Ann Math Artif Intell (2015) 75:89–115
DOI 10.1007/s10472-014-9430-x

Combinatorial properties of support vectors of separating hyperplanes

Peter Veelaert

Published online: 23 September 2014

Abstract In this paper we study the relationship between separating hyperplanes and the Radon partitions of their support vectors. This study is relevant for maximal margin separations, which appear in Support Vector Machines (SVM), as well as for separations that optimize a Chebyshev norm. We propose a new version of the Stiefel exchange algorithm where we exploit the property that each Stiefel exchange is in fact a Radon exchange. Originally, the Stiefel exchange algorithm was developed to find Chebyshev approximations, but we show that it is also suited for finding hyperplane separations. We also show that many important properties in approximation theory are closely related to fundamental results in convex set theory, in particular to Helly's, Radon's and Caratheodory's Theorem. Within this context, we prove a new result that generalizes both Radon's and Caratheodory's Theorem.

Keywords Separating hyperplanes · Radon partition · Enclosing hyperplanes · Maximal margin · Caratheodory's Theorem

Mathematics Subject Classification (2010) 52A37

1 Introduction

One of the basic problems in machine learning is to find an optimal decision surface that separates two sets of data points. Often, as in support vector machines (SVMs), these decision surfaces are hyperplanes. The major challenge is that the separation takes place in some high-dimensional feature space, and that the data sets can be huge, sometimes even too large to store in computer memory.

Given two sets of data points, an SVM constructs the hyperplane of maximal margin, that is the hyperplane that separates the two sets and that has the largest distance to the nearest data points. The separating hyperplane is accompanied by two parallel supporting hyperplanes that contain a small set of data points, called support vectors. Finding the hyperplane

P. Veelaert (✉)
Ghent University, Valentin Vaerwyckweg 1, 9000 Ghent, Belgium
e-mail: peter.veelaert@ugent.be

of maximal margin and the support vectors is a constrained quadratic optimization problem, which in general requires huge data storage and expensive matrix calculations [13]. The properties of the support vectors are important for the development of efficient algorithms, as well as the generalizing behavior of an SVM.

A common approach to obtain acceptable computation times is to decompose the separation problem into subproblems. Chunking methods start with a small arbitrary subset of the data, and solve the problem for the subset. The support vectors of the first problem are then added to a second chunk of data, and the process is repeated [2, 3]. A prerequisite for the efficiency of this method is that the number of support vectors is small compared to the size of the data set. More recent decomposition methods put upper limits on the size of the working set and add or remove data points from the working set depending on how they violate the optimality criterion [11, 12, 19]. In Platt's Sequential Minimal Optimization (SMO) algorithm the working set is reduced to its minimum, that is, two elements [20]. The time complexity of a projected conjugate gradient chunking algorithm scales between linear and cubic in the size of the data set, while the SMO algorithm scales between linear and quadratic time [20].

A further step in this direction is the conversion of the maximal margin separation problem into finding the point in a convex polytope nearest to the origin. The support vectors can be derived from the vertices of the face containing the point nearest to the origin. Keerthi et al discuss several adaptations and extensions of Gilbert's nearest point algorithm and the Mitchell-Demanov-Malozemov algorithm, which makes these algorithms more suitable for the huge computational demands of SVMs [8, 13, 16]. Keerthi et all report computational speeds comparable to SMO.

Although all research on SVMs recognizes the importance of the support vectors, up to now the main focus has been on their statistical distribution over the data set. In this paper we examine the combinatorial properties of support vectors. Our support vector working sets will be elemental subsets, a concept borrowed from robust regression [10, 21]. The main contribution of this paper is the proof that the distribution of the support vectors in an elemental subset corresponds to a Radon partition. According to Radon's Theorem in a d-dimensional space, a set of $d+2$ points can always be uniquely partitioned into two subsets such that the intersection of their convex hulls is non-empty. We will show that if we project the support vectors orthogonally onto the separating hyperplane, we always obtain a Radon partition.

The combinatorial properties of support vectors also hold for other optimization problems. To find the separating hyperplane of maximal margin an SVM determines two parallel supporting hyperplanes at maximal distance from each other. A crucial factor is how this distance is measured. One modification of the separation problem is to compute the supporting hyperplanes such that the difference between their heights is maximized, instead of the margin. This modified separation problem is identical to a Chebyshev (or L_∞) approximation problem, which is a linear programming problem [28]. Again, the optimal hyperplane in the Chebyshev sense is determined by a small set of support vectors, a property known as the minimax property of de la Vallée Poussin [24, 28]. Also in this case we prove that the projections of the support vectors on the separating hyperplane form a Radon partition, but now we project them along the direction of a coordinate axis. In fact, we will show that a Chebyshev separation is also a maximal margin separation, provided its orthogonal projection yields a Radon partition as well. Other interesting relations between L_∞, L_1 and L_2 approximations are discussed in [4].

Also from the algorithmic viewpoint our results shed more light on how a separation algorithm converges towards a solution. A classical method to solve Chebyshev's problem is

Stiefel's exchange algorithm [18, 22]. Osborne and Watson showed that Stiefel's algorithm is equivalent to the simplex method applied to the dual [18, 28]. We will show that Stiefel exchanges are the same as Radon exchanges, which can be defined as point replacements in the elemental subsets that preserve Radon partitions [9]. In particular, we show that a Chebyshev separation always exists if the projections of the convex hulls of the two data sets overlap. This result follows immediately from a generalized version of Caratheodory's Theorem, also proven in this paper. Finally we note that the minimax property of de la Vallée Poussin, which is an essential ingredient of Stiefel's method, follows directly from Helly's Theorem on convex sets. This explains why in general there are $d+1$ support vectors when we separate two data sets in $\mathbb{R}^d$.

This paper is an extension of a previous paper in which a combinatorial separation algorithm was presented for the Chebyshev separation problem [27]. In Section 2 we examine the relation between enclosure and separation problems. Section 3 introduces a fast algorithm for finding enclosures. Section 4 shows how the enclosure algorithm can be used as a combinatorial separation algorithm. Section 5 contains the main result of the paper, where we prove sufficient conditions for the separation algorithm to produce a correct result. In Section 6 we prove a result on Radon partitions for separating hyperplanes that maximize the margin. The time complexity of the algorithms is discussed in Section 7.

2 Separation and enclosure

We first establish the relation between Chebyshev approximations and different kinds of separation and enclosure problems. Let $f_a(p) : \mathbb{R}^d \rightarrow \mathbb{R}$ denote a real function of the form

$$f_a(p) := x_d - (a_0 + a_1x_1 + \cdots + a_{d-1}x_{d-1}), \tag{1}$$

where the a_i represent coefficients and $p = (x_1, \ldots, x_d)$ is a point of $\mathbb{R}^d$. The equation $f_a(p) = 0$ defines a hyperplane in $\mathbb{R}^d$ with coefficients $a = (a_0, \ldots, a_{d-1})$. We first consider the Chebyshev approximation problem [28]. Let S be a finite subset of points p_i in $\mathbb{R}^d$. Let $\epsilon(S)$ denote the minimal value for ϵ for which the system

$$\epsilon \geq x_d - (a_0 + \cdots + a_{d-1}x_{d-1}) \geq -\epsilon \quad (p \in S) \tag{2}$$

is still feasible. Clearly, $\epsilon(S)$ can be found by linear programming, where we have to minimize ϵ, while the a_i and ϵ are subject to linear inequalities. If $a_0, \ldots, a_{d-1}$, and $\epsilon(S)$ represent a solution of (2), then the points of S are tightly enclosed by the hyperplanes $f_a(p) = \epsilon(S)$ and $f_a(p) = -\epsilon(S)$. The problem that we are interested in, however, is not an enclosure but a separation problem. Let S^+, S^- be two finite subsets of $\mathbb{R}^d$. We define $\delta(S^+, S^-)$ as the maximal value of δ for which the system

$$\begin{aligned} x_d - (a_0 + \cdots + a_{d-1}x_{d-1}) &\geq \delta \quad && (p \in S^+) \\ x_d - (a_0 + \cdots + a_{d-1}x_{d-1}) &\leq -\delta \quad && (p \in S^-) \end{aligned} \tag{3}$$

is still feasible. Clearly, the determination of $\delta(S^+, S^-)$ and the a_i is still a linear programming problem. In fact, we can easily convert it into a second kind of enclosure problem. First we rewrite (3) as

$$\begin{aligned} (x_d - \tau) - (a_0 + \cdots + a_{d-1}x_{d-1}) &\geq \delta - \tau \quad && (p \in S^+) \\ (x_d + \tau) - (a_0 + \cdots + a_{d-1}x_{d-1}) &\leq -\delta + \tau \quad && (p \in S^-) \end{aligned} \tag{4}$$

where τ is some real number. Now let $\epsilon = -\delta + \tau$, so that we can derive from S^+ and S^- two new sets:

$$\begin{aligned} T_\tau^+ &= \{q : q = p + (0, \ldots, 0, -\tau),\ \ p \in S^+\} \\ T_\tau^- &= \{q : q = p + (0, \ldots, 0, \tau), \quad\ p \in S^-\}, \end{aligned} \tag{5}$$

If we denote the coordinates of q as $q = (y_1, \ldots, y_d)$, then finding the maximal value for δ in (4), is the same as finding the minimal value for ϵ for which

$$\begin{aligned} y_d - (a_0 + \cdots + a_{d-1}y_{d-1}) &\geq -\epsilon \quad (q \in T_\tau^+) \\ y_d - (a_0 + \cdots + a_{d-1}y_{d-1}) &\leq \quad \epsilon \quad (q \in T_\tau^-) \end{aligned} \tag{6}$$

is still feasible. We shall denote this minimal value as $\epsilon(T_\tau^+, T_\tau^-)$. Clearly,

$$\epsilon(T_\tau^+, T_\tau^-) = \tau - \delta(S^+, S^-) \tag{7}$$

holds for any τ. Thus (6) takes the form of an enclosure problem in which we raised the points of S^- over a distance τ, while we lowered the points of S^+ over the same distance. The main difference between (2) and (6) is that the first system is symmetrical, while the second is not. In the second system there are two different sets, T_τ^+ and T_τ^- that define the upper and lower bounds for ϵ. We will call this problem a *signed enclosure problem* in which we have to compute $\epsilon(T_\tau^+, T_\tau^-)$, while (2) is called an *unsigned enclosure problem* where we compute $\epsilon(S)$, or $\epsilon(S^+ \cup S^-)$ for that matter.

Furthermore, we note that different values for τ yield different values for $\epsilon(T_\tau^+, T_\tau^-)$. However, once we know $\epsilon(T_\tau^+, T_\tau^-)$ for one particular value of τ, we can immediately derive $\delta(S^+, S^-)$ from (7). In the sections that follow, our goal will be first to establish an algorithm for computing $\epsilon(T_\tau^+ \cup T_\tau^-)$, and second to delineate the conditions under which $\epsilon(T_\tau^+, T_\tau^-)$ is equal to $\epsilon(T_\tau^+ \cup T_\tau^-)$.

3 A combinatorial algorithm for unsigned enclosures

3.1 Elemental subsets

In this section we describe a combinatorial algorithm for finding unsigned enclosures, which is based on elemental subsets [26]. Elemental is a term borrowed from robust regression [10, 21]. An elemental subset is any subset of the data that contains the minimum number of points needed to identify the parameters of the model. In our case, an elemental subset contains $d + 1$ points, because there is no unique solution for (2) when $|S| \leq d$.

