SOFT-MARGIN TROPICAL SUPPORT VECTOR MACHINES

SAMUEL BOÏTÉ, THÉO MOLFESSIS
SUPERVISED BY XAVIER ALLAMIGEON, STÉPHANE GAUBERT

Abstract. We develop new methods around soft-margin tropical support vector machines (SVMs) for binary and multi-classification. We explore the use of tropical hyperplanes to partition tropical space, laying the groundwork for our classification approach. We address the challenge of separating overlapping data in tropical space, proposing a method to measure and handle data overlap using non-expansive Shapley operators and a Krasnoselskii-Mann iteration scheme. For separable data, we extend our method to determine hard-margin tropical hyperplanes, ensuring maximal separation. This is further applied to multi-classification scenarios, providing conditions for tropical separability and demonstrating margin optimality. Additionally, the tropical kernel trick is explored as a means to embed data into a higher-dimensional tropical space, transforming decision boundaries into tropical polynomials. This approach is empirically tested on several classic datasets. Finally, we show that tropical hyperplanes emerge as limiting cases of classical hyperplanes on logarithmic paper, making our approach more stable numerically.

1. Introduction

The tropical semifield $\mathbb{R}_{\max}$ is the set of real numbers, completed by $-\infty$ and equipped with the addition $a \oplus b = \max(a, b)$ and the multiplication $a \odot b = a + b$.

Definition 1. A *tropical hyperplane of apex* $a \in \mathbb{R}^d_{max}$ splits $\mathbb{R}^d_{\max}$ depending on where $(x - a)$ reaches its maximum coordinate:

$$H_a := \left\{x \in \mathbb{R}^d_{\max}, \quad (x-a) \text{ reaches its max coordinate at least twice}\right\}.$$

It hence partitions the space between d different *tropical sectors*. A *tropical halfspace* of configuration $I \subset [d]$ is the union of tropical sectors specified in I.

Let $n \in [d]$. We define the *tropical parametrized hyperplane* of configuration $\sigma = \{I^1, \dots, I^n\}$, where $\bigsqcup_{k\in[n]} I^k = [d]$ as:

$$H^\sigma_a := \left\{x \in \mathbb{R}^d_{\max}, \quad \exists k \neq \ell, \quad (x-a) \text{ reaches its max coordinate in } I^k \text{ and } I^\ell\right\}.$$

When $n = 2$, we talk about *tropical signed hyperplanes.* Generally, a tropical parametrized hyperplane is the union of tropical signed hyperplanes defined by all pairs of sectors.

Definition 2. We define the *tropical span* of a finite set of points $X = (x_1, \dots, x_p) \in \mathbb{R}^{d\times p}_{\max}$ as the set of tropical linear combinations of these points.

$$\operatorname{Span}(X) := \left\{\lambda X := \max_{i\in[p]}(x_i + \lambda_i), \quad \lambda \in \mathbb{R}^p\right\}.$$

Given the tropical convexity of these, these spans are often called *tropically convex hulls*, meaning the smallest tropically convex sets containing them.

The tropical classification problem. We want to separate $n \in [d]$ classes of d-dimensional data points $X^1, \dots, X^n$ using a tropical parametrized hyperplane of configuration σ. In the binary setting, we note the classes $X^{\pm}$ consisting of the points with positive (resp. negative) labels. Separating them tropically amounts to separating their spans, noted $V^1, \dots, V^n$ or $V^{\pm}$ in the binary setting.

Definition 3. For all $u \in \mathbb{R}^d_{\max}$, we define the *tropical norm* of u as the largest difference between two of its coordinates:

$$\|u\| := \max u - \min u.$$

The *tropical distance* between u and v in $\mathbb{R}^d_{\max}$ is then deduced as:

$$d(u, v) := \|u - v\|.$$

Definition 4. H^{σ}_{a} is said to *separate point clouds* $(X^k)_{k \in [n]}$ *with a margin of at least* $\nu \geq 0$ when for all $x^k \in X^k$, $k \in [n]$:

(1) x^k is on the correct side of the hyperplane H^{σ}_{u}, that is:

$$\arg\max_{i \in [d]} x^k_i \in I^k.$$

(2) Distance from H^{σ}_{a} to x^k is at least ν:

$$d(H^{\sigma}_{a}, x^k) = \max(x^k - a) - \max_{[d] \setminus I^k}(x^k - a) \geq \nu.$$

When ν is maximal in the previous definition, we say that H^{σ}_{a} *separates with a margin of* ν. When ν is zero, we say that H^{σ}_{u} *separates* $(X^k)_{k \in [n]}$.

