Machine learning and convex optimization with submodular functions

Francis Bach

Sierra project-team, INRIA - Ecole Normale Supérieure

Workshop on combinatorial optimization - Cargese, 2013

Submodular functions - References

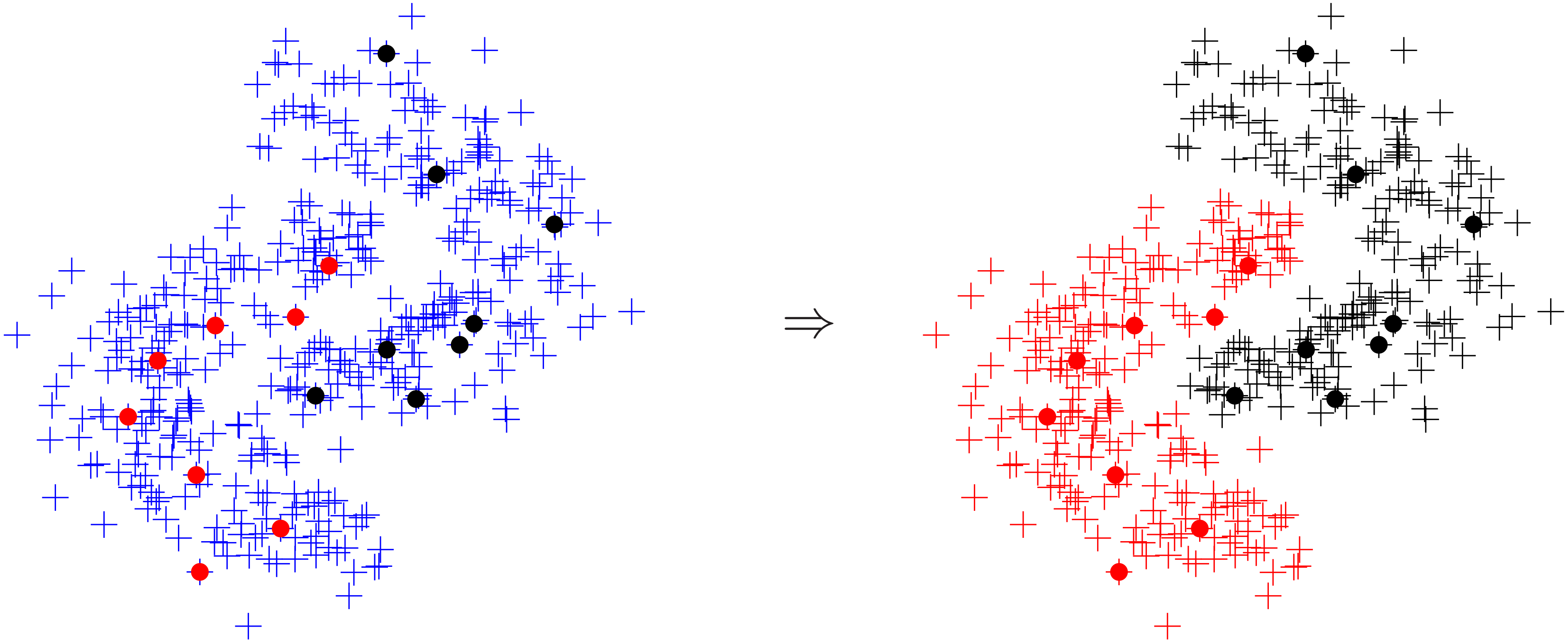
- **References based on combinatorial optimization**
  - *Submodular Functions and Optimization* (Fujishige, 2005)
  - *Discrete convex analysis* (Murota, 2003)

- **Tutorial paper based on convex optimization** (Bach, 2011b)
  - www.di.ens.fr/~fbach/submodular_fot.pdf