Furthermore, we shall add one constraint to ensure that we can actually determine unique values for the parameters of the enclosing hyperplanes. We say that a set of d points p_i in $\mathbb{R}^d$ is in *general position* if there is a unique hyperplane $f_a(p) = 0$ passing through the points p_i, where f_a is defined as in (1).

Definition 1 Let S be a finite set of points in $\mathbb{R}^d$. An elemental subset E is a subset of S with $d + 1$ points, which has at least one d-point subset of points in general position.

The primary importance of elemental subsets stems from the fact that in the special case that E comprises all the points of S, (2) can be solved in an analytical way [25]. To this end, we define the $(d + 1) \times (d + 1)$ matrix

$$M_E := \begin{pmatrix} 1 & x_{11} & \ldots & x_{1d} \\ \ldots & & & \\ 1 & x_{(d+1)1} & \ldots & x_{(d+1)d} \end{pmatrix}, \tag{8}$$

where the entries x_{ij} come from the coordinates of the points $p_i = (x_{i1}, \ldots, x_{id})$ of E. Let C_i denote the cofactors of the last column of M_E, where $1 \le i \le d+1$. These cofactors play an important role with respect to the relative positions of the points in E and the enclosing planes $f_a(p) = \pm\epsilon$.

From now on S is always a finite subset of points in $\mathbb{R}^d$, and E a subset of S with $d+1$ points. We start with a simple observation, which follows immediately from linear algebra.

Lemma 1 *The subset E is an elemental subset if and only if at least one of the cofactors C_i of M_E is non vanishing.*

Proof First we show that if some cofactor $C_i \neq 0$, then there is always a unique hyperplane passing through a d point subset of E. Without loss of generality we may assume $C_{d+1} \neq 0$. Consider the system of d linear equations

$$x_{id} = a_0 + a_1 x_{i1} + \cdots + a_{d-1} x_{i(d-1)} \quad (i = 1, \ldots, d),$$

which defines the hyperplane passing through $p_1, \ldots, p_d$. If at least one $x_{id} \neq 0$, this is a non-homogeneous system, which has a unique solution since the determinant of its coefficient matrix, which is equal to C_{d+1}, is non-zero. If all $x_{id} = 0$ for $i = 1, \ldots, d$, this is a homogeneous system with the unique solution $(a_0, \ldots, a_{d-1}) = (0, \ldots, 0)$. In both cases there is a unique hyperplane $x_d = a_0 + a_1 x_1 + \cdots + a_{d-1} x_{d-1}$ passing through the points $p_i, i = 1, \ldots, d$.

Conversely, suppose $C_{d+1} = 0$. Then the above system either has infinitely many solutions or no solution at all, depending on whether it is homogeneous or not. In neither case it has a unique solution. If all cofactors are zero, none of the corresponding systems has a unique solution, and E is not an elemental subset. □

With Lemma 1 we can also attribute a more geometrical meaning to the notion of general position. Let $\pi(p)$ denote the projection of the point p on the plane $x_d = 0$, i.e., $\pi(p_i) = (x_{i1}, \ldots, x_{i(d-1)}, 0)$, and let $\pi(E)$ denote the set of projected points of E. According to Lemma 1 the uniqueness of the hyperplane depends on the projection $\pi(E)$. To give a specific example, for planes in $\mathbb{R}^3$, there is a unique plane $x_3 = a_0 + a_1 x_1 + a_2 x_2$ passing through 3 distinct points p_1, p_2, p_3, provided $\pi(p_1)$, $\pi(p_2)$, $\pi(p_3)$ are not collinear. More generally, the following conditions are equivalent:

- E is an elemental subset;
- one of the cofactors C_i is non-vanishing;
- the projected points $\pi(p_i)$ affinely span the plane $x_d = 0$;
- E contains at least one d point subset $p_1, \ldots, p_d$ for which there is no affine dependency of the form $\alpha_1 \pi(p_1) + \cdots + \alpha_d \pi(p_d) = 0$ with $\alpha_1 + \cdots + \alpha_d = 1$.

Note that the condition that the points of E span a hyperplane of $\mathbb{R}^d$ is not sufficient. In fact, this hyperplane could be perpendicular to the plane $x_d = 0$, in which case it cannot be of the form $x_d = a_0 + a_1 x_1 + \cdots + a_{d-1} x_{d-1}$.

3.2 Enclosure of elemental subsets

The cofactors not only determine the uniqueness of the hyperplane, but are also essential for the enclosure of E. We start by giving a geometrical interpretation to the value of the determinant of M_E and the cofactors C_i. The absolute value of the determinant is $d!$ times the volume of the d-simplex spanned by the $d+1$ vertices p_i. Similarly, the absolute value of C_i is $(d-1)!$ times the volume of the $(d-1)$-simplex spanned by the d projected vertices $\pi(p_1), \ldots, \pi(p_{i-1}), \pi(p_{i+1}), \ldots, \pi(p_{d+1})$.

For an elemental subset E we now introduce the ratio $\epsilon(E)$ of the volume of the d-simplex over the average volume of the $d-1$ simplices,

$$\begin{aligned}\epsilon(E) &:= \left(\frac{|\det M_E|}{d!}\right) / \frac{(|C_1|+\cdots+|C_{d+1}|)}{d(d-1)!} \\ &= |C_1 x_{1d} + \cdots + C_{d+1} x_{(d+1)d}| / (|C_1| + \cdots + |C_{d+1}|).\end{aligned} \tag{9}$$

We will call $\epsilon(E)$ the height of E. Since by Lemma 1, for an elemental subset at least one of the cofactors is non-zero, the denominator in (9) is always non-zero, and $\epsilon(E)$ is always defined. From (9) it is immediately clear that the height of E is zero when all the points of E lie in a common hyperplane. In fact, even when the points of E are not coplanar, $\epsilon(E)$ will give us a precise measure of how far the points of E lie from a common plane.

Theorem 1 *Let E be an elemental subset for which all the d-point subsets are in general position. Then there is a unique pair of hyperplanes $f_a(p) = \epsilon(E)$, $f_a(p) = -\epsilon(E)$ such that each point p_i in E lies on the hyperplane*

$$f_a(p_i) = sign\,(C_i) sign\,(\det M)\epsilon(E).$$

Furthermore, there is no hyperplane $f_a(p) = 0$ for which $|f_a(p)| < \epsilon$ for all $p \in E$.

The above theorem appears in various forms in the work on Chebyshev approximations [6]. A proof of this particular version can be found in [25, 26]. Geometrically, the theorem states that the height $\epsilon(E)$ is one half times the difference in height between the enclosing hyperplanes.

The requirement in Theorem 1 that all d-point subsets of E must be in general position, is stronger than what we had before. This new requirement is equivalent to the condition that all cofactors C_i are non zero, which is also known as the Haar condition [18, 28]. Although the Haar condition is necessary for Theorem 1, it is not needed to define the height $\epsilon(E)$. As discussed in [28], this has not always been clearly stated in the literature.

Also note that sign $(C_i) \in \{-1, 1\}$ as soon as E satisfies the Haar condition, but this does not prevent the possibility that sign (M_E) can be zero. If this happens, also the height of E will be zero, and all points of E will lie a unique common hyperplane.

On the other hand, to illustrate what happens when the Haar condition is not fulfilled, Fig. 1a and b show a configuration where all points $p_i = (x_i, y_i)$ have the same x coordinate, except for p_5. The elemental subset $E_{145} = \{p_1, p_4, p_5\}$ has matrix

$$M_{145} = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_4 & y_4 \\ 1 & x_5 & y_5 \end{pmatrix},$$

where $\det M_{145} \neq 0$. If we let C_1, C_4, C_5 denote the cofactors of the last column of M_{145}, then, since $x_1 = x_4 \neq x_5$, we have $C_1 \neq 0$, $C_4 \neq 0$, but $C_5 = 0$. As a result the enclosing hyperplanes $f_a(p) = \pm\epsilon(E)$ are not unique, although $\epsilon(E_{145})$ is well defined. Figure 1a and b show two possible enclosures where in each case the vertical height between the two supporting hyperplanes is equal to $\epsilon(E_{145})$. Since the sign of cofactor C_5 is not in $\{-1, 1\}$, the point p_5 can either lie on the lower supporting hyperplane, or on the higher supporting hyperplane.

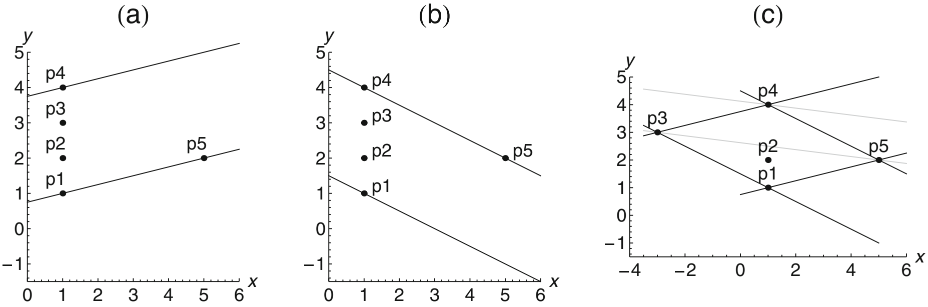

Fig. 1 Three special cases where some of the points have the same x-coordinate. As a result the enclosing (or supporting) hyperplanes are not unique. **c** shows a configuration where $\epsilon(S) = \epsilon(E_{134}) = \epsilon(E_{145})$. However, both E_{134} and E_{145} have some cofactors that are zero, and their enclosing hyperplanes are not unique. Therefore, the enclosing hyperplanes of S are also not unique. On the other hand, the elemental subset E_{345} has a unique pair of enclosing hyperplanes (shown as *gray lines* in **c**), but these hyperplanes do not enclose the entire set S, since $\epsilon(E_{345}) < \epsilon(S)$

3.3 Radon partition of an elemental subset

Geometric meaning cannot only be attributed to the absolute value of the cofactors, but also to their signs. Again according to Theorem 1, when sign (C_i)sign $(\det M)$ is positive, then the point p_i lies on the upper hyperplane $f_a(p) = \epsilon(E)$. When sign (C_i)sign $(\det M)$ is negative, then p_i lies on the lower hyperplane $f_a(p) = -\epsilon(E)$. It is important to note, however, that the cofactors only depend on the positions of the projected points $\pi(p_i)$. More in particular, the signs of the cofactors do not depend on the value of the d-th coordinate x_{id}. Thus the projections single-handedly determine a partition of the elemental subset into two parts. The sign of the determinant then determines to which hyperplane each part belongs. The partition is also invariant for any permutation of the rows of M_E.

For an elemental subset E that satisfies the Haar condition, we let $\{E^+, E^-\}$ denote the partition induced by the signs of the cofactors, that is,

$$\begin{aligned} E^+ &:= \{p_i : p_i \in E \text{ and } C_i > 0\}, \\ E^- &:= \{p_i : p_i \in E \text{ and } C_i < 0\}. \end{aligned} \tag{10}$$

We will show that the projections $\pi(E^+)$ and $\pi(E^-)$ form a Radon partition, whose existence is justified by Radon's Theorem [7]. Let conv S denote the convex hull of a finite set S.

