Network Flow Algorithms for Structured Sparsity

Julien Mairal[1] Rodolphe Jenatton[2]
Guillaume Obozinski[2] Francis Bach[2]

[1]UC Berkeley [2]INRIA - SIERRA Project-Team

Bellevue, ICML Workshop, July 2011

What this work is about

- **Sparse** and **structured** linear models.
- Optimization for group Lasso with overlapping groups.
- Links between sparse regularization and network flow optimization.

Related publications:

[1] J. Mairal, R. Jenatton, G. Obozinski and F. Bach. Network Flow Algorithms for Structured Sparsity. NIPS, 2010.

[2] R. Jenatton, J. Mairal, G. Obozinski and F. Bach. Proximal Methods for Hierarchical Sparse Coding. JMLR, to appear.

[3] R. Jenatton, J. Mairal, G. Obozinski and F. Bach. Proximal Methods for Sparse Hierarchical Dictionary Learning. ICML, 2010.

Part I: Introduction to Structured Sparsity

Sparse Linear Model: Machine Learning Point of View

Let $(y^i, \mathbf{x}^i)_{i=1}^n$ be a training set, where the vectors $\mathbf{x}^i$ are in $\mathbb{R}^p$ and are called features. The scalars y^i are in

- $\{-1, +1\}$ for **binary** classification problems.
- $\mathbb{R}$ for **regression** problems.

We assume there is a relation $y \approx \mathbf{w}^\top \mathbf{x}$, and solve

$$\min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(y^i, \mathbf{w}^\top \mathbf{x}^i)}_{\text{empirical risk}} + \underbrace{\lambda \Omega(\mathbf{w})}_{\text{regularization}} .$$

Sparse Linear Models: Machine Learning Point of View

A few examples:

Ridge regression:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^{n}(y^i - \mathbf{w}^\top \mathbf{x}^i)^2 + \lambda\|\mathbf{w}\|_2^2.$$

Linear SVM:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y^i\mathbf{w}^\top \mathbf{x}^i) + \lambda\|\mathbf{w}\|_2^2.$$

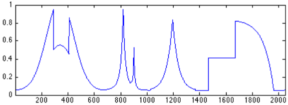
Logistic regression:
$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + e^{-y^i\mathbf{w}^\top \mathbf{x}^i}\right) + \lambda\|\mathbf{w}\|_2^2.$$

The squared ℓ_2-norm induces "**smoothness**" in $\mathbf{w}$. When one knows in advance that $\mathbf{w}$ should be sparse, one should use a **sparsity-inducing** regularization such as the ℓ_1-norm. [Chen et al., 1999, Tibshirani, 1996]

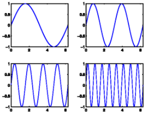
How can one add **a-priori knowledge** in the regularization?

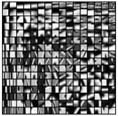
Sparse Linear Models: Signal Processing Point of View

Let $\mathbf{y}$ in $\mathbb{R}^n$ be a signal.



Let $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^p] \in \mathbb{R}^{n \times p}$ be a set of normalized "basis vectors".
We call it **dictionary**.





$\mathbf{X}$ is "adapted" to $\mathbf{y}$ if it can represent it with a few basis vectors—that is, there exists a **sparse vector** $\mathbf{w}$ in $\mathbb{R}^p$ such that $\mathbf{x} \approx \mathbf{X}\mathbf{w}$. We call $\mathbf{w}$ the **sparse code**.

$$\underbrace{\begin{pmatrix} \mathbf{y} \end{pmatrix}}_{\mathbf{y}\in\mathbb{R}^n} \approx \underbrace{\begin{pmatrix} \mathbf{x}^1 & \Big| & \mathbf{x}^2 & \Big| & \cdots & \Big| & \mathbf{x}^p \end{pmatrix}}_{\mathbf{X}\in\mathbb{R}^{n\times p}} \underbrace{\begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_p \end{pmatrix}}_{\mathbf{w}\in\mathbb{R}^p,\ \textbf{sparse}}$$

Sparse Linear Models: the Lasso/ Basis Pursuit

- Signal processing: $\mathbf{X}$ is a dictionary in $\mathbb{R}^{n\times p}$,

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

- Machine Learning:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^{n}(y^i-\mathbf{x}^{i\top}\mathbf{w})^2+\lambda\|\mathbf{w}\|_1=\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}^\top\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_1,$$

with $\mathbf{X} \triangleq [\mathbf{x}^1,\ldots,\mathbf{x}^n]$, and $\mathbf{y} \triangleq [y^1,\ldots,y^n]^\top$.

