Structured sparsity-inducing norms through submodular functions

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Joint work with R. Jenatton, J. Mairal, G. Obozinski

IPAM - February 2013

Outline

- **Introduction: Sparse methods for machine learning**
  - Need for structured sparsity: Going beyond the ℓ_1-norm
- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Unified algorithms and analysis
  - Applications to signal processing and machine learning
- **Extensions**
  - Shaping level sets through symmetric submodular functions
  - ℓ_2-norm relaxation of combinatorial penalties

Sparsity in supervised machine learning

- Observed data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$
  - Response vector $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$
  - Design matrix $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$
- Regularized empirical risk minimization:

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \Omega(w) = \boxed{\min_{w \in \mathbb{R}^p} \; L(y, Xw) + \lambda \Omega(w)}$$

- Norm Ω to promote sparsity
  - square loss + ℓ_1-norm $\Rightarrow$ basis pursuit in signal processing (Chen et al., 2001), Lasso in statistics/machine learning (Tibshirani, 1996)
  - Proxy for interpretability
  - Allow high-dimensional inference: $\boxed{\log p = O(n)}$

Sparsity in unsupervised machine learning

- **Multiple** responses/signals $y = (y^1, \dots, y^k) \in \mathbb{R}^{n \times k}$

$$\min_{w^1, \dots, w^k \in \mathbb{R}^p} \sum_{j=1}^{k} \left\{ L(y^j, Xw^j) + \lambda \Omega(w^j) \right\}$$

Sparsity in unsupervised machine learning

- **Multiple** responses/signals $y = (y^1, \dots, y^k) \in \mathbb{R}^{n \times k}$

$$\min_{w^1,\dots,w^k \in \mathbb{R}^p} \sum_{j=1}^{k} \left\{ L(y^j, Xw^j) + \lambda \Omega(w^j) \right\}$$

- **Only responses are observed** $\Rightarrow$ **Dictionary learning**
  - Learn $X = (x^1, \dots, x^p) \in \mathbb{R}^{n \times p}$ such that $\forall j, \ \|x^j\|_2 \leqslant 1$

$$\min_{X=(x^1,\dots,x^p)} \min_{w^1,\dots,w^k \in \mathbb{R}^p} \sum_{j=1}^{k} \left\{ L(y^j, Xw^j) + \lambda \Omega(w^j) \right\}$$

  - Olshausen and Field (1997); Elad and Aharon (2006); Mairal et al. (2009a)

- **sparse PCA**: replace $\|x^j\|_2 \leqslant 1$ by $\Theta(x^j) \leqslant 1$

Sparsity in signal processing

- **Multiple** responses/signals $x = (x^1, \ldots, x^k) \in \mathbb{R}^{n \times k}$

$$\min_{\alpha^1,\ldots,\alpha^k \in \mathbb{R}^p} \sum_{j=1}^{k} \left\{ L(x^j, D\alpha^j) + \lambda \Omega(\alpha^j) \right\}$$

- **Only responses are observed** $\Rightarrow$ **Dictionary learning**
  - Learn $D = (d^1, \ldots, d^p) \in \mathbb{R}^{n \times p}$ such that $\forall j,\ \|d^j\|_2 \leqslant 1$

$$\min_{D=(d^1,\ldots,d^p)} \min_{\alpha^1,\ldots,\alpha^k \in \mathbb{R}^p} \sum_{j=1}^{k} \left\{ L(x^j, D\alpha^j) + \lambda \Omega(\alpha^j) \right\}$$

  - Olshausen and Field (1997); Elad and Aharon (2006); Mairal et al. (2009a)

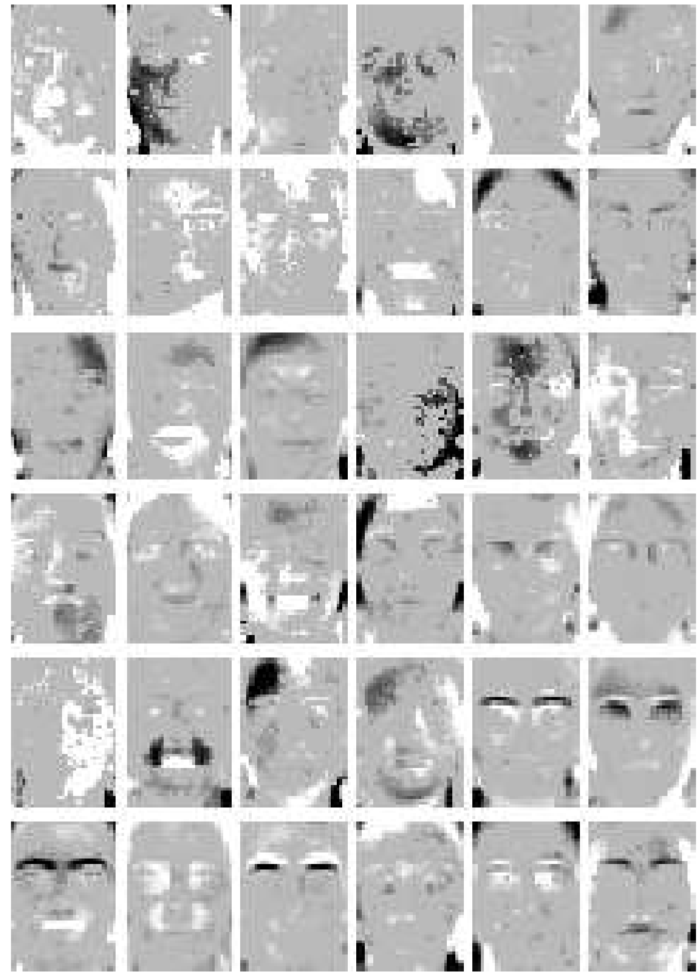
- **sparse PCA**: replace $\|d^j\|_2 \leqslant 1$ by $\Theta(d^j) \leqslant 1$

Why structured sparsity?

- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements "organized" in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)

Structured sparse PCA (Jenatton et al., 2009b)

raw data

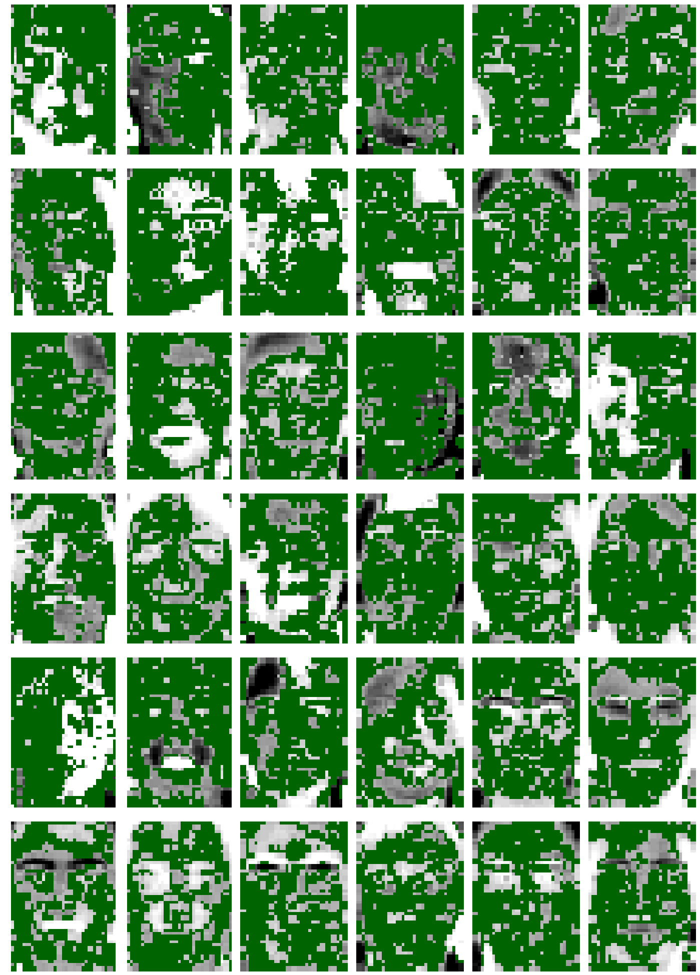


sparse PCA

- Unstructured sparse PCA $\Rightarrow$ many zeros do not lead to better interpretability

Structured sparse PCA (Jenatton et al., 2009b)

raw data

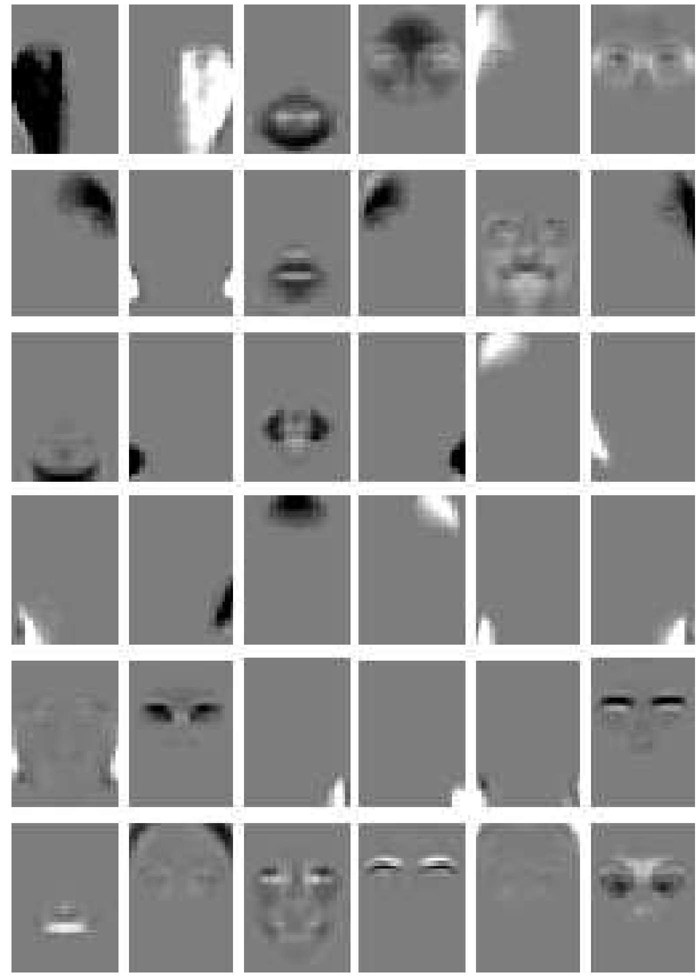


sparse PCA

- Unstructed sparse PCA $\Rightarrow$ many zeros do not lead to better interpretability

Structured sparse PCA (Jenatton et al., 2009b)

raw data

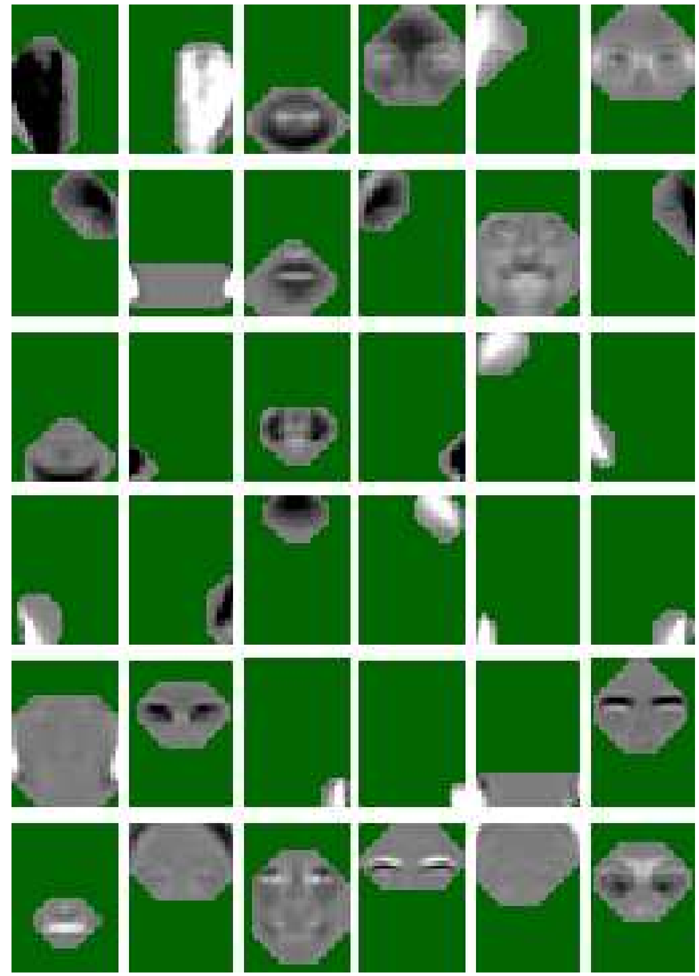


Structured sparse PCA

- Enforce selection of convex nonzero patterns ⇒ robustness to occlusion in face identification

Structured sparse PCA (Jenatton et al., 2009b)

raw data



Structured sparse PCA

- Enforce selection of convex nonzero patterns ⇒ robustness to occlusion in face identification

Why structured sparsity?

- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements "organized" in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)

Why structured sparsity?

- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements "organized" in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)
- **Stability and identifiability**
  - Optimization problem $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda \|w\|_1$ is unstable
  - "Codes" w^j often used in later processing (Mairal et al., 2009c)
- **Prediction or estimation performance**
  - When prior knowledge matches data (Haupt and Nowak, 2006; Baraniuk et al., 2008; Jenatton et al., 2009a; Huang et al., 2009)
- **Numerical efficiency**
  - Non-linear variable selection with 2^p subsets (Bach, 2008)

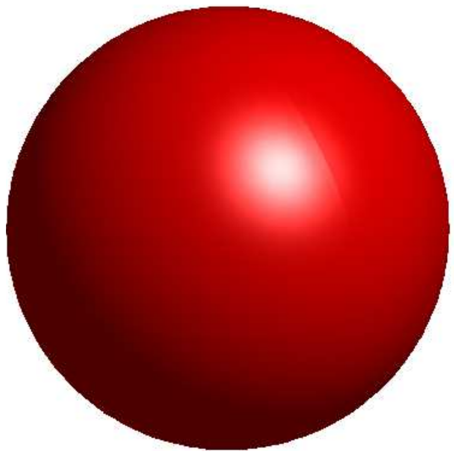
Different types of structured sparsity

- **Enforce specific sets of non-zeros**
  - e.g., group Lasso (Yuan and Lin, 2006)
  - overlapping group Lasso (Jenatton et al., 2009a)
- **Enforce specific level sets**
  - e.g., total variation (Rudin et al., 1992; Chambolle, 2004)
- **Enforce specific matrix factorizations**
  - e.g., nuclear norm (Srebro et al., 2005; Candès and Recht, 2009)
  - Sparse extensions (Bach et al., 2008)

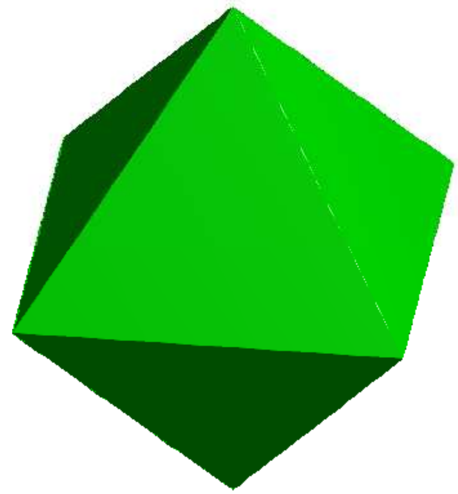
Classical approaches to structured sparsity

- **Many application domains**
  - Computer vision (Cevher et al., 2008; Mairal et al., 2009b)
  - Neuro-imaging (Gramfort and Kowalski, 2009; Jenatton et al., 2011)
  - Bio-informatics (Rapaport et al., 2008; Kim and Xing, 2010)
- **Non-convex approaches**
  - Haupt and Nowak (2006); Baraniuk et al. (2008); Huang et al. (2009)
- **Convex approaches**
  - Design of sparsity-inducing norms

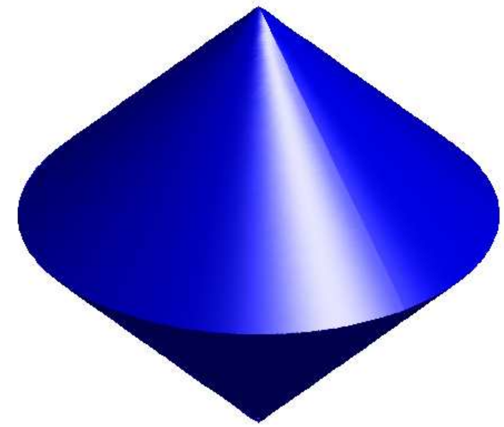
Unit norm balls
Geometric interpretation



$\|w\|_2$



$\|w\|_1$



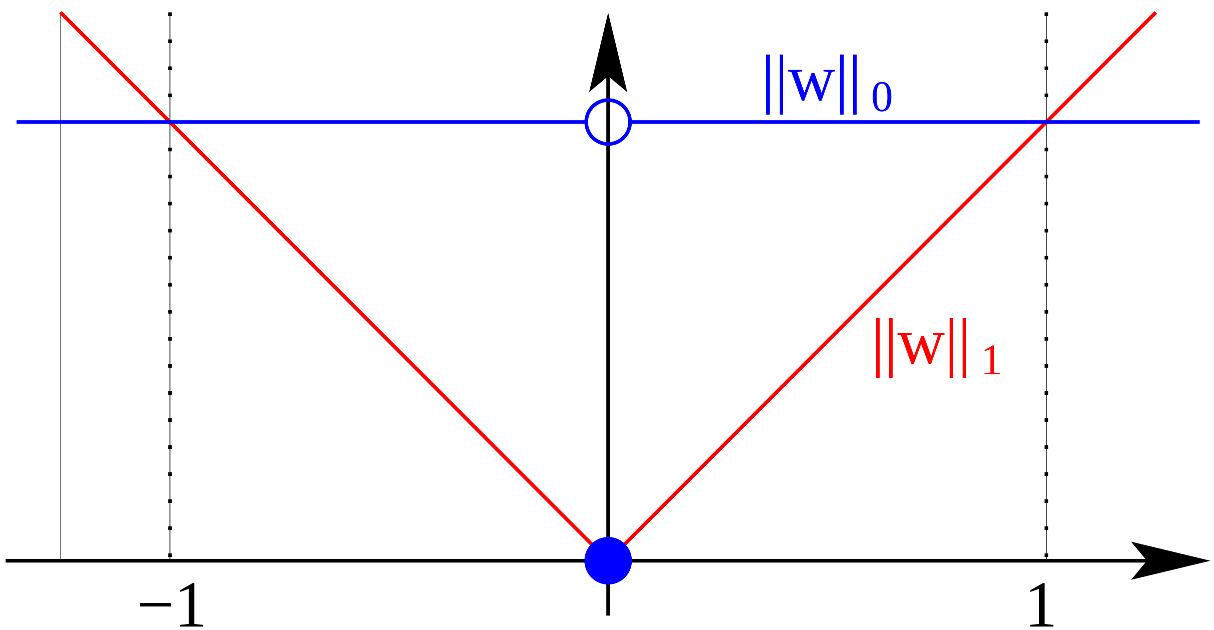
$\sqrt{w_1^2 + w_2^2} + |w_3|$

Outline

- **Introduction: Sparse methods for machine learning**
  - Need for structured sparsity: Going beyond the ℓ_1-norm
- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Unified algorithms and analysis
  - Applications to signal processing and machine learning
- **Extensions**
  - Shaping level sets through symmetric submodular functions
  - ℓ_2-norm relaxation of combinatorial penalties

ℓ_1-norm = convex envelope of cardinality of support

- Let $w \in \mathbb{R}^p$. Let $V = \{1, \dots, p\}$ and $\mathrm{Supp}(w) = \{j \in V, \; w_j \neq 0\}$
- **Cardinality of support**: $\|w\|_0 = \mathrm{Card}(\mathrm{Supp}(w))$
- Convex envelope = largest convex lower bound (see, e.g., Boyd and Vandenberghe, 2004)



- ℓ_1-norm = convex envelope of ℓ_0-quasi-norm on the ℓ_∞-ball $[-1, 1]^p$

Convex envelopes of general functions of the support (Bach, 2010)

- Let $F : 2^V \to \mathbb{R}$ be a **set-function**
  - Assume F is **non-decreasing** (i.e., $A \subset B \Rightarrow F(A) \leqslant F(B)$)
  - Explicit prior knowledge on supports (Haupt and Nowak, 2006; Baraniuk et al., 2008; Huang et al., 2009)

- Define $\Theta(w) = F(\text{Supp}(w))$: How to get its convex envelope?
  1. Possible if F is also **submodular**
  2. Allows **unified** theory and algorithm
  3. Provides **new** regularizers

Submodular functions (Fujishige, 2005; Bach, 2011)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\begin{aligned} & \forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B) \\ \Leftrightarrow \quad & \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing} \end{aligned}$$

Submodular functions (Fujishige, 2005; Bach, 2010)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B)$$
$$\Leftrightarrow \quad \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

- **Intuition 1**: defined like concave functions ("diminishing returns")
  - Example: $F : A \mapsto g(\mathrm{Card}(A))$ is submodular if g is concave

Submodular functions (Fujishige, 2005; Bach, 2010)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\begin{aligned} & \forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B) \\ \Leftrightarrow \ & \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing} \end{aligned}$$

- **Intuition 1**: defined like concave functions ("diminishing returns")
  – Example: $F : A \mapsto g(\mathrm{Card}(A))$ is submodular if g is concave

- **Intuition 2**: behave like convex functions
  – Polynomial-time minimization, conjugacy theory

Submodular functions (Fujishige, 2005; Bach, 2010)

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

$$\forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B)$$
$$\Leftrightarrow \quad \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

- **Intuition 1**: defined like concave functions ("diminishing returns")
  - Example: $F : A \mapsto g(\mathrm{Card}(A))$ is submodular if g is concave

- **Intuition 2**: behave like convex functions
  - Polynomial-time minimization, conjugacy theory

- Used in several areas of signal processing and machine learning
  - Total variation/graph cuts (Chambolle, 2005; Boykov et al., 2001)
  - Optimal design (Krause and Guestrin, 2005)

Submodular functions - Examples

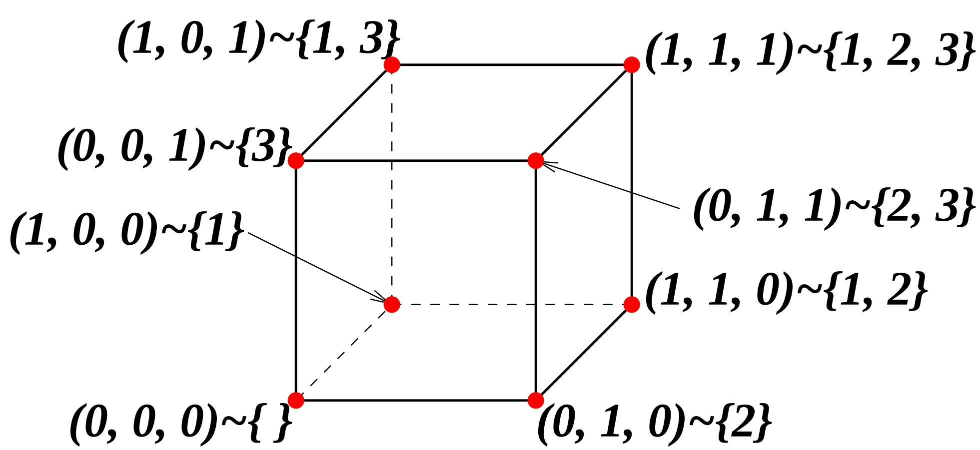
- Concave functions of the cardinality: $g(|A|)$
- Cuts
- Entropies
  - $H((X_k)_{k\in A})$ from p random variables $X_1,\dots,X_p$
  - Gaussian variables $H((X_k)_{k\in A}) \propto \log\det\Sigma_{AA}$
  - Functions of eigenvalues of sub-matrices
- Network flows
  - Efficient representation for set covers
- Rank functions of matroids

Submodular functions - Lovász extension

- Subsets may be identified with elements of $\{0,1\}^p$
- Given any set-function F and w such that $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$, define:

$$f(w) \; = \; \sum_{k=1}^{p} w_{j_k} [F(\{j_1,\dots,j_k\}) - F(\{j_1,\dots,j_{k-1}\})]$$

  - If $w = 1_A$, $f(w) = F(A) \Rightarrow$ extension from $\{0,1\}^p$ to $\mathbb{R}^p$
  - f is piecewise affine and positively homogeneous



Submodular functions - Lovász extension

- Subsets may be identified with elements of $\{0,1\}^p$
- Given any set-function F and w such that $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$, define:

$$f(w) = \sum_{k=1}^{p} w_{j_k} [F(\{j_1,\dots,j_k\}) - F(\{j_1,\dots,j_{k-1}\})]$$

  – If $w = 1_A$, $f(w) = F(A) \Rightarrow$ extension from $\{0,1\}^p$ to $\mathbb{R}^p$
  – f is piecewise affine and positively homogeneous

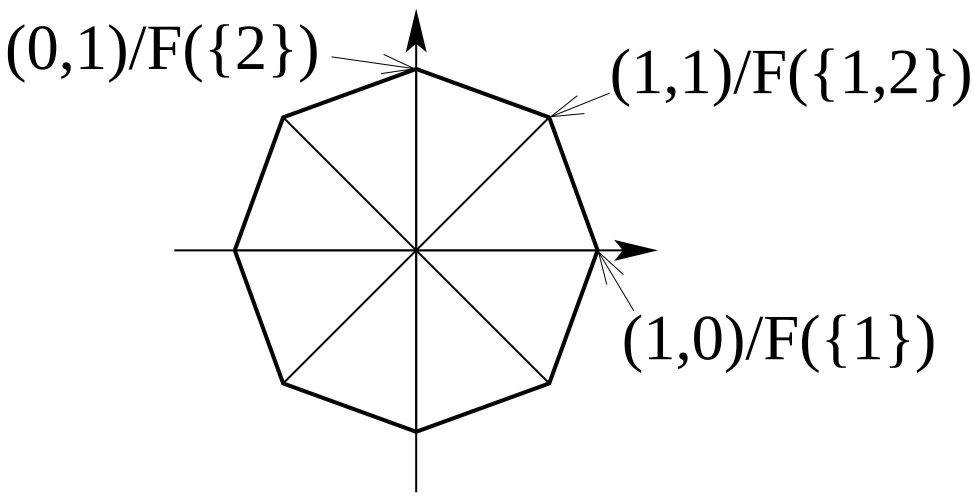
- F is submodular if and only if f is convex (Lovász, 1982)
  – Minimizing $f(w)$ on $w \in [0,1]^p$ equivalent to minimizing F on 2^V
  – Minimizing submodular functions in polynomial time

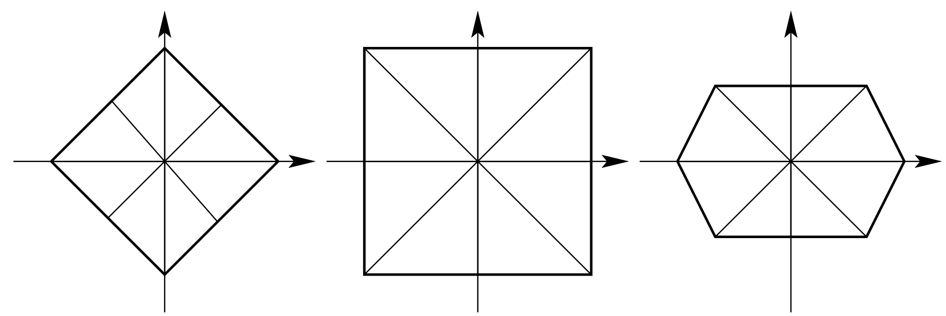
Submodular functions and structured sparsity

