MIT 9.520/6.860, Fall 2018
Statistical Learning Theory and Applications

Class 08: Sparsity Based Regularization

Lorenzo Rosasco

Learning algorithms so far

ERM + explicit ℓ^2 penalty

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \|w\|^2.$$

- Implicit regularization by optimization.
- Regularization with projections/sketching.
- Non linear extension with features/kernels.

What about other norms/penalties?

Sparsity

The function of interest depends on **few building blocks**

Why sparsity?

- Interpretability
- High dimensional statistics, $n \ll d$
- Compression

What is sparsity?

$$f(x) = \sum_{j=1}^{d} x_j w_j$$

Sparse coefficients: few $w_j \neq 0$

Sparsity and dictionaries

More generally consider

$$f(x) = \sum_{j=1}^{p} \phi_j(x) w_j$$

with $\phi_1, \dots, \phi_p$ **dictionary.**

Sparsity and dictionaries (cont.)

The concept of sparsity **depends** on the considered dictionary.

If we let $(\phi_j)_j,(\psi_j)_j$ two dictionaries of lin. indip. features such that

$$f(x) = \sum_j \phi_j \beta_j = \sum_j \psi_j b_j,$$

then $\|f\| = \|\beta\| = \|b\|$.

However, sparsity on $(\phi_j)_j,(\psi_j)_j$ can be very different!

Linear

We stick to linear functions for sake of simplicity.

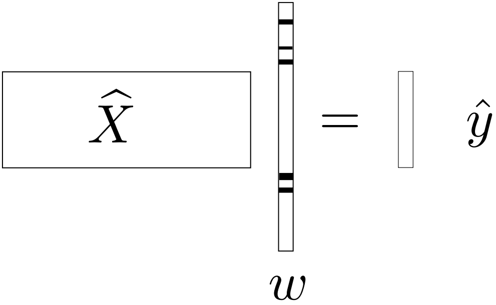
$$f(x) = \sum_{j=1}^{d} x_j w_j.$$

Given data, consider the linear system

$$\widehat{X} w = \widehat{Y}.$$

Linear systems with sparsity

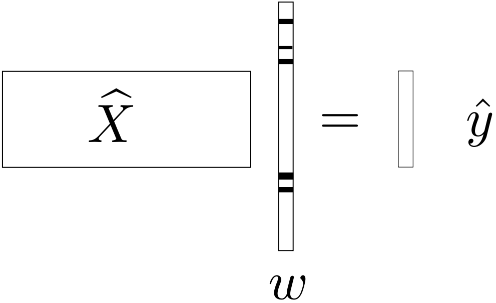
$n \ll d$



There is a solution with $s \ll d$ *non zero entries in unknown locations.*

Best subset selection

Solve for *all* possible columns subsets.



Aka torturing the data until they confess.

Sparse regularization

Best subset selection is equivalent to

$$\min_{w \in \mathbb{R}^d} \cancel{\|w\|} \to {\color{red}\|w\|_0}, \quad \text{subj. to} \quad \widehat{X}w = \widehat{Y},$$

or

$$\min_{w \in \mathbb{R}^d} \frac{1}{n}\|\widehat{X}w - \widehat{Y}\|^2 + \lambda \cancel{\|w\|^2} \to {\color{red}\|w\|_0}$$

ℓ_0-norm

$$\|w\|_0 = \sum_{j=1}^{d} \mathbf{1}_{\{w_j \neq 0\}}$$

Best subset selection

$$\min_{w \in \mathbb{R}^d} \|w\|_0, \qquad \text{subj. to} \qquad \widehat{X}w = \widehat{Y},$$

The problem is combinatorially hard.

Approximate approaches include:

1. Greedy methods.

2. Convex relaxations.

Greedy methods

Initalize, then

- select a variable.
- Compute solution.
- Update.
- Repeat.

Matching pursuit

$$r_0 = \widehat{Y}, \quad w_0 = 0, \quad I_0 = \emptyset$$

for $i = 1$ to T

- Let $\widehat{X}_j = \widehat{X} e_j$, and select $j \in \{1, \dots, d\}$ maximizing [1]

$$a_j = v_j^2 \|\widehat{X}_j\|^2 \quad \text{with} \quad v_j = \frac{r_{i-1}^\top \widehat{X}_j}{\|\widehat{X}_j\|^2},$$

- $I_i = I_{i-1} \cup \{j\}$,
- $w_i = w_{i-1} + v_j e_j$
- $r_i = r_{i-1} - \widehat{X}_j v_j = \widehat{Y} - \widehat{X} w_i$

---

[1] Note that

$$v_j = \arg\min_{v \in \mathbb{R}} \|\hat{X}_j v - r_{i-1}\|^2, \quad \text{and,} \quad \|\hat{X}_j v_j - r_{i-1}\|^2 = \|r_{i-1}\| - a_j$$

Orthogonal Matching pursuit

$$r_0 = \widehat{Y}, \quad w_0 = 0, \quad I_0 = \emptyset$$

for $i = 1$ to T

- Let $\widehat{X}_j = \widehat{X} e_j$, and select $j \in \{1, \dots, d\}$ maximizing

$$a_j = v_j^2 \|\widehat{X}_j\|^2 \quad \text{with} \quad v_j = \frac{r_{i-1}^\top \widehat{X}_j}{\|\widehat{X}_j\|^2},$$

