A unified perspective on convex structured sparsity

Guillaume Obozinski

Laboratoire d'Informatique Gaspard Monge

École des Ponts ParisTech

Joint work with Francis Bach

Conférence Mascot-Num

Ecole Centrale de Nantes, 21 mars 2018

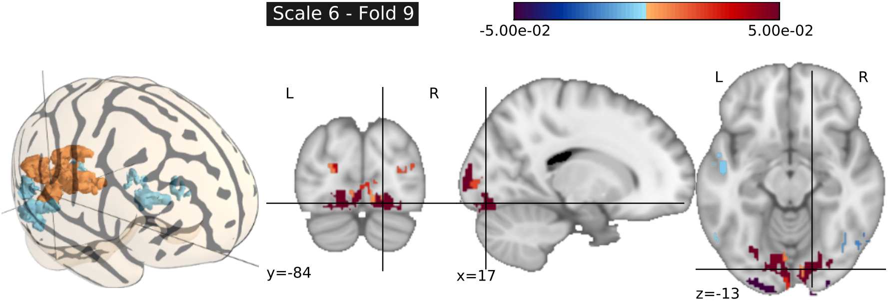
Structured Sparsity

The support is not only **sparse**, but, in addition,
we have prior information about its **structure**.

Examples

- The variables should be selected in groups.
- The variables lie in a hierarchy.
- The variables lie on a graph or network and the support should be localized or densely connected on the graph.

Applications: Difficult inverse problem in Brain Imaging



Jenatton et al. (2011b)

Convex relaxation for classical sparsity

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

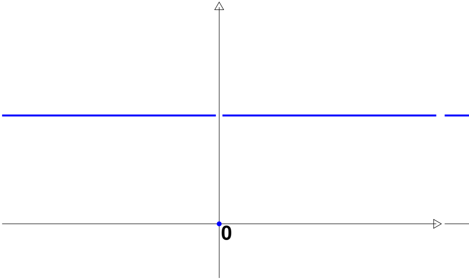
$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$

Lasso

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \|w\|_1$$



Formulation with **combinatorial functions**

Let $V = \{1, \ldots, d\}$.

Let L be some empirical risk such as $L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$.

Given a set function $F : 2^V \mapsto \mathbb{R}_+$ consider

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

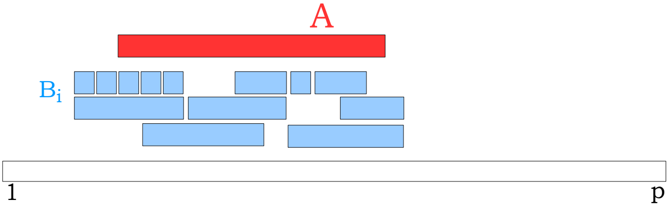
Examples of combinatorial functions

- Use **recursivity** or **counts** of structures (e.g. tree) with DP
- **Block-coding** (Huang et al., 2011)

$$\tilde{G}(A) = \min_{B_i} F(B_1) + \ldots + F(B_k) \quad \text{s.t.} \quad B_1 \cup \ldots \cup B_k \supset A$$

- **Submodular functions**

Block-coding (Huang, Zhang and Metaxas (2009))



$F_+ : 2^V \to \overline{\mathbb{R}}_+$ a positive set function.

$$F_\cup(A) = \min_{\mathcal{S}} \sum_{B \in \mathcal{S}} F_+(B) \quad \text{s.t.} \quad A \subset \bigcup_{B \in \mathcal{S}} B.$$

$\rightarrow$ minimal weighted cover set problem.

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

- $\rightarrow$ Greedy algorithms
- $\rightarrow$ Non-convex methods
- $\rightarrow$ Relaxation

$\lvert A \rvert$	$F(A)$
$L(w) + \lambda \lvert\mathrm{Supp}(w)\rvert$	$L(w) + \lambda\, F(\mathrm{Supp}(w))$
$\downarrow$	$\downarrow ?$
$L(w) + \lambda \lVert w \rVert_1$	$L(w) + \lambda$ ...?...

Penalizing *and* regularizing...

Given a function $F : 2^V \to \bar{\mathbb{R}}_+$, consider for $\nu, \mu > 0$ the combined penalty:

$$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p.$$

Motivations

- Compromise between variable selection and smooth regularization
- Required for functions F allowing large supports
- Interpretable as a *description length* for the parameters w.

