Topographic Dictionary Learning with Structured Sparsity

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

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

San Diego, Wavelets and Sparsity XIV, August 2011

What this work is about

- Group sparsity with overlapping groups.
- Hierarchical, topographic dictionary learning,
- More generally: structured dictionaries of natural image patches.

Related publications:

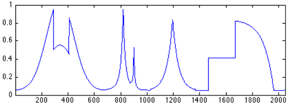
[1] J. Mairal, R. Jenatton, G. Obozinski and F. Bach. Network Flow Algorithms for Structured Sparsity. NIPS, 2010.

[2] R. Jenatton, J. Mairal, G. Obozinski and F. Bach. Proximal Methods for Hierarchical Sparse Coding. JMLR, 2011.

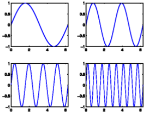
Part I: Introduction to Dictionary Learning

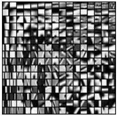
What is a Sparse Linear Model?

Let $\mathbf{x}$ in $\mathbb{R}^m$ be a signal.



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





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

$$\underbrace{\begin{pmatrix} \mathbf{x} \end{pmatrix}}_{\mathbf{x} \in \mathbb{R}^m} \approx \underbrace{\begin{pmatrix} \mathbf{d}^1 & \mathbf{d}^2 & \cdots & \mathbf{d}^p \end{pmatrix}}_{\mathbf{D} \in \mathbb{R}^{m \times p}} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix}}_{\boldsymbol{\alpha} \in \mathbb{R}^p, \textbf{sparse}}$$

The Sparse Decomposition Problem

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \underbrace{\frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2}_{\text{data fitting term}} + \underbrace{\lambda\psi(\boldsymbol{\alpha})}_{\text{sparsity-inducing regularization}}$$

ψ induces sparsity in $\boldsymbol{\alpha}$:

- the ℓ_0 "pseudo-norm". $\|\boldsymbol{\alpha}\|_0 \triangleq \#\{i \text{ s.t. } \boldsymbol{\alpha}_i \neq 0\}$ (NP-hard)
- the ℓ_1 norm. $\|\boldsymbol{\alpha}\|_1 \triangleq \sum_{i=1}^p |\boldsymbol{\alpha}_i|$ (convex),
- ...

This is a selection problem. When ψ is the ℓ_1-norm, the problem is called Lasso [Tibshirani, 1996] or basis pursuit [Chen et al., 1999]

Sparse representations for image restoration

Designed dictionaries

[Haar, 1910], [Zweig, Morlet, Grossman ∼70s], [Meyer, Mallat, Daubechies, Coifman, Donoho, Candes ∼80s-today]...
Wavelets, Curvelets, Wedgelets, Bandlets, ...lets

Learned dictionaries of patches

[Olshausen and Field, 1997, Engan et al., 1999, Lewicki and Sejnowski, 2000, Aharon et al., 2006],...

$$\min_{\boldsymbol{\alpha}^i, \mathbf{D} \in \mathcal{D}} \sum_{i=1}^{n} \underbrace{\frac{1}{2}\|\mathbf{x}^i - \mathbf{D}\boldsymbol{\alpha}^i\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda\psi(\boldsymbol{\alpha}^i)}_{\text{sparsity}}$$

- $\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_0$ ("ℓ_0 pseudo-norm")
- $\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$ (ℓ_1 norm)

Sparse representations for image restoration

Grayscale vs color image patches

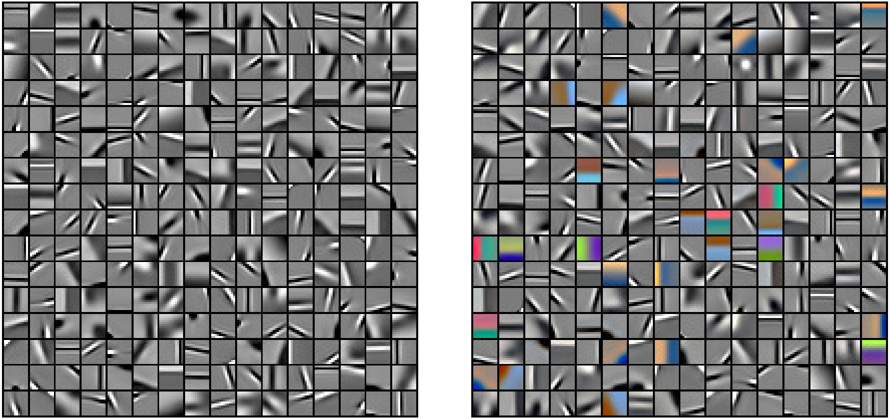


Figure: Left: learned on grayscale image patches. Right: learned on color image patches (after removing the mean color from each patch)

Algorithms

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^{p\times n}\\ \mathbf{D}\in\mathcal{D}}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{x}^i - \mathbf{D}\boldsymbol{\alpha}^i\|_2^2 + \lambda\psi(\boldsymbol{\alpha}^i).$$

How do we optimize that?

