# On the location of separating hyperplanes with $\ell_p$-norms margins

Víctor Blanco
Universidad de Granada

(joint work with J. Puerto and A.M. Rodríguez-Chía)

REDLOCA

Segovia, September 2017

# Outline

# Introduction

Given a set of points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, each of them labeled with a class $y_i \in \{-1, +1\}$, find an hyperplane in $\mathbb{R}^d$ that separate both classes.

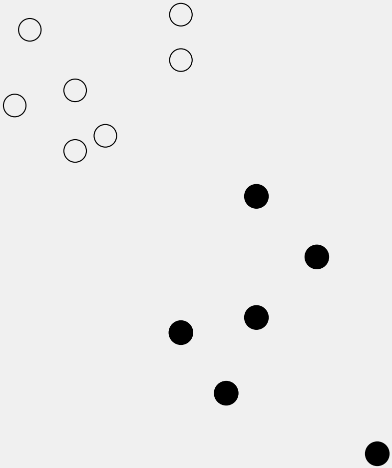
# Introduction

Given a set of points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, each of them labeled with a class $y_i \in \{-1, +1\}$, find an hyperplane in $\mathbb{R}^d$ that separate both classes.

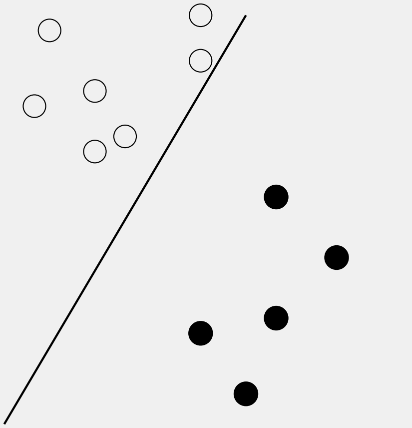
Find $\mathcal{H} = \{z \in \mathbb{R}^d : \omega^t z + b = 0\}$ such that:

- -1 Class belongs to $\{z : \omega^t z + b < 0\}$,
- $+1$ Class belongs to $\{z : \omega^t z + b > 0\}$,

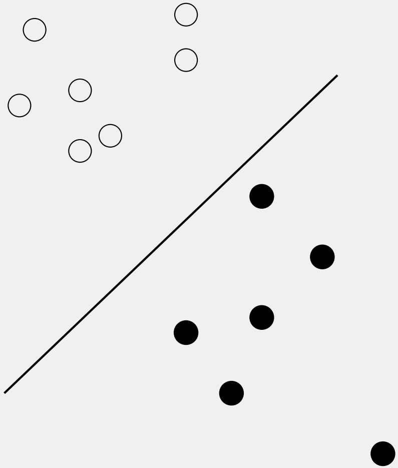
# Introduction



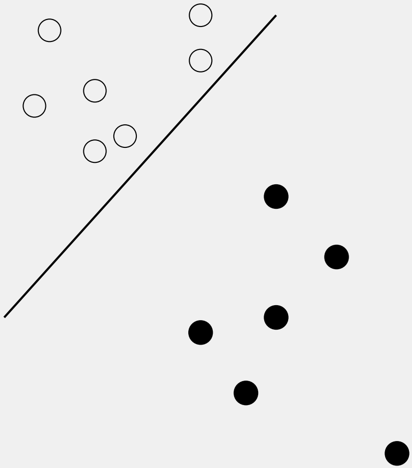
# Introduction



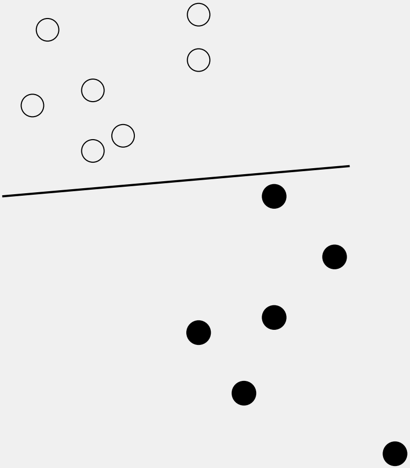
# Introduction



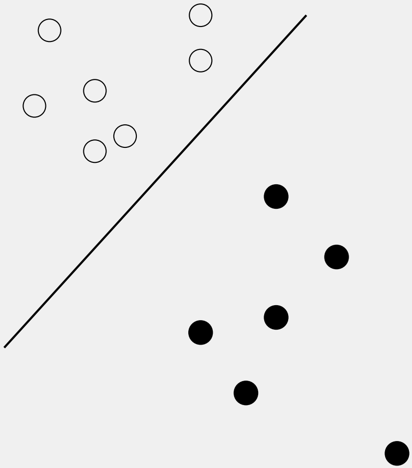
# Introduction



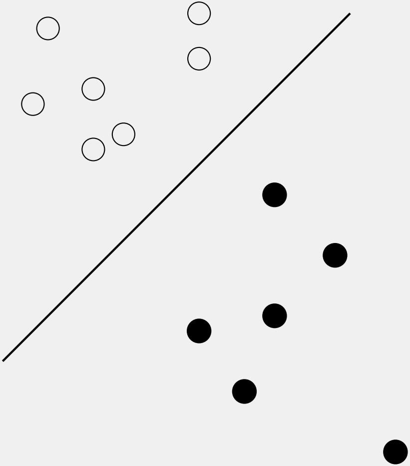
# Introduction



# Introduction



# Introduction

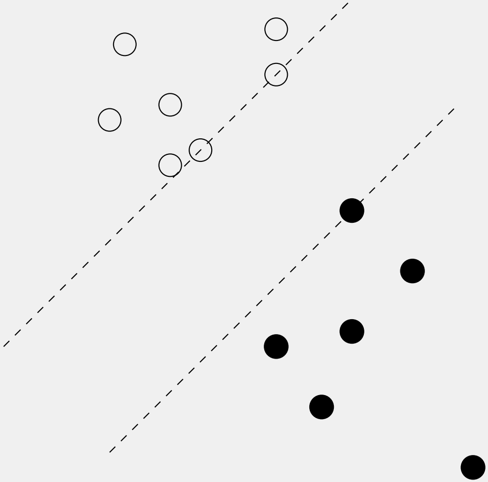


# Support Vector Machines

SVM (Vapnik & Chervonenkis, 63): Hyperplane such that the distance between the classes through $\mathcal{H}$ is maximized:

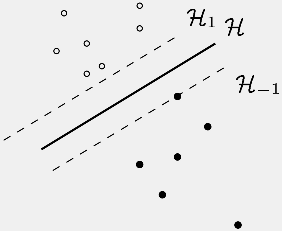
# Support Vector Machines

SVM (Vapnik & Chervonenkis, 63): Hyperplane such that the distance between the classes through $\mathcal{H}$ is maximized:.



