15\, \|x_0\|^2 \gamma_p$, by taking a smaller ball U we can further ensure that $\omega_z(p + x) \subset B_{1.15\, \|x\|^2 \gamma_p, H_2}(0)$ for $x \in U$. So, (ii) holds for ω_z on $p + U$. Furthermore, we may use Claim 2, Inequality (19), and the fact that $\phi(u) < 2.2\, u$ for all $0 < u < c_1$ to deduce

$$\|D\omega_z(p + x_0)\| < 2.2\|x_0\|\gamma_p\Big(1 + 1.15\, c_1\Big) < 2.3\, \|x_0\|\gamma_p,$$

so that ω_z also satisfies (i) on $p + U$, possibly taking an even smaller U.

Finally, since the analytic maps ω_0, defined on $p + x \in B_{r_0,\ker Df(p)}(p)$, and ω_z, defined on $p + x_0 + x$ for $x - x_0 \in U$, coincide by (16) on $B_{r_0,\ker Df(p)}(p) \cap U$, which is non-empty and connected, ω_z is an analytic continuation of ω_0 on $p + U$ around $p + x_0$. Let us denote by U_x this open ball around x for $x \in \mathbb{S}_0 := \{x \in \ker Df(p) \mid \|x\| = r_0\}$ and let $\mathcal{U} := \cup_{x \in \mathbb{S}_0} U_x$. Consider the function $\varphi : \mathbb{S}_0 \to \mathbb{R}$ defined by

$$\varphi(x) = \sup\{t \in \mathbb{R} \mid [x, tx)) \subset \mathcal{U}\}.$$

Note that by construction $t > 1$ for every x. As φ is continuous and $\mathbb{S}$ is compact and connected the image $\varphi(\mathbb{S})$ is a closed interval $[\ell, \ell']$ with $1 \leq \ell \leq \ell'$. Furthermore, there exists $x_* \in \mathbb{S}_0$ such that $\varphi(x_*) = \ell$ and, hence, $\ell \geq \frac{r_0 + r_*}{r_0} > 1$, where r_* is the radius of U_{x_*}. It follows that we can extend ω_0 to the open ball in $\ker Df(p)$ centered at p with radius $r_0 + r_* > r_0$ and both (i) and (ii) hold in this ball, a contradiction. This finishes the proof of Claim 4.

Claim 4 shows that for all $x \in B_{\frac{c_1}{\gamma_p}, \ker Df(p)}(0)$ the point $y = \omega_0(p+x)$ satisfies $(p+x, y) \in \mathcal{Z}$ and $\|y\| \leq \frac{0.0007}{\gamma_p}$. We will next see that it is the only point in H_2 satisfying these two conditions. To do so, for each $x \in \ker Df(p)$, we define $g_x : H_2 \to \mathbb{R}^m$ as the restriction of f to $\{p + x\} \times H_2$ so that $g_x(0) = f(p + x)$. Because of Claim 1, for all $y \in H_2$ such that $\|(x, y)\| < \frac{1 - \frac{\sqrt{2}}{2}}{\gamma_p}$, $Df(p + x, y) \mid_{H_2}$ is invertible. In particular, $Dg_x(0) = Df(p + x)|_{H_2}$ is invertible for $\|x\| < \frac{1 - \frac{\sqrt{2}}{2}}{\gamma_p}$.

Claim 5. *For all $x \in \ker Df(p)$ such that $u = u(x) = \|x\|\gamma(f, p) < 1 - \frac{\sqrt{2}}{2}$, we have $\alpha(g_x, 0) \leq \frac{u}{\psi(u)^2}$.*

To show this claim we adapt the proof of [6, Prop. 3, p. 160]. First we verify that $\gamma(g_0, 0) = \gamma(f|_{\{p\} \times H_2}, p) \leq \gamma(f, p)$. To do this we note that

$$\gamma(f, p) := \max_{k>1} \left\| Df(p)^\dagger \frac{D^k f(p)}{k!} \right\|^{\frac{1}{k-1}} = \max_{k>1} \max_{w_1, \ldots, w_k \in \mathbb{S}^n} \left\| Df(p)^\dagger \frac{D^k f(p)}{k!} (w_1, \ldots, w_k) \right\|$$

and

$$\gamma(f|_{\{p\} \times H_2}, p) : = \max_{k>1} \left\| Df|_{\{p\} \times H_2}(p)^{-1} \frac{D^k f|_{p \times H_2}(p)}{k!} \right\|^{\frac{1}{k-1}} = \max_{k>1} \left\| Df(p)|_{H_2}^{-1} \frac{D^k f(p)|_{H_2}}{k!} \right\|^{\frac{1}{k-1}}$$
$$= \max_{k>1} \max_{v_1, \ldots, v_k \in \mathbb{S}^{m-1}} \left\| Df(p)|_{H_2}^{-1} \frac{D^k f(p)|_{H_2}}{k!} (v_1, \ldots, v_k) \right\|^{\frac{1}{k-1}}.$$

Modulo an orthogonal change of basis (that does not modify norms), we can write $Df(p)^\dagger = \begin{pmatrix} 0 \\ Df(p)|_{H_2}^{-1} \end{pmatrix}$. This proves that $\gamma(g_0, 0) \leq \gamma(f, p)$. Also,

$$\beta(g_x, 0) = \|Dg_x(0)^{-1}\, g_x(0)\| \leq \|Df(p + x)|_{H_2}^{-1}\, Df(p)|_{H_2}\| \|Df(p)|_{H_2}^{-1}\, f(p + x)\|.$$

By [37, Lem. 4.1(10)],

$$\|Df(p + x)|_{H_2}^{-1}\, Df(p)|_{H_2}\| \leq \frac{(1 - u)^2}{\psi(u)}$$

while by the multivariate version of [6, Lem. 4(b), p. 161],

$$\|Df(p)|_{H_2}^{-1}\, f(p + x)\| \leq \frac{\|x\|}{1 - \|x\|\gamma(f|_{\{p\} \times H_2}, p)} \leq \frac{\|x\|}{1 - u},$$

since $\beta(f, p) = 0$. This implies $\beta(g_x, 0) \leq \frac{1-u}{\psi(u)}\|x\|$.

