SEPARATING POINTS BY PIECEWISE-LINEAR FUNCTIONS

THE REAL TROPICAL GEOMETRY OF NEURAL NETWORKS

APPLIED CATS SEMINAR

KTH

Marie-Charlotte Brandenburg
20 February 2024

Georg Loho
FU Berlin | University of Twente

Guido Montúfar
UC LA | MPI MiS

*work in progress

Linear Classifiers and Hyperplane Arrangements

Setup

Given data points $D = \{p_1, \ldots, p_M\} \in \mathbb{R}^d$ in **input space**

- a **linear classifier** is a linear function $f : \mathbb{R}^d \to \mathbb{R}$
- f defines a hyperplane $\{x \in \mathbb{R}^d \mid f(x) = 0\}$ in input space separating $\{p_i \mid f(p_i) > 0\}$ from $\{p_i \mid f(p_i) < 0\}$
- f can be parametrized as $f(x) = \langle s, x \rangle + a$ for some fixed $s \in \mathbb{R}^d, a \in \mathbb{R}$.
- **parameter space of linear classifiers** is $\{(s, a) \mid s \in \mathbb{R}^d, a \in \mathbb{R}\} \cong \mathbb{R}^{d+1}$.

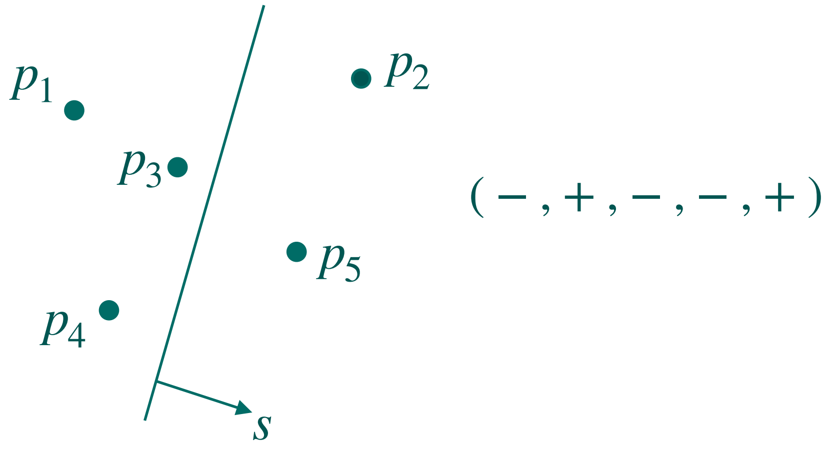
Classification by f **(dichotomy)** :
(sgn($f(p_1)$), …, sgn($f(p_M)$)) $\in \{-,0,+\}^M$

Goal

Subdivide parameter space into cells, in which classifiers have the same classification

Theorem (Cover '64, …)

These cells are chambers in the hyperplane arrangement $\bigcup_{p \in D} (p,1)^{\perp} \subseteq \mathbb{R}^{d+1}$ in parameter space



Linear Classifiers aka Hyperplane Arrangements

Theorem 1. (Function-Counting Theorem.) There are $C(N,d)$ homogeneously linearly separable dichotomies of N points in general position in Euclidean d-space where

$$C(N,d) = 2\sum_{k=0}^{d-1}\binom{N-1}{k}. \qquad (2.5)$$

Thomas M. Cover. “Geometrical and Statistical Properties of Linear Threshold Devices”. PhD Thesis (University of Stanford, 1964)

NUMBER 154

Theorem A. Let E be a Euclidean arrangement of hyperplanes. The number of its regions is

$$c(E) = \sum_{t\in LE} |\mu_{LE}(0,t)| = (-1)^{rE}\chi_{LE}(-1).$$

Thomas Zaslavsky

Facing up to Arrangements:
Face-Count Formulas
for Partitions of Space by Hyperplanes

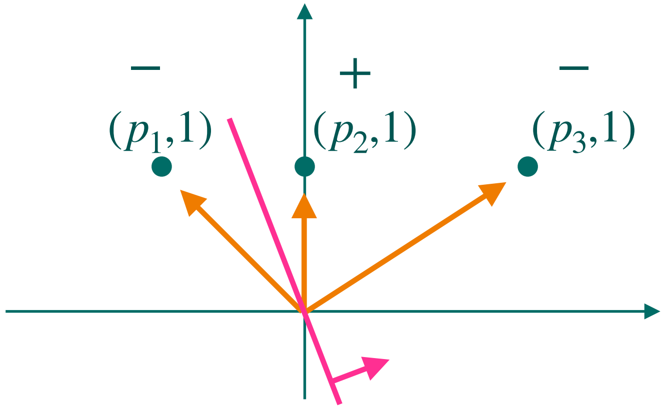
MEMOIRS
OF THE AMERICAN MATHEMATICAL SOCIETY

VOLUME 1 · ISSUE 1 · NUMBER 154 (first of 2 numbers) · JANUARY 1975 · CODEN: MAMCAU

