Manuscript submitted to AIMS Journal

doi:10.3934/xx.xxxxxxx

TVERBERG'S THEOREM AND MULTI-CLASS SUPPORT VECTOR MACHINES

Pablo Soberón[✉*1]

[1]The Graduate Center, City University of New York, USA

[1]Baruch College, City University of New York, USA

(Communicated by Handling Editor)

Abstract. We show how, using linear-algebraic tools developed to prove Tverberg's theorem in combinatorial geometry, we can design new models of multi-class support vector machines (SVMs). These supervised learning protocols require fewer conditions to classify sets of points, and can be computed using existing binary SVM algorithms in higher-dimensional spaces, including soft-margin SVM algorithms. We describe how the theoretical guarantees of standard support vector machines transfer to these new classes of multi-class support vector machines. We give a new simple proof of a geometric characterization of support vectors for largest margin SVMs by Veelaert.

1. **Introduction.** Support vector machines (SVMs) are a supervised learning model for data classification with a wide range of applications [4, 14, 21]. The underlying geometric problem is, given two finite sets A, B of points in $\mathbb{R}^d$, to find a hyperplane separating A and B. A key example are largest margin SVMs, in which the separating hyperplane maximizes the minimum distance to each set. We assume that $\operatorname{conv} A \cap \operatorname{conv} B = \emptyset$ for such a hyperplane to exist. If $\operatorname{conv} A \cap \operatorname{conv} B \neq \emptyset$, minimizing the number of misclassified points by a hyperplane is NP-hard, but one can use adaptations such as soft-margin SVM.

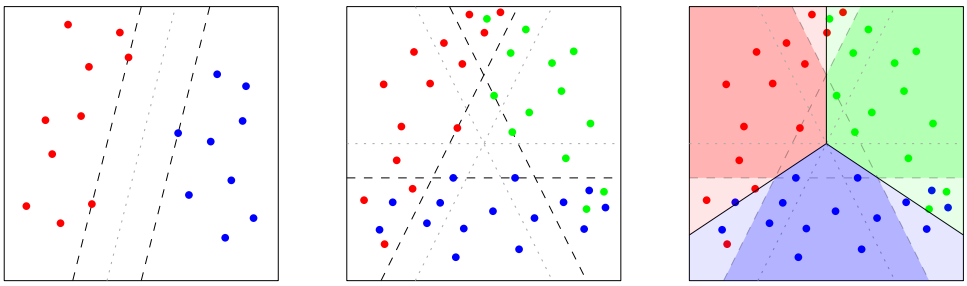
A variation of this model, multi-class support vector machines, arises when we want to classify more than two sets of points. If we want to classify k classes $A_1, \dots, A_k$, the most common approaches are one-versus-all (1vA) and all-versus-all (AvA) models, both of which break the classification problem into many binary support vector machines. In the first, we have to solve for k support vector machines, each separating a single class from the union of the other $k-1$, In the second, we solve for $\binom{k}{2}$ support vector machines separating each pair of classes. Some optimization methods aggregate several SVMs into a single optimization problem in a higher-dimensional space, which can then be adjusted to be easier to solve [11, 8]. Multiple other models for multi-class SVMs have been proposed [10, 15].

Many combinatorial properties of SVMs are related to classic results in discrete geometry, such as Radon's theorem [23, 1]. Radon's theorem states that *given $d+2$ points in $\mathbb{R}^d$, there exists a partition of them into two sets whose convex*

2020 *Mathematics Subject Classification.* Primary: 52A35, 52C35; Secondary: 62R07, 62R40.

Key words and phrases. Support vector machine, Tverberg's theorem, Multiclass vector machine, Sarkaria's transformation, Randomized linear programming.

The author is supported by NSF CAREER award no. 2237324, NSF award no. 2054419 and a PSC-CUNY Trad B award.



(1) (2) (3)

FIGURE 1. (1) An example of an SVM, we emphasize the support hyperplanes parallel to the generated hyperplane for each class. (2) An example of an multi-class SVM under the proposed model. Notice that it is not possible to separate any two classes of points with a hyperplane. (3) The half-spaces in part (2) can be used to classify space using convex regions. The model can distinguish regions where it is ambiguous.

hulls intersect [16]. A well-known generalization of Radon's theorem is Tverberg's theorem, in which we seek to split a set of points into several subsets whose convex hulls intersect. Tverberg proved that *given* $(k-1)(d+1)+1$ *points in* $\mathbb{R}^d$, *there exists a partition of them into* k *sets whose convex hulls intersect* [22]. The case $k = 2$ is Radon's theorem. There is active research around Tverberg's theorem, as it has led to important developments in discrete geometry and topological combinatorics [5, 3, 7].

A far-reaching tool to prove variations of Tverberg's theorem is a linear-algebraic technique devised by Sarkaria [18] and simplified by Bárány and Onn [6]. In addition to leading to one of the simplest known proofs of Tverberg's theorem, this technique is highly malleable and can be used to prove a multitude of variations of Tverberg's theorem.

The goal of this manuscript is to show a link between Tverberg's theorem and multi-class SVMs via the linear algebra techniques mentioned above. The existence of a connection between multi-class SVMs and Tverberg's theorem was conjectured by Adams et al. [1], when they linked Radon's theorem to binary SVMs. To have a multi-class SVM that does not missclassify any points, the (1vA) model requires each class to be separable from the union of the other $k-1$, and the (AvA) model requires any two classes to be separable. We propose a new type of multi-class SVM which uses a weaker condition. Applying our model for $k = 2$ leads to classic SVMs. Of course, since we do not ask any two A_i, A_j to be separable, potential miss-classifications are unavoidable. We only require $\bigcap_{i=1}^{k} \operatorname{conv}(A_i) = \emptyset$. The output will be a family of k closed half-spaces $H_1, \ldots, H_k$ such that

- For each $i = 1, \ldots, k$ we have have $A_i \subset H_i$ and
- the half-spaces satisfy $\bigcap_{i=1}^{k} H_i = \emptyset$.