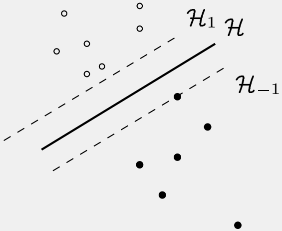
# Support Vector Machines

- Consider $\mathcal{H}$ and shifted hyperplanes $\mathcal{H}_1 = \{\mathrm{z} : \omega^{\mathrm{t}}\mathrm{x} + \mathrm{b} = 1\}$ and $\mathcal{H}_{-1} = \{\mathrm{z} : \omega^{\mathrm{t}}\mathrm{x} + \mathrm{b} = -1\}$.



# Support Vector Machines

- Consider $\mathcal{H}$ and shifted hyperplanes $\mathcal{H}_1 = \{z : \omega^t x + b = 1\}$ and $\mathcal{H}_{-1} = \{z : \omega^t x + b = -1\}$.
- Each observation should verify $y_i(\omega^t x_i + b) \geq 1$ (Separation).

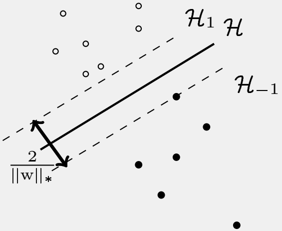


# Support Vector Machines

- Consider $\mathcal{H}$ and shifted hyperplanes $\mathcal{H}_1 = \{z : \omega^t x + b = 1\}$ and $\mathcal{H}_{-1} = \{z : \omega^t x + b = -1\}$.
- Each observation should verify $y_i(\omega^t x_i + b) \geq 1$ (Separation).
- Choose a norm $||\cdot||$ to measure the distances between both hyperplanes, then (Mangasarian, 99):

$$D(\mathcal{H}_1, \mathcal{H}_{-1}) = \frac{2}{||\omega||_*}$$

(where $||\cdot||_*$ is the dual norm of $||\cdot||$).



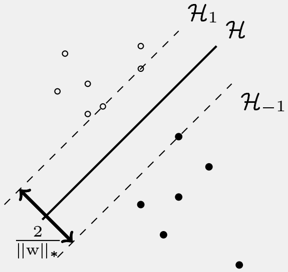
# Support Vector Machines

- Consider $\mathcal{H}$ and shifted hyperplanes $\mathcal{H}_1 = \{z : \omega^t x + b = 1\}$ and $\mathcal{H}_{-1} = \{z : \omega^t x + b = -1\}$.
- Each observation should verify $y_i(\omega^t x_i + b) \geq 1$ (Separation).
- Choose a norm $\|\cdot\|$ to measure the distances between both hyperplanes, then (Mangasarian, 99):

$$D(\mathcal{H}_1, \mathcal{H}_{-1}) = \frac{2}{\|\omega\|_*}$$

(where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$).
- Solve

$$\max_{y_i(\omega^t x_i + b) \geq 1} \frac{1}{\|\omega\|_*} \equiv \min_{y_i(\omega^t x_i + b) \geq 1} \|\omega\|_* .$$



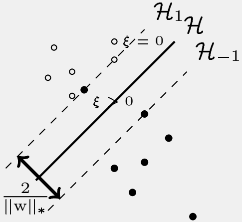
# Support Vector Machines

If points are non-linearly separable case: soft margin constraints:

# Support Vector Machines

If points are non-linearly separable case: soft margin constraints:
$\xi_i = \mathtt{max}\{0, 1 - y_i(\omega^t x_i + b)\}$ (Hinge Loss)

$$\begin{aligned} \mathtt{min} \quad & \|w\|_* + C \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & y_i(\omega^t x_i + b) \geq 1 - \xi_i, \forall i = 1, \ldots, n, \\ & \xi_i \geq 0, \forall i = 1, \ldots, n, \\ & \omega \in \mathbb{R}^d, b \in \mathbb{R}. \end{aligned}$$




Minimization of the risk incurred applying SVM to outsample data and the one of classifying the insample data.

# $\ell_p$-SVM

- Standard SVM = $\ell_2$-SVM.
- Successfully applied to classify data of different nature (Finance, Medicine, Biology, etc).
- $\ell_1$ and $\ell_\infty$ explored (Bradley & Mangasarian, 1998; Pedroso & Murata, 2001, Bennet and Bredensteiner 2000).
- Geometry under $\ell_p$-SVMs (Ikeda & Murata; 2005; Liu et. al, 2007).
- Different norms for different classes ($\ell_p$-SVM-$\ell_q$).

But very few is known about the optimization problems, transformation of data, use of kernels and about actual applications to classify databases.

# $\ell_p$-SVM

- Standard SVM = $\ell_2$-SVM.
- Successfully applied to classify data of different nature (Finance, Medicine, Biology, etc).
- $\ell_1$ and $\ell_\infty$ explored (Bradley & Mangasarian, 1998; Pedroso & Murata, 2001, Bennet and Bredensteiner 2000).
- Geometry under $\ell_p$-SVMs (Ikeda & Murata; 2005; Liu et. al, 2007).
- Different norms for different classes ($\ell_p$-SVM-$\ell_q$).

But very few is known about the optimization problems, transformation of data, use of kernels and about actual applications to classify databases.

- SOCP Formulations for the primal problem for $\ell_p$-SVMs ($p \geq 1$)..
- Formulate dual problem as polynomial optimization problems in homogeneous polynomials.
- Extend the theory under the Kernel Trick through Multidimensional Kernels.
- Apply $\ell_p$-SVM to real standard benchmarking problems.

# $\ell_p$-SVMs

Let $p = \frac{r}{s} > 1$, with $r, s \in \mathbb{Z}_+$ and $\mathsf{gcd}(r, s) = 1$.
We are given a set of n points in $\mathbb{R}^d$, x, and their classes $y \in \{-1, 1\}^n$.

$$x_{i\cdot} = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d.$$
$$x_{\cdot j} = (x_{1j}, \dots, x_{nj}) \in \mathbb{R}^n.$$

# $\ell_p$-SVMs

Let $p = \frac{r}{s} > 1$, with $r, s \in \mathbb{Z}_+$ and $\mathsf{gcd}(r, s) = 1$.
We are given a set of n points in $\mathbb{R}^d$, x, and their classes $y \in \{-1, 1\}^n$.

$$x_{i\cdot} = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d.$$
$$x_{\cdot j} = (x_{1j}, \ldots, x_{nj}) \in \mathbb{R}^n.$$

Let q such that $\frac{1}{p} + \frac{1}{q} = 1$: $\|\cdot\|_{p^*} = \|\cdot\|_q$.