- $I_i = I_{i-1} \cup \{j\}$,
- $w_i = \arg\min_w \|\widehat{X} M_{I_i} w - \widehat{Y}\|^2$, where $(M_{I_i} w)_j = \delta_{j \in I_i} w_j$
- $r_i = \widehat{Y} - \widehat{X} w_i$

Convex relaxation

Lasso (statistics) or Basis Pursuit (signal processing)

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{X}w - \widehat{Y}\|^2 + \lambda \|w\|^2 \rightarrow \|w\|_1$$

ℓ_1-norm

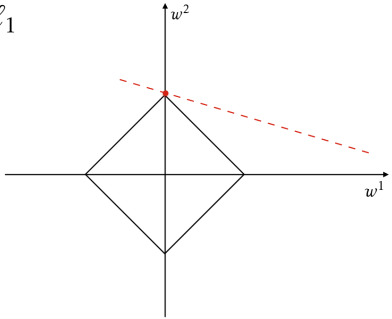
$$\|w\|_1 = \sum_{i=1}^{d} |w_i|.$$

Next, we discuss modeling + optimization aspects.

The geometry of sparsity

ℓ_1

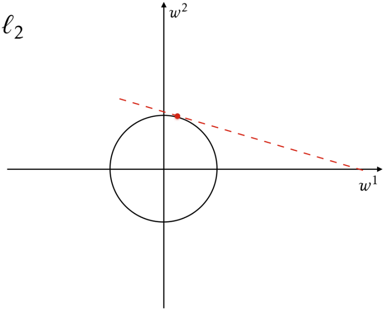
$$\min \|w\|_1, \quad \text{s.t. } \widehat{X}w = \widehat{Y}$$



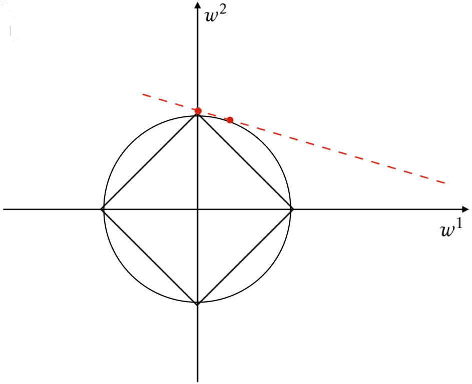
Ridge regression and sparsity

ℓ_2

Replace $\|w\|_1$ with $\|w\|$?



ℓ_1 vs ℓ_2



Unlike ridge-regression, ℓ_1 regularization leads to sparsity!

Optimization for sparse regularization

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{X} w - \widehat{Y}\|^2 + \lambda \|w\|_1$$

- Convex but not smooth

Optimization

- Could be solved via the subgradient method
- Objective function is composite

$$\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{n}\|\widehat{X}w - \widehat{Y}\|^2}_{\text{convex smooth}} + \lambda \underbrace{\|w\|_1}_{\text{convex}}$$

Proximal methods

$$\min_{w \in \mathbb{R}^d} E(w) + R(w)$$

Let

$$\mathrm{Prox}_R(w) = \min_{v \in \mathbb{R}^d} \frac{1}{2}\|v - w\|^2 + R(v)$$

and, for $w_0 = 0$

$$w_t = \mathrm{Prox}_{\gamma R}(w_{t-1} - \gamma \nabla E(w_{t-1}))$$

Proximal Methods (cont.)

$$\min_{w \in \mathbb{R}^d} E(w) + R(w)$$

Let $R : \mathbb{R}^p \to \mathbb{R}$ convex continuous and $E : \mathbb{R}^p \to \mathbb{R}$ differentiable, convex and such that

$$\|\nabla E(w) - \nabla E(w')\| \le L\|w - w'\|$$

(e.g. $\sup_w \| \underbrace{H(w)}_{\text{hessian}} \| \le L$), Then for $\gamma = 1/L$,

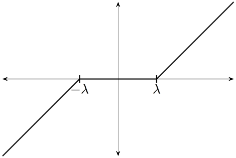
$$w_t = \text{Prox}_{\gamma R}(w_{t-1} - \gamma \nabla E(w_{t-1}))$$

converges to a minimizer of $E + R$.

Soft thresholding

$$R(w) = \lambda \|w\|_1$$

$$(\mathrm{Prox}_{\lambda\|\cdot\|_1}(w))_j = \begin{cases} w_j - \lambda & w_j > \lambda \\ 0 & w_j \in [-\lambda, \lambda] \\ w_j + \lambda & w_j < -\lambda \end{cases}$$



ISTA

$$w_{t+1} = \text{Prox}_{\gamma\lambda\|\cdot\|_1}(w_t - \frac{\gamma}{n}\widehat{X}^\top(\widehat{X}w_t - \widehat{Y}))$$

$$(\text{Prox}_{\gamma\lambda\|\cdot\|_1}(w))^j = \begin{cases} w^j - \gamma\lambda & w^j > \gamma\lambda \\ 0 & w^j \in [-\gamma\lambda, \gamma\lambda] \\ w^j + \gamma\lambda & w^j < -\gamma\lambda \end{cases}$$

Small coefficients are set to zero!