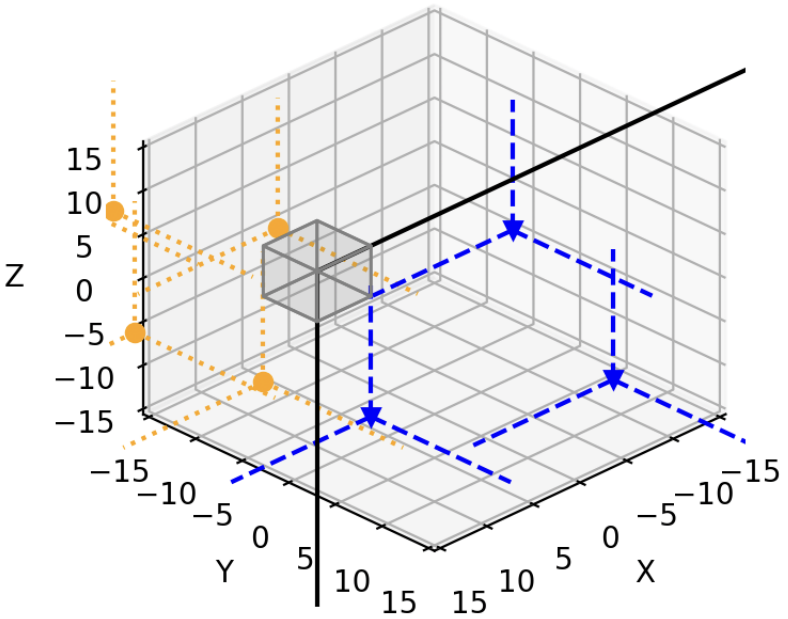


Figure 1.1. Tropical Binary Classification Example

2. Tropical Binary Classification

For the moment, let's confine ourselves to binary classification.

2.1. **Tropical projections and Shapley operators.** In this section, we see that Shapley operators are the appropriate framework for describing and tropically separating finite point clouds, in particular but not exclusively.

Definition 5. A *Shapley operator* on $\mathbb{R}_{\max}$ is a map $T : \mathbb{R}^d_{\max} \longrightarrow \mathbb{R}^p_{\max}$ such that T is non-decreasing and that for all $x \in \mathbb{R}^n_{\max}$ and $\lambda \in \mathbb{R}_{\max}$, $T(\lambda + x) = \lambda + T(x)$. T is said to be *non-expansive* if it is 1-Lipschitz.

Definition 6. A Shapley operator T is said to be *diagonal free* when $T_i(x)$ is independent of x_i for all $i \in [n]$. That is, when for all $i \in [n]$, and for all $x, y \in \mathbb{R}_{\max}$ such that $x_j = y_j$ for all $j \neq i$, we have $T_i(x) = T_i(y)$.

Definition 7. Let V a tropically convex, compact and nonempty subset of $\mathbb{R}^d_{\max}$. We define the *projection* of x on V as:

$$P_V(x) := \max\{y \in V, \quad y \leq x\}.$$

When $V = \mathrm{Span}(X)$ with X in $\mathbb{R}^{d\times p}_{\max}$, we will write the projection as P_X.

Proposition 8. *Let $X \in \mathbb{R}^{d\times p}_{\max}$. P_X is nondecreasing and for $x \in \mathbb{R}^d_{max}$, $P_X(x) \leq x$. Moreover, we know from Maclagan et al.* [6] *that:*

$$\forall i \in [p], \quad P_X(x) = \max_{j\in[p]} \left\{ X_{ij} + \min_{k\in[d]}(-X_{kj} + x_k) \right\}.$$

Remark 9. Tropical projections, and Shapley operators in general, are closely tied to mean payoff games: for instance, let's define T as

$$\forall i \in [n], \quad T_i(x) = \max_{j\in[p]} \left\{ A_{ji} + \min_{k\in[n]}(-B_{jk} + x_k) \right\}.$$

Here, $T_i(0)$ is the payoff, recieved by player MIN, of one round of a game with perfect information, where player MAX starts from their state i, transitions to their opponent MIN's state j by receiving A_{ji}, who in turn chooses a MAX state k by recieving B_{jk}.

Therefore, $(T^m(0))_i$ is the *value* of such a game after m consecutive rounds, having started in the MIN state i. The *escape rate* of the game defined by T is defined by:

$$\chi_T := \lim_{m\to+\infty} \frac{T^m(0)}{m}.$$

It is well known that this limit does exist and coincides with the *mean payoff* of the game [4]. Under certain conditions, χ_T is the unique eigenvalue of the Shapley operator T.

Definition 10. Let T be a non-expansive Shapley operator. We define

$$\mathcal{S}(T) = \{x \in \mathbb{R}^d, \quad x \leq T(x)\}.$$

Proposition 11. *Let P_V be the tropical projection on V. Then:*

$$\mathcal{S}(P_V) = V.$$

Proof. As $P_V(x) \leq x$ for any x, $x \leq P_V(x)$ (i.e $x \in \mathcal{S}(T)$) is equivalent to $x = P(x)$ (i.e $x \in V$). □

Remark 12. Operator P can be slightly tweaked to make it *diagonal-free* (DF), a good property that will be of interest to us in the future. In the modified game, the MAX player is prevented from replying to his opponent eye-for-an-eye:

$$[P_{\mathrm{DF}}(x)]_i := \max_{j\in[p]} \left\{ X_{ij} + \min_{k\neq i}(-X_{kj} + x_k) \right\}.$$

Beware, this is not a projector anymore!