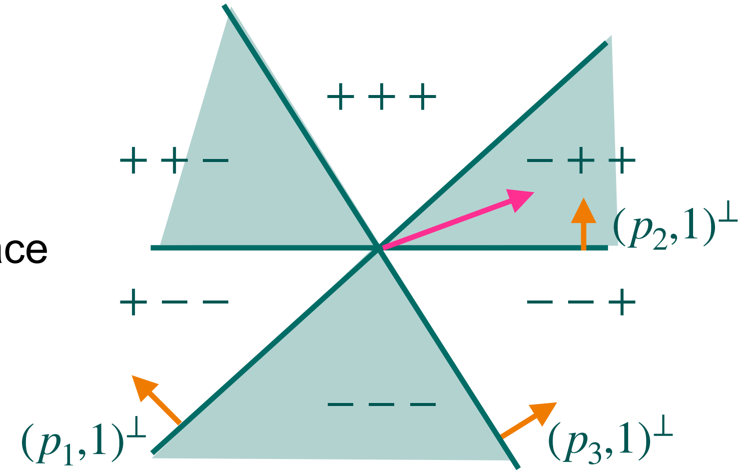
Thomas Zaslavsky. “Facing up to Arrangements”. PhD Thesis (MIT, 1974) and *Memoirs of the AMS* 1, 154 (1975)

Linear Classifiers and Hyperplane Arrangements

input space $\times\ \mathbb{R}$



parameter space



Goal

Subdivide parameter space into cells, in which classifiers have the same classification

Theorem (Cover '64, ...)

These cells are chambers in the hyperplane arrangement $\bigcup_{p\in D}(p,1)^\perp \subseteq \mathbb{R}^{d+1}$ in parameter space

$\{+++,\ ++-,\ +--,\ ---,\ --+, -++\}$
are the **dichotomies** of the data set

Fix a labelling $D = D^+ \sqcup D^-$

f **makes a mistake** at $p \in D^+$ if $f(p) < 0$
f **makes a mistake** at $p \in D^-$ if $f(p) > 0$

$0/1-$**loss function** counts number of mistakes of f

LINEAR CLASSIFIERS AND HYPERPLANE ARRANGEMENTS

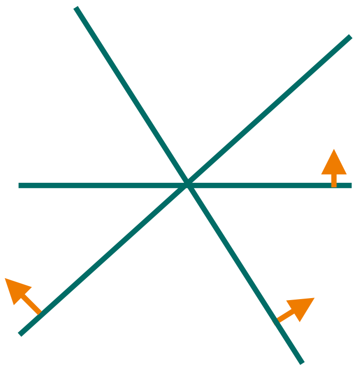
THEOREM

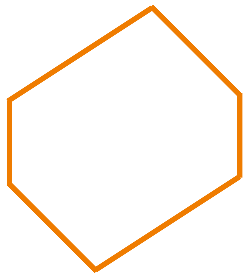
Let $D \subset \mathbb{R}^d$ be a finite data set. Then

(i) the hyperplane arrangement $\mathscr{H}_D = \bigcup_{p \in D} (1,p)^\perp$ subdivides the parameter space into regions according to the represented dichotomies,

(ii) $\mathscr{H}_D$ induces the normal fan of the zonotope $P_D = \sum_{p \in D} \text{conv}(\mathbf{0}, (1,p))$,

(iii) the dichotomies are the maximal covectors of the underlying realizable oriented matroid.





$\{+++,\ ++-,\ +--,\ ---,\ --+, -++\}$

What happens for piecewise-linear functions?

MOTIVATION: RELU NEURAL NETWORKS

A (feed-forward) **neural network** is a function $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$ which is an alternating composition

$$f = T^{(L)} \circ \sigma \circ \ldots \circ \sigma \circ T^{(1)} \circ \sigma \circ T^{(0)}$$

of affine linear functions

$$T^{(l)} : \mathbb{R}^{d_l} \to \mathbb{R}^{d_{l+1}}, T^{(l)}(x) = A^{(l)}x + b^{(l)}$$

and fixed functions $\sigma : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$.

Today: $d_{L+1} = 1$.