Back to inverse problems

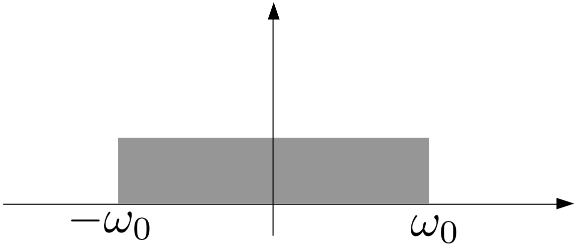
$$\widehat{X}w = \widehat{Y}$$

If x_i are i.i.d. gaussian vectors, $\|w\|_0 = s$ and

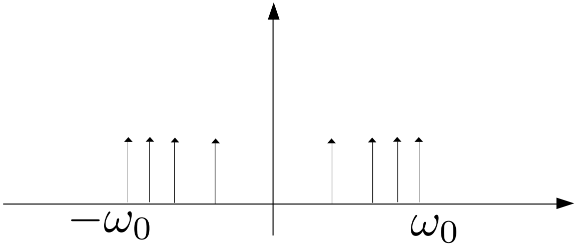
$$n \geq 2s \log \frac{d}{s}$$

then ℓ_1 regularization recovers w with high probability.

Sampling theorem



Classically $2\omega_0$ samples needed



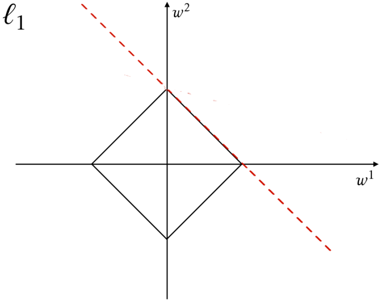
LASSO

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{X} w - \widehat{Y}\|^2 + \lambda \|w\|_1$$

- Interpretability: variable selection!

Variable selection and correlation

$$\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{n}\|\widehat{X}w - \widehat{Y}\|^2 + \lambda\|w\|_1}_{\text{~~strictly~~ convex}}$$

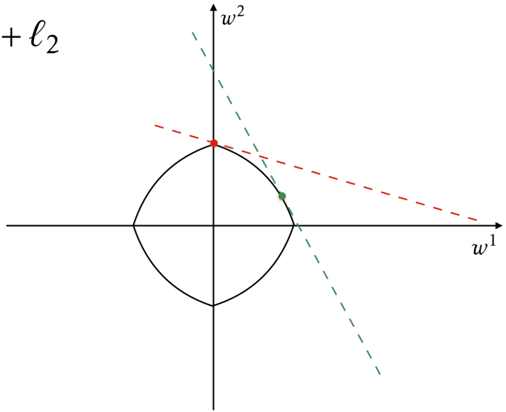


Cannot handle correlations between the variables

Elastic net regularization

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{X}w - \widehat{Y}\|^2 \; + \; \lambda(\alpha\|w\|_1 + (1-\alpha)\|w\|^2)$$

$\ell_1 + \ell_2$



ISTA for elastic net

$$w_{t+1} = \text{Prox}_{\gamma\lambda\alpha\|\cdot\|_1}(w_t - \gamma \frac{2}{n}\widehat{X}^\top(\widehat{X}w_t - \widehat{Y}) - \gamma\lambda(1-\alpha)w_{t-1})$$

$$(\text{Prox}_{\gamma\lambda\alpha\|\cdot\|_1}(w))^j = \begin{cases} w^j - \gamma\lambda\alpha & w^j > \gamma\lambda\alpha \\ 0 & w^j \in [-\gamma\lambda\alpha, \gamma\lambda\alpha] \\ w^j + \gamma\lambda\alpha & w^j < -\gamma\lambda\alpha \end{cases}$$

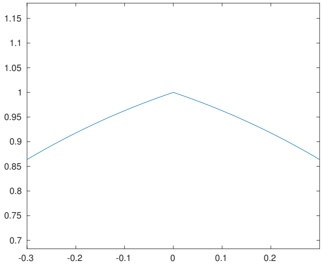
Small coefficients are set to zero!

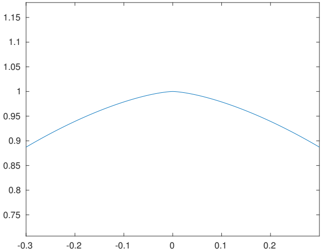
Grouping effect

Strong convexity

$\Longrightarrow$ All relevant (possibly correlated) variables are selected

Elastic net and ℓ_p norms





$$\frac{1}{2}\|w\|_1 + \frac{1}{2}\|w\|^2 = 1$$

$$(\sum_{j=1}^{d} |w_j|^p)^{1/p} = 1$$

ℓ_p norms are similar to elastic net but they are smooth (no “kink”!)

Summary

- Sparsity
- Geometry
- Computations
- Variable selection and elastic net