- alternate between $\mathbf{D}$ and $\boldsymbol{\alpha}$ [Engan et al., 1999], or other variants [Elad and Aharon, 2006]
- online learning [Olshausen and Field, 1997, Mairal et al., 2009, Skretting and Engan, 2010]

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

Part II: Introduction to Structured Sparsity (Let us play with ψ)

Group Sparsity-Inducing Norms

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \underbrace{\psi(\boldsymbol{\alpha})}_{\text{sparsity-inducing norm}}$$

The most popular choice for ψ:

- The ℓ_1 norm, $\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$.
- 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 $q = \infty$

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

- 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]

Structured Sparsity with Overlapping Groups

$$\psi(\boldsymbol{\alpha}) = \sum_{g \in \mathcal{G}} \|\boldsymbol{\alpha}_g\|_q$$

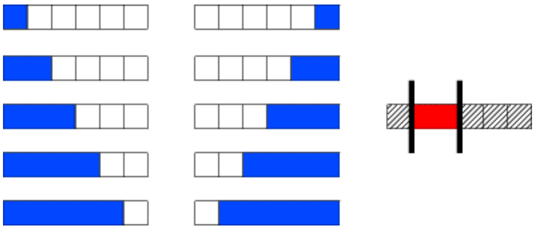
What happens when the groups overlap? [Jenatton et al., 2009]

- 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]

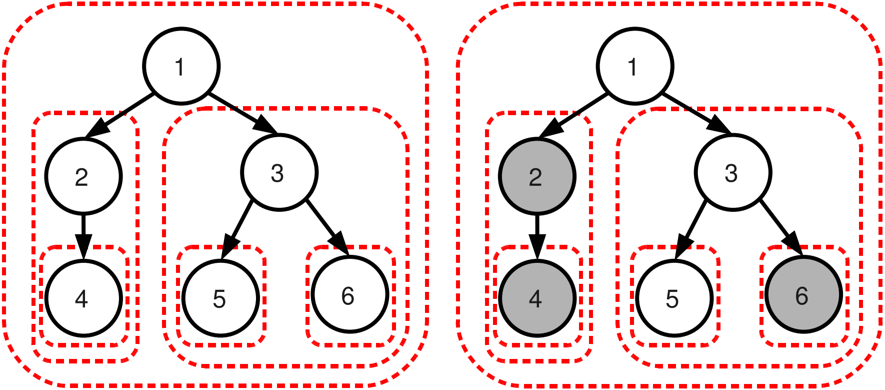
Selection of contiguous patterns on a sequence, $p = 6$.



- $\mathcal{G}$ 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.**

Algorithms/Difficulties

[Jenatton et al., 2010, Mairal et al., 2011]

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{g \in \mathcal{G}} \|\boldsymbol{\alpha}_g\|_q.$$

The function is convex non-differentiable; the sum is a sum of simple **non-separable** regularizers.

How do we optimize that?

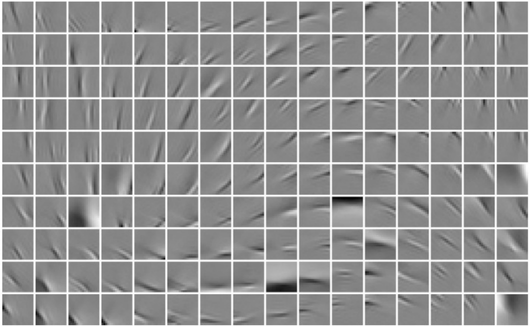
- hierarchical norms: **same complexity as ℓ_1** with proximal methods.
- general case: Augmenting Lagrangian Techniques.
- general case with ℓ_∞-norms: proximal methods combine with network flow optimization.