- Let $F : 2^V \to \mathbb{R}$ be a **non-decreasing submodular set-function**
- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\mathrm{Supp}(w))$ on the ℓ_∞-ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F

Submodular functions and structured sparsity

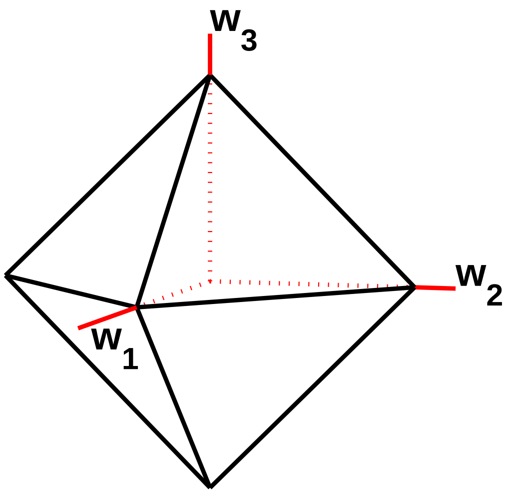
- Let $F : 2^V \to \mathbb{R}$ be a **non-decreasing submodular set-function**
- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\text{Supp}(w))$ on the ℓ_∞-ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F
- **Sparsity-inducing properties**: Ω is a polyhedral norm



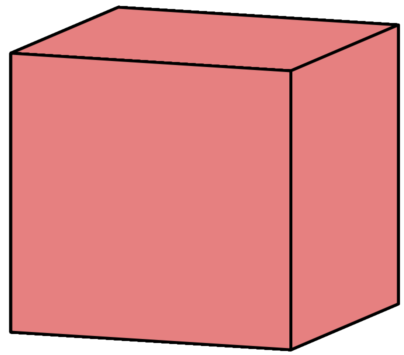


  - A if stable if for all $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$
  - With probability one, stable sets are the only allowed active sets

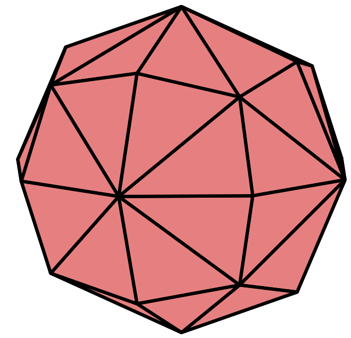
Polyhedral unit balls



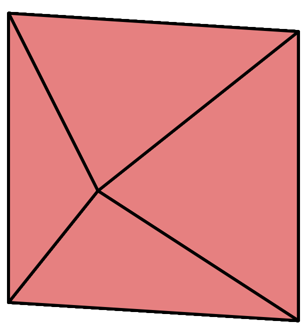
$F(A) = |A|$
$\Omega(w) = \|w\|_1$



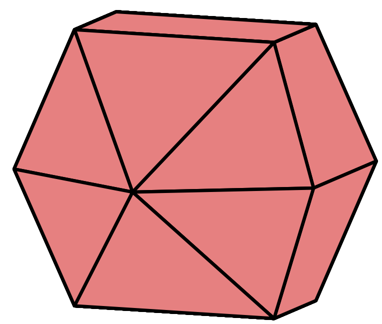
$F(A) = \min\{|A|, 1\}$
$\Omega(w) = \|w\|_\infty$



$F(A) = |A|^{1/2}$
all possible extreme points



$F(A) = 1_{\{A \cap \{1\} \neq \varnothing\}} + 1_{\{A \cap \{2,3\} \neq \varnothing\}}$
$\Omega(w) = |w_1| + \|w_{\{2,3\}}\|_\infty$



$F(A) = 1_{\{A \cap \{1,2,3\} \neq \varnothing\}}$
$+ 1_{\{A \cap \{2,3\} \neq \varnothing\}} + 1_{\{A \cap \{3\} \neq \varnothing\}}$
$\Omega(w) = \|w\|_\infty + \|w_{\{2,3\}}\|_\infty + |w_3|$

Submodular functions and structured sparsity

Examples

- **From $\Omega(w)$ to $F(A)$:** provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_\infty$$

  - ℓ_1-ℓ_∞ norm $\Rightarrow$ sparsity at the group level
  - Some w_G's are set to zero for some groups G

$$\big(\mathrm{Supp}(w)\big)^c = \bigcup_{G \in \mathbf{H}'} G \quad \text{for some } \mathbf{H}' \subseteq \mathbf{H}$$

Submodular functions and structured sparsity
Examples

- **From $\Omega(w)$ to $F(A)$:** provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_\infty \;\Rightarrow\; F(A) = \mathrm{Card}\big(\{G \in \mathbf{H},\ G \cap A \neq \varnothing\}\big)$$

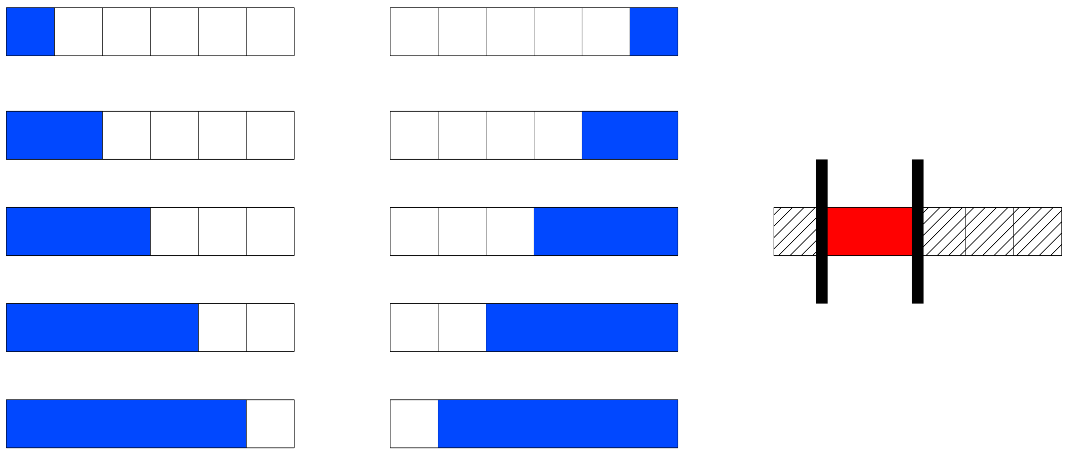
  - ℓ_1-ℓ_∞ norm $\Rightarrow$ sparsity at the group level
  - Some w_G's are set to zero for some groups G

$$\big(\mathrm{Supp}(w)\big)^c = \bigcup_{G \in \mathbf{H}'} G \quad \text{for some } \mathbf{H}' \subseteq \mathbf{H}$$

  - Justification not only limited to allowed sparsity patterns

Selection of contiguous patterns in a sequence

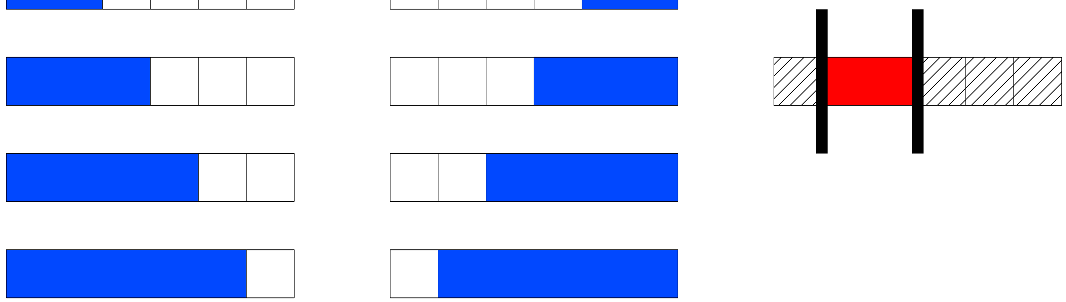
- Selection of contiguous patterns in a sequence



- $\mathbf{H}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a **contiguous pattern**

Selection of contiguous patterns in a sequence

- Selection of contiguous patterns in a sequence

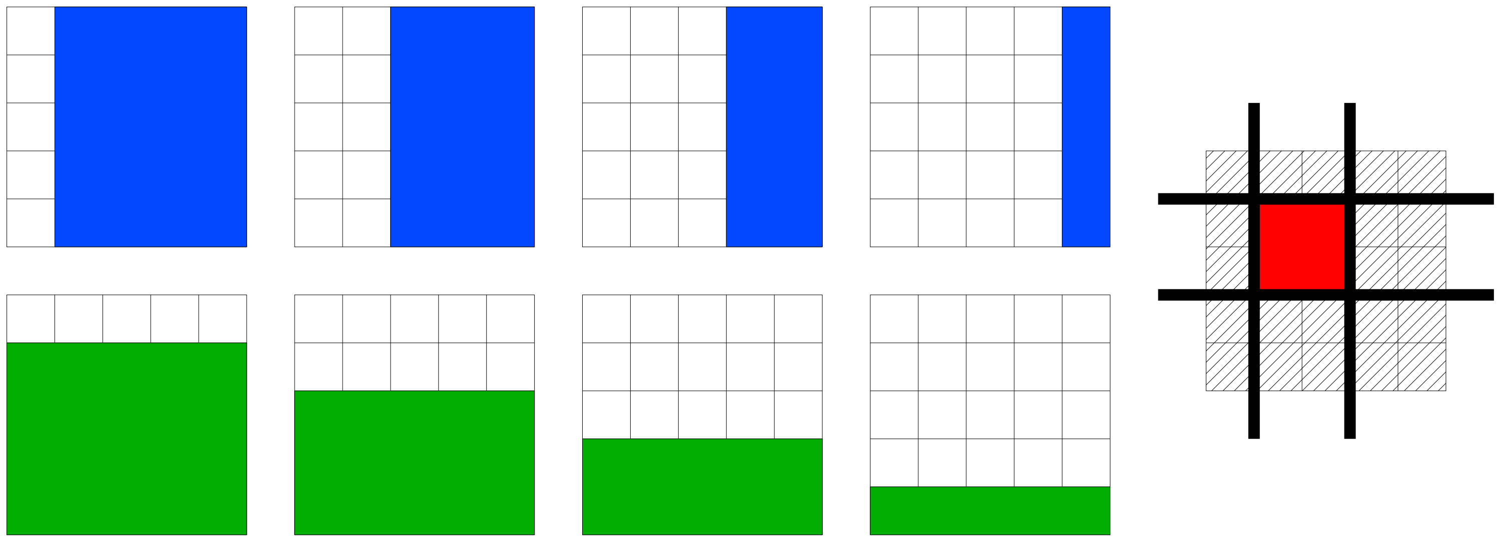


- $\mathbf{H}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a **contiguous pattern**

- $\sum_{G\in\mathbf{H}} \|w_G\|_\infty \Rightarrow F(A) = p - 2 + \text{Range}(A)$ if $A \neq \varnothing$

Other examples of set of groups $\mathbb{H}$

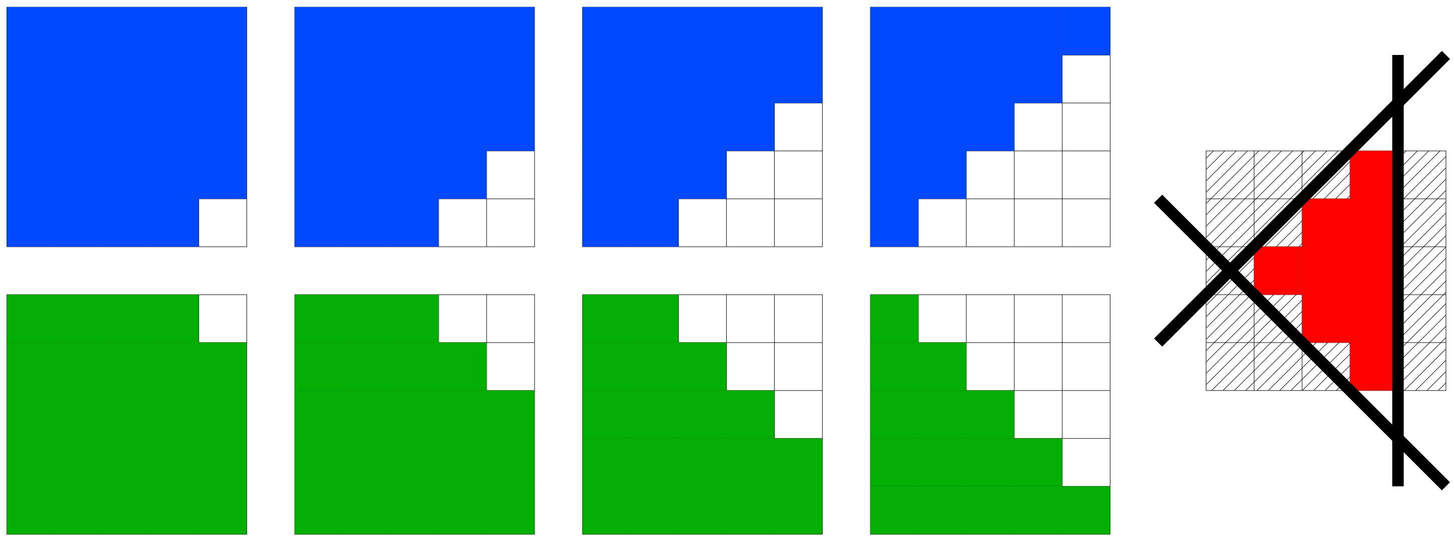
- Selection of rectangles on a 2-D grids, $p = 25$



  - $\mathbf{H}$ is the set of blue/green groups (with their not displayed complements)
  - Any union of blue/green groups set to zero leads to the selection of a rectangle

Other examples of set of groups $\mathbb{H}$

- Selection of diamond-shaped patterns on a 2-D grids, $p = 25$.