- **Slides for this lecture**
  - www.di.ens.fr/~fbach/fbach_cargese_2013.pdf

Submodularity (almost) everywhere
Clustering

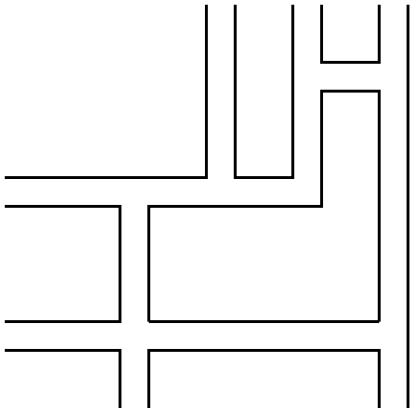
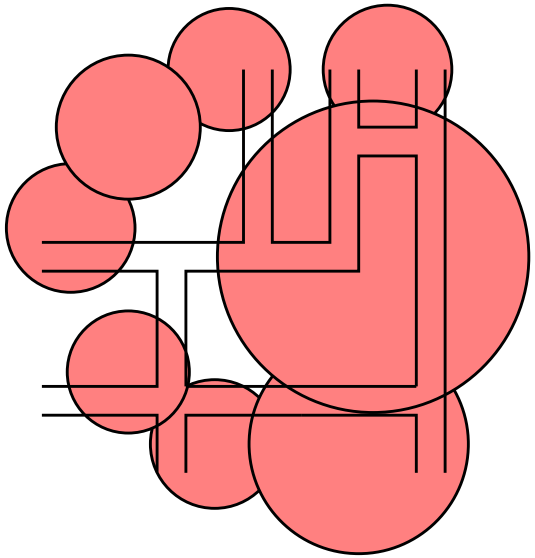
- Semi-supervised clustering



- Submodular function minimization

Submodularity (almost) everywhere
Sensor placement

- Each sensor covers a certain area (Krause and Guestrin, 2005)
  - Goal: maximize coverage

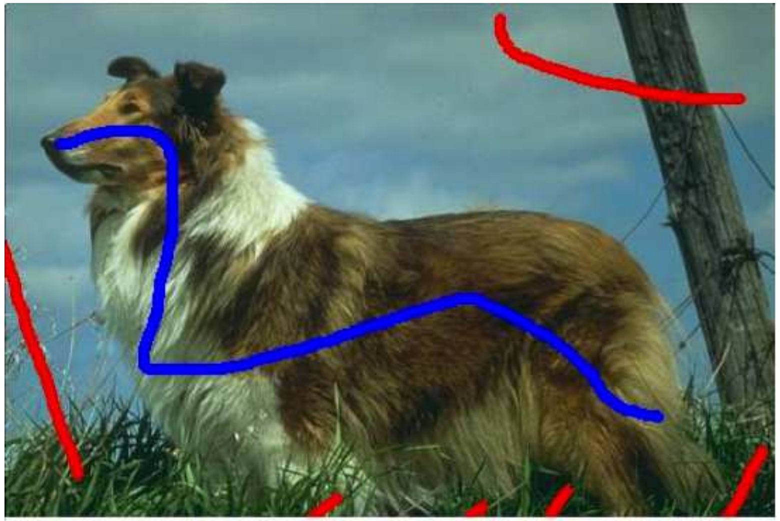



- Submodular function maximization

- Extension to experimental design (Seeger, 2009)

