Best approximation on semi-algebraic sets and $k$-border rank approximation of symmetric tensors

Shmuel Friedland* and Małgorzata Stawiska†

November 6, 2013

Abstract

In the first part of this paper we study a best approximation of a vector in Euclidean space $\mathbb{R}^n$ with respect to a closed semi-algebraic set $C$ and a given semi-algebraic norm. Assuming that the given norm and its dual norm are differentiable we show that a best approximation is unique outside a hypersurface. We then study the case where $C$ is an irreducible variety and the approximation is with respect to the Euclidean norm. We show that for a general point in $x \in \mathbb{R}^n$ the number of critical points of the distance function of $x$ to $C$ is bounded above by a degree of a related dominant map. If $C$ induces a smooth projective variety $V_P \subset P(\mathbb{C}^n)$ then this degree is the top Chern number of a corresponding vector bundle on $V_P$. We then study the problem when a best $k(\geq 2)$-border rank approximation of a symmetric tensor is symmetric. We show that under certain dimensional conditions there exists an open semi-algebraic set of symmetric tensors for which a best $k$-border rank is unique and symmetric.

Keywords: best $C$-approximation, semi-algebraic sets, critical points, top Chern number, tensors, symmetric tensors, best border rank $k$ approximation.

2010 Mathematics Subject Classification. 14A25, 14C17, 14P10, 15A69, 41A52, 41A65, 53A45.

1 Introduction

Let $\nu : \mathbb{R}^n \to [0, \infty)$ be a norm on $\mathbb{R}^n$. In many applications one needs to approximate a given vector $x \in \mathbb{R}^n$ by a point $y$ in a given closed subset $C \subset \mathbb{R}^n$. Assume that we measure the approximation of $y$ to $x$ by $\nu(x-y)$. Then the distance of $x$ to $C$ with respect to the norm $\nu$ is defined as $\text{dist}_\nu(x, C) := \min\{\nu(x-y), \ y \in C\}$. A point $y^* \in C$ is called a best $\nu$-approximation of $x$ if $\nu(x-y^*) = \text{dist}_\nu(x, C)$. Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^n$ and let $\text{dist}(x, C)$ denote the distance $\text{dist}_{\|\cdot\|}(x, C)$. We call a best $\|\cdot\|$-approximation a best $C$-approximation, or briefly a best approximation.

*Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, Illinois 60607-7045, USA, email: friedlan@uic.edu. This work was supported by NSF grant DMS-1216393

†Mathematical Reviews, 416 Fourth Street, Ann Arbor, MI 48103-4816, USA, email: stawiska@umich.edu
One of important applications is an approximation of a given matrix $A \in \mathbb{R}^{p \times q}$ by a matrix of rank at most $k$ in the Frobenius norm, i.e., the Euclidean norm on $\mathbb{R}^{p \times q}$. This classical problem has a well-known solution, namely, the singular value decomposition (SVD) of $A$ gives a best rank $k$ approximation \[15\]. Recall that finding SVD decomposition with in $\varepsilon$ precision has a polynomial time algorithm. On the other hand, a similar problem for tensors, i.e., approximating a given $d$-mode tensor for $d \geq 3$, is much more difficult \[18\].

In general, the numerical methods for finding best approximation to $x$ in $C$ are based on finding a local minimum of the function $\nu(x - y), y \in C$. For example, to find a best rank one approximation of tensors one uses an alternating least squares method \[12\]. Most of these methods at most will converge to a local minimum point of $f_{k, \nu}(y) := \nu(x - y), y \in C$. We use the abbreviation $f_k$ for the function $f_{x, \| \cdot \|}$.

In a recent paper, the first named author and G. Ottaviani \[13\] considered the above approximation problem in the following special setting. Let $m = (m_1, \ldots, m_d) \in \mathbb{N}^d$. For a field $F$ let $F^m \equiv F^{m_1 \times \cdots \times m_d} := \otimes_{i=1}^d F^{m_i}$. Let $\mathcal{R}(1, m), \mathcal{R}_C(1, m)$ be the closed sets of all decomposable tensors $\otimes_{i=1}^d x_i$ in $\otimes_{i=1}^d \mathbb{R}^{m_i}, \otimes_{i=1}^d \mathbb{C}^{m_i}$ respectively. For $C = \mathcal{R}(1, m)$ a best approximation to $T \in \mathbb{R}^m$ is called a best rank one approximation. Critical points of $f_T(x) := \|T - x\|, x \in \mathcal{R}(1, m)$ were characterized by L.-H. Lim \[20\], which gave rise to the definition of singular vectors and singular values for tensors. The notion of singular vectors can be naturally extended to complex-valued tensors $T \in \mathbb{C}^m$. It was shown in \[13\] that there exists a variety $U \subset \mathbb{C}^m$ such that each $T \in \mathbb{C}^m \setminus U$ has exactly $\delta(\mathcal{R}_C(1, m))$ critical points. Furthermore, a closed formula for $\delta(\mathcal{R}_C(1, m))$ is given in \[13\]. Moreover, a best rank approximation is unique for almost all $T \in \mathbb{R}^m$.

Let $m^{\times d} := (m_1, \ldots, m_d)$ and let $S(d, m, \mathbb{R}) \subset \mathbb{R}^{m \times d}$ denote the $d$-mode symmetric tensors. Recall Banach’s theorem \[2\], which can be stated as follows: Every symmetric tensor $S \in S(d, m, \mathbb{R})$ has a symmetric best rank one approximation. The Banach theorem was re-proved in \[11\]. It is shown in \[13\] that for almost all $S \in S(m, d, \mathbb{R})$ a best rank one approximation is unique and symmetric.

Let $\mathcal{R}(k, m) \subset \mathbb{R}^m$ be the closure of all tensors of rank at most $k$, i.e., all tensors of border rank at most $k$. For $C = \mathcal{R}(k, m)$ we call a best approximation by a best $k$-border rank approximation. It is natural to ask if Banach’s theorem holds for $C = \mathcal{R}(k, m^{\times d}), k \geq 2$. I.e., does there exist a symmetric best $k$-border rank approximation to every symmetric tensor $S \in S(d, m, \mathbb{R})$? We give a partial positive answer to this problem for certain values of $k$. Namely, there exists function $N(m, d) \in \mathbb{N}$, given by \[12,14\], such that for each $k \leq N(d, m)$ there exists an open semi-algebraic set in $S(d, m, \mathbb{R})$ for which a best $k$-border rank approximation is unique and symmetric.

We now summarize the contents of the paper. In \[2\] we discuss the best $\nu$-approximation to a closed subset $C \subset \mathbb{R}^n$. We show that $\text{dist}_{\nu}(. \cup C)$ is Lipschitz on $\mathbb{R}^n$. Assuming that $\nu$ and its dual norm $\nu^*$ are differentiable we show that for a.a. $x \in \mathbb{R}^n$ the $\nu$ approximation of $y^* \in C$ is unique. In \[3\] we assume that $C$ is a semi-algebraic set and the norm $\nu$ is a semi-algebraic function. We show that the function
distν(·, C) is a semi-algebraic function. Furthermore, if ν and ν* are differentiable then the set of x ∈ ℜn which do not have a unique best ν approximation is a semi-algebraic set of dimension less than n. In §4 we discuss the case where C is an irreducible algebraic variety in ℜn and ν(·) = ∥ · ∥. For each x ∈ ℜn we characterize the critical points of the function f_x on C. The notion of critical points of f_x extends naturally to z ∈ ℂn with respect to C_C ⊂ ℂn, the corresponding complex irreducible variety in ℂn. Assume for simplicity of the exposition that C_C is smooth. Let Σ(C_C) ⊂ ℂn × C_C be the variety of the points (z, y) where y is a critical point of f_z. Then Σ(C_C) is an irreducible variety of dimension n. Furthermore, the projection of Σ(C_C) on its first n coordinates is a dominant map. Let δ(C_C) be the degree of this map. Hence there exists a subvariety U ⊂ ℂn such that for each z ∈ ℂn \ U, f_z has exactly δ(C_C) critical points. In particular, for each x ∈ ℂn \ U the real function f_x has at most δ(C_C) critical points in C. The case where C_C has singular points is also discussed. In §5 we assume that C_C is a homogeneous variety such that the projective variety (C_C)_P is smooth. Suppose furthermore that (C_C)_P intersects the projective standard quadric Qn := {∑ni=1 x_i^2 = 0, x = (x_1, ..., x_n) ⊤ ∈ ℂn} transversally. Then we define a vector bundle C over (C_C)_P of rank dim(C_C)_P. We show that the top Chern number of C is δ(V). In §6 we consider a nontrivial finite group of orthogonal matrices acting as linear transformations on ℜn and keeping the irreducible variety C fixed. Let U ⊂ ℂn be the subspace of a the fixed points of G. Assume that dim U ∈ {1, ..., n − 1}. Let U_R := U ∩ ℜn. (The complex dimension of U is equal to the real dimension of U_R.) We study the question when there exists a best C-approximation of x ∈ U_R \ C in C ∩ U_R. (This problem is a generalization of the problem when a best k-border rank approximation of a symmetric tensor S is symmetric.) We show that under the assumption that C ∩ U_R contains smooth points of C there exists an open semi-algebraic set O ⊂ ℜn of dimension n containing all smooth point of C lying in C ∩ U_R such that for each x ∈ O ∩ U_R the best C approximation of x is unique and lies in C ∩ U_R. In §7 we first discuss the notion of symmetric tensor rank of symmetric tensor over a field ℱ. Then we define the notion of a generic tensor in ℱm×d given as a sum of r rank one tensors, and of a generic symmetric tensor S(m, d, ℱ) given as a linear combination of r symmetric rank one tensors. Using Kruskal’s theorem we show that for certain values of r the generic tensors have rank r. In §8 we show that for an integer k, 2 ≤ k ≤ N(m, d), there exists an open semi-algebraic set O ⊂ S(m, d, ℱ) containing all generic symmetric tensors of rank k such that a best k-border rank approximation for each S ∈ O is unique and symmetric.

After we completed this paper we became aware of [3, 8] which are related to some of the results of our paper.

2 Uniqueness of a best ν-approximation

Let ν be a norm on ℜn. Let

\[ B_ν := \{ x ∈ ℜ^n, ν(x) ≤ 1 \}, \quad S_ν := \{ x ∈ ℜ^n, ν(x) = 1 \}, \tag{2.1} \]

denote respectively the unit ball and the unit sphere with respect to ν. It is well known that all norms on ℜn are equivalent, i.e.,
$\kappa_1(\nu) \|x\| \leq \nu(x) \leq \kappa_2(\nu) \|x\|$ for all $x \in \mathbb{R}^n$, where $0 < \kappa_1(\nu) \leq \kappa_2(\nu)$. \hfill (2.2)

Recall that the dual norm $\nu^*$ is defined as $\nu^*(x) = \max_{y \in S_\nu} y^\top x$. Since $\nu$ is a convex function on $\mathbb{R}^n$ it follows that the hyperplane $y^\top z = 1, z \in \mathbb{R}^n$ is a supporting hyperplane of $B_\nu$ at $x \in S_\nu$ if and only if $y^\top x = 1$ and $y \in S_{\nu^*}$. Furthermore, $\nu$ is differentiable at $x$ if and only if the supporting hyperplane at $x$ is unique \cite{22}. We say that $\nu$ is differentiable if it is differentiable on $\mathbb{R}^n \setminus \{0\}$. Since $\nu$ is homogeneous, $\nu$ is differentiable if and only if at each $x \in S_\nu$ the supporting hyperplane is unique. Assume that $\nu$ is differentiable. For $x \neq 0$ denote by $\partial \nu(x)$ the differential of $\nu$ at $x$. We view $\partial \nu(x)$ as a row vector in $\mathbb{R}^n$. So the directional derivative of $\nu$ at $x$ in the direction $u \in \mathbb{R}^n$ is given as $\partial \nu(x)u$. Note that the $\ell_p$-norm on $\mathbb{R}^n$, $\|x_1, \ldots, x_n\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$, is differentiable for $p \in (1, \infty)$.