We describe how the half-spaces can be used to split $\mathbb{R}^d$ into k convex regions, each corresponding to an A_i. The model can also be used to distinguish the regions of ambiguity. The subdivision of $\mathbb{R}^d$ is most natural when $k \leq d+1$.

(AvA)	$\binom{k}{2}\tau(n/k, n/k; d)$
(1vA)	$k \cdot \tau(n/k, n-(n/k); d)$
(Simple TSVM)	$\tau(1, n-1; (d+1)(k-1))$
(TSVM)	$O\Big(n \cdot \tau\Big(1, (d+1)(k-1)+1; (d+1)(k-1)+1\Big)\Big)$ (randomized) $\tau((n/k)^k, 1; d(k-1)+1)$ (deterministic)

TABLE 1. Any linear SVM algorithm can be applied to the computation of our multi-class SVM, including soft-margin SVMs. If we denote $\tau(a, b; d)$ the complexity of computing an SVM with $a+b$ data points in $\mathbb{R}^d$ (one class with a points and one with b points), then the computational complexity in terms of n of our results can be described as listed above. We assume our original set of points has k classes with n/k points each. We include the complexity of a naive approach to (AvA) and (1vA) for comparison. The randomized algorithm for (TSVM) also has constant factors that depend on the product dk but not n. Statistical guarantees would be the same as those running a linear SVM with the parameters above. (Simple TSVM) and deterministic (TSVM) are equivalent to running a single SVM, while randomized (TSVM) is equivalent to running $O(n)$ binary SVMs.

We describe in Table 1 the complexity of computing these multi-class SVM. We compare directly with the complexity of computing a single SVM, to highlight the influence of the dimension.

Tverberg's theorem is a challenging algorithmic problem [13]. One key difference between the problem addressed in this manuscript and the problem of finding Tverbreg partitions is that when training SVMs the labels are assigned beforehand. Tverberg's theorem has also been applied to multi-class logistic regression [9].

Since the constructions are based on Sarkaria's linear-algebraic technique, we can deduce several combinatorial properties of these multi-class SVMs. The model (simple TSVM) is invariant under orthogonal transformations, but not under translations. The model (TSVM) is invariant under any isometry of $\mathbb{R}^d$. To prove these properties, a closer look at Sarkaria's method is needed, so the arguments presented here may be useful in the classic context of variations of Tverberg's theorem.

We also discuss the existence and properties of support vectors. It is known that for any two separable sets $A, B \subset \mathbb{R}^d$, there exist $A' \subset A, B' \subset B$ such that $|A' \cup B'| \leq d+1$ and such that the largest-margin SVM induced by A, B is the same as the largest-margin SVM induced by A', B'. For (TSVM) and (simple TSVM) a similar property holds. For any k-tuples of sets $A_1, \ldots, A_k$, there is a $(k-1)(d+1)$-subset of $A_1 \cup \ldots \cup A_k$ that induces the same (TSVM). The same holds for (simple TSVM). In either case we call this $(k-1)(d+1)$-subset the *support vectors* of the multi-class SVM.

The manuscript is organized as follows. First, we present in Section 2 a new proof of a characterization of critical points in largest-margin SVMs. In Section3 we describe the linear-algebraic tools needed for our constructions. In Section 4 we describe the models (TSVM) and (simple TSVM), and their main properties. In Section5 we discuss the induced partitions of $\mathbb{R}^d$ and finally in Section 6 we study

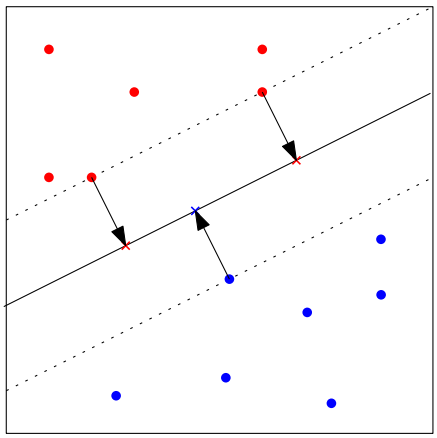


FIGURE 2. An example of a largest-margin SVM with two sets. If we project the support vectors onto the separating hyperplane, the convex hulls of the projections of different sides intersect.

how the model behaves when we apply orthogonal transformations to the sets of points.

2. **Projection of support vectors.** Given two finite sets of points A, B in $\mathbb{R}^d$ that are linearly separable, let H be a separating hyperplane at maximal distance from A and B. We denote by ε this distance, so

$$\varepsilon = \operatorname{dist}(H, \operatorname{conv} A) = \operatorname{dist}(H, \operatorname{conv} B)$$

We say that a point in $a \in A$ is a support vector if $\operatorname{dist}(a, H) = \varepsilon$, and similarly for points in B. We assign labels to the points so that the points of A assigned *positive* and the points of B are assigned *negative*.

One interesting property about the projections of the support vectors is that the convex hulls of the projections onto the separating hyperplane of each side intersect, as in Figure 2. This was proven independently by Veelaert and by Adams, Carr, and Farnell [23, 1]. One of the proofs involves the Karush–Kuhn–Tucker theorem and the other Householder transformations. We present an elementary proof.