Proposition 13. *P_{DF} is a diagonal-free Shapley map that equivalently describes X.*

Proof. P_{DF} is a Shapley operator for the same reasons P is, and its formula directly proves it is diagonal free. We then note that $P \leq P_{\mathrm{DF}}$, therefore $\mathcal{S}(F) \subset \mathcal{S}(F_{\mathrm{DF}})$. Let $x \in \mathcal{S}(P_{\mathrm{DF}})$, and $i \in [n]$. There exists $j \in [p]$ such that

$$x_i \leq -X_{ij} + \min_{k\neq i}(-X_{kj} + x_k).$$

That inequality holds for $k = i$, hence $x_i \leq P_i(x)$ and finally $\mathcal{S}(P) = \mathcal{S}(P_{\mathrm{DF}})$. □

2.2. **Separating finite overlapping data.** Assuming that the data overlap, we want to find a way to transform them slightly to make them separable by a tropical hyperplane. This would give us a measure of their non-separability.

2.2.1. *Measuring data overlap.* We want a simple criteria for computing the size of the intersection between convex hulls.

Lemma 14. *Let $(T^i)_{i\in[n]}$ be non-expansive Shapley operators. We have:*

$$\bigcap_{i\in[n]} \mathcal{S}(T^i) = \mathcal{S}\left(\bigwedge_{i\in[n]} T^i\right),$$

According to [1], when we have non-expansive Shapley operators, the notions of Collatz-Wielandt numbers, spectral radus and game value are identical . We can then state the following theorem:

Theorem 15. *(Allamigeon, Gaubert et al.* [4]*) V contains a Hilbert ball of positive radius if and only if the spectral radius of F, defined as*

$$\rho(T) = \sup\{\mu \in \mathbb{R}, \quad \exists z \in \mathbb{R}^d, \quad T(z) = \mu + z\},$$

is strictly positive. In this case, $\rho(F)$ is the inner radius inrad$(V^+ \cap V^-)$, i.e supremum of the radii of the Hilbert balls contained in it.

We can then apply the following Krasnoselskii-Mann algorithm to compute the eigenpair in pseudo-polynomial time. Given an initial point a^0, we iteratively compute:

$$\begin{cases} z^{k+1} & = \frac{a^k + T(a^k)}{2} \\ a^{k+1} & = z^{k+1} - \max_{i\in[d]} z_i^{k+1} \cdot \mathbf{1}_d \end{cases}$$

The following convergence theorem follows from [5]:

Corollary 16. As T is a non-expansive Shapley operator, the Krasnoselskii-Mann algorithm converges in pseudo-polynomial time.

The eigenpair we search is $\left(a^\infty, 2 \cdot \max_{i\in[d]} z_i^\infty\right)$.

2.2.2. *Separating two classes of data.* We place ourselves back in the binary classification setting. We have described the corresponding Shapley operators corresponding to our classes as diagonal-free maps. We now define a process for separating overlapping data. Assuming $V^+ \cap V^- \neq \emptyset$, let (a, λ) be the eigenpair of $T := T^+ \wedge T^-$ approximated by the previously described iteration algorithm.

Proposition 17. *We project all points of $X^\pm$ located at a distance less than λ from H_a, onto H_a. Then the intersection of new convex hulls $W^\pm$ is of empty interior.*

Proof. Let's denote X the cloud consisting of all points (regardless of sign), and for $x_j \in X$, we note y_j its sign, s_j its sector and d_j the second argmax of $(x_j - a)$. For each point $x_j = X_{\cdot j} \in X$ at distance less than λ of H_a, the transformation consists in setting

$$W_{kj} := \begin{cases} X_{kj} & \text{if } k \neq s_j \\ X_{s_j j} - d(x_j, H_a) & \text{at } s_j \end{cases}$$

so that w_j is projected on the hyperplane. As $T^\pm$ is non-expansive and diagonal-free, let's remark that for $x^\pm \in V^\pm$:

$$x_i^\pm \leq T^\pm(x^\pm)_i = \big(T^\pm(x^\pm) - T^\pm(a)\big)_i + T^\pm(a)_i,$$

hence

$$x_i^\pm \leq \max(x^\pm - a)_{\neq i} + T^\pm(a)_i. \tag{2.1}$$

Let $i \in [d]$. If $x_j \in X$ is not in the i-th sector, then for k different from the sector of x_j, by definition:

$$(w_j - a)_k \leq (x_j - a)_{d_j} \leq (w_j - a)_{s_j},$$

hence

$$W_{ij} - \max_{k \neq i} (W_{kj} - a_k) \leq a_i.$$

Otherwise, x_j is in the i-th sector and:

$$\max_{k \neq i} (W_{kj} - a_k) = X_{d_j j} - a_{d_j},$$

thus

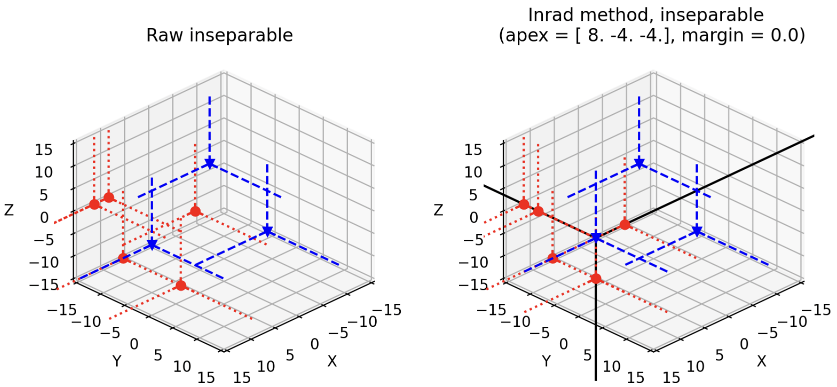
$$W_{ij} - \max_{\neq i} (w_j - a) = (w_j - a)_{s_j} - (x_j - a)_{d_j} + a_{s_j} \geq a_{s_j} = a_i,$$

with equality iff $d(x_j, H_a) \leq \lambda$.

Suppose by symmetry that $T(a)_i = T^+(a)_i = \lambda + a_i$. We also have $T^-(a)_i \geq \lambda + a_i$. Then, using the proof of Theorem 22 in [2], we know that there exists $j^+, j^- \in [p]$ such that $x_{j^+} \in X^+$ and $x_{j^-} \in X^-$ are in sector i, with x_{j^+} being at distance λ from H_a and x_{j^-} at distance greater than λ. Therefore, $W_{ij^+} - \max_{\neq i} \big(w_{j^+} - a\big) = a_i$ and $W_{ij^-} - \max_{\neq i} \big(w_{j^-} - a\big) \geq a_i$. Moreover, for any j such that $x_j \in X^+$ is in sector i, eq. 2.1 gives $d(x_j, H_a) \leq \lambda$.

Let $Q^\pm$ be the *diagonal-free* projections over transformed point clouds, and $Q = Q^+ \wedge Q^-$. We've just shown that $Q(a)_i = Q^+(a)_i = a_i$, and finally $Q(a) = a$. □

Example 18. Here is what the transformation yields with a toy inseparable dataset:



Raw inseparable

Inrad method, inseparable
(apex = [8. -4. -4.], margin = 0.0)

Figure 2.1. Separating convex hulls

Remark 19. In order to strictly separate the sets, we would have to deal with the branching points of null interior.

In pseudo-polynomial time, we are thus able to compute a distance of our point clouds to separability. It gives a bound on the distance under which some points have to be moved to separate hulls.

2.3. Optimal hard-margin in the separable case.

Assuming that data is now separable, we prove that the previous method can be extended to give us separating hyperplanes with maximal margin.

Remark 20. When V^+ and V^- are disjoint, the spectral radius λ of T is strictly negative. From [3], we may see $-\lambda$ as the eigenvalue an operator T^{dual} in dual space, as T is itself a finitely-generated Shapley operator. This would give us a complementary inner radius interpretation, and makes H_a a great candidate in the separable case.

Let (a, λ) the eigenpair of T approximated by previous algorithm, verifying $T(a) = \lambda + a$. Let's define the sectors:

$$I^\pm := \{i \in [d], \quad T^\pm(a)_i > \lambda + a_i\},$$

and the corresponding configuration $\sigma = \{I^\pm\}$.

Proposition 21. *H_a^σ, given the sectors defined above, separates V^+ and V^- with a margin of $-\lambda$. Moreover, this margin is optimal in the case where $T^\pm$ are of the form $T^\pm(x) = P_{V^\pm}(x) = \sup_{v \in V^\pm}(v_i + \min(-v + x))$, which is in particular the case when separating finite point clouds.*

Proof. As $T^\pm$ is non-expansive, let's first remark that for $x^\pm \in V^\pm$:

$$x_i^\pm \leq T^\pm(x^\pm)_i = \left(T^\pm(x^\pm) - T^\pm(a)\right)_i + T^\pm(a)_i,$$

hence

$$x_i^{\pm} \leq \max(x^{\pm} - a) + T^{\pm}(a)_i. \tag{2.2}$$

For instance, let $i \in [d] \setminus I^+$. Then $T^+(a)_i = \lambda + a_i$, so for $x^+ \in V^+$, using equation 2.2:

$$x_i^+ - a_i \leq \max(x^+ - a) + \lambda.$$

In particular, $x_i^+ - a_i < \max(x^+ - a)$ and any element of V^+ can't belong to any of sectors in $[d] \setminus I^+$ with respect to H_a, from which the sectors $I^{\pm}$ are well-defined. Finally,

$$d(H_a^{I^+}, x^+) = \max(x^+ - a) - \max(x^+ - a)_{[d]\setminus I^+} \geq -\lambda,$$

and the margin comes from the fact that this applies to any element of V^+.