The following are generalizations of the results in \cite{13 \& 6}.

**Theorem 2.1** Let $\nu$ be a differentiable norm on $\mathbb{R}^n$ and $C \subseteq \mathbb{R}^n$ be a nonempty closed set. Let $U \subset \mathbb{R}^n$ be a subspace with $\dim U \in \{1, \ldots, n\}$ and such that $U$ is not contained in $C$. Let $d(x), x \in U$ be the restriction of $\text{dist}_\nu(\cdot, C)$ to $U$. Then

1. $|\text{dist}_\nu(x, C) - \text{dist}_\nu(z, C)| \leq \nu(x - z)$ for all $x, z \in \mathbb{R}^n$. \hfill (2.3)

Hence $\text{dist}_\nu(x, C)$ is Lipschitz.

2. The function $d(\cdot)$ is differentiable a.e. in $U$.

3. Let $x \in U \setminus C$ and assume that $d(\cdot)$ is differentiable at $x$. Let $\partial d(x)$ denote the differential at $x$ which is viewed as linear functional on $U$. Let $y^* \in C$ be a best $\nu$-approximation to $x$. Then

$$\partial d(x)(u) = \partial \nu(x - y^*)u \text{ for each } u \in U. \hfill (2.4)$$

If $z^*$ is another best $\nu$-approximation to $x$ then $\partial \nu(x - z^*) - \partial \nu(x - y^*)$ is orthogonal to $U$.

Suppose furthermore that $\nu^*$ is differentiable. Then at each point where $\text{dist}_\nu(x, C)$ is differentiable a best $\nu$-approximation is unique.

**Proof.** Let $y^*$ be a best $\nu$-approximation to $x \in \mathbb{R}^n$. Then

$$\text{dist}_\nu(z, C) \leq \nu(z - y^*) = \nu(z - x + x - y^*) \leq \nu(z - x) + \nu(x - y^*) = \nu(z - x) + \text{dist}_\nu(x, C) \Rightarrow \text{dist}_\nu(z, C) - \text{dist}_\nu(x, C) \leq \nu(z - x) \Rightarrow$$

$$\text{dist}_\nu(x, C) - \text{dist}_\nu(z, C) \leq \nu(x - z) = \nu(z - x).$$

This shows \cite{23}. (2.2) yields that $\text{dist}(\cdot, C)$ is Lipschitz.

Let $U \subset \mathbb{R}^n$ be a subspace of $\mathbb{R}^n$ which is not contained in $C$. Let $d(\cdot)$ be the restriction of $\text{dist}(\cdot, C)$ to $U$. As $d(\cdot)$ is Lipschitz, Rademacher’s theorem yields that
it is differentiable almost everywhere in \( U \). Assume that \( d(\cdot) \) is differentiable at \( x \in U \setminus C \). Hence \( x - y^* \neq 0 \). Fix \( u \in U \) and let \( t \in \mathbb{R} \). Then

\[
d(x + tu) = d(x) + \partial d(x)(tu) + to(t) \leq \nu(x + tu - y^*) = \nu(x - y^*) + \partial \nu(x - y^*)tu + to(t).
\]

Recall that \( d(x) = \nu(x - y^*) \). Assume first that \( t > 0 \). Subtract \( d(x) \) from both sides of this inequality and divide by \( t \). Let \( t \searrow 0 \) to deduce \( \partial d(x)(u) \leq \partial \nu(x - y^*)u \). By using the same arguments for \( t < 0 \) we deduce that \( \partial d(x)(u) \geq \partial \nu(x - y^*)u \). This establishes (2.4).

Suppose that \( z^* \) is another best \( \nu \)-approximation to \( x \). Hence in equality (2.4) we can replace \( y^* \) by \( z^* \). Therefore \( \partial \nu(x - z^*) - \partial \nu(x - y^*) \) is orthogonal to \( U \).

Assume now that \( U = \mathbb{R}^n \) and the norms \( \nu \) and \( \nu^* \) are differentiable. Suppose that \( \text{dist}(\cdot, C) \) is differentiable at \( x \in \mathbb{R}^n \setminus C \). Assume to the contrary that \( x \) has two best \( \nu \)-approximations \( y^*, z^* \). So \( \partial \nu(x - y^*) - \partial \nu(x - z^*) \) is orthogonal to \( U \). Hence \( \partial \nu(x - y^*) = \partial (x - z^*) \). Clearly \( \text{dist}(x, C) = \nu(x - y^*) = \nu(x - z^*) > 0 \).

Let \( x_1 = \frac{1}{\nu(x - y^*)} (x - y^*) \), \( x_2 = \frac{1}{\nu(x - z^*)} (x - z^*) \in S_\nu \). So \( \partial \nu(x_1) = \partial \nu(x_2) := w \).

Recall that \( w \in S_\nu \) and hyperplanes \( x_i^\top v = 1 \), \( v \in \mathbb{R}^n \) are supporting hyperplanes of \( B_\nu \) at \( w \). The assumption that \( \nu^* \) is differentiable yields that \( x_1 = x_2 \). So \( x - y^* = x - z^* \) which contradicts the assumption that \( y^* \neq z^* \).

}\]

### 3 Semi-algebraic sets

Recall that a set \( S \subset \mathbb{R}^n \) is called semi-algebraic if it is a finite union of basic semi-algebraic set given by a finite number of polynomial equalities \( p_i(x) = 0 \), \( i \in \{1, \ldots, \lambda \} \) and polynomial inequalities \( q_j(x) > 0 \), \( j \in \{1, \ldots, \lambda' \} \). The fundamental result about semi-algebraic sets (Tarski-Seidenberg theorem) is that a semi-algebraic set \( S \subset \mathbb{R}^n \) can be described by a quantifier-free first order formula (with parameters in \( \mathbb{R} \) considered as a real closed field). It follows that the projection of a semi-algebraic set is semi-algebraic. The class of semi-algebraic sets is closed under finite unions, finite intersections and complements.

A function \( f : \mathbb{R}^n \to \mathbb{R} \) is called semi-algebraic if its graph \( G(f) = \{(x, f(x)) : x \in \mathbb{R}^n \} \) is semi-algebraic. We call a norm \( \nu \) semi-algebraic if the function \( \nu(\cdot) \) is semi-algebraic. Clearly, the norm \( \|(x_1, \ldots, x_n)^\top\|_a := (\sum_{i=1}^n |x_i|^a)^{\frac{1}{a}}, a \geq 1 \) is semi-algebraic if \( a \) is rational. Indeed, assume that \( a = \frac{b}{c} \) where \( b \geq c \geq 1 \) are coprime integers. Then

\[
G(\|\cdot\|_a) = \{(x_1, \ldots, x_n, t)^\top : x_i = \pm y_i^c, y_i \geq 0, i = 1, \ldots, n, t = s^c, s \geq 0, \sum_{i=1}^n y_i^b - s^b = 0 \}.
\]

The following result is well known in the case when \( \nu \) is the Euclidean norm \( \| \cdot \| \). [§1.1] We will sketch the proof in the general case.

**Lemma 3.1** Let \( C \subset \mathbb{R}^n \) be a nonempty closed semi-algebraic set and let \( \nu \) be a semi-algebraic function. Then the function \( f(\cdot) := \text{dist}_\nu(\cdot, C) \) is semi-algebraic.

**Proof.** The graph of \( f \) is characterized as

\[
\{(x, t) \in \mathbb{R}^{n+1} : t \geq 0, \ \forall y \in C \ t \leq \nu(x - y), \ \forall \varepsilon > 0 \ \exists y^* \in C : t + \varepsilon > \nu(x - y^*)\}.
\]
This is a finite intersection of semi-algebraic sets. For example, the set \( \{(x, t) \in \mathbb{R}^{n+1} : \forall y \in C \ t \leq \nu(x - y)\} \) is the complement in \( \mathbb{R}^{n+1} \) of the set which is the projection onto \( \mathbb{R}^{n+1} \) of the set \( B \subset \mathbb{R}^{n+1} \times \mathbb{R}^n \),

\[
B = (\mathbb{R}^{n+1} \times C) \cap u^{-1}(-\infty, 0),
\]

where the function \( u : \mathbb{R}^{n+1} \times \mathbb{R}^n \mapsto \mathbb{R}, \quad u(x, t, y) = \nu(x - y) - t \) is semi-algebraic. Since preimages of semi-algebraic sets by semi-algebraic maps are semi-algebraic, \( B \) is semi-algebraic, and so is its projection. A similar argument applies to other sets in the intersection characterizing the graph of \( f \).

For our next theorem, we will need the following results, proved as parts of Theorem 3.3 (Fact 1 and Fact 2) in [9]. Recall that a semi-algebraic set is called smooth if it is an open subset of the set of smooth points of some algebraic set.

**Proposition 3.2** Let \( X \subset \mathbb{R}^n \) be a semi-algebraic set and let \( f : X \mapsto \mathbb{R}^p \) be a semi-algebraic map. Then there is a Whitney semi-algebraic stratification \( X = \bigcup \Delta_i \) such that the graph \( f|_{\Delta_i} \) is a smooth semi-algebraic set for each \( i \).

**Proposition 3.3** Let \( X \subset \mathbb{R}^n \) be a smooth semi-algebraic set and let \( f : X \mapsto \mathbb{R}^p \) be a map whose graph is a smooth semi-algebraic set. Then the set of points in \( X \) where \( f \) is not differentiable is contained in a closed semi-algebraic set with dimension less than the dimension of \( X \).

Now we can prove an approximation result.

**Theorem 3.4** Let \( C \subset \mathbb{R}^n \) be a semi-algebraic set. Assume that \( \nu \) is a semi-algebraic norm such that \( \nu \) and \( \nu^* \) are differentiable. Then the set of all points \( x \in \mathbb{R}^n \setminus C \), denoted by \( S(C) \), at which the \( \nu \)-approximation to \( x \) in \( C \) is not unique is a semi-algebraic set which does not contain an open set. In particular \( S(C) \) is contained in some hypersurface \( H \subset \mathbb{R}^n \).

