Sparsity in Learning

Y. Grandvalet

Heudiasyc, CNRS & Université de Technologie de Compiègne

Statistical Learning

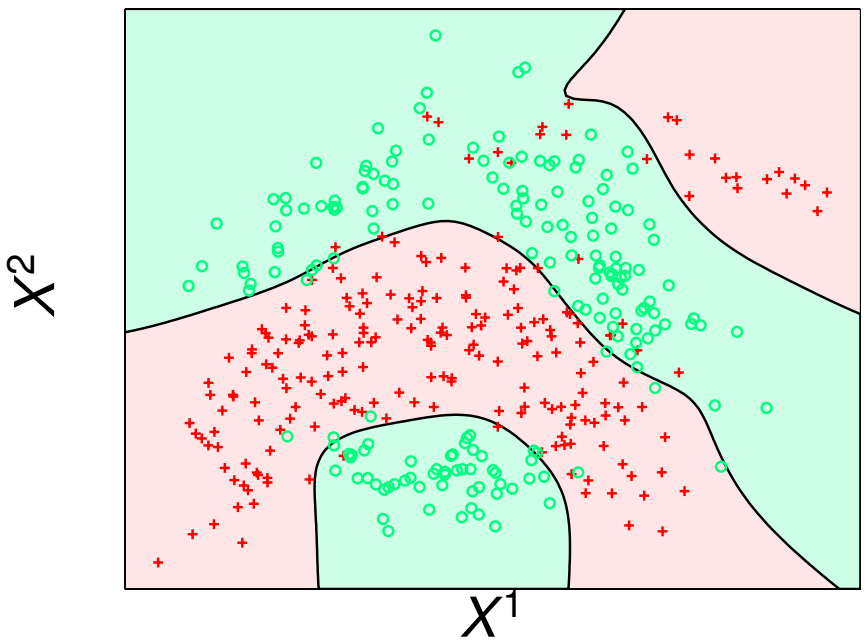
- **Regression**
- **Classification**
- Clustering

Statistical Learning

Generalize from examples

Given a training sample, $\{(x_i, y_i)\}_{i=1}^n$ adjust $\widehat{f} \in \mathcal{F}$, such that $\widehat{f}(x_i) \simeq y_i$.

Choose $\mathcal{F}$ not too small, nor too large, so that $\widehat{f}$ reaches a trade-off between fit and smoothness



Learning Algorithm

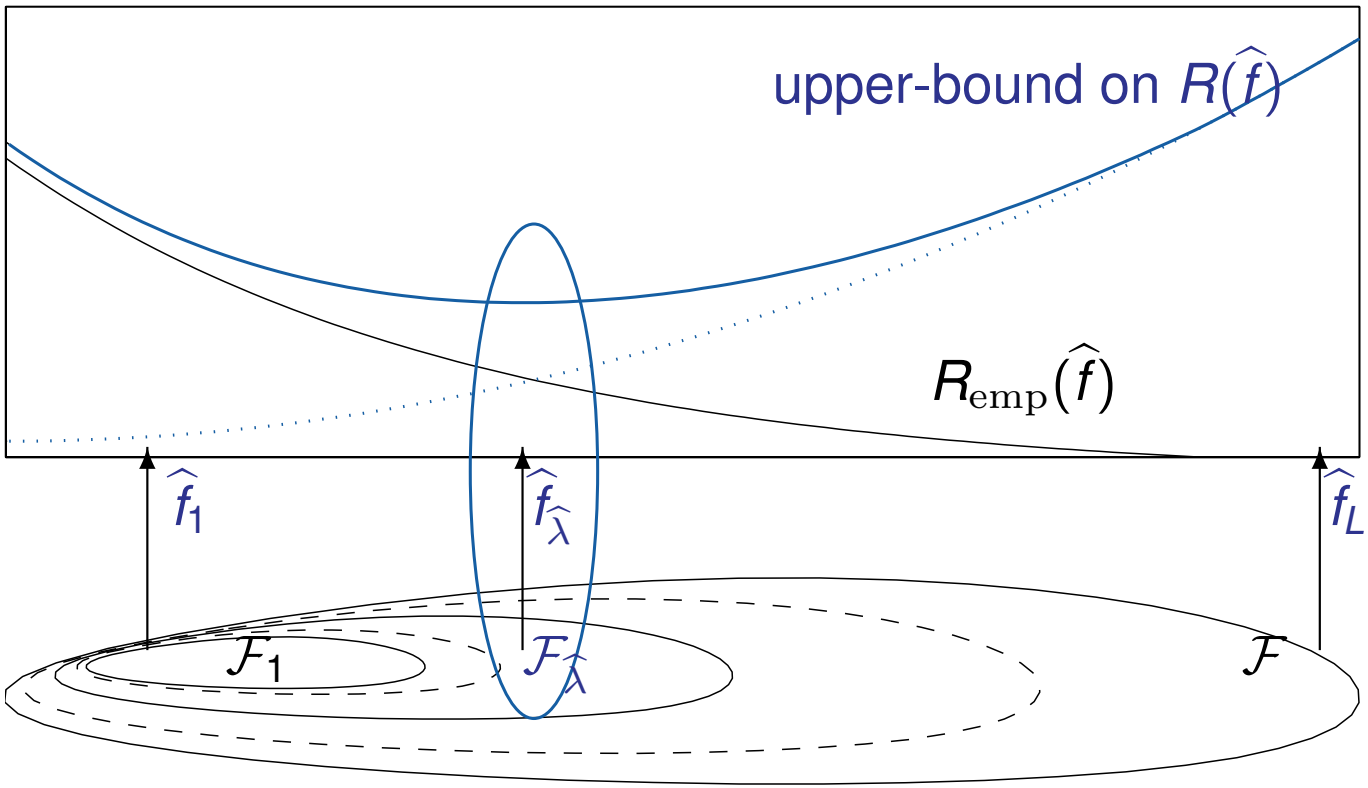
A 3 steps process

Structural Risk Minimization: choose $\mathcal{F}$ and $\widehat{f}$

1. Define a nested family of models $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \mathcal{F}_\lambda \ldots \subset \mathcal{F}_L$
2. Fit to data: $\widehat{f}_\lambda = \underset{f \in \mathcal{F}_\lambda}{\operatorname{Argmin}} R_{\text{emp}}(f),\ \lambda = 1 \ldots, L$
3. Select model $\mathcal{F}_{\widehat{\lambda}}$ by estimating the expected loss of $\widehat{f}_\lambda$