If $\sigma(x) = \max(0,x)$ (coordinate-wise) then f is a **ReLU network (Rectified Linear Unit)** with **depth** L and **architecture** $(d_0, d_1, \ldots, d_{L+1})$

$\to f$ is a piecewise linear function

$\to f$ is a **tropical rational function**

$$\begin{aligned} f(x) &= g - h \\ &= \max_{i\in[n]} \left(a_i + \langle s_i, x\rangle\right) - \max_{j\in[m]} \left(b_j + \langle t_j, x\rangle\right) \\ &= \bigoplus\nolimits_{i\in[n]} a_i \odot x^{\odot s_i} \oslash \bigoplus\nolimits_{j\in[m]} b_j \odot x^{\odot t_j} \end{aligned}$$

for some $n, m \in \mathbb{N}$.

THEOREM (ARORA-BASU-MIANJY-MUKHERJEE, ZHANG-NAITZAT-LIM, '18)

- $f\colon \mathbb{R}^d \to \mathbb{R}$ is a tropical rational function $f = g - h \iff f$ can be represented by a ReLU neural network.
- $g - h$ can be represented by a ReLU NN with depth $\min\left(\lceil \log_2(d+1)\rceil + 1, \max(\lceil \log_2(n)\rceil, \lceil \log_2(m)\rceil) + 2\right)$.

Parameter space of tropical rational functions

Fix number of terms n, m of functions

$$f(x) = \max_{i\in[n]} \left(a_i + \langle s_i, x\rangle\right) - \max_{j\in[m]} \left(b_j + \langle t_j, x\rangle\right)$$

$f = f_\theta$ is defined through its parameters

$$\theta = \begin{pmatrix} a_1 & \dots & a_n & b_1 & \dots & b_m \\ s_1 & \dots & s_n & t_1 & \dots & t_m \end{pmatrix}.$$

Parameter space of trop. rational functions:

$$\Theta(d,n,m) = \left\{\theta : a_i, b_j \in \mathbb{R},\ s_i, t_j \in \mathbb{R}^d\right\}$$
$$\cong \mathbb{R}^{(d+1)\times(n+m)}$$

$\mathrm{ReLU}(d_0, d_1, \dots, d_{L+1})$ = set of piecewise linear functions represented by a **ReLU network with architecture** $(d_0, d_1, \dots, d_{L+1})$

Theorem (B.-Loho-Montúfar):

Let $d, d_1, \dots, d_L \in \mathbb{N}$. Then

- There exist $n, m \in \mathbb{N}$ such that $\mathrm{ReLU}(d, d_1, \dots, d_L, 1)$ can be embedded into $\Theta(d, n, m)$
- This embedding is a basic semialgebraic set, i.e. described by polynomial inequalities.
- n, m can be chosen as $\log_2(m) \leq \sum_{k=1}^{L} 2^{L-k} \prod_{l=k}^{L} d_l$ and $n = 2m$.

Choosing $n, m = 1$ recovers the linear case

From Linear to Piecewise Linear Classification

Linear function	Tropical rational function
Separation by hyperplane	Signed tropical hypersurface
Polyhedral cone of perfect classifiers	Perfect classification fan
Chambers in a hyperplane arrangement	Classification fan
Arrangement of hyperplanes	Arrangement of indecision surfaces
Covectors of oriented matroids	Activation patterns
Zonotope	Activation polytope

} input space

} parameter space $\Theta(d, n, m)$

Decision Boundaries

$$g - h = \max_{i=1,\dots,n} \left(a_i + \langle s_i, x\rangle\right) - \max_{j=1,\dots,m} \left(b_j + \langle t_j, x\rangle\right)$$

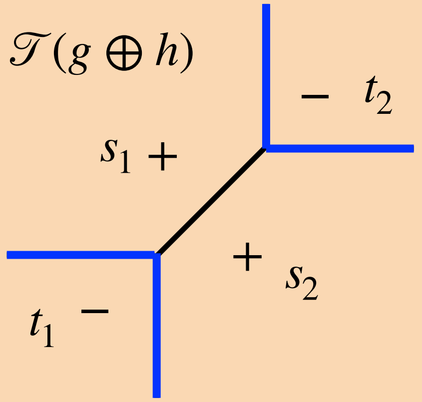
Decision boundary:

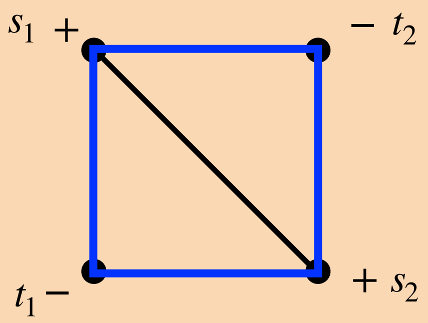
$\mathcal{B}(g-h) = \{x \in \mathbb{R}^d \mid g(x) - h(x) = 0\}$

→ Polyhedral complex with $\leq n \cdot m$ linear pieces

$g - h : \mathbb{R}^2 \to \mathbb{R}$

$$g(x) - h(x) = \max\left(a_1 + \langle s_1, x\rangle, a_2 + \langle s_2, x\rangle\right) - \max\left(b_1 + \langle t_1, x\rangle, b_2 + \langle t_2, x\rangle\right),$$





How to construct the decision boundary

- $g \oplus h = \max_{i \in [n],\, j \in [m]} \left(a_i + \langle s_i, x\rangle, b_j + \langle t_j, x\rangle\right)$
- Subdivide $\mathbb{R}^d$ into
  $\{x \mid g(x) \oplus h(x) = a_i + \langle s_i, x\rangle\}$ with label "+" and
  $\{x \mid g(x) \oplus h(x) = b_j + \langle t_j, x\rangle\}$ with label "–"
- Tropical hypersurface $\mathcal{T}(g \oplus h)$ is the codim-1 skeleton
- Decision Boundary $\mathcal{B}(g-h)$ is the sign-mixed subcomplex of $\mathcal{T}(g \oplus h)$
- Dual: Regular subdivision of signed Newton polytope $\mathcal{N}(g \oplus h)$. Decision boundary is dual to sign-mixed edges.

From Linear to Piecewise Linear Classification

Linear function	Tropical rational function
Separation by hyperplane	Signed tropical hypersurface
Polyhedral cone of perfect classifiers	Perfect classification fan
Chambers in a hyperplane arrangement	Classification fan
Arrangement of hyperplanes	Arrangement of indecision surfaces
Covectors of oriented matroids	Activation patterns
Zonotope	Activation polytope

input space

parameter space $\Theta(d, n, m)$

CLASSIFICATION BY TROPICAL RATIONAL FUCNTIONS

Given data points $D = \{p_1, \ldots, p_M\} \in \mathbb{R}^d$.

$$\theta = \begin{pmatrix} a_1 & \ldots & a_n & b_1 & \ldots & b_m \\ s_1 & \ldots & s_n & t_1 & \ldots & t_m \end{pmatrix} \in \Theta(d, n, m)$$

determines

$$f_\theta = \max_{i \in [n]} \left(a_i + \langle s_i, x \rangle\right) - \max_{j \in [m]} \left(b_j + \langle t_j, x \rangle\right)$$

Fix a target labelling $D = D^+ \cup D^-$.
θ defines a **perfect classifier** f_θ if and only if

$$\max_{i \in [n]} a_i + \langle s_i, p \rangle \geq \max_{j \in [m]} b_j + \langle t_j, p \rangle \; \forall p \in D^+$$

$$\max_{i \in [n]} a_i + \langle s_i, p \rangle \leq \max_{j \in [m]} b_j + \langle t_j, p \rangle \; \forall p \in D^-$$

$\Longrightarrow$ the set of perfect classifiers $\Sigma \subset \Theta(d, n, m)$ is a union of polyhedral cones (pure, non-complete polyhedral fan)

Ranging over all target labelling yields a complete polyhedral fan: **classification fan**

$\rightarrow$ How many cones does Σ have?
$\rightarrow$ How many connected components?

THEOREM (B.-LOHO-MONTÚFAR-TSERAN):

Σ has $\leq n^{|D^+|} m^{|D^-|}$ maximal cones. This bound is attained $\Longleftrightarrow$ D^+, D^- are separable by a hyperplane and both and D^+ and D^- are affinely independent sets.