**Proof.** Let \( f(x) = \text{dist}_\nu(x, C) \). Since \( f \) is semi-algebraic, the graph of \( G(f) \) is a semi-algebraic set. Hence

\[
G(f) \times C := \{(x^T, t, y^T) \in \mathbb{R}^{2n+1}, x \in \mathbb{R}^n, t = \text{dist}_\nu(x, C), y \in C\} \quad (3.1)
\]

is semi-algebraic. Let

\[
T(f) := \{(x^T, t, y^T, x^T, t, z^T)^T \in \mathbb{R}^{2(2n+1)}, (x^T, t, y^T, t, z^T, )^T, (x^T, t, z^T, ) \in G(f) \times C, \nu(x - y) = \nu(x - z) = t, \|y - z\|^2 > 0\}. \quad (3.2)
\]

Clearly, \( T(f) \) is semi-algebraic. It is straightforward to see that \( S(C) \) is the projection of \( T(f) \) on the first \( n \) coordinates, so \( S(C) \) is semi-algebraic. Theorem [2.1] along with Propositions [3.2] and [3.3] yields that \( S(C) \) does not contain an open set. Hence each basic set of \( S(C) \) is contained in a hypersurface. Therefore \( S(C) \) is contained in a finite union of hypersurfaces which is a hypersurface \( H \).
4 The case of an irreducible variety

Let \( p \in \mathbb{F}[\mathbb{F}^n] \). Denote by \( Z(p) \) the zero set of \( p \) in \( \mathbb{F}^n \). Recall that \( V \subset \mathbb{F}^n \) is a variety (an algebraic set) if there exists a finite number of polynomials \( p_1, \ldots, p_m \in \mathbb{F}[\mathbb{F}^n] \) so that \( V = \cap_{i=1}^m Z(p_i) \). Assume that \( \mathbb{F} = \mathbb{R} \) or \( \mathbb{C} \) and \( V \subset \mathbb{F}^n \) is a variety. If \( V = V_1 \cup V_2 \), where \( V_1 \), \( V_2 \) are strict subvarieties of \( V \), \( V \) is called reducible. Consider the following conditions hold: For each \( y \) \( V \) has a unique decomposition as a finite union of irreducible varieties. Assume that \( V \) is irreducible. Consider the Jacobian \( D(V)(y) = ([\frac{\partial p_i(y)}{\partial x_j}]_{i,j=1}^m) \in \mathbb{C}^{n \times m} \). Denote \( C \subset \mathbb{R}^n \) be a variety. So \( C = \cap_{i=1}^m Z(p_i) \), where \( p_i \in \mathbb{R}[\mathbb{R}^n], i = 1, \ldots, m \). Denote \( C_C := \{ z \in \mathbb{C}^n, p_i(z) = 0, i = 1, \ldots, m \} \). In this section we assume that \( C \) is irreducible. Since \( (\cdot, \cdot) \) is an inner product on \( \mathbb{R}^n \) we deduce that

\[
(C_C)_s \cap (C \setminus \text{Sing } C) = \emptyset. \tag{4.2}
\]

Fix an \( x \in \mathbb{R}^n \) and consider the function \( g_x(y) := \|x - y\|^2 \) restricted to \( C \). We will now study the critical points of \( g_x \).

**Lemma 4.1** Let \( C = \cap_{i=1}^m Z(p_i) \subset \mathbb{R}^n \) be an irreducible smooth variety. Fix an \( x \in \mathbb{R}^n \). Then \( y \) is a critical point of \( g_x = \|x - y\|^2 \) on \( C \) if and only if it satisfies the condition: All minors of order \( n - d + 1 \) of the matrix \( [D(C)(y), x - y] \in \mathbb{R}^{n \times (m+1)} \) are zero.

**Proof.** Note that \( y \) is a critical point of \( g_x \) if and only if \( x - y \) is orthogonal to the tangent space of \( C \) at \( y \). I.e., \( y \) is a critical point if and only if \( x - y \in \text{range } D(C)(y) \). Since \( \text{rank } D(C)(y) = n - d \) we deduce the lemma.

For a general irreducible complex variety \( V \subset \mathbb{C}^n \) of complex dimension \( d < n \) we can define the following:

**Definition 4.2** Let \( V = \cap_{i=1}^m Z(p_i) \subset \mathbb{C}^n \) be an irreducible variety of dimension \( d \in \{1, \ldots, n-1\} \). For each \( x \in \mathbb{C}^n \) we say that \( y \in V \) is a semi-critical point corresponding to \( x \) if all minors of order \( n - d + 1 \) of the matrix \( [D(V)(y), x - y] \) are zero.
Observe that a singular point of \( V \) is a semi-critical point of \( g_s \). Denote by 
\[ \Sigma_1(V) \subset \mathbb{C}^n \times V \subset \mathbb{C}^{2n} \]
the variety of the set of pairs \((x, y) \in \mathbb{C}^n \times V\) where \( x \) and \( y \) satisfy the conditions of Definition 4.2. For \( k \in \{1, \ldots, N\} \) denote by \( \pi_k : \mathbb{C}^N \to \mathbb{C}^k \)
the projection of \( \mathbb{C}^N \) onto the first \( k \) coordinates.

**Lemma 4.3** Let \( V \) be an irreducible complex variety. For \( y \in V \) denote by 
range \( D(V)(y) \subset \mathbb{C}^n \) the column subspace of the matrix \( D(V)(y) \). Let \( \Sigma_0(V) \subset \mathbb{C}^n \times V \) be the set all pairs \((x, y) \in \mathbb{C}^n \times (V \setminus \text{Sing } V)\), where \( x \in y + \text{range } D(V)(y) \). Then 
\[ \Sigma_0(V) \]
is a quasi-algebraic subset of the irreducible variety \( \Sigma(V) \subset \mathbb{C}^n \times V \subset \mathbb{C}^{2n} \) of dimension \( n \), which is an irreducible component of the variety \( \Sigma_1(V) \subset \mathbb{C}^n \times V \subset \mathbb{C}^{2n} \). Each point in \( \Sigma_0(V) \) is a smooth point of \( \Sigma(V) \). If \( V \) is singular then all other irreducible components of \( \Sigma(V) \) are the irreducible components of \( \mathbb{C}^n \times \text{Sing } V \).

**Proof.** Recall that a quasi-algebraic set is a set of the form \( A \setminus B \), where \( A \) and \( B \) are algebraic. Clearly \( \mathbb{C}^n \times \text{Sing } V \) is a subvariety \( \Sigma_1(V) \). Furthermore, 
\[ \Sigma_1(V) \setminus \mathbb{C}^n \times \text{Sing } V = \Sigma_0(V) \]
So \( \Sigma_0(V) \) is quasi-algebraic. Since \( V \setminus \text{Sing } V \) is a manifold of dimension \( d \) it follows that \( \Sigma_0(V) \) is a manifold of dimension \( n \). Hence the Zariski closure of \( \Sigma_0(V) \), which is equal to the closure of \( \Sigma_0(V) \) in the standard topology in \( \mathbb{C}^n \), is an irreducible variety \( \Sigma(V) \). The dimension of \( \Sigma(V) \) is \( n \) and each point in \( \Sigma_0(V) \) is a smooth point of \( \Sigma(V) \). Observe next that if \( V \) is not smooth then the dimension of each irreducible component of \( \mathbb{C}^n \times \text{Sing } V \) has dimension at least \( n \). Hence no irreducible component of \( \mathbb{C}^n \times \text{Sing } V \) is a subvariety of \( \Sigma(V) \).

Thus 
\[ \Sigma_1(V) = \Sigma(V) = \Sigma_0(V) \]
if \( V \) is smooth. \( \square \)

If \( V \) has singular points then the decomposition of \( \Sigma_1(V) \) into its irreducible components consists of \( \Sigma(V) \) and the irreducible components of \( \mathbb{C}^n \times \text{Sing } V \).

In the above notation we also have:

**Theorem 4.4** Let \( C \subset \mathbb{R}^n \) be an irreducible variety of dimension \( d \in \{1, \ldots, n-1\} \). Assume that \( V = C \). Let \( V_s \) be given by \( \{1,1\} \). Then \( V_s \) is a strict quasi-algebraic set of \( V \). For each \( y \in V \setminus (V_s \cup \text{Sing } V) \) the map \( \pi_n : \Sigma(V) \to \mathbb{C}^n \)
\( \pi_n \) is locally 1-to-1 at \((y, y)\). In particular, \( \pi_n \) is dominant. Let
\[ U := \{ x \in \mathbb{C}^n, (x, y) \in \Sigma(V), y \in V_s \cup \text{Sing } (V) \}; \]
\[ X := \pi_n^{-1}(\mathbb{C}^n \setminus U) \cap \Sigma(V). \]

Then \( \pi_n^{-1}(x) \) consists of exactly \( \delta(V) \) distinct points \((x, y_1(x)), \ldots, (x, y_{\delta(V)}(x)) \in \Sigma(V) \)
for each \( x \in \mathbb{C}^n \setminus U \), i.e., \( X \) is \( \delta(V) \)-covering of \( \mathbb{C}^n \setminus U \). Hence the field of rational functions over \( \Sigma(V) \) is a \( \delta(V) \)-extension of the field of rational functions over \( \mathbb{C}^n \).

**Proof.** Clearly the map \( \pi_n : \Sigma(V) \to \mathbb{C}^n \) is a polynomial map. Assume that \( y_0 \) is a smooth point of \( V \). We show that \( \pi_n \) is locally 1-to-1 at \((y_0, y_0)\) if and only if \( y_0 \not\in V_s \). The local coordinates of \( V \) can be identified with the local coordinates \( z \in \mathbb{C}^d \) in the neighborhood of the origin. So for \( y \) in a neighborhood of \( y_0 \in V \) one has \( y(z) = y_0 + u(z) \). Let \( F(y) \) be a fixed \( n \times (n-d) \)-submatrix of \( D(V)(y) \) of rank \( n-d \) in the neighborhood of \( y_0 \) in \( V \). So the range \( D(V)(y(z)) \)
is \( \{ F(y(z))w, w \in \mathbb{C}^{n-d} \} \) for \( y(z) = y_0 + u(z) \). Note that \( u(z) = Az + O(\|z\|^2) \)
and $F(y(z)) = B + O(||z||)$. Then the projection of $\Sigma(V)$ in the neighborhood of $(y_0, y_0)$ onto the first $n$ coordinates is $y_0 + Az + Bw + O(||z||^2 + ||w||^2)$. Hence the Jacobian of $\pi_n$ at $(y_0, y_0)$ is the matrix $[A B]$. So $\pi_n$ is locally 1-to-1 if and only if rank $[A B] = n$, i.e., range $D(V)(y_0) \cap (\text{range } D(V)(y_0))^\perp = \{0\}$. Thus $V_s$ is the set of all points in $V \setminus \text{Sing } V$ which satisfy $\det [A B] = 0$. Hence $V_s$ is a quasi-algebraic.

(1.2) yields that $V_s$ is a strict quasi-algebraic subset of $V$. Furthermore, $\pi_n$ is locally 1-to-1 at $(y_0, y_0)$ for each $y_0 \in C \setminus \text{Sing } C$. Therefore $\pi_n$ is a dominant map. The degree of this map is $\delta(V)$. So $\pi_n : X \rightarrow C^n \setminus U$ is a covering map. It is degree is $\delta(V)$. Thus $\pi_n^{-1}(x)$ consists of exactly $\delta(V)$ distinct points $(x, y_1(x)), \ldots, (x, y_{\delta(V)}) \in \Sigma(V)$ for each $x \in C^n \setminus U$. Therefore the fields of rational functions over $\Sigma(V)$ is a finite extension of the field of rational functions over $C^n$ of degree $\delta(V)$.