Let's finally prove that the margin is maximal in the case where $T^{\pm}$ are of the form $T^{\pm}(x) = P_{V^{\pm}}(x) = \sup_{v \in V^{\pm}} (v_i + \min(-v + x))$. Let $i \in [d] \setminus I^+$. Then, for $\varepsilon > 0$, we can find $v \in V^+$ such that

$$T^+(a)_i - \varepsilon \leq v_i - \max(v - a) \leq T^+(a)_i,$$

giving us

$$\lambda - \varepsilon \leq v_i - a_i - \max(v - a) \leq \lambda.$$

Maximizing over all $i \in [d] \setminus I^+$ yields that v is at most at distance $-\lambda + \varepsilon$ of H_a^I, hence the optimality. □

Example 22. Here is what the algorithm gives with a toy separable dataset:

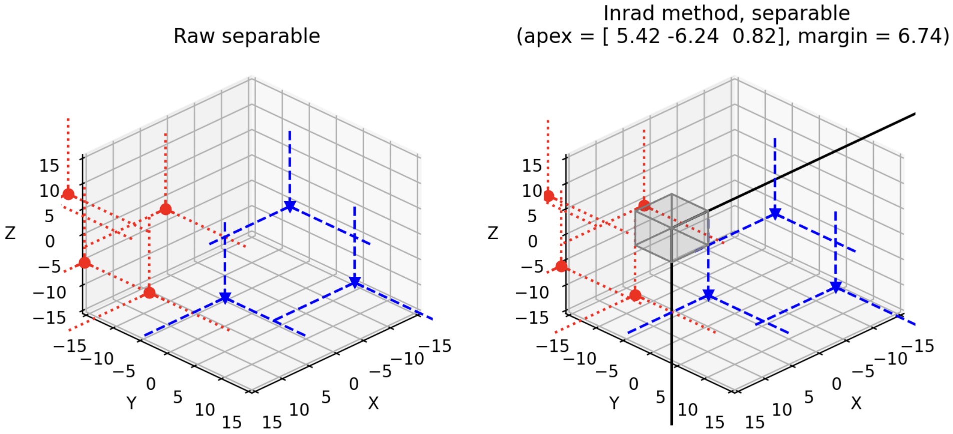


FIGURE 2.2. Optimal-margin separable hyperplane

3. Topics Being Explored

3.1. **Hard-Margin Multi-Classification.** In this section, we consider convex hulls of point clouds of n classes, noted V^k for $k \in [n]$, and described by Shapley operators T^k. We give a sufficient condition for these classes to be tropically separable in their whole, meaning that there exists a tropical signed hyperplane such that each V^k belong to sectors of $I^k \subset [d]$ with $I^k \cap I^l = \emptyset$ for $k \neq l \in [n]$. We then adapt the previous results to this case.

We now consider the Shapley operator :

$$T := \bigvee_{1 \leq k < l \leq n} T^k \wedge T^l$$

We remark that in the case $n = 2$, T is the same as previously defined.

Let (a, λ) the eigenpair of T approximated by the Kranoselskii-Mann algorithm, verifying $T(a) = \lambda + a$. Let's define the sectors:

$$I^k := \{i \in [d], \quad T^k(a)_i > \lambda + a_i\}.$$

Given the definition of T, it is clear that $I^k \cap I^l = \emptyset$ for $k \neq l \in [n]$, and we have the following result :

Proposition 23. *If $\lambda < 0$, the tropical parametrized hyperplane H_a^σ, given the configuration $\sigma = \{I^k\}_{k \in [n]}$, separates V^k and V^l for all $k \neq l \in [n]$ with a margin of $-\lambda$. Moreover, this margin is optimal in the case where all T^k are of the form*

$$T^k(x) = P_{V^k}(x) = \sup_{v \in V^k} \left(v_i + \min(-v + x)\right),$$

which is in particular the case when separating finite point clouds.

Proof. Let $k \in [n]$ and $i \in [d] \setminus I_k$. Using the same reasoning as in the proof of Proposition 19, we obtain that for all $x^k \in V^k$,

$$d(H_a^\sigma, x^k) = \max(x^k - a) - \max(x^k - a)_{[d] \setminus I_k} \geq -\lambda,$$

hence the margin. Let's now prove the optimality in the case where all T^k are of the form $T^k(x) = P_{V^k}(x) = \sup_{v \in V^k} (v_i + \min(-v + x))$. For all $i \in [d]$, there are two distinct classes $k \neq l \in [n]$ such that $T^k(a)_i \wedge T^l(a)_i = \lambda + a_i$. As $I^k \cap I^l = \emptyset$, we can suppose by symmetry that $i \in [d] \setminus I_k$. Then using the same argument as in the proof of optimality in Proposition 19, we know that for all $\varepsilon > 0$ there is a point $v^k \in V^k$ such that $\max(v^k - a) - (v_i^k - a_i) \leq -\lambda + \varepsilon$, which means that

$$d(H_a^\sigma, x^k) = \max(x^k - a) - \max(x^k - a)_{[d] \setminus I_k} \leq -\lambda + \varepsilon.$$

As this holds for every sector $i \in [d]$, we have proven the optimality. □

3.2. **Adding Features: Tropically Polynomial Decision Boundaries.** In the classical setting, the kernel trick consists in mapping the training points in a higher-dimensional space in which we expect the data to become easily linearly separable. In this paragraph, we adapt this idea to integer combinations of features in the tropical setting.