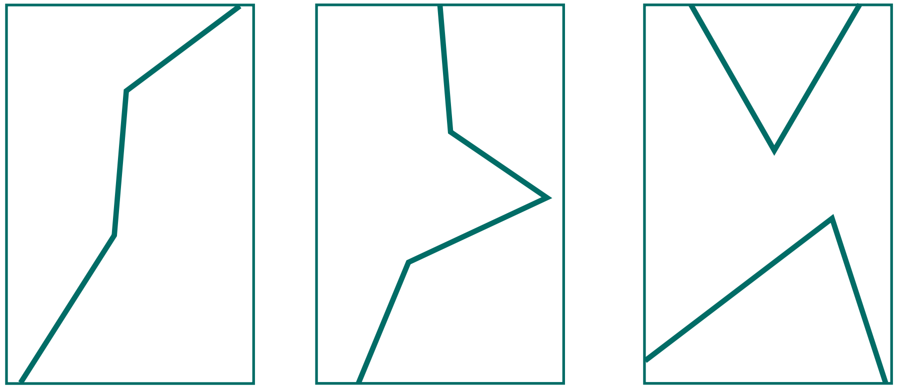
Perfect Classification

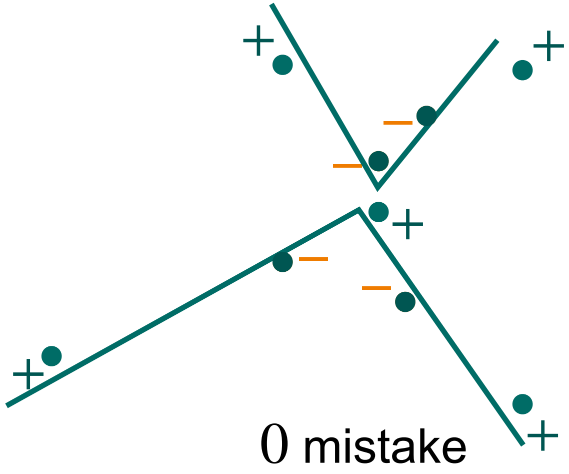
Classify 9 points in $\mathbb{R}^2$ in general position by piecewise linear functions (tropical rational functions) with $n = m = 2$ pieces.

Parameter space $\cong \mathbb{R}^{12}$, subdivided into 41680 12-dimensional polyhedral cones.

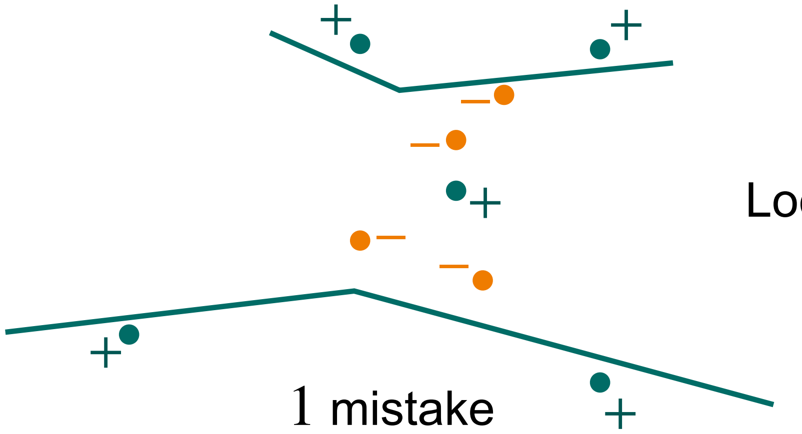
Fix a labelling $D = D^+ \sqcup D^-$.

16 cones make 0 mistakes, 8 connected components
304 cones make 1 mistake, 28 connected components





Global minimum



but no path to a cone with 0 mistakes (through codimension 1)

Local minimum

PERFECT CLASSIFICATION

THEOREM (B.-LOHO-MONTÚFAR):

- The perfect classification fan is not always connected (even if the data points are in general position).
- The sublevel sets of the 0/1-loss function are not always connected (even if the data points are in general position).

CHARACTERIZATIONS OF CLASSIFICATION FAN:

The perfect classification fan w.r.t $D^+ \sqcup D^-$ is

- the set of solutions to the linear inequalities

$$\max_{i\in[n]} a_i + \langle s_i, p\rangle \geq \max_{j\in[m]} b_j + \langle t_j, p\rangle \ \forall p \in D^+$$

$$\max_{i\in[n]} a_i + \langle s_i, p\rangle \leq \max_{j\in[m]} b_j + \langle t_j, p\rangle \ \forall p \in D^-$$

- the set of solutions of a system of tropical polynomial inequalities (tropical semialgebraic set)
- $\bigcap_{p\in D} \mathcal{S}^{\mathrm{label}(p)}(p)$
- the collection of all cones of the activation fan with compatible activation pattern