  – It is possible to extend such settings to 3-D space, or more complex topologies

Sparse Structured PCA
(Jenatton, Obozinski, and Bach, 2009b)

- Learning **sparse and structured dictionary elements**:

$$\min_{W \in \mathbb{R}^{k \times n}, X \in \mathbb{R}^{p \times k}} \frac{1}{n} \sum_{i=1}^{n} \|y^i - Xw^i\|_2^2 + \lambda \sum_{j=1}^{p} \Omega(x^j) \text{ s.t. } \forall i,\ \|w^i\|_2 \leq 1$$

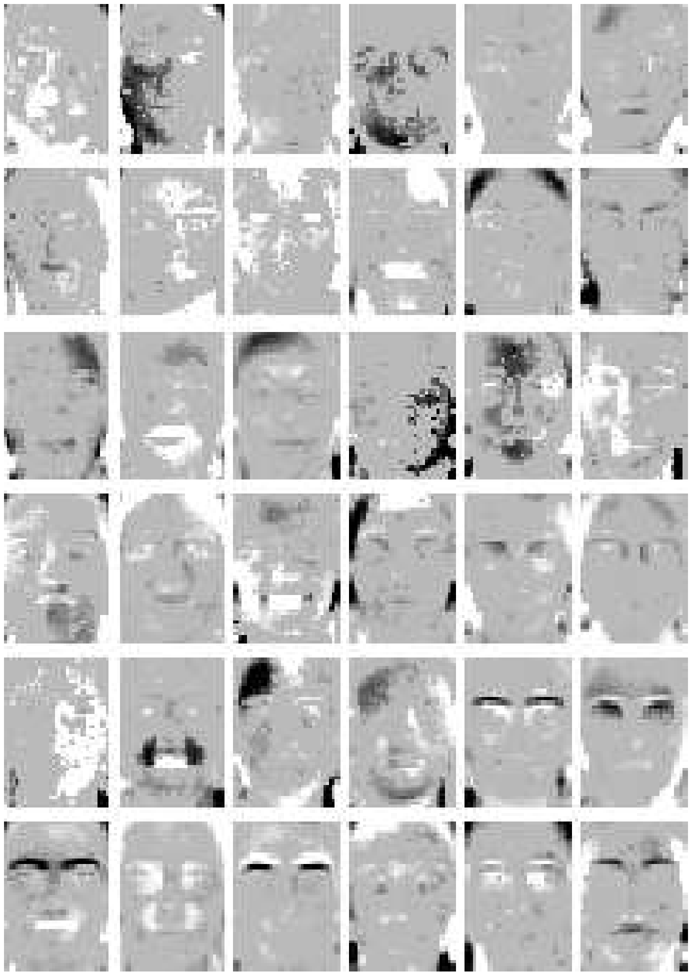
Application to face databases (1/3)

raw data

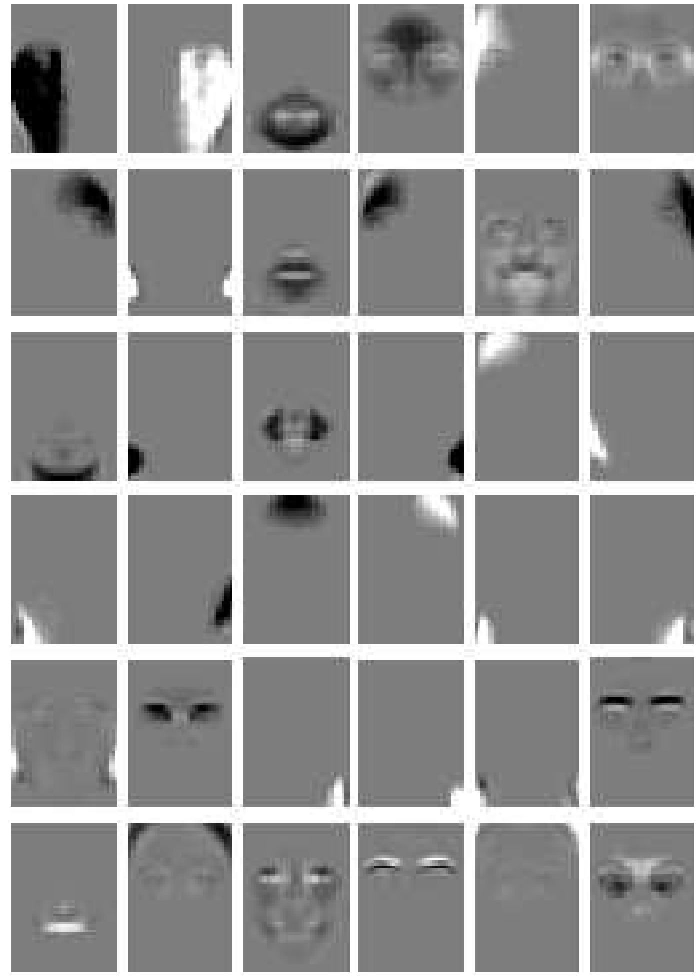
(unstructured) NMF

- NMF obtains partially local features

Application to face databases (2/3)



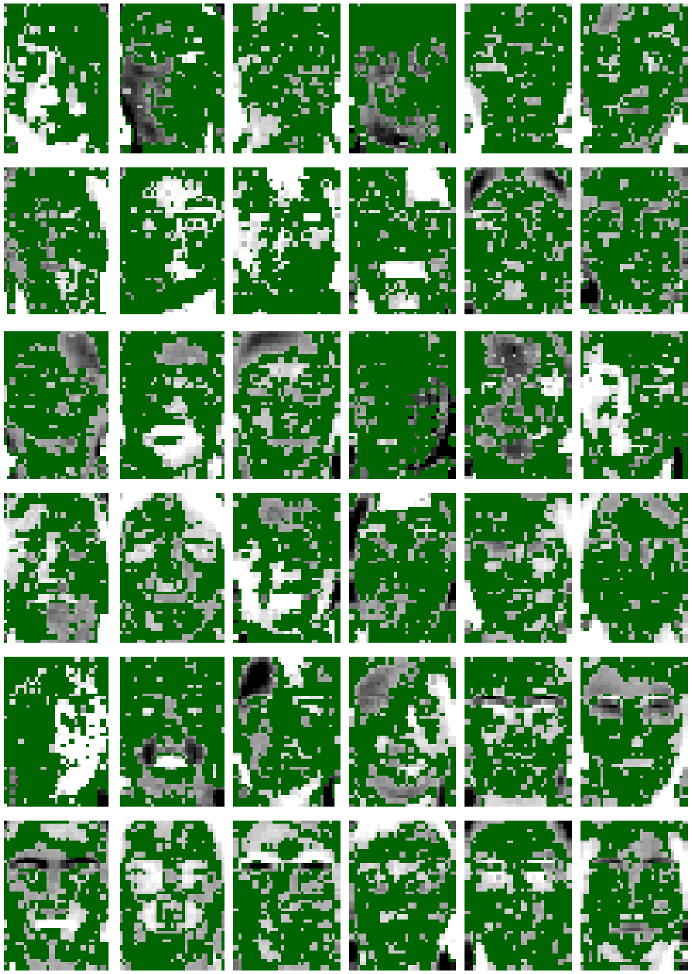
(unstructured) sparse PCA



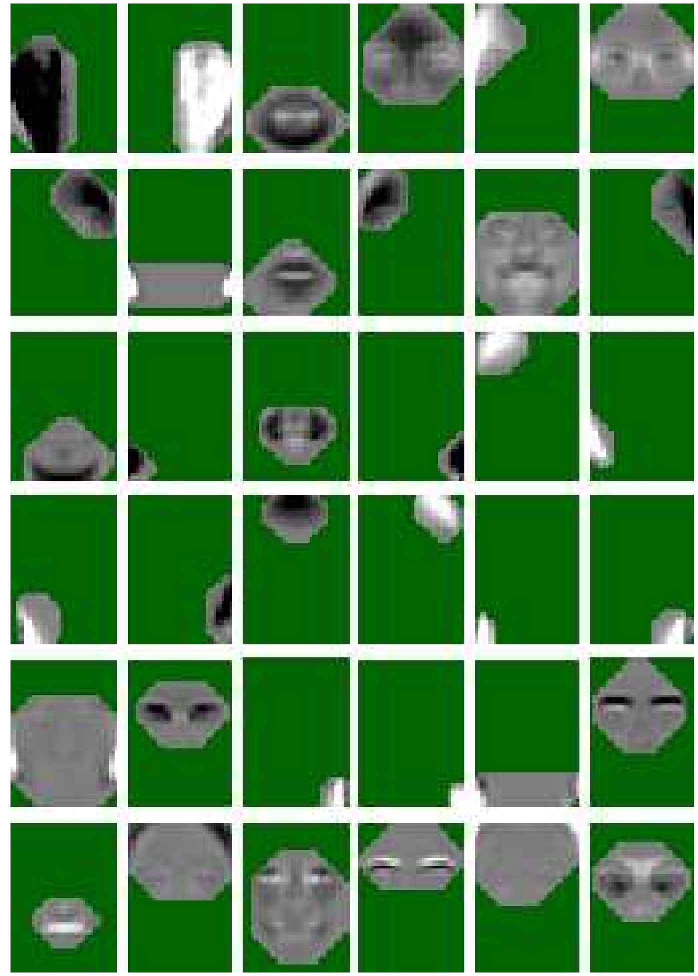
Structured sparse PCA

- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion

Application to face databases (2/3)



(unstructured) sparse PCA

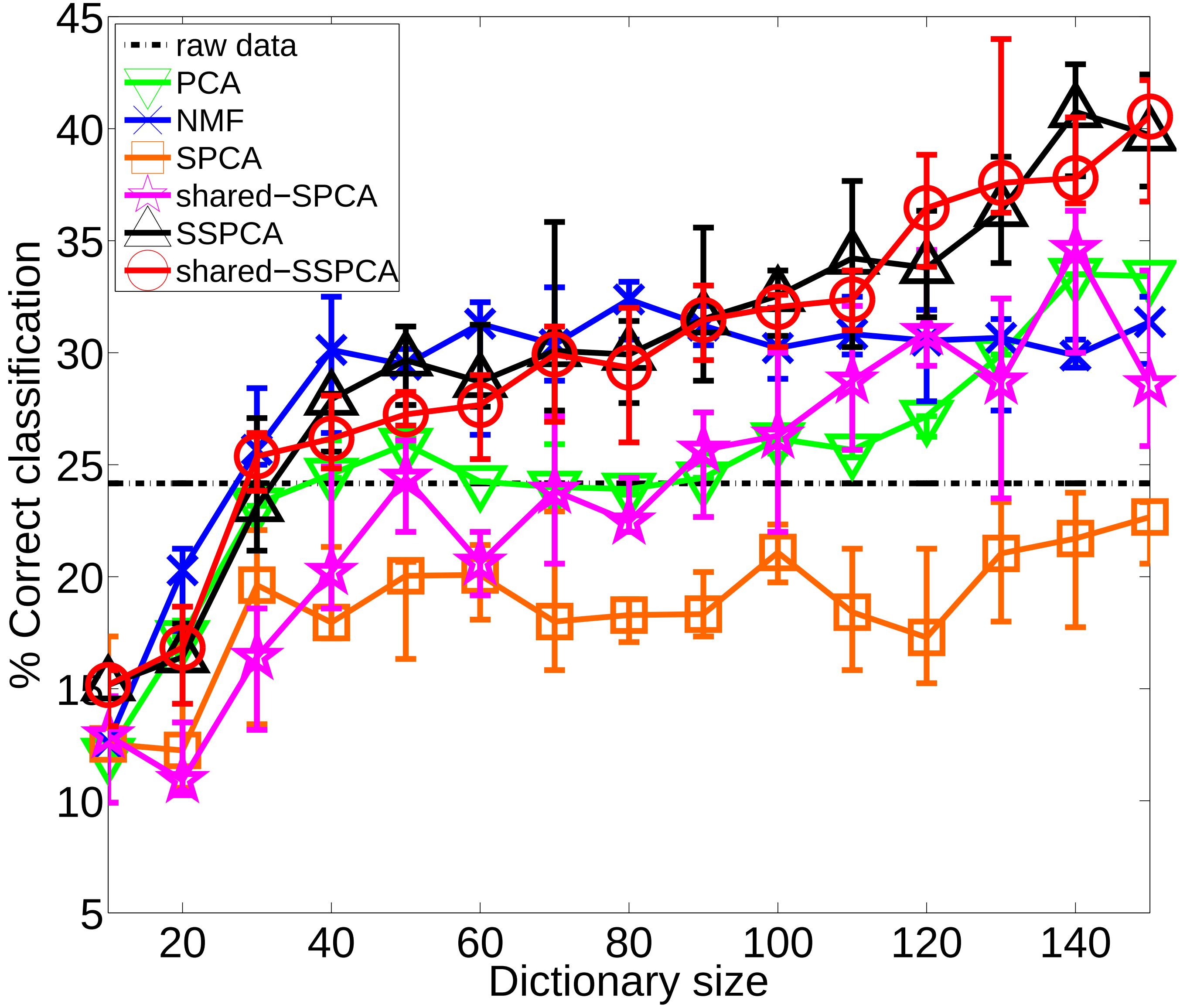


Structured sparse PCA

- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion

Application to face databases (3/3)

- Quantitative performance evaluation on classification task



Dictionary learning vs. sparse structured PCA
Exchange roles of X and w

- Sparse structured PCA (**structured dictionary elements**):