Proof. Recall that the projection of $\pi_n^{-1}(x) \cap \Sigma(V_i)$ onto the last $n$ coordinates is a set of cardinality $\delta(V_i)$ in $C^n$ for $z \in C^n \setminus U_i$. Hence $|\pi_n^{-1}(z) \cap \Sigma(V_i)| = \delta(V_i)$ for $z \in C^n \setminus U_i$ for $i \in \{0, \ldots, k\}$.

Let $C \subset \mathbb{R}^n$ be an irreducible variety of dimension $d \in \{1, \ldots, n-1\}$. Recall the stratification of $C$ into a union of smooth quasi-algebraic sets. With this stratification we associate the corresponding irreducible varieties in $\mathbb{R}^n$ as follows: Denote $C_0 := C$. We associate with $C$ the following directed tree $T(C)$. The vertices of the tree are labeled by nonnegative integers from 0 to $k \geq 0$. Assume that $i < j$. We say that a vertex $j$ is a descendant of the vertex $i$ (and that there is a directed edge from $i$ to $j$) if $C_j$ is an irreducible component of $\text{Sing } C_i$. Denote by $N\{i\} \subseteq \{i+1, \ldots, k\}$ the set of all descendants of $i$. Each $i \in \{0, \ldots, k\}$ corresponds to an irreducible subvariety $C_i$ of $C$. The vertex $i$ is a leaf of $T(C)$, i.e., $i$ has no descendants, if and only if $C_i$ is smooth. Otherwise, $\cup_{j \in N\{i\}} C_j$ is the decomposition of $\text{Sing } C_i$ into a union of its irreducible components. Then

$$C = \cup_{i \in \{0, \ldots, k\}} C_i \setminus \text{Sing } C_i,$$

(4.5)

is the smooth stratification of $C$.

**Definition 4.5** Let $C \subset \mathbb{R}^n$ be an irreducible variety. Consider the stratification (4.5). Let $V_i := (C_i)_C$ for $i = 0, \ldots, k$. Denote $V_0$ by $V$. A point $y \in V$ is called a **critical point corresponding to** $x \in C^n$ if $(x, y) \in \Sigma(V_i)$ for some $i \in \{0, \ldots, k\}$.

We now explain briefly our definition. Assume that $y \in C \setminus \text{Sing } C$. Then $y$ is a critical point corresponding to $x \in \mathbb{R}^n$ if and only if $y$ is a semi-critical point as defined in Definition 1.2. Assume that $y \in \text{Sing } C$. So $y$ is in $C_i$ (one of the irreducible component of $\text{Sing } C$). Suppose furthermore that $y_i \in C_i \setminus \text{Sing } C_i$. Then $y$ can be viewed as a critical point of $g_x|C$ if $y$ is a critical point of $g_x|C_i$.

**Theorem 4.6** Let $C \subset \mathbb{R}^n$ be a real irreducible variety. Assume that the smooth stratification of $C$ is given by (4.5). Let $V_i := (C_i)_C$ for $i = 0, \ldots, k$. Let $U_i \subset C^n$ be the variety defined in Theorem 1.2 for $V = V_i$. Let $W := \cup_{i \in \{0, \ldots, k\}} U_i$. For each $x \in \mathbb{R}^n \setminus W$ let $Y(x)$ be the projection of $\cup_{i \in \{0, \ldots, k\}} \pi_n^{-1}(x) \cap (\mathbb{R}^n \times C)$ onto the last $n$ coordinates. Then $Y(x)$ is a nonempty finite set of $C$ of cardinality at most $\sum_{i=0}^k \delta(V_i)$. Furthermore

$$\text{dist}(x, C) = \min_{y \in Y(x)} \|x - y\|.$$  

(4.6)

Proof. Recall that the projection of $\pi_n^{-1}(x) \cap \Sigma(V_i)$ onto the last $n$ coordinates is a set of cardinality $\delta(V_i)$ in $C^n$ for $z \in C^n \setminus U_i$. Hence $|\pi_n^{-1}(z) \cap \Sigma(V_i)| = \delta(V_i)$ for $z \in C^n \setminus U_i$ for $i \in \{0, \ldots, k\}$.
Let \( x \in \mathbb{R}^n \setminus W \). Then \( |Y(x)| \leq \sum_{i=0}^k \delta(V_i) \). Assume that \( \text{dist}(x, C) = \|x - v\| \) for some \( v \in C \). Clearly, \( v \in C_i \setminus \text{Sing} C_i \) for some \( i \in \{0, \ldots, k\} \). Hence \( \text{dist}(x, C) = \text{dist}(x, C_i) = \|x - v\| \). Therefore \( v \) is a critical point of \( \|x - u\| \), where \( u \in C_i \). Thus \( (x, v) \in \Sigma_0(V_i) \subset \Sigma(V_i) \). So \( v \in Y(x) \). Furthermore, (4.6) holds.

**Theorem 4.7** Let \( C \subset \mathbb{R}^n \) be an irreducible variety. Let \( V = C_C \) and assume that \( V_s \) is defined as in Theorem 4.4. Let \( y \in V \setminus (V_s \cup \text{Sing} V) \). Then there exists an open neighborhood \( O(y) \subset \mathbb{C}^n \) such that for each \( x \in O(y) \) there exists a unique point \( y(x) \) such that \( \Sigma(C_C) \cap ([x] \times O(y)) = \{x, y(x)\} \). In particular, if \( y \in C \setminus \text{Sing} C \) and \( x \in O(y) \cap \mathbb{R}^n \) then \( \text{dist}(x, C) = \|x - y(x)\| \).

**Proof.** Theorem 4.4 claims that for each \( y \in V \setminus (V_s \cup \text{Sing} V) \) the projection \( \pi_n \) is locally 1-to-1 at \( (y, y) \). Hence there exists an open neighborhood \( O(y) \subset \mathbb{C}^n \) of \( y \) such that \( \pi_n: (O(y) \times V) \cap \Sigma(V) \to O(y) \) is a diffeomorphism.

Assume that \( y \in C \setminus \text{Sing} C \). Theorem 4.4 yields that \( y \in V \setminus (V_s \cup \text{Sing} V) \). The proof of Theorem 4.4 implies that the Jacobian of \( \pi_n \) at \( (y, y) \) is a real matrix. Hence \( y(x) \in C \) for each \( x \in O(y) \cap \mathbb{R}^n \). Choose \( O(y) \), a small enough open neighborhood of \( y \), so that \( \text{dist}(x, C) = \|x - y(x)\| \) for each \( x \in O(y) \cap \mathbb{R}^n \). \( \square \)

## 5 \( \delta(V) \) as the top Chern number

Assume that \( V \subset \mathbb{C}^n \) is a homogeneous irreducible variety. Then \( V \) induces a projective variety \( V_{\mathbb{P}} \subset \mathbb{P}(\mathbb{C}^n) \). Note that \( \dim V_{\mathbb{P}} = \dim V - 1 = d - 1 \). It is not difficult to show that \( \Sigma(V) \) is also a homogeneous variety in \( \mathbb{C}^n \times V \). Recall that \( V_{\mathbb{P}} \) is smooth if and only if \( V \setminus \{0\} \) is smooth. Assume that \( V_{\mathbb{P}} \) is smooth. We show that the number \( \delta(V) \) is equal to the top Chern number of a certain vector bundle associated with \( V_{\mathbb{P}} \) under suitable conditions. Results of this type are discussed in \( [3, 13, 8] \).

View \( V_{\mathbb{P}} \) as \( V' := V \setminus \{0\} \) where we identify \( y \in V \) with \( ty, t \in C \setminus \{0\} \). Thus \( [y] \in V_{\mathbb{P}} \) is the line spanned by \( y \) in \( \mathbb{C}^n \). View \( \mathbb{C}^n \) as a trivial \( n \)-dimensional vector bundle over \( V' \). Let \( y \in V', t \in \mathbb{C} \setminus \{0\} \). For \( y \in V' \) denote by \( T_{V', y} \subset \mathbb{C}^n \) the tangent space of \( V \) at \( y \). Then \( T_{V', y} = T_{V', ty} \). Furthermore, \( y \in T_{V', y} \). In what follows we assume that we chose \( y \in V' \) as a representative for \( [y] \). Let \( B \) be the direct sum of the tangent bundle of \( V_{\mathbb{P}} \) with the tautological line bundle over \( \mathbb{P}(\mathbb{C}^n) \), denoted as \( O(-1) \). Note that the rank of \( B \) is \( d \). Denote by \( B' \) the dual vector bundle of linear transformations from \( B \) to \( \mathbb{C} \). Let \( \psi \in H^0(B') \) be the following global section in \( B' \): \( \psi_Y \in (T_{V_{\mathbb{P}}} \oplus O(-1))[y], \psi_Y(z) = y^\top z, z \in T_{V', y} \). Clearly, \( \psi_Y = 0 \) if and only if \( y \) belongs to the normal bundle of \( V' \) at \( y \). In particular \( y^\top y = 0 \).

Let \( Q_n \) be the standard hyperquadric \( Q_n = \{z \in \mathbb{C}^n, z^\top z = 0\} \). We say that \( V_{\mathbb{P}} \) and \( (Q_{n})_{\mathbb{P}} \) intersect transversally if for each \( y \in V' \cap Q'_n \) the intersection of the normal bundles of \( V' \) and \( Q'_n \) at \( y \) is \( \{0\} \).

**Proposition 5.1** Let \( V \subset \mathbb{C}^n \) be a homogeneous variety. Assume that \( V_{\mathbb{P}} \) is smooth. Let \( B' \) and \( \psi \in H^0(B') \) be respectively the vector bundle and the global section defined above. Then \( \psi \) vanishes nowhere if and only if \( V_{\mathbb{P}} \) and \( (Q_{n})_{\mathbb{P}} \) intersect transversally.
Proof. Clearly, the normal bundle of \((Q_n)\| y\) at \(y\) is given by \(\text{span}(y)\). Assume that \(y \in V' \cap Q_n'\). Then the normal bundles of \(V'\) and \(Q_n'\) at \(y\) intersect transversally if and only if \(y\) is not in the normal bundle of \(V'\). That is, \(\psi_y \neq 0\). □

We now give two simple examples of smooth projective varieties that intersect \(Q_n\) transversally. Let \(C \subset \mathbb{R}^n\) be a linear subspace and denote \(V = C_C \subset \mathbb{C}^n\). It is straightforward to see that that \(V_p\) and \((Q_n)\| p\) intersect transversally. (Choose an orthonormal basis in \(C\).) Consider next a quadric of the form \(K := \{\sum_{i=1}^n a_i x_i^2 = 0, x = (x_1, \ldots, x_n) \in \mathbb{C}^n\}\), where \(a_i \neq 0, i = 1, \ldots, n\). Then \(K_p\) and \((Q_n)\| p\) intersect transversally if and only if \(a_i \neq a_j\) for \(i \neq j\).