Also, in the same way that we verified that $\gamma(g_0, 0) \leq \gamma(f, p)$ we can check that $\gamma(g_x, 0) \leq \gamma(f, p + x)$, and therefore, as in the proof of [6, Prop. 3, p. 162], one gets

$$\gamma(g_x, 0) \leq \frac{\gamma(f, p)}{\psi(u)(1 - u)}. \qquad (20)$$

Multiplying $\beta(g_x,0)$ and $\gamma(g_x,0)$ we conclude that as long as $u=\|x\|\gamma(f,p)<1-\frac{\sqrt{2}}{2}$ we have

$$\alpha(g_x,0)\le\frac{u}{\psi(u)^2}.$$

This proves Claim 5.

Claim 6. *Recall $c_1=0.024$. For all $x\in\ker Df(p)$ with $\|x\|\le\frac{c_1}{\gamma_p}$, there is at most one zero of the map g_x in the ball $\|y\|<\frac{0.044}{\gamma_p}$.*

For $0\le u\le c_1$ one has $0.905\le\psi(u)\le 1$ and $\frac{u}{\psi(u)^2}<0.03$. Thus, by Claim 5, $\alpha(g_x,0)<0.03$ for all $x\in\ker Df(p)$ with $\|x\|\le\frac{c_1}{\gamma_p}$. The second statement in Theorem 2.5 applied to g_x tells us that 0 converges to a zero of g_x and that all points in the ball of radius $\frac{0.05}{\gamma(g_x,0)}$ converge to the same zero. This implies that there is at most one zero of g_x in the ball of radius

$$\frac{0.05}{\gamma(g_x,0)}\underset{(20)}{\ge}\frac{0.05\,\psi(u)(1-u)}{\gamma(f,p)}\ge\frac{0.05\,\psi(0.024)(1-0.024)}{\gamma(f,p)}\ge\frac{0.044}{\gamma(f,p)}$$

which proves Claim 6.

We can now finish the proof of the proposition. Since $B_{\frac{c_1}{\gamma_p}}(p)\subset B_{\frac{c_1}{\gamma_p},\ker Df(p)}(p)\times B_{\frac{0.044}{\gamma_p},H_2}(0)$, it follows from Claims 4 and 6 that $\mathcal{Z}\cap B_{\frac{c_1}{\gamma_p}}(p)$ is included in the graph $\mathrm{Gr}(\omega_0)$ of ω_0. We finally restrict $\mathcal{Z}\cap B_{\frac{c_1}{\gamma_p}}(p)$ to $\mathcal{Z}\cap B_{\frac{c_1}{\gamma_p},T}(p)$, and therefore ω_0 restricts to $\omega: B_{\frac{c_1}{\gamma_p},H_1}(p)\to H_2$, as explained at the beginning of the proof. The bounds for $\|D\omega(p+x)\|$ and $\|\omega(p+x)\|$ follow from Claim 4. □

Lemma 6.4. *Let ω be the map of Proposition 6.3 and define the following continuous map*

$$\Phi: B_{\frac{c_1}{\gamma_p},H_1}(0)\subset H_1\longrightarrow H_1,\quad \Phi(x)=\frac{x}{\|(p+x,\omega(p+x))\|}.$$

Then Φ is a bijection onto its image and satisfies

(i) $\|\Phi(x)\|\ge 0.9997\|x\|$,

(ii) $\|D\Phi(x)^{-1}\|\le 1.0013$.

Proof. If we define the map

$$S: B_{\frac{c_1}{\gamma_p},H_1}(0)\to\mathbb{R},\ S(x)=\|(p+x,\omega(p+x))\|, \tag{21}$$

then $\Phi(x)=\frac{x}{S(x)}$, which implies that Φ maps rays to themselves. To see that Φ is bijective, it is therefore sufficient to see that it is monotone increasing along rays, so we study its derivative along rays and show it is positive.

Let $x=tv$ with v a unit vector and differentiate $\Phi(tv)=\frac{tv}{S(tv)}$ w.r.t. t to obtain

$$\frac{d\Phi}{dt}(tv)=\frac{(1+\|\omega(p+tv)\|^2-t\langle D\omega(p+tv)v,\omega(p+tv)\rangle)}{S(tv)^3}v.$$

As we have

$$\begin{aligned} t|\langle D\omega(p+tv)v, \omega(p+tv)\rangle| &\leq t\|D\omega(p+tv)\|\|\omega(p+tv)\| \underset{\text{Prop. 6.3}}{\leq} 2.3\, t^2\gamma_p\|\omega(p+tv)\| \\ &\underset{t\leq\frac{c_1}{\gamma_p}}{\leq} 2.3\,\frac{c_1^2}{\gamma_p}\|\omega(p+tv)\| < 2\|\omega(p+tv)\| \end{aligned}$$

since $c_1 = 0.024$ and $\gamma_p \geq 1$, it follows that

$$\begin{aligned} 1+\|\omega(p+tv)\|^2 - t\langle D\omega(p+tv)v, \omega(p+tv)\rangle &> 1+\|\omega(p+tv)\|^2 - 2\|\omega(p+tv)\| \\ &= (1-\|\omega(p+tv)\|)^2 \geq 0, \end{aligned}$$

since $\|\omega(p+tv)\| \leq 1.15t^2\gamma_p \leq 1.15c_1^2 < 1$ by Proposition 6.3 for $|t| \leq \frac{c_1}{\gamma_p}$. This shows that Φ restricted to $\{tv\}_{|t|\leq\frac{c_1}{\gamma_p}}$ is strictly monotone, as wanted.

