Machine Learning using a Constrained Convex Splitting Method

Michel Barlaud ,
Wafa Belhajali , Patrick Combettes , Lionel Fillatre

Laboratoire I3S Université de Nice sophia Antipolis,
North Carolina State University

December 22, 2016

History

- 1990 Wavelet Transform
  On board Image Processing of Pleiades Satellites
  JPEG 2000
- 1994 3D Scintigraphy (3D Scanner)
- 2006 ...2016 Machine Learning
  ANR IcosHD, ANR Phasequant
  Boosting, Proximal methods
- 2009 Applications in biology

Context of this talk

- Feature selection using sparsity induction
- Feature selection using grouping induction
- Constrained Projection algorithm
- Applications in biology

Classification and regression minimizing Empirical surrogate risk

- Each sample x_i is annotated with a label y_i with values in $\{-1,+1\}$
- Linear classifier $w\colon \mathbb{R}^d \to \mathbb{R}\colon x \mapsto \langle x \mid w\rangle = x^\top w$,
- Let define empirical surrogate risk

$$\Phi\colon \mathbb{R}^d \to \mathbb{R}\colon w \mapsto \frac{1}{m}\sum_{i=1}^{m} \phi\big(y_i\langle x_i \mid w\rangle\big), \tag{1}$$

- Note that regression learns $w \in \mathbb{R}^d$ by minimizing the classical quadratic loss

$$\mathbb{R}^d \to \mathbb{R}\colon w \mapsto \frac{1}{m}\sum_{i=1}^{m} \big|y_i - \langle x_i \mid w\rangle\big|^2. \tag{2}$$

Constraint Classification and regression

- Minimize

$$\Phi : \mathbb{R}^d \to \mathbb{R} : w \mapsto \frac{1}{m} \sum_{i=1}^{m} \phi\big(y_i \langle x_i \mid w \rangle\big), \tag{3}$$

- with constraint

$$\Psi : \mathbb{R}^d \to \mathbb{R} : w \mapsto \Psi(w) \tag{4}$$

Which constraint for sparsity or grouping induction?

- Sparsity induction
  $\Psi(w) = \|w\|_1$
- Grouping induction: Elastic net
  $\Psi(w) = \lambda(\alpha\|w\|_1 + (1-\alpha)\|w\|_2)$.
- Grouping induction: Fused Lasso
- Graph Grouping induction: pairwise ℓ_∞
  $\Psi(w) = \sum_{i<j} \max(|\omega_i|, |\omega_j|)$.
- Directed acyclic Graphs (Biology)

Classical penalty Classification

- Minimize a weighted objective criterion,

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \;\; \Phi(w) + \lambda \Psi(w), \tag{5}$$

- Efficient proximal methods
- Compute λ ?
  Lasso path: Homotopy algorithm
  Complexity issue bound 3^d (intractable for large d).

Constrained classification with sparsity or grouping induction

- Let β be the Lipschitz constant of $\nabla\Phi$ and let Ψ convex
- Sparsity induction $\Psi(w) = \|w\|_1$
- Grouping induction $\Psi\colon w \mapsto \sum_{i<j} \max(|\omega_i|, |\omega_j|)$.

$$\Phi\colon \mathbb{R}^d \to \mathbb{R}\colon w \mapsto \frac{1}{m}\sum_{i=1}^{m} \phi\big(y_i\langle x_i \mid w\rangle\big), \tag{6}$$

$$C = \left\{ w \in \mathbb{R}^d \;\middle|\; \Psi(w) \leqslant \eta \right\} \tag{7}$$

$$\underset{w \in C}{\text{minimize}}\;\; \Phi(w). \tag{8}$$

Splitting algorithm

- Splitting algorithm:

$$\text{for } n = 0, 1, \ldots \\ \left\lfloor \begin{array}{l} v_n = w_n - \gamma_n \nabla \Phi(w_n) \\ w_{n+1} = P_C(v_n) + a_n. \end{array} \right.$$

- w_n converges to a solution if

$$\sum_{n \in \mathbb{N}} \|a_n\| < +\infty, \quad \text{and} \quad \sup_{n \in \mathbb{N}} \gamma_n < \frac{2}{\beta}. \tag{9}$$

- How do we implement this projection?