Theorem 2 *(Radon's Theorem) Let $S = \{p_1, \ldots, p_r\} \subset \mathbb{R}^d$ be a finite set, and let $\{S_1, S_2\}$ be a partition of S, i.e., $S = S_1 \cup S_2$, $|S_1| \geq 1$, $|S_2| \geq 1$.*

(a) If $r \geq d + 2$ then the partition can be chosen such that conv $S_1 \cup$ conv $S_2 \neq \emptyset$.
(b) If $r = d + 2$ and any $d + 1$ points of S are affinely independent, then the partition in (a) is unique.

The condition in (b) is the Haar condition. Furthermore, when $|S_1 + S_2| = d + 2$ the partition $\{S_1, S_2\}$ is the so-called Radon partition [29]. We have formulated Radon's Theorem here in a form stronger than usual (compare with [29]), as we shall need this strong version later on .

We can now prove that the partition $\{\pi(E^+), \pi(E^-)\}$, as induced by the signs of cofactors, is also a Radon partition. What is more, the intersection of the convex spans of E^+ and E^- can be expressed by the cofactors.

Lemma 2 *Let C_i be the cofactors of the last column of the matrix M_E. Then we have*

$$\sum_{1\le i\le d+1} \pi(p_i)C_i = 0$$

and

$$\sum_{1\le i\le d+1} C_i = 0.$$

Furthermore, if E satisfies the Haar condition, then $\{\pi(E^+), \pi(E^-)\}$ is the unique Radon partition of $\pi(E)$.

Proof To prove the first part it suffices to replace the last column of M_E by the elements x_{ij} for some $j = 1, \ldots, d-1$. Since the determinant of a matrix with two identical columns is zero, $\sum_{1\le i\le d+1} x_{ij}C_i = 0$ for each $j = 1, \ldots, d-1$. It follows that $\sum_{1\le i\le d+1} \pi(p_i)C_i = 0$. Second, if we replace the elements of the last column of M_E all by 1, we again obtain a matrix with two identical columns, and therefore $\sum_{1\le i\le d+1} C_i = 0$.

Finally, assume E satisfies the Haar condition, or equivalently, that all cofactors are non-zero. Then we have

$$\sum_{p_i\in E^+} |C_i|\pi(p_i) = \sum_{p_i\in E^-} |C_i|\pi(p_i) \tag{11}$$

while

$$\sum_{p_i\in E^+} |C_i| = \sum_{p_i\in E^-} |C_i| \neq 0.$$

When we divide each $|C_i|$ in (11) by $\sum_{p_i\in E^+} |C_i|$, we obtain on the left side a convex combination of the points in $\pi(E^+)$ that coincides on the right side with a convex combination of the points in $\pi(E^-)$. Therefore, $\{\pi(E^+), \pi(E^-)\}$ must be a Radon partition. Since E satisfies the Haar condition, all d points subsets of $\pi(E)$ are affinely independent and the Radon partition is unique. □

Thus the Radon partition of $\pi(E)$ is the same as the partition $\{\pi(E^+), \pi(E^-)\}$, with E^+ and E^- defined by (10). By Theorem 1 this partition also coincides with the distribution of the points over the two enclosing hyperplanes of E. It is important to emphasize that the partition $\{E^+, E^-\}$ only depends on the relative positions of the points in the projected set $\pi(E)$. The partition does not depend on the x_{id} coordinate of the points p_i.

Unsigned enclosures are optimal in the following sense. There is no signed enclosure that is tighter than the unsigned enclosure of Theorem 1 [25, 26]. To formulate this more precisely, given a partition $\{S^+, S^-\}$ of S, for each elemental subset E in $S^+\cup S^-$, we define

$$\begin{aligned} E_S^+ &:= E\cap S^+, \\ E_S^- &:= E\cap S^-. \end{aligned}$$

Thus $\{E_S^+, E_S^-\}$ is the partition of E as enforced by $\{S^+, S^-\}$, while $\{E^+, E^-\}$ is the partition for which $\{\pi(E^+), \pi(E^-)\}$ is a Radon partition. The following result is proven in [25].

Theorem 3 *Let E be an elemental subset of S that satisfies the Haar condition. Then for any partition $\{S^+, S^-\}$ of S, we have $\epsilon(E) \leq \epsilon(E_S^+, E_S^-)$.*

From Theorems 1 and 3 we conclude that $\epsilon(E_S^+, E_S^-)$ takes its minimal value when the partition $\{E_S^+, E_S^-\}$ coincides with the partition as defined in (10). If we enforce a partition that differs from the partition induced by the cofactor signs, the enclosure will be less tight. So in general, the height of the unsigned enclosure will not be the same as the height of a signed enclosure. In Sections 4 and 5 we will establish conditions under which both heights are equal nevertheless.

3.4 Unsigned enclosure of arbitrary sets

Up to now, all results refer to the unsigned enclosure of elemental subsets. Let S be a (large) finite set of points that contains at least one elemental subset. We extend our definition of the height of a set as follows,

$$\epsilon(S) := \max_{E \subseteq S} \epsilon(E), \tag{12}$$

where the maximum of $\epsilon(E)$ is taken over all elemental subsets E in S.

With respect to unsigned enclosures, we have the following result by de la Vallée Poussin [24], which is very effective in Chebyshev approximations.

Theorem 4 *Let S be a finite set of points and let E be an elemental subset for which $\epsilon(E) = \epsilon(S)$. Let $f_a(p) = 0$ be a hyperplane such that $|f_a(p)| \leq \epsilon(E)$, for all $p \in E$. Then $|f_a(p)| \leq \epsilon(E)$, for all $p \in S$. Furthermore, for any $\epsilon < \epsilon(E)$ there is no hyperplane $f_a(p) = 0$ such that $|f_a(p)| \leq \epsilon$, for all $p \in S$.*

Geometrically, this means that the height of S is the maximum of the heights of all its elemental subsets. Originally, Theorem 4 was proven within the framework of Chebyshev approximations [18, 24]. However, as shown in [25], this theorem also readily follows from Helly's theorem on convex sets.

Helly's Theorem states that in $\mathbb{R}^d$ a finite collection of convex sets has a non-empty intersection if the intersection of every $d+1$ of these sets has a non-empty intersection. More specifically, since each inequality in (2) defines a convex set, this system of inequalities has a solution if each subsystem with $d + 1$ inequalities has a solution. As a matter of fact, each elemental subset defines a subsystem

$$\epsilon \geq x_d - (a_0 + \cdots + a_{d-1}x_{d-1}) \geq -\epsilon \quad (p \in E)$$

with $d + 1$ inequalities, and a minimal value $\epsilon(E)$ for which this subsystem is still feasible. Therefore, as a consequence of Helly's Theorem, the elemental subset E in S for which $\epsilon(E)$ is maximal defines the smallest value for ϵ such that the larger system (2) still has a solution. In fact, if we introduce $\epsilon = \max_{E \subseteq S} \epsilon(E)$ in (2), then all its $(d + 1)$-inequality subsystems will be feasible, and therefore the entire system will have a solution.

Whether the hyperplane in Theorem 4 is unique depends on the number of elemental subsets E that yield the same maximal height $\epsilon(S)$, as well as their cofactors. For example, in Fig. 1c, we have $\epsilon(S) = \epsilon(E_{145}) = \epsilon(E_{134})$. Furthermore, we note that in (12) the maximum is taken only over the elemental subsets of S. For example, in Fig. 1a we have

$$\epsilon(S) = \max\{\epsilon(E_{125}), \epsilon(E_{135}), \ldots, \epsilon(E_{345})\},$$

where the sets $\{p_1, p_2, p_3\}, \ldots, \{p_2, p_3, p_4\}$ are not included, since they are not elemental subsets.

3.5 The Haar assumption

In the sequel, to simplify the exposition we will avoid the special cases described in Fig. 1, and we will assume that all the d point subsets of S are in general position. In other words, we assume that each subset of $d + 1$ points is an elemental subset that satisfies the Haar condition. With respect to Fig. 1 this means that no two points can have the same x coordinate.

In practice, when working with real data sets, the Haar condition can be enforced by guarding the sign tests of the cofactors. When the value $|C_i|$ is less than some small positive real number, we add a small perturbation to the position of some of the data points. This has the additional advantage that all floating point computations will be more reliable [15].

3.6 Exchange algorithm for unsigned enclosures

With Theorem 4 we have a result that defines the height of S in terms of the heights of the elemental subsets. This result has an immediate corollary [26]. It suffices to note that an elemental subset can only give rise to enclosing hyperplanes if its height is maximal. Or conversely, if a point $q \in S$ is not enclosed by the supporting hyperplanes of some elemental subset E, then $E \cup \{q\}$ must have a height that is larger than the height of E (since otherwise the supporting hyperplanes of E would also enclose $E \cup \{q\}$). On the other hand, $E \cup \{q\}$ must contain at least one elemental subset that has the same height as $E \cup \{q\}$. Thus, we have the following result, which is also known as the Stiefel exchange property [18, 26].

Corollary 1 (**Stiefel exchange**) *Let E be an elemental subset of S. Let $f_a(p) = \epsilon(E)$ and $f_a(p) = -\epsilon(E)$ denote the two hyperplanes that enclose the points of E. If there is a point $q \in S$ for which $|f_a(q)| > \epsilon(E)$ then the set $E \cup \{q\}$ contains at least one elemental subset D for which $\epsilon(D) > \epsilon(E)$.*

The Stiefel exchange property states that when a point q is not enclosed by the supporting hyperplanes of E, then there is a point p in E such that if replace p by q, we obtain a new elemental subset that has a larger height than E. However, Corollary 1 does not specify how we can identify the point that has to be replaced.

The selection of p again depends on the properties of Radon partitions. The following lemma shows that we can use the cofactors C_i to select p. The proof is an adaptation of the proof of the Radon Exchange Lemma in [9]. Here we use the additional fact that the convex dependences in a Radon partition can be expressed by the cofactors of M_E, as in (11). We give a detailed proof because the way in which we select p will also be used in the exchange algorithm that follows.

Lemma 3 (**Radon exchange**) *Let E be an elemental subset of S, and let $\{\pi(E^+), \pi(E^-)\}$ be the Radon partition of $\pi(E)$. If q is any point of S, then there is a point $p \in (E^+ \cup E^-)$ such that*

$$\{\pi(E^+ \cup \{q\} \setminus \{p\}), \pi(E^- \setminus \{p\})\}$$

is a Radon partition.

Proof Let I^+ denote the set of indices of the points p_i contained in E^+, and let I^- denote the indices of the points contained in I^-. Then, by Lemma 2,

$$\sum_{i\in I^+} |C_i|\pi(p_i) = \sum_{i\in I^-} |C_i|\pi(p_i), \tag{13}$$

where

$$\sum_{i\in I^+} |C_i| = \sum_{i\in I^-} |C_i|.$$

Since E is an elemental subset, $\pi(q)$ is affinely dependent on the $d+1$ points $\pi(p_i)$, and there must be α_i, β_i such that

$$\begin{aligned} \pi(q) + \textstyle\sum_{i\in I^+} \alpha_i\pi(p_i) &= \textstyle\sum_{i\in I^-} \beta_i\pi(p_i), \\ 1 + \textstyle\sum_{i\in I^+} \alpha_i &= \textstyle\sum_{i\in I^-} \beta_i, \end{aligned} \tag{14}$$

where α_i and β_i are not necessarily positive. Let μ be the minimum of the ratios $\alpha_i/|C_i|$, $i \in I^+$, and $\beta_i/|C_i|$, $i \in I^-$. We subtract (13) μ times from (14) to obtain an expression of the form

$$\pi(q) + \sum_{i\in I^+} \lambda_i\pi(p_i) = \sum_{i\in I^-} \delta_i\pi(p_i),$$

where $\lambda_i = \alpha_i - \mu|C_i|$, $\delta_i = \beta_i - \mu|C_i|$. Furthermore, $\lambda_i \ge 0$, $\delta_i \ge 0$, and some λ_i or δ_i will be zero, while $1 + \sum_{i\in I^+} \lambda_i = \sum_{i\in I^-} \delta_i$. Hence, q together with the points for which λ_i is strictly positive on the left side, and the points for which δ_i is strictly positive on the right side will form a Radon partition. □

In the above lemma we add q to E^+. This preference for E^+ is completely arbitrary, and we may as well add q to E^-. This will only change the choice of the point p that must be removed.