Theorem 2.1. *Given a separable set of points in $\mathbb{R}^d$ with two labels, the convex hulls of the projections of the negative and positive support vectors onto the induced largest margin SVM intersect.*

Proof. Let L be the set of labeled points. Let H be the separating hyperplane at maximum distance from the labeled sets, and let S_+, S_- be the positive and negative support vectors, respectively. Let $\varepsilon > 0$ be the distance of the support vectors to H. This means that

$$\begin{aligned} \operatorname{dist}(x, H) &= \varepsilon \qquad \text{for } x \in S_+ \cup S_- \\ \operatorname{dist}(x, H) &> \varepsilon \qquad \text{for } x \in L \setminus (S_+ \cup S_-). \end{aligned}$$

Let P_+ be the orthogonal projection of S_+ onto H, and P_- be the orthogonal projections of S_- onto H. We assume that $\operatorname{conv} P_+ \cap \operatorname{conv} P_- = \emptyset$ and look for a contradiction.

Since $\operatorname{conv} P_+ \cap \operatorname{conv} P_- = \emptyset$, there exists a co-dimension one affine subspace H' of H that separates P_+ and P_-. Notice that H' is a co-dimension two affine subspace of $\mathbb{R}^d$.

Let us project all the points into the two-dimensional subspace $(H')^\perp$. We denote this projection by π. In $(H')^\perp$, $\pi(H)$ is a line ℓ and $\pi(H')$ is a point p. Let ℓ_2 be the orthogonal line to ℓ through p.

The lines ℓ and ℓ_2 split $(H')^\perp$ into four quadrants. Since $\operatorname{conv} P_+ \cap \operatorname{conv} P_- = \emptyset$, the points of $\pi(S_+)$ and those of $\pi(S_-)$ are separated by ℓ_2. They are also separated by ℓ by construction, so $\pi(S_+)$ and $\pi(S_-)$ are in opposite quadrants.

This means that we can rotate ℓ slightly around p so that its distance to each point in $\pi(S_+)$ and $\pi(S_-)$ increases. If the angle of rotation is small enough, the distance of $\pi^{-1}(\ell)$ to the rest of the points in L remains strictly larger than ε. This contradicts H being the largest-margin SVM. □

3. **Linear-algebraic tools.** In this section, we introduce Sarkaria's construction to tackle Tverberg-type problems. Suppose we are given k sets $A_1, \ldots, A_k$ in $\mathbb{R}^d$. We introduce $v_1, \ldots, v_k$, which are the vertices of a regular simplex in $\mathbb{R}^{k-1}$ centered at the origin. We further assume that each v_i is a unit vector. A crucial property of this k-tuple is that its linear dependences are precisely the linear combinations in which all coefficients are equal. For each i we associate v_i to A_i. Given $x \in A_i$ we first append a coordinate 1 to make it into a vector in $\mathbb{R}^{d+1}$, $\bar{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}$. Then, we take the tensor product with its corresponding v_i, defining $S(x) = \bar{x} \otimes v_i = \bar{x} v_i^T \in \mathbb{R}^{(d+1)(k-1)}$.

In this manuscript we treat $\mathbb{R}^{(d+1)(k-1)}$ as the set of $(d+1)(k-1) \times 1$ vectors and as the set of $(d+1) \times (k-1)$ matrices interchangeably.

Finally, for $i = 1, \ldots, k$, we define the set

$$Y_i = \{S(x) : x \in A_i\} \subset \mathbb{R}^{(d+1)(k-1)}.$$

The main difference with our approach and the one by Bárány and Onn is that for each point in $A_1 \cup \ldots \cup A_k$ we already know to which class it belongs, so it yields a unique point in the higher-dimensional space. When one wants to prove Tverberg's theorem, we have to assign classes to unlabeled sets of points, so each point in $\mathbb{R}^d$ is represented by a k-tuple in $\mathbb{R}^{(d+1)(k-1)}$.

The main reason why this transformation can be used to study Tverberg-type problems and why we can use it in the context of multi-class SVMs is the following lemma.

Lemma 3.1. *Let k, d be positive integers. Let $A_1, \ldots, A_k$ be finite sets of points in $\mathbb{R}^d$ such that $\bigcap_{i=1}^k \operatorname{conv}(A_i) = \emptyset$. Then, for $Y_1, \ldots, Y_k$ defined as above, $0 \notin \operatorname{conv}\left(\bigcup_{i=1}^k Y_i\right)$.*

Proof. Let $A = \bigcup_{i=1}^k A_i$ and $Y = \bigcup_{i=1}^k Y_i$. We prove the contrapositive. Assume that the origin in $\mathbb{R}^{(d+1)(k-1)}$ is in the convex hull of Y. We want to show that the convex hulls of the sets A_i intersect. Then, for each $x \in A$ there exists a non-negative coefficient $\alpha(x)$ such that $\sum_{x \in A} \alpha(x) = 1$ and

$$0 = \sum_{x \in A} \alpha(x) S(x) = \sum_{x \in A_1} \alpha(x) S(x) + \ldots + \sum_{x \in A_k} \alpha(x) S(x)$$

$$= \sum_{x \in A_1} \alpha(x)(\bar{x} \otimes v_1) + \ldots + \sum_{x \in A_K} \alpha(x)(\bar{x} \otimes v_k)$$
$$= \left(\sum_{x \in A_1} \alpha(x)\bar{x} \right) \otimes v_1 + \ldots + \left(\sum_{x \in A_k} \alpha(x)\bar{x} \right) \otimes v_k.$$

If we look at the linear dependences of $v_1, \ldots, v_k$ in $\mathbb{R}^{k-1}$, we can see that $\beta_1 v_1 + \ldots + \beta_k v_k = 0$ if an only if $\beta_1 = \ldots = \beta_k$. This carries through the tensor product and we have $\sum_{x \in A_1} \alpha(x)\bar{x} = \ldots = \sum_{x \in A_k} \alpha(x)\bar{x}$.