$$\begin{aligned} \rho^* = \min \|\omega\|_q^q + C\sum_{i=1}^{n} \xi_i & \\ \text{s.t. } y_i(\omega^t x_{i\cdot} + b) \geq 1 - \xi_i, & \quad \forall i = 1, \ldots, n, \quad (\ell_p - \text{SVM}) \\ \xi_{\geq} 0, \omega \in \mathbb{R}^d, b \in \mathbb{R} & \quad (1) \end{aligned}$$

# $\ell_p$-SVMs

$$
\begin{aligned}
\min \quad & t + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad & y_i(\omega^t x_{i\cdot} + b) \geq 1 - \xi_i, & \forall i = 1, \ldots, n, \\
& t \geq \|\omega\|_q^q, \\
& \xi \geq 0, \omega \in \mathbb{R}^d, b \in \mathbb{R}
\end{aligned}
$$

# $\ell_p$-SVMs

$$
\begin{aligned}
\min \quad & t + C\sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad & y_i(\omega^t x_{i\cdot} + b) \geq 1 - \xi_i, && \forall i = 1, \ldots, n, \\
& t \geq \|\omega\|_q^q, \\
& \xi_{\geq} 0, \omega \in \mathbb{R}^d, b \in \mathbb{R}
\end{aligned}
$$

Constraint $t \geq \|\omega\|_q^q$ can be rewritten as ($q = \frac{r}{r-s}$):

$$
\begin{cases}
v_j \geq |\omega_j| & \forall j = 1, \ldots, d, \\
t \geq \displaystyle\sum_{j=1}^{d} u_j, \\
u_j^{r-s} \geq v_j^r, & \forall j = 1, \ldots, d,
\end{cases}
$$

Polynomial constraints in the form $u_j^{r-s} \geq v_j^r$ can be efficiently rewritten as SOC-constraints (B., Puerto, ElHaj, 2014).

# The Dual Problem

$$
\begin{aligned}
&\min ||\omega||_q^q + C\sum_{i=1}^{n}\xi_i \\
&\text{s.t. } y_i(\omega^t x_{i\cdot} + b) \geq 1 - \xi_i, \qquad \forall i = 1, \ldots, n, \qquad \text{(PRIMAL)} \\
&\qquad \xi_i \geq 0, \qquad \forall i = 1, \ldots, n.
\end{aligned}
$$

# The Dual Problem

$$\min ||\omega||_q^q + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } y_i(\omega^t x_{i\cdot} + b) \geq 1 - \xi_i, \qquad \forall i = 1, \ldots, n, \qquad \text{(PRIMAL)}$$

$$\xi_i \geq 0, \qquad \forall i = 1, \ldots, n.$$

$$\max \left( \frac{1}{q^p} - \frac{1}{q^{p-1}} \right) \sum_{j=1}^{d} \left| \sum_{i=1}^{n} \alpha_i y_i x_{ij} \right|^p + \sum_{i=1}^{n} \alpha_i \qquad \text{(LAG-DUAL)}$$

$$\text{s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,$$

$$0 \leq \alpha_i \leq C, \quad \forall i = 1, \ldots, n.$$

# The Dual Problem

$$\min ||\omega||_q^q + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.}\ \ y_i(\omega^t x_{i\cdot} + b) \geq 1 - \xi_i, \qquad \forall i = 1, \ldots, n, \qquad \text{(PRIMAL)}$$

$$\xi_i \geq 0, \qquad \forall i = 1, \ldots, n.$$

$$\max \left(\frac{1}{q^p} - \frac{1}{q^{p-1}}\right) \sum_{j=1}^{d} \left| \sum_{i=1}^{n} \alpha_i y_i x_{ij} \right|^p + \sum_{i=1}^{n} \alpha_i \qquad \text{(LAG-DUAL)}$$

$$\text{s.t.} \sum_{i=1}^{n} \alpha_i y_i = 0,$$

$$0 \leq \alpha_i \leq C, \quad \forall i = 1, \ldots, n.$$

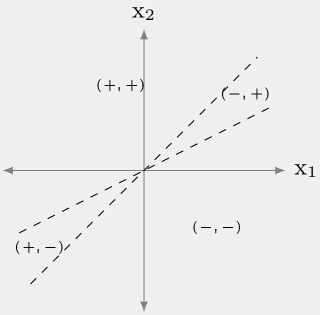
LAG-DUAL reformulated as a polynomial optimization problem.

# Alternative Dual Formulation

Consider the arrangement $\left\{ \sum_{i=1}^{n} \alpha_i y_i x_{ij} = 0 \right\}_{j=1}^{d}$ and subdivide the space into cells, such that each cell C is univocally defined by the signs of the expressions $\sum_{i=1}^{n} \alpha_i y_i x_{ij}$: $s_j$, for $j = 1, \ldots, d$: For each $\alpha$ in a cell:

# Alternative Dual Formulation

Consider the arrangement $\left\{ \sum_{i=1}^{n} \alpha_i y_i x_{ij} = 0 \right\}_{j=1}^{d}$ and subdivide the space into cells, such that each cell C is univocally defined by the signs of the expressions $\sum_{i=1}^{n} \alpha_i y_i x_{ij}$: $s_j$, for $j = 1, \ldots, d$: For each $\alpha$ in a cell:



# Alternative Dual Formulation

Consider the arrangement $\left\{ \sum_{i=1}^{n} \alpha_i y_i x_{ij} = 0 \right\}_{j=1}^{d}$ and subdivide the space into cells, such that each cell C is univocally defined by the signs of the expressions $\sum_{i=1}^{n} \alpha_i y_i x_{ij}$: $s_j$, for $j = 1, \ldots, d$: For each $\alpha$ in a cell:

$$\sum_{j=1}^{d} \left| \sum_{i=1}^{n} \alpha_i y_i x_{ij} \right|^r = \sum_{j=1}^{d} \mathcal{S}_{\alpha,j}^r \left( \sum_{i=1}^{n} \alpha_i y_i x_{ij} \right)^r = \sum_{\gamma \in \mathbb{N}_r^n} c_\gamma \alpha^\gamma y^\gamma \sum_{j=1}^{d} s_j^r x_{\cdot j}^\gamma$$

where $c_\gamma = \binom{(\sum_{i=1}^n \gamma_i)}{\gamma_1, \ldots, \gamma_n} = \frac{(\sum_{i=1}^n \gamma_i)!}{\gamma_1! \cdots \gamma_n!}$, and $\mathbb{N}_a^n := \{\gamma \in \mathbb{N}^n : \sum_{i=1}^n \gamma_i = a\}$.