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\rightarrow$ scale invariance.

Definition (Homogeneous extension of a function g)

$$g_h : x \mapsto \inf_{\lambda > 0} \frac{1}{\lambda} g(\lambda x).$$

Proposition

The tightest convex positively homogeneous lower bound of a function g is the convex envelope of g_h.

Leads us to consider:

$$\begin{aligned} \mathrm{pen}_h(w) &= \inf_{\lambda > 0} \frac{1}{\lambda} \big(\mu\, F(\mathrm{Supp}(\lambda w)) + \nu\, \|\lambda w\|_p^p \big) \\ &\propto \Theta(w) := \|w\|_p\, F(\mathrm{Supp}(w))^{1/q} \qquad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1. \end{aligned}$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A\subset V, A\neq\varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

Proposition

The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p\, F(\mathrm{Supp}(w))^{1/q}$.

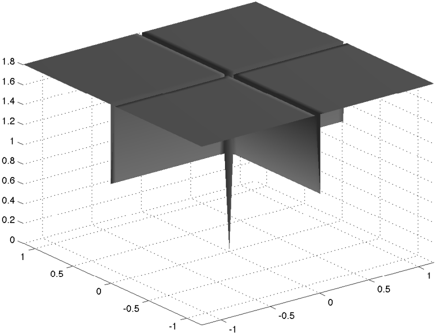
Proof.

Denote $\Theta(w) = \|w\|_p\, F(\mathrm{Supp}(w))^{1/q}$:

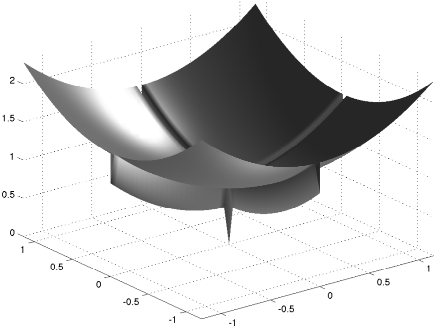
$$\begin{aligned}
\Theta^*(s) &= \max_{w\in\mathbb{R}^d} w^\top s - \|w\|_p\, F(\mathrm{Supp}(w))^{1/q} \\
&= \max_{A\subset V} \max_{w_A\in\mathbb{R}^A} w_A^\top s_A - \|w_A\|_p\, F(A)^{1/q} \\
&= \max_{A\subset V} \iota_{\{\|s_A\|_q \leqslant F(A)^{1/q}\}} = \iota_{\{\Omega_p^*(s)\leqslant 1\}}
\end{aligned}$$

□

Graphs of the different penalties

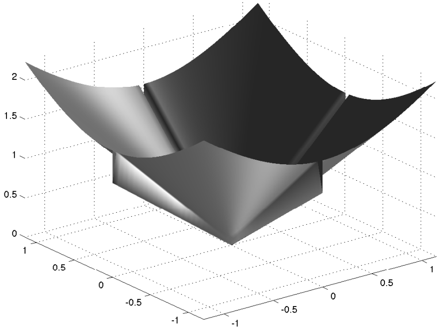


$F(\mathrm{Supp}(w))$

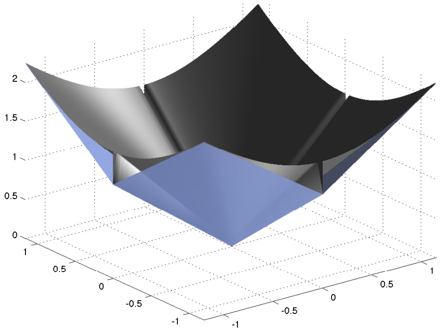


$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_2^2$

Graphs of the different penalties



$\Theta(w) = \sqrt{F(\mathrm{Supp}(w))}\|w\|_2$

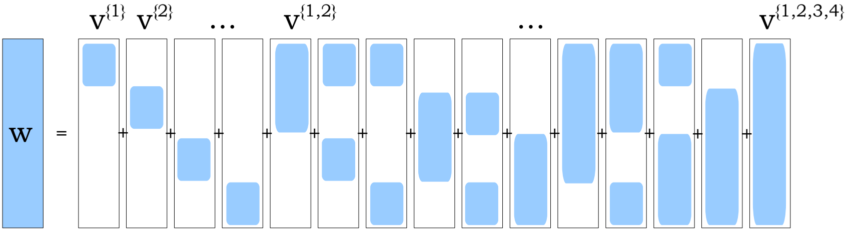


$\Omega^F(w)$

A large latent group Lasso (Jacob et al., 2009)

