Sparsity Based Regularization

Lorenzo Rosasco

9.520 Class 11

March 11, 2009

About this class

Goal To introduce sparsity based regularization with emphasis on the problem of variable selection. To discuss its connection to sparse approximation and describe some of the methods designed to solve such problems.

Plan

- sparsity based regularization: finite dimensional case
- introduction
- algorithms
- theory

Sparsity Based Regularization?

- **interpretabilty of the model**: a main goal besides good prediction is detecting the most discriminative information in the data.
- **data driven representation**: one can take a large, redundant set of measurements and then use a data driven selection scheme.
- **compression**: it is often desirable to have parsimonious models, that is models requiring a (possibly very) small number of parameters to be described.

More generally if the target function is sparse enforcing sparsity of the solution *may* be a way to avoid overfitting.

A Useful Example

Biomarker Identification

Set up:

- n patients belonging to 2 groups (say two different diseases)
- p measurements for *each* patient quantifying the expression of p genes

Goal:

- learn a classification rule to predict occurrence of the disease for future patients
- detect which are the genes responsible for the disease

$p \gg n$ paradigm

typically n is in the order of tens and p of thousands....

Some Notation

Measurement matrix

Let X be the $n \times p$ measurements matrix.

$$X = \begin{pmatrix} x_1^1 & \dots & \dots & \dots & x_1^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^1 & \dots & \dots & \dots & x_n^p \end{pmatrix}$$

- n is the number of examples
- p is the number of variables
- we denote with X^j, $j = 1, \dots, p$ the columns of X

For each patient we have a response (output) $y \in R$ or $y = \pm 1$.
In particular we are given the responses for the training set

$$Y = (y_1, y_2, \dots, y_n)$$

Approaches to Variable Selection

So far we still have to define what are "*relevant*" variables. Different approaches are based on different way to specify what is relevant.

- Filters methods.
- Wrappers.
- Embedded methods.

We will focus on the latter class of methods.

(see "Introduction to variable and features selection" Guyon and Elisseeff '03)

Embedded Methods

The selection procedure is **embedded** in the training phase.

An intuition

what happens to the generalization properties of empirical risk minimization as we discard variables?

- if we keep all the variables we probably overfit,
- if we take just a few variables we are likely to oversmooth (in the limit we have a single variable classifier).

We are going to discuss this class of methods in detail.

Sparse Linear Model

Suppose the output is a linear combination of the variables

$$f(x) = \sum_{i=1}^{p} \beta^i x^i = \langle \beta, x \rangle$$

each coefficient β^i can be seen as a weight on the i-th variable.

Sparsity

We say that a function is *sparse* if most coefficients in the above expansion are zero.

Solving a BIG linear system

In vector notation we can can write the problem as a linear system of equation

$$Y = X\beta.$$

The problem is *ill-posed*.

Sparsity

Define the ℓ_0-norm (not a real norm) as

$$\|\beta\|_0 = \#\{i = 1, \ldots, p \mid \beta^i \neq 0\}$$

It is a measure of how "complex" is f and of how many variables are important.

ℓ_0 Regularization

If we assume that a few variables are meaningful we can look for

$$\min_{\beta \in \mathbb{R}^P} \{ \frac{1}{n} \sum_{j=1}^{n} V(y_j, \langle \beta, x_j \rangle) + \lambda \|\beta\|_0 \}$$

Best subset selection is hard!!

$\Rightarrow$ This is as difficult as trying all possible subsets of variables.

Can we find meaningful approximations?

Approximate solutions

Two main approaches

Approximations exist for various loss functions usually based on:

1. Convex relaxation.
2. Greedy schemes.

We mostly discuss the first class of methods (and consider the square loss).

Convex Relaxation

A natural approximation to ℓ_0 regularization is given by:

$$\frac{1}{n}\sum_{j=1}^{n} V(y_j, \langle \beta, x_j \rangle) + \lambda \|\beta\|_1$$

where $\|\beta\|_1 = \sum_{i=1}^{p} |\beta^i|$.