To show the bounds (i–ii) we first note that for any x with $\|x\| < \frac{c_1}{\gamma_p}$, by Proposition 6.3, we have

$$S(x) = (1+\|x\|^2+\|\omega(p+x)\|^2)^{\frac{1}{2}} \leq \sqrt{1+\frac{c_1^2}{\gamma_p^2}+1.15^2\frac{c_1^4}{\gamma_p^2}} \leq \sqrt{1+c_1^2+1.15^2c_1^4} \leq 1.0003,$$

and hence $\|\Phi(x)\| = \frac{\|x\|}{S(x)} \geq 0.9997\|x\|$. This shows (i).

Also, for any $y \in H_1$,

$$DS(x)\,y = \frac{\langle x, y\rangle + \big\langle \omega(p+x), D\omega(p+x)y\big\rangle}{S(x)}.$$

As $\langle x, y\rangle \leq \|x\|\|y\|$ and by Proposition 6.3,

$$\big\langle\omega(p+x), D\omega(p+x)y\big\rangle \leq \|\omega(p+x)\|\big\|D\omega(p+x)\big\|\|y\| \leq 2.65\|x\|^3\gamma_p^2\|y\| \underset{\|x\|<\frac{c_1}{\gamma_p}}{\leq} 2.65\,c_1^2\|x\|\|y\|,$$

we deduce that

$$\big\|DS(x)y\big\| \leq \frac{1}{S(x)}(1+2.65c_1^2)\|x\|\|y\| \leq \frac{1.0016\|x\|\|y\|}{S(x)}.$$

So,

$$\|DS(x)\| \leq \frac{1.0016\|x\|}{S(x)} \leq 1.0016\|x\| \tag{22}$$

since $S(x) \geq 1$.

We now use that $\Phi(x) = \frac{x}{S(x)}$ to derive that, for any $y \in H_1$,

$$D\Phi(x)y = \frac{S(x) - xDS(x)}{S(x)^2}y$$

and, consequently,

$$\big\|D\Phi(x)y\big\| \geq \frac{S(x)-1.0016\|x\|^2}{S(x)^2}\|y\| \underset{S(x)\geq 1}{\geq} \frac{1-1.0016\,c_1^2}{S(x)^2}\|y\| \geq \frac{1-1.0016\,c_1^2}{1.0003^2}\|y\|$$

the last inequality since $S(x) \leq 1.0003$. Therefore,

$$\big\|D\Phi(x)y\big\| \geq 0.9988\|y\|.$$

It follows that the smallest singular value σ of $D\Phi(x)$ satisfies $\sigma \geq 0.9988$ and therefore

$$\big\|D\Phi(x)^{-1}\big\| = \frac{1}{\sigma} \leq 0.0013.$$

This shows (ii). □

In what follows, we denote by ϕ the map $\phi : \mathbb{R}^{n+1} \setminus \{0\} \to \mathbb{S}_n$, $\phi(z) = \frac{z}{\|z\|}$, as we did in Section 2.5.

Lemma 6.5. *For all $\varepsilon \in (0,1)$, $\phi(\mathcal{Z} \cap B_{\varepsilon,T}(p)) = \mathcal{M}_{\mathbb{S}} \cap \phi(B_{h(\varepsilon)}(p))$, where $h(\varepsilon) := \frac{\varepsilon}{\sqrt{1+\varepsilon^2}}$.*

Proof. The worst possible situation corresponds to a point $z \in \mathcal{Z} \cap B_{\varepsilon,T}(p)$ with $\|z-p\| = \varepsilon$. This situation is depicted in Figure 2.

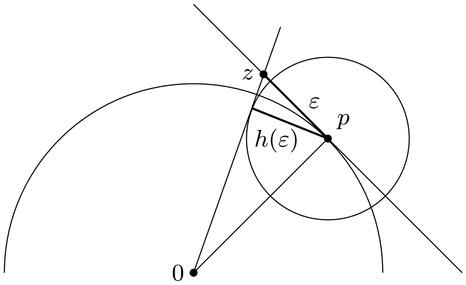

Figure 2

If α denotes the angle at the origin in the figure, then $\varepsilon = \tan(\alpha)$ and $h(\varepsilon) = \sin(\alpha)$ so that $h(\varepsilon) = \sin\arctan(\varepsilon) = \frac{\varepsilon}{\sqrt{1+\varepsilon^2}}$. □

Proposition 6.6. *Define $c_2 = 0.023$ and let Φ be the map defined in Lemma 6.4. Then $\mathcal{M}_{\mathbb{S}} \cap \phi\Big(B_{\frac{c_2}{\gamma_p}}(p)\Big)$ is contained in the graph of a real analytic map*

$$\vartheta \colon p + \Phi\Big(B_{\frac{c_1}{\gamma_p},H_1}(0)\Big) \subset p + H_1 \to H_3$$

satisfying $\vartheta(p) = 0$, $\|D\vartheta(p+x)\| \leq 3.4\|x\|\gamma_p$ and $\|\vartheta(p+x)\| \leq 1.7\|x\|^2\gamma_p$, for all $x \in \Phi\Big(B_{\frac{c_1}{\gamma_p},H_1}(0)\Big)$. Moreover, $B_{\frac{c_2}{\gamma_p},H_1}(0) \subset \Phi\Big(B_{\frac{c_1}{\gamma_p},H_1}(0)\Big)$.