Submodularity (almost) everywhere

Graph cuts and image segmentation

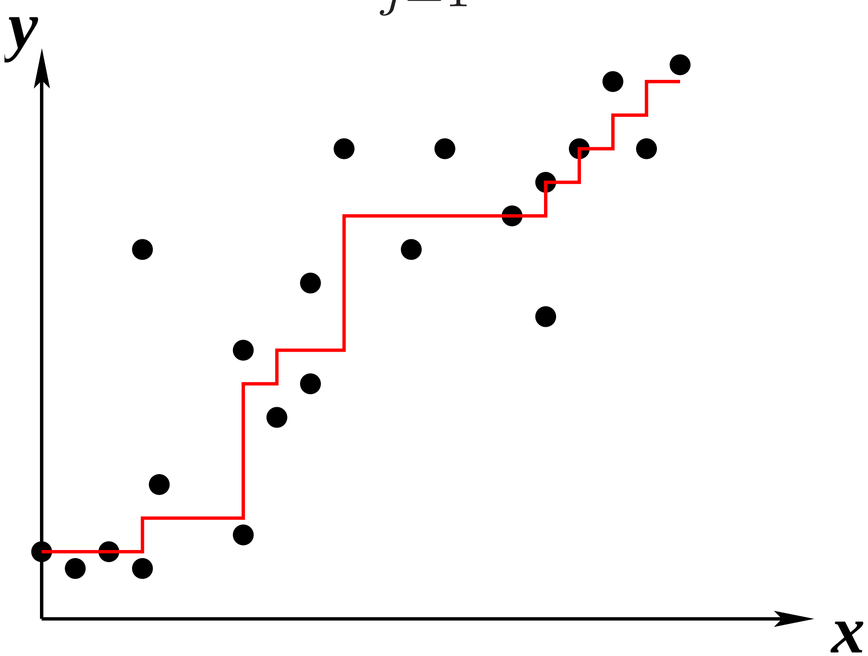


- Submodular function minimization

Submodularity (almost) everywhere

Isotonic regression

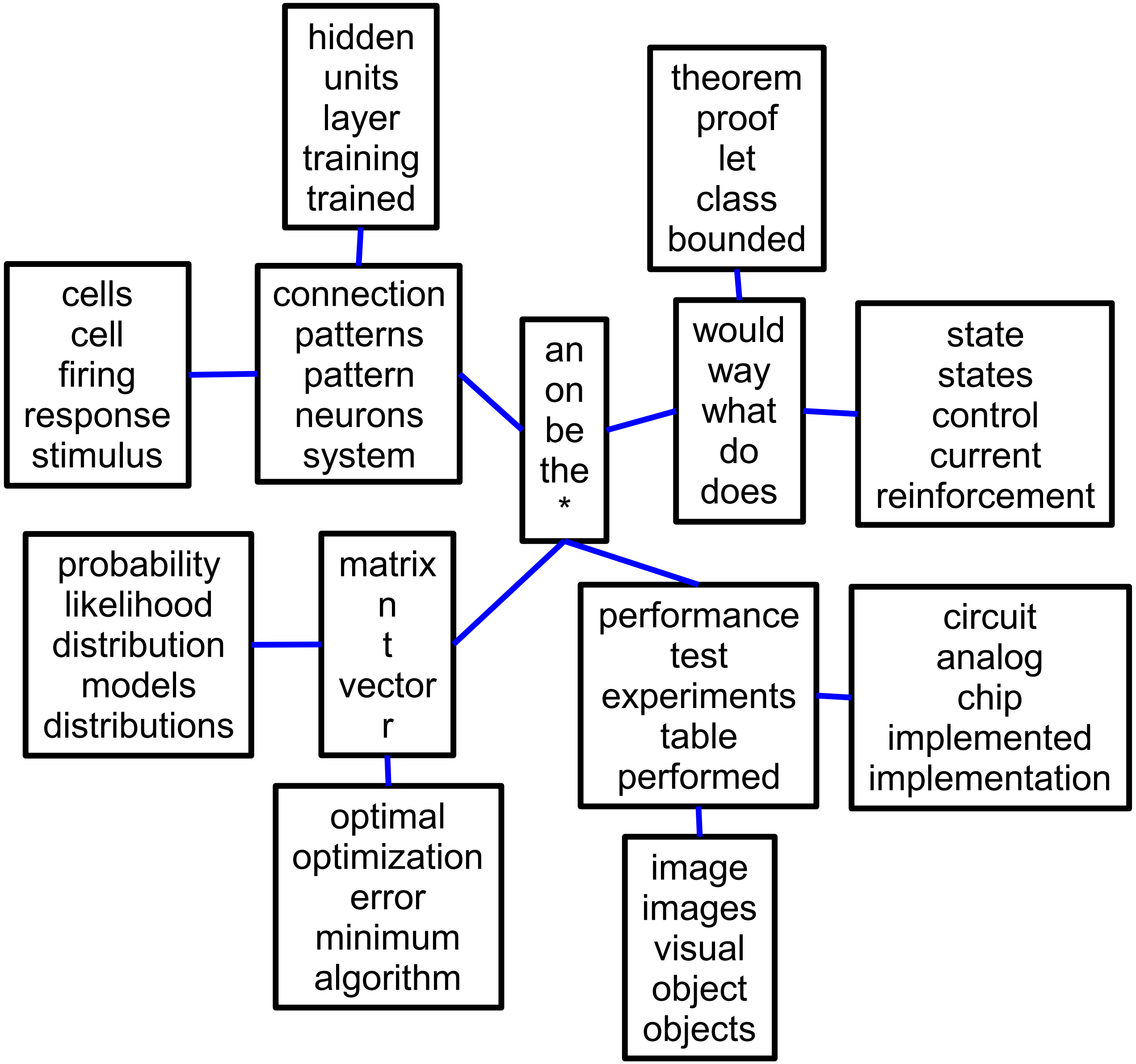
- Given real numbers x_i, $i = 1, \dots, p$
  - Find $y \in \mathbb{R}^p$ that minimizes $\frac{1}{2}\sum_{j=1}^{p}(x_i - y_i)^2$ such that $\forall i, y_i \leqslant y_{i+1}$