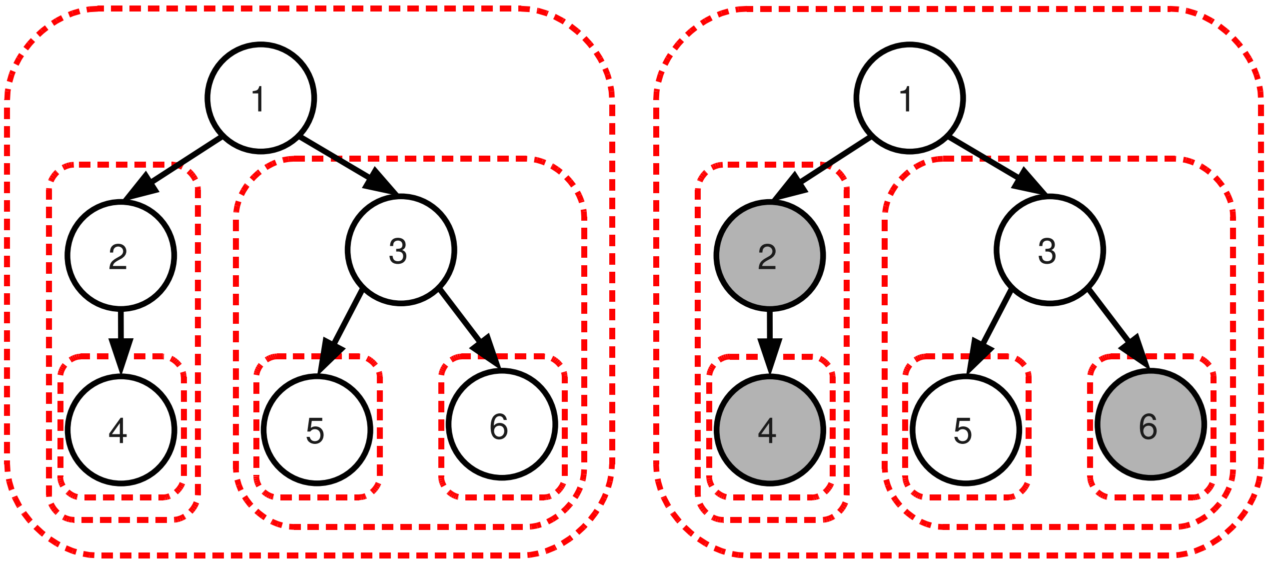
$$\min_{W\in\mathbb{R}^{k\times n}, X\in\mathbb{R}^{p\times k}} \frac{1}{n}\sum_{i=1}^{n}\|y^i - Xw^i\|_2^2 + \lambda\sum_{j=1}^{k}\Omega(x^j) \text{ s.t. } \forall i,\ \|w^i\|_2 \leq 1.$$

- Dictionary learning with **structured sparsity for codes** w:

$$\min_{W\in\mathbb{R}^{k\times n}, X\in\mathbb{R}^{p\times k}} \frac{1}{n}\sum_{i=1}^{n}\|y^i - Xw^i\|_2^2 + \lambda\Omega(w^i) \text{ s.t. } \forall j,\ \|x^j\|_2 \leq 1.$$

Hierarchical dictionary learning (Jenatton, Mairal, Obozinski, and Bach, 2010)

- Structure on codes w (not on dictionary X)
- Hierarchical penalization: $\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_2$ where groups G in $\mathbf{H}$ are equal to set of descendants of some nodes in a tree

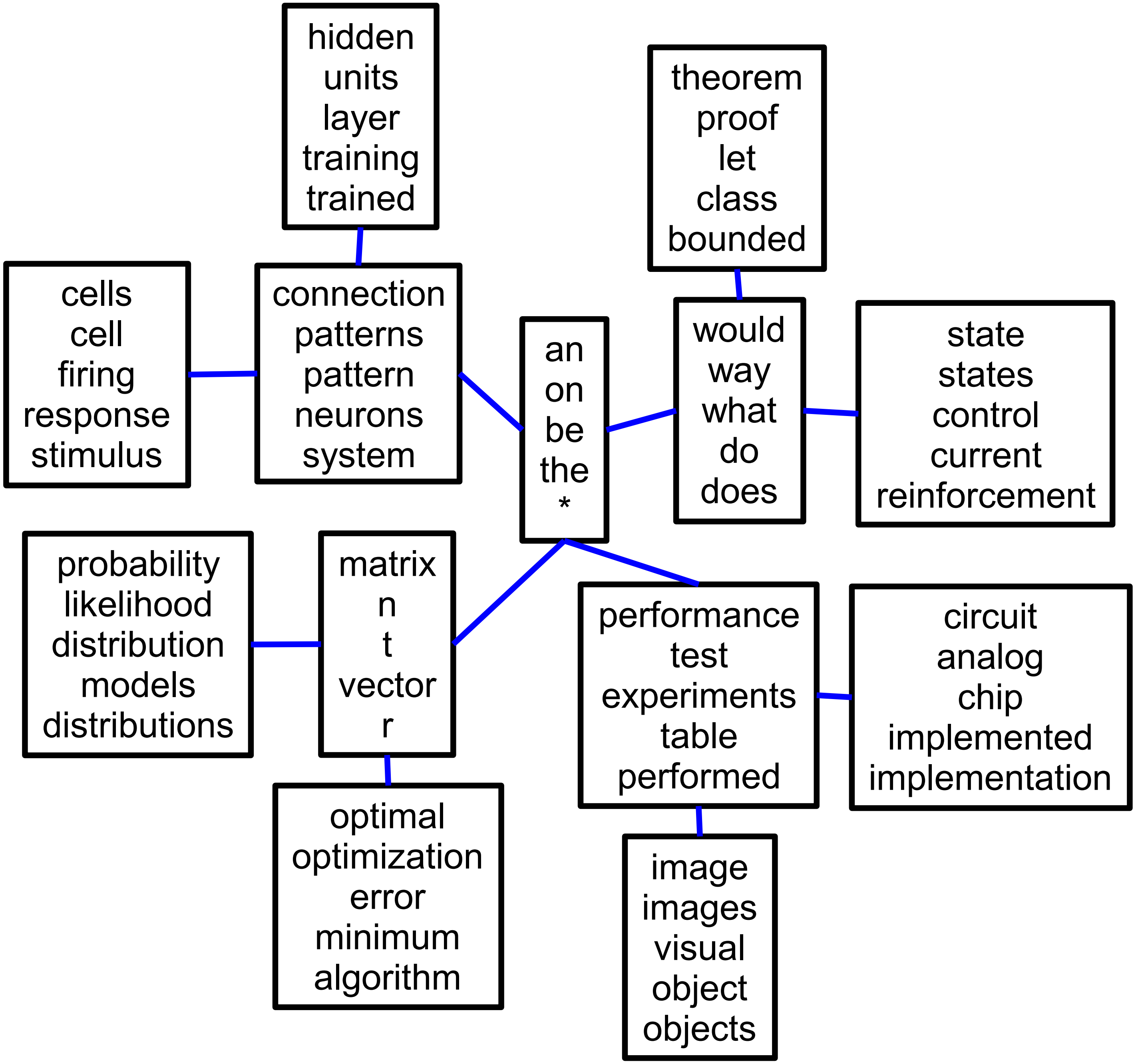


- Variable selected after its ancestors (Zhao et al., 2009; Bach, 2008)
  - Corresponds to $F(A) =$ cardinality of set of ancestors of A

Hierarchical dictionary learning
Modelling of text corpora

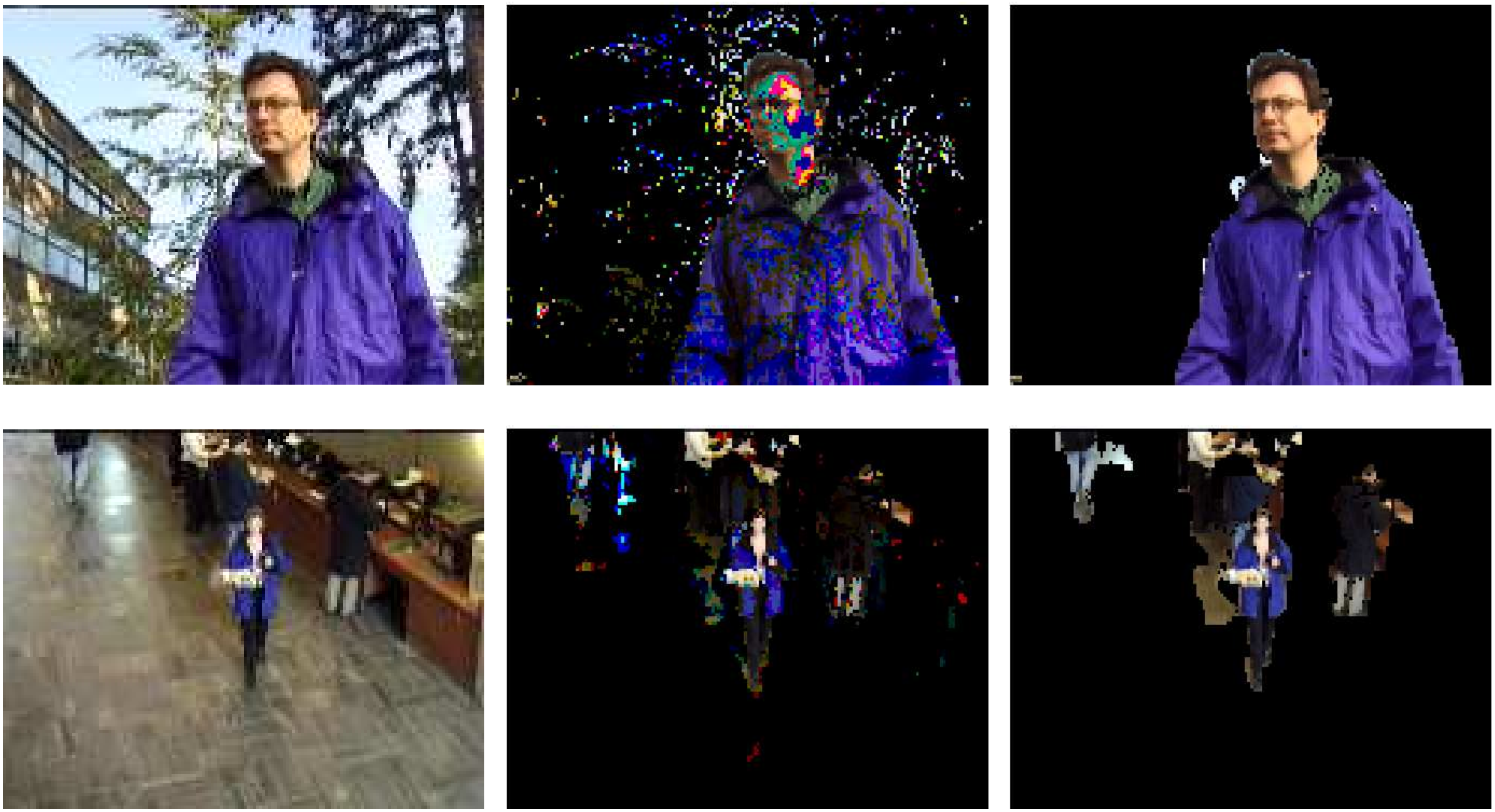
- Each document is modelled through word counts
- Low-rank matrix factorization of word-document matrix
- Probabilistic topic models (Blei et al., 2003)
  - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
  - **Can we achieve similar performance with simple matrix factorization formulation?**

Modelling of text corpora - Dictionary tree



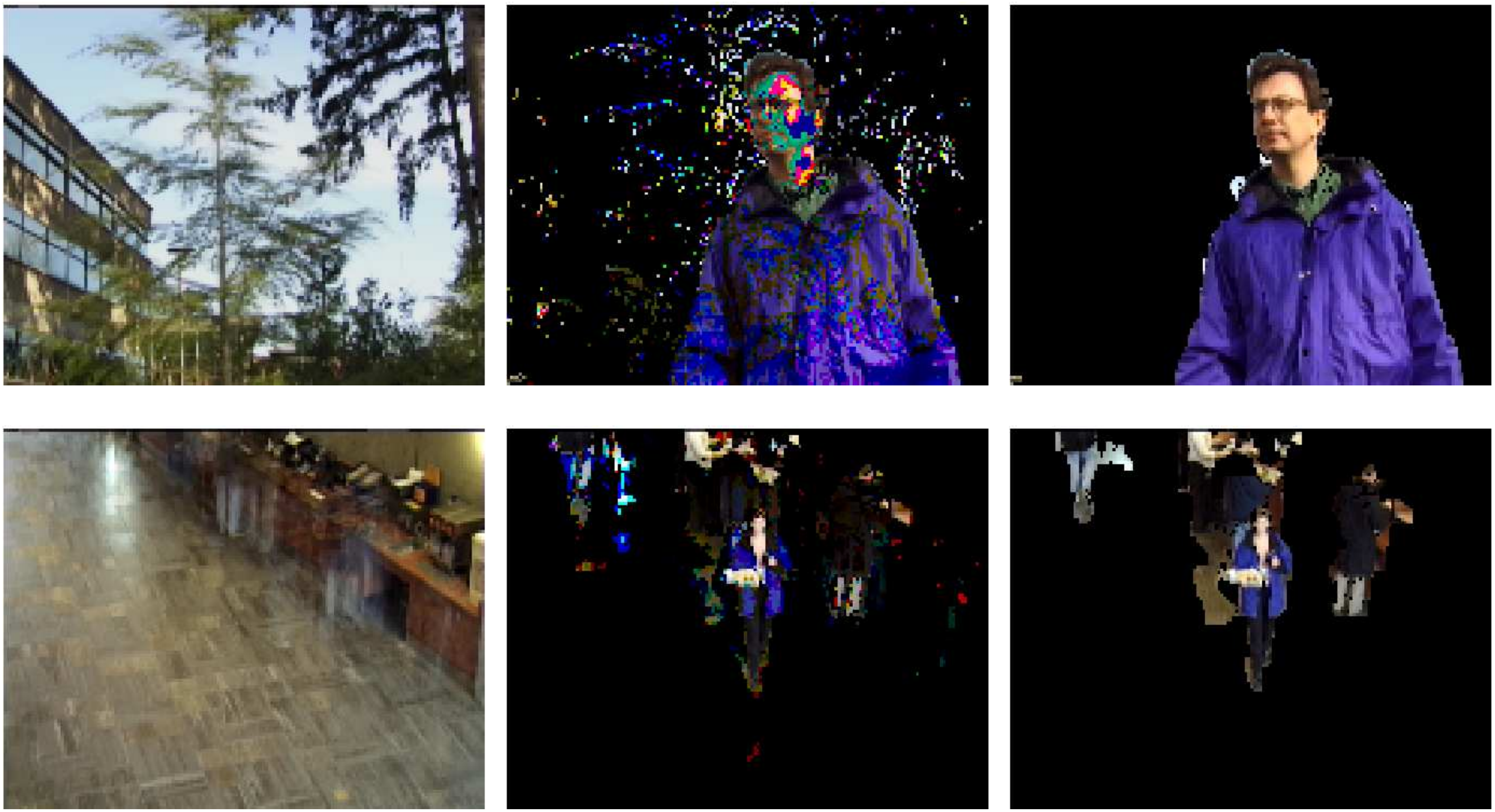
Application to background subtraction (Mairal, Jenatton, Obozinski, and Bach, 2010)