It remains to show that the height $\epsilon(E)$ will increase when we replace p by q. The proof of the lemma below is again based on classical results [6, 18, 22], but with the adaptation that we express the growth of height in terms of cofactors.

Lemma 4 *Let E be an elemental subset of S. Let $f_a(p) = 0$ be the hyperplane for which $|f_a(p_i)| = \epsilon(E)$, for all $p_i \in E$. Let q be any point of S such that $f_a(q) > \epsilon(E)$. Let p be selected as in Lemma 3, and define*

$$\begin{aligned} D^+ &:= E^+ \cup \{q\} \setminus \{p\} \\ D^- &:= E^- \setminus \{p\}. \end{aligned}$$

Then $D = D^+ \cup D^-$ is an elemental subset whose height is larger than the height of E.

Proof Let λ denote the column vector $(C_1, \ldots, C_{d+1})^T$ of the cofactors. Let $b_0, \ldots, b_{d-1}$ be the coefficients of any hyperplane, and let $\theta = (b_0, \ldots, b_{d-1}, -1)^T$ denote the column vector of the coefficients b_i extended with -1. We define the column vector

$$r := M_E\theta$$

whose components r_i are called the residuals of the points p_i. By Lemma 2, we have $\sum_i C_i \pi(p_i) = 0$. Hence we have

$$\begin{aligned} |\lambda^T r| &= |\lambda^T M_E \theta| \\ &= |(0, \ldots, 0, \textstyle\sum_i C_i x_{id})\, (b_0, \ldots, b_{d-1}, -1)^T| \\ &= |-\sum_i C_i x_{id}| \\ &= \epsilon(E) \sum_i |C_i|, \end{aligned}$$

where we used (9) in the last step. On the other hand $\lambda^T r = \sum C_i r_i$, and we find the following relation for the residuals r_i:

$$\epsilon(E) \sum_i |C_i| = |\sum_i C_i r_i|. \tag{15}$$

Now we apply (15) to the elemental subset D, but where the coefficients in θ are those of the best fitting hyperplane of the elemental subset E. Note that with respect to this hyperplane the residuals r_i of the points p_i are all equal to $\pm\epsilon(E)$. Moreover, the signs of the residuals are either equal to the signs of the corresponding cofactors, or equal to the opposite signs. Let p_k be the point in E that has been replaced by q. Let D_i denote the cofactors of the last column of M_D, where M_D is the matrix M_E in which we replaced the k-th row by the row vector of q. Then (15) takes the form

$$\epsilon(D) \sum_i |D_i| = \sum_{i \neq k} |D_i| \epsilon(E) + |D_k| r_q,$$

where r_q is the residual of q. Hence, we have

$$\epsilon(D) = \epsilon(E) + \frac{|D_k|(r_q - \epsilon(E))}{\sum_i |D_i|}, \tag{16}$$

which means that $\epsilon(D) > \epsilon(E)$ as long as $r_q > \epsilon(E)$. This will be the case when $f_a(q) > \epsilon(E)$. □

Although the above lemma was stated for a point q that lies above the supporting hyperplanes, the result is equally true for any point q that lies below the supporting hyperplanes, provided we add q to E^- instead of E^+.

The above results form the basis for an exchange algorithm for unsigned enclosures [26]. The basic idea is to increase the height of an elemental subset by replacing one of its points until we have found an elemental subset of maximal height. This technique is also known as the Stiefel exchange method [22, 28]. In the version given below we add an important improvement, however. We use the cofactors of M_E to select the point p that has to be replaced, which speeds up the computation significantly, because we do not have to compute all the heights of the $d+2$ elemental subsets of $E \cup \{q\}$.

Exchange algorithm for unsigned enclosures.

Input: A finite subset S of $\mathbb{R}$.

Output: An elemental subset E such that $\epsilon(E) = \epsilon(S)$.

1. Select an arbitrary initial elemental subset E of S.
2. Use the cofactors of M_E to compute the height $\epsilon(E)$ and the Radon partition of $\pi(E)$.
3. Compute the best fit f_a of E by solving the linear system
   $$x_{id} - (a_0 + a_1 x_{i1} + \cdots + a_{d-1} x_{(d-1)}) = \text{sign}\,(C_i)\text{sign}\,(\det M_E)\epsilon(E),$$
   $(x_{i1}, \ldots, x_{i(d-1)}) \in E$, for the d unknowns $a_0, \ldots a_{d-1}$. Although this system has $d + 1$ equations, its rank is equal to d.
4. Process all points of S until a point q is found at a distance further than $\epsilon(E)$ from the best fit.
5. If no such point is found, return the current elemental subset E.
6. Perform a Radon exchange. Replace E by a new elemental subset D of $E \cup \{q\}$ such that $\epsilon(D) > \epsilon(E)$. If $\pi(E \cup \{q\})$ satisfies the Haar condition, the point p that has to be replaced by q can be found by the Radon exchange computation used in the proof of Lemma 3. If $\pi(E \cup \{q\})$ does not satisfy the Haar condition, D can still be found by computing the height of all the $d + 1$ point subsets of $E \cup \{q\}$. According to Corollary 1, at least one of the $d + 1$-point subsets of $E \cup \{q\}$ will have a height larger than E.
7. Proceed with step 2.

Note that although in our theoretical results we made the assumption that all elemental subsets of S satisfy the Haar condition, in the above exchange algorithm we have taken into account the possibility that some elemental subsets do not meet this requirement. An alternative approach would be to apply small perturbations to the data set so that the Haar condition can be met.

Also note that the point q with the largest residual will not always give the largest increase in height, because this increase also depends on the cofactors of the new elemental subset D, as specified in (16). In practice, however, one simply chooses the point with largest residual because it takes as much time to compute the cofactors of D as it takes to do one more iteration. Figure 2 illustrates how the enclosure algorithm iterates towards a solution for a simple example.

The time complexity of the exchange algorithm will be further addressed in Section 7.

4 Signed separations from unsigned enclosures

The exchange algorithm determines the unsigned enclosure and height $\epsilon(S)$ of a finite set S. However, to solve a separation problem, we must either determine the maximal value of δ in (3), or after raising the points of S^- and lowering the points of S^+, determine the minimal value of ϵ in (6). Both problems are signed problems with asymmetrical constraints.

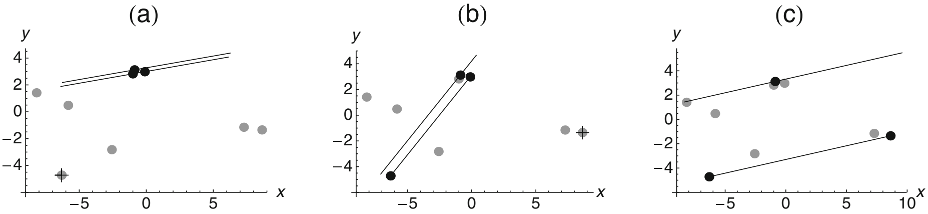

Fig. 2 Subsequent iterations in the enclosure algorithm. In **a** we start with an arbitrary elemental subset and the two supporting lines that enclose it. If we extend the subset E with any point p that is not enclosed, i.e, the crossed point in **a**, then one of the elemental subsets $E \cup \{p\}$ will have a larger height than E. This enables us to replace E by a new elemental subset, shown in **b**, which contains p and two previous points. The process continues as long as not all the points of S are enclosed by the supporting lines E. A solution is found after three iterations, as shown in **c**

Figure 3 shows the difference between a signed and an unsigned enclosure. The dark points are the result of raising of some set S^-; the light points depict the lowering of S^+. Figure 3a shows the unsigned enclosure. One of the supporting lines contains points of S^- as well as S^+, and therefore it cannot be the supporting line of a signed enclosure. Figure 3b shows a signed enclosure.

Figure 4 shows a configuration of lowered and raised points for which the signed and unsigned enclosure do coincide. Each supporting line only contains points either from S^- or from S^+. In accordance with what we expect for an unsigned enclosure, the distribution of the support vectors over the two support lines yields a Radon partition when we project the points upon the x_1-axis. This means that the convex hulls of the projected sets $\pi(S^-)$ and $\pi(S^+)$ must intersect.

In the sequel we will show that for separations of raised and lowered points the opposite is also true, which is the main result of this paper. We will prove that the unsigned and signed enclosures coincide when the intersection of the convex hulls of $\pi(S^-)$ and $\pi(S^+)$ is non-empty, and provided that the distance τ over which we lower and raise the points is sufficiently large.

The following lemma states that an elemental subset E for which the Radon partition of $\pi(E)$ does not coincide with the partition imposed by S^+ and S^-, cannot be an enclosing subset of S when τ is sufficiently large. This is the first step in proving that we can obtain separations from unsigned enclosures if the points are sufficiently raised or lowered.

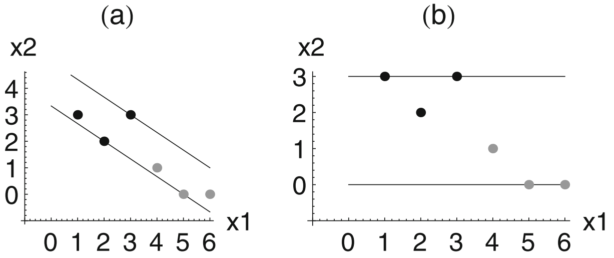

Fig. 3 **a** Unsigned enclosure, and **b** signed enclosure. The signed and unsigned enclosures are distinct. Also note that $\epsilon(S) \leq \epsilon(S^+, S^-)$ and that projections of the support vectors of the unsigned enclosure form a Radon partition

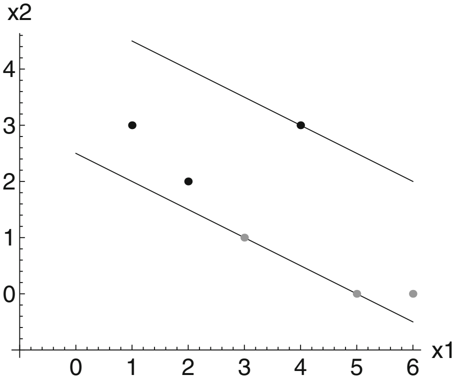

Fig. 4 The signed and unsigned enclosure coincide. In Section 5 we will prove that a sufficient condition for signed and unsigned enclosures to coincide is that the convex hulls of the projected sets have a non-empty intersection

Lemma 5 *Let $\{S^+, S^-\}$ be a partition of S, such that $S^+ \cup S^-$ contains at least one elemental subset E for which the Radon partition of $\pi(E)$ coincides with the partition $\{\pi(E_S^+), \pi(E_S^-)\}$. Let τ be a positive real number, and let T_τ^+ and T_τ^- denote the sets derived from S^+ and S^- as in (5). Let $T_\tau = T_\tau^+ \cup T_\tau^-$ and let $E_\tau^{\max}$ be the elemental subset in T_τ with the largest height, i.e.,*

$$E_\tau^{\max} := argmax_{E \subset T_\tau} \epsilon(E).$$

Then there exists a lower bound τ_0 such that

$$\{\pi(E_\tau^{\max} \cap T_\tau^+), \pi(E_\tau^{\max} \cap T_\tau^-)\} \tag{17}$$

is a Radon partition of $\pi(E_\tau^{\max})$ for all $\tau \geq \tau_0$.

