An Active Set Algorithm for Structured Sparsity-Inducing Norms

Rodolphe Jenatton **Jean-Yves Audibert** **Francis Bach**
Willow project, INRIA - Ecole Normale Supérieure

OPT Workshop, NIPS, 12th December 2009

INRIA

LETTRES
SCIENCES
ECOLE NORMALE SUPERIEURE

IMAGINE
UNIVERSITÉ PARIS-EST

École des Ponts
ParisTech

CSTB
le futur en construction

Outline

- Sparsity-inducing norms.
- **Structured** sparsity-inducing norms.
- Active set algorithm.

Sparsity-inducing norms (1/2)

$$\min_{w \in \mathbb{R}^p} \quad \overbrace{L(w)}^{\text{data fitting term}} + \mu \underbrace{\Omega(w)}_{\textbf{sparsity-inducing norm}}$$

Standard approach to enforce sparsity in learning procedures:

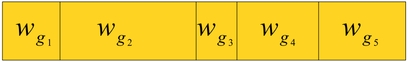
- Regularizing by a **sparsity-inducing norm** Ω.
- Some w_j's are set to zero, depending on the regularization parameter $\mu \geq 0$.

The most popular choice for Ω:

- The ℓ_1 norm, $\|w\|_1 = \sum_{j=1}^{p} |w_j|$.
- For the square loss, Lasso (Tibshirani, 1996).
- However, ℓ_1 just about **cardinality**!

Sparsity-inducing norms (2/2)

Another popular choice for Ω:



$$\mathcal{G} = \{g_1, g_2, g_3, g_4, g_5\}$$

- The ℓ_1-ℓ_2 norm,

$$\sum_{g \in \mathcal{G}} \|w_g\|_2 = \sum_{g \in \mathcal{G}} \Big(\sum_{j \in g} w_j^2 \Big)^{1/2}, \text{ with } \mathcal{G} \text{ a } \textbf{partition} \text{ of } \{1, \ldots, p\}.$$

- The ℓ_1-ℓ_2 norm sets to zero **groups of non-overlapping variables** (as opposed to single variables for the ℓ_1 norm).
- For the square loss, group Lasso (Yuan and Lin, 2006).
- However, ℓ_1-ℓ_2 encodes **fixed/static prior information**:
  - Require to know in advance how to group the variables !

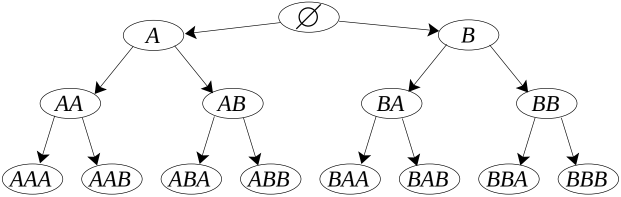
Questions:

- What happen if $\mathcal{G}$ is not a partition anymore?
- Why is structure important?

When structure can help

Hierarchical structure:

- Descriptors of images organized in a pyramid.
- Tree of the substrings of a finite alphabet (e.g., in bioinformatics/text-processing).



Contiguous/Convex-like structure:

- Contiguous sequences in time-series.
- Brain activation areas in MEG/EEG.

Structured sparsity-inducing norms (1/2)

For a more general set of groups $\mathcal{G}$ (in the power set of $\{1,\dots,p\}$):

When penalizing by the ℓ_1-ℓ_2 norm,

$$\sum_{g\in\mathcal{G}} \|w_g\|_2 = \sum_{g\in\mathcal{G}} \Big(\sum_{j\in g} w_j^2 \Big)^{1/2}$$

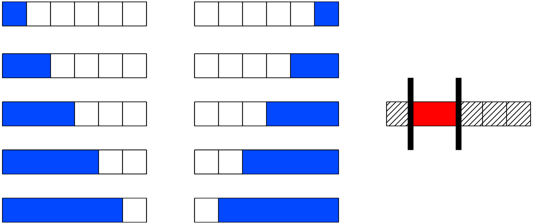
- The ℓ_1 norm induces sparsity at the group level:
  - Some w_g's are set to zero.
- Inside the groups, the ℓ_2 norm does not promote sparsity.
- Intuitively, the zero pattern of w is given by

$$\bigcup_{g\in\mathcal{G}'} g \quad \text{for some } \mathcal{G}' \subseteq \mathcal{G}.$$

(see proof in Jenatton et al., 2009a)

Examples of set of groups $\mathcal{G}$ (1/2)