If we choose the square loss

$$\frac{1}{n}\sum_{j=1}^{n}(y_j - \langle \beta, x_j \rangle)^2 = \|Y - X\beta\|_n^2$$

such a scheme is called **Basis Pursuit** or **Lasso** algorithms.

What is the difference with Tikhonov regularization?

- We have seen that Tikhonov regularization is a good way to avoid overfitting.
- Lasso provides sparse solution Tikhonov regularization doesn't.

Why?

Tikhonov Regularization

Of course we can go back to:

$$\min_{\beta \in \mathbb{R}^p} \{ \frac{1}{n} \sum_{j=1}^{n} V(y_j, \langle \beta, x_j \rangle) + \lambda \sum_{i=1}^{p} |\beta^i|^2 \}$$

How about sparsity?

$\Rightarrow$ in general all the β^i in the solution will be different from zero.

Constrained Minimization

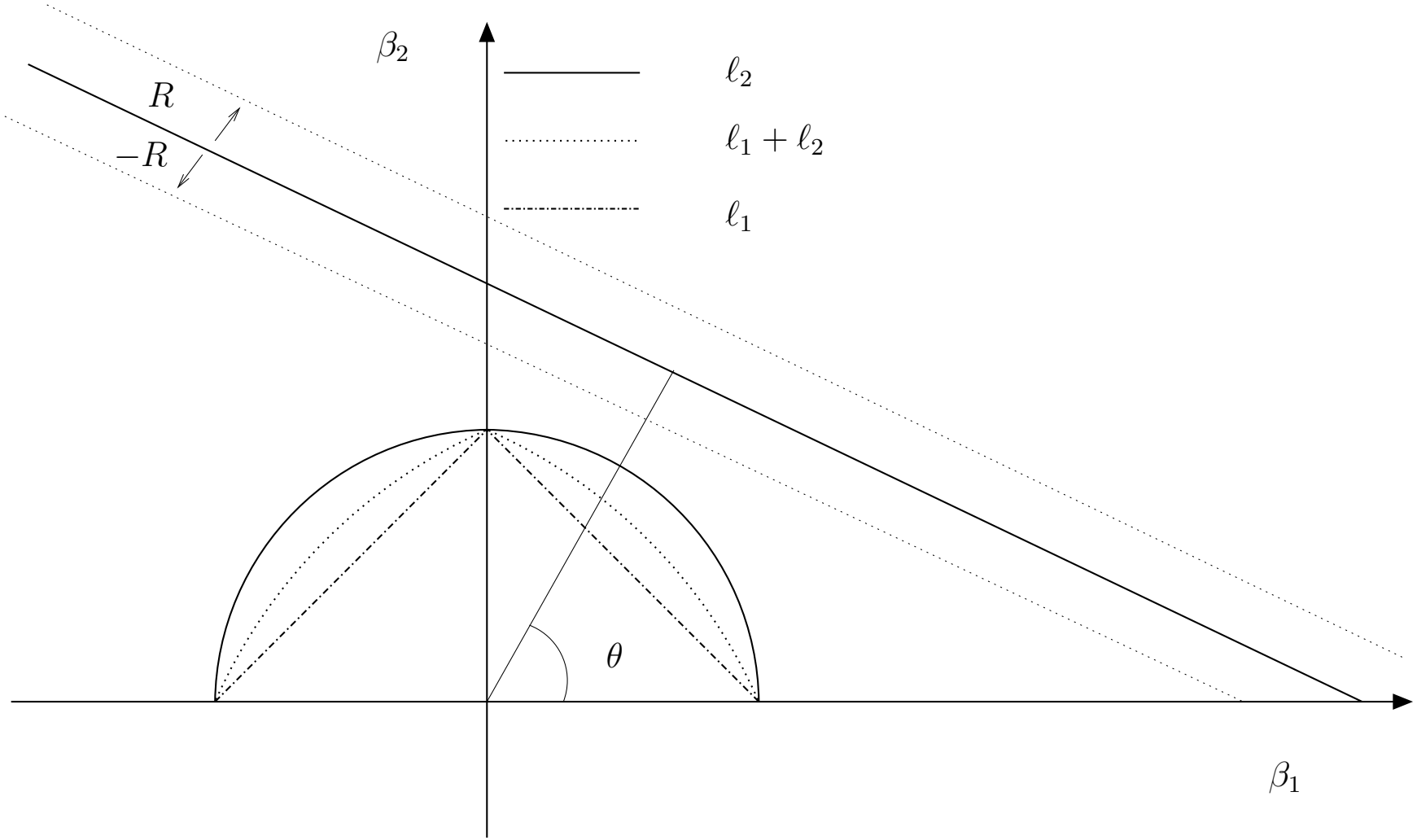
Consider

$$\min_{\beta}\{\sum_{i=1}^{p}|\beta^i|.\}$$

subject to

$$\|Y - X\beta\|_n^2 \leq R.$$