This is an equality in $\mathbb{R}^{d+1}$. If we look at the last coordinate, we have $\sum_{x \in A_1} \alpha(x) = \ldots = \sum_{x \in A_k} \alpha(x)$. Since the total sum of the coefficients was one, each of the sums above must be $1/k$. If we look at the first d coordinates and multiply each equation by k, we have

$$\sum_{x \in A_1} (k\alpha(x))x = \ldots = \sum_{x \in A_k} (k\alpha(x))x.$$

and each of the terms above is a convex combination. This means that the convex hulls of the A_i have non-empty intersection, $\bigcap_{i=1}^{k} \operatorname{conv}(A_i) \neq \emptyset$. □

Therefore, the origin in $\mathbb{R}^{(d+1)(k-1)}$ can be separated from Y by a hyperplane. We can find this hyperplane with any existing algorithm for SVMs, which is the central point of this manuscript. If we have a hyperplane separating a set from the origin in $\mathbb{R}^{(d+1)(k-1)}$, we want to obtain a set of half-spaces as described in the introduction.

In other words, we need to be able to map hyperplanes in $\mathbb{R}^{(d+1)(k-1)}$ into k-tuples of hyperplanes in $\mathbb{R}^d$ explicitly. This has been done recently [17]. We describe the process below.

Let $\Pi : \mathbb{R}^{d+1} \to \mathbb{R}^d$ be the function that erases the last coordinate. For each $y \in \mathbb{R}^{(d+1)(k-1)}$ let us think of y as a $(d+1) \times (k-1)$ matrix. For $i \in [k]$ we define the function

$$f_i : \mathbb{R}^{(d+1)(k-1)} \to \mathbb{R}^d$$
$$y \mapsto \Pi(yv_i)$$

where yv_i is considered as a product of matrices.

Lemma 3.2. *If $x \in A_i$, then $f_i(S(x)) = x$.*

Proof. A simple computation shows that

$$f_i(S(x)) = f_i(\bar{x}v_i^T) = \Pi(\bar{x}v_i^T v_i) = \Pi(\bar{x}) = x.$$

The third equality follows since v_i is a unit vector. □

For each $i \in [k]$, consider the d-dimensional affine subspace $U_i = \{\bar{x} \otimes v_i : x \in \mathbb{R}^d\} \subset \mathbb{R}^{(d+1)(k-1)}$. Given a half-space H in $\mathbb{R}^{(d+1)(k-1)}$, consider the k half-spaces in $\mathbb{R}^d$ defined by $H_i = f_i(U_i \cap H)$ for $i \in [k]$.

Lemma 3.3. *Let H be a closed half-space in $\mathbb{R}^{(k-1)(d+1)}$. If $0 \notin H$ then $\bigcap_{i=1}^{k} H_i = \emptyset$.*

Proof. As before, let's consider $\mathbb{R}^{(d+1)(k-1)}$ as the set of $(d+1) \times (k-1)$ matrices. Each closed half-space H can be defined using a linear functional and a constant. Using the Frobenius product, we can express H using a $(k-1) \times (d+1)$ matrix M and a constant λ such that

$$H = \{S \in \mathbb{R}^{(d+1)(k-1)} : \operatorname{tr}(SM) \geq \lambda\}.$$

Since the origin is not contained in H, we can assume that $\lambda > 0$. Suppose on the contrary that there exists an $x \in \mathbb{R}^d$ so that $x \in \bigcap_{i=1}^k f_i(U_i \cap H)$ and we look for a contradiction. In other words, for $i = 1, \ldots, k$ we have $\bar{x} \otimes v_i \in H$, so $\operatorname{tr}(\bar{x} v_i^T M) \geq \lambda > 0$.

If we write each of the k inequalities as i varies and add them, we have

$$0 < k\lambda \leq \sum_{i=1}^{k} \operatorname{tr}(\bar{x} v_i^T M) = \operatorname{tr}\left(\bar{x}\left(\sum_{i=1}^{k} v_i^T\right) M\right) = 0.$$

The last equality follows as $\sum_{i=1}^k v_i = 0$. This is the contradiction we wanted. □

Now we have all the ingredients to define a multi-class SVM. Another important subspace for our computations is the following $d(k-1)$-dimensional space

$$R = \{y \in \mathbb{R}^{(d+1)(k-1)} : \text{the last row of } y, \text{ as a } (d+1) \times (k-1) \text{ matrix, is zero}\}.$$

This subspace has been used previously to prove some variations of Tverberg's theorem with some coloring conditions added to the set [20]. Some particular translates of R will also be useful. For $i = 1, \ldots, k$ we define $R_i = \{S \in \mathbb{R}^{(d+1)(k-1)} : \text{the last row of } S \text{ is } v_i^T\}$.

Notice that $U_i \subset R_i$ for each $i = 1, \ldots, k$.

Lemma 3.4. *Let $z_1, z_2, \ldots, z_k \in \mathbb{R}^{(d+1)(k-1)}$ such that $z_i \in U_i$ for each $i = 1, \ldots, k$. The only point in $R \cap \operatorname{conv}\{z_1, \ldots, z_k\}$ is the barycenter of the set $\{z_1, \ldots, z_k\}$.*

Proof. Consider each z_i as a $(d+1) \times (k-1)$ matrix. Suppose that $\lambda_1 z_1 + \ldots + \lambda_k z_k$ is a convex combination in R. If we look at the last row of this linear combination we have $\lambda_1 v_1 + \ldots + \lambda_k v_k = 0$. This means that $\lambda_1 = \ldots = \lambda_k$, as we wanted. □

4. **Construction and basic properties of multi-class SVM.** We are now ready to formalize the multiclass SVMs described in the introduction. Given k sets $A_1, \ldots, A_k$ in $\mathbb{R}^d$ whose convex hulls do not all overlap, we seek a family of k half-spaces $H_1, \ldots, H_k$ such that $A_i \subset H_i$ for each $i = 1, \ldots, k$ and so that the half-spaces $H_1, \ldots, H_k$ do not all intersect. For the following definition we need the subspaces $U_i = \{\bar{x} \otimes v_i : x \in \mathbb{R}^d\}$ and their associated functions f_i defined above.

