Other regularizations, sparsity and the Lasso

MATH-412 - Statistical Machine Learning

A diverse set of regularization approaches

Regularizers are not necessarily quadratic and not necessarily norms :

$$\min_{f \in S} \widehat{\mathcal{R}}(f) + \lambda \Omega(f) \qquad \text{with e.g.} \qquad \Omega(f) = \int (f''(\mathbf{x}))^2 d\mathbf{x} \quad \text{or} \quad \int |f'(\mathbf{x})| d\mathbf{x}$$

It is possible to couple the regularization of different tasks :

$$\min_{f_1, f_2, \dots, f_K \in S} \sum_k \widehat{\mathcal{R}}_{(k)}(f_k) + \lambda \sum_k \|f_k - \bar{f}\|^2 + \mu \|\bar{f}\|^2$$

Even when the predictor has a linear parameterisation, there are various options

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \widehat{\mathcal{R}}(\boldsymbol{w}) + \lambda \Omega(\boldsymbol{w}) \qquad \text{with} \qquad \Omega(\boldsymbol{w}) = \dots$$

$$\|\boldsymbol{w}\|_q \quad , \quad \|\boldsymbol{w}\|_1 \quad , \quad \|\boldsymbol{w}\|_1 + \eta \|\boldsymbol{w}\|_2^2 \quad , \quad \sum_{j=1}^{p-1} (w_{j+1} - w_j)^2 \quad , \quad \sum_{j=1}^{p-1} |w_{j+1} - w_j| \quad , \quad \text{etc.}$$

Best subset selection

The number of effectively used variables is often denoted

$$\|\boldsymbol{w}\|_0 := \#\{j \mid w_j \neq 0\} = \sum_{j=1}^{p} 1_{\{w_j \neq 0\}}$$

Best subset selection formulation

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \widehat{\mathcal{R}}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_0$$

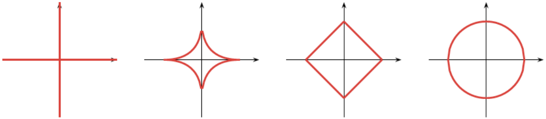
- Compromise between fitting and $\#$ of variables in the model
- The problem is NP-hard to solve in general
- Can be solved by exhaustive search amongst 2^p models for p small

Lasso (Least Absolute Shrinkage and Selection Operator)

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \widehat{\mathcal{R}}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1$$

- No closed form solution
- Convex but non-differentiable optimization problem
- Can nonetheless be solved by efficient algorithms

The general approach extends to $q < 1$ with quasi-norms but then the problem is not convex anymore.



(a) ℓ_0-ball. (b) $\ell_{0.5}$-ball. (c) ℓ_1-ball. (d) ℓ_2-ball.

Lasso regression : Constrained vs regularized problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{w}\|_1$$

vs

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{w}\|_1\le C$$

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} f(\boldsymbol{w})+\lambda\, g(\boldsymbol{w})$$

vs

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} f(\boldsymbol{w}) \quad \text{s.t.} \quad g(\boldsymbol{w})\le C$$

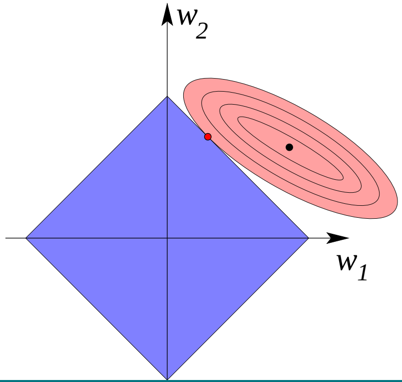
Proposition

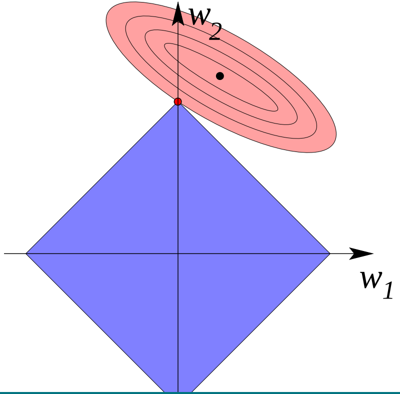
If f and g are convex, then for any value of λ there is a value of C such that both problem have the same solution and vice versa.