Geometry of the Problem



Back to ℓ_1 regularization

We focus on the square loss so that we now have to solve

$$\min_{\beta \in \mathbb{R}^p} \|Y - \beta X\|^2 + \lambda \|\beta\|_1 .$$

- Though the problem is no longer hopeless it is nonlinear.
- The functional is convex but not *strictly* convex, so that the solution is not unique.
- One possible approach relies on linear (or quadratic) programming techniques.
- Using convex analysis tools we can get a simple iterative algorithms.

An Iterative Thresholding Algorithm

It can be proved that the following iterative algorithm converges to the solution β^λ of ℓ_1 regularization as the number of iteration increases.

Set $\beta_0^\lambda = 0$ and let

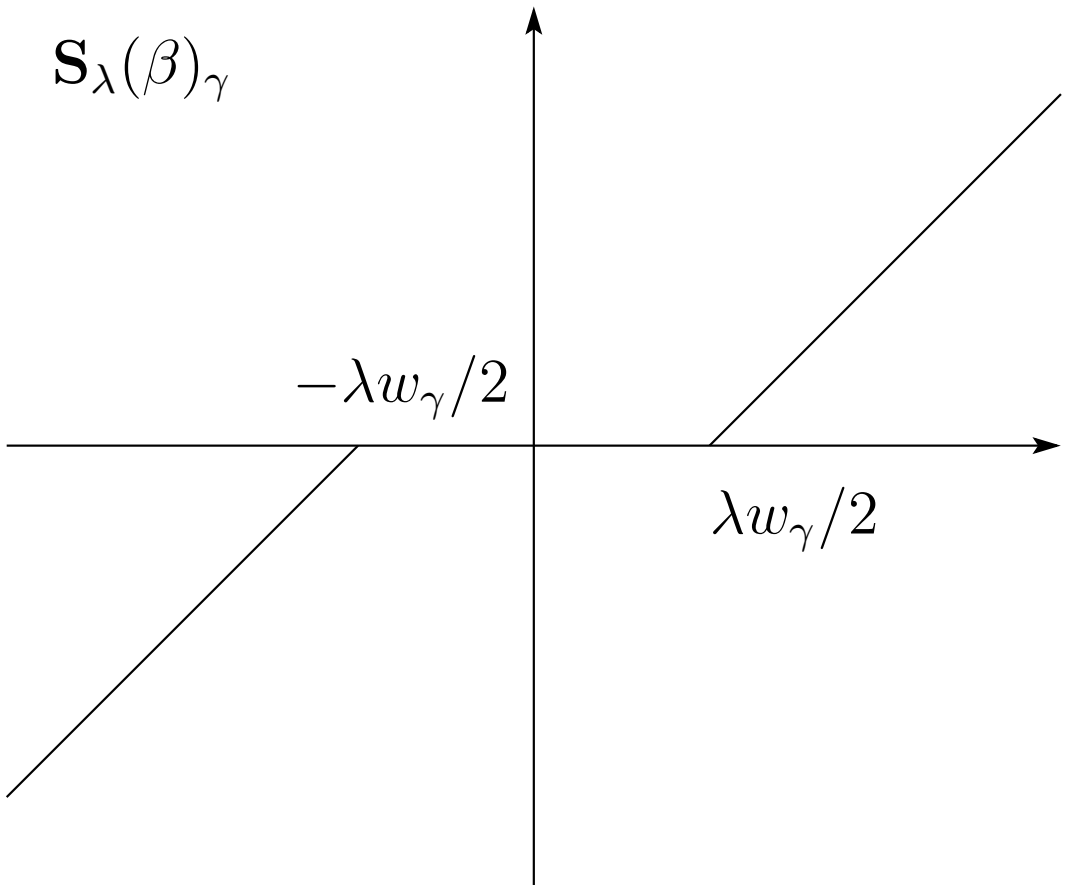
$$\beta_t^\lambda = S_\lambda[\beta_{t-1}^\lambda + \tau X^T(Y - X\beta_{t-1}^\lambda)]$$