Definition 4.1 (Simple TSVM)**.** Let $A_1, \ldots, A_k$ be finite families of points in $\mathbb{R}^d$ whose convex hulls do not intersect. We define the multi-class support vector machine (Simple TSVM) as a family of k closed half-spaces $H_1, \ldots, H_k$ obtained as follows. First, for each $x \in A_i$ construct the point $S(x) = \bar{x} v_i^T \in \mathbb{R}^{(d+1)(k-1)}$. Let Y be the collection of all points obtained this way. Find H the closed half-space in $\mathbb{R}^{(d+1)(k-1)}$ that contains Y and whose distance from the origin is maximal. For $i = 1, \ldots, k$ the half-space H_i is defined as $H_i = f_i(U_i \cap H)$.

The computation of (simple TSVM) consists of finding the distance from Y to the origin. We can also think of this as finding the largest-margin SVM in $\mathbb{R}^{(d+1)(k-1)}$ that separates the origin from Y, and then doubling the distance to the origin. The discussion in the previous section shows that this multi-class support vector machine satisfies the desired properties. For a **soft-margin** version, it suffices to compute in $\mathbb{R}^{(d+1)(k-1)}$ an SVM with one class equal to Y and the other equal to $\{0\}$. If we denote by $\tau(a, b; d)$ the complexity of an algorithm to compute an SVM with data points in $\mathbb{R}^d$ and two classes of size a and b, then the complexity of computing (simple TSVM) is $\tau(n, 1; (d+1)(k-1))$, where n is the number $|A_1| + \cdots + |A_k|$

of data points. Any other performance metrics we have for an SVM transfer to (simple TSVM) if we do the change of parameter as outlined above.

For the second type of multi-class SVM, we consider the following alternative definition. Recall that in the space of $(d+1)\times(k-1)$ matrices we denoted by R the subspace where the last row is equal to zero.

Definition 4.2 (TSVM)**.** Let $A_1,\dots,A_k$ be finite families of points in $\mathbb{R}^d$ whose convex hulls do not intersect. We define the multi-class support vector machine (TSVM) as a family of k closed half-spaces $H_1,\dots,H_k$ obtained as follows. First, for each $x\in A_i$ construct the point $S(x)=\bar{x}v_i^T\in\mathbb{R}^{(d+1)(k-1)}$. Let Y be the collection of all points obtained this way and consider $P=R\cap\operatorname{conv}(Y)$. Compute p the closest point of P to the origin, and H the closed half-space in $\mathbb{R}^{(d+1)(k-1)}$ that contains Y, whose boundary hyperplane contains p, and whose distance from the origin is maximal. The half-spaces H_i are defined as $H_i=f_i(U_i\cap H)$.

The definition above assumes that P is not empty, which is a consequence of the proof of Theorem 4.4 below. Even though this definition is more involved it has two big advantages. First, it is stable under translations of the sets of points. Second, in the case $k=2$ it is precisely a largest-margin SVM. We prove these properties in the next section. Just like SVM have critical points, any (TSVM) is fixed by a small set of points.

Theorem 4.3. *Let $A_1,\dots,A_k$ be k finite sets in $\mathbb{R}^d$ such that $\bigcap_{i=1}^k\operatorname{conv}A_i=\emptyset$. We can find subsets $A_1'\subset A_1,\dots,A_k'\subset A_k$ such that $A_1',\dots,A_k'$ induces the same (simple TSVM) as $A_1,\dots,A_k$ and such that $|A_1'|+\ldots+|A_k'|\leq(d+1)(k-1)$*

Proof. We follow the construction in Definition 4.1. Since $0\notin\operatorname{conv}(Y)$ the closest point to the origin in $\operatorname{conv}(Y)$ must be in a face of the polytope $\operatorname{conv}(Y)$. This face K can have dimension at most $(d+1)(k-1)-1$. Recall p is the closest point of P to the origin. By Carathéodory's theorem, we can choose a set of at most $(d+1)(k-1)$ points in $Y\cap K$ whose convex hull contains p. The subsets of $A_1,\dots,A_k$ that induced this subset in $Y\cap K$ satisfy the condition we wanted. □

Theorem 4.4. *Let $A_1,\dots,A_k$ be k finite sets in $\mathbb{R}^d$ such that $\bigcap_{i=1}^k\operatorname{conv}A_i=\emptyset$. We can find subsets $A_1'\subset A_1,\dots,A_k'\subset A_k$ such that $A_1',\dots,A_k'$ induces the same (TSVM) as $A_1,\dots,A_k$ and such that $|A_1'|+\ldots+|A_k'|\leq(d+1)(k-1)$*

Proof. The proof is similar to the previous theorem. If we look for the minimal face of $\operatorname{conv}Y$ containing p, it has dimension at most $(k-1)(d+1)-1$, so the same application of Carathéodory's theorem yields the result. The only additional detail to check is that $R\cap\operatorname{conv}(Y)\neq\emptyset$. This holds because for every choice $x_1\in A_1,\dots,x_k\in A_k$, the baryceneter of the point $S(x_1),\dots,S(x_k)$ is in $\operatorname{conv}(Y)\cap R$. □

We denote the subsets obtained by Theorem 4.3 and Theorem 4.4 as the *support vectors* of a (simple TSVM) or (TSVM), respectively.