Proof. Write $\mathcal{B} = \Phi\Big(B_{\frac{c_1}{\gamma_p},H_1}(0)\Big)$. By Lemma 6.4, Φ is a bijection onto $\mathcal{B}$. Let ω be the map defined in Proposition 6.3. We recall $H_3 = H_2 + \langle p\rangle$ and π_3 is the projection onto H_3 and define

$$\begin{array}{rcl} \vartheta : p + \mathcal{B} & \to & H_3 \\ p + x & \mapsto & \big(\pi_3\phi(\mathrm{Id},\omega)\big)(p + \Phi^{-1}(x)) - p, \end{array}$$

where $\phi(p+x',y) := \frac{(p+x',y)}{\|(p+x',y)\|}$ for $(x',y) \in H_1 \times H_2$.
Note that $\vartheta(p) = 0$.
For $x' := \Phi^{-1}(x)$ we have

$$\vartheta(p+x) = \pi_3 \phi(p+x', \omega(p+x')) - p.$$

Also note that $x = \Phi(x') = \frac{x'}{\|(p+x',\omega(p+x'))\|} = \pi_1 \phi(p+x', \omega(p+x'))$ implies that, for each $x \in \mathcal{B}$ (or, equivalently, for each $x' \in \Phi^{-1}(\mathcal{B}) = B_{\frac{c_1}{\gamma_p},H_1}(0))$,

$$\begin{aligned} \big(p+x, \vartheta(p+x)\big) \quad &= \quad \big(p+x, \pi_3\phi(p+x',\omega(p+x')) - p\big) = \big(x, \pi_3\phi(p+x',\omega(p+x'))\big) \\ &= \quad \big(\pi_1,\pi_3)\phi(p+x',\omega(p+x')\big) = \phi\big(p+x',\omega(p+x')\big) \end{aligned} \tag{23}$$

modulo the identification $H_1 \times H_3 = H_1 \oplus H_3 = \mathbb{R}^{n+1}$. Identity (23) shows that $\mathrm{Gr}(\vartheta) = \phi(\mathrm{Id},\omega)(B_{\frac{c_1}{\gamma_p},H_1}(p))$.

Now, from Proposition 6.3 we know that

$$\mathcal{Z} \cap B_{\frac{c_1}{\gamma_p},T}(p) \subseteq \mathrm{Gr}(\omega) = (\mathrm{Id},\omega)(B_{\frac{c_1}{\gamma_p},H_1}(p))$$

and therefore, by Lemma 6.5,

$$\mathcal{M}_{\mathbb{S}} \cap \phi\Big(B_{\frac{c_1}{\sqrt{\gamma_p^2+c_1^2}}}(p)\Big) = \phi(\mathcal{Z} \cap B_{\frac{c_1}{\gamma_p},T}(p)) \subseteq \phi(\mathrm{Id},\omega)(B_{\frac{c_1}{\gamma_p},H_1}(p)) = \mathrm{Gr}(\vartheta).$$

As $\gamma_p \geq 1$ we have $\gamma_p^2 + c_1^2 \leq 1.0006\gamma_p^2$ and therefore $\frac{c_1}{\sqrt{\gamma_p^2+c_1^2}} \geq \frac{c_2}{\gamma_p}$ for $c_2 := 0.023$. This shows that $\mathcal{M}_{\mathbb{S}} \cap \phi\Big(B_{\frac{c_2}{\gamma_p}}(p)\Big) \subset \mathrm{Gr}(\vartheta)$.

We now show the bounds. By definition, for all $x \in \mathcal{B}$ one has $\vartheta(p+x) = (\psi_3 \circ \Phi^{-1})(x)$, where $\psi_3 : B_{\frac{c_1}{\gamma_p},H_1}(0) \to H_3$ is defined as

$$\psi_3(x') := \pi_3\phi(\mathrm{Id},\omega)(p+x') - p = \frac{\omega(p+x')}{S(x')} - \Big(1 - \frac{1}{S(x')}\Big)p,$$

where $S(x')$ is defined in (21). Hence, for $x \in \mathcal{B}$,

$$D\vartheta(p+x) = D\psi_3(\Phi^{-1}(x)) \circ D\Phi^{-1}(x). \tag{24}$$

For $x' \in B_{\frac{c_1}{\gamma_p},H_1}(0)$ and any $y \in H_1$ we have

$$D\psi_3(x')y = \bigg(\frac{D\omega(p+x')y}{S(x')} - \frac{DS(x')y\,\omega(p+x')}{S(x')^2}, -\frac{DS(x')y}{S(x')^2}\bigg)^t.$$

Therefore,

$$\begin{aligned} \|D\psi_3(x')\| \qquad &\leq \qquad \frac{\|D\omega(p+x')\|}{S(x')} + \frac{\|DS(x')\|\|\omega(p+x')\|}{S(x')^2} + \frac{\|DS(x')\|}{S(x')^2} \\ &\underset{S(x')\geq 1}{\leq} \qquad 2.3\|x'\|\gamma_p + 1.0016\|x'\|1.15\|x'\|^2\gamma_p + 1.0016\|x'\| \\ &\underset{\|x'\|\leq\frac{c_1}{\gamma_p},\gamma_p\geq 1}{\leq} \qquad \|x'\|\gamma_p\big(2.3 + 1.0016\cdot 1.15\cdot c_1^2 + 1.0016\big) \ \leq \ 3.303\|x'\|\gamma_p \end{aligned}$$

by Proposition 6.3 and Inequality (22).