Choosing $\mathcal{F}$ amounts to choose a parameter

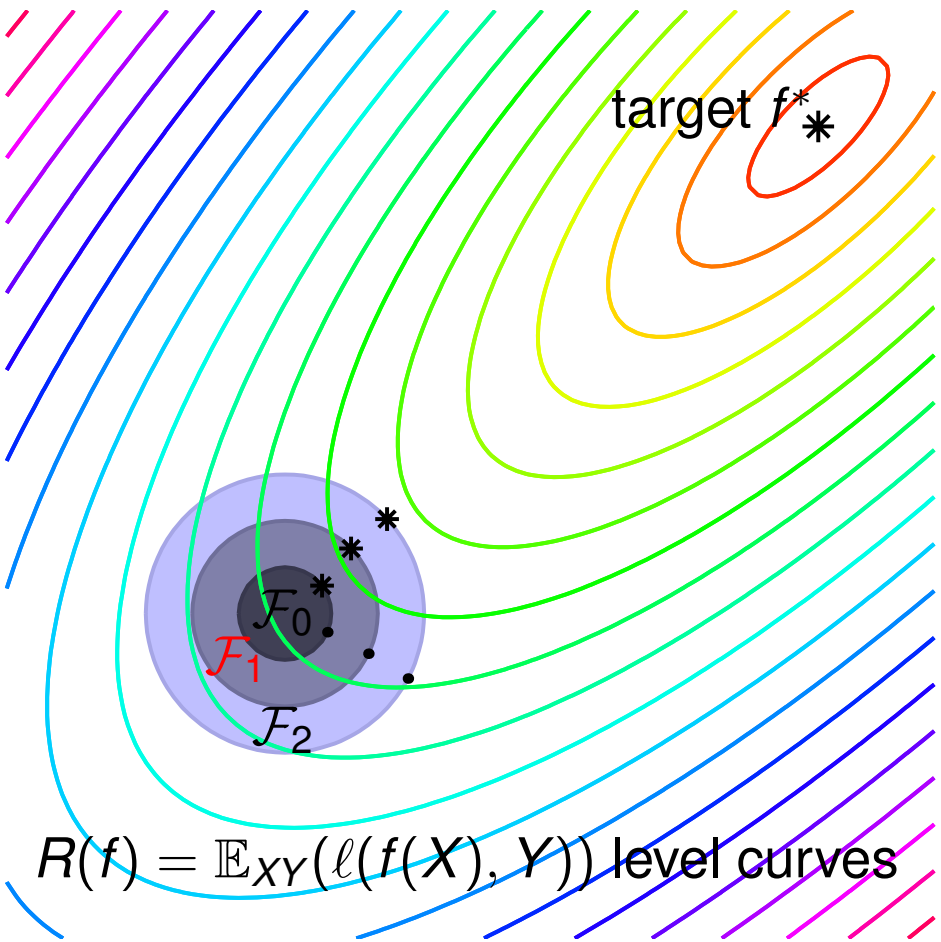
Structural Risk Minimization



3. Minimize $\widehat{R}(\widehat{f}_\lambda)$

Structural Risk Minimization

Approximation/estimation trade-off



Parsimonious use of data

We consider the data table :

$$\mathbf{X} = \begin{pmatrix} x_1^t \\ \vdots \\ x_i^t \\ \vdots \\ x_n^t \end{pmatrix} = \left(\mathbf{x}^1 \dots \mathbf{x}^j \dots \mathbf{x}^d\right)$$

This table can be reduced

1. in rows $\Rightarrow$ suppress some examples: compression $\Rightarrow$ loss function
2. in columns $\Rightarrow$ suppress variables: Occam's razor $\Rightarrow$ model selection
3. in rows and columns
4. in rank (PCA, PLS, . . .)

Why ignoring some variables...

since the Bayes error may only decreases with more variables ?

- Means to implement Structural Risk Minimization
  - Penalize to stabilize
  - Parsimony is sometimes a "reasonable prior"
- Computational efficiency:
  - Iteratively solve problem of increasing size
  - Exact regularization paths
  - Fast evaluation
- Interpretability 
  - Understanding the underlying phenomenon
  - Acceptability

Three categories of methods

1. "Filter" approach
   - Variables "filtered" by a criterion (Fisher, Wilks, mutual information)
   - Learning proceeds after the treatement
2. "Wrapper" approach
   - Heuristic search of subsets of variables
   - Subset selection is determined by the learning algorithm performance
   - no feedback
3. **"Embedded"**
   - Variable selection mechanism incorporated in the learning algorithm
   - All variables processed during learning, some will not influence the solution

Embedded Subset Selection

For linear models

$$f(x;\beta) = \beta_0 + \sum_{j=1}^{d} \beta_j x^j \quad ,$$

Subset selection aims at solving the problem

$$\begin{cases} \min\limits_{\beta} & \frac{1}{n}\sum\limits_{i=1}^{n} \ell(f(x_i;\beta), y_i) \\ \text{s. t.} & \|\beta\|_0 \leq d' < d \end{cases} \quad ,$$

wher d' is the number of desired variables

NP-hard problem

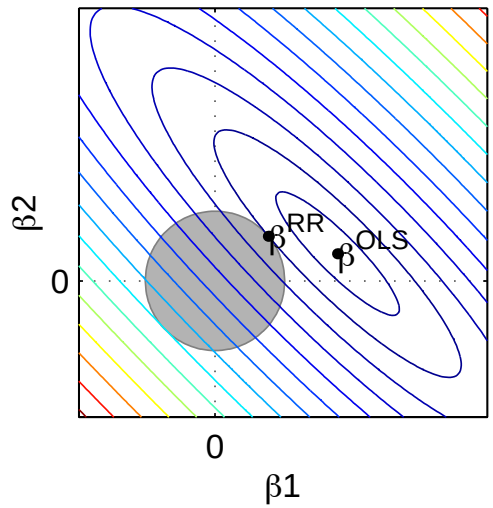
Relaxation
Soft-thresholding

Relax "hard" subset selection

$$\begin{cases} \min\limits_{\beta} & \frac{1}{n}\sum\limits_{i=1}^{n} \ell(f(x_i;\beta), y_i) \\ \text{s. t.} & \|\beta\|_p \leq c \end{cases} .$$

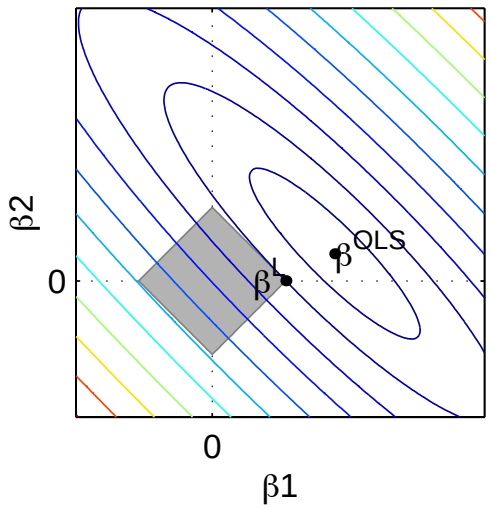
Sparse solution for $p \leq 1$
Convex optimization problem (if ℓ convex) for $p \geq 1$

Sparsity – Convexity Trade-off



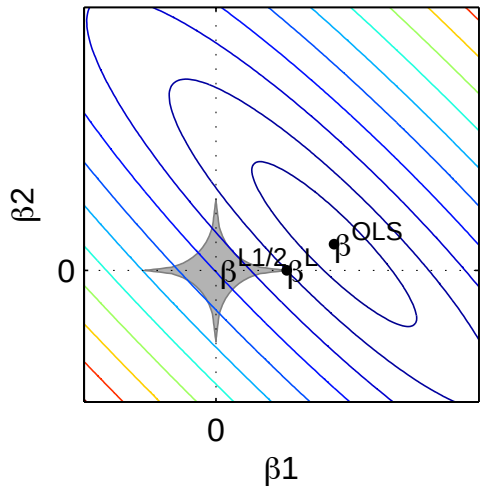
$\sum_{j=1}^{d} |\beta_j|^2$

ridge (weight decay)