Useful tool in signal processing, machine learning, statistics, neuroscience,…as long as one wishes to **select** features.

Group Sparsity-Inducing Norms

$$\min_{\mathbf{w}\in\mathbb{R}^p} \overbrace{f(\mathbf{w})}^{\text{data fitting term}} + \lambda \underbrace{\Omega(\mathbf{w})}_{\text{sparsity-inducing norm}}$$

The most popular choice for Ω:

- The ℓ_1 norm, $\|\mathbf{w}\|_1 = \sum_{j=1}^{p} |\mathbf{w}_j|$.
- However, the ℓ_1 norm encodes poor information, just **cardinality**!

Another popular choice for Ω:

- The ℓ_1-ℓ_q norm [Turlach et al., 2005], with $q = 2$ or $q = \infty$

$$\sum_{g\in\mathcal{G}} \|\mathbf{w}_g\|_q \quad \text{with } \mathcal{G} \text{ a } \textbf{partition} \text{ of } \{1, \ldots, p\}.$$

- The ℓ_1-ℓ_q norm sets to zero **groups of non-overlapping variables** (as opposed to single variables for the ℓ_1 norm).

Structured Sparsity with Overlapping Groups

Warning: Under the name "structured sparsity" appear in fact significantly different formulations!

1. non-convex
   - zero-tree wavelets [Shapiro, 1993]
   - sparsity patterns are in a predefined collection: [Baraniuk et al., 2010]
   - select a union of groups: [Huang et al., 2009]
   - structure via Markov Random Fields: [Cehver et al., 2008]
2. convex
   - tree-structure: [Zhao et al., 2009]
   - non-zero patterns are a union of groups: [Jacob et al., 2009]
   - **zero patterns are a union of groups: [Jenatton et al., 2009]**
   - other norms: [Micchelli et al., 2010]

Sparsity-Inducing Norms

$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_q$$

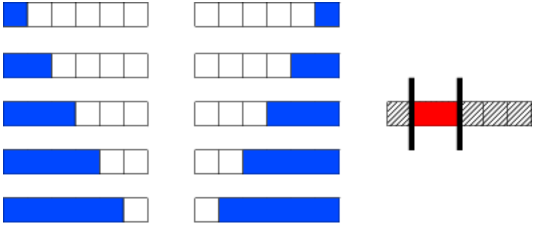
What happens when the groups overlap? [Jenatton et al., 2009]

- Inside the groups, the ℓ_2-norm (or ℓ_∞) does not promote sparsity.
- Variables belonging to the same groups are encouraged to be set to zero together.

Examples of set of groups $\mathcal{G}$

[Jenatton et al., 2009]

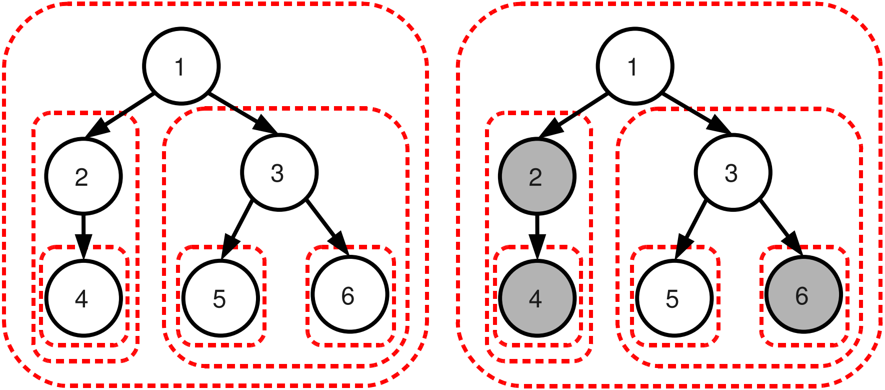
Selection of contiguous patterns on a sequence, $p = 6$.



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

Hierarchical Norms

[Zhao et al., 2009]



A node can be active only if its **ancestors are active**.

The selected patterns are **rooted subtrees.**

Part II: How do we optimize these cost functions?