Going back to (24), using that $D\Phi^{-1}(x) = (D\Phi(\Phi^{-1}(x)))^{-1}$, the above inequality and Lemma 6.4(i,ii), we obtain for any $x \in \mathcal{B}$,

$$\begin{aligned}\|D\vartheta(p+x)\| &\leq \|D\psi_3(\Phi^{-1}(x))\|\,\|D\Phi^{-1}(x)\| \\ &\leq \|D\psi_3(\Phi^{-1}(x))\|\big\|(D\Phi(\Phi^{-1}(x)))^{-1}\big\| \;\leq\; 3.303\|\Phi^{-1}(x)\|\gamma_p \cdot 1.0013 \\ &\leq 3.303 \cdot 1.0004\|x\|\gamma_p \cdot 1.0013 \;\leq\; 3.4\|x\|\gamma_p.\end{aligned}$$

Now, we deduce that

$$\|\vartheta(p+x)\| \leq 1.7\|x\|^2\gamma_p.$$

the same way we deduced the bound for $\|\omega(p+x)\|$ in Proposition 6.3.

Finally, Lemma 6.4 also implies that $B_{\frac{c_2}{\gamma_p},H_1}(0) \subset \Phi\Big(B_{\frac{c_1}{\gamma_p},H_1}(0)\Big)$, since for $\|x'\| = \frac{c_1}{\gamma_p}$, $\|\Phi(x')\| \geq \frac{0.9997c_1}{\gamma_p} \geq \frac{c_2}{\gamma_p}$. □

Lemma 6.7. *Let $\varphi : H_1 \to H_3$ be any linear map and $E \subset H_1 \times H_3$ be the graph of φ. Then,*

(i) $E^\perp = \{(-\varphi^*(v), v) \mid v \in H_3\} \subset H_1 \times H_3$.

(ii) *Let $w \in H_3 \cap \Big((p+x, \vartheta(p+x)) + E^\perp\Big)$ for ϑ the map of Proposition 6.6 and $x \in \Phi\Big(B_{\frac{c_1}{\gamma_p},H_1}(0)\Big) \subset H_1$. Then $\|w-p\| \geq \frac{\|x\|}{\|\varphi\|} - \|\vartheta(p+x)\|$.*

Proof. (i) For all $x \in H_1$ and $v \in H_3$ we have

$$\langle (x,\varphi(x)), (-\varphi^*(v), v)\rangle = \langle x, -\varphi^*(v)\rangle + \langle \varphi(x), v\rangle = -\langle x, \varphi^*(v)\rangle + \langle x, \varphi^*(v)\rangle = 0.$$

This shows that the linear space $\{(-\varphi^*(v), v) \mid v \in H_3\}$, of dimension $\dim(H_3)$, is included in $E^\perp$. The reverse inclusion follows as both spaces have the same dimension.

(ii) As $w \in \big((p+x, \vartheta(p+x)) + E^\perp\big) = \big((x, p+\vartheta(p+x)) + E^\perp\big)$, we use Lemma 6.7(i) to deduce the existence of $v \in H_3$ such that $w = (x, p+\vartheta(p+x)) + (-\varphi^*(v), v) \in H_1 \times H_3$. Hence, since $w \in H_3$, $x - \varphi^*(v) = 0$, i.e., $x = \varphi^*(x)$, and $w = p + \vartheta(p+x) + v$, i.e., $w - p = \vartheta(p+x) + v$. We deduce

$$\|v\| \geq \frac{\|x\|}{\|\varphi^*\|} = \frac{\|x\|}{\|\varphi\|}$$

and, consequently, $\|w-p\| \geq \|v\| - \|\vartheta(p+x)\| \geq \frac{\|x\|}{\|\varphi\|} - \|\vartheta(p+x)\|$. □

Proof of Theorem 2.9. We show that for all points $p, q \in \mathcal{M}_{\mathbb{S}}$ the normals N_p and N_q of $\mathcal{M}_{\mathbb{S}}$ at p and q, i.e., the normal spaces to their tangent planes at $\mathcal{M}_{\mathbb{S}}$, either do not intersect or, if they do, the intersection points lie outside $B_{\frac{c_2}{2\gamma_p}}(p) \cap B_{\frac{c_2}{2\gamma_p}}(q)$. Therefore,

$$\tau(f) \geq \min_{p\in\mathcal{M}_{\mathbb{S}}} \frac{c_2}{2\,\gamma_p} = \frac{c_2}{2\,\max_{p\in\mathcal{M}_{\mathbb{S}}}\gamma_p} \geq \frac{1}{87\,\Gamma(f)},$$

since $\mathcal{M}_{\mathbb{S}}$ is compact and $c_2 = 0.023$.

To prove this statement, we take p to be the point in the preceding development (which is arbitrary on $\mathcal{M}_{\mathbb{S}}$) and divide by cases.

(i) If $\|q-p\| \geq \frac{c_2}{\gamma_p}$, then $B_{\frac{c_2}{2\gamma_p}}(p) \cap B_{\frac{c_2}{2\gamma_p}}(q) = \emptyset$, which implies that the normals N_p and N_q cannot intersect at any point in the intersection of these two balls.
(ii) If $\|q-p\| < \frac{c_2}{\gamma_p}$, then $q \in \mathcal{M}_{\mathbb{S}} \cap \phi\Big(B_{\frac{c_2}{\gamma_p}}(p)\Big)$ is in the hypothesis of Proposition 6.6. Let $x_0 \in \Phi\Big(B_{\frac{c_1}{\gamma_p},H_1}(0)\Big) \subset H_1$ be such that $q = (p+x_0, \vartheta(p+x_0))$. Then $\Big(\frac{c_2}{\gamma_p}\Big)^2 > \|q-p\|^2 = \|x_0\|^2 + \|\vartheta(p+x_0)\|^2 \geq \|x_0\|^2$ implies $x_0 \in B_{\frac{c_2}{\gamma_p},H_1}(0)$, and hence, by the last statement in Proposition 6.6, $p+x_0$ belongs to the domain of ϑ and we may consider its derivative

$$\varphi := D\vartheta(p+x_0) : H_1 \to H_3.$$

Then the graph $E := \mathrm{Gr}(\varphi)$ is a linear subspace of $\mathbb{R}^{n+1}$ and the normal N_q to E at $q = (p+x_0, \vartheta(p+x_0))$ equals $(p+x_0, \vartheta(p+x_0)) + E^{\perp}$. Analogously the normal N_p of $\mathcal{M}_{\mathbb{S}}$ at p equals $p + H_1^{\perp} = p + H_3 = H_3$.