Projection algorithm

- Let first computes the so-called subgradient projection of p_k onto D

$$p_{k+1/2} = p_k + \frac{\eta - \Psi(p_k)}{\|\nabla\Psi(p_k)\|^2}\nabla\Psi(p_k) \tag{10}$$

Subgradient

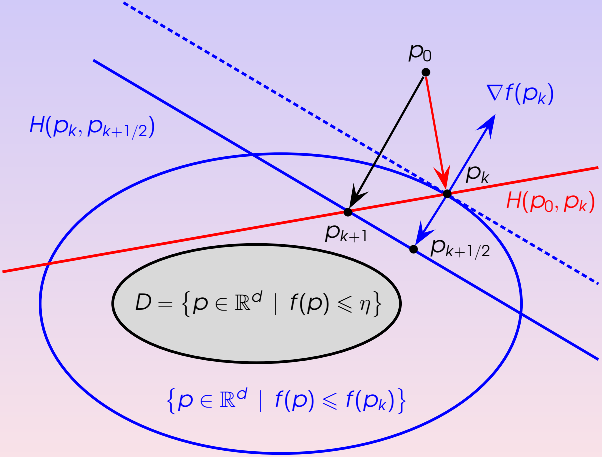
Let $\Psi(w) = \|w\|_1$, a subgradient s of Ψ at a point $(\xi_i)_{1\leqslant i\leqslant d} \in \mathbb{R}^d$ is $s = (\text{sign}(\xi_i))_{1\leqslant i\leqslant d}$

$$\text{sign}\colon \xi \mapsto \begin{cases} 1 & \text{if } \ \xi > 0 \\ 0 & \text{if } \ \xi = 0 \\ -1 & \text{if } \ \xi < 0. \end{cases} \tag{11}$$

In the case of pairwise ℓ^∞ constraint, the ith component of a subgradient of φ at a point $(\xi_i)_{1\leqslant i\leqslant d} \in \mathbb{R}^d$ is given by

$$\sigma_i = \sum_{j>i} \begin{cases} \text{sign}(\xi_i) & \text{if } \ |\xi_i| \geqslant |\xi_j| \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Projection step of the algorithm

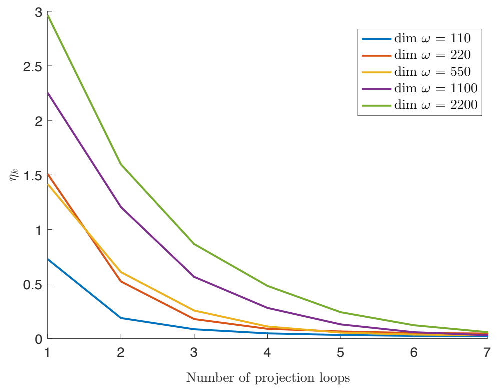


Algorithm

```
for $n = 0, 1, \ldots$
    $v_n = w_n - \frac{\gamma_n}{m} \sum_{i=1}^{m} y_i \phi'\big(y_i \langle x_i \mid w \rangle\big) x_i$
    $p_0 = v_n$
    for $k = 0, 1, \ldots, K_n - 1$
        $\eta_k = \eta - \varphi(p_k)$
        if $\eta_k \geqslant 0$
            terminate.
        $p_{k+1/2} = p_k + \eta_k s_k / \|s_k\|^2$
        Intersection computation (Haugazeau)
            $p_{k+1} = Q(p_0, p_k, p_{k+1/2})$
    $w_{n+1} = p_{K_n}$.                                        (13)
```