$\sum_{j=1}^{d} |\beta_j|$

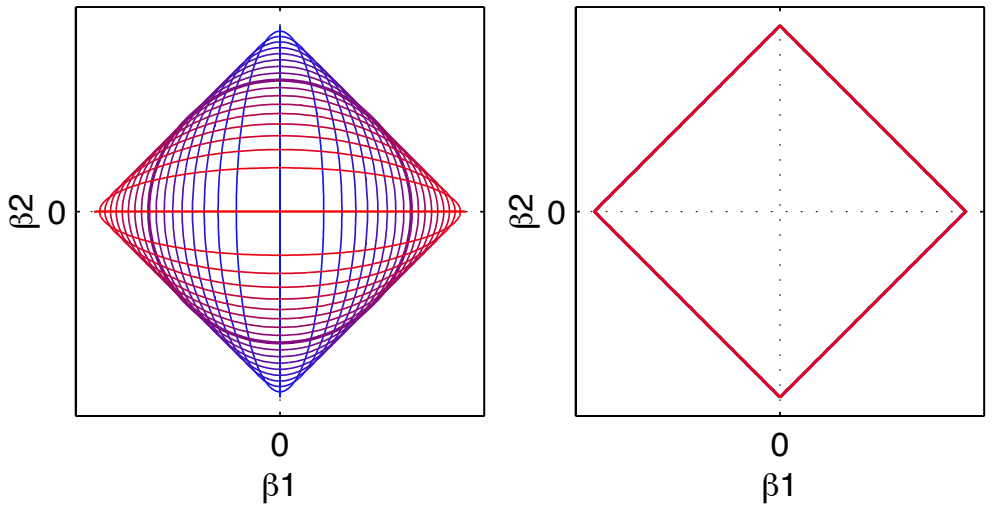
LASSO



$\sum_{j=1}^{d} |\beta_j|^{1/2}$

Geometric Insight on Adaptivity

Variational formulation

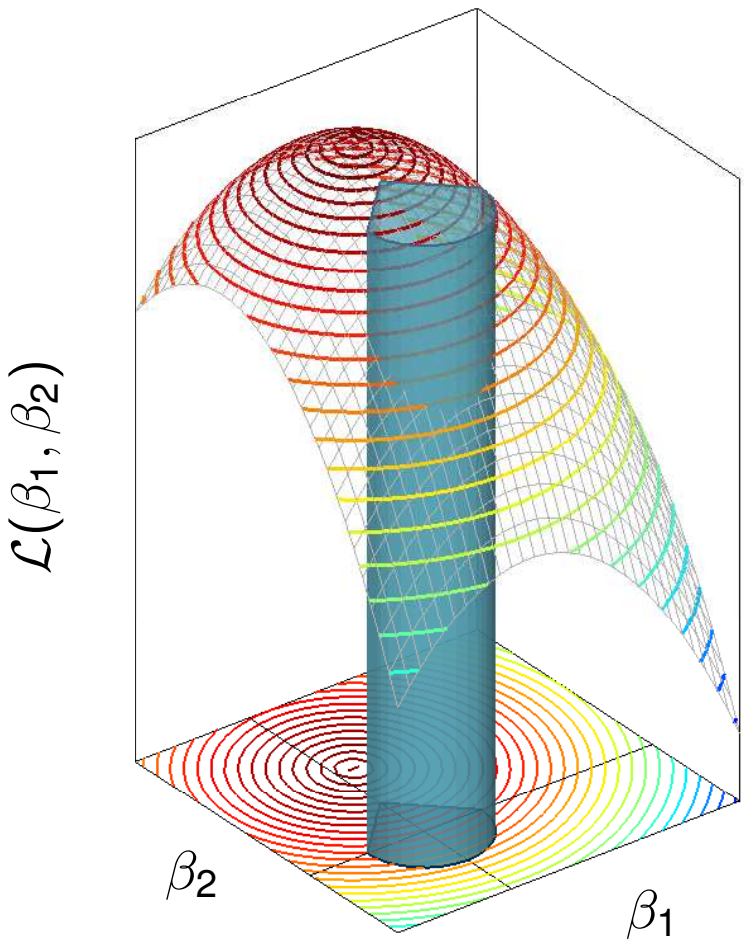


$$\begin{cases} \min\limits_{\beta} & \frac{1}{n}\sum\limits_{i=1}^{n}\ell(f(x_i;\beta),y_i) \\ \text{s. t.} & \sum\limits_{j=1}^{d}|\beta_j\| \leq c \end{cases} \Leftrightarrow \begin{cases} \min\limits_{\beta,s} & \frac{1}{n}\sum\limits_{i=1}^{n}\ell(f(x_i;\beta),y_i) \\ \text{s. t.} & \sum\limits_{j=1}^{d}\frac{\beta_j^2}{s_j} \leq c^2 \\ & \sum_{j=1}^{d} s_j \leq 1 \ , s_j \geq 0 \ \ j=1,\ldots,d \end{cases}$$

Adaptive ridge penalty

Geometric Insight on Sparsity

Constrained Optimization



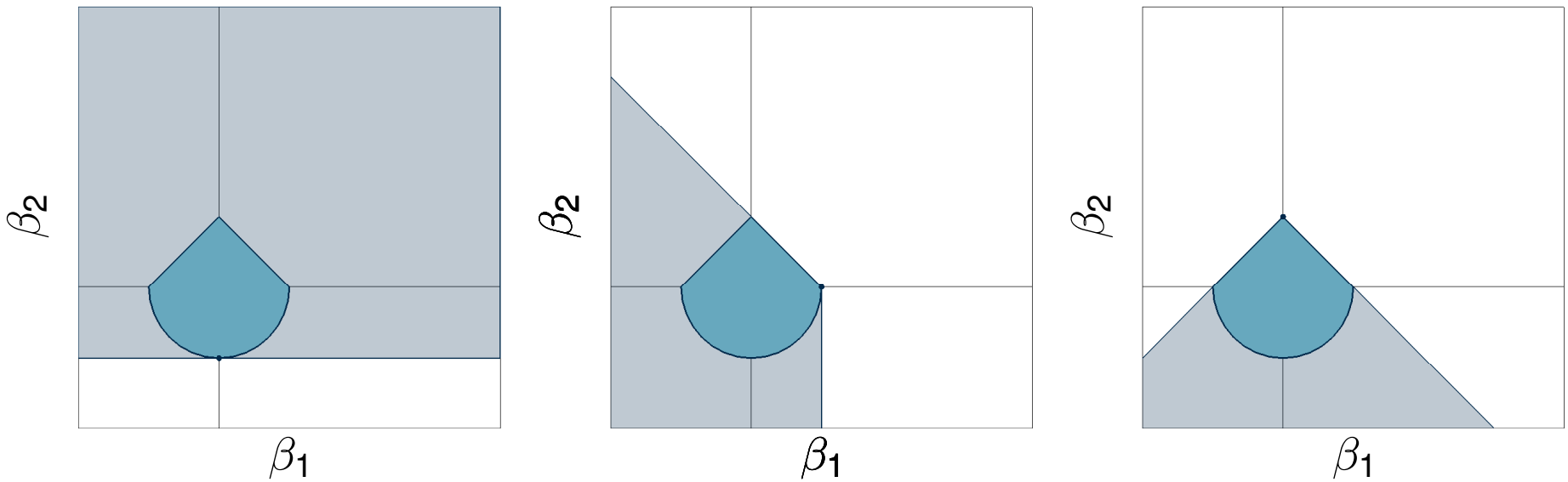
$$\max_{\beta_1,\beta_2} \mathcal{L}(\beta_1, \beta_2) - \lambda\Omega(\beta_1, \beta_2)$$

$$\Leftrightarrow \begin{cases} \max\limits_{\beta_1,\beta_2} & \mathcal{L}(\beta_1, \beta_2) \\ \text{s.t.} & \Omega(\beta_1, \beta_2) \leq c \end{cases}$$

Geometric Insight on Sparsity
Supporting Hyperplane

An hyperplane supports a set iff

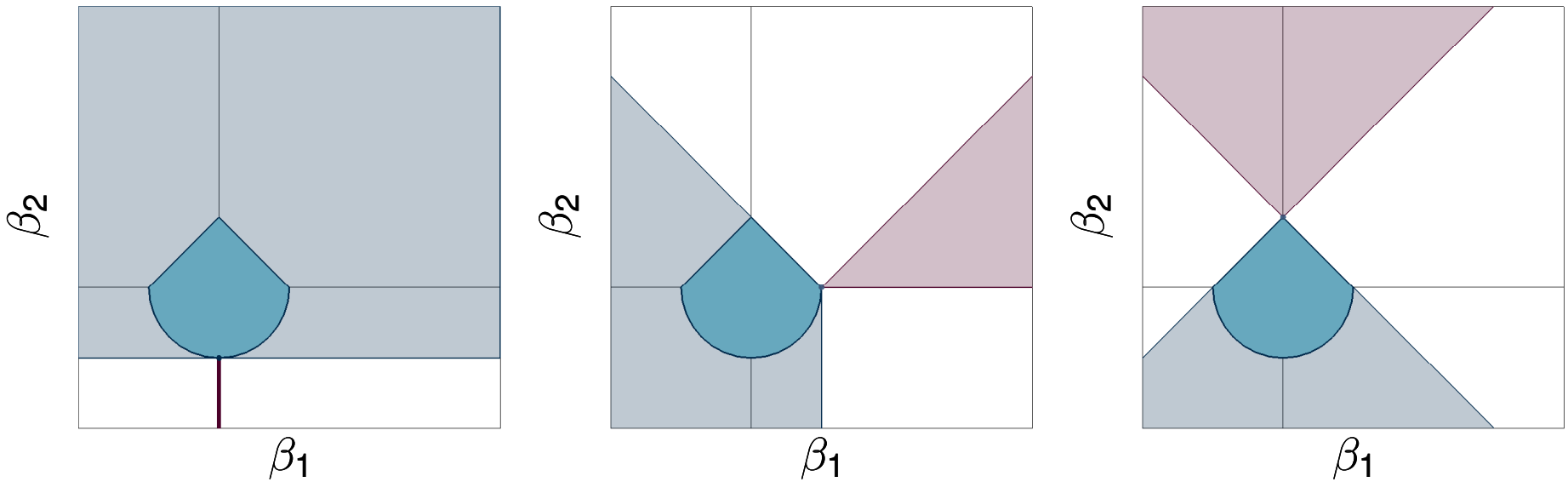
- the set is contained in one half-space
- the set has at least one point on the hyperplane