Suppose now that $N_q = (p+x_0, \vartheta(p+x_0)) + E^{\perp}$ intersects $N_p = H_3$ at a point w. Applying Lemma 6.7(ii) and Proposition 6.6 we obtain

$$\begin{aligned} \|w-p\| &\geq \frac{\|x_0\|}{\|D\vartheta(p+x_0)\|} - \|\vartheta(p+x_0)\| \geq \frac{\|x_0\|}{3.4\,\gamma_p\|x_0\|} - 1.7\,\gamma_p\|x_0\|^2 \\ &\geq \frac{1}{3.4\,\gamma_p} - \frac{1.7\,c_2^2}{\gamma_p} = \frac{c_2}{2\gamma_p}\Big(\frac{1}{1.7\,c_2} - 3.4\,c_2\Big) \geq \frac{c_2}{2\gamma_p} \end{aligned}$$

the third inequality as $\|x_0\| \leq \frac{c_2}{\gamma_p}$. This shows that N_p and N_q do not intersect in $B_{\frac{c_2}{2\gamma_p}}(p)$. □

7 On numerical stability

In this last section we deal with the numerical stability of our algorithms. Part (iv) of Theorem 1.1 claims that our algorithms are numerically stable. We now give a precise meaning to this claim.

Numerical stability refers to the effects of finite-precision arithmetic in the final result of a computation. During the execution of such computation real numbers x are systematically replaced by approximations $\mathtt{fl}(x)$ satisfying that

$$\mathtt{fl}(x) = x(1+\delta), \qquad \text{with } |\delta| \leq \varepsilon_{\mathsf{mach}}$$

where $\varepsilon_{\mathsf{mach}} \in (0,1)$ is the *machine precision.* If the algorithm is computing a function $\varphi : \mathbb{R}^p \to \mathbb{R}^q$ a common definition of stability says that the algorithm is *forward stable* when, for sufficiently small $\varepsilon_{\mathsf{mach}}$ and for each input $a \in \mathbb{R}^p$, the computed point $\widetilde{\varphi(a)} \in \mathbb{R}^q$ satisfies

$$\left\|\widetilde{\varphi(a)} - \varphi(a)\right\| \leq \varepsilon_{\mathsf{mach}}\, \|\varphi(a)\|\, \mathsf{cond}(a) P(p,q). \tag{25}$$

Here P is a polynomial (which in practice should be of small degree) and $\mathsf{cond}(a)$ is the condition number of a given by

$$\mathsf{cond}(a) := \lim_{\delta \to 0} \sup_{\|\widetilde{a}-a\| \leq \delta} \frac{\|\varphi(\widetilde{a}) - \varphi(a)\|}{\|\widetilde{a}-a\|} \frac{\|a\|}{\|\varphi(a)\|}. \tag{26}$$

We observe that $\mathsf{cond}(a)$ depends on φ and a but not on the algorithm and that inequality (25) is satisfied in first order whenever the algorithm is *backward stable*, that is, whenever it satisfies that

$$\widetilde{\varphi(a)} = \varphi(\widetilde{a}), \qquad \text{for some } \widetilde{a} \text{ satisfying } \|\widetilde{a} - a\| \leq \|a\|\varepsilon_{\mathsf{mach}}\, P(p,q). \tag{27}$$

These notions are appropriate for a continuous function φ (such as in matrix inversion, the solution of linear systems of equations, the computation of eigenvalues, ...) but not for discrete-valued problems: if the range of φ is discrete (as in deciding the feasibility of a linear program, counting the number of solutions of a polynomial system, or computing Betti numbers) then definition (26) becomes meaningless (see [10, Overture, §6.1, and §9.5] for a detailed exposition of these issues). For these, a now common definition of condition number, pioneered by Jim Renegar [28, 29, 30], consists of identifying the set Σ of ill-posed inputs and taking the condition of a as the relativized inverse of the distance from a to Σ. That is, one takes

$$\mathscr{C}(a) := \frac{\|a\|}{\operatorname{dist}(a,\Sigma)}. \tag{28}$$

Proposition 2.2 shows that our condition number $\kappa(f)$ is bounded by such an expression (with respect to the set of ill-posed inputs $\Sigma_{\mathbb{R}}$).

The idea of stability changes together with the definition of condition. The issue now is not the one underlying (25) —given $\varepsilon_{\mathsf{mach}}$, how good is the computed value— but a different one: *how small does $\varepsilon_{\mathsf{mach}}$ need to be to ensure that the computed output is correct?* The answer to this question depends on the condition of the input at hand, a quantity that is generally not known a priori, and stability results can be broadly divided in two classes. In a *fixed-precision* analysis the algorithm runs with a pre-established machine precision and the users have no guarantee that the returned output is correct. They only know that if the input a is well conditioned (i.e., smaller than a bound depending on $\varepsilon_{\mathsf{mach}}$) then the answer is correct. In a *variable-precision* analysis the algorithm has the capacity to adjust its machine precision during the execution and returns an output which is guaranteed to be correct. Needless to say, not all algorithms may be brought to a variable-precision analysis. But in the last decades a number of problems such as feasibility for semialgebraic systems [19] or for linear programs [18], real zero counting of polynomial systems [15], or the computation of optimal bases for linear programs [11] have been given such analysis.