$$\mathcal{V} = \{v = (v^A)_{A \subset V} \in \left(\mathbb{R}^V\right)^{2^V} \text{ s.t. } \operatorname{Supp}(v^A) \subset A\}$$

$$\Omega_p(w) \quad = \quad \min_{v \in \mathcal{V}} \sum_{A \subset V} F(A)^{\frac{1}{q}} \|v^A\|_p \quad \text{s.t.} \quad w = \sum_{A \subset V} v^A,$$

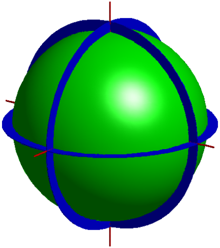


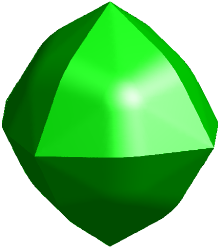
Some simple examples

	F	Ω_p
	$\lvert A \rvert$	$\lVert w \rVert_1$
	$1_{\{A \neq \varnothing\}}$	$\lVert w \rVert_p$
If $\mathcal{G}$ is a partition:	$\sum_{B \in \mathcal{G}} 1_{\{A \cap B \neq \varnothing\}}$	$\sum_{B \in \mathcal{G}} \lVert w_B \rVert_p$
If $\mathcal{G}$ is **not** a partition:	$\sum_{B \in \mathcal{G}} 1_{\{A \cap B \neq \varnothing\}}$	**new**: Overlap count Lasso

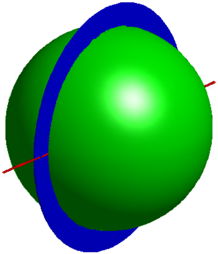
Combinatorial norms as atomic norms

$F(A) = |A|^{1/2}$





$F(A) =$
$1_{\{A \cap \{1,2,3\} \neq \varnothing\}}$
$+ 1_{\{A \cap \{2,3\} \neq \varnothing\}}$
$+ 1_{\{A \cap \{3\} \neq \varnothing\}}$



$\Theta_2^F(w)$ $\Omega_2^F(w)$

Relation between combinatorial functions and norms

Name	$F(A)$	Norm Ω_p
cardinality	$\|A\|$	Lasso (ℓ_1)
nb of groups	$\sum_{B\in\mathcal{G}} 1_{\{A\cap B\neq\varnothing\}}$	Group Lasso (ℓ_1/ℓ_p)
nb of groups	$\delta_A, A\in\mathcal{G}, +\infty$ else	Latent group Lasso
max. nb of el./group	$\max_{B\in\mathcal{G}} \|A\cap B\|$	Exclusive Lasso (ℓ_p/ℓ_1)
constant	$1_{\{A\neq\varnothing\}}$	ℓ_p-norm
func. of cardinality	$h(\|A\|)$, h sublinear	
	$1_{\{A\neq\varnothing\}} \vee \frac{\|A\|}{k}$	k-support norm ($p=2$)
func. of cardinality	$h(\|A\|)$, h concave	OWL (for $p=\infty$)
	$\lambda_1\|A\| + \lambda_2\left[\binom{d}{k} - \binom{d-\|A\|}{k}\right]$	OSCAR ($p=\infty, k=2$)
	$\sum_{i=1}^{\|A\|} \Phi^{-1}\left(1-\frac{qi}{2d}\right)$	SLOPE ($p=\infty$)
chain length	$h(\max(A))$	wedge penalty

Is the relaxation "faithful" to the original function

Consider $V = \{1, \ldots, p\}$ and the function

$$F(A) = \mathrm{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

What is its convex relaxation?

- Easy to show that $|A|$ must have the same relaxation.

$\Rightarrow$ $\Omega_p^F(w) = \|w\|_1$

The relaxation fails

$\Rightarrow$ What are the good functions F?