$\mathcal{S}_{\alpha,j}^r = \text{sg}\left( \sum_{i=1}^{n} \alpha_i y_i x_{ij} \right)^r$.

# Alternative Dual Formulation

For $p \in \mathbb{N}$, and sign-patterns of the cell, s:

$$\max f_s(\alpha) := \left(\frac{1}{q^p} - \frac{1}{q^{p-1}}\right) \sum_{\gamma \in \mathbb{N}_r^n} c_\gamma \alpha^\gamma y^\gamma \sum_{j=1}^{d} s_j^r x_{\cdot j}^\gamma + \sum_{i=1}^{n} \alpha_i \tag{2}$$

$$\text{s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0, \tag{3}$$

$$s_j \sum_{i=1}^{n} \alpha_i y_i x_{ij} \geq 0, \qquad \forall j = 1, \ldots, d, \tag{4}$$

$$0 \leq \alpha_i \leq C, \qquad \forall i = 1, \ldots, n. \tag{5}$$

# From dual solution to hyperplane

Let $\bar{\alpha}$ optimal for a subdivision:
For $i_0$ such that $0 < \bar{\alpha}_{i_0} < C$:

$$b = y_{i_0} - \frac{1}{q^{p-1}} \sum_{j=1}^{d} \mathcal{S}^{p}_{\bar{\alpha},j} \left( \sum_{i=1}^{n} \bar{\alpha}_i y_i x_{ij} \right)^{p-1} x_{i_0 j}.$$

and the induced hyperplane is:

$$\frac{1}{q^{r-1}} \sum_{\gamma \in \mathbb{N}^{n}_{r-1}} c_\gamma \bar{\alpha}^{\gamma} y^{\gamma} \sum_{j=1}^{d} \mathcal{S}^{r}_{\bar{\alpha},j} x^{\gamma}_{.j} z_j + b \quad = \quad 0.$$

for all $z \in \mathbb{R}^d$.

# Kernels

Let $\Phi : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{D}}$. Can we manage the dual problem without the explicit knowledge of $\Phi$?

# Kernels

Let $\Phi : \mathbb{R}^d \to \mathbb{R}^D$. Can we manage the dual problem without the explicit knowledge of $\Phi$? For the Euclidean case, YES:

$$\begin{aligned} &\max \; -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k y_i y_k \Phi(x_{i\cdot})^t \cdot \Phi(x_{k\cdot}) + \sum_{i=1}^{n}\alpha_i \\ &\text{s.t. } \sum_{i=1}^{n}\alpha_i y_i = 0, \\ &\qquad 0 \le \alpha_i \le C, \qquad \forall i = 1, \dots, n. \end{aligned}$$

Only the products $\Phi(x_{i\cdot})^t \cdot \Phi(x_{k\cdot})$ are needed! (Kernel trick)

# Kernels

Let $\Phi : \mathbb{R}^d \to \mathbb{R}^D$. Can we manage the dual problem without the explicit knowledge of $\Phi$? For the Euclidean case, YES:

$$\max \; -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k y_i y_k \Phi(x_{i\cdot})^t \cdot \Phi(x_{k\cdot}) + \sum_{i=1}^{n}\alpha_i$$

$$\text{s.t.} \; \sum_{i=1}^{n}\alpha_i y_i = 0,$$

$$0 \le \alpha_i \le C, \qquad \forall i = 1, \ldots, n.$$

Only the products $\Phi(x_{i\cdot})^t \cdot \Phi(x_{k\cdot})$ are needed! (Kernel trick)

Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $\Big(K(x_{i\cdot}, x_{j\cdot})\Big)_{i,j} \succ 0$. Then, there exists $\Phi : \mathbb{R}^d \to \mathbb{R}^D$ with $K(x_{i\cdot}, x_{j\cdot}) = \Phi(x_{i\cdot})^t \cdot \Phi(x_{k\cdot})$. (Mercer, 1909)

Also, the optimal $\ell_2$-SVM is $\sum_{i=1}^{n}\alpha_i^* y_i K(x_{i\cdot}, z) + b^* = 0$, $\forall z \in \mathbb{R}^d$.

NO NEED TO KNOW $\Phi$ NOT EVEN D.

# $\ell_p$-Kernels

We are given a data set $[x] = (x_{1\cdot}, \ldots, x_{n\cdot})$ together with their classification patterns $y = (y_1, \ldots, y_n)$ and $r \in \mathbb{N}$.

$$H_y = \{\alpha \in [0, C]^n : \sum_{i=1}^{n} \alpha_i y_i = 0\} \quad \text{and} \quad S : 2^{H_y} \to 2^{\{-1,1\}^D}$$

$$S(R) := \left\{ s = (s_1, \ldots, s_D) \in \{-1, 1\}^D : s_j = \mathrm{sg}(\sum_{i=1}^{n} \alpha_i y_i \Phi_j(x_{i\cdot})), \alpha \in R, \forall j \right\}.$$

## Definition

The family of sets $\{R_k\}_{k \in \mathcal{K}} \subseteq 2^{H_y}$ is called a suitable subdivision of $H_y$ if:

1. $\mathcal{K}$ is finite.
2. $\{R_k\}_{k \in \mathcal{K}}$ is a subdivision of $H_y$ and,
3. $S(R_k) = \{s_{R_k}\}$ for some $s_{R_k} \in \{-1, 1\}^D$ and for all $k \in \mathcal{K}$.

## Definition

Given a suitable partition $\{R_k\}_{k\in\mathcal{K}} \subseteq 2^{H_y}$ and $(\gamma, \lambda) \in \mathbb{N}_r^{n+1}$, $\lambda \in \{0, 1\}$, the operator

$$K[x]_{R_k,\gamma,\lambda}(z) := \sum_{j=1}^{D} s_{R_k,j}^{r} \Phi_j(x)^{\gamma} \Phi_j(z)^{\lambda}, \forall z \in \mathbb{R}^d, \tag{6}$$

is called a r-order Kernel function of $\Phi$ valid for each element $\alpha$ in $R_k$.

## Proposition

The separating hyperplane and the objective function can be rewritten for the $\Phi$-transformed data using the Kernel function.