Geometric intuition for the Lasso

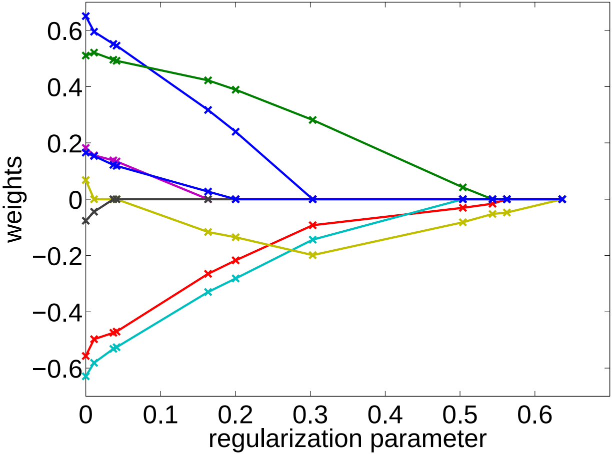
Consider the constrained problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{w}\|_1 \leq C$$





Lasso regression has piecewise linear paths



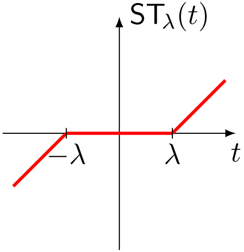
Lasso regression with orthogonal design

- Assume $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}$
- Then solving the Lasso is equivalent to solving

$$\min_{w} \tfrac{1}{2}\|\hat{\boldsymbol{c}} - \boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \quad \text{for} \quad \hat{\boldsymbol{c}} = \tfrac{1}{n}\boldsymbol{X}^\top\boldsymbol{y}$$

- $\hat{c}_j = \frac{1}{n}\boldsymbol{y}^\top\mathbf{x}^{(j)}$
- Equivalent to solve $\forall j$

$$\begin{aligned} \hat{w}_j &= \arg\min_{v\in\mathbb{R}} \ \tfrac{1}{2}v^2 - v\,\hat{c}_j + \lambda|v| \\ &= \mathsf{ST}_\lambda(\hat{c}_j) \end{aligned}$$



with the *soft-thresholding* operator :

$$\mathsf{ST}_\lambda(t) := (|t|-\lambda)_+ \ \mathsf{sign}(t)$$

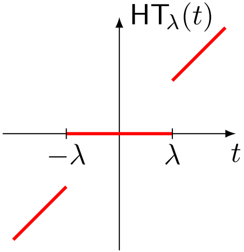
Best subset selection with orthogonal design

- Assume $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}$
- Then solving the Lasso is equivalent to solving

$$\min_{w} \frac{1}{2}\|\hat{\boldsymbol{c}} - \boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_0 \quad \text{for} \quad \hat{\boldsymbol{c}} = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{y}$$

- $\hat{c}_j = \frac{1}{n}\boldsymbol{y}^\top\mathbf{x}^{(j)}$
- Equivalent to solve $\forall j$

$$\begin{aligned} \hat{w}_j &= \arg\min_{v\in\mathbb{R}} \; \frac{1}{2}v^2 - v\,\hat{c}_j + \lambda\, 1_{\{v\neq 0\}} \\ &= \mathsf{HT}_\lambda(\hat{c}_j) \end{aligned}$$



with the *hard-thresholding* operator :

$$\mathsf{HT}_\lambda(t) := t\, 1_{\{|t|>\lambda\}}$$

Tackling the ℓ_0 constrained problem for p large...

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 \qquad \text{s.t.} \qquad \|\boldsymbol{w}\|_0 \leq k$$

- The problem is NP-hard : what if p is large ?