Proof Let $E = \{p_1, \ldots, p_{d+1}\}$ be an elemental subset of S. Let $I = \{1, \ldots, d+1\}$ denote the set of indices of the points in E. We partition I into four different subsets:

$$\begin{array}{lll} I^{++} := \{i : p_i \in E \cap S^+ & \text{and} & C_i > 0\} \\ I^{--} := \{i : p_i \in E \cap S^- & \text{and} & C_i < 0\} \\ I^{+-} := \{i : p_i \in E \cap S^+ & \text{and} & C_i < 0\} \\ I^{-+} := \{i : p_i \in E \cap S^- & \text{and} & C_i > 0\}. \end{array} \tag{18}$$

Since $C_i \neq 0$ for all i, we have $I = I^{++} \cup I^{--} \cup I^{+-} \cup I^{-+}$. We will use C_I as a shorthand for the value of $\sum_{i \in I} |C_i|$.

The height of E can then be written as

$$\epsilon(E) = \left(|\sum_{i \in I^{++}} x_{id} C_i + \sum_{i \in I^{--}} x_{id} C_i + \sum_{i \in I^{+-}} x_{id} C_i + \sum_{i \in I^{-+}} x_{id} C_i| \right) / C_I. \tag{19}$$

There are only two ways in which the signs of the cofactors can coincide with the partition (17), either $I^{+-} = I^{-+} = \emptyset$, or $I^{++} = I^{--} = \emptyset$. First assume that $I^{+-} = I^{-+} = \emptyset$. Then (19) is equal to

$$\epsilon(E) = \left(|\sum_{i \in I^{++}} x_{id} C_i + \sum_{i \in I^{--}} x_{id} C_i| \right) / C_I.$$

When we raise the points of S^- by τ, while we lower the points of S^+ also by τ the height will change:

$$\epsilon(E_\tau) = \left(|\sum_{i \in I^{++}} (x_{id} - \tau)C_i + \sum_{i \in I^{--}} (x_{id} + \tau)C_i| \right) / C_I \tag{20}$$

where E_τ contains the raised and lowered points of E. Since $C_i > 0$ for $i \in I^{++}$ and $C_i < 0$ for $i \in I^{--}$, we obtain

$$\epsilon(E_\tau) = \left(\left| \sum_{i \in I} x_{id} C_i - \tau \sum_{i \in I} |C_i| \right| \right) / C_I.$$

Hence, for τ sufficiently large, i.e, $\tau > | \sum_{i \in I} x_{id} C_i | / C_I$, the height $\epsilon(E_\tau)$ increases as

$$\epsilon(E_\tau) = \tau - \epsilon(E). \tag{21}$$

The case $I^{++} = I^{--} = \emptyset$ leads to a similar result. Also in this case $\epsilon(E_\tau)$ increases as in (21).

Now let $F = \{q_1, \ldots, q_{d+1}\}$ be a second elemental subset for which $\{F_S^+, F_S^-\}$ is not a Radon partition. We again define index sets as in (18). In this case we have $I^{-+} \cup I^{+-} \neq \emptyset$ as well as $I^{++} \cup I^{--} \neq \emptyset$. We denote the coordinates of the points of F as $q_i = (y_{i1}, \ldots, y_{id})$. If we let F_τ denote the set of lowered and raised points of F, we have

$$\epsilon(F_\tau) = \left(\left| \sum_{i \in I^{++}} (y_{id} - \tau)C_i + \sum_{i \in I^{--}} (y_{id} + \tau)C_i + \sum_{i \in I^{+-}} (y_{id} - \tau)C_i + \sum_{i \in I^{-+}} (y_{id} + \tau)C_i \right| \right) / C_I.$$

Since, $C_i > 0$ for $i \in I^{++} \cup I^{-+}$ and $C_i < 0$ for $i \in I^{--} \cup I^{+-}$ we obtain

$$\epsilon(F_\tau) = \left(\left| \sum_{i \in I} y_{id} C_i - \tau \left(\sum_{i \in I^{++} \cup I^{--}} |C_i| - \sum_{i \in I^{+-} \cup I^{-+}} |C_i| \right) \right| \right) / C_I.$$

Because in this case $I^{-+} \cup I^{+-} \neq \emptyset$ and $I^{++} \cup I^{--} \neq \emptyset$, it follows that

$$0 \leq \left| \sum_{i \in I^{++} \cup I^{--}} |C_i| - \sum_{i \in I^{+-} \cup I^{-+}} |C_i| \right| / C_I < 1.$$

Hence, for τ sufficiently large, the height $\epsilon(F_\tau)$ increases as

$$\epsilon(F_\tau) = c_0 + c_1 \tau, \tag{22}$$

where c_0, c_1 are constants and $0 \leq c_1 < 1$. Thus, by comparing (21) with (22), we conclude that when τ is sufficiently large, $\epsilon(E_\tau)$ will always be larger than $\epsilon(F_\tau)$. Hence, the partition in (17) is a Radon partition for τ sufficiently large. □

We have shown that for τ sufficiently large, the height of those elemental subsets E for which $\{\pi(E_S^+), \pi(E_S^+)\}$ is a Radon partition will dominate the height of the other elemental subsets. In essence this means that the exchange algorithm can also be of use to compute signed enclosures. However, Lemma 5 can only be applied if there is at least one elemental subset for which the Radon partition coincides with the partition imposed by $\{S^+, S^-\}$. At this moment it is not clear yet when such an elemental subset is available. In the next section we will settle this question.

5 Sufficient conditions for the existence of Radon partitions

In this section we will derive more explicit conditions that ensure that $S^+ \cup S^-$ contains at least one elemental subset E for which the Radon partition $\{E^+, E^-\}$ coincides with the partition $\{E_S^+, E_S^-\}$ as imposed by $\{S^+, S^-\}$. This correspondence is needed for the application of Lemma 5. The conditions formulated here are more general than those in [27].

Our proof will be based on the following construction. If the convex hulls of $\pi(S^+)$ and $\pi(S^-)$ have a non-empty intersection, we can find always subsets $P \subseteq S^+$, and $Q \subseteq S^-$, such that $\operatorname{conv}\pi(P) \cap \operatorname{conv}\pi(Q) \neq \emptyset$, and $|P| + |Q| = d + 1$. In fact, $E = P \cup Q$ will be an elemental subset for which $\{\pi(E_S^+), \pi(E_S^-)\}$ is a Radon partition.

According to Caratheodory's theorem if a point p lies in the convex hull of a subset S of $\mathbb{R}^k$, then there is a subset P of S with $k + 1$ or fewer points such that p lies in the convex hull of P. Caratheodory's theorem was used to prove the main result of [27]. Here, to prove a more general result, we need the following extension of Caratheodory's theorem. Note that for the sake of notational simplicity, we formulate this theorem in $\mathbb{R}^d$, but later we will apply it on the projected sets in $\mathbb{R}^{d-1}$.

Theorem 5 *Let P, Q be two finite sets in $\mathbb{R}^d$ such that conv $P \cap$ conv $Q \neq \emptyset$. Then there are subsets $P' \subset P$, $Q' \subset Q$ such that*

(*a*) $|P'| \geq 1$, $|Q'| \geq 1$,
(*b*) $|P'| + |Q'| \leq d + 2$,
(*c*) *conv* $P' \cap$ *conv* $Q' \neq \emptyset$.

Proof Let p_i denote the points of P, q_i denote the points of Q. Since conv $P \cap$conv $Q \neq \emptyset$, there is a point s that can be written as a convex combination of both sets, i.e.,

$$s = \sum_{i=1}^{n} \alpha_i p_i = \sum_{i=1}^{m} \beta_i q_i \tag{23}$$

where $\sum_{i=1}^{n} \alpha_i = 1$, with $\alpha_i \geq 0$, and $\sum_{i=1}^{m} \beta_i = 1$, with $\beta_i \geq 0$. If $n + m \leq d + 2$, we are done. So we may assume that $n + m > d + 2$. Now suppose one of the convex combinations only contains one point, e.g,

$$p_j = \sum_{i=1}^{m} \beta_i q_i .$$

Then we can choose $P' = \{p_j\}$ and by Caratheodory's original theorem [29], p_j is a convex combination of a subset Q' that contains at most $d + 1$ points, which proves the theorem for this particular case.

Therefore, from now on we assume that $n + m > d + 2$, as well as $n \geq 2$, and $m \geq 2$. The theorem is obviously true for $d = 1$. Therefore we may also restrict ourselves to the case $d \geq 2$. It follows that at least one of n or m is larger than 2. Without loss of generality we assume $n > 2$. Since $n + m - 2 > d$, the $n + m - 2$ points $(p_3 - p_1), \ldots, (p_n - p_1)$,

$(q_1 - p_1), \ldots, (q_m - p_1)$ are linearly dependent. Hence there are real numbers μ_{1i}, γ_{1i} not all zero, such that

$$\sum_{i=3}^{n} \mu_{1i}(p_i - p_1) + \sum_{i=1}^{m} \gamma_{1i}(q_i - p_1) = 0.$$

If μ_{11} is defined as $\mu_{11} := -(\sum_{i=3}^{n} \mu_{1i} + \sum_{i=1}^{m} \gamma_{1i})$, then

$$\mu_{11} p_1 + \sum_{i=3}^{n} \mu_{1i} p_i + \sum_{i=1}^{m} \gamma_{1i} q_i = 0 \tag{24}$$

with

$$\mu_{11} + \sum_{i=3}^{n} \mu_{1i} + \sum_{i=1}^{m} \gamma_{1i} = 0.$$

Since at least one of the coefficients in (24) is non-zero, without loss of generality we may assume that $\mu_{11} \neq 0$. If this was not yet the case, we can always relabel the points of P and Q, and, if necessary, also swap the roles of P and Q.

Likewise, since the $n + m - 2$ points $(p_3 - p_2), \ldots, (p_n - p_2)$, $(q_1 - p_2), \ldots, (q_m - p_2)$ are also linearly dependent, we can find real numbers μ_{2i}, γ_{2i} not all zero, such that

$$\mu_{22} p_2 + \sum_{i=3}^{n} \mu_{2i} p_i + \sum_{i=1}^{m} \gamma_{2i} q_i = 0 \tag{25}$$

with

$$\mu_{22} + \sum_{i=3}^{n} \mu_{2i} + \sum_{i=1}^{m} \gamma_{2i} = 0.$$

Now, we introduce

$$\begin{aligned} \mu_{1s} &:= \mu_{11} + \textstyle\sum_{i=3}^{n} \mu_{1i} \\ \mu_{2s} &:= \mu_{22} + \textstyle\sum_{i=3}^{n} \mu_{2i}. \end{aligned}$$

Then, we also have $-\mu_{1s} = \sum_{i=1}^{m} \gamma_{1i}$, and $-\mu_{2s} = \sum_{i=1}^{m} \gamma_{2i}$.