There are Supporting Hyperplane at all points of convex sets:
Generalize tangents

Geometric Insight on Sparsity
Dual Cone

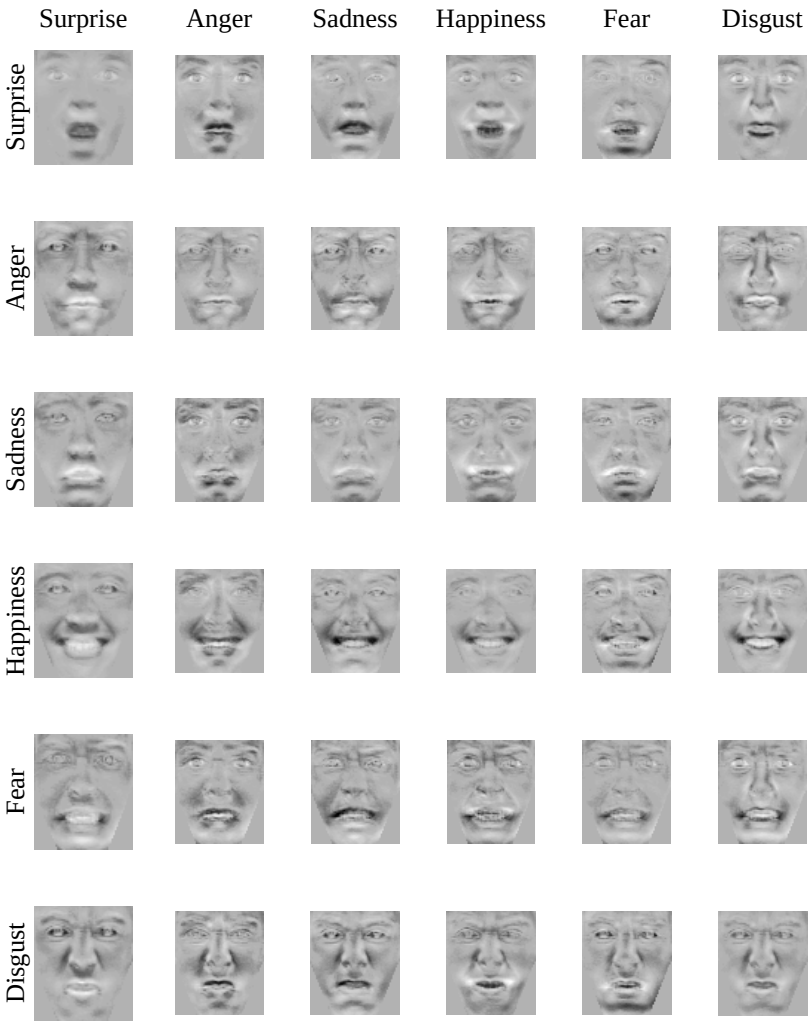
Generalizes normals

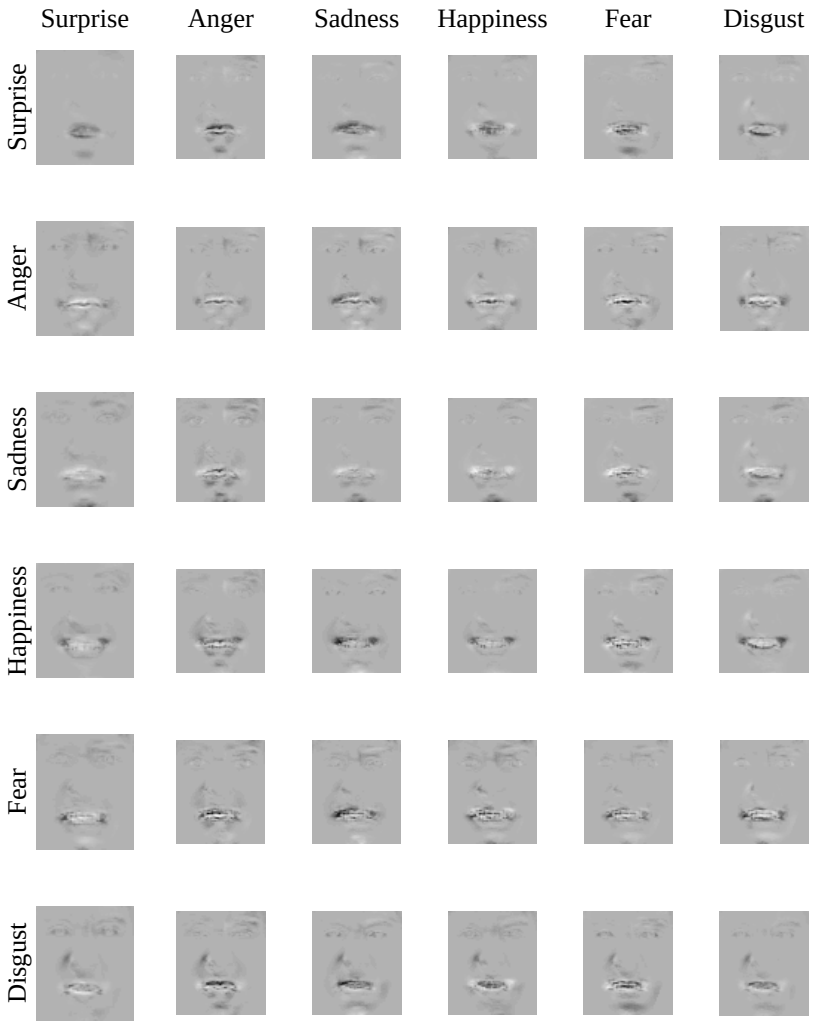


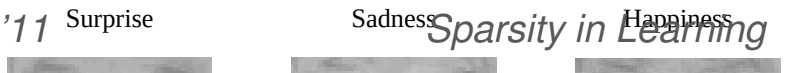
Shape of dual cones $\Rightarrow$ sparsity pattern