# $\ell_p$-Kernels

- For even r, the sign coefficients are no longer needed: $\{H_y\}$ (with $|\mathcal{K}| = 1$) is a suitable subdivision of $H_y$. In such a case, given a transformation $\Phi : \mathbb{R}^d \to \mathbb{R}^D$, the kernel function becomes:

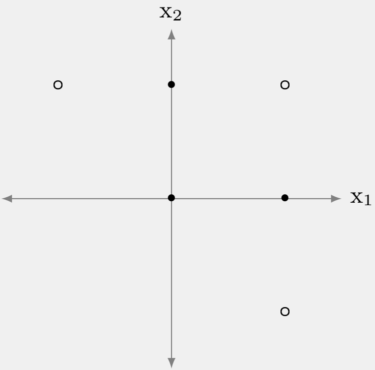
$$K[x]_{H_y,\gamma,\lambda}(z) := \sum_{j=1}^{D} \Phi_j(x)^\gamma \Phi_j(z)^\lambda, \qquad \forall z \in \mathbb{R}^d,$$

  for $(\gamma, \lambda) \in \mathbb{N}_r^n$ and $\lambda \in \{0, 1\}$, but being it independent of $\alpha$.

- For the Euclidean case, the usual definition of kernel is $K(z, z') = \Phi(z)^t \Phi(z')$ which is independent of the observations. However, for solving the dual problem, one only uses $K(x_{i\cdot}, x_{k\cdot})$ for $i_1, i_2 = 1, \ldots, n$, while for classifying an arbitrary observation z, one uses $K(x_{i\cdot}, z)$. (Extra information never used in K is required!!)

# Example

Let us consider six points in the plane $[x] = \big((0,0), (0,1), (1,0), (1,1), (1,-1), (-1,1)\big)$ with patterns $y = (1,1,1,-1,-1,-1)$.




Take $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$, $\Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$.

# Example

$H_y = \{\alpha \in [0, C]^6 : \sum_{i=1}^{6} \alpha_i y_i = 0\} = \{\alpha : \alpha_1 + \alpha_2 + \alpha_3 = \alpha_4 + \alpha_5 + \alpha_6\}$,

and $sg(\sum_{i=1}^{n} \alpha_i y_i \Phi_j(x_{i\cdot}))$ are:

$j = 1$ $sg(\alpha_3 - \alpha_4 - \alpha_5 - \alpha_6) = sg(-\alpha_1 - \alpha_2) = -1.$

$j = 2$ $sg(-\sqrt[r]{2}\alpha_4 + \sqrt[r]{2}\alpha_5 + \sqrt[r]{2}\alpha_6) = sg(\alpha_5 + \alpha_6 - \alpha_4).$

$j = 3$ $sg(\alpha_2 - \alpha_4 - \alpha_5 - \alpha_6) = sg(-\alpha_1 - \alpha_3) = -1.$

For odd r:

$$R_1 = \{\alpha \in H_y : \alpha_5 + \alpha_6 \geq \alpha_4\} \text{ and } R_2 = \{\alpha \in H_y : \alpha_5 + \alpha_6 \leq \alpha_4\}.$$

with $S(R_1) = \{(-1, 1, -1)\}$ while $S(R_2) = \{(-1, -1, -1)\}$.

# Example

$$K[x]_{R_k,\gamma,\lambda}(z) = \begin{cases} -\Phi_1(x)^\gamma\Phi_1(z)^\lambda + \Phi_2(x)^\gamma\Phi_2(z)^\lambda - \Phi_3(x)^\gamma\Phi_3(z)^\lambda, & \text{if } k = 1, \\ -\Phi_1(x)^\gamma\Phi_1(z)^\lambda - \Phi_2(x)^\gamma\Phi_2(z)^\lambda - \Phi_3(x)^\gamma\Phi_3(z)^\lambda, & \text{if } k = 2. \end{cases}$$

being then:

$$K[x]_{R_k,\gamma,\lambda}(z) = \begin{cases} -\left(x_{\cdot 1}^\gamma z_1^\lambda - x_{\cdot 2}^\gamma z_2^\lambda\right)^2, & \text{if } k = 1, \\ -\left(x_{\cdot 1}^\gamma z_1^\lambda + x_{\cdot 2}^\gamma z_2^\lambda\right)^2, & \text{if } k = 2. \end{cases}$$

For even r:

$$K[x]_{R_k,\gamma,\lambda}(z) = \left(x_{\cdot 1}^\gamma z_1^\lambda + x_{\cdot 2}^\gamma z_2^\lambda\right)^2,$$

for $k = 1, 2$, $(\gamma, \lambda) \in \mathbb{N}_r^{n+1}$ and $\lambda \in \{0, 1\}$.

# Kernels and Tensors

Given $z \in \mathbb{R}^d$, and for any $k \in \mathcal{K}$, the kernel operator $K[x]_{R_k,\gamma,\lambda}$ induces a r-order $(n+1)$-dimensional real tensor namely $\mathbb{K}^k = (\mathbb{K}^k_{i_1\ldots i_r})^n_{i_1,\ldots,i_r=1}$ with $\mathbb{K}^k_{i_1\ldots i_r} \in \mathbb{R}$ such that

$$\mathbb{K}^k_{i_1\ldots i_r} = \begin{cases} K[x]_{R_k,\gamma_0,0}(z) & \text{if } i_1,\ldots,i_r < n+1, \\ K[x]_{R_k,\gamma_1,1}(z) & \text{if there exists } s \in \{1,\ldots,r\} \text{ such that } i_s = n+1. \end{cases}$$

being $(\gamma_0,\lambda) = \sum_{l=1}^r e_{i_l}$ with $\lambda = 0$ and $(\gamma_1,\lambda) = \sum_{l=1}^r e_{i_l}$ with $\lambda = 1$ .

# Kernels and Tensors

## Theorem

Let $\{R_k\}_{k\in\mathcal{K}}$ be a suitable subdivision of $H_y$ consisting of semialgebraic sets and $\mathbb{K}^k$, for $k \in \mathcal{K}$, be a r-order $n+1$-dimensional symmetric tensor such that each $\mathbb{K}^k$ can be decomposed as:

$$\mathbb{K}^k = \sum_{j=1}^{h} \psi_{kj} v_j \otimes \overset{r}{\cdots} \otimes v_j, \forall k \in \mathcal{K},$$

satisfying, either

1. r is even and $\psi_j := \psi_{kj} \geq 0$, or
2. r is odd and $\mu_j := |\psi_{kj}|$ and $\text{sg}(\psi_{kj}) = \text{sg}\left( \sum_{i=1}^{n} \alpha_i y_i \sqrt[r]{\mu_j} v_{ji} \right)$, for all $k \in \mathcal{K}$.

Then, $\left( \{R_k\}_{k\in\mathcal{K}}, \{\mathbb{K}^k\}_{k\in\mathcal{K}} \right)$ induces a r-order kernel function.