From Linear to Piecewise Linear Classification

Linear function	Tropical rational function
Separation by hyperplane	Signed tropical hypersurface
Polyhedral cone of perfect classifiers	Perfect classification fan
Chambers in a hyperplane arrangement	Classification fan
Arrangement of hyperplanes	Arrangement of indecision surfaces
Covectors of oriented matroids	Activation patterns
Zonotope	Activation polytope

input space

parameter space $\Theta(d, n, m)$

Indecision Surfaces

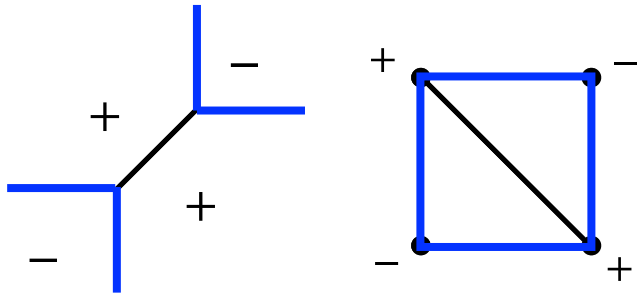
Linear case: Hyperplane arrangement $\bigcup_{p \in D} (1,p)^\perp$,
$(1,p)^\perp = \{(a,s) \mid a + \langle s, p \rangle = 0\}$

The **indecision surface** of a data point $p \in D$ is
$\mathcal{S}(p) = \{\theta \in \Theta(d,n,m) \mid (g_\theta - h_\theta)(p) = 0\}$.

$\mathcal{S}(p)$ consists of a_i, s_i, b_j, t_j such that

$$\max_{i \in [n]} \left(a_i + \langle s_i, p \rangle\right) - \max_{j \in [m]} \left(b_j + \langle t_j, p \rangle\right) = 0$$

$$\max_{i \in [n]} \left\langle \begin{pmatrix} a_i \\ s_i \end{pmatrix}, \begin{pmatrix} 1 \\ p \end{pmatrix} \right\rangle - \max_{j \in [m]} \left\langle \begin{pmatrix} b_j \\ t_j \end{pmatrix}, \begin{pmatrix} 1 \\ p \end{pmatrix} \right\rangle = 0$$



Theorem (B.-Loho-Montúfar-Tseran)

The indecision surface is the sign-mixed subcomplex of the normal fan of a simplex $\Delta(p)$ with signs

$$\Delta(p) = \text{conv}\left(\underbrace{\overset{+}{\begin{pmatrix} 1 & 0 & \dots & 0 \\ p & 0 & \dots & 0 \end{pmatrix}}, \dots, \overset{+}{\begin{pmatrix} 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & p & 0 & \dots & 0 \end{pmatrix}}}_{n \text{ times}}, \underbrace{\overset{-}{\begin{pmatrix} 0 & \dots & 0 & 1 & \dots & 0 \\ 0 & \dots & 0 & p & \dots & 0 \end{pmatrix}}, \dots, \overset{-}{\begin{pmatrix} 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & p \end{pmatrix}}}_{m \text{ times}} \right).$$

INDECISION SURFACES

$\text{label}(p) = +$ if $p \in D^+$,

$\text{label}(p) = -$ if $p \in D^-$

$\mathcal{S}(p) = \{\theta \in \Theta(d,n,m) \mid (g_\theta - h_\theta)(p) = 0\}$

$\mathcal{S}(p)$ subdivides the parameter space into
$\mathcal{S}^+(p) = \{\theta \mid (g_\theta - h_\theta)(p) \geq 0\}$ and
$\mathcal{S}^-(p) = \{\theta \mid (g_\theta - h_\theta)(p) \leq 0\}$.

CHARACTERIZATIONS OF CLASSIFICATION FAN:

The perfect classification fan w.r.t $D^+ \sqcup D^-$ is

- the set of solutions to the linear inequalities

$$\max_{i\in[n]} a_i + \langle s_i, p\rangle \geq \max_{j\in[m]} b_j + \langle t_j, p\rangle \ \forall p \in D^+$$

$$\max_{i\in[n]} a_i + \langle s_i, p\rangle \leq \max_{j\in[m]} b_j + \langle t_j, p\rangle \ \forall p \in D^-$$

- the set of solutions of a system of tropical polynomial inequalities (tropical semialgebraic set)
- $\bigcap_{\mathbf{p}\in\mathbf{D}} \mathcal{S}^{\text{label}(\mathbf{p})}(\mathbf{p})$
- the collection of all cones of the activation fan with compatible activation pattern

From Linear to Piecewise Linear Classification

Linear function	Tropical rational function	
Separation by hyperplane	Signed tropical hypersurface	} input space
Polyhedral cone of perfect classifiers	Perfect classification fan	parameter space $\Theta(d, n, m)$
Chambers in a hyperplane arrangement	Classification fan	
Arrangement of hyperplanes	Arrangement of indecision surfaces	
Covectors of oriented matroids	Activation patterns	
Zonotope	Activation polytope	