Expression Recognition

Logistic Regression

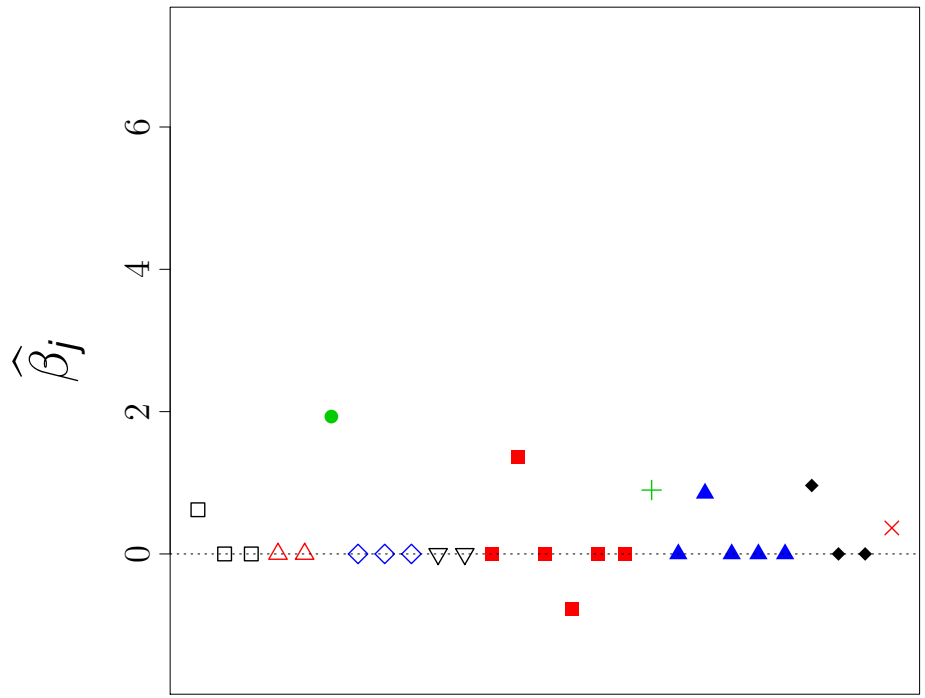






Prediction of Response to Chemotherapy

Logistic Regression

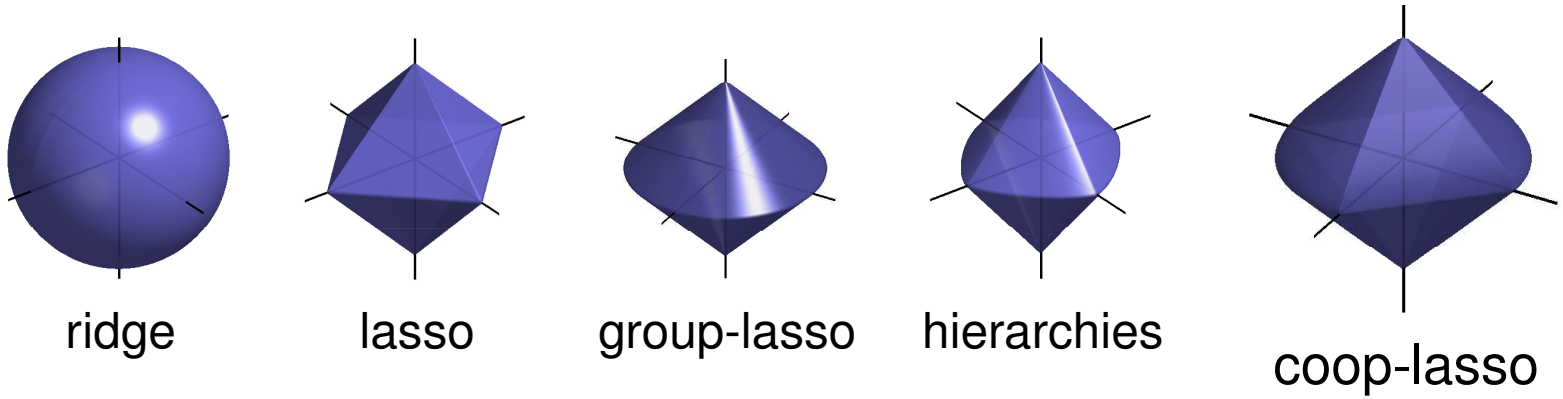


probe sets/genes

No coherent pattern

Ball crafting

Group sparsity



- Additive models (Grandvalet & Canu 1999, Bakin, 1999)
  - Adaptive metric $\Rightarrow$ 1 or 2 hyper-parameters (compared to d)
  - Ease to implementation, interpretability
- Multiple/Composite Kernel Learning (Lanckriet *et al.*, 2004, Szafranski *et al.*, 2010)
  - Adaptive metric: "learn the kernel" $\Rightarrow$ 1 hyper-parameter
  - CKL takes into account a group structure on kernels
- Sign-coherent groups
  - Multi-task learning for pathway inference (Chiquet *et al.*, 2010)
  - Prediction from cooperative features (Chiquet *et al.*, 2011)

Group-Lasso

$$\begin{cases} \min\limits_{\beta} & \frac{1}{n}\sum\limits_{i=1}^{n} \ell(f(x_i;\beta), y_i) \\ \text{s. t.} & \sum\limits_{k=1}^{K} \left(\sum\limits_{j \in \mathcal{G}_k} \beta_j^2 \right)^{1/2} \leq c \end{cases},$$

where $\{\mathcal{G}_k\}_{k=1}^{K}$ forms a partion of $\{1, \ldots, d\}$

Sparse solution groupwise
No sign-coherence

Coop-Lasso

$$\begin{cases} \min\limits_{\beta} & \frac{1}{n}\sum\limits_{i=1}^{n} \ell(f(x_i;\beta), y_i) \\ \text{s. t.} & \sum\limits_{k=1}^{K} \left(\sum\limits_{j\in\mathcal{G}_k} [\beta_j]_+^2 \right)^{1/2} + \left(\sum\limits_{j\in\mathcal{G}_k} [\beta_j]_-^2 \right)^{1/2} \leq c \end{cases},$$