First, suppose that $\mu_{1s} = 0$, then we have found a combination

$$\sum_{i=1}^{n} \mu_{1i} p_i = \sum_{i=1}^{m} (-\gamma_{1i}) q_i, \tag{26}$$

with $\sum_{i=1}^{n} \mu_{1i} = 0$, as well as $\sum_{i=1}^{m} (-\gamma_{1i}) = 0$, and where at least one of the μ_{1i} or $(-\gamma_{1i})$ is not zero. Then for any τ

$$s = \sum_{i=1}^{n} (\alpha_i - \tau \mu_{1i}) p_i \tag{27}$$

and

$$s = \sum_{i=1}^{m} (\beta_i + \tau \gamma_{1i}) q_i \tag{28}$$

still represent convex combinations, provided $(\alpha_i - \tau \mu_{1i}) \geq 0$ and $(\beta_i + \tau \gamma_{1i}) \geq 0$. Furthermore, at least one of the μ_{1i} or $(-\gamma_{1i})$ is positive. Let κ denote the maximum of all μ_{1i} and $(-\gamma_{1i})$. We define

$$\lambda := \min\{\alpha_1/\kappa, \ldots, \beta_1/\kappa, \ldots\}.$$

Then we have $\lambda > 0$, and

$$\begin{aligned} \alpha_i - \lambda \mu_{1i} &\geq 0 \quad (i = 1, \ldots, n) \\ \beta_i + \lambda \gamma_{1i} &\geq 0 \quad (i = 1, \ldots, m). \end{aligned}$$

For at least one α_i or β_i the equality will hold. Therefore, in at least one of the two convex combinations (27) or (28) at least one of the coefficients will be zero. In other words, at least one of the convex combinations can be written with one point less. This process can be repeated as long as $n+m > d+2$, and the same process can be used when $\mu_{2s} = 0$.

We now consider the remaining case that $\mu_{1s} \neq 0$ and $\mu_{2s} \neq 0$. Then there is a non-zero scalar ρ such that $\mu_{2s} - \rho\mu_{1s} = 0$. Subtracting (24) ρ times from (25) we find

$$-\rho\mu_{11}p_1 + \mu_{22}p_2 + \sum_{i=3}^{n}(\mu_{2i} - \rho\mu_{1i})p_i = -\sum_{i=1}^{m}(\gamma_{2i} - \rho\gamma_{1i})q_i. \tag{29}$$

Since $-\mu_{1s} = \sum_{i=1}^{m}\gamma_{1i}$, and $-\mu_{2s} = \sum_{i=1}^{m}\gamma_{2i}$, we also have

$$\sum_{i=1}^{m}(\gamma_{2i} - \rho\gamma_{1i}) = 0,$$

and therefore

$$-\rho\mu_{11} + \mu_{22} + \sum_{i=3}^{n}(\mu_{2i} - \rho\mu_{1i}) = 0.$$

Since $\mu_{11} \neq 0$, $\rho \neq 0$, we have $\rho\mu_{11} \neq 0$, in other words, at least one of the coefficients in (29) is non-zero.

Hence, we can rewrite (29) as

$$\sum_{i=1}^{n}\mu_i' p_i = \sum_{i=1}^{m}(-\gamma_i')q_i, \tag{30}$$

with $\sum_{i=1}^{n}\mu_i' = 0$, and $\sum_{i=1}^{m}(-\gamma_i') = 0$, and where at least one of the coefficients is non-zero. Since the above expression has the same form and properties as (26), we can proceed as before and subtract the coefficients in (30) multiplied by an appropriate constant from α_i and β_i to eliminate a point in one of the convex combinations of (23). □

Theorem 5 encompasses Caratheodory's original theorem as a special case, if we let P consist of a single point. In addition, if we let $P = Q$ it implies part (a) of Radon's Theorem (Theorem 2).

Theorem 6 *Let $\{S^+, S^-\}$ be a partition of S such that*

$$conv\,\pi(S^+) \cap conv\,\pi(S^-) \neq \emptyset.$$

Then $S^+ \cup S^-$ contains an elemental subset E for which the Radon partition of $\pi(E)$ coincides with the partition $\{\pi(E_S^+), \pi(E_S^-)\}$.

Proof It suffices to combine the strong form of Radon's Theorem with the extension of Caratheodory's Theorem. First, we apply Theorem 5 to the $(d-1)$-dimensional space defined by the hyperplane $x_d = 0$. The projection $\pi(S^+ \cup S^-)$ lies in this hyperplane. Hence, there are two sets $P^+ \subset S^+$ and $P^- \subset S^-$ such that $P^+ \cup P^-$ contains $d+1$ points, and $\mathrm{conv}\,\pi(P^+) \cap \mathrm{conv}\,\pi(P^-) \neq \emptyset$. Now choose $E = P^+ \cup P^-$. We then have

$$\mathrm{conv}\,\pi(E_S^+) \cap \mathrm{conv}\,\pi(E_S^-) \neq \emptyset.$$

On the other hand, we assume that all subsets with $d+1$ points satisfy the Haar condition. Hence by Radon's Theorem (Theorem 2), there is only one way to partition E into two sets E^+ and E^- such that

$$\operatorname{conv}\pi(E^+)\cap\operatorname{conv}\pi(E^-)\neq\emptyset.$$

It follows that the Radon partition of $\pi(E)$ coincides with the partition $\{\pi(E_S^+), \pi(E_S^-)\}$. □

Note that the previous lemma formalizes our observation made in Figs. 3 and 4. In fact, the main result of this paper now follows immediately.

Theorem 7 *Let $\{S^+, S^-\}$ be a partition of S such that $\operatorname{conv}\pi(S^+)\cap\operatorname{conv}\pi(S^-)\neq\emptyset$. Let T_τ denote the set of raised and lowered points of S, that is, $T_\tau := T_\tau^+\cup T_\tau^-$. Then there is a lower bound τ_0 such that for all $\tau\geq\tau_0$*

$$\max_{E\subset T_\tau}\epsilon(E)=\epsilon(T_\tau)=\epsilon(T_\tau^+, T_\tau^-)=\tau-\delta(S^+, S^-).$$

Proof By Theorem 6 there is at least one elemental subset E in S for which the imposed partition $\{E_S^+, E_S^-\}$ is a Radon partition. The same must be true for the elemental subsets of T_τ, for any value of τ.

Let $E_\tau^{\max}$ denote the elemental subset of maximal height in T_τ, or more precisely, $\epsilon(E_\tau^{\max}) := \max_{E\subset T_\tau}\epsilon(E)$. By definition the height of $E_\tau^{\max}$ is the same as the height of T_τ, that is, $\epsilon(E_\tau^{\max})=\epsilon(T_\tau)$. This makes sense since by Theorem 4, the hyperplanes that enclose $E_\tau^{\max}$ also enclose T_τ. All these results hold for any value of τ.

However, when τ is sufficiently large, then by Lemma 5 we also have $\epsilon(E_\tau^{\max}) = \epsilon(E_\tau^{\max}\cap T_\tau^+, E_\tau^{\max}\cap T_\tau^-)$. By Theorem 1 the hyperplanes that enclose $E_\tau^{\max}$ also enclose T_τ. Since each of these enclosing hyperplanes either contains points from $E_\tau^{\max}\cap T_\tau^+$ or from $E_\tau^{\max}\cap T_\tau^-$, but not from both sets, the signed enclosure of $\{E_\tau^{\max}\cap T_\tau^+, E_\tau^{\max}\cap T_\tau^-\}$ is also a signed enclosure of $\{T_\tau^+, T_\tau^-\}$. It follows that $\epsilon(T_\tau^+, T_\tau^-)\leq\epsilon(E_\tau^{\max}\cap T_\tau^+, E_\tau^{\max}\cap T_\tau^-)$. On the other hand, by Theorem 3 we always have $\epsilon(T_\tau)\leq\epsilon(T_\tau^+, T_\tau^-)$. As a result, we have $\epsilon(T_\tau)=\epsilon(T_\tau^+, T_\tau^-)$, and therefore also $\epsilon(E_\tau^{\max})=\epsilon(T_\tau^+, T_\tau^-)$. The equality $\epsilon(T_\tau^+, T_\tau^-)=\tau-\delta(S^+, S^-)$ is the same as (7). □

Based on the previous results we can now propose an algorithm for signed separations. The signed separation algorithm is based on the enclosure algorithm, but where we displace the points over τ, and give a new interpretation of the output. The basic idea is illustrated in Fig. 5.

An essential characteristic of the signed separation algorithm is that we do not have to substitute an actual value for τ when we compute the heights of the elemental subsets. Let E denote an elemental subset of S. Without loss of generality we may assume that the first k points of E belong to S^+ and the remaining $d+1-k$ points belong to S^-. Let E_τ be the elemental subset obtained after raising the points of S^-, and lowering the points of S^+. To compute $\epsilon(E_\tau)$ we must compute the cofactors of

$$M_{E_\tau} := \begin{pmatrix} 1 & x_{11} & \dots & x_{1d}-\tau \\ \dots & & & \\ 1 & x_{k1} & \dots & x_{kd}-\tau \\ 1 & x_{(k+1)1} & \dots & x_{(k+1)d}+\tau \\ \dots & & & \\ 1 & x_{(d+1)1} & \dots & x_{(d+1)d}+\tau \end{pmatrix}.$$

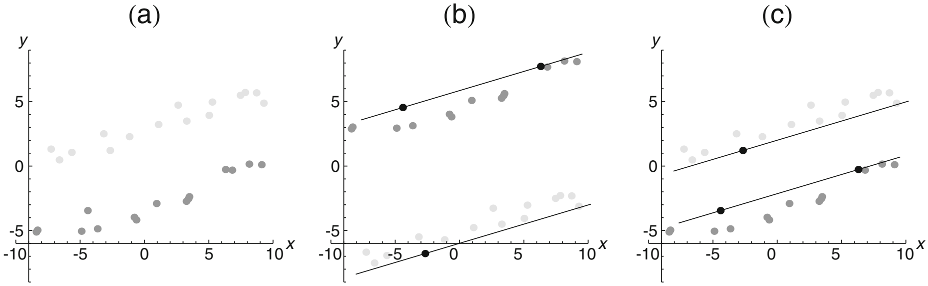

Fig. 5 How the separation algorithm uses enclosure to find hyperplanes that maximize the vertical distance. Two finite sets S^+ and S^- are shown in **a**. To find a separation we lower the points of S^+, raise the points of S^-, and use the enclosure algorithm to find the enclosing hyperplanes, as in **b**. The separating hyperplanes are then found by translating the points and the hyperplanes back to their original position, shown in **c**. In this illustration we used small displacements. When the height comparisons are correctly implemented in the algorithm, the displacements will always be sufficiently large

Since τ only appears in the last column of the matrix M_{E_τ}, the height $\epsilon(E_\tau)$ will always be of the form $|a + b\tau|$, where a, b are real numbers derived from the cofactors which do not depend on τ. When comparing heights or when computing the coefficients of the supporting hyperplanes, we simply take the limit for $\tau \to \infty$. To be precise, when comparing the heights of two elemental subsets E_τ and E'_τ, with heights $|a+b\tau|$ and $|a'+b'\tau|$, respectively, it suffices to compare the number pairs (a, b) and (a', b'). We define $|a + b\tau| < |a' + b'\tau|$ whenever $0 \leq |b| < |b'|$. If $|b| = |b'|$, we define $|a + b\tau| < |a' + b'\tau|$ whenever $|a| < |a'|$. This avoids the computation of very large numbers, and at the same time it guarantees that the outcome of the height comparisons always agrees with an optimal choice of τ. Furthermore, as demonstrated in the proof of Lemma (5), for the resulting elemental subset E we can only have $|b| = 1$ when the imposed partition $\{\pi(E_S^+), \pi(E_S^-)\}$ coincides with a Radon partition. In all other cases, $|b| < 1$. Hence, the computed value of $|b|$ can be used to determine whether such an elemental subset is present or not.