In the remaining part of this section we assume that \(V_p\) and \((Q_n)\| p\) intersect transversally. For each \(x \in \mathbb{C}^n\) we define a global section \(\phi_x \in H^0(B')\) as follows: \(\phi_x(y)\) is given by \(\phi_x(y)(z) = x^\top z, z \in T_{y} V\). Let \(\wedge^2 B'\) be the the exterior product of order 2 of the vector bundle \(B'\) with itself. That is, \(\wedge^2 B'_{y}\) is spanned by all \(f \wedge g\) for \(f, g \in B'_{y}\). Note that rank \(\wedge^2 B' = \binom{n}{2}\). Fix the section \(\psi \in H^0(B')\) that was defined above. Denote by \(C \subset \wedge^2 B\) the subbundle induced by all global sections \(F_x := \phi_x \wedge \psi \in H^0(\wedge^2 B')\). Since \(\psi\) vanishes nowhere it follows that \(C\) has rank \(\dim V_p = d - 1\). We can view \(F_x\) as a global section in \(C\). Let \(W := \{F_x, x \in \mathbb{C}^n\}\). For each \(y \in V'\) let \(W_y\) be the subspace generated by \(F_{x,y} := F_x \mid y\). Then \(W_y = C_{y}\).

Recall that we can associate with \(C\) the corresponding Chern classes \(c_j(C), j = 1, \ldots, d - 1\) \([16]\). Since rank \(C = \dim V_p\) it follows that the top Chern class of \(C\), \(c_{\dim V_p}(C)\), is of the from \(\gamma(C) \omega\). Here \(\omega \in H^{2 \dim V_p}(V_p, \mathbb{Z})\) is the volume form on \(V_p\) such that \(\omega\) is a generator of \(H^{2 \dim V_p}(V_p, \mathbb{Z})\) and \(\gamma(C) \in \mathbb{Z}\). The number \(\gamma(C)\) is called the top Chern number of \(C\).

**Theorem 5.2** Let \(C \subset \mathbb{R}^n\) be an irreducible variety. Denote \(V = C_C\) and assume that \(V\) is a homogenous variety such that \(V_p\) is smooth. Let \(\delta(V)\) be the number of nonzero semi-critical points for \(x \in \mathbb{C}^n \setminus U\), (see Definition 4.2 and Theorem 4.7) Assume that \(V_p\) and \((Q_n)\| p\) intersect transversally. Let \(C\) be the vector bundle over \(V_p\) of rank \(\dim V_p\) defined above. Then \(\delta(V)\) equals to the top Chern number of \(C\):

\[
\delta(V) = \gamma(C). \tag{5.1}
\]

**Proof.** Our theorem follows from “Bertini type” theorem or Generic Smoothness Theorem \([13]\) Corol. III 10.7 and \([14]\) Example 3.2.16]. See \([13]\) §2.5 for more details. Let \(W \subset H^0(C)\) be defined as above. Since rank \(C = \dim V_p\) and \(W\) generates \(C\), \([13]\) Theorem 2] yields that a generic \(F_x \in W\) has \(\gamma(C)\) zeros. We now show that a generic \(F_x\) has \(\delta(V)\) zeros.

Suppose first that \(C\) is a subspace. As we pointed out above, \(V_p\) and \((Q_n)\| p\) intersects transversally. Let \(T : \mathbb{R}^n \to C\) be the orthogonal projection on \(C\). Then \(\Sigma(V) = \{x, T x, x \in \mathbb{C}^n\}\), i.e., \(\delta(V) = 1\). Let \(\tilde{V} : = \{x \in \mathbb{C}^n, x^\top y = 0, \forall y \in V\}\) be the orthogonal complement of \(V\). Let \(x \notin \tilde{V}, y \in V'\). Suppose that \(F_{x,y} = 0\). This means that the linear functionals \(\phi_x, \phi_y : V \to C\) are linearly dependent. Since \(\phi_y\) is a nonzero linear functional it follows that \(x - ay\) is orthogonal to \(V\). Hence \(T(x) = ay\). Note that \(a \neq 0\) as \(x \notin \tilde{V}\). Vice versa, if \(T(x) = ay\) then \(F_x\) vanishes at \(y\). Hence \(F_{x,y}\) vanishes only at \(y\).
Assume now that $C$ is not a subspace. Then $0 \in \text{Sing } V$. Let $(V_F)^\vee \subset P(\mathbb{C}^n)$ be the dual variety of $V_F$ [17]. That is, $[z] \in P(\mathbb{C}^n)$ is in $(V_F)^\vee$ if and only if there is a point $y \in V'$ such that $z \in \text{range } D(V)(y)$. Denote by $V^\vee \subset \mathbb{C}^n$ the homogeneous variety induced by $(V_F)^\vee$. (So $(V_F)^\vee = (V_F)^\vee$.) Let $U$ be the variety defined in Theorem 4.4. Suppose that $x \not\in U \cup V \cup V^\vee$. Let $y_i(x) \in V, i = 1, \ldots, \delta(V)$ be defined as in Theorem 4.4. Since $\mathbb{C}^n \times \{0\}$ is an irreducible component of $\Sigma_0(V)$ of dimension $n$, (see Lemma 4.3), we may assume that $y_i(x) \neq 0$ for $i = 1, \ldots, \delta(V)$. By definition $x - y_i(x) \in \text{range } D(V)(y_i(x))$. Hence $\phi - \psi$ vanish at $y_i(x)$. Therefore $F_x$ vanishes at each $y_i(x)$. Vice versa, suppose that $F_x$ vanishes at $y \in V'$. Since $\psi$ vanishes nowhere it follows that $\phi - a\psi$ vanishes at $y$. Suppose that $a = 0$. Then $x \in \text{range } D(V)(y)$, i.e., $[x] \in (V_F)^\vee$, contrary to our assumption. So $a \neq 0$. Hence $a y \in \{y_1(x), \ldots, y_{\delta(V)}(x)\}$. So $F_x$ vanishes exactly at $[y_i(x)], i = 1, \ldots, \delta(V)$. \hfill \Box

6 The distance function from an invariant subspace

**Theorem 6.1** Let $C \subset \mathbb{R}^n$ be an irreducible variety. Let $\mathcal{G} \subset \mathbb{R}^{n \times n}$ be a finite nontrivial group of orthogonal matrices, $|\mathcal{G}| > 1$, which acts as a finite group of linear transformations on $\mathbb{C}^n$, i.e., $x \mapsto Ax$ for $A \in \mathcal{G}$. Assume that the following conditions are satisfied:

1. $U \subset \mathbb{C}^n$ and $U_R \subset \mathbb{R}^n$ are the subspaces of all complex respectively real fixed points of $\mathcal{G}$ of complex respectively real dimension $\ell \in \{1, \ldots, n - 1\}$;

2. $C$ is fixed by $\mathcal{G}$;

3. $U_R \cap C$ is a strict subset of $U_R$ and $C$.

Denote $V = C_C, C_U := V \cap U$. Then $V$ is fixed by $\mathcal{G}$ and $U \cap V$ are strict subsets of $U$ and $V$.

Let $V_s \subset \mathbb{C}^n$ be the quasi-algebraic set given by (4.4). Let $y \in C_U \setminus (V_s \cup \text{Sing } V)$. Then there exists a neighborhood $O(y) \subset \mathbb{C}^n$ of $y$ such that for each $x \in O(y) \cap U$, the point $y(x)$ given by Theorem 4.4 is in $C_U$.

Denote $\Phi := (U \times V) \cap \Sigma(V)$. Then $\Phi$ contains a finite number of irreducible components $\Phi_1, \ldots, \Phi_k$, each of dimension $\ell$, such that the following conditions hold. Let $y \in C_U \setminus (V_s \cup \text{Sing } V)$. Let $O(y) \subset \mathbb{C}^n$ be a neighborhood of $y$ defined as above. Then for each $x \in O(y) \cap U$ and $y(x) \in O(x) \cap C_U$ the point $(x, y(x))$ is in some $\Phi_i$. Furthermore, for each $i \in \{1, \ldots, k\}$ there exists $y \in C_U \setminus (V_s \cup \text{Sing } V)$ such that for each $x \in O(y) \cap U$ and $y(x) \in O(x) \cap C_U$ the point $(x, y(x))$ is in $\Phi_i$. Hence each $\Phi_i$ lies in $U \times C_U$.

**Proof.** Since $C$ is invariant under the action of $\mathcal{G}$ it follows that $V$ is invariant under the action of $\mathcal{G}$. As $U_R \cap C$ is a strict subset of $U_R$ and $C$, it follows that $U \cap V$ is a strict subset of $U$ and $V$. Assume that the group $\mathcal{G}$ acts on $\mathbb{C}^n \times \mathbb{C}^n$ as follows: $(x, y) \mapsto (Ax, Ay)$. Since $\mathcal{G}$ is a group of orthogonal matrices, it follows that $\Sigma_0(V) \cap (\mathbb{R}^n \times C)$ is invariant under the action of $\mathcal{G}$. Hence $\Sigma(V)$ is invariant under the action of $\mathcal{G}$. So $(x, z) \in \Sigma(V) \iff (Ax, Az) \in \Sigma(V)$ for each $A \in \mathcal{G}$. Assume that $y \in C_U \setminus (V_s \cup \text{Sing } V)$. Without loss of generality we can assume that the neighborhood $O(y) \subset \mathbb{C}^n$ given in Theorem 4.7 is invariant under the
action of \( \mathcal{G} \). Suppose that \( x \in O(y) \cap U \). Then \( (x, y(x)) \in (O(y) \times O(y)) \cap \Sigma(V) \). Clearly \( A(x, Ay(x)) = (x, Ay(x)) \in (O(y) \times O(y)) \cap \Sigma(V) \). Theorem 4.7 yields that \( Ay(x) = y(x) \). Hence \( y(x) \in C_U \).

For \( x \in O(y) \cap U \) the point \( (x, y(x)) \) is an isolated point of \( \Phi \). Hence the set of all points \( (x, y(x)) \), where \( x \in O(y) \cap U \), lies in \( U \times C_U \). Furthermore, these points belong to an irreducible component of \( \Phi \), which is denoted by \( \Phi_i \). The dimension of this variety is \( \ell \). Since this variety intersects \( U \times C_U \) in an open set, it follows that \( \Phi_i \subset U \times C_U \). Let \( \Phi_1, \ldots, \Phi_k \) be all the irreducible components of \( \Phi \) obtained in this way. \( \square \)

Assume that the assumptions of Theorem 6.1 hold. An interesting and nontrivial problem is to give conditions such that for each \( x \in U_R \) there exists a best \( C \)-approximation in \( C \cap U_R \). We now give sufficient conditions so that this property holds. Recall that a set \( Q \in R^n \) or \( C^n \) is called constructible if it is a finite union of quasi-algebraic sets. Chevalley’s theorem states that the image of a constructible set by a polynomial map is a constructible set.

**Theorem 6.2** Let the assumptions of Theorem 6.1 hold. Let \( \Phi_1, \ldots, \Phi_k \) be the irreducible components of \( \Phi \) defined in Theorem 6.1. Let \( \Phi = \Phi' \cup (\bigcup_{i=1}^k \Phi_i) \) be the decomposition of \( \Phi \) into its irreducible components. Assume the following conditions.

First, the equality

\[
\text{dist}(x, C) = \text{dist}(x, \text{Sing } C) \text{ for } x \in U_R \tag{6.1}
\]

defines a semi-algebraic \( Q \) of dimension less than \( \ell \) in \( U_R \). (\( Q = \emptyset \) if \( \text{Sing } C = \emptyset \).) Second, the constructible set \( \pi_n(\Phi' \cap (R^n \times R^n)) \) has dimension less than \( \ell \), i.e., the closure of \( \pi_n(\Phi' \cap (R^n \times R^n)) \) in \( U_R \) is a strict subvariety of \( U_R \). Then for each \( x \in U_R \) there exists a \( C \)-best approximation which is in \( C \cap U_R \).