Different strategies

$$\min_{\mathbf{w} \in \mathbb{R}^p} \; f(\mathbf{w}) + \lambda \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_q$$

- generic methods: QP, CP, subgradient descent.
- Augmented Lagrangian, ADMM [Mairal et al., 2011, Qi and Goldfarb, 2011]
- Nesterov smoothing technique [Chen et al., 2010]
- hierarchical case: proximal methods [Jenatton et al., 2010a]
- **for $q=\infty$: proximal gradient methods with network flow optimization.** [Mairal et al., 2010]
- also proximal gradient methods with inexact proximal operator [Jenatton et al., 2010a, Liu and Ye, 2010]
- for $q=2$, reweighted-ℓ_2 [Jenatton et al., 2010b, Micchelli et al., 2010]

First-order/proximal methods

$$\min_{\mathbf{w}\in\mathbb{R}^p} \; f(\mathbf{w}) + \lambda\Omega(\mathbf{w})$$

- f is strictly convex and differentiable with a Lipshitz gradient.
- Generalizes the idea of gradient descent

$$\mathbf{w}^{k+1} \leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{f(\mathbf{w}^k) + \nabla f(\mathbf{w}^k)^\top(\mathbf{w}-\mathbf{w}^k)}_{\text{linear approximation}} + \underbrace{\frac{L}{2}\|\mathbf{w}-\mathbf{w}^k\|_2^2}_{\text{quadratic term}} + \lambda\Omega(\mathbf{w})$$

$$\leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{w} - (\mathbf{w}^k - \frac{1}{L}\nabla f(\mathbf{w}^k))\|_2^2 + \frac{\lambda}{L}\Omega(\mathbf{w})$$

When $\lambda = 0$, $\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \frac{1}{L}\nabla f(\mathbf{w}^k)$, this is equivalent to a classical gradient descent step.

First-order/proximal methods

- They require solving efficiently the **proximal operator**

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{w})$$

- For the ℓ_1-norm, this amounts to a soft-thresholding:

$$\mathbf{w}_i^\star = \text{sign}(\mathbf{u}_i)(\mathbf{u}_i - \lambda)^+.$$

- There exists accelerated versions based on Nesterov optimal first-order method (gradient method with "extrapolation") [Beck and Teboulle, 2009, Nesterov, 2007, 1983]
- suited for large-scale experiments.

Tree-structured groups

Proposition [Jenatton, Mairal, Obozinski, and Bach, 2010a]

- If $\mathcal{G}$ is a *tree-structured* set of groups, i.e., $\forall g, h \in \mathcal{G}$,

$$g \cap h = \emptyset \text{ or } g \subset h \text{ or } h \subset g$$

- For $q = 2$ or $q = \infty$, we define Prox_g and Prox_Ω as

$$\text{Prox}_g : \mathbf{u} \to \arg\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{u} - \mathbf{w}\| + \lambda \|\mathbf{w}_g\|_q,$$

$$\text{Prox}_\Omega : \mathbf{u} \to \arg\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{u} - \mathbf{w}\| + \lambda \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_q,$$

- If the groups are sorted from the leaves to the root, then

$$\text{Prox}_\Omega = \text{Prox}_{g_m} \circ \ldots \circ \text{Prox}_{g_1}.$$

$\rightarrow$ Tree-structured regularization : Efficient linear time algorithm.

General Overlapping Groups for $q = \infty$

Dual formulation [Jenatton, Mairal, Obozinski, and Bach, 2010a]

The solutions $\mathbf{w}^\star$ and $\boldsymbol{\xi}^\star$ of the following optimization problems

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{u} - \mathbf{w}\| + \lambda \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_\infty, \qquad \texttt{(Primal)}$$

$$\min_{\boldsymbol{\xi} \in \mathbb{R}^{p \times |\mathcal{G}|}} \frac{1}{2}\|\mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g\|_2^2 \ \text{ s.t. } \ \forall g \in \mathcal{G},\ \|\boldsymbol{\xi}^g\|_1 \leq \lambda \ \text{ and } \ \boldsymbol{\xi}_j^g = 0 \text{ if } j \notin g, \qquad \texttt{(Dual)}$$

satisfy

$$\mathbf{w}^\star = \mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^{\star g}. \qquad \texttt{(Primal-dual relation)}$$

The dual formulation has more variables, but **no overlapping constraints**.

General Overlapping Groups for $q = \infty$

[Mairal, Jenatton, Obozinski, and Bach, 2010]

First Step: Flip the signs of u

The dual is equivalent to a **quadratic min-cost flow problem**.

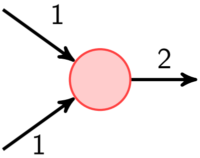
$$\min_{\boldsymbol{\xi} \in \mathbb{R}_+^{p \times |\mathcal{G}|}} \frac{1}{2} \|\mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g\|_2^2 \ \text{ s.t. } \ \forall g \in \mathcal{G}, \sum_{j \in g} \boldsymbol{\xi}_j^g \leq \lambda \ \text{ and } \ \boldsymbol{\xi}_j^g = 0 \ \text{ if } j \notin g,$$

Quick introduction to network flows

References:

- Ahuja, Magnanti and Orlin. Network Flows, 1993
- Bertsekas. Network Optimization, 1998.