$\rightarrow$ Good functions are *Lower Combinatorial Envelopes* (LCE)

- Submodular functions are LCEs !

Min-cover *vs* Overlap count functions

Given a collection of sets $\mathcal{G}$ with weights $(d_B)_{B\in\mathcal{G}}$...
... two natural functions to consider:

Min-cover

$$F_{\cup}(A) := \inf_{\mathcal{S}\subset\mathcal{G}} \Big\{ \sum_{B\in\mathcal{S}} d_B \;\mid\; A \subset \bigcup_{B\in\mathcal{S}} B \Big\} :$$

- $F_{\cup,-}$ is the corresponding *fractional* min-cover value

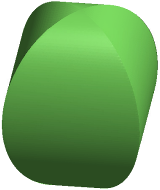
Overlap count

$$F_{\cap}(A) = \sum_{B\in\mathcal{G}} d_B \, 1_{\{A\cap B\neq\varnothing\}}$$

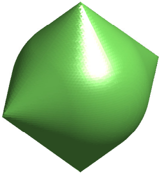
- counting the number of set of $\mathcal{G}$ intersected
- "maximal cover" by elements of $\mathcal{G}$
- $F_{\cap}$ is a *submodular* function (as a sum of submodular functions).

Latent group Lasso vs Overlap count Lasso vs ℓ_1/ℓ_p

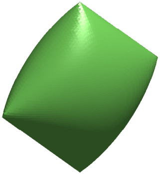
$$\mathcal{G} = \{\{1,2\}\{2,3\}\}.$$



$\Omega_2^{F_\cup}(w) \leq 1$



$\Omega_2^{F_\cap}(w) \leq 1$



$\|w_{\{1,2\}}\|_2 + \|w_{\{2,3\}}\|_2 \leq 1$

$$F_\cap(A) = 1_{\{A\cap\{1,2\}\neq\varnothing\}} + 1_{\{A\cap\{2,3\}\neq\varnothing\}},$$

$$F_\cup(A) = \min_{\delta,\delta'} \left\{\delta + \delta' \mid 1_A \leq \delta\, 1_{\{1,2\}} + \delta'\, 1_{\{2,3\}}\right\}.$$

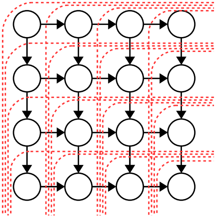
Hierarchical sparsity

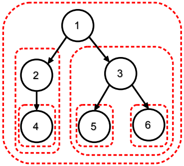
Consider a DAG, with

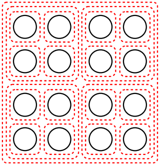
- A_i, D_i ancestors/descendants sets of i including itself.
- Significant literature: Zhao et al. (2009); Yuan et al. (2009); Jenatton et al. (2011c); Mairal et al. (2011); Bien et al. (2013); Yan and Bien (2015) and many others...
- e.g. formulations with ℓ_1/ℓ_p-norms (Zhao et al., 2009; Jenatton et al., 2011c)

$$\Omega(w) = \sum_{i \in V} \|w_{D(i)}\|_2, \quad \text{with}$$

efficient algorithms for *tree-structured* groups.







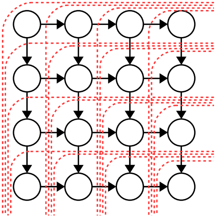
Combinatorial functions for strong hierarchical sparsity

Consider a DAG, with

- A_i, D_i ancestors/descendants sets of i including itself.

Strong hierarchical sparsity:

"A node can be selected only if **all** its ancestors are selected".

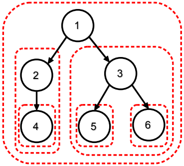


Overlap count with D_i:

$$F_\cap(B) := \sum_{i \in V} d_i \, 1_{\{B \cap D_i \neq \varnothing\}} = \sum_{i \in A_B} d_i,$$

vs **Min-cover with A_i:**

$$F_\cup(B) := \inf_{I \subset V} \Big\{ \sum_{i \in I} f_i \;\mid\; B \subset \bigcup_{i \in I} A_i \Big\}.$$



Results for different types of graphs

Chains

- Families $F_\cap$ and $F_\cup$ are equivalent
- Norms and prox can be computed using algorithms for isotonic regression.