Input | ℓ_1-norm | Structured norm



Application to background subtraction (Mairal, Jenatton, Obozinski, and Bach, 2010)

Background ℓ_1-norm Structured norm



Submodular functions and structured sparsity
Examples

- **From** $\Omega(w)$ **to** $F(A)$: provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_\infty \quad \Rightarrow \quad F(A) = \mathrm{Card}\big(\{G \in \mathbf{H},\ G \cap A \neq \varnothing\}\big)$$

  - Justification not only limited to allowed sparsity patterns

Submodular functions and structured sparsity
Examples

- **From** $\Omega(w)$ **to** $F(A)$: provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)

$$\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_\infty \quad \Rightarrow \quad F(A) = \mathrm{Card}\big(\{G \in \mathbf{H},\ G \cap A \neq \varnothing\}\big)$$

  - Justification not only limited to allowed sparsity patterns

- **From** $F(A)$ **to** $\Omega(w)$: provides new sparsity-inducing norms
  - $F(A) = g(\mathrm{Card}(A)) \ \Rightarrow \ \Omega$ is a combination of **order statistics**
  - **Non-factorial priors** for supervised learning: Ω depends on the eigenvalues of $X_A^\top X_A$ and not simply on the cardinality of A

Non-factorial priors for supervised learning

- **Joint variable selection and regularization**. Given support $A \subset V$,

$$\min_{w_A \in \mathbb{R}^A} \frac{1}{2n}\|y - X_A w_A\|_2^2 + \frac{\lambda}{2}\|w_A\|_2^2$$

  – Minimizing with respect to A will always lead to $A = V$

- **Information/model selection criterion** $F(A)$

$$\min_{A \subset V} \min_{w_A \in \mathbb{R}^A} \frac{1}{2n}\|y - X_A w_A\|_2^2 + \frac{\lambda}{2}\|w_A\|_2^2 + F(A)$$

$$\Leftrightarrow \quad \min_{w \in \mathbb{R}^p} \frac{1}{2n}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 + F(\mathrm{Supp}(w))$$

Non-factorial priors for supervised learning

- Selection of subset A from design $X \in \mathbb{R}^{n \times p}$ with ℓ_2-penalization
- **Frequentist analysis** (Mallow's C_L): $\operatorname{tr} X_A^\top X_A (X_A^\top X_A + \lambda I)^{-1}$
  - Not submodular
- **Bayesian analysis** (marginal likelihood): $\log\det(X_A^\top X_A + \lambda I)$
  - Submodular (also true for $\operatorname{tr}(X_A^\top X_A)^{1/2}$)

p	n	k	submod.	ℓ_2 vs. submod.	ℓ_1 vs. submod.	greedy vs. submod.
120	120	80	40.8 ± 0.8	-2.6 ± 0.5	**0.6 ± 0.0**	**21.8 ± 0.9**
120	120	40	35.9 ± 0.8	**2.4 ± 0.4**	**0.3 ± 0.0**	**15.8 ± 1.0**
120	120	20	29.0 ± 1.0	**9.4 ± 0.5**	-0.1 ± 0.0	**6.7 ± 0.9**
120	120	10	20.4 ± 1.0	**17.5 ± 0.5**	-0.2 ± 0.0	-2.8 ± 0.8
120	20	20	49.4 ± 2.0	0.4 ± 0.5	**2.2 ± 0.8**	**23.5 ± 2.1**
120	20	10	49.2 ± 2.0	0.0 ± 0.6	1.0 ± 0.8	**20.3 ± 2.6**
120	20	6	43.5 ± 2.0	**3.5 ± 0.8**	**0.9 ± 0.6**	**24.4 ± 3.0**
120	20	4	41.0 ± 2.1	**4.8 ± 0.7**	-1.3 ± 0.5	**25.1 ± 3.5**

Unified optimization algorithms

- **Polyhedral norm** with up to $O(2^p p!)$ faces and $O(3^p)$ extreme points
  - Not suitable to linear programming toolboxes
- **Subgradient** ($w \mapsto \Omega(w)$ non-differentiable)
  - subgradient may be obtained in polynomial time $\Rightarrow$ too slow

Unified optimization algorithms

- **Polyhedral norm** with up to $O(2^p p!)$ faces and $O(3^p)$ extreme points
  - Not suitable to linear programming toolboxes
- **Subgradient** ($w \mapsto \Omega(w)$ non-differentiable)
  - subgradient may be obtained in polynomial time $\Rightarrow$ too slow
- **Proximal methods**
  - $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda\Omega(w)$: differentiable + non-differentiable
  - Efficient when proximal operator is easy to compute

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

  - See, e.g., Beck and Teboulle (2009); Combettes and Pesquet (2010); Bach et al. (2011) and references therein

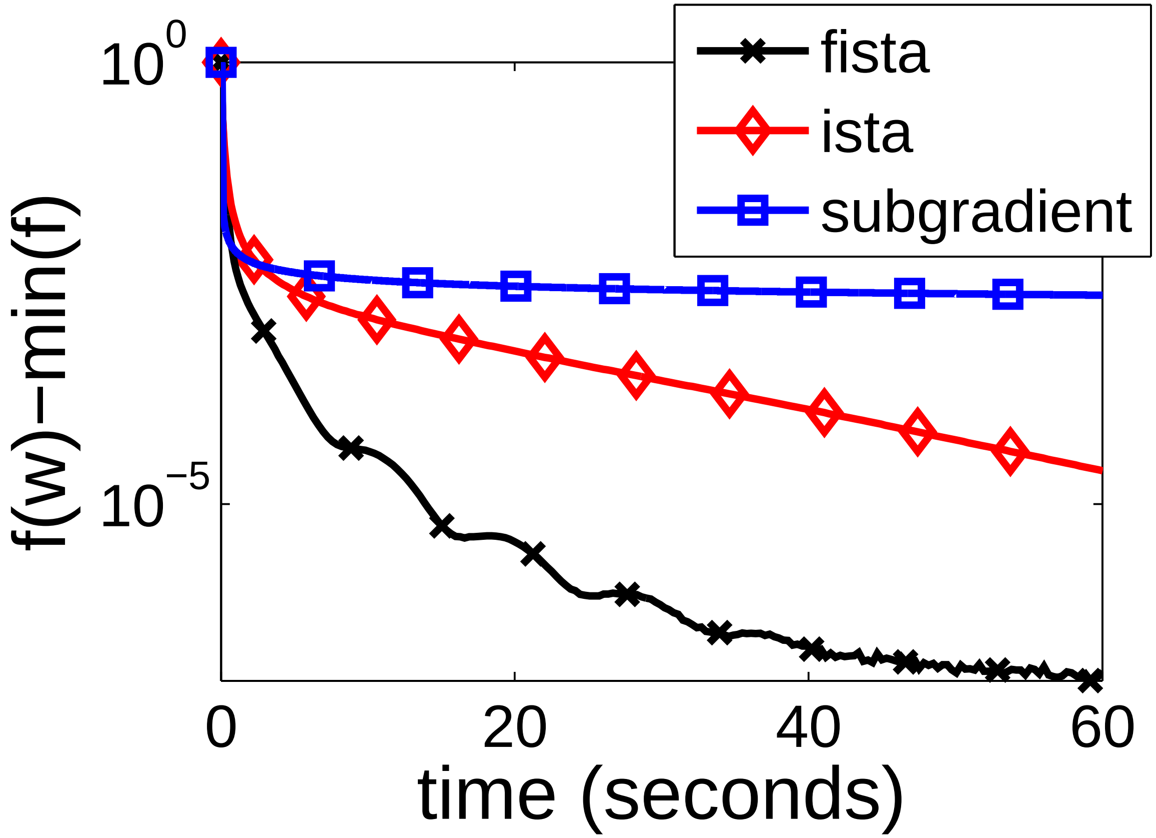
Proximal methods for Lovász extensions

- **Proposition** (Chambolle and Darbon, 2009): let w^* be the solution of $\min_{w\in\mathbb{R}^p} \frac{1}{2}\|w-z\|_2^2 + \lambda f(w)$. Then the minimal and maximal solutions of
$$\min_{A\subset V} \lambda F(A) + \sum_{j\in A}(\alpha - z_j)$$
are $\{w^* > \alpha\}$ and $\{w^* \geqslant \alpha\}$.
  – May be extended to penalization by $f(|w|)$ (Bach, 2011)

- **Parametric submodular function optimization**
  – General divide-and-conquer strategy (Groenevelt, 1991)
  – Efficient only when submodular minimization is efficient (see, e.g., Mairal et al., 2010)
  – Otherwise, minimum-norm-point algorithm (a.k.a. Frank Wolfe)

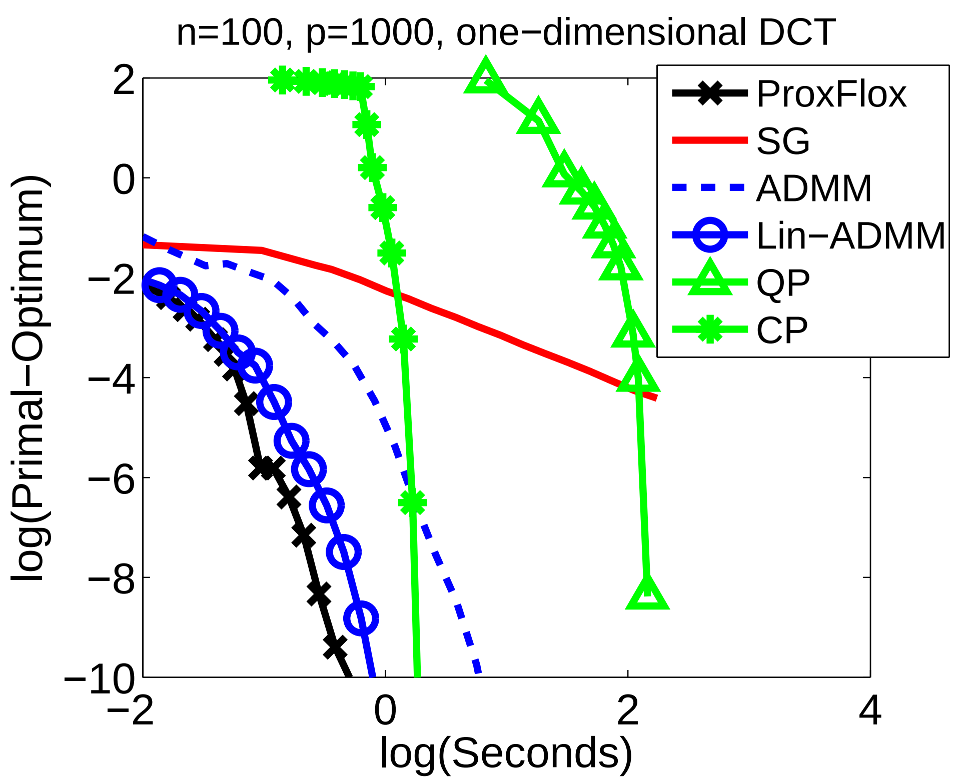
Comparison of optimization algorithms

- Synthetic example with $p = 1000$ and $F(A) = |A|^{1/2}$
- ISTA: proximal method
- FISTA: accelerated variant (Beck and Teboulle, 2009)



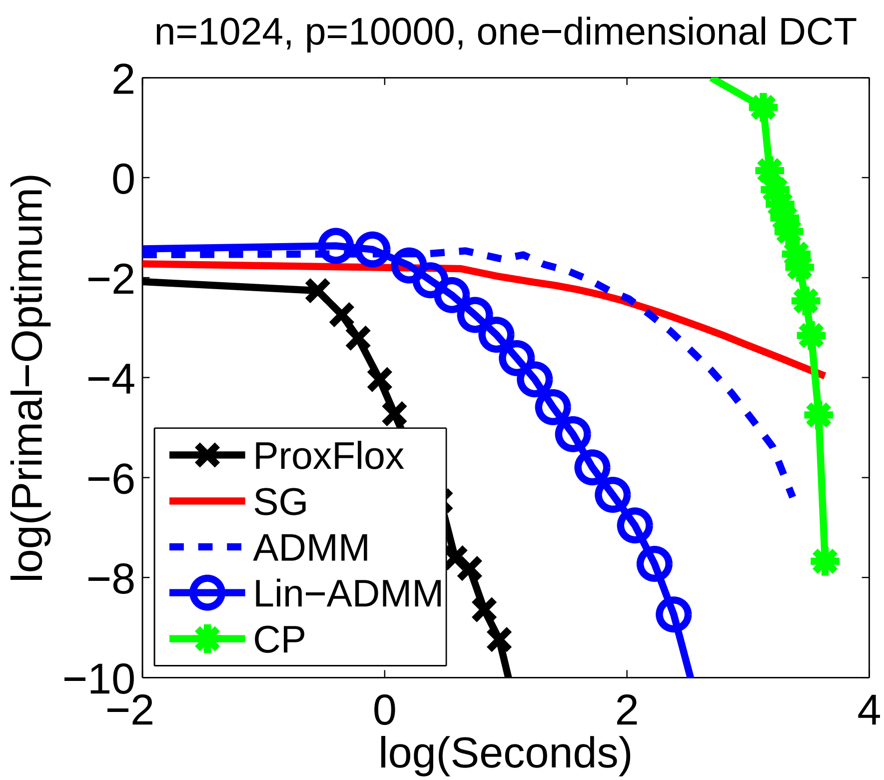
Comparison of optimization algorithms
(Mairal, Jenatton, Obozinski, and Bach, 2010)
Small scale