A flow is a non-negative function on arcs that respects conservation constraints (Kirchhoff's law)

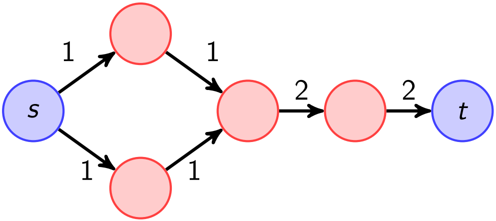


Quick introduction to network flows

References:

- Ahuja, Magnanti and Orlin. Network Flows, 1993
- Bertsekas. Network Optimization, 1998

A flow is a non-negative function on arcs that respects conservation constraints (Kirchhoff's law)



Flows usually go from a source node s to a sink node t.

Quick introduction to network flows

For a graph $G = (V, E)$:

- An arc (u, v) in E might have capacity constraints.
- Sending the maximum amount of flow in a network under capacity constraints is called **maximum flow problem**.
- This problem is dual to the **minimum cut problem**: finding a partition (V_s, V_t) of V, with $s \in V_s$ and $t \in V_t$ with minimal capacity (sum of capacities of all arcs going from V_s to V_t). [Ford and Fulkerson, 1956]
- it is a **linear program**, but there exists efficient dedicated algorithms [Goldberg and Tarjan, 1986] ($|V| = 1\,000\,000$ is "fine").
- Finding a flow that minimizes a linear cost is called a **minimum cost flow problem**.

General Overlapping Groups for $q = \infty$

Example: $\mathcal{G} = \{g = \{1, \ldots, p\}\}$

$$\min_{\boldsymbol{\xi}^g \in \mathbb{R}_+^p} \frac{1}{2}\|\mathbf{u} - \boldsymbol{\xi}^g\|_2^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} \boldsymbol{\xi}_j^g \leq \lambda.$$

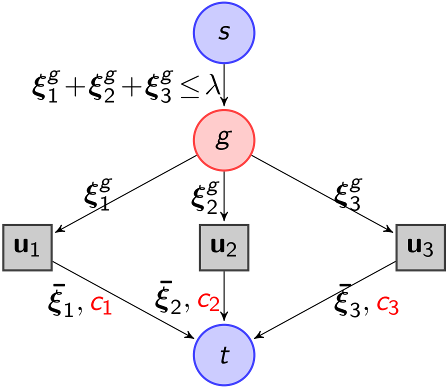


Figure: $\mathcal{G} = \{g = \{1, 2, 3\}\}$, $\forall j,\ c_j = \frac{1}{2}(\mathbf{u}_j - \bar{\boldsymbol{\xi}}_j)^2$.

General Overlapping Groups for $q = \infty$

Example with two overlapping groups

$$\min_{\boldsymbol{\xi} \in \mathbb{R}_+^{p \times |\mathcal{G}|}} \frac{1}{2} \|\mathbf{u} - \sum_{g \in \mathcal{G}} \boldsymbol{\xi}^g\|_2^2 \text{ s.t. } \forall g \in \mathcal{G}, \sum_{j \in g} \xi_j^g \leq \lambda \text{ and } \xi_j^g = 0 \text{ if } j \notin g,$$

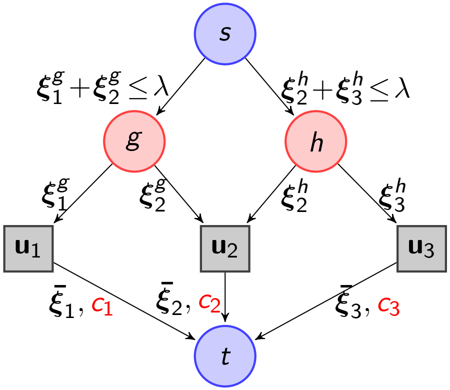


Figure: $\mathcal{G} = \{g = \{1,2\}, h = \{2,3\}\}$, $\forall j,\ c_j = \frac{1}{2}(\mathbf{u}_j - \bar{\xi}_j)^2$.

General Overlapping Groups for $q = \infty$

[Mairal, Jenatton, Obozinski, and Bach, 2010]

Main ideas of the algorithm: Divide and conquer

1. Solve a relaxed problem in linear time.
2. Test the feasability of the solution for the "non-relaxed" problem with a max-flow.
3. If the solution is feasible, it is optimal and stop the algorithm.
4. If not, find a minimum cut and removes the arcs along the cut.
5. Recursively process each subgraph defined by the cut.