**Proof.** Clearly \( Q \) is a closed semi-algebraic set. Since \( U_R \setminus Q \) is an open set which is dense in \( U_R \) it follows that for each \( x \in U_R \)

\[
\text{dist}(x, C) = \min\{\|x - y\|, \ x \in U_R, (x, y) \in \Phi \cap (R^n \times R^n)\}. \tag{6.2}
\]

The assumption on \( \Phi' \) means that

\[
\pi_n((\Phi' \setminus \Phi') \cap (R^n \times R^n)) = \pi_n((\bigcup_{i=1}^k (\Phi_i' \setminus \Phi_i') \cap (R^n \times R^n)) \subset U_R \times (C \cap U_R)
\]

is a constructible set in \( U_R \) which is dense in \( U_R \). Observe next that for each \( x \in \pi_n((\Phi' \setminus \Phi') \cap (R^n \times R^n)) \setminus Q \) each \( C \)-best approximation is in \( C \cap U_R \). As \( \pi_n((\Phi' \setminus \Phi') \cap (R^n \times R^n)) \setminus Q \) is dense in \( U_R \) it follows that for each \( x \in U_R \) there exists a \( C \)-best approximation which is in \( C \cap U_R \). \( \square \)

Two cases where the conclusion of Theorem 6.2 hold are discussed in §8. We close this section with the following simpler result:

**Proposition 6.3** Let the assumptions of Theorem 6.1 hold. Then there exists a semi-algebraic open set \( O \subset R^n \) containing all smooth points of \( C \) lying in \( C \cap U_R \) such that for each \( x \in O \) a best \( C \) approximation of \( x \) is unique. Furthermore, for each \( x \in O \cap U_R \) a best \( C \) approximation lies in \( C \cap U_R \).
Theorems 4.7 and 6.1 yield that \( O \) of a semi-algebraic set is semi-algebraic. So \( \text{Closure}(x) \) is semi-algebraic. Let \( V \) be defined as in Theorem 4.4. Theorem 4.4 claims that \( V \cap (C \setminus \text{Sing}) \neq \emptyset \). The arguments of the proof of Theorem 4.4 yield that \( O_2 \) is semi-algebraic.

Since \( U_R \cap C \) is a strict subset \( U_R \) and \( C \) it follows that \( U_R \cap (C \setminus \text{Sing}) \neq \emptyset \). Let \( V_x \) be defined as in Theorem 4.4. Theorem 4.4 claims that \( V_x \cap (C \setminus \text{Sing}) = \emptyset \). Theorems 4.7 and 6.1 yield that \( O_2 \) contains a neighborhood of each \( y \in U_R \cap (C \setminus \text{Sing}) \). Let \( O_3 := \mathbb{R}^n \setminus O_2 \). So \( O_3 \) is semi-algebraic. Recall that a closure of a semi-algebraic set is semi-algebraic. So \( \text{Closure}(O_3) \) is semi-algebraic. Hence \( O := \mathbb{R}^n \setminus \text{Closure}(O_3) \), the interior of \( O_2 \), is a set satisfying the conditions of the proposition.

\[ \square \]

7 On the rank of generic tensors in \( \mathbb{F}^{m \times d} \)

Let \( \mathbb{F}^{m \times d} \supset S(m, d, \mathbb{F}) \) be the space of \( d \)-mode, \( (d \geq 2) \), tensors whose each mode has \( m \) coordinates, and the subspace of \( d \)-mode symmetric tensors over a field \( \mathbb{F} \).

For \( x_1, \ldots, x_d \in \mathbb{F}^m \) denote by \( \otimes^d_{i=1} x_i \) the tensor product \( x_1 \otimes \ldots \otimes x_d \in \mathbb{F}^{m \times d} \). If \( x_i = u \) for \( i = 1, \ldots, d \) then \( \otimes^d u := \otimes^d_{i=1} x_i \). Let \( T \in \mathbb{F}^{m \times d} \setminus \{0\} \) and \( S \subset S(m, d, \mathbb{F}) \setminus \{0\} \). Consider the following decompositions of \( T \) and \( S \) into rank one tensors:

\[
T = \sum_{j=1}^r \otimes^d_{i=1} x_{i,j}, \quad x_{i,j} \in \mathbb{F}^m \setminus \{0\}, \quad i = 1, \ldots, d, \quad j = 1, \ldots, r \tag{7.1}
\]

\[
S = \sum_{j=1}^s t_j \otimes^d u_j, \quad t_j \in \mathbb{F} \setminus \{0\}, \quad u_j \in \mathbb{F}^m \setminus \{0\}, \quad j = 1, \ldots, s \tag{7.2}
\]

The minimal \( r \) and \( s \) for which the above equalities holds for \( T \) are called the rank and symmetric rank of \( T \) and \( S \) respectively, which are denoted by rank \( T \) and \( \text{rank} S \). (The rank and the symmetric rank of zero tensor is zero.) Note that if \( \mathbb{F} \) is algebraically closed that in (7.2) we can assume that each \( t_j = 1 \). For \( \mathbb{F} = \mathbb{R} \) we can assume that each \( t_j = \pm 1 \).

Let \( e_i := (\delta_{i1}, \ldots, \delta_{im})^\top \), \( i = 1, \ldots, m \), be the standard basis in \( \mathbb{F}^m \). Then \( \otimes^d_{j=1} e_{i,j}, \quad i_1, \ldots, i_d = 1, \ldots, m \) is a standard basis in \( \mathbb{F}^{m \times d} \). Hence any \( T \in \mathbb{F}^{m \times d} \) has a decomposition (7.1). So rank \( T \) is well defined.

However the symmetric rank of a symmetric tensor \( S \in S(m, d, \mathbb{F}) \) may be not defined for a field of a finite characteristic \( p \geq 2 \). We now show that such \( S \) exist for any finite field \( \mathbb{F}_p \) with \( p^l \) elements and a corresponding \( d \).

Denote by \( S_d \) the group of permutations on \( \{1, \ldots, d\} \). Let \( 1 \leq i_1 \leq \ldots \leq i_d \leq m \). Denote by \( \text{orb}(i_1, \ldots, i_d) \) the orbit of the multiset \( \{i_1, \ldots, i_d\} \subset \{1, \ldots, m\}^d \) under the action of \( S_d \), i.e., this orbit is a union of all ordered distinct multisets \( \{\sigma(i_1), \ldots, \sigma(i_d)\}, \sigma \in S_d \). Note that the number of such orbits is \( \binom{m+d-1}{d} \). To each such orbit we associate a symmetric tensor \( S(i_1, \ldots, i_d) := \sum_{(j_1, \ldots, j_d) \in \text{orb}(i_1, \ldots, i_d)} \otimes_{k=1}^d e_{i_k} \). Clearly the set of all these symmetric tensors form a basis in \( S(m, d, \mathbb{F}) \). Hence
dim $S(m, d, \mathbb{F}) = \binom{m+d-1}{d}$. (The following two results are probably known, and we give their proof for completeness.)

**Proposition 7.1** Let $p \geq 2$ be a prime, $\mathbb{F}_{p^l}$ be a field with $p^l$ elements and $m \geq 2$ be an integer. Then for each integer $d$ satisfying

$$\binom{m+d-1}{d} > \frac{p^{ml}-1}{p^l-1}$$

(7.3)

there exists $S \in S(m, d, \mathbb{F}_{p^l})$ such that $S$ is not a linear combination of rank one symmetric tensors.

**Proof.** The number of nonzero elements in $\mathbb{F}^m_{p^l}$ is $p^{ml} - 1$. Each nonzero vector $x \in \mathbb{F}^m_{p^l}$ generates a line with $p^l - 1$ nonzero elements of the form $tx, t \in \mathbb{F}_{p^l} \setminus \{0\}$. Hence the subspace $U \subset S(m, d, \mathbb{F}_{p^l})$ generated by all vectors of the form $\otimes^d u, u \in \mathbb{F}^m_{p^l}$ is of dimension $\frac{p^{ml}-1}{p^l-1}$ at most. Assume that (7.3) holds. So dim $U < \dim S(m, d, \mathbb{F}_{p^l})$. Hence there exists $S \in S(m, d, \mathbb{F}_{p^l}) \setminus U$. □

**Proposition 7.2** Let $\mathbb{F}$ be a field with at least $d$ elements. Then for each $S \in S(m, d, \mathbb{F})$ (7.2) holds.

**Proof.** For $u, v \in \mathbb{F}^m$ and a variable $t \in \mathbb{F}$ let

$$\otimes^d (tu + v) = \sum_{k=0}^{d} t^k S_{k,d-k}(u, v).$$

So $S_{k,d-k}(u, v)$ is the symmetric tensor induced by $(\otimes^k u) \otimes (\otimes^{d-k} v)$. Clearly $S_{0,d}(u, v) = \otimes^d v, S_{d,0}(u, v) = \otimes^d u$. We claim that $S_{k,d-k}(u, v)$ has a decomposition (7.2) for $k = 1, \ldots, d - 1$. Consider the polynomial

$$\otimes^d (tu + v) - t^d \otimes^d u = \sum_{k=0}^{d-1} t^k S_{k,d-k}(u, v).$$

Recall that the Vandermonde matrix $[\frac{j}{i}]_{i,j=0}^{d-1}$ is an invertible matrix for a subset $\{\tau_0, \ldots, \tau_{d-1}\}$ of $\mathbb{F}$ of cardinality $d$. Hence each $S_{k,d-k}(u, v), k = 0, \ldots, d-1$ can be expressed as a linear combination of $\otimes^d (\tau_i u + v), i = 0, \ldots, d-1$ and $\otimes^d u$.

More generally, consider the following polynomial

$$\mathcal{X}(t_1, \ldots, t_{d-1}) := \otimes^d (e_m + \sum_{j=1}^{m-1} t_j e_i) - \sum_{j=1}^{m-1} t^d_j \otimes^d e_j$$

in $d - 1$ variables $t_1, \ldots, t_{m-1}$. The coefficients the monomial $t_1^{i_1} \cdots t_{d-1}^{i_{d-1}}$ is the tensor $S(i_1, \ldots, i_d)$ in the standard basis of $S(m, d, \mathbb{F})$ described above. Vice versa, each vector in the standard basis of $S(m, d, \mathbb{F})$, except $\otimes^d e_j, j = 1, \ldots, m - 1$, is a coefficient of the corresponding monomial $t_1^{j_1} \cdots t_{d-1}^{j_{d-1}}$. View $\mathcal{X}(t_1, \ldots, t_{m-1})$ as a polynomial in $t_{m-1}$:

$$\mathcal{X}(t_1, \ldots, t_{m-1}) = \sum_{k=0}^{d-1} t_{m-1}^{k} \mathcal{X}_k(t_1, \ldots, t_{m-2}).$$
Find the polynomials $X_k(t_1, \ldots, t_{m-2})$ for $k = 0, \ldots, d-1$ using the above procedure. Continue this procedure to obtain all vectors in the standard basis of $S(m, d, F)$, except $\otimes^d e_j$, $j = 1, \ldots, m-1$ as linear combinations of vectors $X(t_1, \ldots, t_{m-1})$, where $t_1, \ldots, t_{m-1} \in \{\tau_0, \ldots, \tau_{d-1}\}$. Then (7.2) holds. \end{proof}

For the case where $F$ is an infinite field see [1].