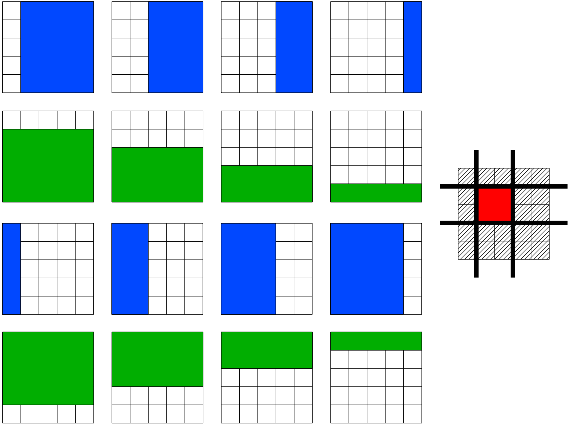
Selection of contiguous patterns on a sequence, $p = 6$:



- $\mathcal{G}$ is the set of blue groups.
- Any union of blue groups set to zero leads to the selection of a contiguous pattern.

Examples of set of groups $\mathcal{G}$ (2/2)

Selection of rectangles on a 2-D grid, $p = 25$.



- $\mathcal{G}$ is the set of blue/green groups.
- Any union of blue/green groups set to zero leads to the selection of a rectangle.

Structured sparsity-inducing norms (2/2)

To sum up, given $\mathcal{G}$, the variables set to zero by Ω belong to

$$\Big\{ \bigcup_{g \in \mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \Big\},\ \text{i.e., are } \textbf{a union of elements of } \mathcal{G}.$$

$\mathcal{G} \rightarrow$ **Zero patterns** $\mathcal{Z}$:

- Generating the **union-closure** of $\mathcal{G}$.

Zero patterns $\mathcal{Z} \rightarrow \mathcal{G}$:

- Design groups $\mathcal{G}$ from any **union-closed set of zero patterns** . . .
- . . . or from any **intersection-closed set of non-zero patterns**.
  (see result from set theory, e.g., Doignon and Falmagne, 1998)

Norm design, in form of allowed zero patterns by Ω.

Overview of other work on structured sparsity

- Specific hierarchical structure (Szafranski et al., 2007; Zhao et al., 2008; Bach, 2008).
- **Union-closed** (as opposed to intersection-closed) family of nonzero patterns (Baraniuk et al., 2008; Jacob et al., 2009).
- Nonconvex penalties based on information-theoretic criteria with greedy optimization (Huang et al., 2009).
- Structure expressed through a Bayesian prior (see, e.g., He and Carin, 2009).

Optimization

$$\min_{w \in \mathbb{R}^p} L(w) + \frac{\lambda}{2}[\Omega(w)]^2.$$

- Data fitting term L, continuously differentiable and convex.

Hard problem:

- Standard tricks for Lasso/group Lasso do not apply (e.g., subgradient, proximal or projection-based methods).

Options to deal with this nonsmooth convex problem:

- *Small scale:* Second-Order Cone Programming (SOCP), time complexity $O(p^{3.5} + |\mathcal{G}|^{3.5})$.
- *Small/medium scale:* variational equalities ("η-trick"), then projected gradient descent or alternating optimization scheme.

These approaches do not take advantage of sparsity. . .

Active set algorithm outline

Active set algorithm (Lee et al., 2007; Szafranski et al., 2007; Roth and Fischer, 2008; Bach, 2008; Obozinski et al., 2009):

- Start with $J = \varnothing$.
- Solve sequence of problems reduced to a **small set of active variables** $J \subseteq \{1, \dots, p\}$:

$$\min_{w_J \in \mathbb{R}^{|J|}} L_J(w_J) + \frac{\lambda}{2} \left[\Omega_J(w_J)\right]^2 .$$

- The active set is increased at each iteration, while global optimality is checked.

Checking global optimality...

Optimality, from reduced problem to global problem:

- Is $w = \binom{w_J}{0_{J^c}}$ optimal ?
- Check the **global duality gap**:

$$\frac{1}{2\lambda}\left\{\, [\Omega^*(\kappa)]^2 + \lambda w_J^\top \nabla L_J(w_J)\right\},\ \text{with } \kappa = \nabla L(w).$$

- Needs to compute the **dual norm** $\Omega^*(\kappa) = \max_{\Omega(u)\leq 1} u^\top \kappa$.
- **But computation as hard as the initial problem !**

Main technical contribution:

- lower/upper bounds on Ω^* for necessary/sufficient optimality conditions.