The algorithm converges to the solution.

Related works:

- network flow optimization and total-variation [Chambolle and Darbon, 2009].
- similar algorithms exist in the optimization literature of submodular functions [Groenevelt, 1991].

Part III: Applications of Structured Sparsity

Background Subtraction

Given a video sequence, how can we remove foreground objects?

$$\underbrace{\mathbf{x}}_{\text{frame}} \approx \underbrace{\mathbf{X}\mathbf{w}}_{\text{linear combination of background frames}} + \underbrace{\mathbf{e}.}_{\text{foreground}}$$

Solved by

$$\min_{\mathbf{w}\in\mathbb{R}^p,\mathbf{e}\in\mathbb{R}^m} \frac{1}{2}\|\mathbf{x} - \mathbf{X}\mathbf{w} - \mathbf{e}\|_2^2 + \lambda_1\|\mathbf{w}\| + \lambda_2\Omega(\mathbf{e}).$$

Same idea as Wright et al. [2009] for robust face recognition, where $\Omega = \ell_1$.

Same task as Cehver et al. [2008], Huang et al. [2009] who used structured sparsity + background subtraction.

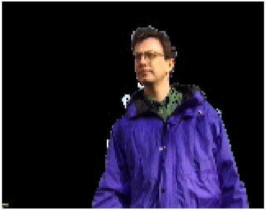
We are going to use **overlapping groups** with 3×3 neighborhoods to add spatial consistency.

Background Subtraction

(a) input

(b) estimated background

(c) foreground, ℓ_1

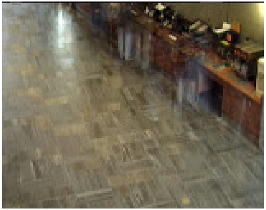


(d) foreground, ℓ_1+struct

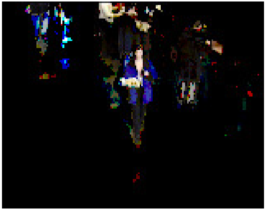
(e) other example

Background Subtraction

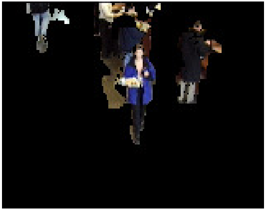
(a) input



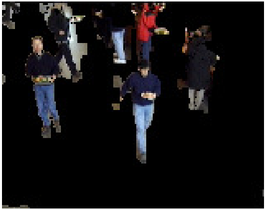
(b) estimated background



(c) foreground, ℓ_1



(d) foreground, ℓ_1+struct



(e) other example

Speed Benchmark

[Mairal, Jenatton, Obozinski, and Bach, 2011]

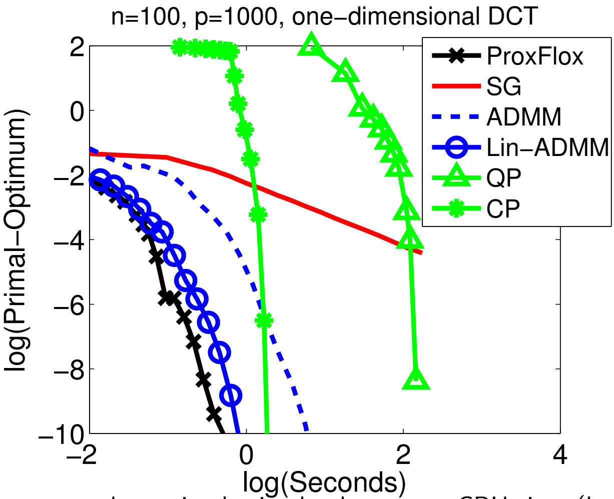


Figure: Distance to the optimal primal value versus CPU time (log-log scale).

Speed Benchmark

[Mairal, Jenatton, Obozinski, and Bach, 2011]

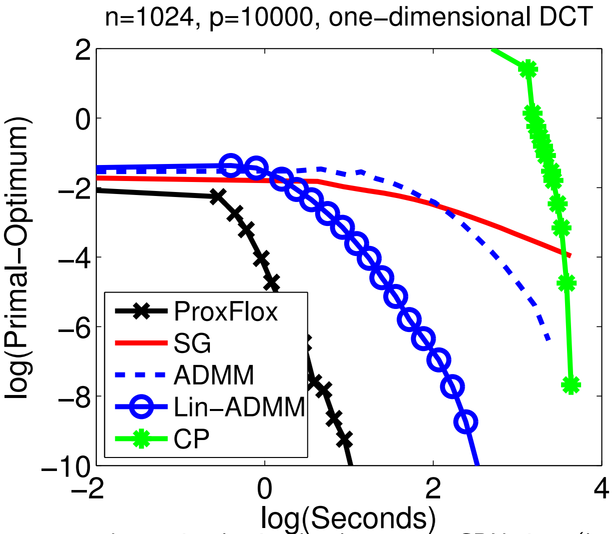


Figure: Distance to the optimal primal value versus CPU time (log-log scale).

