RegML 2016
Class 5
Sparsity based regularization

Lorenzo Rosasco
UNIGE-MIT-IIT

June 30, 2016

Learning from data

Possible only under assumptions $\rightarrow$ regularization

$$\min_{w} \widehat{\mathcal{E}}(w) + \lambda R(w)$$

- Smoothness
- **Sparsity**

Sparsity

$\boxed{\text{The function of interest depends on } \textbf{few building blocks}}$

Why sparsity

- Interpretability
- High dimensional statistics
- Compression

What is sparsity?

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

Sparse coefficients: few $w_j \neq 0$

Sparsity and dictionaries

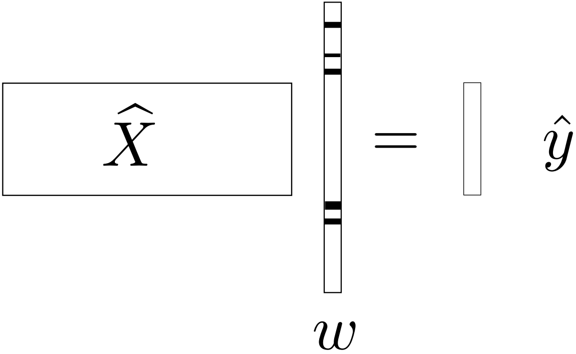
More generally consider

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

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

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

Linear inverse problem



$n < d$ more variables than observations

Sparse regularization

$$\min_w \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda \cancel{\|w\|_2^2} \longrightarrow \|w\|_0$$

ℓ_0-norm

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

Best subset selection

Best subset selection

$$\min_w \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda\|w\|_0$$

as hard as trying all possible subsets...

Best subset selection

$$\min_{w} \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda\|w\|_0$$

as hard as trying all possible subsets...

1. Greedy methods
2. Convex relaxations

Greedy methods

Initalize, then

- Select a variable
- Compute solution
- Update
- Repeat

Matching pursuit

$$r_0 = \hat{y}, \quad w_0 = 0, \quad I_0 = \emptyset$$

for $i = 1$ to T

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

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

---

[1] Note that

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

Matching pursuit

$$r_0 = \hat{y}, \quad w_0 = 0, \quad I_0 = \emptyset$$

for $i = 1$ to T

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

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

- $I_i = I_{i-1} \cup \{j\}$,

[1] Note that

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

Matching pursuit

$$r_0 = \hat{y}, \quad w_0 = 0, \quad I_0 = \emptyset$$

for $i = 1$ to T

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

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

- $I_i = I_{i-1} \cup \{j\}$,
- $w_i = w_{i-1} + v_j e_j$

[1] Note that

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

Matching pursuit

$$r_0 = \hat{y}, \quad w_0 = 0, \quad I_0 = \emptyset$$

for $i = 1$ to T

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

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

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

[1] Note that

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

Orthogonal Matching pursuit

$$r_0 = \hat{y}, \quad w_0 = 0, \quad I_0 = \emptyset$$

for $i = 1$ to T

- Select $j \in \{1, \dots, d\}$ which maximizes

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

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

Convex relaxation

$$\min_w \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda \cancel{\|w\|_2^2} \quad \|w\|_1$$

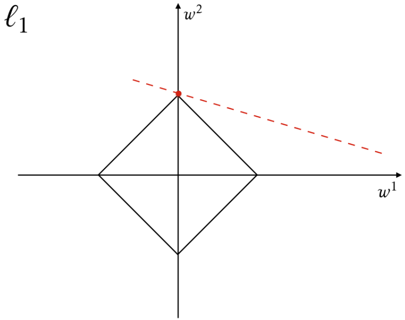
ℓ_1-norm

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

- Modeling
- Optimization

The problem of sparsity

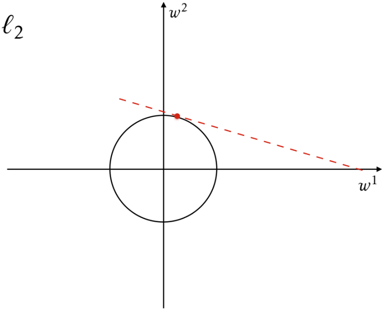
$$\min \|w\|_1, \quad \text{s.t.} \quad \hat{X}w = \hat{y}$$

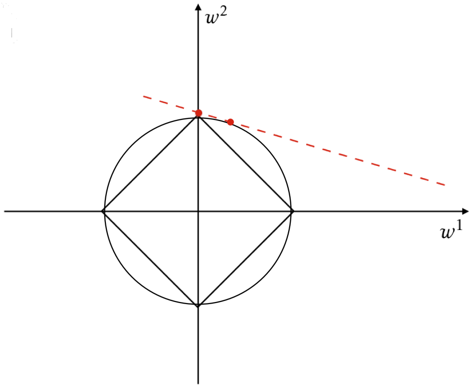


Ridge Regression and sparsity

ℓ_2

Replace $\|w\|_1$ with $\|w\|_2$?