where $\{\mathcal{G}_k\}_{k=1}^K$ forms a partion of $\{1,\dots,d\}$

Sparse solution groupwise
Favors sign-coherence

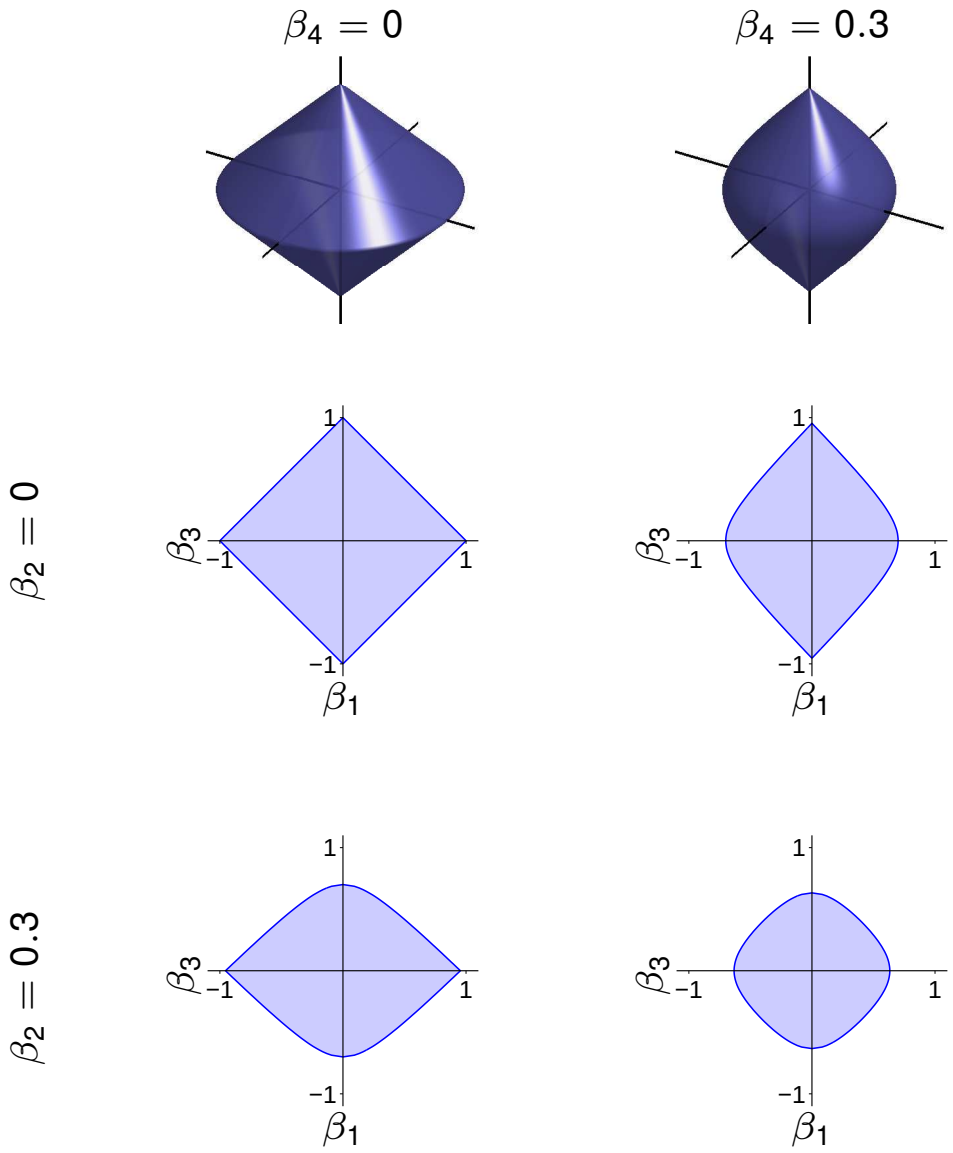
Group-LASSO balls

Admissible set

- $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^\intercal$,
- $\mathcal{G}_1 = \{1, 2\}, \mathcal{G}_2 = \{3, 4\}$.

Unit ball

$$\sum_{k=1}^{2} \left(\sum_{j \in \mathcal{G}_k} \beta_j^2 \right)^{1/2} \leq 1$$



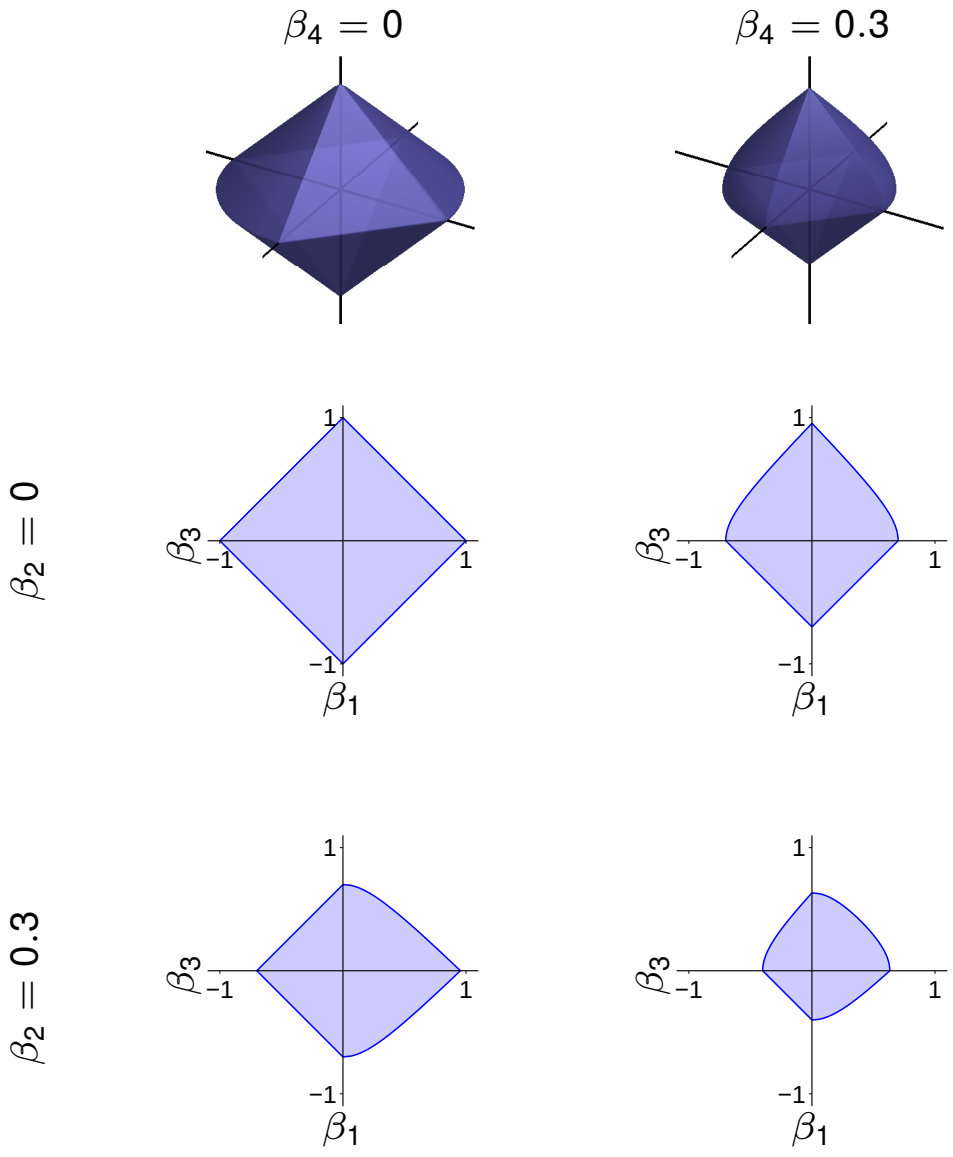
Coop-lasso balls

Admissible set

- $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^\intercal$,
- $\mathcal{G}_1 = \{1, 2\}, \mathcal{G}_2 = \{3, 4\}$.

Unit ball

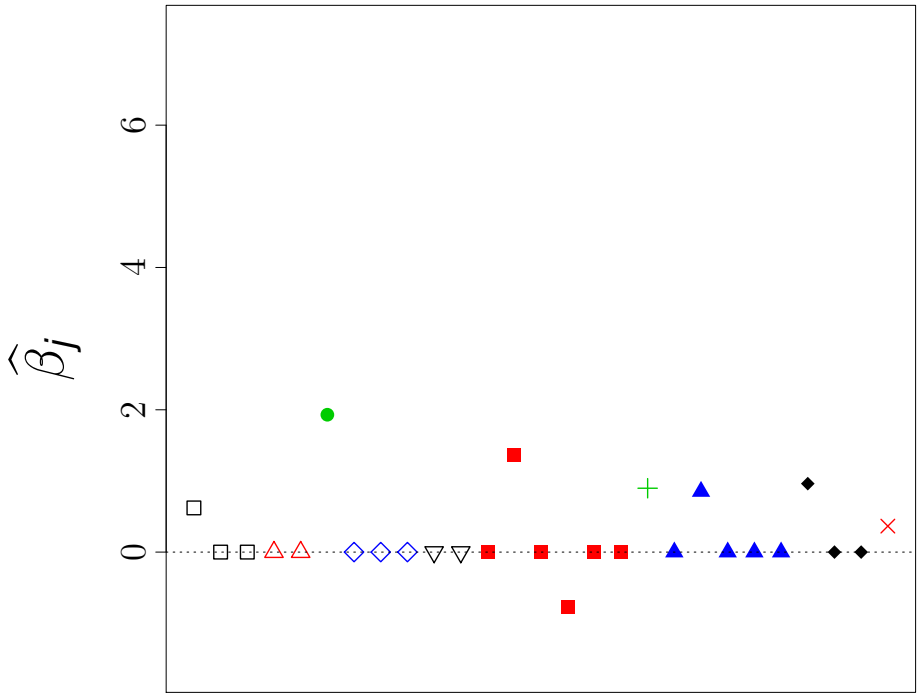
$$\sum_{k=1}^{2} \left(\sum_{j \in \mathcal{G}_k} [\beta_j]_+^2 \right)^{1/2} + \left(\sum_{j \in \mathcal{G}_k} [\beta_j]_-^2 \right)^{1/2} \leq 1$$



Prediction of Response to Chemotherapy

Logistic Regression

Lasso



probe sets/genes
No coherent probe activation

group-Lasso

Why ignoring some examples ?

"There is no data like more data"

All examples convey information about $P(Y|X)$, but not necessarily in the neighborhood of $P(Y = 1|X) = \frac{1}{2}$.