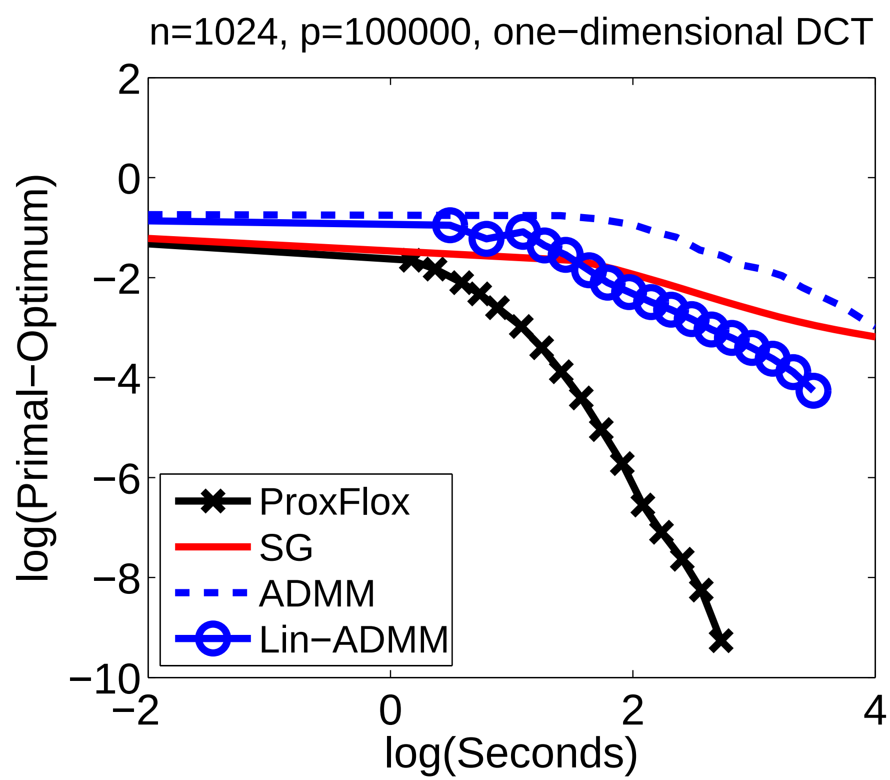
- Specific norms which can be implemented through network flows



Comparison of optimization algorithms (Mairal, Jenatton, Obozinski, and Bach, 2010) Large scale

- Specific norms which can be implemented through network flows





Unified theoretical analysis

- **Decomposability**
  - Key to theoretical analysis (Negahban et al., 2009)
  - **Property**: $\forall w \in \mathbb{R}^p$, and $\forall J \subset V$, if $\min_{j \in J} |w_j| \geqslant \max_{j \in J^c} |w_j|$, then $\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c})$
- **Support recovery**
  - Extension of known sufficient condition (Zhao and Yu, 2006; Negahban and Wainwright, 2008)
- **High-dimensional inference**
  - Extension of known sufficient condition (Bickel et al., 2009)
  - Matches with analysis of Negahban et al. (2009) for common cases

Support recovery - $\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|y - Xw\|_2^2 + \lambda \Omega(w)$

- **Notation**
  - $\rho(J) = \min_{B \subset J^c} \frac{F(B \cup J) - F(J)}{F(B)} \in (0, 1]$ (for J stable)
  - $c(J) = \sup_{w \in \mathbb{R}^p} \Omega_J(w_J) / \|w_J\|_2 \leqslant |J|^{1/2} \max_{k \in V} F(\{k\})$

- **Proposition**
  - Assume $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, I)$
  - J = smallest stable set containing the support of w^*
  - Assume $\nu = \min_{j, w_j^* \neq 0} |w_j^*| > 0$
  - Let $Q = \frac{1}{n} X^\top X \in \mathbb{R}^{p \times p}$. Assume $\kappa = \lambda_{\min}(Q_{JJ}) > 0$
  - Assume that for $\eta > 0$, $\boxed{(\Omega^J)^*[(\Omega_J(Q_{JJ}^{-1} Q_{Jj}))_{j \in J^c}] \leqslant 1 - \eta}$
  - If $\lambda \leqslant \frac{\kappa\nu}{2c(J)}$, $\hat{w}$ has support equal to J, with probability larger than
    $$1 - 3P\Big(\Omega^*(z) > \frac{\lambda\eta\rho(J)\sqrt{n}}{2\sigma}\Big)$$
  - z is a multivariate normal with covariance matrix Q

Consistency - $\min_{w\in\mathbb{R}^p} \frac{1}{2n}\|y - Xw\|_2^2 + \lambda\Omega(w)$

- **Proposition**
  - Assume $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, I)$
  - J = smallest stable set containing the support of w^*
  - Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{p\times p}$.
  - Assume that $\forall\Delta$ s.t. $\Omega^J(\Delta_{J^c}) \leqslant 3\Omega_J(\Delta_J)$, $\Delta^\top Q\Delta \geqslant \kappa\|\Delta_J\|_2^2$
  - Then $\boxed{\Omega(\hat{w} - w^*) \leqslant \frac{24c(J)^2\lambda}{\kappa\rho(J)^2}}$ and $\boxed{\frac{1}{n}\|X\hat{w} - Xw^*\|_2^2 \leqslant \frac{36c(J)^2\lambda^2}{\kappa\rho(J)^2}}$

    with probability larger than $1 - P\big(\Omega^*(z) > \frac{\lambda\rho(J)\sqrt{n}}{2\sigma}\big)$
  - z is a multivariate normal with covariance matrix Q

- **Concentration inequality** (z normal with covariance matrix Q):
  - $\mathcal{T}$ set of stable inseparable sets
  - Then $P(\Omega^*(z) > t) \leqslant \sum_{A\in\mathcal{T}} 2^{|A|}\exp\big(-\frac{t^2F(A)^2/2}{1^\top Q_{AA}1}\big)$

Outline

- **Introduction: Sparse methods for machine learning**
  - Need for structured sparsity: Going beyond the ℓ_1-norm
- **Structured sparsity through submodular functions**
  - Relaxation of the penalization of supports
  - Unified algorithms and analysis
  - Applications to signal processing and machine learning
- **Extensions**
  - Shaping level sets through symmetric submodular functions
  - ℓ_2-norm relaxation of combinatorial penalties

Symmetric submodular functions (Bach, 2011)

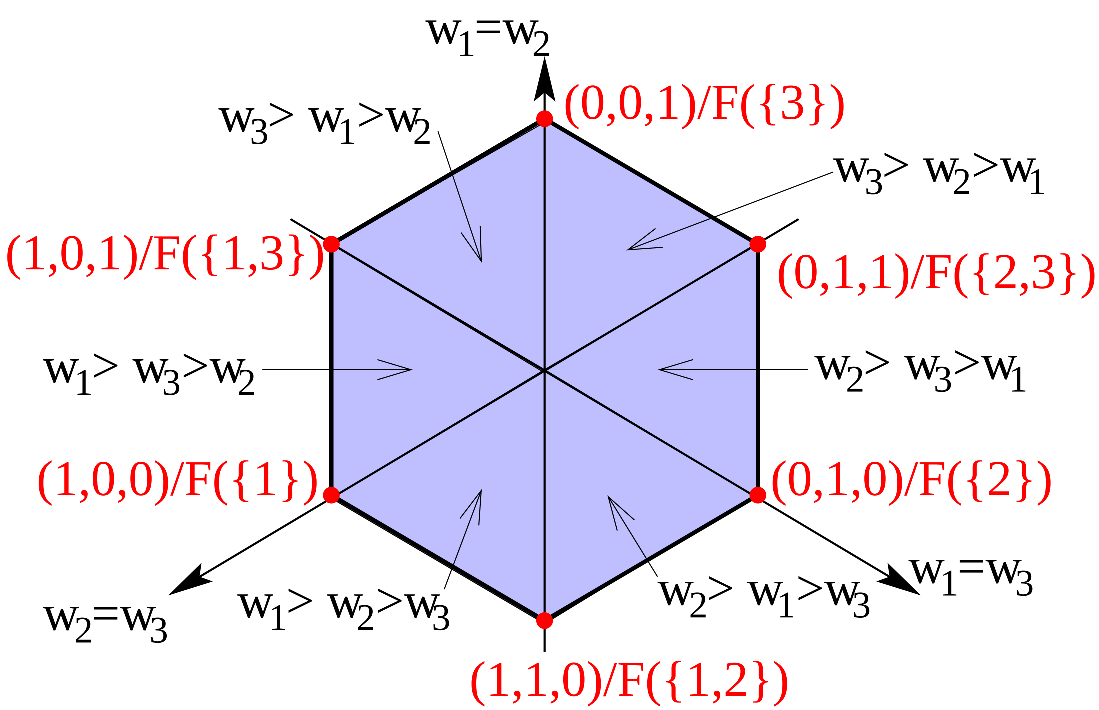
- Let $F : 2^V \to \mathbb{R}$ be a **symmetric submodular set-function**

- **Proposition**: The Lovász extension $f(w)$ is the convex envelope of the function $w \mapsto \max_{\alpha \in \mathbb{R}} F(\{w \geqslant \alpha\})$ on the set $[0,1]^p + \mathbb{R}1_V = \{w \in \mathbb{R}^p, \ \max_{k \in V} w_k - \min_{k \in V} w_k \leqslant 1\}$.

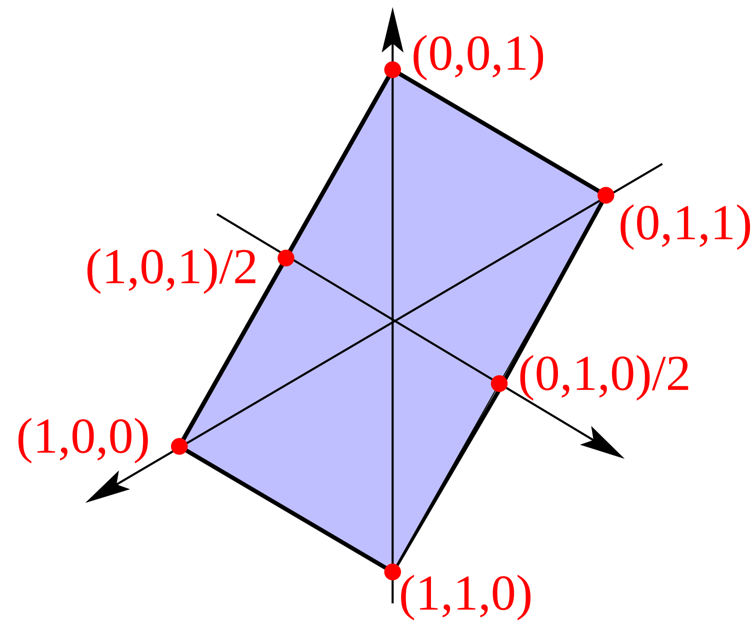
Symmetric submodular functions (Bach, 2011)

- Let $F : 2^V \to \mathbb{R}$ be a **symmetric submodular set-function**

- **Proposition**: The Lovász extension $f(w)$ is the convex envelope of the function $w \mapsto \max_{\alpha \in \mathbb{R}} F(\{w \geqslant \alpha\})$ on the set $[0,1]^p + \mathbb{R}1_V = \{w \in \mathbb{R}^p, \ \max_{k \in V} w_k - \min_{k \in V} w_k \leqslant 1\}$.