Trees

- Families $F_\cap$ and $F_\cup$ are different
- Norms and prox for $F_\cap$ can be computed using a *decomposition algorithm*.
- No efficient algorithm known for $F_\cup$.

DAGs

- Norms and prox for $F_\cap$ can be computed using general connexion with isotonic regressions on DAGs.
- No efficient algorithm known for $F_\cup$.

Sublinear functions of the cardinality

$$F(A) = \sum_{k=1}^{d} f_k \, 1_{\{|A|=k\}},$$

and F_- must be sublinear.

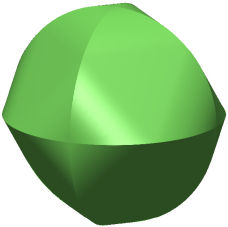
Let $|s|_{(1)} \geq \ldots \geq |s|_{(d)}$ be the reverse order statistics of the entries of s. Then

$$\Omega_p^*(w) = \max_{1 \leq j \leq d} \frac{1}{f_j^{1/q}} \left[\sum_{i=1}^{j} |s|_{(i)}^q \right]^{1/q}$$

First example

$$F_+(A) = \begin{cases} 1 & \text{if } |A| = k \\ \infty & \text{o.w.} \end{cases}$$

recovers the k-support norm of Argyriou et al. (2012) ($p = 2$).



Concave functions of the cardinality

If $k \mapsto f_k$ is concave then we have

$$\Omega_\infty(w) = \sum_{i=1}^{d} (f_i - f_{i-1})\,|w|_{(i)}.$$

Ordered weighted Lasso (OWL) (Figueiredo and Nowak, 2014)

Examples

- OSCAR (Bondell and Reich, 2008): $= \lambda_1 \|w\|_1 + \lambda_2 \Omega(w)$ with

$$\Omega(w) = \sum_{i<j} \max\big(|w_i|, |w_j|\big) \qquad \text{obtained with} \quad f_k = \binom{d}{2} - \binom{d-k}{2}$$

- SLOPE (Bogdan et al., 2015): $f_k = \sum_{i=1}^{k} \Phi^{-1}\Big(1 - \dfrac{qi}{2d}\Big)$

Computations and extensions of OWL

Since F is submodular, Ω_∞^F is a linear function of $|w|$ if the order of the coefficients is fixed. Computational problem can therefore be reduced to the case of the chain.

Proposition (Figueiredo and Nowak, 2014)

In the $p = \infty$ case the proximal operator can be computed efficiently via *isotonic regression* and PAVA.

Proposition (ℓ_p-OWL norms)

Norm definitions and efficient computations of norms and proximal operators can be naturally extended to Ω_p^F via *isotonic regression* and PAVA.

An example: penalizing the range

Structured prior on support (Jenatton et al., 2011a):

- the support is an interval of $\{1, \ldots, p\}$

Natural associated penalization:

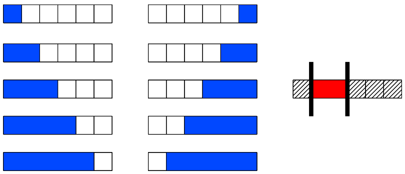
$F(A) = \mathrm{range}(A) = i_{\max}(A) - i_{\min}(A) + 1.$

$\rightarrow$ F is not submodular...

$\rightarrow$ $G(A) = |A|$

But $F(A) := d - 1 + range(A)$ is submodular !

In fact $F(A) = \sum_{B \in \mathcal{G}} 1_{\{A \cap B \neq \varnothing\}}$ for B of the form:



Jenatton et al. (2011a) considered $\Omega(w) = \sum_{B \in \mathcal{B}} \|w_B \circ d_B\|_2$.