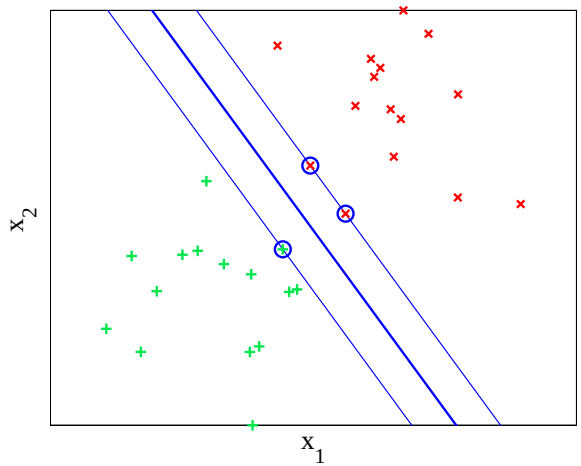
- Learning speed
  - A gradient step is $O(nd)$
  - A second order step is in $O(nd^2 + d^3)$ and requires $O(d^2)$ of memory
  - For kernel methods $n = d \ldots$
- Evaluation speed
  - For kernel methods $O(n)$ per test example

All examples are processed during learning, but some of them may not influence the solution

Support Vector Machines
Separable Case

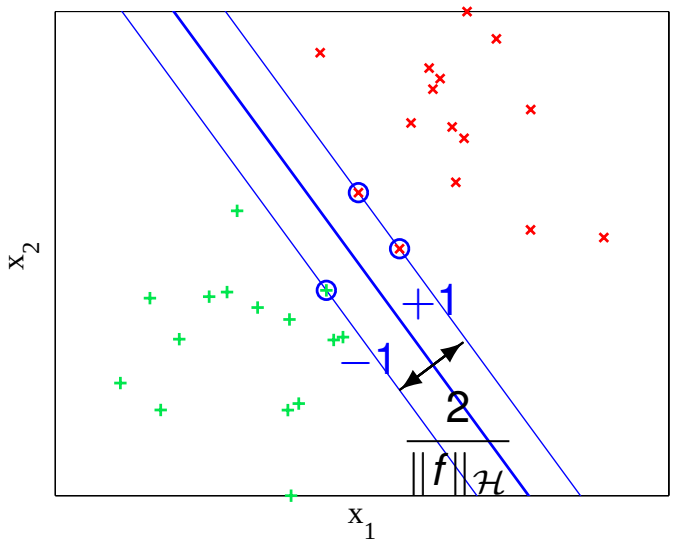
Motivation: separate with minimal capacity

Separating hyperplane $w^t x + b = 0$ with maximal margin *i.e.* maximizing $min_{x_i} |w^t x_i + b|$ with $\|w\|^2 = 1$.



$$f(x) + b = w^t x + b \underset{\text{class}-1}{\overset{\text{class}+1}{\lessgtr}} 0$$

Optimization problem
Separable Case

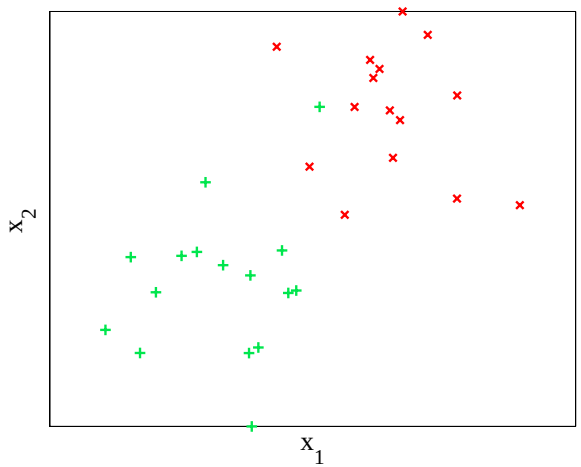


If $y_i(f(x_i)+b) \geq 1$, maximize the margin $\Leftrightarrow$ maximize $\frac{2}{\|f\|_{\mathcal{H}}}$

$$\begin{cases} \min\limits_{f,b} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 \\ \text{s. t.} & y_i(f(x_i)+b) \geq 1 \qquad i = 1,\dots,n \end{cases}$$

Sparse solution: "support examples" at margin

Support Vector Machines
Non Separable Case



$$\begin{cases} \min\limits_{f,b} & \frac{1}{2}\|f\|^2_{\mathcal{H}} \\ \text{s. t.} & y_i(f(x_i)+b) \geq 1 \quad i = 1,\dots,n \end{cases}$$

Empty admissible set $\Rightarrow$ **relaxation**

Relaxation
Soft Margins

1. Relaxation by adding slack variables

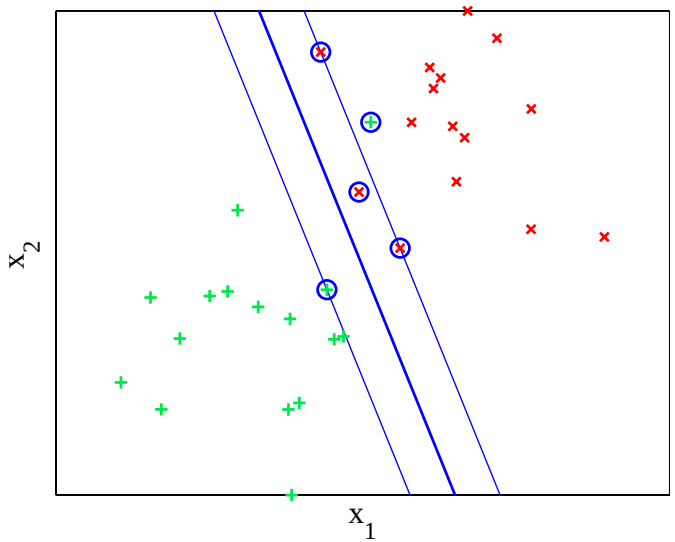
$$\begin{cases} \min\limits_{f,b,\xi} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 \\ \text{s. t.} & y_i(f(x_i)+b) \geq 1-\xi_i \quad i=1,\dots,n \\ & \xi_i \geq 0 \quad i=1,\dots,n \end{cases}$$

2. Loss pertaining to discrepancy

$$\begin{cases} \min\limits_{f,b,\xi} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C\sum\limits_{i=1}^{n}\xi_i^p \\ \text{s. t.} & y_i(f(x_i)+b) \geq 1-\xi_i \quad i=1,\dots,n \\ & \xi_i \geq 0 \quad i=1,\dots,n \end{cases}$$

$p=0 \Rightarrow$ **misclassification** $\Rightarrow$ **NP-hard**
$p \geq 1 \Rightarrow$ **convex problem**
$p=1 \Rightarrow$ **sparse problem: hinge loss**