$F(A) = 1_{|A| \in \{1,2\}}$

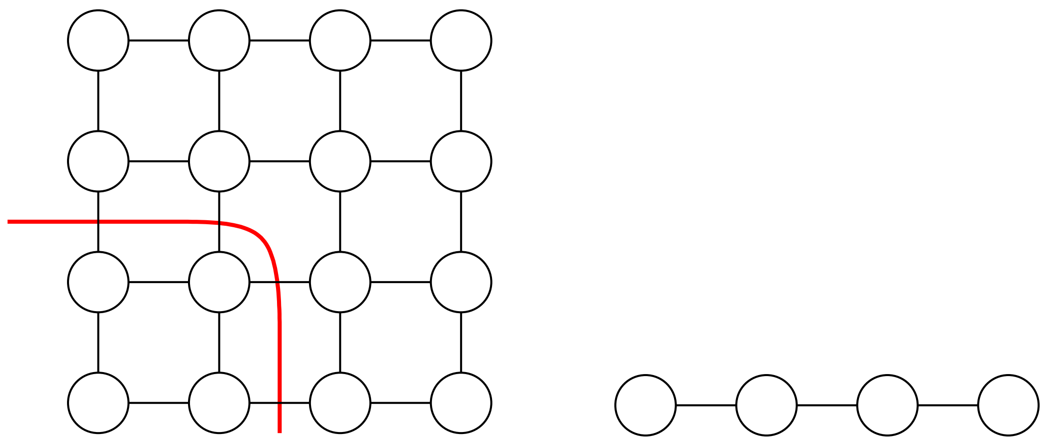


$F(A) = |1_{1 \in A} - 1_{2 \in A}| + |1_{2 \in A} - 1_{3 \in A}|$

Symmetric submodular functions - Examples

- **From $\Omega(w)$ to $F(A)$:** provides new insights into existing norms
  - **Cuts - total variation**

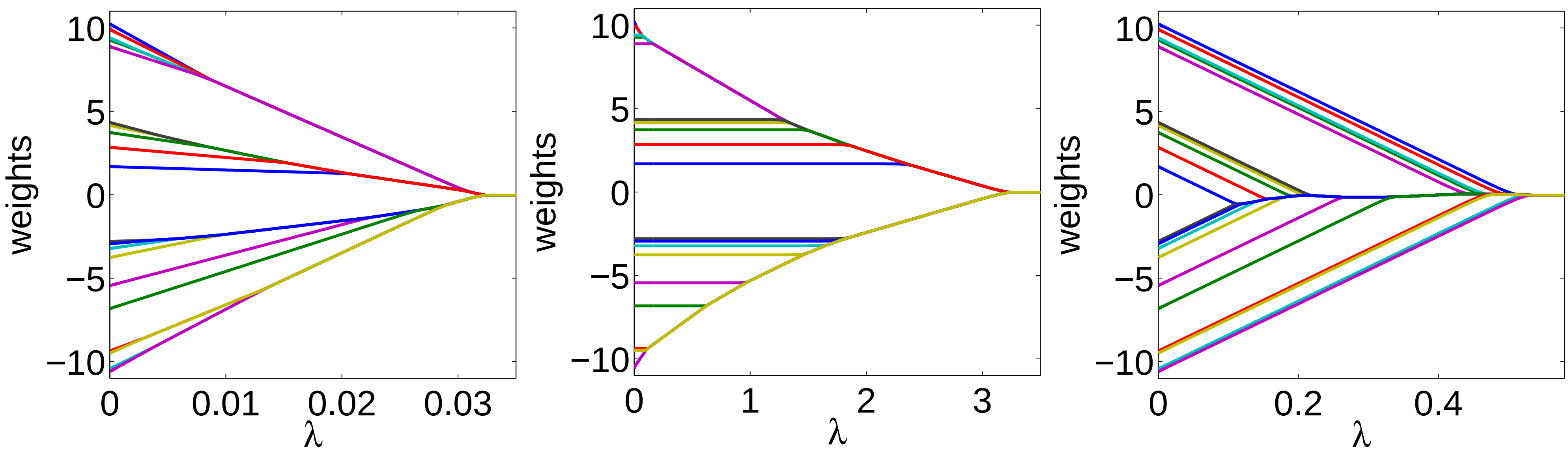
$$F(A) = \sum_{k\in A, j\in V\backslash A} d(k,j) \quad \Rightarrow \quad f(w) = \sum_{k,j\in V} d(k,j)(w_k - w_j)_+$$



  - NB: graph may be directed

Symmetric submodular functions - Examples

- **From $F(A)$ to $\Omega(w)$**: provides new sparsity-inducing norms
  - $F(A) = g(\mathrm{Card}(A)) \Rightarrow$ priors on the size and numbers of clusters



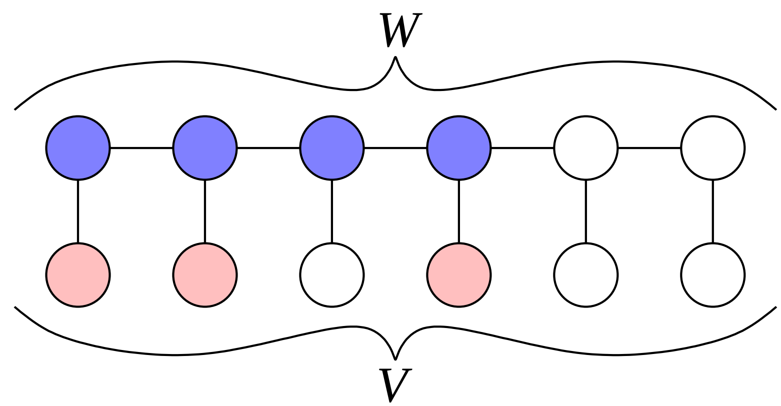
$|A|(p - |A|)$ $\qquad$ $1_{|A| \in (0,p)}$ $\qquad$ $\max\{|A|, p - |A|\}$

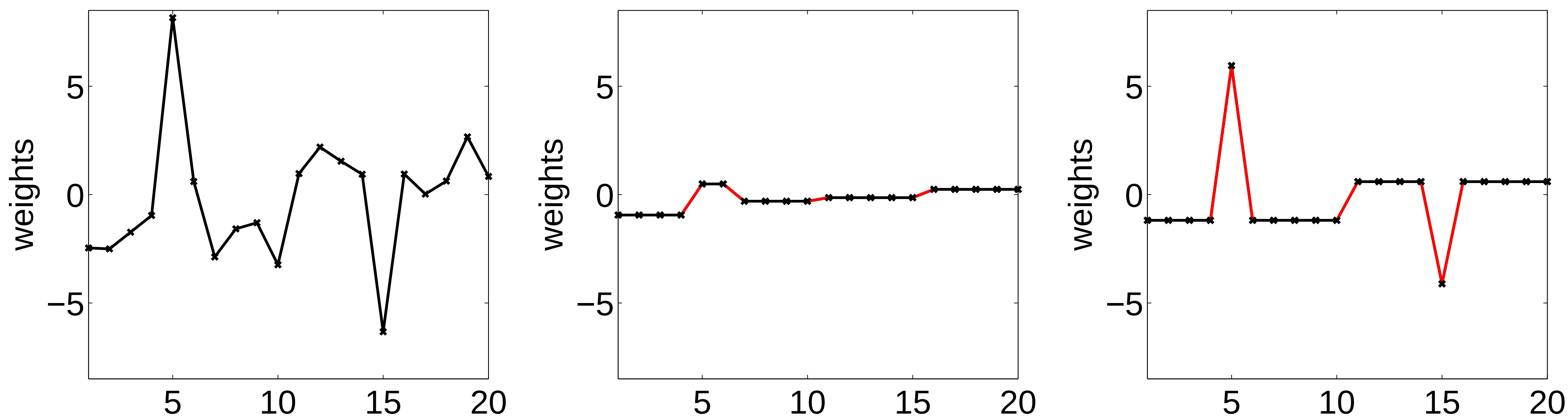
  - Convex formulations for clustering (Hocking, Joulin, Bach, and Vert, 2011)

Symmetric submodular functions - Examples

- **From $F(A)$ to $\Omega(w)$**: provides new sparsity-inducing norms
  - **Regular functions** (Boykov et al., 2001; Chambolle and Darbon, 2009)

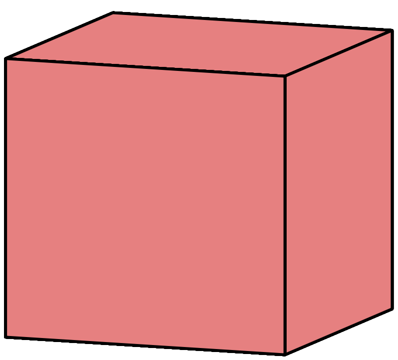
$$F(A) = \min_{B \subset W} \sum_{k \in B,\ j \in W \backslash B} d(k,j) + \lambda |A \Delta B|$$



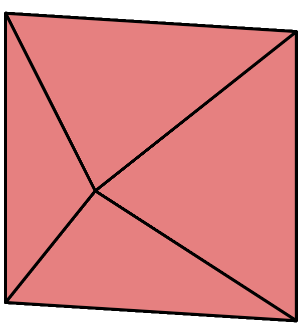


ℓ_2-relaxation of combinatorial penalties (Obozinski and Bach, 2012)

- **Main result** of Bach (2010):
  - $f(|w|)$ is the convex envelope of $F(\mathrm{Supp}(w))$ on $[-1,1]^p$
- **Problems**:
  - Limited to submodular functions
  - Limited to ℓ_∞-relaxation: undesired artefacts



$F(A) = \min\{|A|, 1\}$
$\Omega(w) = \|w\|_\infty$



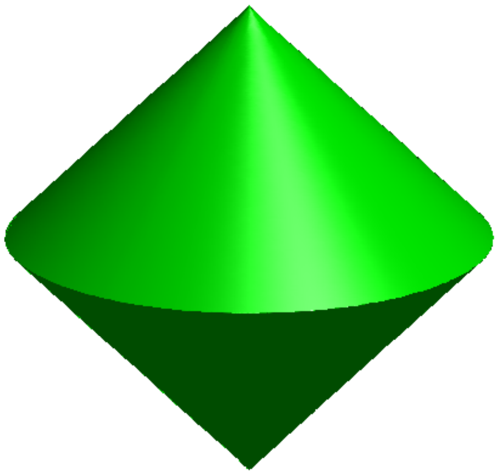
$F(A) = 1_{\{A \cap \{1\} \neq \varnothing\}} + 1_{\{A \cap \{2,3\} \neq \varnothing\}}$
$\Omega(w) = |w_1| + \|w_{\{2,3\}}\|_\infty$

ℓ_2-relaxation of submodular penalties (Obozinski and Bach, 2012)

- F a nondecreasing submodular function with Lovász extension f
- Define $\Omega_2(w) = \min_{\eta \in \mathbb{R}_+^p} \frac{1}{2} \sum_{i \in V} \frac{|w_i|^2}{\eta_i} + \frac{1}{2} f(\eta)$
  - NB: general formulation (Micchelli et al., 2011; Bach et al., 2011)
- **Proposition 1**: Ω_2 is the convex envelope of $w \mapsto F(\text{Supp}(w)) \|w\|_2$
- **Proposition 2**: Ω_2 is the *homogeneous* convex envelope of $w \mapsto \frac{1}{2} F(\text{Supp}(w)) + \frac{1}{2} \|w\|_2^2$
- **Jointly penalizing and regularizing**
  - Extension possible to ℓ_q, $q > 1$

From ℓ_∞ to ℓ_2

Removal of undesired artefacts



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

$$\Omega_2(w) = |w_3| + \|w_{\{1,2\}}\|_2$$

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

Tightness of relaxation

What information of F is kept after the relaxation?

- **Extension to any set-function F**
  - Can always be defined
- Does it always do anything useful?
  - Lower-combinatorial envelope G (Obozinski and Bach, 2012)
- Some functions are not attainable
  - Many set-functions lead to $G(A) = |A|$ and the ℓ_1-norm
  - Example: $F(A) = |A| 1_{A \in \mathcal{A}}$
  - Convexification is not always useful

Conclusion

- **Structured sparsity for machine learning and statistics**
  - Many applications (image, audio, text, etc.)
  - May be achieved through structured sparsity-inducing norms
  - Link with submodular functions: unified analysis and algorithms

    Submodular functions to encode discrete structures

Conclusion

- **Structured sparsity for machine learning and statistics**
  - Many applications (image, audio, text, etc.)
  - May be achieved through structured sparsity-inducing norms
  - Link with submodular functions: unified analysis and algorithms

    Submodular functions to encode discrete structures

- **On-going work on structured sparsity**
  - Norm design beyond submodular functions
  - Instance of general framework of Chandrasekaran et al. (2010)
  - Links with greedy methods (Haupt and Nowak, 2006; Baraniuk et al., 2008; Huang et al., 2009)
  - Achieving $\log p = O(n)$ algorithmically (Bach, 2008)

References

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

F. Bach. Structured sparsity-inducing norms through submodular functions. In *NIPS*, 2010.

F. Bach. Convex analysis and optimization with submodular functions: a tutorial. Technical Report 00527714, HAL, 2010.

F. Bach. Learning with Submodular Functions: A Convex Optimization Perspective. 2011. URL `http://hal.inria.fr/hal-00645271/en`.

F. Bach. Shaping level sets with submodular functions. In *Adv. NIPS*, 2011.

F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. Technical Report 0812.1869, ArXiv, 2008.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2011.

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. Technical report, arXiv:0808.3572, 2008.

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.

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

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

D. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16:106, 2004.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.

E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

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

A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1):89–97, 2004.

A. Chambolle. Total variation minimization and a class of binary MRF models. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 136–152. Springer, 2005.

A. Chambolle and J. Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84(3):288–307, 2009.

V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. *Arxiv preprint arXiv:1012.0621*, 2010.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

P.L. Combettes and J.C. Pesquet. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter Proximal Splitting Methods in Signal Processing. New York: Springer-Verlag, 2010.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

A. Gramfort and M. Kowalski. Improving M/EEG source localization with an inter-condition sparse prior. In *IEEE International Symposium on Biomedical Imaging*, 2009.

H. Groenevelt. Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *European Journal of Operational Research*, 54(2):227–236, 1991.

J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.

T. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath: an algorithm for clustering using convex fusion penalties. In *Proc. ICML*, 2011.

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

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Submitted to ICML*, 2010.

R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion. Multi-scale mining of fmri data with hierarchical structured sparsity. Technical report, Preprint arXiv:1105.0363, 2011. In submission to SIAM Journal on Imaging Sciences.

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

S. Kim and E. P. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proc. UAI*, 2005.

L. Lovász. Submodular functions and convexity. *Mathematical programming: the state of the art, Bonn*, pages 235–257, 1982.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. Technical report, arXiv:0908.0050, 2009a.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2272–2279. IEEE, 2009b.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *Advances in Neural Information Processing Systems (NIPS)*, 21, 2009c.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *NIPS*, 2010.

C.A. Micchelli, J.M. Morales, and M. Pontil. Regularizers for structured sparsity. *Arxiv preprint arXiv:1010.0556*, 2011.

S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of ℓ_1-ℓ_∞-regularization. In *Adv. NIPS*, 2008.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. 2009.

G. Obozinski and F. Bach. Convex relaxation of combinatorial penalties. Technical report, HAL, 2012.

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.

F. Rapaport, E. Barillot, and J.-P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382, Jul 2008.

L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, 2005.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B*, 58(1):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 and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.