As mentioned above, to compute (Simple TSVM) we need to compute an SVM in a $(k-1)(d+1)$-dimensional space with $|A_1|+\ldots+|A_k|+1$ points. A direct approach to compute (TSVM) would be to first find the vertices of $\operatorname{conv}(Y)\cap R$ and solve the induced SVM. We know R is a linear subspace of co-dimension $k-1$, so the vertices of $\operatorname{conv}(Y)\cap R$ should be the intersection of the $(k-1)$-skeleton of $\operatorname{conv}(Y)$ with R. Due to Lemma 3.4, this is a subset of the barycenters of k-tuples with one element in each Y_i. Therefore, we can compute these barycenters and then

compute an SVM in R. This leads us to solve an SVM in a $(k-1)d$-dimensional space with $|A_1| \cdot \ldots \cdot |A_k| + 1$ points.

Theorem 4.4 shows that computing a TSVM can be treated as a linear programming type problem, as in the framework of Sharir and Welzl [19]. This is a randomized approach to problems which are combinatorially similar to linear programming problems, so that they can be solved in expected linear time in the input, which is a signficant reduction over brute-force approaches. This means that for fixed k, d we can compute (TSVM) with a randomized algorithm in expected time linear in $|A_1|+\ldots+|A_k|$. We describe the process in Algorithm 1, before translating back to $\mathbb{R}^d$.

The key idea to compute this is to order the points randomly. At any point, we have computed (TSVM) for the first $t-1$ points and we kept track of the support vector of this TSVM. When including the t-th point, if we don't need to adjust the current halfspace H generated by (TSVM), we keep going. Otherwise, we adjust our guess for the support vectors and run the algorithm again for the first t points. The computations of Sharir and Welzl bound the expected number of times we need to rerun this procedure, and end up with an expected running time linear on the input. For deeper explanations, we recommend references on linear-programming type algorithms and violator spaces [12, 2].

Algorithm 1 Computing TSVM

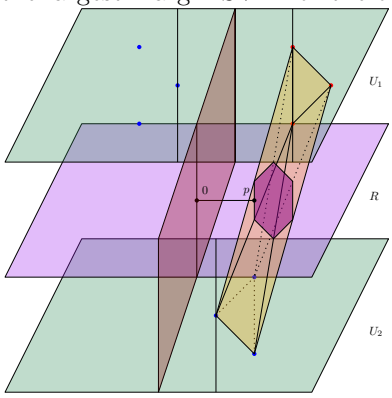
procedure TSVM(Family Y, Tuple Y')
    Order Y randomly as $p_1, \ldots, p_{|Y|}$ where the first $|Y'|$ elements are Y'. The tuple Y' must have $(k-1)(d+1)$ points, and are the candidates for the support vectors of the TSVM.
    Find the TSVM for Y', denoted H. This is a half-space in $\mathbb{R}^{(d+1)(k-1)}$ that does not contain the origin.
    for each $p_t \in Y$ **do**
        Check $p_t \in H$.
        if $p_t \notin H$ **then**
            Find the TSVM H' for $Y' \cup \{p_t\}$.
            Let Y'' be the $(d+1)(k-1)$-tuple whose TSVM is H'.
            $H = \text{TSVM}(\{p_1, \ldots, p_t\}, Y'')$
    return H

The model (TSVM) generalizes largest-margin SVMs when $k = 2$. This is the main motivation to use the subspace R in the computation. Let us prove that this is indeed the case.

Theorem 4.5. *For $k = 2$, the multiclass SVM (TSVM) gives the two support hyperplanes of the largest-margin SVM of A_1 and A_2.*

Proof. Notice that $k = 2$ is the only case when $U_i = R_i$ for all values of i. Additionally, each U_i is a translate of R. In this case we also have $v_1 = 1, v_2 = -1$ in $\mathbb{R}^1$. Therefore $R_1 = \{\bar{x} \in \mathbb{R}^{d+1} : x \in \mathbb{R}^d\}$ and $R_2 = \{-\bar{x} \in \mathbb{R}^{d+1} : x \in \mathbb{R}^d\}$. Let p be the closest point of $\text{conv}(Y) \cap R$ to the origin, and H be the affine hyperplane in $\mathbb{R}^{d+1}$ through p from Definition 4.2. Let H' be the hyperplane through the origin of $\mathbb{R}^{d+1}$, parallel to H, and let $H'' = H' \cap U_1$. Notice that $\|p\|$ is the distance between H and H'. Since a translate of p lies in U_1 (consider p as a directional vector, not just a point), $H'' + p$ contains the support vectors of A_1 in U_1. The same holds

FIGURE 3. This figure shows the process to find (TSVM) for two sets of points. First we embed the sets in U_1, then we reflect A_2 across the origin to obtain their representatives in U_2. We take the convex hull of Y_1 and Y_2 and intersect it with R, which in the figure gives us a hexagon. We take the closest point p to the origin in $\text{conv}(Y) \cap R$ and construct a hyperplane parallel to the facet containing p of $\text{conv}(Y)$ through the origin. This hyperplane intersects U_1 in the largest-margin SVM for the original sets.



for $(H' \cap U_2) + p$ for the support vectors in A_2, so $H'' - p$ contains the support vectors of A_2 in U_1. This means that the (TSVM) induced by A_1, A_2 is an SVM at common distance $\|p\|$ from each side.