Greedy methods

Principle : w is estimated by increasing the support greedily. At each iteration

1. **Selection step :** A new coordinate is included in the support of $\boldsymbol{w}$
2. **Fitting step :** The new coefficient and possibly old ones are re-optimized

Forward selection (regression)

Initialization :

- $\hat{S} = \varnothing$ (estimate of support)

Repeat :

1. **Selection Step :**
   - $$j \leftarrow \arg\min_{j'} \min_{\boldsymbol{w}_{\hat{S}\cup\{j'\}}} \|\boldsymbol{y} - \boldsymbol{X}_{\hat{S}\cup\{j'\}} \boldsymbol{w}_{\hat{S}\cup\{j'\}}\|_2^2, \qquad \hat{S} \leftarrow \hat{S} \cup \{j\}$$
2. **Fitting Step :**
   - $$\hat{\boldsymbol{w}}_{\hat{S}} \leftarrow \arg\min_{\boldsymbol{w}_{\hat{S}}} \|\boldsymbol{y} - \boldsymbol{X}_{\hat{S}} \boldsymbol{w}_{\hat{S}}\|_2^2$$

Backward selection :

- Symmetric by removing variables one by one
- Not recommended if the number of variables is large, because starting from an overfitted situation.

Orthogonal Matching Pursuit (regression)

Initialization :

- $\hat{S} = \varnothing$ (estimate of support)
- $\boldsymbol{r} \leftarrow \boldsymbol{y}$ (residuals)

Repeat :

1. **Selection Step :**
   - $j \leftarrow \arg\max_{j'} |\langle \mathbf{x}^{(j')}, \boldsymbol{r} \rangle|, \qquad \hat{S} \leftarrow \hat{S} \cup \{j\}$
2. **Fitting Step :**
   - $\hat{\boldsymbol{w}}_{\hat{S}} \leftarrow \arg\min_{\boldsymbol{w}_{\hat{S}}} \|\boldsymbol{y} - \boldsymbol{X}_{\hat{S}} \boldsymbol{w}_{\hat{S}}\|_2^2 \qquad \boldsymbol{r} \leftarrow \boldsymbol{y} - \boldsymbol{X}_{\hat{S}} \hat{\boldsymbol{w}}_{\hat{S}}$

Comparing Lasso and other strategies for linear regression

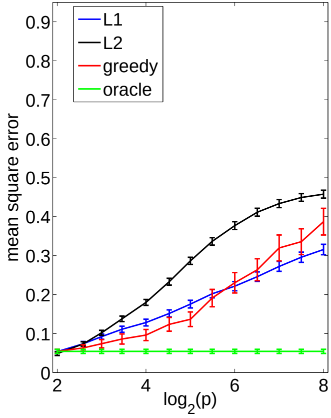
Comparing :

$$\begin{aligned}
&\text{Ridge regression :} && \min_{\boldsymbol{w}\in\mathbb{R}^p} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \\
&\text{Lasso :} && \min_{\boldsymbol{w}\in\mathbb{R}^p} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \\
&\text{OMP/FS :} && \min_{\boldsymbol{w}\in\mathbb{R}^p} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_0
\end{aligned}$$

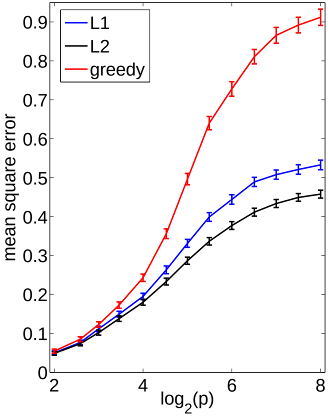
- Each method builds a path of solutions from 0 to ordinary least-squares solution
- Regularization parameters selected on the test set

Simulation results

- $\boldsymbol{X}$ = i.i.d. Gaussian design, $n = 64$, $p \in [2, 256]$,
- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{\varepsilon}$, $\|\boldsymbol{w}^*\|_0 = 4$, $w_j^* \in \{-1, 0, 1\}$, $\sigma^2 = 1$.



Sparse



Rotated (non sparse)

Note ℓ_1 stability to non-sparsity