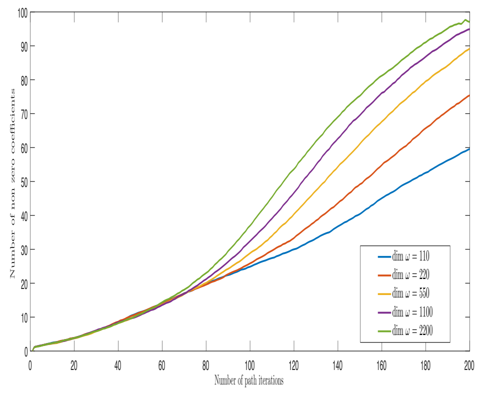
Convergence of the projection loop

- Convergence of projection loop graph



Convergence of the Lasso path

- Convergence of projection loop is faster than convergence of Lasso

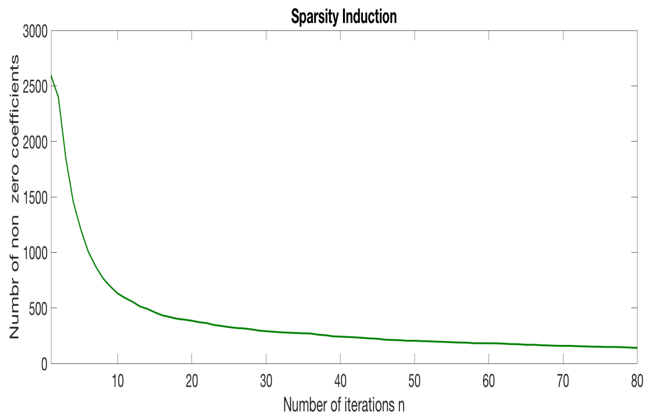


Application to Computational Biology

- Classification of cells : **SATT grant ”Cellid”** (2011) and **ANR** project “Phasequant” (2014)
  with Phasics, Tiro and Morpheme
- Predict relapse in early stage lung adenocarcinoma (2014) using Genomic RNAseq data set
  with Pr B. Mari and A. Paquet IPMC
  Results: Complex signature
- Predict response to treatment (amisulpride) in Psychiatric disorder (2014), **SATT grant 2016**
  with Pr N. Glaichenhaus IPMC and INSERM Creteil, (European Project “Optimize”)
  2016 Dassault price
  Preliminary result: IL 15 and CCL2 Biomarker signature
- Single cells
  with Pr P Barbry IPMC

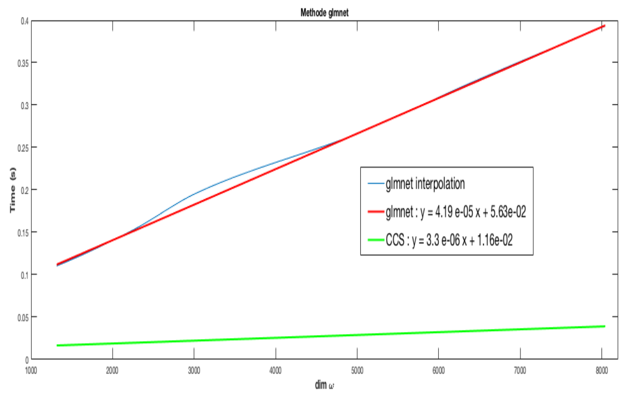
Sparsity induction: Convergence of $\|w\|_0$ norm

- $\|w\|_0$ is a monotone decreasing function of the number of iterations.



Time Comparaison CCS versus glmnet

- Breast Cancer data base: $n = 295$ samples, $d = 8041$ genes



- 10 times faster than Lagrangian methods

AUC Comparaison CCS versus glmnet

- Breast Cancer data base: $n = 295$ samples, $d = 8041$ genes

Table: Breast Cancer AUC comparisons

	ℓ^1 penalty (glmnet)	Group lasso	ℓ^1 CCS
AUC (%)	64.5	66.7	72.3

- 8 % accuracy improvment over Lagrangian methods

Conclusion

- Our CCS method outperforms Lagrangian approaches
  8 % accuracy improvement over glmnet
  10 times faster than glmnet
- Current work: Speed up using Random Coordinate Descent algorithm
  (with JC Regin and G Perez)
  27 times faster than glmnet
- Future work: Extension to Constraint Matrix Framework
  (with JB Caillau)

References

- M. Barlaud, W Belhajali, P. L. Combettes, and L. Fillatre, Classification and regression using a constrained convex splitting method,
  June 2015, http://arxiv.org/abs/1506.02196v3
- *democode (Matlab mex mac) available at http://www.i3s.unice.fr/ barlaud/code.html*
- Thanks J.B Caillau, J.C Regin and G Perez for fruitful discussions