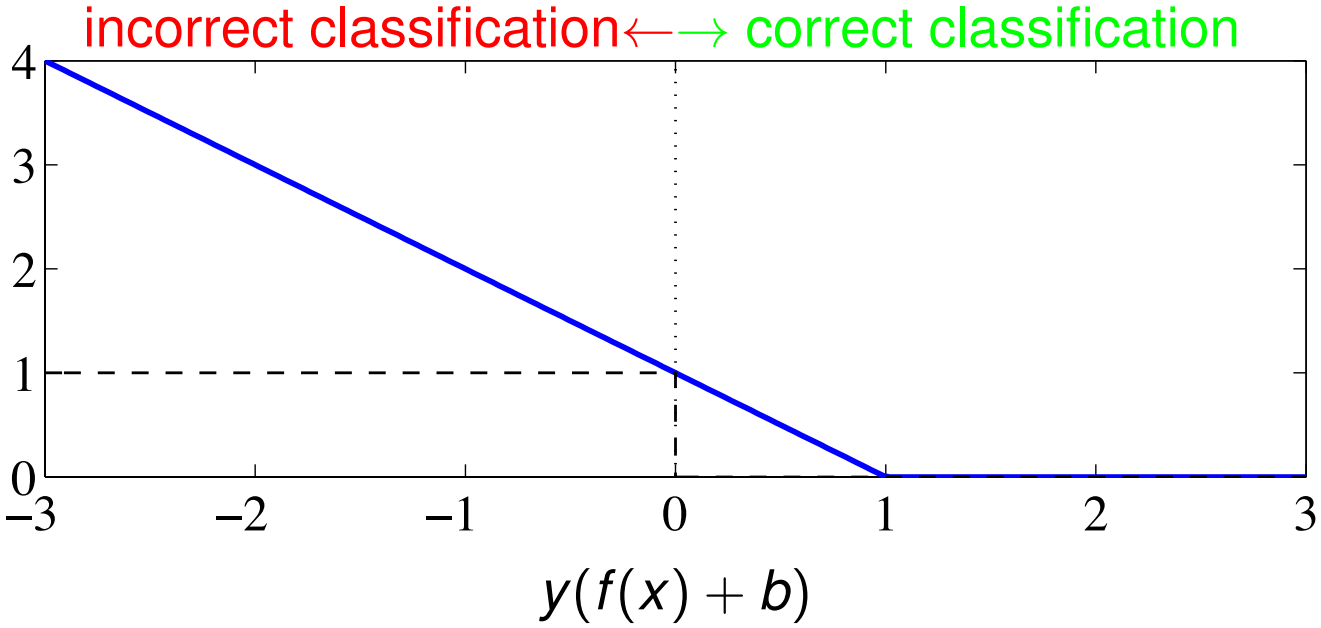
L1 Support Vector Machines



Lagrangian Formulation

$$\min_{f,b} \quad \frac{1}{2}\|f\|_{\mathcal{H}}^2 \quad + C\sum_{i=1}^{n}[1 - y_i(f(x_i) + b)]_+$$

Hinge Loss Function $[1 - y(f(x) + b)]_+$



- upper-bound of misclassification → decision-oriented
- convex and piecewise linear in $f(x)$ → "easy" to optimize
- constant for $y(f(x) + b) > 1$ → sparse

Asymptotically, we only recover $P(Y = 1|X = x)$ at $\frac{1}{2}$ (Bartlett & Tewari, 2004)

Asymptotic sparsity

Limiting sparsity (Steinwart, 2004)

When SVMs are consistant, the ratio of support vectors is

$\mathbb{E}_X\left(2\min(P(Y=1|X), P(Y=-1|X))\right)$ for $\ell(f(x),y) = [1-y(f(x)+b)]_+$

$P_X((0 < (P(Y=1|X) < 1)$ for $\ell(f(x),y) = [1-y(f(x)+b)]_+^2$

Estimation of probabilities (Bartlett & Tewari, 2004)

When SVMs are consistant, one can recover $P(Y=1|X=x)$

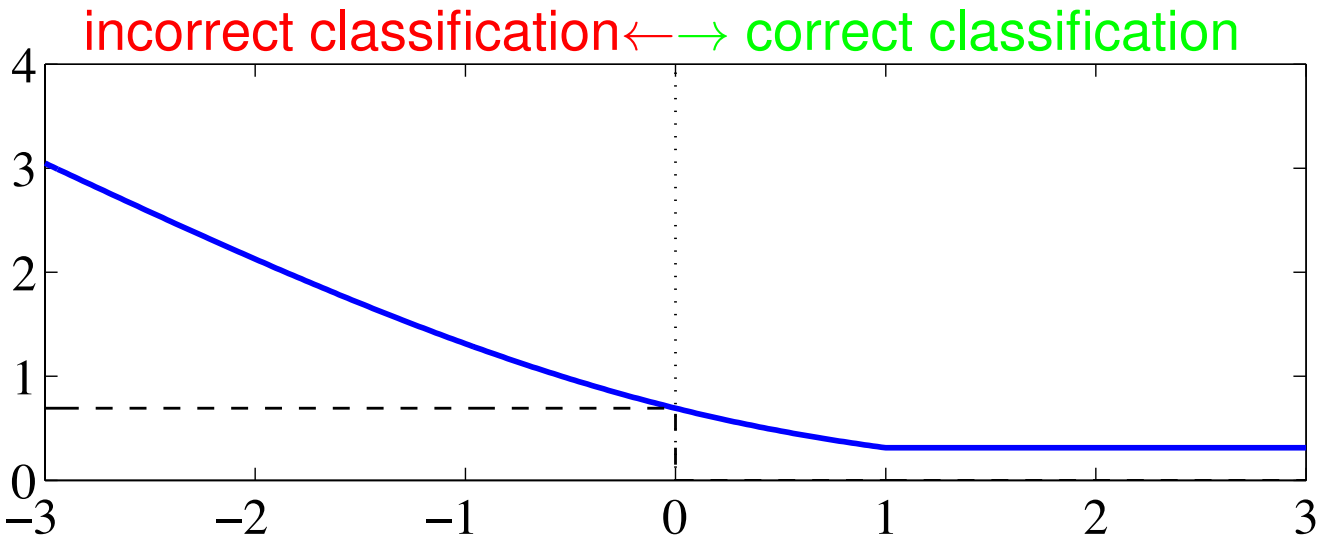
At $P(Y=1|X=x) = \frac{1}{2}$ for $\ell(f(x),y) = [1-y(f(x)+b)]_+$

At $0 < P(Y=1|X=x) < 1$ for $\ell(f(x),y) = [1-y(f(x)+b)]_+^2$

There is a sparsity-estimation range trade-off

Loss function $\max(-\ln(\pi_g), \ln(1 + exp(-y(f(x) + b))))$

Truncated Neg-log-likelihood



$y(f(x) + b)$ (pour $\pi_+ = \pi_-$)

- upper-bound of misclassification $\rightarrow$ decision-oriented
- convex in $f(x)$ $\rightarrow$ "easy" to optimize
- constant for $y(f(x) + b) > 1$ $\rightarrow$ sparse

We recover $P(Y = 1|X = x)$ in $[1 - \pi_-, \pi_+]$

Some losses

- Truncated Neg-log-likelihood (Hérault & Grandvalet, 2007)
  - Estimates $P(Y = 1|X = x)$ in $[\pi_-, \pi_+]$
  - $\Rightarrow$ Estimation of gray zones
  - $\Rightarrow$ Binary classifiers for Muli-class classification
- Asymetric hinge $\Rightarrow$ (Veropoulos *et al.*,1999)
  - $P(Y = 1|X = x)$ at $\{\pi_0\}$
  - $\Rightarrow$ unbalanced classification losses
- Double hinge (Bartlett & Wegkamp, 2008)
  - $P(Y = 1|X = x)$ at $\{\pi_-, \pi_+\}$
  - $\Rightarrow$ Reject option

Conclusions

- Sparsity generating methods are derived similarly for variables and examples
  - Start from a NP-hard problem
  - Relax to the convexity limit
- They are adressed by similar optimization problems whose objective functions are
  - Convex
  - Non-smooth
  - Piecewise linear
- Optimisation algorithms :
  - Active sets
  - Fast Iterative Shrinkage/Threshold Algorithm

A few open questions

- How to index models to enhance model selection?
  - Lagrange parameter (C, λ)
  - Number of non-zero "slack variables" (ξ_i, β_j)
  - Magnitude of parameters ($\|f\|_{\mathcal{H}}$, $\sum_j |\beta_j|$)
  - Fit ($\sum_i \xi_i$, $\sum_i \ell(f(x_i), y_i)$)
- What Is necessary for stability w.r.t. sample perturbations?
  - Prevailing concensus: convex methods are stable, combinatoric are unstable
  - What about non-convex losses/penalties such as Ψ-learning, adaptive Lasso, clipped estimates?
- Structural Risk Minimization...
  1. defines a data-dependent path in $\mathcal{F}$
  2. follows the path or sample the trajectory according to λ
  3. selects a point on the path

  It is a descent algorithm on R_{emp} from a controlled initial point
  Is it possible to characterize the properties of learning algorithms from these trajectories rather than from the whole $\mathcal{F}$?