Speed Benchmark

[Mairal, Jenatton, Obozinski, and Bach, 2011]

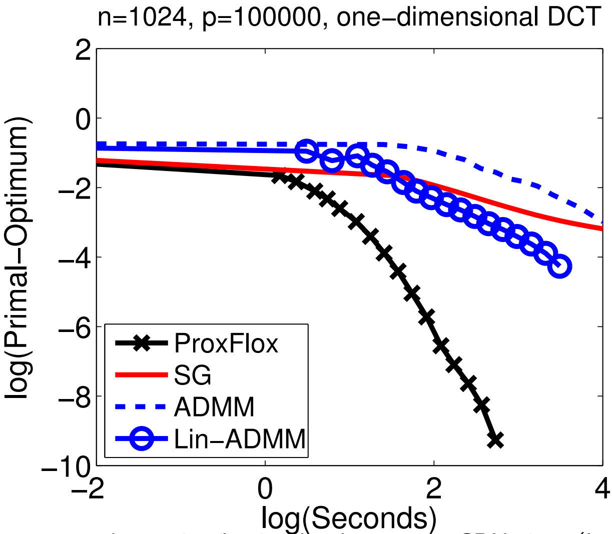


Figure: Distance to the optimal primal value versus CPU time (log-log scale).

Structured Dictionary Learning

$$\min_{\mathbf{X}\in\mathcal{X},\mathbf{W}\in\mathbb{R}^{p\times n}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{y}^i - \mathbf{X}\mathbf{w}^i\|_2^2 + \lambda\Omega(\mathbf{w}^i).$$

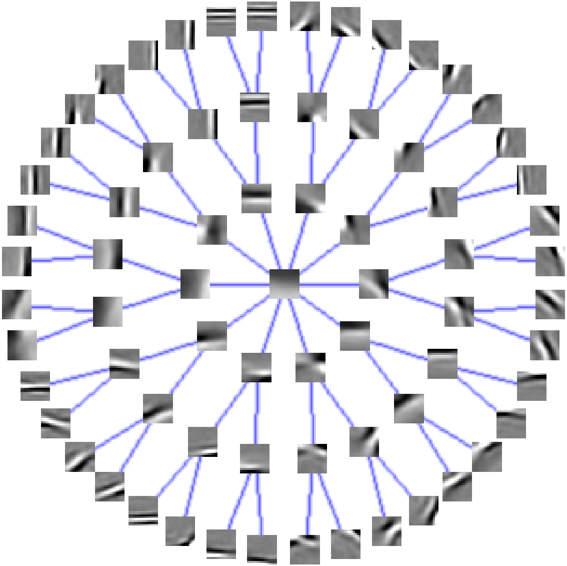
- structure **X**? [Jenatton et al., 2010b]
- structure **W**? [Kavukcuoglu et al., 2009, Jenatton et al., 2010a, Mairal et al., 2011]

Optimization

- Alternate minimization between **X** and **W**.
- online learning techniques [Olshausen and Field, 1997, Mairal et al., 2009].

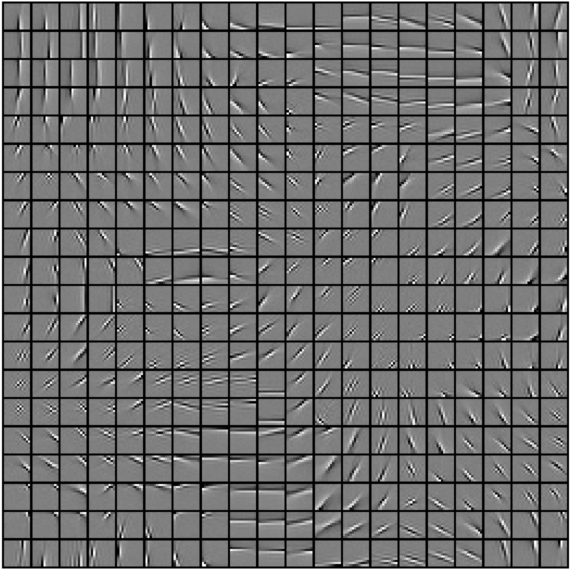
Hierarchical Dictionary Learning

[Jenatton, Mairal, Obozinski, and Bach, 2010a]