Algorithm for signed separations

Input: A finite subset $S = S^+ \cup S^-$ on $\mathbb{R}^d$ such that any subset of $d + 1$ points satisfies the Haar condition.

Output: Either an elemental subset E that defines the separating hyperplane that maximizes the separation distance, or the conclusion that no such elemental subset exists.

1. Use the enclosure algorithm to compute the elemental subset E_τ of largest height where the points of S^- have been lowered over a distance τ and the points of S^- have been raised over τ.
2. Let $|a + b\tau|$ denote the height of E_τ. If $|b| = 1$, we have found a subset for which the Radon partition $\{\pi(E^+), \pi(E^-)\}$ coincides with the imposed partition $\{\pi(E_\tau \cap T_\tau^+), \pi(E_\tau \cap T_\tau^-)\}$. This solves the signed separation problem. If $|b| < 1$, then $\operatorname{conv}\pi(T_\tau^+) \cap \operatorname{conv}\pi(T_\tau^-) = \operatorname{conv}\pi(S^+) \cap \operatorname{conv}\pi(S^-) = \emptyset$. In that case, there are no supporting hyperplanes that separate S^+ from S^- such that the projection $\{\pi(E_S^+), \pi(E_S^-)\}$ of the support vectors forms a Radon partition.

After applying the signed separation algorithm, there are three possible outcomes for the separation problem. First, the algorithm may return an elemental subset that defines a hyperplane that separates the given data sets S^+ and S^- in an optimal way.

Second, even when the algorithm returns an elemental subset E, after inspection we may find that the optimal hyperplane that separates E^+ from E^- does not separate S^+ from S^-. The reason is that the algorithm actually computes an enclosure of displaced points, which may not correspond to a separation after the displacement is undone. This situation invariably occurs when S^+ and S^- are not linearly separable, or in other words, when their convex hulls intersect. The hyperplane defined by E will then be the hyperplane for which the enclosure of the points that are at the wrong side of the separating hyperplane is minimal. This is similar to a separation with a soft margin by an SVM in the case of inseparable data sets [3].

Third, the outcome of the algorithm may be that no appropriate elemental subset can be found. Or in other words, the unsigned separation problem does not provide a solution for the signed problem, due to a peculiarity of the data set. This peculiarity, however, also opens the way for a fall-back strategy. Since the outcome now indicates that the convex hulls of $\pi(S^+)$ and $\pi(S^-)$ do not intersect, we can use the same algorithm to separate the projected sets $\pi(S^+)$ and $\pi(S^-)$ in $\mathbb{R}^{(d-1)}$. Any hyperplane in $\mathbb{R}^{(d-1)}$ of the form $x_{d-1} = a_0 + a_1 x_1 + \cdots + a_{d-2} x_{d-2}$ that separates $\pi(S^+)$ and $\pi(S^-)$ can always be embedded into a larger hyperplane orthogonal to $x_d = 0$ that separates S^+ and S^- in $\mathbb{R}^d$.

6 Maximal margin separation

In the previous sections we described a separation algorithm to find hyperplanes that maximize the distance to the nearest data points. The distance is measured in the sense of Chebyshev approximations, that is, along a fixed coordinate axis. As explained in Section 2 Chebyshev enclosures and separations give rise to linear programming problems. Since the separation/enclosure algorithms proceed by applying Radon exchanges to an elemental subset, they may be called combinatorial methods. Support vector machines, on the other hand, find hyperplanes that maximize the margin measured orthogonally to the hyperplane. This is a quadratic programming problem, which is usually solved by iterative methods [3].

Because quadratic programming poses challenges that are not present in linear programming, at first sight there appears to be a fundamental difference between a maximal margin problem and the separation problem of the previous sections. This is also confirmed by the examples. In general the hyperplane found by our combinatorial algorithm will not be the same as the maximal margin hyperplane, as illustrated in Fig. 6.

Nonetheless, we will show that the criterion for optimality turns out to be almost the same for both problems. We will show that a separating hyperplane H found by the combinatorial algorithm is also a separation with maximal margin if the orthogonal projection of the elemental subset E on H coincides with a Radon partition. Hence, the main difference is that we must project the support vectors orthogonally on the separating hyperplane instead of projecting them along the x_d-axis.

Suppose we have two data sets S^+ and S^- of points $p = (x_1, \ldots, x_d)$. To solve the maximal margin separation, we have to minimize

$$a_1^2 + \cdots + a_d^2$$

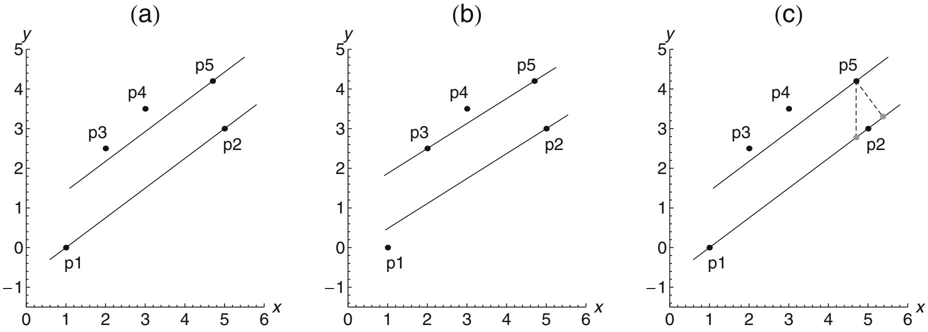

Fig. 6 Two data sets $S^- = \{p_1, p_2\}$ and $S^- = \{p_3, p_4, p_5\}$ that have to be separated. **a** indicates the combinatorial solution, while **b** shows the optimal supporting hyperplanes that maximize the margin, which is different. The reason why the combinatorial solution is not the same as the optimal solution is illustrated in **c**. If we project p_5 along the y-axis on the line $p_1 p_2$ then the projection falls within the convex hull of p_1 and p_2. However, if we project p_5 perpendicularly upon the line $p_1 p_2$, then the projection falls outside the convex hull of p_1 and p_2. Hence, p_1, p_2 and p_5 cannot be support vectors for an optimal hyperplane where we measure the distance perpendicularly. In Figure **b** the orthogonal projection of p_2 upon the line $p_3 p_5$ falls within the convex hull of p_3 and p_5

subject to the conditions

$$\begin{aligned} a_0 + a_1 x_1 + \cdots + a_d x_d &\geq 1 \quad (p \in S^+) \\ a_0 + a_1 x_1 + \cdots + a_d x_d &\leq 1 \quad (p \in S^-). \end{aligned}$$

The solution of this quadratic programming problem will give us the parameters $a_0, \ldots, a_d$ of the optimal hyperplane that separates S^+ and S^- with the largest margin [3].

The next theorem defines an optimality criterion for maximal margin separations, which is again based on Radon partitions. Although the result can be proven almost immediately, we will give a constructive proof based on Householder transformations. A Householder transformation is a linear transformation that corresponds to a reflection about a hyperplane. We will use it to map the separating plane onto the plane $x_d = 0$.

Theorem 8 *Let $\{S^+, S^-\}$ be a partition of S, and let E denote the elemental subset that maximizes $\epsilon(E)$. Let $f_a(p) = \epsilon(E)$ and $f_a(p) = -\epsilon(E)$ be the hyperplanes as defined in Theorem 1. If the orthogonal projections of E_S^+ and E_S^- on the hyperplane $f_a(p) = 0$ coincide with a Radon partition, then $f_a(p) = 0$ is a separating hyperplane that maximizes the margin.*

Proof Let n_1 denote the normal vector of the hyperplane H, let n_2 denote the normal vector of the plane $x_d = 0$, and define $v = (n_1 + n_2)/(||n_1 + n_2)||)$. Then v is the normal vector of the hyperplane H' that bisects the hyperplanes H and x_d. Let $P = I - 2vv^T$ be a Householder transformation matrix, where I is the identity matrix and v^T denotes the transpose of v. The Householder transformation matrix represents a reflection R about H'. We will denote the image of the transformation of a set S as $R(S)$, and of a point p as $R(p)$. If we apply the transformation R to H, $R(H)$ will coincide with the horizontal plane $x'_d = 0$. As a result, for each point p of $S^+ \cup S^-$, the distance measured between $R(p)$ along the axis orthogonal to the plane $x'_d = 0$ will now coincide with the real shortest distance between $R(p)$ and $R(H)$. Furthermore, the projections of $R(E_S^+)$ and $R(E_S^-)$ on $R(H)$ will coincide with a Radon partition if and only if the orthogonal projection E_S^+ and E_S^- on H

was a Radon partition. Hence, because of Theorem 1, the elemental subset $R(E)$ defines the supporting planes with maximal margin. Since a reflection preserves all distances, it follows that E defines the supporting planes with maximal margin. □

Theorem 8 is restrictive in the sense that it only states that the Chebyshev separation yields a maximal margin separation when the orthogonal projection of the support points defines a Radon partition. This does not exclude the possibility that there can be maximal margin separations that are not generated by elemental subsets. It is perfectly possible that there are less than $d + 1$ support vectors. In this case, it is not difficult to see that the projected support vectors will form a Radon partition in an affine subspace of the separating hyperplane. There can also be more than $d+1$ support vectors. In that case the set of support vectors will contain a subset of $d + 1$ points that projects onto a Radon partition.

7 Time and space complexity

The separation algorithm has the same time complexity as the enclosure algorithm, the only adaptation being that we lower and raise points. Therefore, we only examine the time complexity of the exchange algorithm for enclosures.

Number of radon exchanges To determine the time complexity of the exchange algorithm, we first look at the number of needed Radon exchanges. Suppose the data set consists of n points. In $\mathbb{R}^d$, this means that there are $O(m)$ elemental subsets, where $m = \binom{n}{d+1}$. Since the goal of the enclosure algorithm is to find the elemental subset with the largest height, the time complexity depends on the speed with which the height will increase at each exchange. Let $\epsilon_1, \ldots, \epsilon_m$ represent the heights of all the elemental subsets in increasing order. Let ϵ_i be the height of the current elemental subset E in the exchange algorithm. When we replace one point of E by a new point, the new height will be one of $\epsilon_i, \ldots, \epsilon_m$. We will now assume that it is equally probable that ϵ_j is larger than median of $\epsilon_i, \ldots, \epsilon_m$, than that it is smaller. Under this assumption, the number of elemental subsets with height larger than the current subset is halved at each exchange. The algorithm reaches the maximal height after $O(\log \binom{n}{d+1}) = O((d + 1) \log n)$ replacements.

Although the above assumption has not been proven, simulations with random data sets confirm that for d fixed the average number of replacements is bounded by $k \log_2 n$, where k does not increase faster than d [27]. Figures 7 and 8 illustrate the outcome of these experiments.