**Corollary 7.3** Let $m, d \geq 2$ be integers. Let $F$ be a field with at least $d$ elements. Then there exists $k = (m + d - 1) d$ vectors $u_1, \ldots, u_k \in F^m$ such that $\otimes^d u_1, \ldots, \otimes^d u_k$ form a basis in $S(m, d, F)$.

Assume that $F$ is an infinite field. Clearly

$$\text{rank } S \leq \text{rank } S \text{ for } S \in S(m, d, F). \quad (7.4)$$

**Example 7.4** Let $F$ be any field of characteristic 2. Let $S = e_1 \otimes e_2 + e_2 \otimes e_1 \in S(2, 2, F)$. Then

$$\text{rank } S = 2, \quad \text{rank } S = 3. \quad (7.5)$$

**Proof.** Clearly $\text{rank } S = 2$. The equality $S = e_1 \otimes e_1 + e_2 \otimes e_2 + (e_1 + e_2) \otimes (e_1 + e_2)$ yields that $\text{rank } S \leq 3$. Assume now that $\text{rank } S = 2$. So $S = au \otimes u + bv \otimes v$. Since $\text{rank } S = 2$ we deduce that $u, v \in F^2$ are linearly independent and $a, b \neq 0$. Let $u = (u_1, u_2)\top, v = (v_1, v_2)\top$. Clearly

$$au_1^2 + bv_1^2 = au_2^2 + bv_2^2 = 0.$$ 

If $u_1 = 0$ then $v_1 = 0$ which contradicts the assumption that $u, v$ are linearly independent. Hence $u_1, u_2, v_1, v_2 \neq 0$. Since the characteristic of $F$ is 2 it follows

$$\frac{a}{b} = \frac{v_1^2}{u_1^2} = \frac{v_2^2}{u_2^2} \Rightarrow 0 = v_1^2 u_2^2 + v_2^2 u_1^2 = (v_1 u_2 + v_2 u_1)^2$$

So $u, v$ are linearly dependent contrary to our assumption. \end{proof}

It is an open problem if equality holds in (7.4) for each symmetric tensor $S$ for $F = \mathbb{C}$, or more generally over any algebraically closed (or an infinite) field $F$ of characteristic different from 2. (This is true for $d = 2$.) [5, Proposition 5.5] shows that equality holds in (7.4) for $F = \mathbb{C}$ and $\text{rank } T \leq 2$.

**Definition 7.5** Let $F$ be a field, $m \geq 2, d \geq 3$ be integers. Assume that $T \in F^{m \times d} \setminus \{0\}, S \in S(m, d, F) \setminus \{0\}$ have decompositions (7.1) and (7.2) respectively, which can be rewritten as follows:

$$T = \sum_{j=1}^r (\otimes^a x_{i,j}) \otimes (\otimes^{a+b} x_{i,j}) \otimes (\otimes^{a+b+c} x_{i,j}), \quad (7.6)$$

$$S = \sum_{j=1}^s t_j (\otimes^a u_j) \otimes (\otimes^b u_j) \otimes (\otimes^c u_j). \quad (7.7)$$

Here $a, b, c$ are positive integers such that $a + b + c = d$. Then $T$ and $S$ of are called $(a, b, c)$-generic if the following conditions hold for $T$ and $S$ respectively.
1. For $\mathcal{T}$: any $\min(m^a, r)$ vectors out of $\otimes_{i=1}^{d} x_{i,1}, \ldots, \otimes_{i=1}^{d} x_{i,r}$ are linearly independent; any $\min(m^b, r)$ vectors out of $\otimes_{i=a+1}^{a+b} x_{i,1}, \ldots, \otimes_{i=a+1}^{a+b} x_{i,r}$ are linearly independent; any $\min(m^c, r)$ vectors out of $\otimes_{i=a+b+1}^{d} x_{i,1}, \ldots, \otimes_{i=a+b+1}^{d} x_{i,r}$ are linearly independent.

2. For $\mathcal{S}$: if $\mathbb{F}$ is algebraically closed then each $t_j = 1$ and for each $t_j = \pm 1$; any $\min((m+a-1), s)$ vectors out of $\{\otimes^a u_1, \ldots, \otimes^a u_s\}$ are linearly independent; any $\min((m+b-1), s)$ vectors out of $\{\otimes^b u_1, \ldots, \otimes^b u_s\}$ are linearly independent; any $\min((m+c-1), s)$ vectors out of $\{\otimes^c u_1, \ldots, \otimes^c u_s\}$ are linearly independent.

Assume that $\mathbb{F} = \mathbb{C}$, $d = 3$ and $r \leq m$. Then the Definition 7.35 of a generic symmetric tensor coincides with the definition in [5].

**Theorem 7.6** Let $\mathbb{F}$ be a field, $m \geq 2, d \geq 3$ be integers. Assume that the decompositions (7.6-7.7) of $\mathcal{T} \in \mathbb{F}^{m \times d}$ and $\mathcal{S} \in \mathbb{S}(m, d, \mathbb{F})$ are $(1, \lceil \frac{d-1}{2} \rceil, \lceil \frac{d-1}{2} \rceil)$-generic.

1. Let $d = 2b + 1 \geq 3$. Then

$$\text{rank } \mathcal{T} = r \text{ if } r \leq m^b + \frac{m-2}{2}, \quad (7.8)$$

$$\text{rank } \mathcal{S} = \text{rank } \mathcal{S} = s \text{ if } r \leq m^b - 1 + \frac{m-2}{2}, \quad (7.9)$$

2. Let $d = 2b + 2 \geq 4$. Then

$$\text{rank } \mathcal{T} = r \text{ if } r \leq m^b + m - 2, \quad (7.10)$$

$$\text{rank } \mathcal{S} = \text{rank } \mathcal{S} = s \text{ if } r \leq m^b - 1 + m - 2. \quad (7.11)$$

In all the above cases representation of $\mathcal{T}$ and $\mathcal{S}$ as a sum of rank one tensors is unique up to a permutation of the summands.

**Proof.** The proof of the theorem uses Kruskal’s theorem for 3-mode tensors $\mathbb{F}^{n \times p \times q} := \mathbb{F}^n \otimes \mathbb{F}^p \otimes \mathbb{F}^q$ [19]. (See [21] for a short proof of Kruskal’s theorem.) Let $x_1, \ldots, x_l \in \mathbb{F}^n$ be nonzero vectors. Form the matrix $X = [x_1 \ldots x_l] \in \mathbb{F}^{n \times l}$, (whose columns are $x_1, \ldots, x_l$). Then Kruskal rank of $X$, denoted as $\text{rank } X$, is the maximal integer $k \leq l$ such that any $k$ vectors out of $\{x_1, \ldots, x_l\}$ are linearly independent. Let $\mathcal{X} \in \mathbb{F}^{n \times p \times q}$ and assume that

$$\mathcal{X} = \sum_{i=1}^{r} y_i \otimes z_i \otimes w_i, \quad y_i \in \mathbb{F}^n, z_i \in \mathbb{F}^p, w_i \in \mathbb{F}^q, i = 1, \ldots, r. \quad (7.12)$$

Form the matrices $Y = [y_1 \ldots y_r], Z = [z_1 \ldots z_r], W = [w_1 \ldots w_r]$. Let $\kappa_1 = \text{rank } Y, \kappa_2 = \text{rank } Z, \kappa_3 = \text{rank } W$. We call $(\kappa_1, \kappa_2, \kappa_3)$ the Kruskal ranks of the decomposition (7.12). Kruskal’s theorem claims that if

$$\kappa_1 + \kappa_2 + \kappa_3 \geq 2r + 2 \quad (7.13)$$

17
then rank \( X = r \). Furthermore, the decomposition of \( X \) to a sum of \( r \) rank one tensors is unique up to a permutation of the summands.

We now prove our theorem. Assume that the decompositions of tensors \( T \) and \( S \) given by (7.6, 7.7) are \((a, b, c)\) generic. (For \( S \) we use the identity \( t_j \otimes u_j \otimes (\otimes u_j) = (t_j \otimes u_j) \otimes (\otimes u_j) \) for \( j = 1, \ldots, r \).) Then the Kruskal ranks of these decompositions of \( T \) and \( S \), denoted by \((\alpha_1, \alpha_2, \alpha_3)\) and \((\beta_1, \beta_2, \beta_3)\) respectively, are:

\[
\alpha_1 = \min(m^a, r), \alpha_2 = \min(m^b, r), \alpha_3 = \min(m^c, r),
\]

\[
\beta_1 = \min\left(\frac{m + a - 1}{a}\right), \beta_2 = \min\left(\frac{m + b - 1}{b}\right), \beta_3 = \min\left(\frac{m + c - 1}{c}\right).
\]

For \( r = 1 \) the theorem is trivial. In what follows we assume that \( r \geq 2 \). Suppose first that \( d = 2b + 1 \). Assume first that \( S \) is \((1, b, b)\) generic. Suppose first that \( r \leq m \). Then \( \alpha_1 = \alpha_2 = \alpha_3 = r \). Again, Kruskal’s inequality holds. We deduce the theorem in this case for \( S \). Assume that \( m \geq 4 \) and \( r > \left(\frac{m + b - 1}{b}\right) \). Then \( \alpha_1 = m, \alpha_2 = \alpha_3 = \left(\frac{m + b - 1}{b}\right) \). (Recall that \( \dim S(m, d, F) = \left(\frac{m + b - 1}{b}\right) \).) Then Kruskal inequality holds if and only if \( r \leq \left(\frac{m + b - 1}{b}\right) + \frac{m - 2}{2} \). Hence the theorem holds in this case for \( S \) too.

Assume now that \( d = 2b + 2 \). The above arguments apply for \( S \) and \( r \leq \left(\frac{m + b - 1}{b}\right) \). Assume that \( m > 2 \) and \( \left(\frac{m + b - 1}{b}\right) < r \leq \left(\frac{m + b}{b + 1}\right) \). Then \( \alpha_1 = m, \alpha_2 = \left(\frac{m + b - 1}{b}\right), \alpha_3 = r \). Kruskal’s inequality yields that \( r \leq \left(\frac{m + b - 1}{b}\right) + m - 2 \). Hence the theorem holds in this case too for \( S \). Similar arguments yield the theorem for \( T \).

The upper bound on \( r \) in Theorem 7.6 for \( S \) can be replaced by

\[
N(m, d) = \begin{cases} 
(m^d - 1) + \frac{m - 2}{2} & \text{for an odd } d \geq 3 \\
(m^d - 1) + m - 2 & \text{for an even } d \geq 4
\end{cases}
\]

(7.14)

We used the identity \( \binom{n}{k} = \binom{n}{n-k} \).