3.2.1. *Tropical kernel trick.* If $\mathcal{A} \subset \mathbb{Z}^d$ is a set of vectors, we define the *veronese embedding* of x as

$$\mathrm{ver}(x) := (\langle x, \alpha\rangle)_{\alpha\in\mathcal{A}} \in \mathbb{R}^{\mathcal{A}},$$

allowing us to map our point space into a larger space, made up of various integer combinations of features. Noting $s \in \mathbb{N}$ a scale parameter, and Δ_d the d-dimensional simplex, we take

$$\mathcal{A}_s := (s\Delta_d) \cap \mathbb{Z}^d.$$

Proposition 24. *Applying the previous method to point clouds ver(C_i) for each class i yields a classifier whose decision boundaries, when seen in the initial vector space, are tropical polynomials. (cite ?)*

Feedforward neural networks with rectified linear units are, modulo trivialities, nothing more than tropical rational maps [7], i.e differences of tropical polynomials. Thus, the framework we have chosen is close to that of dense neural networks, and should yield similar, yet inferior results for classic datasets.

3.2.2. *Testing our approach on classic datasets.* We therefore consider:

The *Iris flower dataset.* It includes measurements such as sepal length, sepal width, petal length, and petal width, across three species of Iris (Iris setosa, Iris virginica, and Iris versicolor). This dataset is commonly used for classification tasks and testing clustering algorithms.

The *Breast Cancer Wisconsin dataset.* It contains features computed from a digitized image of a fine needle aspirate of a breast mass. The features describe characteristics of the cell nuclei present in the image, and the dataset is primarily used for binary classification tasks to distinguish between malignant and benign tumors.

The *Wine quality and type dataset.* It comprises physicochemical tests (like alcohol content, acidity, sugar level, etc.) and sensory information (quality score) for various samples of red and white wines. This dataset is often employed for regression tasks to predict wine quality or for classification tasks to differentiate between wine types.

The *FIFA 2022 cards dataset.* It includes skill ratings of players featured in the game and playing positions. This dataset can be used to predict positions on the field.

These datasets have features that are comparable with each other, which makes them suitable for the tropical framework – although this should work in any case with the kernel trick, since we're describing objects analogous to DNNs.

So as not to bother with branches in the non-separable case, we assign the sectors to the majority population, which in particular settles these borderline cases.

Dataset	Groups	d	p
Iris	Setosa, Virginica, Versicolor	4	150
Cancer	Malign, Benign	31	570
Wine	Red (bad), Red (good), White (bad), White (good)	10	6500
FIFA	Striker, Center back, Center mid, Goal	34	19240

FIGURE 3.1. Datasets description

The sets A_s grow exponentially with the dimension and polynomially with s, and encode the complexity of the mimicked neural network. s will therefore be a relevant hyperparameter of overfitting or underfitting.

Dataset	$\mathcal{A}_1$		$\mathcal{A}_2$		$\mathcal{A}_3$		$\mathcal{A}_4$	
	1v1	*1vR*	*1v1*	*1vR*	*1v1*	*1vR*	*1v1*	*1vR*
Iris	70%	87%	93%	93%	90%	87%	87%	87%
Cancer	90%	89%	90%	91%	88%	86%	$\cdots$	$\cdots$
Wine	67%	56%	72%	74%	78%	85%	87%	92%
FIFA	78%	69%	79%	77%	$\cdots$	$\cdots$	$\cdots$	$\cdots$

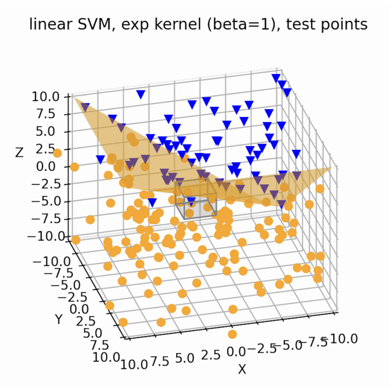
FIGURE 3.2. Accuracies for one-vs-one and one-vs-rest classifiers

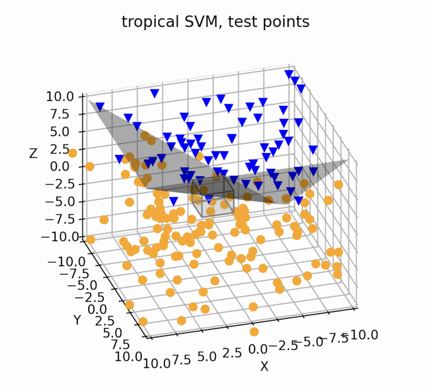
In the table, the dots indicate overfitting, or that the dimension has become too large for the calculations to succeed. The addition of complexity by this kernel method seems particularly effective for the wine dataset.