# Solving the dual problem

$$\max f_s(\alpha) := \left(\frac{1}{q^p} - \frac{1}{q^{p-1}}\right) \sum_{\gamma \in \mathbb{N}_r^n} c_\gamma \alpha^\gamma y^\gamma \sum_{j=1}^{d} s_j^r x_{\cdot j}^\gamma + \sum_{i=1}^{n} \alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0,$$

$$s_j \sum_{i=1}^{n} \alpha_i y_i x_{ij} \geq 0, \qquad \forall j = 1, \ldots, d,$$

$$0 \leq \alpha_i \leq C, \qquad \forall i = 1, \ldots, n.$$

# Solving the dual problem

**Theorem (Lasserre's Relaxation)**

Let $t \geq t_0 = \lceil \frac{r}{2} \rceil$ and

$$\begin{aligned} \rho_t^* = \inf_{w} \;& L_w(-\tilde{f}) \\ \text{s.t.}\;& M_t(w) \succeq 0, \\ & M_{t-1}((\tilde{g}_0)w) \succeq 0, \\ & M_{t-\lceil \frac{r}{2} \rceil}(\tilde{g}_j w) \succeq 0,\; j = 1, \ldots, d, \\ & M_{t-1}(\tilde{g}_{d+j} w) \succeq 0,\; j = 1, \ldots, d \\ & M_{t-1}(\tilde{\ell}_i w) \succeq 0,\; i = 1, \ldots, n, \\ & L_w(w_0) = 1. \end{aligned}$$

The sequence $\{\rho_t^*\}_{t \geq t_0}$ of optimal values of the hierarchy of problems above satisfies

$$\lim_{t \to +\infty} -\rho_t^* \downarrow \max_{\alpha, \gamma \in H} \tilde{f}(\alpha, \gamma).$$

# Example

$$K[x]_{R_k,\gamma,\lambda}(z) = \begin{cases} -\left(x_{\cdot 1}^{\gamma} z_1^{\lambda} - x_{\cdot 2}^{\gamma} z_2^{\lambda}\right)^2, & \text{if } k = 1, \\ -\left(x_{\cdot 1}^{\gamma} z_1^{\lambda} + x_{\cdot 2}^{\gamma} z_2^{\lambda}\right)^2, & \text{if } k = 2. \end{cases}$$

where $R_1 = \{\alpha \in H_y : \alpha_5 + \alpha_6 \geq \alpha_4\}$ and $R_2 = \{\alpha \in H_y : \alpha_5 + \alpha_6 \leq \alpha_4\}$..

The following two problems have to be solved, for $r = 3$.

$$\max \left( \frac{1}{\left(\frac{3}{2}\right)^3} - \frac{1}{\left(\frac{3}{2}\right)^2} \right) \sum_{\gamma \in \mathbb{N}_3^6} c_\gamma \alpha^\gamma y^\gamma K[x]_{R_k,\gamma,0}(z) + \sum_{i=1}^{6} \alpha_i$$

$$\text{s.t. } \alpha \in R_k.$$

for $k = 1, 2$.

# Example

which for k = 1:

$$\rho_1^* = \max \frac{-4}{27} \Big(-\alpha_2^3 + 3\alpha_2^2\alpha_4 + 3\alpha_2^2\alpha_5 + 3\alpha_2^2\alpha_6 - 3\alpha_2\alpha_4^2 - 6\alpha_2\alpha_4\alpha_5 - 6\alpha_2\alpha_4\alpha_6 - 3\alpha_2\alpha_5^2 - 6\alpha_2\alpha_5\alpha_6 -$$
$$3\alpha_2\alpha_6^2 - \alpha_3^3 + 3\alpha_3^2\alpha_4 + 3\alpha_3^2\alpha_5 + 3\alpha_3^2\alpha_6 - 3\alpha_3\alpha_4^2 - 6\alpha_3\alpha_4\alpha_5 - 6\alpha_3\alpha_4\alpha_6 - 3\alpha_3\alpha_5^2 - 6\alpha_3\alpha_5\alpha_6 -$$
$$3\alpha_3\alpha_6^2 + 12\alpha_4^2\alpha_5 + 12\alpha_4^2\alpha_6 + 4\alpha_5^3 + 12\alpha_5^2\alpha_6 + 12\alpha_5\alpha_6^2 + 4\alpha_6^3\Big) + \sum_{i=1}^{6} \alpha_i$$

$$\begin{aligned} \text{s.t.}\quad & \alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 - \alpha_5 - \alpha_6 = 0, \\ & \alpha_5 + \alpha_6 \geq \alpha_4, \\ & 0 \leq \alpha_i \leq 10, \forall i = 1, \ldots, 6. \end{aligned}$$

# Example

Since $r = 3$, $t_0 \geq 2$. $M_2(w) \in \mathbb{R}^{28\times 28}$ is:

$$
\begin{array}{ccccccccc}
1 & \alpha_1 & \alpha_2 & \cdots & \alpha_6 & \alpha_1^2 & \alpha_1\alpha_2 & \cdots & \alpha_6^2
\end{array}
$$

$$
\begin{bmatrix}
w_{000000} & w_{100000} & w_{010000} & \cdots & w_{000001} & w_{200000} & w_{110000} & \cdots & w_{000002} \\
w_{100000} & w_{200000} & & \cdots & w_{100001} & w_{300000} & w_{120000} & \cdots & w_{100002} \\
& & \ddots & & & & & & \\
w_{000002} & & & & & & & & w_{000004}
\end{bmatrix}
\begin{array}{c}
1 \\ \alpha_1 \\ \vdots \\ \alpha_6^2
\end{array}
$$

in which 210 different variables are involved.

# Example

The semidefinite problem to solve is:

$$
\begin{aligned}
\min\ & L_w(-\tilde{f}) \\
\text{s.t.}\ & M_2(w) \succeq 0, \\
& w_{100000} + w_{010000} + w_{001000} - w_{000100} - w_{000010} - w_{000001} = 0, \\
& w_{000010} + w_{000001} \geq w_{000100}, \\
& 0 \leq w_{100000} \leq 10, \\
& 0 \leq w_{010000} \leq 10, \\
& 0 \leq w_{001000} \leq 10, \\
& 0 \leq w_{000100} \leq 10, \\
& 0 \leq w_{000010} \leq 10, \\
& 0 \leq w_{000001} \leq 10,
\end{aligned}
$$

where in $L_w(-\tilde{f})$ each term $\boldsymbol{\alpha}^\gamma$ is mapped into $w_\gamma$, for $\boldsymbol{\gamma} \in \mathbb{N}_r^n$.