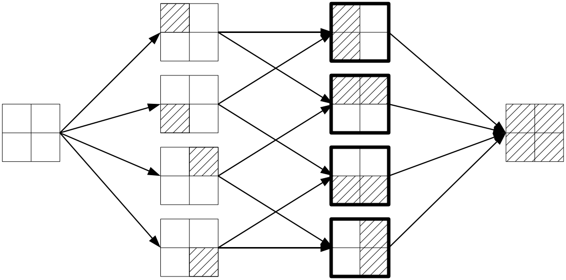
Difficulty "hidden" in usual cases:

- Lasso, ℓ_∞ norm.
- (Non-overlapping) group Lasso, block ℓ_∞-ℓ_2 norm.

Growth of the active set

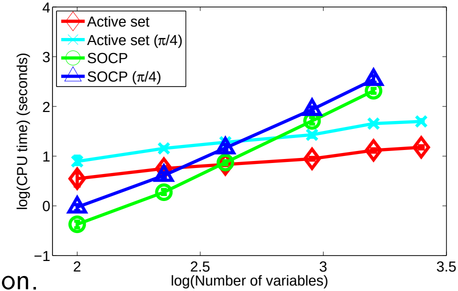
How is the active set growing?

- Ω defines a set of allowed nonzero patterns (e.g., rectangles)...
- ...naturally ordered, **by inclusion**, in a directed acyclic graph (DAG).
- **Active set algorithm = "walk" in this DAG**.



- Next active variables given by the necessary/sufficient optimality conditions.

Algorithmic complexity



Complexity:

- s, active set size after optimization.
- If SOCP is used as a *black-box* solver.
- If $|\mathcal{G}| = O(\sqrt{p})$, e.g., the rectangles.
- **Complexity** in $O(s \max\{p^{1.75}, s^{3.5}\})$, versus $O(p^{3.5})$.

Caveats:

- No backward steps.
- If optimality conditions not tight enough, $|J| \leq s$.
- If $s \approx p$, active set strategy is more expensive.

Example of application, *dictionary learning*

Goal: learning simultaneously U, V such that $X \approx UV^\top$

- $X \in \mathbb{R}^{n \times p}$, n data points in $\mathbb{R}^p$.
- $U \in \mathbb{R}^{n \times r}$, decomposition coefficients.
- $V \in \mathbb{R}^{p \times r}$, r dictionary elements (the columns V^k of V).

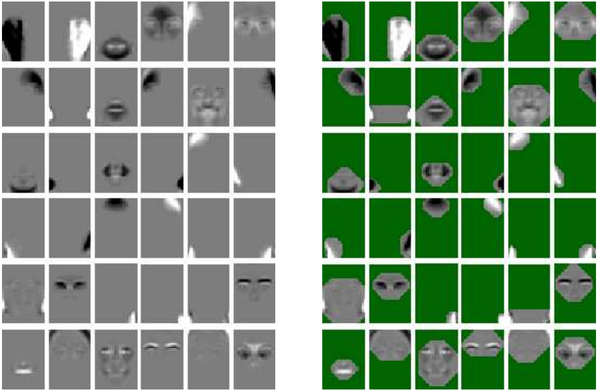


Figure: 36 learned dictionaries, AR face dataset (Martinez and Kak, 2001)

Conclusion

Structured sparsity-inducing norms:

- Sparsity-inducing norms can encode higher-order structure:
  - Not just **cardinality** or **fixed** group information.
- The structure prior is expressed in terms of **allowed nonzero patterns** by Ω.

Optimization:

- Take advantage of sparsity for computational purpose.
- Key quantity for optimization, **dual norm** Ω^*.
- Active set algorithm valid for any *black-box* solver.

Future directions:

- Can be used in other signal processing/learning tasks, as soon as structure information about the sparse decomposition is known.

  e.g., multi-task learning or matrix-factorization (Jenatton et al., 2009b).

References I

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008.

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. Technical report, arXiv:0808.3572, 2008. Submitted to IEEE Transactions on Information Theory.

J. P. Doignon and J. C. Falmagne. *Knowledge Spaces*. Springer-Verlag, 1998.

L. He and L. Carin. Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57:3488–3497, 2009.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlaps and graph Lasso. In *Proceedings of the 26th International Conference on Machine learning*, 2009.

R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, arXiv:0904.3523, 2009a.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. Technical report, arXiv:0909.1440, 2009b.

H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, 2007.

References II

A. M. Martinez and A. C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 1–22, 2009.

V. Roth and B. Fischer. The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th International Conference on Machine learning*, 2008.

M. Szafranski, Y. Grandvalet, and P. Morizet-Mahoudeaux. Hierarchical penalization. *Advances in neural information processing systems*, 20, 2007.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67, 2006.

P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 2008. To appear.