- Submodular convex optimization problem

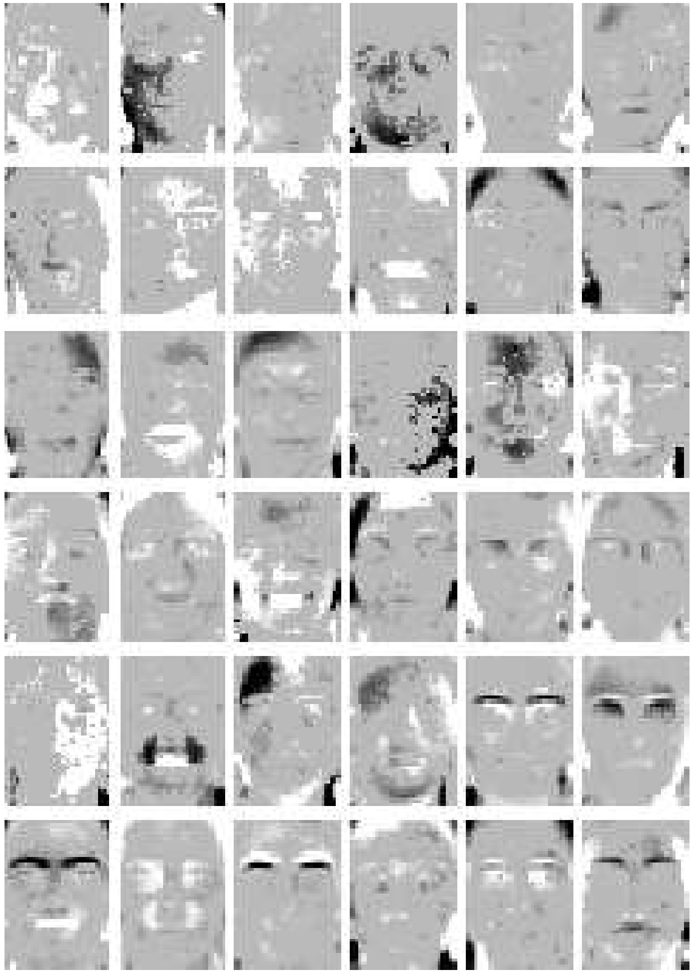
Submodularity (almost) everywhere

Structured sparsity - I



Submodularity (almost) everywhere
Structured sparsity - II

raw data



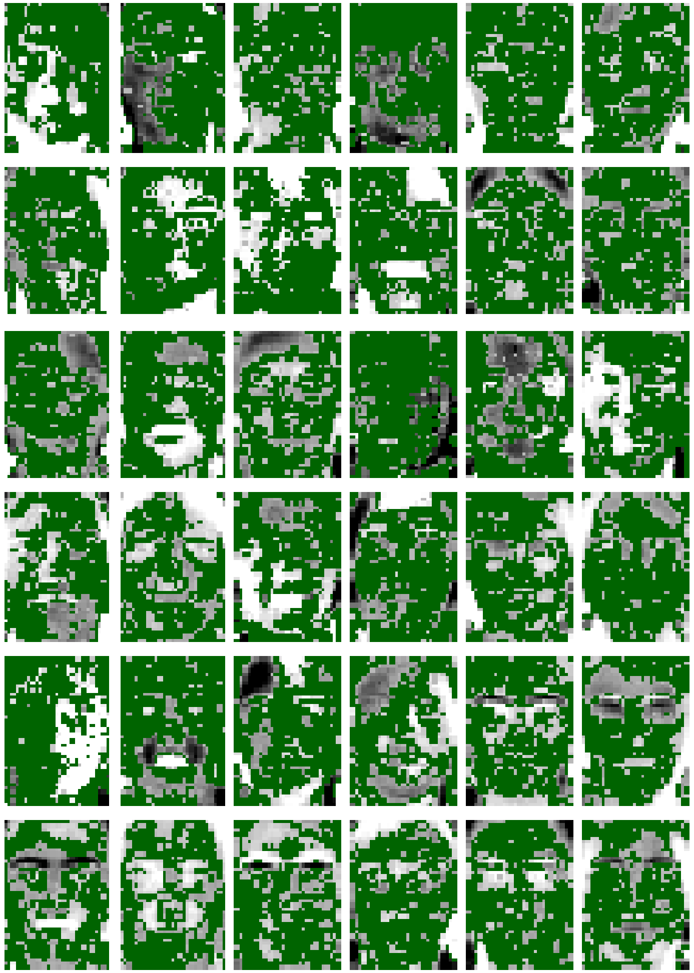
sparse PCA

- No structure: many zeros do not lead to better interpretability

Submodularity (almost) everywhere

Structured sparsity - II