# Example

Solving the above problem, we get $\rho^* = -5.6569$ and:

$$\mathrm{w}_{100000} = 0, \mathrm{w}_{010000} = \mathrm{w}_{001000} = \mathrm{w}_{000100} = 2.1213, \mathrm{w}_{000010} = \mathrm{w}_{000001} = 1.0611$$

The solution verifies the rank condition, certifying that the obtained solution is optimal:

$$\alpha^* = (0, 2.1213, 2.1213, 2.1213, 1.0611, 1.0611).$$

# Example

$$b = y_2 - \frac{1}{\left(\frac{3}{2}\right)^2} \sum_{\gamma \in \mathbb{N}_2^6} c_\gamma (\alpha^*)^\gamma y^\gamma K[x]_{R_1,\gamma,1}(x_{2\cdot}) = 3.0000$$

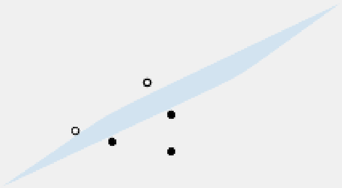
and the classifier:

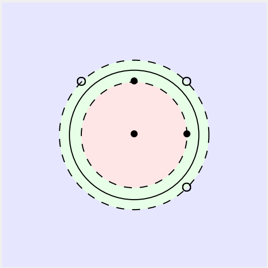
$$H(z) = sg\left( \frac{1}{q^{r-1}} \sum_{\gamma \in \mathbb{N}_2^6} c_\gamma {\alpha^*}^\gamma y^\gamma K[x]_{R_1,\gamma,1}(z) + b \right)$$

Evaluating the six points, we get:

$$H(x_{1\cdot}) = H(x_{2\cdot}) = H(x_{3\cdot}) = 1, \quad H(x_{4\cdot}) = H(x_{5\cdot}) = H(x_{6\cdot}) = -1$$

# Example





# Avoiding kernels: Schauder Bases

- Urysonh's Lemma $\Rightarrow$ Class 1 and Class -1 can be separated by a continuous function ($\mathbb{R}^d$ is topologically normal).
- Stone-Wiertrass Theorem $\Rightarrow$ Polynomials are dense over continuous functions.

For each $\varepsilon > 0$:

$$\exists N \in \mathbb{N} \text{ and } \omega_\gamma \in \mathbb{R} : y_i \left( \sum_{\gamma \in \mathbb{N}^n_N} \omega_\gamma x^\gamma + b \right) \geq 1$$

Transformations (for a given N):

- $\Phi_\gamma(z) = z^\gamma$, $\gamma \in \mathbb{N}^n_N$.
- $\widetilde{\Phi}_\gamma(z) = e^{\delta^2 ||z||^2} \dfrac{\sqrt{2\delta}}{\gamma_1! \cdots \gamma_n!} z^\gamma$, $\gamma \in \mathbb{N}^n_N$ (approximate exp functions).

Other Schauder bases!: orthogonal polynomials, trigonometric, etc

# Experiments

Datasets (UCI repository):

- `cleveland`: heart disease (303 obs., 13 features).
- `housing`: prices of Boston houses (303 obs., 13 features).
- `gc`: loan defaulters (1000 obs., 21 features)
- `colon`: cancerous colon tissues (62 obs., 2002 features)

Models were coded in `Python` 3.6, and solved using `Gurobi` 7.51.
A 10-fold cross validation scheme is used and the Accuracy is reported:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{n}} \quad \text{(Proportion of well-classified points)}$$

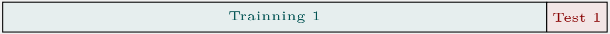
# Experiments

Datasets (UCI repository):

- **cleveland**: heart disease (303 obs., 13 features).
- **housing**: prices of Boston houses (303 obs., 13 features).
- **gc**: loan defaulters (1000 obs., 21 features)
- **colon**: cancerous colon tissues (62 obs., 2002 features)

Models were coded in **Python** 3.6, and solved using **Gurobi** 7.51.
A 10-fold cross validation scheme is used and the Accuracy is reported:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{n}} \quad \text{(Proportion of well-classified points)}$$




$\mathrm{ACC}_1$

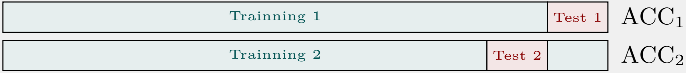
# Experiments

Datasets (UCI repository):

- **cleveland**: heart disease (303 obs., 13 features).
- **housing**: prices of Boston houses (303 obs., 13 features).
- **gc**: loan defaulters (1000 obs., 21 features)
- **colon**: cancerous colon tissues (62 obs., 2002 features)

Models were coded in **Python** 3.6, and solved using **Gurobi** 7.51.
A 10-fold cross validation scheme is used and the Accuracy is reported:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{n}} \quad \text{(Proportion of well-classified points)}$$



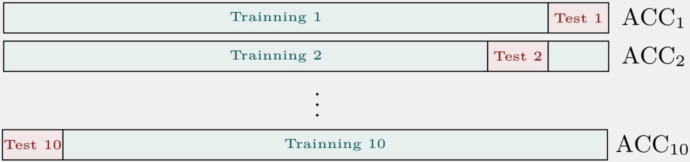
# Experiments

Datasets (UCI repository):

- **`cleveland`**: heart disease (303 obs., 13 features).
- **`housing`**: prices of Boston houses (303 obs., 13 features).
- **`gc`**: loan defaulters (1000 obs., 21 features)
- **`colon`**: cancerous colon tissues (62 obs., 2002 features)

Models were coded in **`Python`** 3.6, and solved using **`Gurobi`** 7.51.
A 10-fold cross validation scheme is used and the Accuracy is reported:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{n}} \quad \text{(Proportion of well-classified points)}$$



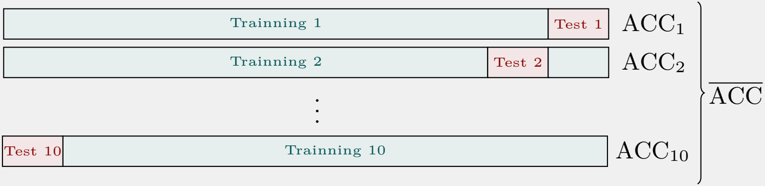
# Experiments

Datasets (UCI repository):

- **`cleveland`**: heart disease (303 obs., 13 features).
- **`housing`**: prices of Boston houses (303 obs., 13 features).
- **`gc`**: loan defaulters (1000 obs., 21 features)
- **`colon`**: cancerous colon tissues (62 obs., 2002 features)

Models were coded in **`Python`** 3.6, and solved using **`Gurobi`** 7.51.
A 10-fold cross validation scheme is used and the Accuracy is reported:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{n}} \quad \text{(Proportion of well-classified points)}$$