Unlike ridge-regression, ℓ_1 regularization leads to sparsity!

Sparse regularization

$$\min_w \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda\|w\|_1$$

- Called Lasso or Basis Pursuit
- Convex but not smooth

Optimization

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

$$\min_w \underbrace{\frac{1}{n}\|\hat{X}w - \hat{y}\|^2}_{\text{convex smooth}} + \lambda \underbrace{\|w\|_1}_{\text{convex}}$$

Proximal methods

$$\min_w E(w) + R(w)$$

Let

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

and, for $w_0 = 0$

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

Proximal Methods (cont.)

$$\min_w E(w) + R(w)$$

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

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

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

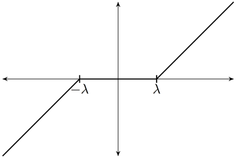
$$w_t = \text{Prox}_{\gamma R}(w_{t-1} - \gamma \nabla E(w_{t-1}))$$

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

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

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

Small coefficients are set to zero!

Back to inverse problems

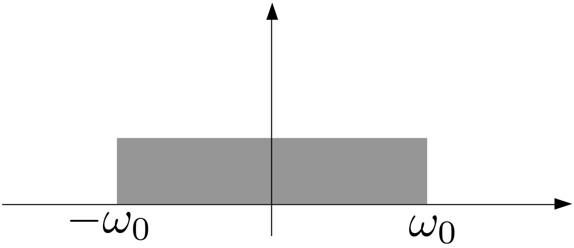
$$\hat{X} w^* + \delta = \hat{y}$$

If x_i i.i.d. random and

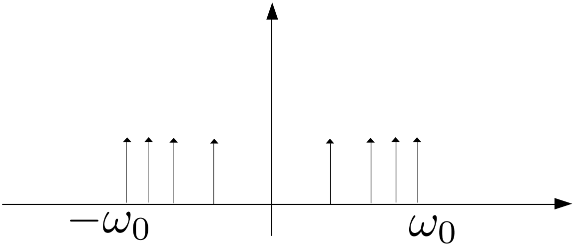
$$n \geq 2s \log \frac{d}{s}$$

then ℓ_1 regularization reaches w^*

Sampling theorem



$2\omega_0$ samples needed



LASSO

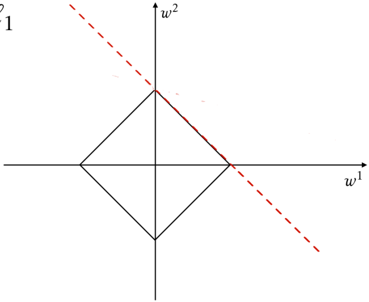
$$\min_{w} \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda\|w\|_1$$

- Interpretability: variable selection!

Variable selection and correlation

$$\min_w \underbrace{\frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda\|w\|_1}_{\cancel{\text{strictly}}\ \text{convex}}$$

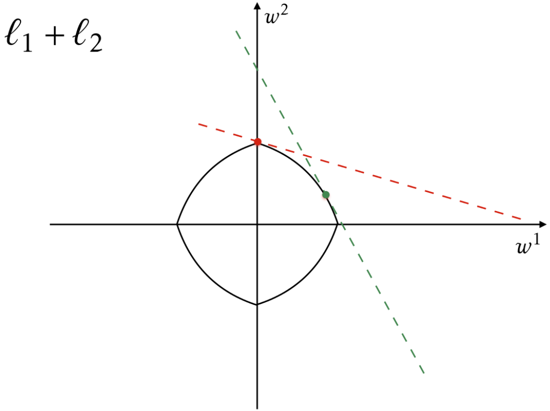
ℓ_1



Cannot handle correlations between the variables

Elastic net regularization

$$\min_w \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 \ + \ \lambda(\alpha\|w\|_1 + (1-\alpha)\|w\|^2)$$



ISTA for elastic net

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

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

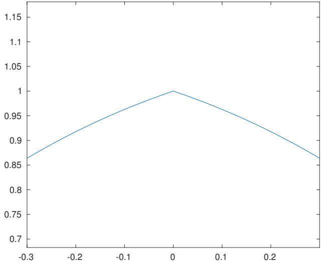
Small coefficients are set to zero!

Grouping effect

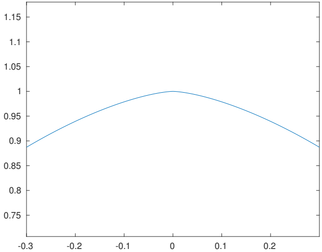
Strong convexity

$\Longrightarrow$ 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”!)

This Class

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

Next Class

- Structured Sparsity