Experiments

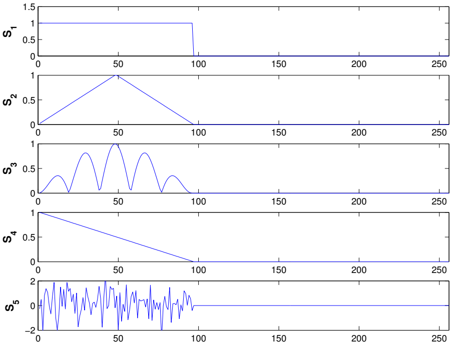


Figure: Signals

S_1 constant

S_2 triangular shape

S_3 $x \mapsto |\sin(x)\sin(5x)|$

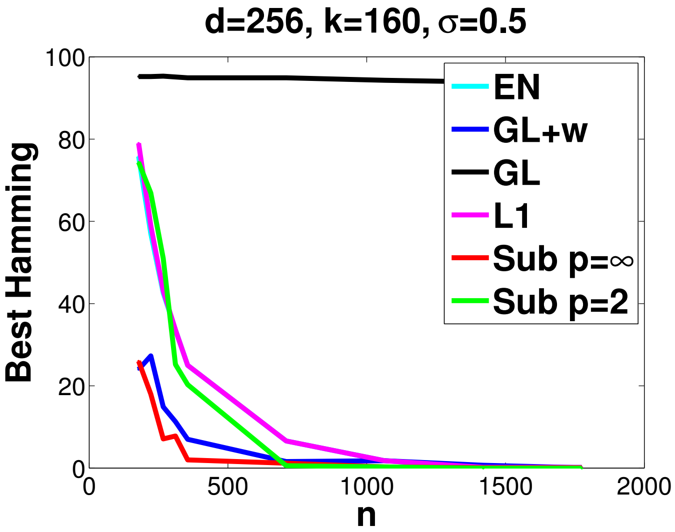
S_4 a slope pattern

S_5 i.i.d. Gaussian pattern

Compare:

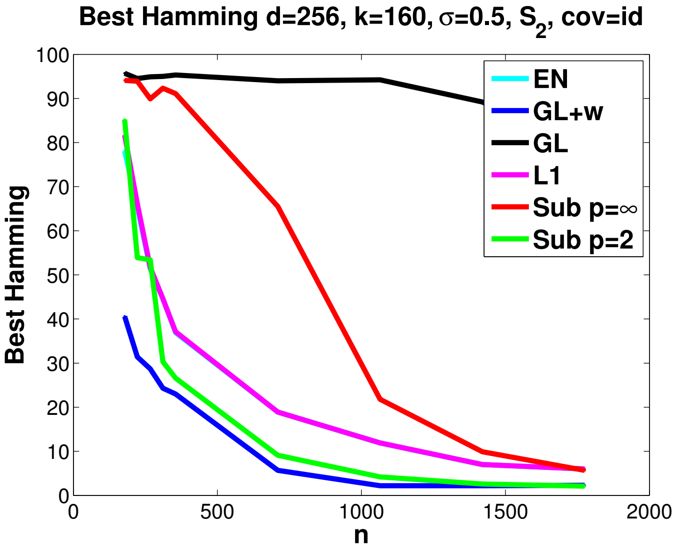
- Lasso
- Elastic Net
- Naive ℓ_2 group-Lasso
- Ω_2 for $F(A) = d - 1 + \mathrm{range}(A)$
- Ω_∞ for $F(A) = d - 1 + \mathrm{range}(A)$
- The weighted ℓ_2 group-Lasso of (Jenatton et al., 2011a).