The height $\epsilon(E)$ and the enclosing hyperplanes $f_a(E) = \pm\epsilon(E)$ After each replacement of the current elemental subset by a new subset E we have to compute the new height $\epsilon(E)$. The most straightforward approach is to compute $\epsilon(E)$ from (9). Both the cofactors and the determinant can be computed in $O(d^3)$ time by matrix inversion of M_E. Once the cofactors and $\epsilon(E)$ are known, the separating hyperplane of E can be computed in $O(d^3)$ time by solving the linear system

$$x_d^j - (a_0 + \cdots + a_{d-1}x_{d-1}^j) = \epsilon(E), \quad (x_1, \ldots, x_d) \in E$$

for the d unknowns a_i, $i = 0, \ldots, d - 1$. This system has $(d + 1)$ equations, but its rank is equal to d, and we can find a unique solution by discarding one of the equations. In particular, if $C_j \neq 0$, we can discard the j-th equation. Each elemental subset has at least one such cofactor.

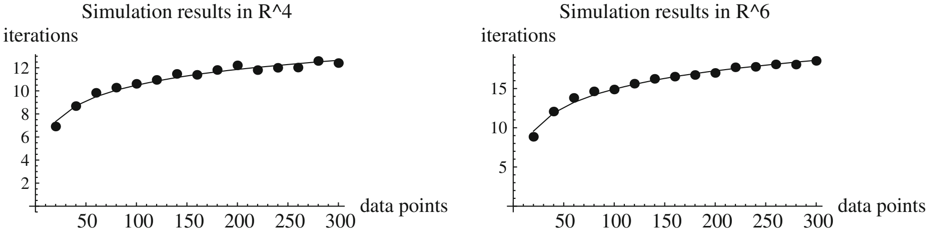

Fig. 7 Time complexity for varying size of the data set. From the time complexity analysis, the expected number of iteration grows as $(d+1)\log n$. The experimental results show the average number of iterations that were needed to find the enclosing elemental subset. The average was taken over 100 experiments for each $n = 10, 20, \ldots$. The smooth curves show plots of the best fit of the form $a + b\log_2(n)$ to the experimental results, where n represents the number of data points. In $\mathbb{R}^4$ a good fit was $a + 1.35\log_2(n)$ (*shown left*), in $\mathbb{R}^6$ a good fit was $a + 2.31\log_2(n)$. In each case, the rate b is smaller than the expected value $d + 1$

A point outside the enclosing hyperplanes For each elemental subset we have to evaluate, in the worst case, all n points to find a new point q outside the enclosing hyperplanes of the current elemental subset, which yields time complexity $O(n)$.

Selecting an appropriate subset in $E \cup \{p\}$ Once a new point q has been found outside the enclosing hyperplanes of E, we must exchange q for a point p. As explained in the proof of Lemma 3, for the selection of p we need the cofactors C_i and an affine dependency relation between $\pi(q)$ and points of $\pi(E)$. The latter can be found by solving a linear system with $d+1$ unknowns (Eq. 14). Hence, p can be found in $O(d^3)$ time.

To summarize, the number of replacements is of the order $O((d+1)\log n)$. Since $\epsilon(E)$, $f_a(E), q, p$ have to be determined at each iteration of the algorithm, the total time complexity is $O((d^3+n)((d+1)\log n)) = O((d^4+nd)\log n)$. This means that in its present form the enclosure algorithm scales reasonably well for the number of points n, but badly for the dimension d. The main asset, however, is that the space complexity is $O(d^2+n)$. When d is small, the algorithm can handle very large data sets. Moreover, we do not need random access to the data points. To find a new exchange point q the data points can be examined one by one in arbitrary order.

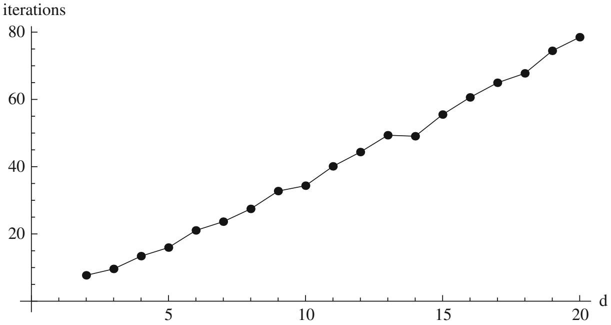

Fig. 8 The average number of iterations when d increases, and n is kept fixed ($n = 1000$). For each d, the average number of iterations was taken over 100 simulations. The graph confirms that the time complexity scales as $O((d+1)\log n)$, for n fixed

8 Conclusion

The primary goal of this paper was to examine the relationship between support vectors and Radon partitions, both for maximal margin separations as for Chebyshev separations and enclosures. To establish these relationships all the cornerstones of combinatorial convexity made their appearance, i.e., Radon's, Caratheodory's and Helly's Theorem.

We proposed an improved version of the Stiefel exchange algorithm, suited to compute enclosures as well as separations in the Chebyshev sense. Although this algorithm is partially based on classical methods, it also includes new elements such as Radon exchanges based on cofactors.

To derive the time complexity of the enclosure/separation algorithms we assumed that at each exchange the increase in height is uniformly distributed between the minimal and maximal increase. With this assumption in mind, the separation algorithm has space complexity $O(d^2+n)$ and time complexity $O((d^4+nd)\log n)$, where n is the number of data points, and d the dimension of the space in which the separation takes place. Although the assumption remains unproven, the simulations confirm that it is plausible.

Especially with respect to the number of dimensions d, the proposed separation algorithm is slow when compared to learning algorithms for Support Vector Machines where time complexities have been reported of the order $O(d^2n+d^3)$, $O(dn^2+n^3)$, or $O(dn)$ [5, 12, 23]. One reason is that the exchange algorithm solves a slightly different problem, and that it does not make use of any of kind of heuristics or active subsampling [3, 17]. Furthermore, the linear time complexity of the iterative SVM algorithms often depends strongly on the parameter settings, such as the required accuracy of the approximation, the sparseness of the feature vectors, and the soft margin parameter [1, 14].

An important question that remains is how we can exploit Radon partitions to understand and improve maximal margin algorithms. For example, in [13] the maximal margin separation problem is first transformed into a minimum norm problem, by taking the Minkowski difference of the sets S^+ and S^-. Each point in the Minkowski corresponds to a vector $p^+ - p^-$, with $p+ \in S^+$ and $p^- \in S^-$. In the minimum norm problem we have to find the point in the Minkowski difference that lies closest to the origin. In [13] this problem is solved iteratively. At each iteration step, a new point is found that lies closer to the origin along a certain line segment. An interesting question, therefore, is to examine which points of the original sets are involved at each iteration step, knowing that the final result has to correspond to a Radon partition. A similar question arises for the GJK algorithm that finds the smallest distance between two polytopes in 3D space [8].

References

1. Bialon, P.: A linear support vector machine solver for a large number of training examples. Control Cybern. **38**(1), 281–300 (2009)
2. Boser, B.E., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Haussler, D. (ed.) COLT, pp. 144–152. ACM (1992)
3. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 121–167 (1998)
4. Cadzow, J.A.: Minimum l_1, l_2, and l_∞ norm approximate solutions to an overdetermined system of linear equations. Digital Signal Process. **12**(4), 524–560 (2002)
5. Chapelle, O.: Training a support vector machine in the primal. Neural Comput. **19**(5), 1155–1178 (2007)
6. Cheney, E.W.: Introduction to Approximation Theory. McGraw-Hill (1966)
7. Ewald, G.: Combinatorial Convexity and Algebraic Geometry. Springer (1996)

8. Gilbert, E.G., Johnson, D.W., Keerthi, S.S.: A fast procedure for computing the distance between complex objects in three-dimensional space. IEEE J Robot Autom **4**, 193–203 (1988)
9. Goodman, J., Pollack, R.: Hadwiger's transversal theorem in higher dimensions. J. Amer. Math. Soc. **1**, 301–309 (1988)
10. Hawkins, D.M., Bradu, D., Kass, G.: Location of several outliers in multiple regression data using elemental sets. Technometrics **26**, 197–208 (1984)
11. Joachims, T.: Advances in Kernel Methods - Support Vector Learning, chap. Making Large-Scale SVM Learning Practical, pp. 169–184. MIT Press (1999)
12. Joachims, T.: Training linear svms in linear time. In: Eliassi-Rad, T., Ungar, L.H., Craven, M., Gunopulos D. (eds.) KDD, pp. 217–226. ACM (2006)
13. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: A fast iterative nearest point algorithm for support vector machine classifier design. IEEE Trans Neural Netw **11**, 124–136 (2000)
14. Keerthi, S.S., Chapelle, O., DeCoste, D.: Building support vector machines with reduced classifier complexity. J. Mach. Learn. Res. **7**, 1493–1515 (2006)
15. Mehlhorn, K., Osbild, R., Sagraloff, M.: Reliable and efficient computational geometry via controlled perturbation. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP (1), Lecture Notes in Computer Science, vol. 4051, pp. 299–310. Springer (2006)
16. Mitchell, B.F., Amd, V.N., Malozemov, V.F.D.: Finding the point of a polyhedron closest to the origin. SIAM J. Control **12**, 19–26 (1974)
17. Musicant, D.R., Feinberg, A.: Active set support vector regression. IEEE Trans Neural Netw **15**(2), 268–275 (2004)
18. Osborne, M.R., Watson, G.A.: On the best linear Chebyshev approximation. Comput. J. **10**, 172–177 (1967)
19. Osuna, E., Freund, R., Girosi, F.: An improved training algorithm for support vector machines. In: Principe, J., Abd L.G., Morgan, N., Wilson, E. (eds.) Proceedings of the 1997 IEEE Workshop on Neural Networks for Signal Processing, pp. 276–285 (1997)
20. Platt, J.: Advaces in Kernel Methods: Support Vector Machines, chap. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. MIT Press (1998)
21. Rousseeuw, P., Leroy, A.: Robust Regression and Outlier Detection. Wiley, New York (1987)
22. Stiefel, E.: über diskrete und lineare tschebyscheff-approximationen. Numerische Mathematik **1**, 1–28 (1959)
23. Tsang, I.W., Kwok, J.T., Cheung, P.M.: Core vector machines: Fast SVM training on very large data sets. J. Mach. Learn. Res. **6**, 363–392 (2005)
24. de la Vallée Poussin, C.J.: Sur la methode de l' approximation minimum. Societe Scientifique de Bruxellles. Annales Memoires **35**, 1–16 (1911)
25. Veelaert, P.: Constructive fitting and extraction of geometric primitives. Graph. Models Image Process. **59**(4), 233–251 (1997)
26. Veelaert, P.: Digital Geometry Algorithms, Lecture Notes in Computational Vision and Biomechanics, vol. 2, chap. Separability and Tight Enclosure of Point Sets, pp. 215–243. Springer (2012)
27. Veelaert, P.: Fast combinatorial algorithm for tightly separating hyperplanes. In: Barneva, R.P., Brimkov, V.E., Aggarwal, J.K. (eds.) IWCIA, Lecture Notes in Computer Science, vol. 7655, pp. 31–44. Springer (2012)
28. Watson, G.A.: Approximation in normed linear spaces. J. Comput. Appl. Math. **121**, 1–36 (2000)
29. Ziegler, G.M.: Lectures on Polytopes. Graduate Texts in Mathematics, vol. 152. Springer-Verlag, New York (1995)