Classification By Tropical Polynomials

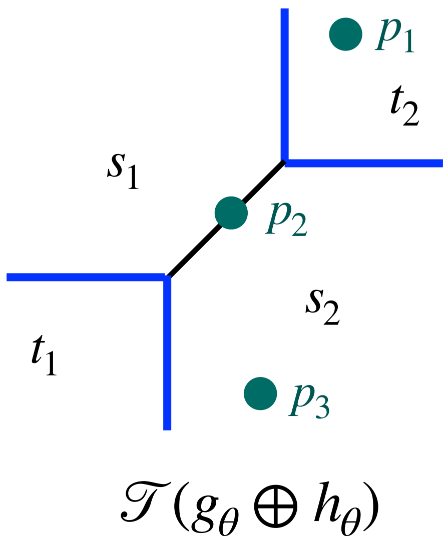
Given data points $D = \{p_1, \ldots, p_M\} \in \mathbb{R}^d$,
parameter $\theta \in \Theta(d, n, m)$,
function $f_\theta(x) = g_\theta(x) - h_\theta(x)$.

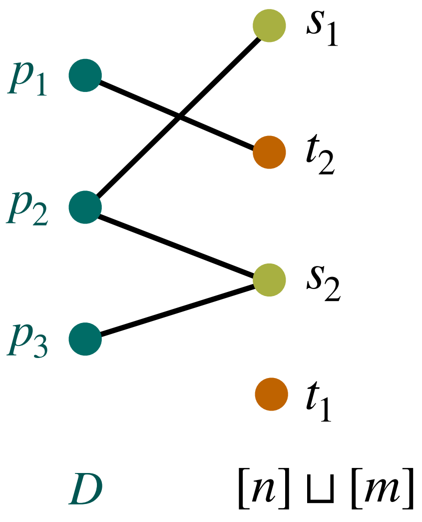
Recall:

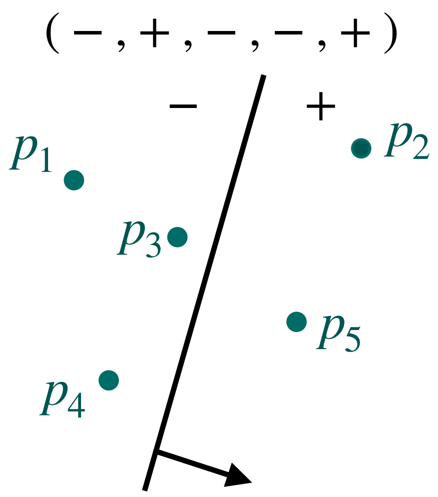
$$g_\theta \oplus h_\theta = \max_{i \in [n],\ j \in [m]} \Big(a_i + \langle s_i, x \rangle, b_j + \langle t_j, x \rangle \Big)$$

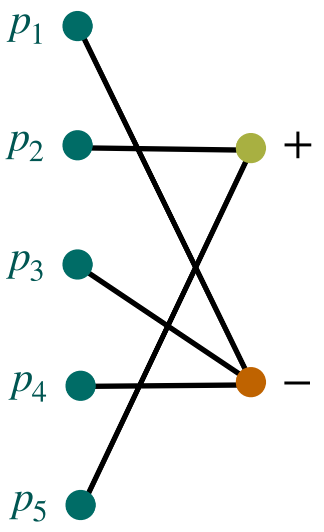
p lies in the region s_k of $\mathcal{T}(g_\theta \oplus h_\theta)$
$\iff g_\theta(p) \oplus h_\theta(p) = a_k + \langle s_k, p \rangle$
$\iff p$ **activates** the term s_k

activation pattern: bipartite graph $G_\theta = (V, E)$
$V = D \sqcup [N]$
$E = \{pk \mid p$ activates term k of $g_\theta \oplus h_\theta\}$
$\rightarrow$ generalization of covectors of oriented matroids









Classification by Tropical Polynomials

activation cone of a bipartite graph G: $C(G) = \{\theta \mid G = G_\theta$ is activation pattern$\}$

activation fan = collection of all nonempty activation cones (over all bipartite graphs)

→ complete fan in $\Theta(d, n, m)$

A data point $p \in \mathbb{R}^d$ defines a $(N-1)$-dimensional simplex

$$\Delta(p) = \text{conv}\left(\begin{pmatrix} 1 & 0 & \dots & 0 \\ p & 0 & \dots & 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & p & 0 & \dots & 0 \end{pmatrix}, \begin{pmatrix} 0 & \dots & 0 & 1 & \dots & 0 \\ 0 & \dots & 0 & p & \dots & 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & p \end{pmatrix}\right)$$

activation polytope $= \sum_{p \in D} \Delta(p)$ → generalization of zonotope

Theorem (B.-Loho-Montúfar-Tseran)