8 Approximation of real symmetric tensors

Let \( R(k, m^{x_d}) \subset \mathbb{R}^{m^{x_d}}, R_C(k, m^{x_d}) \subset \mathbb{C}^{m^{x_d}} \) be the closure of all tensors of rank at most \( k \), i.e., all tensors of border rank at most \( k \) in \( \mathbb{R}^{m^{x_d}}, \mathbb{C}^{m^{x_d}} \) respectively. In this section we consider the best approximation problem in \( \mathbb{R}^{m^{x_d}} \) for \( C := R(k, m^{x_d}) \). Clearly, \( C_C = R_C(k, m^{x_d}) \). Observe next that the symmetric group \( S_d \) of order \( d \) acts on \( \mathbb{C}^{m^{x_d}} \) as follows: For \( \sigma \in S_d \) and \( T = [t_{i_1,\ldots,i_d}] \in \mathbb{C}^{m^{x_d}} \) we define \( \sigma(T) = [t_{i_{\sigma(1)},\ldots,i_{\sigma(d)}}] \). Clearly, the action of each \( \sigma \) preserves the Hilbert-Schmidt norm on \( \mathbb{R}^{m^{x_d}} \). It is straightforward to see that \( C \) and \( C_C \) are invariant under the action of \( S_d \). Furthermore \( S(m, d, F) \) is the set of the fixed points in \( \mathbb{F}^{m^{x_d}} \) for \( \mathbb{F} = \mathbb{R}, \mathbb{C} \) respectively.
A natural question is if for each $S \in S(m,d,\mathbb{R})$ there exists a best $k$-border rank approximation which is symmetric. This is a special case of the approximation problem discussed in [3].

For $d = 2$, i.e., the space of real symmetric matrices, the answer to this question is positive [15]. Let $C_k$ be the set of real matrices of rank at most $k$. Then the set of singular points of $C_k$ is $\text{Sing} C_k = C_{k-1}$. A best $k$-rank approximation of $B \in S(m,2,\mathbb{R})$ is $A \in C_k$ which has the same $k$ maximal singular values and corresponding left and right singular vectors as $B$. Hence the set $Q$ defined in Theorem 6.2 is $C_{k-1}$. Assume now that the $m$ eigenvalues of $B$, $\lambda_1, \ldots, \lambda_m$ satisfy the condition $|\lambda_i| \neq |\lambda_j|$ for $i \neq j$. (I.e., all singular values of $B$ are distinct.) So $B = \sum_{i=1}^{m} \lambda_i u_i u_i^\top$, where $u_i^\top u_j = \delta_{ij}, i,j = 1,\ldots,m$. Then $A$ is a critical point of the function $\text{tr}(B - X)^\top(B - X)$, $X \in C_k$ if and only if $A = \sum_{i \in \Omega} \lambda_i u_i u_i^\top$ for any subset $\Omega$ of $\{1, \ldots, m\}$ of cardinality $k$. So each real critical point of $B$ is symmetric. Hence the assumptions of Theorem 6.2 hold in this case.

For $k = 1$ and any $d \geq 3$ the answer to this question is also positive. I.e., every symmetric tensor $S \in S(d,m,\mathbb{R})$ has a symmetric best rank one approximation. This result is implied by Banach’s theorem [2]. The Banach theorem was re-proved in [4, 11]. The results in [13] yield that the set of symmetric tensors which do not have a unique best rank one approximation has zero Lebesgue measure. We now give an improved version of this result.

**Theorem 8.1** Let $m \geq 2, d \geq 3$ be integers. Then there exists a semi-algebraic set $W \subset S(m,d,\mathbb{R})$ which does not contain an open set with the following properties. $S \in S(m,d,\mathbb{R})$ does not have a unique best rank one approximation if and only if $S \in W$. Furthermore, for each $S \in S(m,d,\mathbb{R}) \setminus W$ the unique best rank one approximation is symmetric. In particular, there exists a hypersurface $H \subset S(m,d,\mathbb{R})$ such that for each $S \in S(m,d,\mathbb{R}) \setminus H$ the unique best rank one approximation is symmetric.

**Proof.** Let $W \subset S(m,d,\mathbb{R})$ be the set of all $S \in S(m,d,\mathbb{R})$ that do not have a unique best rank one approximation. The arguments of the proof of Theorem 3.4 imply that $W$ is semi-algebraic. Let $\mathcal{S} \subset S(m,d,\mathbb{R}) \setminus W$. Then $\mathcal{S}$ has a unique best rank one approximation. [11] claims that $\mathcal{S}$ has a best rank one approximation which is symmetric. Hence $\mathcal{S}$ has a unique best rank one approximation which is symmetric.

We now show that $W$ does not contain an open set. Recall that the function $\text{dist}(\cdot, \mathcal{R}(1,m^{\times d})) : \mathbb{R}^{m^{\times d}} \to \mathbb{R}$ is semi-algebraic. Let $d(\cdot) : S(m,d,\mathbb{R}) \to \mathbb{R}$ be the restriction of $\text{dist}(\cdot, \mathcal{R}(1,m^{\times d}))$ to $S(m,d,\mathbb{R})$. Clearly, $d(\cdot)$ is semi-algebraic. Proposition 3.3 yields that $d(\cdot)$ is not differentiable on a semi-algebraic set $W' \subset S(m,d,\mathbb{R})$ of dimension less than $\text{dim} S(m,d,\mathbb{R})$.

Let $\mathcal{S} \subset S(m,d,\mathbb{R}) \setminus W'$. Hence function $d(\cdot)$ is differentiable at $\mathcal{S}$. The arguments of [13, §7] yield that the set of all best rank one approximation of $\mathcal{S}$ is the orbit of one best rank one approximation $\otimes_{i \in [d]} x_i$ under the action of the symmetric group $S_d$ as defined above. That is, all best rank one approximations are of the form $\otimes_{i \in [d]} x_{\sigma(i)}$ for $\sigma \in S_d$. [11] claims that $\mathcal{S}$ has a best rank one approximation which is symmetric. Hence the orbit of $\otimes_{i \in [d]} x_i$ consists of one symmetric tensor. In particular, $\mathcal{S}$ has a unique best rank one approximation which is symmetric. Therefore $W' \supset W$. Hence $W$ does not contain an open set. The arguments of the proof of Theorem 3.4 imply that $W$ is contained in a hypersurface $H$. □
For rank one approximation of tensors in $S(m, d, \mathbb{R})$ the assumptions of Theorem 6.2 are equivalent to the following statement. There exists an open quasi-algebraic set $W \subset S(m, d, \mathbb{R})$ such that for each $B \in W$ every real rank one tensor $A$ which is a critical point of $\|B - X\|^2$, $X \in \mathcal{R}(1, m \times d)$ is symmetric.

For $k \geq 2$ and $d \geq 3$ we have the following weaker result.

**Theorem 8.2** Let $m \geq 2, d \geq 3$ be integers. Assume that $2 \leq k \leq N(m, d)$. Then there exists an open semi-algebraic set $O \subset S(m, d)$ containing all $(1, \lfloor d - 1 \rfloor, \lceil d - 1 \rceil)$-generic real symmetric tensors of rank $k$ such that for each $S \in O$ a best $k$-border rank approximation is symmetric and unique.

**Proof.** Let $S$ be an $(a, b, c)$ symmetric generic tensor given by (7.7), where $a = 1, b = \lfloor \frac{d-1}{2} \rfloor, c = \lceil \frac{d-1}{2} \rceil, r = k \in [2, N(m, d)]$ and each $t_j = \pm 1$. We claim that $S$ is a smooth point of $\mathcal{R}(k, m \times d)$. Recall the proof of Theorem (7.6). Kruskal’s theorem yields that any $T \in \mathcal{R}(k, m \times d)$ in a suitably small neighborhood of $S$ must be of the form (7.6), such that each $(\otimes_{i=1}^{a+b} x_{ij}) \otimes (\otimes_{i=a+1}^{a+b+c} x_{ij})$ is in the neighborhood of $t_j(\otimes_{i=a}^{a+b} u_{ij}) \otimes (\otimes_{i=a+b+1}^{a+b+c} u_{ij})$. Hence $T$ is $(a, b, c)$-generic and its rank is $k$. Therefore $S$ is a smooth point of $\mathcal{R}(k, m \times d)$. Theorem 4.7 and Proposition 6.3 imply our theorem. \(\square\)

**References**

[1] J. Alexander and A. Hirschowitz, Polynomial interpolation in several variables, *J. Algebraic Geom.*, 4 (1995), 201–222.

[2] S. Banach, Über homogene Polynome in $(L^2)$, *Studia Math.* 7 (1938), 36–44.

[3] F. Catanese and C. Trifogli, Focal loci of algebraic varieties I, *Communications in Algebra* 28 (2000), 6017–6057.

[4] B. Chen, S. He, Z. Li, and S. Zhang, Maximum block improvement and polynomial optimization, *SIAM J. Optimization*, 22 (2012), 87–107.

[5] P. Comon, G. Golub, L.-H. Lim, and B. Mourrain, Symmetric tensors and symmetric tensor rank, *SIAM Journal on Matrix Analysis and Applications* 30 (2008), 1254–1279.

[6] M. Coste, *Real Algebraic Sets*, Lecture notes, 2005, [http://perso.univ-rennes1.fr/michel.coste/polyens/RASroot.pdf](http://perso.univ-rennes1.fr/michel.coste/polyens/RASroot.pdf).

[7] V. De Silva and L.-H. Lim, Tensor rank and the ill-posedness of the best low-rank approximation problem, *SIAM Journal on Matrix Analysis and Applications*, 30 (2008), 1084–1127.

[8] J. Draisma, E. Horobet, G. Ottaviani, B. Sturmfels and R.R. Thomas, The Euclidean distance degree of an algebraic variety, [arXiv:1309.0049](https://arxiv.org/abs/1309.0049).

[9] A. H. Durfee, Neighborhoods of algebraic sets, *Transactions of the American Mathematical Society*, 276 (1983), 517–530.
[10] S. Friedland, On the generic rank of 3-tensors, *Linear Algebra and its Applications*, 436 (2012) 478–497.

[11] S. Friedland, Best rank one approximation of real symmetric tensors can be chosen symmetric, *Front. Math. China*, 8 (2013), 19–40.

[12] S. Friedland, V. Mehrmann, R. Pajarola, S.K. Suter, On best rank one approximation of tensors, *Numerical Linear Algebra with Applications*, 2013, to appear.

[13] S. Friedland and G. Ottaviani, The number of singular vector tuples and uniqueness of best rank one approximation of tensors, [arXiv:1210.8316](http://arxiv.org/abs/1210.8316).

[14] W. Fulton, *Intersection Theory*, Springer, Berlin 1984.

[15] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 4th Edition, The Johns Hopkins University Press, 2013.

[16] R. Hartshorne, *Algebraic Geometry*, Graduate Texts in Mathematics 52, Springer, 1977, New York.

[17] A. Holme: The geometric and numerical properties of duality in projective algebraic geometry, *Manuscripta Math.* 61 (1988), 145–162.

[18] C.J. Hillar and L.-H. Lim, Most tensor problems are NP-hard, *Journal of the ACM*, (2013), to appear.

[19] J.B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions with application to arithmetic complexity and statistics, *Linear Algebra Appl.* 18 (1977), 95–138.

[20] L.-H. Lim, Singular values and eigenvalues of tensors: a variational approach, *Proceedings of the IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing* (CAMSAP ’05), 1 (2005), 129–132.

[21] J.A. Rhodes, A concise proof of Kruskals theorem on tensor decomposition, *Linear Algebra Appl.* 432 (2010), 1818–1824.

[22] R. T. Rockafellar, *Convex Analysis*, Princeton Univ. Press 1970.