Constant signal



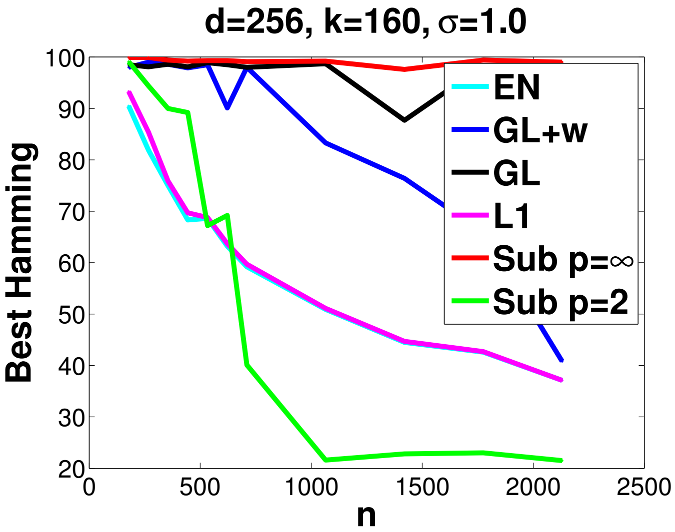
- $d = 256$
- $k = 160$
- $\sigma = .5$

Triangular signal



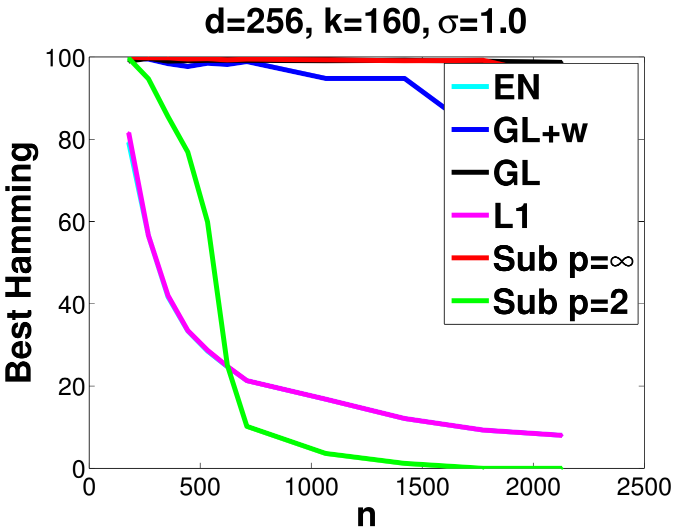
- $d = 256$
- $k = 160$
- $\sigma = .5$

$(x_1, x_2) \mapsto |\sin(x_1)\sin(5x_1)\sin(x_2)\sin(5x_2)|$ signal in 2D



- $d = 256$
- $k = 160$
- $\sigma = .5$

i.i.d Random signal in 2D



- $d = 256$
- $k = 160$
- $\sigma = .5$

Summary

- A convex relaxation for functions penalizing
  (a) the support via a general set function
  (b) the ℓ_p norm of the parameter vector w.
- Retrieves a large fraction of the norms used (Lasso, group Lasso, Exclusive Lasso, OSCAR, OWL, SLOPE, etc).
- Generic efficient algorithms for chains/trees/graphs-OCL
- Open: efficient prox computation for tree/DAG for $F_\cap$
  - Alternative fast column generation/FCFW algorithm (Vinyes and Obozinski, 2017).
- Did not talk about general support recovery and fast rates convergence that can be obtained based on generalization of the irrepresentability condition/restricted eigenvalue condition.

References I

Argyriou, A., Foygel, R., and Srebro, N. (2012). Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems 25*, pages 1466–1474.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732.

Bien, J., Taylor, J., Tibshirani, R., et al. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). SLOPE: adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 9(3):1103–1140.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123.

Figueiredo, M. and Nowak, R. D. (2014). Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. Technical Report 1409.4005, arXiv.

Groenevelt, H. (1991). Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *Eur. J Oper. Res.*, 54(2):227–236.

Huang, J., Zhang, T., and Metaxas, D. (2011). Learning with structured sparsity. *The JMLR*, 12:3371–3412.

Jacob, L., Obozinski, G., and Vert, J. (2009). Group lasso with overlap and graph lasso. In *ICML*.

References II

Jenatton, R., Audibert, J., and Bach, F. (2011a). Structured variable selection with sparsity-inducing norms. *JMLR*, 12:2777–2824.

Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., and Thirion, B. (2011b). Multi-scale mining of fmri data with hierarchical structured sparsity. *Arxiv preprint arXiv:1105.0363*.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2011c). Proximal methods for hierarchical sparse coding. *JMLR*, 12:2297–2334.

Mairal, J., Jenatton, R., Obozinski, G., and Bach, F. (2011). Convex and network flow optimization for structured sparsity. *JMLR*, 12:2681–2720.

Vinyes, M. and Obozinski, G. (2017). Fast column generation for atomic norm regularization. In *Artificial Intelligence and Statistics*.

Wainwright, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1- constrained quadratic programming. *IEEE Transactions on Information Theory*, 55:2183–2202.

Yan, X. and Bien, J. (2015). Hierarchical sparse modeling: A choice of two regularizers. *arXiv preprint arXiv:1512.01631*.

Yuan, M., Joseph, V. R., and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4):1738–1757.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497.