Similarly, given a separating hyperplane $\tilde{H}$ for A_1, A_2 at distance ε from each set, we can embed $\mathbb{R}^d$ in U_1 and then reflect the embedding of A_2 with respect to the origin in $\mathbb{R}^{d+1}$ so that it lies in U_2. If we extend $\tilde{H}$ through the origin in $\mathbb{R}^{d+1}$, we have a hyperplane through the origin at distance ε from the convex hull of the embedding of A_1 in U_1 and A_2 in U_2. The largest margin SVM must therefore coincide with the one induced by (TSVM). □

An illustration of the ideas behind this proof is shown in Figure 3.

5. **Subdivision of ambient space and potential classification errors.** In each of Definition 4.1 and Definition 4.2 we use a half-space H in $\mathbb{R}^{(d+1)(k-1)}$ that does not contain the origin to generate the corresponding half-spaces $H_1, \dots, H_k$ in $\mathbb{R}^d$.

In each case, we can introduce a half-space H' that is a translate of H and whose boundary contains the origin. Notice that the half-spaces $H'_i = f_i(U_i \cap H')$ for $i = 1, \dots k$ have non-empty intersection but their interiors have an empty intersection. This is a direct consequence of Lemma 3.3 because H' contains the origin and the interior of H' does not.

As an illustration, for $k = 2$ the two half-spaces H'_1, H'_2 from (TSVM) share their boundary, which is precisely the largest-margin SVM for A_1, A_2.

Let $T = \bigcap_{i=1}^k H'_k$ and $\Delta = \mathbb{R}^d \setminus \left(\bigcup_{i=1}^k H_i^\circ\right)$, where X° denotes the interior of X for any $X \subset \mathbb{R}^d$. Now for $i = 1, \dots, k$ we define the convex sets $M_i = \{p + tq : p \in T,\ t \geq 0,\ q \in \Delta \cap H_i\}$.

Intuitively, Δ is the polytopal region not contained in the union of the H_i. The set T is an affine subspace inside Δ. If $k \leq d+1$, the set Δ is constructed by making

a simplex in the orthogonal complement of T and extending it in the directions of T. The set M_i is formed by taking all possible rays that start at T and go in the direction of a point of H_i in the boundary of Δ. The case when Δ is a simplex is perhaps the most illustrative one, since in this case T is a point and we simply take the cones from T towards each of the facets of Δ. This case looks like Figure 1 (3).

As mentioned before, the condition needed to generate (TSVM) or (simple TSVM) is that the convex hulls of the sets A_i do not all overlap. If the convex hulls of fewer of these sets overlap, any model that subdivides $\mathbb{R}^d$ into convex pieces is bound to miss-label some data. We minimize the mislabelings with our constructions.

6. **Equivariance.** In this section we describe how the multi-class SVMs we introduced interact with transformations of the set of points. It is clear that if we apply the same affine transformation to the sets of points $A_1, \ldots, A_r$ and the half-spaces $H_1, \ldots, H_r$ the containments are preserved, but we are interested to see if the algorithms to obtain $H_1, \ldots, H_r$ behave as expected with these transformations.

Theorem 6.1. *Let M be an orthogonal linear transformation of $\mathbb{R}^d$. Let $H_1, \ldots, H_k$ be the (simple TSVM) induced by $A_1, \ldots, A_k$. Then $(MH_1, \ldots, MH_k)$ is the (simple TSVM) induced by $MA_1, \ldots, MA_k$.*

Proof. First notice that M can be extended to $\mathbb{R}^{d+1}$ by acting on the first d coordinates and leaving the last coordinate fixed. This is also an orthogonal transformation. We denote this transformation by M_2, so $\overline{Mx} = M_2\bar{x}$. Finally, we denote by M_3 the transformation on $\mathbb{R}^{(d+1)(k-1)}$ that multiplies every column of a $(d+1) \times (k-1)$ matrix by M_2, so $y \mapsto M_2 y$ as a product of matrices.

This last transformation is also orthogonal. To see this, we first show that it preserves the dot product between vectors in U_i and U_j for any (possibly equal) i and j. We use a known factorization for the dot product of tensor products, as shown below.

$$\begin{aligned}\langle \bar{x} \otimes v_i, \bar{y} \otimes v_j \rangle &= \langle \bar{x}, \bar{y} \rangle \langle v_i, v_j \rangle = \langle M_2\bar{x}, M_2\bar{y} \rangle \langle v_i, v_j \rangle = \\ &= \langle (M_2\bar{x}) \otimes v_i, (M_2\bar{y}) \otimes v_j \rangle = \langle M_3(\bar{x} \otimes v_i), M_3(\bar{y} \otimes v_j) \rangle\end{aligned}$$

Consider the union of an affine basis for each of $U_1, \ldots, U_{k-1}$. This set of $(d+1)(k-1)$ vectors forms a basis of $\mathbb{R}^{(d+1)(k-1)}$, and M_3 preserves the dot product between any two of these vectors. Therefore M_3 preserves the dot product in $\mathbb{R}^{(d+1)(k-1)}$ and is therefore orthogonal.

If we consider the set Y (as in the definition of (simple TSVM) generated by $A_1, \ldots, A_k$, then $M_3 Y$ will be the set generated by $MA_1, \ldots, MA_k$. Since M_3 is orthogonal, it will preserve the distance from Y, it will send the closest point to the origin on Y to the closest point on the origin to $M_3 Y$. The transformation M_3 will also map the half-space H containing Y furthest from the origin to the half-space containing $M_3 Y$ furthest from the origin, proving our claim.