where τ is a normalization constant ensuring $\tau \|X\| \leq 1$ and the map S_λ is defined component-wise as

$$S_\lambda(\beta^i) = \begin{cases} \beta^i + \lambda/2 & \text{if } \beta^i < -\lambda/2 \\ 0 & \text{if } |\beta^i| \leq \lambda/2 \\ \beta^i - \lambda/2 & \text{if } \beta^i < \lambda/2 \end{cases}$$

(see Daubechies et al.'05)

Thresholding Function



Algorithmics Aspects

Set $\beta_0^\lambda = 0$
for t=1:tmax

$$\beta_t^\lambda = S_\lambda[\beta_{t-1}^\lambda + \tau X^T(Y - X\beta_{t-1}^\lambda)]$$

- The algorithm we just described is very easy to implement but can be quite heavy from a computational point of view.
- The number of iteration t can be stopped when a certain precision is reached.
- The complexity of the algorithm is $O(tp^2)$ for each value of the regularization parameter.
- The regularization parameter controls the degree of sparsity of the solution.

Some Remarks

- **About Uniqueness**: the solution of ℓ_1 regularization is not unique. Note that the various solution have the **same prediction properties** but **different selection properties**.
- **Correlated Variables**: If we have a group of correlated variables the algorithm is going to select just one of them. This can be bad for interpretability but maybe good for compression.

ℓ_q regularization?

Consider a more general penalty of the form

$$\|\beta\|_q = (\sum_{i=1}^{p} |\beta^i|^q)^{1/q}$$

(called bridge regression in statistics).
It can be proved that:

- $\lim_{q \to 0} \|\beta\|_q \to \|\beta\|_0$,
- for $0 < q < 1$ the norm is **not** a convex map,
- for $q = 1$ the norm **is** a convex map and is **strictly** convex for $q > 1$.

Elastic Net Regularization

One possible way to cope with the previous problems is to consider

$$\min_{\beta \in \mathbb{R}^p} \|Y - \beta X\|^2 + \lambda(\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2).$$