Also implemented in the toolbox SPAMS

Part III: Learning Structured Dictionaries

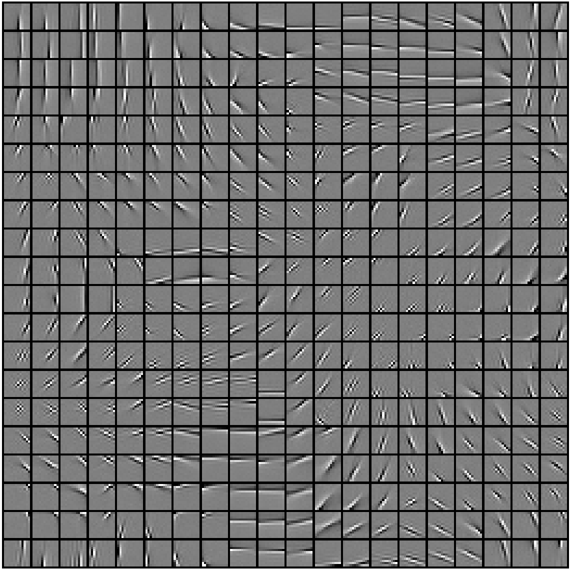
Topographic Dictionary Learning

- [Kavukcuoglu et al., 2009]: organize the dictionary elements on a 2D-grids and use ψ with $e \times e$ overlapping groups.
- [Garrigues and Olshausen, 2010]: sparse coding + probabilistic model to model lateral interactions.
- topographic ICA by Hyvärinen et al. [2001]:



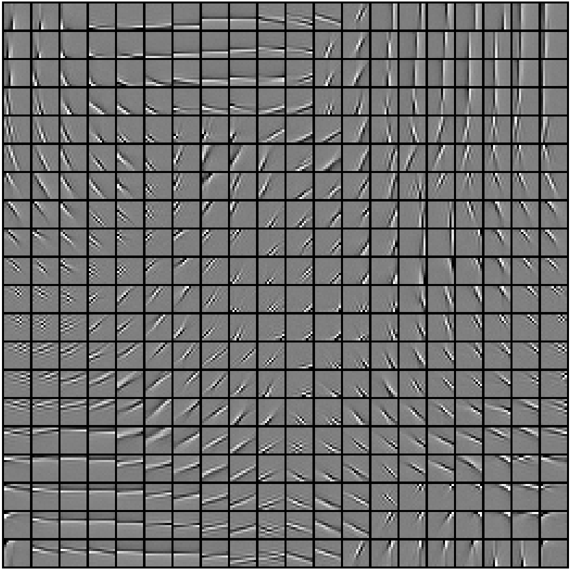
Topographic Dictionary Learning

[Mairal, Jenatton, Obozinski, and Bach, 2011], 3×3-neighborhoods



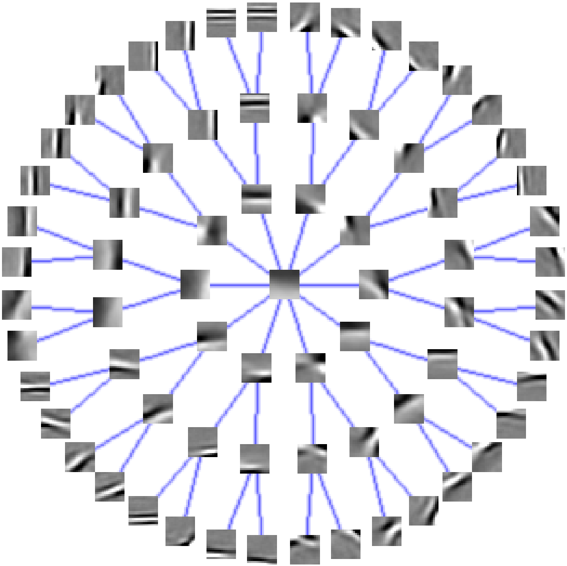
Topographic Dictionary Learning

[Mairal, Jenatton, Obozinski, and Bach, 2011],4 $\times$ 4-neighborhoods



Hierarchical Dictionary Learning

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



Conclusion / Discussion

- Structured sparsity is a natural framework for learning structured dictionaries...
- ...and has efficient optimization tools.
- other applications in natural language processing, bio-informatics, neuroscience...

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

M. Aharon, M. Elad, and A. M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, November 2006.

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.

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.

References II

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

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 54(12):3736–3745, December 2006.

K. Engan, S. O. Aase, and J. H. Husoy. Frame based signal compression using method of optimal directions (MOD). In *Proceedings of the 1999 IEEE International Symposium on Circuits Systems*, volume 4, 1999.

P. Garrigues and B. Olshausen. Group sparse coding with a laplacian scale mixture prior. In *Advances in Neural Information Processing Systems*, 2010.

A. Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69:331–371, 1910.

References III

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

A. Hyvärinen, P. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

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.

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)*, 2010.

References IV

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

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

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. 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 V

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.

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

K. Skretting and K. Engan. Recursive least squares dictionary learning algorithm. *IEEE Transactions on Signal Processing*, 58(4):2121–2130, 2010.

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

References VI

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

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

First-order/proximal methods

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

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

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

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

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

First-order/proximal methods

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

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

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

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

- 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, 2010]

- 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

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

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

- 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 $q = \infty$

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

Dual formulation

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

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

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

satisfy

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

The dual formulation has more variables, but is equivalent to **quadratic min-cost flow problem**.