# Experiments

| cleveland | | $\ell_{1.5}$ | | | | $\ell_2$ | | | | $\ell_3$ | | | | $\ell_4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | deg | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ |
| $\Phi$ | 1 | 84.61% | 83.33% | 0.02 | 100% | 85.15% | 83.48% | 0.01 | 100% | 85.11% | 83.16% | 0.01 | 100% | 85.11% | 82.84% | 0.01 | 100% |
| | 2 | 95.23% | 76.67% | 0.43 | 98.09% | 93.33% | 81.58% | 0.04 | 98.95% | 93.58% | 81.57% | 0.40 | 94.48% | 94.02% | 82.57% | 0.44 | 88.86% |
| | 3 | 100% | 76.67% | 2.92 | 99.82% | 99.67% | 78.53% | 0.14 | 98.82% | 99.41% | 75.60% | 2.87 | 84.84% | 99.34% | 74.93% | 5.49 | 72.02% |
| | 4 | 100% | 76.67% | 19.01 | 98.44% | 99.74% | 79.21% | 0.47 | 97.54% | 99.67% | 76.92% | 22.50 | 81.88% | 99.67% | 76.56% | 28.00 | 72.00% |
| $\widetilde{\Phi}$ | 1 | 84.61% | 83.34% | 0.02 | 100% | 85.15% | 83.48% | 0.01 | 100% | 85.11% | 83.16% | 0.01 | 100% | 85.11% | 82.84% | 0.01 | 100% |
| | 2 | 90.84% | 81.12% | 0.02 | 96.48% | 88.71% | 83.86% | 0.04 | 99.62% | 89.29% | 85.17% | 0.29 | 89.62% | 89.81% | 83.85% | 0.29 | 75.71% |
| | 3 | 96.33% | 80.01% | 2.74 | 98.95% | 93.44% | 81.55% | 0.14 | 97.46% | 93.69% | 81.57% | 2.85 | 73.64% | 93.95% | 80.18% | 5.62 | 58.00% |
| | 4 | 85.71% | 85.02% | 7.43 | 59.61% | 85.29% | 84.18% | 0.19 | 55.32% | 85.33% | 83.80% | 11.12 | 14.07% | 85.48% | 85.06% | 28.95 | 6.19% |

| housing | | $\ell_{1.5}$ | | | | $\ell_2$ | | | | $\ell_3$ | | | | $\ell_4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | deg | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ |
| $\Phi$ | 1 | 90.54% | 86.27% | 0.03 | 100% | 88.10% | 84.36% | 0.02 | 100% | 88.25% | 85.16% | 0.02 | 100% | 88.56% | 85.36% | 0.01 | 100% |
| | 2 | 92.15% | 80.03% | 0.43 | 99.04% | 92.31% | 80.02% | 0.14 | 99.05% | 94.14% | 80.03% | 0.42 | 96.67% | 94.93% | 78.85% | 0.22 | 90.57% |
| | 3 | 96.82% | 80.35% | 6.90 | 99.82% | 97.34% | 79.81% | 0.51 | 97.27% | 98.24% | 80.00% | 6.13 | 74.84% | 98.60% | 80.95% | 9.57 | 57.36% |
| | 4 | 97.54% | 80.39% | 30.05 | 98.44% | 98.37% | 78.63% | 1.59 | 95.30% | 98.90% | 77.78% | 31.69 | 68.32% | 99.23% | 79.99% | 45.09 | 50.82% |
| $\widetilde{\Phi}$ | 1 | 87.03% | 85.15% | 0.05 | 100% | 88.10% | 84.36% | 0.02 | 100% | 88.25% | 85.16% | 0.02 | 100% | 88.56% | 85.36% | 0.01 | 100% |
| | 2 | 86.59% | 83.23% | 0.92 | 100% | 87.42% | 82.94% | 0.11 | 99.24% | 88.84% | 82.95% | 0.48 | 88.48% | 89.53% | 83.53% | 0.25 | 75.14% |
| | 3 | 89.89% | 79.35% | 3.31 | 99.64% | 91.50% | 80.21% | 0.25 | 88.30% | 93.30% | 79.82% | 4.29 | 54.52% | 94.01% | 80.03% | 4.47 | 37.38% |
| | 4 | 86.15% | 82.19% | 12.01 | 99.41% | 88.95% | 81.59% | 0.17 | 20.31% | 90.58% | 83.36% | 20.98 | 7.56% | 90.80% | 82.37% | 14.43 | 4.23% |

# Experiments

| gc | | $\ell_{1.5}$ | | | | $\ell_2$ | | | | $\ell_3$ | | | | $\ell_4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | deg | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ |
| $\Phi$ | 1 | 78.54% | 76.20% | 0.04 | 99.58% | 78.53% | 76.20% | 0.05 | 99.58% | 78.53% | 76.20% | 0.04 | 99.58% | 78.53% | 76.20% | 0.02 | 99.58% |
| | 2 | 93.00% | 67.70% | 3.32 | 99.75% | 92.98% | 67.40% | 0.50 | 99.69% | 93.04% | 67.60% | 2.50 | 98.15% | 93.03% | 67.50% | 0.92 | 96.62% |
| | 3 | 100% | 68.90% | 98.58 | 99.65% | 100% | 70.20% | 3.14 | 96.76% | 100% | 70.50% | 94.12 | 78.20% | 100% | 71.90% | 85.86 | 60.93% |
| $\widetilde{\Phi}$ | 1 | 78.26% | 78.75% | 0.04 | 100% | 78.25% | 78.63% | 0.05 | 100% | 78.33% | 78.88% | 0.04 | 100% | 78.35% | 79.00% | 0.02 | 99.48% |
| | 2 | 81.15% | 75.22% | 2.13 | 100% | 79.23% | 74.44% | 0.45 | 99.97% | 77.83% | 75.00% | 2.37 | 97.62% | 77.29% | 74.38% | 2.96 | 90.23% |
| | 3 | 98.24% | 76.57% | 48.40 | 99.99% | 96.36% | 77.88% | 2.75 | 99.82% | 92.78% | 79.00% | 63.64 | 91.69% | 76.72% | 76.75% | 577.01 | 3.39% |

| colon | | $\ell_{1.5}$ | | | | $\ell_2$ | | | | $\ell_3$ | | | | $\ell_4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Order | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ | $ACC^{Tr}$ | $ACC^{Te}$ | Time | NonZ |
| $\Phi$ | 1 | 100% | 80.48% | 14.61 | 99.44% | 100% | 80.48% | 0.05 | 89.74% | 100% | 80.48% | 15.73 | 64.54% | 100% | 82.14% | 20.30 | 46.14% |
| $\widetilde{\Phi}$ | 1 | 100% | 85.71% | 8.77 | 25.68% | 100% | 85.71% | 0.17 | 20.63% | 100% | 85.71% | 7.94 | 13.89% | 100% | 85.71% | 22.04 | 10.59% |

# Thank you!

vblanco@ugr.es