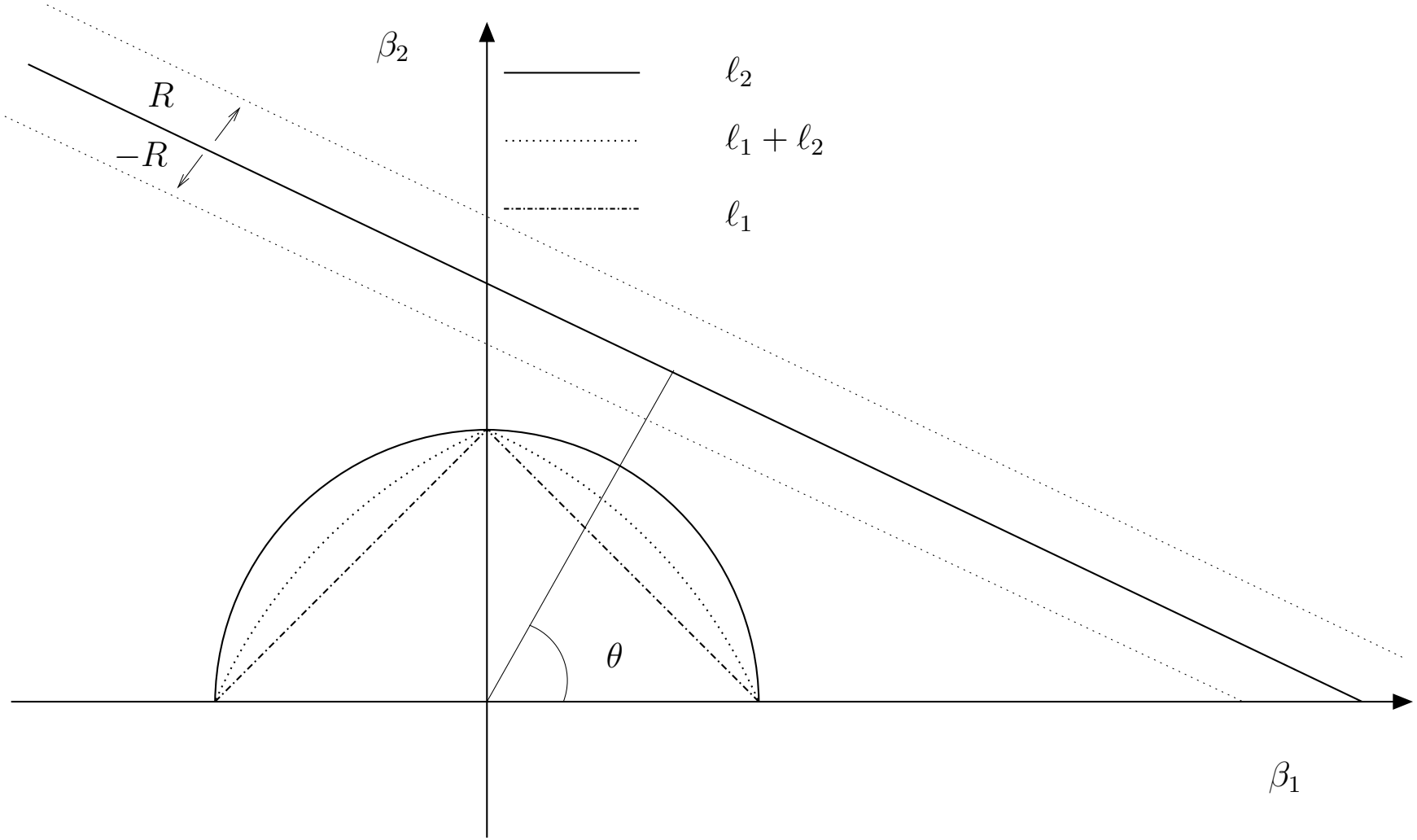
- λ is the regularization parameter.
- α controls the amount of sparsity and smoothness.

(Zhu. Hastie '05; De Mol, De Vito, Rosasco '07)

Elastic Net Regularization (cont.)

- The ℓ_1 term promotes sparsity and the ℓ_2 term smoothness.
- The functional is strictly convex: the solution is unique.
- A whole group of correlated variables is selected rather than just one variable in the group.

Geometry of the Problem



What about Theory?

The same kind of approaches were considered in different domains for different (but related) purposes.

- Machine Learning.
- Fixed design regression.
- Compressed sensing.

Similar theoretical questions but different settings (deterministic vs stochastics, random design vs fixed design).

What about Theory? (cont.)

Roughly speaking, the results in fixed design regression prove that:
If $Y = X\beta^* + \xi$, where X is an n by p matrix, $\xi \sim \mathcal{N}(0, \sigma^2 I)$, β^* has at most s non zero coefficients and $s \leq n/2$ then

$$\left\| \beta^\lambda - \beta^* \right\| \leq Cs\sigma^2 \log p.$$