Topographic Dictionary Learning

[Mairal, Jenatton, Obozinski, and Bach, 2011]



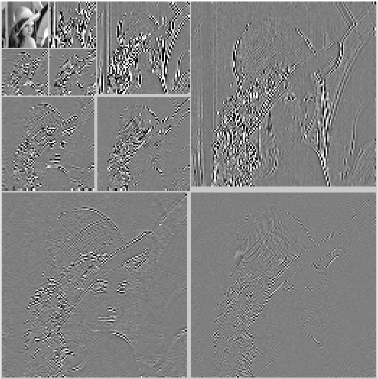
Wavelet denoising with structured sparsity

[Mairal, Jenatton, Obozinski, and Bach, 2011]

Classical wavelet denoising [Donoho and Johnstone, 1995]:

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

When $\mathbf{X}$ is orthogonal, the solution is obtained via **soft-thresholding**.



Wavelet denoising with hierarchical norms

[Mairal, Jenatton, Obozinski, and Bach, 2011]

Benchmark on a database of 12 standard images:

	PSNR			IPSNR vs. ℓ_1		
σ	ℓ_1	Ω_{tree}	Ω_{grid}	ℓ_1	Ω_{tree}	Ω_{grid}
5	35.67	35.98	**36.15**	$0.00 \pm .0$	$0.31 \pm .18$	$\mathbf{0.48 \pm .25}$
10	31.00	31.60	**31.88**	$0.00 \pm .0$	$0.61 \pm .28$	$\mathbf{0.88 \pm .28}$
25	25.68	26.77	**27.07**	$0.00 \pm .0$	$1.09 \pm .32$	$\mathbf{1.38 \pm .26}$
50	22.37	23.84	**24.06**	$0.00 \pm .0$	$1.47 \pm .34$	$\mathbf{1.68 \pm .41}$
100	19.64	21.49	**21.56**	$0.00 \pm .0$	$1.85 \pm .28$	$\mathbf{1.92 \pm .29}$

CUR Matrix Decomposition

[Mairal, Jenatton, Obozinski, and Bach, 2011]

CUR matrix decomposition [Mahoney and Drineas, 2009]

Let $\mathbf{X}$ in $\mathbb{R}^{n \times p}$. The goal is to find an low-rank approximation:

$$\mathbf{X} \approx \mathbf{CUR},$$

where $\mathbf{C}$ and $\mathbf{R}$ are respectively subsets of columns and rows of $\mathbf{X}$.

Bien et al. [2010] uses the Group Lasso for decomposing $\mathbf{X} \approx \mathbf{CW}$.
We use here structured sparsity:

$$\min_{\mathbf{W} \in \mathbb{R}^{p \times n}} \frac{1}{2}\|\mathbf{X} - \mathbf{XWX}\|_F^2 + \lambda_{\text{row}} \sum_{i=1}^{n} \|\mathbf{W}^i\|_\infty + \lambda_{\text{col}} \sum_{j=1}^{p} \|\mathbf{W}_j\|_\infty.$$

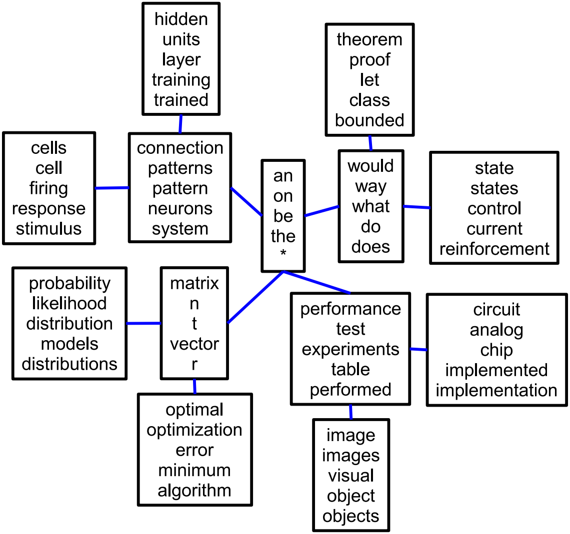
The performance is experimentally similar to the sampling procedure of Mahoney and Drineas [2009].