- The activation fan is the normal fan of the activation polytope.
- The activation fan coincides with the classification fan
- The perfect classification fan consists of all cones with compatible activation pattern

ACTIVATION PATTERNS

THEOREM (ACTIVATION PATTERNS)

Let $\mathscr{G}$ be the set of activation patterns of the activation fan. Then $\mathscr{G}$ satisfies

- **(Zero)** $K_{N,D} \in \mathscr{G}$
- **(Symmetry)** $G \in \mathscr{G} \implies$ any graph isomorphic to G under the action of S_N is contained in $\mathscr{G}$ →
- **(Composition)** $G, H \in \mathscr{G} \implies G \circ H \in \mathscr{G}$
- **(Elimination)** If $G, H \in \mathscr{G}$, $p \in D$ then there exists a graph $F \in \mathscr{G}$ with $N(p;F) = N(p;G) \cup N(p;H)$ →
- **(Boundary)** For each $i \in [N]$ and G the bipartite graph with edges $E(G) = \{pi \mid p \in D\}$ holds $G \in \mathscr{G}$ →
- **(Comparability)** For any $p \in D$, $G, H \in \mathscr{G}$, the comparability graph $\mathrm{CG}^p_{G,H}$ is acyclic

DEFINITION (TROPICAL ORIENTED MATROID)

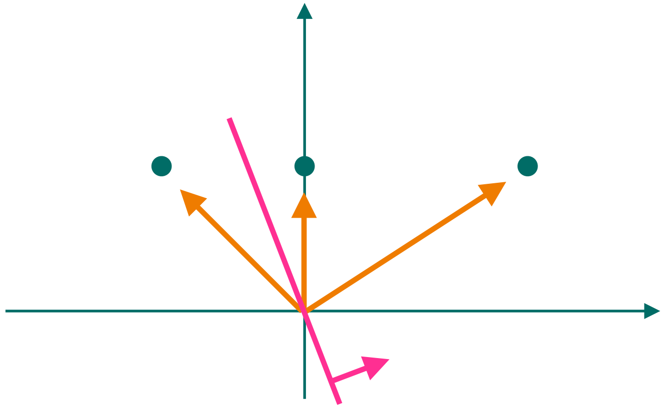
A tropical oriented matroid is a pair $([M], \mathscr{T})$, where $\mathscr{T} \subseteq \{(A_1, \ldots, A_M) \mid A_i \subseteq [N], i \in [M]\}$ are tropical covectors satisfying

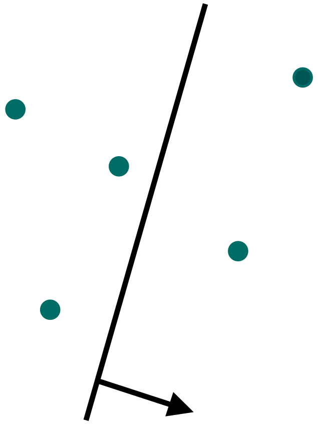
- **(Elimination)** If $A, B \in T$ and $j \in [D]$ then there exists a type $C \in T$ with $C_j = A_j \cup B_j$ and $C_k \in \{A_k, B_k, A_k \cup B_k\}$ for all $k \in [D]$.
- **(Boundary)** For each $j \in [N]$ holds $(\{j\}, \ldots, \{j\}) \in T$
- **(Comparability)** The comparability graph $\mathrm{CG}_{A,B}$ of any two types A and B in T is acyclic.
- **(Surrounding)** If $A \in T$ the any refinement is also in T.

DEFINITION (ORIENTED MATROID)

An oriented matroid is a pair $([M], \mathscr{C})$, where $\mathscr{C} \subseteq \{-,0,+\}^{[M]}$ are covectors satisfying

- **(Zero)** $(0,\ldots,0) \in \mathscr{C}$
- **(Symmetry)** $C \in \mathscr{C} \implies -C \in \mathscr{C}$
- **(Composition)** if $C, D \in \mathscr{C}$ then $(C \circ D) \in \mathscr{C}$
- **(Elimination)** if $C, D \in \mathscr{C}$ and $i \in S(C,D)$ then there exists some $Z \in \mathscr{C}$ such that $Z_i = 0$ and $Z_j = (C \circ D)_j \, \forall j \in [M] \setminus S(C,D)$.





THANK YOU