We note that irrelevant features are driven out of the decision process by a correspondingly large coordinate in the apex, preventing this feature from being maximal. This could guide us towards heuristics to remove superfluous features, combat overfitting and reduce training time.

3.3. **Linear Hyperplanes Look Tropical on Log Paper.** Let X_{ij} be our point clouds, and for $\beta > 0$, let's define $x^\beta = (x_{ij}^\beta := e^{\beta X_{ij}})_{ij}$. We will show that separating different classes from x^β using a linear SVM, when β tends toward infinity, yields a separating hyperplane that converges towards a tropical hyperplane in the initial space.

Our method, by directly outputting an optimal-margin separating tropical hyperplane in pseudo-polynomial time, is expected to achieve better results. If β is small, we might indeed be far away from the limiting tropical hypersurface; conversely, as β approaches infinity, numerical error becomes predominant.


linear SVM, exp kernel (beta=1), test points


tropical SVM, test points

3.3.1. *Tropical hyperplanes are limiting classical hyperplanes.* Under the assumption that the x_{ij}^{β} are linearly separable ***(strong assumption because it has to hold for all values of*** β***)***, we compute a support vector classifier without intercept term, whose separation surface's equation is:

$$H^{\beta}: \quad w^{\beta} \cdot x = 0,$$

where all positive (resp. negative) points verify $w^{\beta} \cdot x \geq 1$ (resp. $w^{\beta} \cdot x \leq -1$). We write

$$w_i^{\beta} = \sigma_i e^{\beta W_i^{\beta}},$$

where $\sigma_i \in \{\pm 1\}$ and $W_i^{\beta} \in \mathbb{R} \cup \{-\infty\}$. For now, we simplify the study by considering a fixed W_i value and $w_i^{\beta} = \sigma_i e^{\beta W_i}$.

Lemma 25. *(Maslov's sandwich) For $\beta > 0$ and $I \subset [d]$ we have:*

$$0 \leq \beta^{-1} \log \left(\sum_{i \in I} e^{\beta(W_i + X_{ij})} \right) - \max_{i \in I}(W_i + X_{ij}) \leq \beta^{-1} \log d.$$

Proof. Let $\beta > 0$ and $I \subset [d]$. We have:

$$\exp\left\{\beta \max_{i \in I} (W_i + X_{ij})\right\} \leq \sum_{i \in I} e^{\beta(W_i + X_{ij})} \leq d \cdot \exp\left\{\beta \max_{i \in I} (W_i + X_{ij})\right\},$$

hence the result by taking the logarithm and dividing by β. □

Proposition 26. *Defining H^{trop} as the hyperplane of apex $(-W_i)_{i \in [d]}$, signed using $I^+ := \{i \in [d], \quad \sigma_i > 0\}$ and $I^- := [d] \backslash I^+$, we have:*

$$d_H \left(\log H^{\beta}, H^{trop}\right) \leq \beta^{-1} \log d,$$

where d_H is the tropical Haussdorf distance. Hence $\log H^{\beta} \longrightarrow H^{trop}$ as $\beta \longrightarrow +\infty$.

Proof. Let $\beta > 0$ and $X \in H^{\beta}$. By writing the inequalities from previous lemma with I^+ and I^-, and as

$$\beta^{-1} \log \left(\sum_{i \in I^+} e^{\beta(W_i + X_{ij})} \right) = \beta^{-1} \log \left(\sum_{i \in I^-} e^{\beta(W_i + X_{ij})} \right),$$

substracting the first to second inequality yields:

$$-\beta^{-1} \log d \leq \max_{i \in I^+}(W_i + X_{ij}) - \max_{i \in I^-}(W_i + X_{ij}) \leq \beta^{-1} \log d.$$

hence $d(X, H^{\mathrm{trop}}) \leq \beta^{-1} \log d$.

Reciprocally, let $Y \in H^{\mathrm{trop}}$ and $X = Y + \delta \mathbf{1}_{I^+}$, with δ to be defined later. To have $Y \in H^{\beta}$, we have to ensure that:

$$w^{\beta} \cdot y = 0,$$

which amounts to

$$\sum_{i \in [d]} \sigma_i e^{\beta W_i} e^{\beta X_i} = 0.$$

By separating the positive and negative terms, and taking the logarithm, we get

$$\beta \delta + \log \left(\sum_{i \in I^+} e^{\beta(W_i + X_{ij})} \right) = \log \left(\sum_{i \in I^-} e^{\beta(W_i + X_{ij})} \right),$$

which similarily yields

$$\delta \leq \left| \beta^{-1} \log \left(\sum_{i \in I^+} e^{\beta(W_i + X_{ij})} \right) - \beta^{-1} \log \left(\sum_{i \in I^-} e^{\beta(W_i + X_{ij})} \right) \right| \leq \beta^{-1} \log d,$$

hence $d(Y, H^\beta) \leq \beta^{-1} \log d$. □

Remark 27. We note L the order of magnitude of our data points, and B a typical number at which the computer starts having numerical errors or overflow (typically, for C integer calculations, $B = 2^{16} - 1$). We want to have $\beta L \ll \log B$ (that is, no numerical error), and $\beta^{-1} \log d \ll L$ (good convergence towards tropical hyperplane), i.e $d \ll B$. By directly computing logarithms, for instance, our method should be very suitable for $d \gtrsim 65535$ dimensions using C integers.