Hierarchical Topic Models for text corpora

[Jenatton, Mairal, Obozinski, and Bach, 2010a]

- Each document is modeled through word counts
- Low-rank matrix factorization of word-document matrix
- Probabilistic topic models such as Latent Dirichlet Allocation [Blei et al., 2003]
- Organise the topics in a tree.
- Previously approached using non-parametric Bayesian methods (Hierarchical Chinese Restaurant Process and nested Dirichlet Process): [Blei et al., 2010]
- **Can we achieve similar performance with simple matrix factorization formulation?**

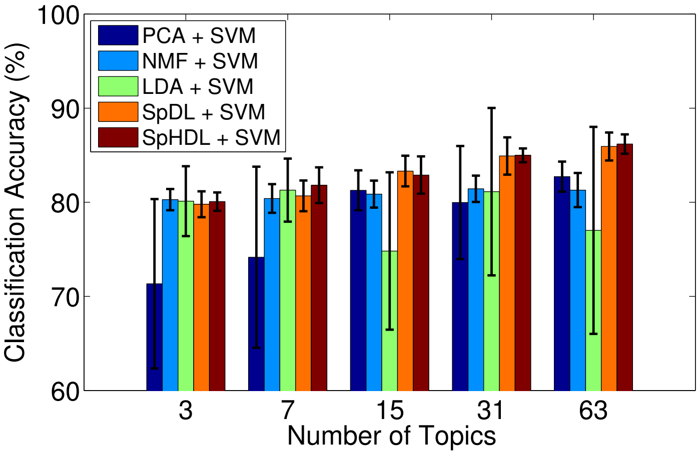
Tree of Topics



Classification based on topics

Comparison on predicting newsgroup article subjects

- 20 newsgroup articles (1425 documents, 13312 words)



Conclusion / Discussion

- Network Flow Optimization can handle structured sparse regularization functions based on the ℓ_∞-norm.
- Hierarchical norms lead to the same complexity as the Lasso.
- We have presented new applications to matrix factorization, dictionary learning, topic modelling...

Code SPAMS available: `http://www.di.ens.fr/willow/SPAMS/`**, now open-source!**

SPAMS toolbox (open-source)

- C++ interfaced with Matlab.
- proximal gradient methods for ℓ_0, ℓ_1, **elastic-net, fused-Lasso, group-Lasso, tree group-Lasso, tree-ℓ_0, sparse group Lasso, overlapping group Lasso...**
- ...for **square, logistic, multi-class logistic** loss functions.
- handles sparse matrices,
- provides duality gaps.
- also coordinate descent, block coordinate descent algorithms.
- fastest available implementation of **OMP** and **LARS**.
- dictionary learning and matrix factorization (NMF, sparse PCA).
- fast projections onto some convex sets.

Try it! `http://www.di.ens.fr/willow/SPAMS/`

References I

R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 2010. to appear.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

J. Bien, Y. Xu, and M. W. Mahoney. CUR from a sparse optimization viewpoint. In *Advances in Neural Information Processing Systems*, 2010.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

D. Blei, T. Griffiths, and M. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010.

References II

V. Cehver, M. F. Duarte, C. Hegde, and R. G. Baraniuk. Sparse signal recovery usingmarkov random fields. In *Advances in Neural Information Processing Systems*, 2008.

A. Chambolle and J. Darbon. On total variation minimization and surface evolution using parametric maximal flows. *International Journal of Computer Vision*, 84(3), 2009.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1999.

X. Chen, Q. Lin, S. Kim, J.G. Carbonell, and E.P. Xing. An efficient proximal gradient method for general structured sparse learning. *Preprint arXiv:1005.4717*, 2010.

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

References III

L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8(3):399–404, 1956.

A. V. Goldberg and R. E. Tarjan. A new approach to the maximum flow problem. In *Proc. of ACM Symposium on Theory of Computing*, pages 136–146, 1986.

H. Groenevelt. Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *Europeran Journal of Operations Research*, pages 227–236, 1991.

J. Huang, Z. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

References IV

R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, 2009. preprint arXiv:0904.3523v1.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010a.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *AISTATS*, 2010b.

K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *Proceedings of CVPR*, 2009.

J. Liu and J. Ye. Fast overlapping group lasso. *Preprint arXiv:1009.0306*, 2010.

References V

M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697, 2009.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *Preprint arXiv:1104.1872*, 2011.

C. A. Micchelli, J. M. Morales, and M. Pontil. A family of penalty functions for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010.

References VI

Y. Nesterov. A method for solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, 27:372–376, 1983.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, CORE, 2007.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37: 3311–3325, 1997.

Z. Qi and D. Goldfarb. Structured sparsity via alternating directions methods. *Preprint arXiv:1105.0728*, 2011.

J.M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12): 3445–3462, 1993.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

References VII

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 210–227, 2009.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. 37(6A):3468–3497, 2009.