In all these cases, it is shown that the finest precision $\varepsilon^*_{\mathsf{mach}}$ used by the algorithm satisfies

$$\varepsilon^*_{\mathsf{mach}} = \frac{1}{(p\,\mathscr{C}(a))^{\mathcal{O}(1)}} \tag{29}$$

where p is the *size* of the input and $\mathscr{C}(a)$ is the condition number defined in (28). We can (and will) consider algorithms satisfying (29) to be *stable* as this bound implies that the number of bits in the mantissa of the floating-point numbers occurring in the computation with input $a \in \mathbb{R}^p$ is bounded by $\mathcal{O}(\log_2 p + \log_2 \mathscr{C}(a))$.

It is in this sense that our algorithms are stable.

Proposition 7.1. *The algorithms in Propositions 4.3 and 4.4 computing the homology groups of spherical and projective sets, respectively, can be modified to work with variable-precision and satisfy the following. Their cost, for an input $f \in \mathcal{H}_{\boldsymbol{d}}[m]$, remain*

$$(nD\kappa(f))^{\mathcal{O}(n^2)}$$

and the finest precision $\varepsilon^*_{\mathsf{mach}}$ used by the algorithm is

$$\varepsilon^*_{\mathsf{mach}} = \frac{1}{(nD\kappa(f)\log N)^{\mathcal{O}(1)}}.$$

Sketch of proof. A key observation for the needed modification is that only the routine $\mathsf{Covering}$ needs to work with finite precision. Indeed, we can modify this routine to return a pair $\{\mathcal{X}, \varepsilon\}$ where all numbers, coordinates of points x in $\mathcal{X}$ and ε, are rational numbers (expressed as quotients of integers in binary form). Furthermore, we can do so such that the differences $\|x - \widetilde{x}$ and $|\varepsilon - \widetilde{\varepsilon}|$ between the real objects and their rational approximations are small. Sufficiently small actually for Proposition 2.6 to apply to $(\widetilde{\mathcal{X}}, \widetilde{\varepsilon})$ (recall that Remark 2.7 gives us plenty of room to do so).

From this point on, the computation of the nerve $\mathcal{N}$ and then of the homology groups of either $\mathcal{M}_{\mathbb{S}}$ or $\mathcal{M}_{\mathbb{P}}$ is done symbolically (i.e., with infinite precision). The complexity of the whole procedure, that is, its cost, which now takes account of the size of the rational numbers occuring during the computation, remains within the same general bound in the statement.

We therefore only need to show that a variable-precision version of $\mathsf{Covering}$ can be devised that returns an output with rational components and that satisfies the bounds in the statement. This version is constructed, essentially, as the variable-precision version of the algorithm for counting roots in §5.2 of [15] is constructed in §6.3 of that paper. We do not give all the details here since these do not add anything new to our understanding of the algorithm: we just "make room" for errors by weakening the desired inequalities by a factor of 2; in our case, the inner loop of the algorithm becomes

```
for all $x \in \mathcal{G}_\eta$
    if $\overline{\alpha}(f,x) \le \frac{\alpha_0}{2}$ and $\frac{1}{1000\,\overline{\gamma}(f,x)} \ge r$ and $4.4\,\overline{\beta}(f,x) < r$ then
        $\mathcal{X} := \mathcal{X} \cup \{x\}$
    elsif $\|f(x)\| \ge 2\delta(f,\eta)$ then do nothing
    elsif go to (*)
return the pair $\{\mathcal{X}, \varepsilon\}$ and halt
end for
```

Also, as Proposition 2.6 does neither require the points of $\mathcal{X}$ to belong to the sphere, nor a precise value for ε, there is no harm in returning points (with rational coefficients) close to the sphere and to work with a good (rational) approximation ε of $3.5\sqrt{\mathsf{sep}(\eta)}$. □

We close this section by recalling that the biggest mantissa required in a floating-point computation with input f has $\mathcal{O}(\log_2(nD\kappa(f)\log N))$ bits. If f is randomly drawn from $\mathbb{S}^{N-1}$ this is a random variable. Using the second bound in Theorem 5.1 along with Propositions 2.2 and 5.3 it follows that the expectation for the number of bits in this longest mantissa is of the order of

$$\mathcal{O}\big(n\log_2(Dm) + \log_2 N + \log_2 n\big).$$

This is a relatively small quantity compared with (and certainly polynomially bounded in) the size N of input f.

References

[1] E.L. Allgower and K. Georg. *Numerical Continuation Methods.* Springer-Verlag, 1990.

[2] D. Amelunxen and M. Lotz. Average-case complexity without the black swans. To appear at *J. Compl.*. Available at `arXiv:1512.09290`, 2016.

[3] S. Basu. Computing the top Betti numbers of semialgebraic sets defined by quadratic inequalities in polynomial time. *Found. Comput. Math.*, 8(1):45–80, 2008.

[4] S. Basu, R. Pollack, and M.-F. Roy. Computing the first Betti number of a semi-algebraic set. *Found. Comput. Math.*, 8(1):97–136, 2008.

[5] A. Björner. Topological methods. In R. Graham, M. Grotschel, and L. Lovasz, editors, *Handbook of Combinatorics*, pages 1819–1872. North-Holland, Amsterdam, 1995.

[6] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and Real Computation.* Springer-Verlag, 1998.

[7] L. Blum, M. Shub, and S. Smale. On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines. *Bulletin of the Amer. Math. Soc.*, 21:1–46, 1989.

[8] P. Bürgisser and F. Cucker. Counting complexity classes for numeric computations II: Algebraic and semialgebraic sets. *J. Compl.*, 22:147–191, 2006.