3.3.2. *Finding tropical apex.* In the classical hard-margin setting, admissible w vectors verify $w^T x^+ \geq 1$ (resp. $w^T x^- \leq -1$) for positive (resp. negative) vectors. Thus they belong to the polytope P^β where

$$P^\beta := \left\{ w \in \mathbb{R}^d, \quad w^T x^\beta_{j_+} \geq 1 \text{ and } w^T x^\beta_{j_-} \leq -1, \forall j_+, j_- \in J^+, J^- \right\}.$$

We hope that $L_\sigma(P^\beta) := \left\{ (\beta^{-1} \log(\sigma_i w_i))_{i \in [d]} \right\}$ converges towards the corresponding limiting tropical polytope $P^{\text{trop}}_\sigma := P^{\text{trop}}_{\sigma,+} \cap P^{\text{trop}}_{\sigma,-}$, where

$$P^{\text{trop}}_{+,\sigma} := \left\{ W \in (\mathbb{R} \cup \{-\infty\})^d, \quad \max_{i, \sigma_i = 1} (W_i + X^\sigma_{ij}) \geq \max_{i, \sigma_i = -1} (W_i + X^\sigma_{ij}) \vee 0, \forall j \in J^+ \right\}.$$

and

$$P^{\text{trop}}_{-,\sigma} := \left\{ W \in (\mathbb{R} \cup \{-\infty\})^d, \quad \max_{i, \sigma_i = -1} (W_i + X^\sigma_{ij}) \leq \max_{i, \sigma_i = 1} (W_i + X^\sigma_{ij}) \vee 0, \forall j \in J^- \right\}.$$

Theorem 28. *(Cite the corresponding article) When points x^β_{ij} are in a general position,* ***(clarify what this means as β approaches infinity)***

$$\lim_{\beta \to +\infty} L_\sigma(P^\beta) = P^{trop}_\sigma,$$

with respect to the Haussdorf distance.

We proved the convergence of the logarithm of linear hypersurfaces towards a tropical limiting hypersurface, when the apex is fixed. However, in practice, there is an underlying double limit here: for each β value, we compute an apex in the exponentialized space and we hope it will converge towards a fixed apex in the initial space: let's consider $w^\beta_i = \sigma_i e^{\beta W^\beta_i}$ again.

Theorem 29. *(Cite the corresponding article)*

$$W^\infty := \lim_{\beta \to +\infty} \beta^{-1} \log w^\beta \in \underset{W \in P^{trop}}{\arg\min} \left(\max w \right).$$

3.3.3. *Computing limiting margin (experimental).* In the classical setting, the margin is $\|w\|^{-1}$, described by the vector $w / \|w\|^2$. Conjecturing that this vector, when mapped back to the initial space, gives a meaningful approximation of the margin (prove it), we can compute the norm of this limiting vector. For $\beta > 0$:

$$\beta^{-1} \log \left(\frac{w^\beta}{\|w^\beta\|^2} \right) = \beta^{-1} \log w^\beta - 2 \log \|e^{\beta W^\beta}\|^{\beta^{-1}}.$$

Where the term $\log\|e^{\beta W^\beta}\|^{\beta^{-1}} \to \max\left|w^\beta\right|$, and $\beta^{-1}\log w^\beta \to W^\infty$. Hence:

$$\left\|\beta^{-1}\log\left(\frac{w^\beta}{\|w^\beta\|^2}\right)\right\| \longrightarrow \max W^\infty - \min W^\infty$$

would give a good approximation of the limiting margin.

References

[1] Marianne Akian, Stephane Gaubert, and Alexander Guterman. Tropical polyhedra are equivalent to mean payoff games. *International Journal of Algebra and Computation*, 22(01):1250001, Feb 2012.

[2] Marianne Akian, Stéphane Gaubert, Yang Qi, and Omar Saadi. Tropical linear regression and mean payoff games: or, how to measure the distance to equilibria. *SIAM J. Discret. Math.*, 37:632–674, 2021.

[3] Marianne Akian, Stephane Gaubert, and Sara Vannucci. Ambitropical geometry, hyperconvexity and zero-sum games, 2021.

[4] Xavier Allamigeon, Stephane Gaubert, Ricardo D. Katz, and Mateusz Skomra. Condition numbers of stochastic mean payoff games and what they say about nonarchimedean semidefinite programming. 2018.

[5] R. Cominetti, J. A. Soto, and J. Vaisman. On the rate of convergence of krasnosel'skiĭ-mann iterations and their connection with sums of bernoullis. *Israel Journal of Mathematics*, 199(2):757–772, Aug 2013.

[6] Diane Maclagan and Bernd Sturmfels. *Introduction to Tropical Geometry*. American Mathematical Society, Apr 2015.

[7] Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks, 2018.