□

Theorem 6.2. *Let M be an orthogonal linear transformation of $\mathbb{R}^d$ and $H_1, \ldots, H_k$ be the (TSVM) induced by $A_1, \ldots, A_k$. Then $(MH_1, \ldots, MH_k)$ is the (TSVM) induced by $MA_1, \ldots, MA_k$.*

Proof. We follow the ideas used in the proof of Theorem 6.1. We notice that M_3 fixes R. Therefore, the restriction of M_3 to R is an orthogonal transformation. This means that for any half-space H in $\mathbb{R}^{(d+1)(k-1)}$, we have $(M_3 H) \cap R = M_3(H \cap R)$.

Again, if H is the half-space in $\mathbb{R}^{(d+1)(k-1)}$ that induces our (TSVM), we have that M_3H is the half-space for the new set of points.

Now, if we consider the (TSVM) induced by $A_1, \ldots, A_k$, we have to find the half-space H in $\mathbb{R}^{(d+1)(k-1)}$ farthest from the origin that contains Y. Clearly, M_3H is the farthest half-space from the origin that contains M_3Y. For $i = 1, \ldots, k$, we also have $f_i(U_i \cap (M_3H)) = Mf_i(U_i \cap H)$. □

Theorem 6.3. *Let q be a vector in $\mathbb{R}^d$. Let X be the set of support vectors of the (TSVM) induced by $A_1, \ldots, A_k$. Then $X + q$ is the set of support vectors of the (TSVM) induced by $A_1 + q, \ldots, A_k + q$.*

Proof. To find the (TSVM) induced by $A_1, \ldots, A_k$ we need to compute $\operatorname{conv}(Y) \cap R$. Notice that this set is invariant under translations of $A = \bigcup_{i=1}^{k} A_i$. This is because for any points $x_1, \ldots, x_k$ the barycenter of $\{\bar{x}_1 \otimes v_1, \ldots, \bar{x}_k \otimes v_k\}$ is the same as the barycenter of $\{\overline{(x_1 + q)} \otimes v_1, \ldots, \overline{(x_k + q)} \otimes v_k\}$. Since $\operatorname{conv}(Y) \cap R$ does not change, the set of support vectors remains the same. □

Acknowledgments. The author thanks Henry Adams for helpful comments, and the two anonymous referees for their detailed corrections and suggestions.

REFERENCES

[1] Henry Adams, Elin Farnell, and Brittany Story. Support vector machines and Radon's theorem. *Found. Data Sci.*, 4(4):467–494, 2022.

[2] N. Amenta, J. De Loera, and P. Soberón. Helly's theorem: New variations and applications. In Heather Harrington, Mohamed Omar, and Matthew Wright, editors, *Algebraic and Geometric Methods in Discrete Mathematics*, pages 55–96. American Mathematical Society, 2017.

[3] Pavle V. M. Blagojević and Günter M. Ziegler. Beyond the Borsuk–Ulam Theorem: The Topological Tverberg Story. volume 34 of *A Journey Through Discrete Mathematics*, pages 273 – 341. Springer, Cham, 2017.

[4] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

[5] Imre Bárány, Pavle V. M. Blagojević, and Günter M. Ziegler. Tverberg's Theorem at 50: Extensions and Counterexamples. *Notices of the American Mathematical Society*, 63:732 – 739, 2016.

[6] Imre Bárány and Shmuel Onn. Colourful linear programming. volume 1084 of *Integer Programming and Combinatorial Optimization*, pages 1 – 15. Springer Berlin Heidelberg, 1996.

[7] Imre Bárány and Pablo Soberón. Tverberg's theorem is 50 years old: A survey. *Bulletin of the American Mathematical Society*, 55(4):459 – 492, 2018.

[8] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

[9] Jesus A De Loera and Thomas Hogan. Stochastic tverberg theorems with applications in multiclass logistic regression, separability, and centerpoints of data. *SIAM Journal on Mathematics of Data Science*, 2(4):1151–1166, 2020.

[10] Kai-Bo Duan and S Sathiya Keerthi. Which is the best multiclass svm method? an empirical study. In *International workshop on multiple classifier systems*, pages 278–285. Springer, 2005.

[11] Voitčch Franc and Václav Hlaváč. Multi-class Support Vector Machine. *Object recognition supported by user interaction for service robots*, 2:236–239, 2002.

[12] Bernd Gärtner, Jiří Matoušek, Leo Rüst, and Petr Škovroň. Violator spaces: Structure and algorithms. *Discrete Applied Mathematics*, 156(11):2124–2141, 2008.

[13] Sariel Har-Peled and Mitchell Jones. Journey to the center of the point set. *ACM Transactions on Algorithms (TALG)*, 17(1):1–21, 2020.

[14] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[15] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.
[16] Johann Radon. Mengen konvexer Körper, die einen gemeinsamen Punkt enthalten. *Mathematische Annalen*, 83(1):113 – 115, 1921.
[17] Sherry Sarkar and Pablo Soberón. Tolerance for colorful Tverberg partitions. *arXiv*, 2020.
[18] Karanbir S. Sarkaria. Tverberg's theorem via number fields. *Israel journal of mathematics*, 79(2):317 – 320, 1992.
[19] Micha Sharir and Emo Welzl. A combinatorial bound for linear programming and related problems. volume 577 of *Annual Symposium on Theoretical Aspects of Computer Science*, pages 567 – 579, 1992.
[20] Pablo Soberón. Equal coefficients and tolerance in coloured tverberg partitions. *Combinatorica*, 35(2):235–252, 2015.
[21] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
[22] Helge Tverberg. A generalization of Radon's theorem. *J. London Math. Soc*, 41(1):123 – 128, 1966.
[23] Peter Veelaert. Combinatorial properties of support vectors of separating hyperplanes. *Annals of Mathematics and Artificial Intelligence*, 75(1-2):89–115, 2015.