[9] P. Bürgisser and F. Cucker. Exotic quantifiers, complexity classes, and complete problems. *Found. Comput. Math.*, 9:135–170, 2009.

[10] P. Bürgisser and F. Cucker. *Condition*, volume 349 of *Grundlehren der mathematischen Wissenschaften.* Springer-Verlag, Berlin, 2013.

[11] D. Cheung and F. Cucker. Solving linear programs with finite precision: II. Algorithms. *J. Compl.*, 22:305–335, 2006.

[12] G.E. Collins. *Quantifier elimination for real closed fields by cylindrical algebraic deccomposition*, volume 33 of *Lect. Notes in Comp. Sci.*, pages 134–183. Springer-Verlag, 1975.

[13] F. Cucker. Approximate zeros and condition numbers. *J. Compl.*, 15:214–226, 1999.

[14] F. Cucker, H. Diao, and Y. Wei. Smoothed analysis of some condition numbers. *Numer. Lin. Alg. Appl.*, 13:71–84, 2006.

[15] F. Cucker, T. Krick, G. Malajovich, and M. Wschebor. A numerical algorithm for zero counting. I: Complexity and accuracy. *J. Compl.*, 24:582–605, 2008.

[16] F. Cucker, T. Krick, G. Malajovich, and M. Wschebor. A numerical algorithm for zero counting. II: Distance to ill-posedness and smoothed analysis. *J. Fixed Point Theory Appl.*, 6:285–294, 2009.

[17] F. Cucker, T. Krick, G. Malajovich, and M. Wschebor. A numerical algorithm for zero counting. III: Randomization and condition. *Adv. Applied Math.*, 48:215–248, 2012.

[18] F. Cucker and J. Peña. A primal-dual algorithm for solving polyhedral conic systems with a finite-precision machine. *SIAM J. Optim.*, 12:522–554, 2002.

[19] F. Cucker and S. Smale. Complexity estimates depending on condition and round-off error. *Journal of the ACM*, 46:113–184, 1999.

[20] Carlos D'Andrea, Teresa Krick, and Martín Sombra. Heights of varieties in multiprojective spaces and arithmetic Nullstellensätze. *Ann. Sci. Éc. Norm. Supér. (4)*, 46(4):549–627 (2013), 2013.

[21] J. Demmel. The probability that a numerical analysis problem is difficult. *Math. Comp.*, 50:449–480, 1988.

[22] H. Edelsbrunner and J.L. Harer. *Computational topology.* American Mathematical Society, Providence, RI, 2010. An introduction.

[23] E. Kostlan. Complexity theory of numerical linear algebra. *J. of Computational and Applied Mathematics*, 22:219–230, 1988.

[24] M. Lotz. On the volume of tubular neighborhoods of real algebraic varieties. *Proc. Amer. Math. Soc.*, 143(5):1875–1889, 2015.

[25] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39:419–441, 2008.

[26] V. Noferini and A. Townsend. Numerical instability of resultant methods for multidimensional rootfinding. To appear at *SIAM J. Num. Analysis.* Available at `arXiv:1507.00272`.

[27] J. Renegar. On the computational complexity and geometry of the first-order theory of the reals. Part I. *Journal of Symbolic Computation*, 13:255–299, 1992.

[28] J. Renegar. Some perturbation theory for linear programming. *Math. Program.*, 65:73–91, 1994.

[29] J. Renegar. Incorporating condition measures into the complexity theory of linear programming. *SIAM J. Optim.*, 5:506–524, 1995.

[30] J. Renegar. Linear programming, complexity theory and elementary functional analysis. *Math. Program.*, 70:279–351, 1995.

[31] P. Scheiblechner. On the complexity of deciding connectedness and computing Betti numbers of a complex algebraic variety. *J. Complexity*, 23(3):359–379, 2007.

[32] P. Scheiblechner. Castelnuovo-Mumford regularity and computing the de Rham cohomology of smooth projective varieties. *Found. Comput. Math.*, 12(5):541–571, 2012.

[33] M. Shub and S. Smale. Complexity of Bézout's Theorem I: geometric aspects. *Journal of the Amer. Math. Soc.*, 6:459–501, 1993.

[34] M. Shub and S. Smale. Complexity of Bézout's Theorem II: volumes and probabilities. In F. Eyssette and A. Galligo, editors, *Computational Algebraic Geometry*, volume 109 of *Progress in Mathematics*, pages 267–285. Birkhäuser, 1993.

[35] M. Shub and S. Smale. Complexity of Bézout's Theorem III: condition number and packing. *Journal of Complexity*, 9:4–14, 1993.

[36] M. Shub and S. Smale. Complexity of Bézout's Theorem V: polynomial time. *Theoret. Comp. Sci.*, 133:141–164, 1994.

[37] M. Shub and S. Smale. Complexity of Bézout's Theorem IV: probability of success; extensions. *SIAM J. of Numer. Anal.*, 33:128–148, 1996.

[38] S. Smale. Newton's method estimates from data at one point. In R. Ewing, K. Gross, and C. Martin, editors, *The Merging of Disciplines: New Directions in Pure, Applied, and Computational Mathematics.* Springer-Verlag, 1986.

[39] A. Storjohann. Nearly optimal algorithms for computing Smith normal forms of integer matrices. In *Proceedings of the International Symposium on Symbolic and Algebraic Computation (ISSAC'96)*, pages 267–274. ACM Press, 1996.

[40] H.R. Wüthrich. Ein Entscheidungsverfahren für die Theorie der reell-abgeschlossenen Körper. In E. Specker and V. Strassen, editors, *Komplexität von Entscheidungsproblemen*, volume 43 of *Lect. Notes in Comp. Sci.*, pages 138–162. Springer-Verlag, 1976.