raw data



sparse PCA

- No structure: many zeros do not lead to better interpretability

Submodularity (almost) everywhere
Structured sparsity - II

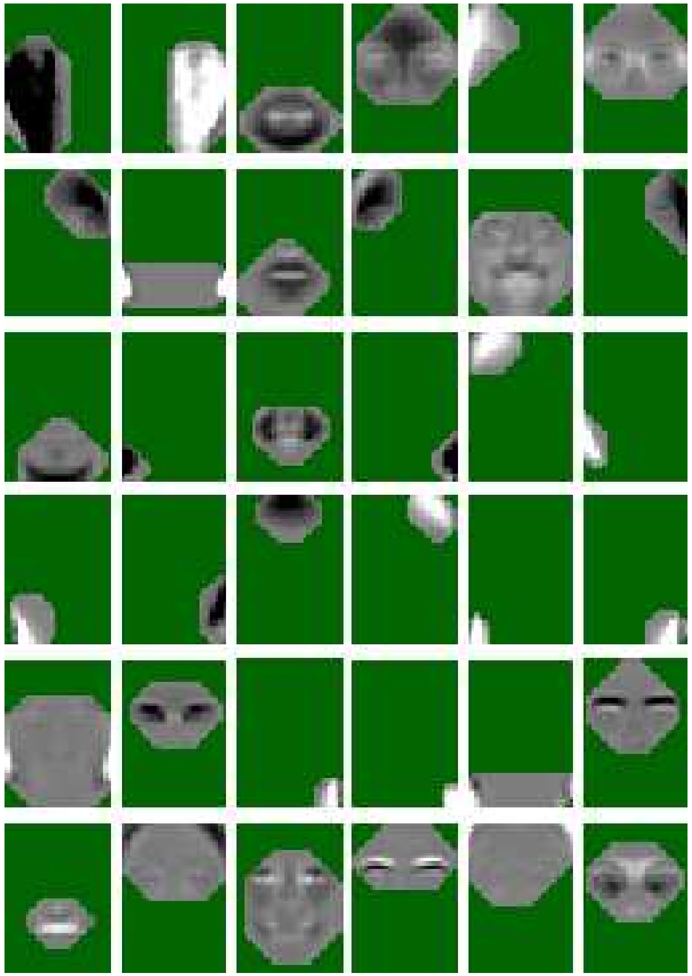
raw data

Structured sparse PCA

- Submodular convex optimization problem

Submodularity (almost) everywhere
Structured sparsity - II

raw data



Structured sparse PCA

- Submodular convex optimization problem

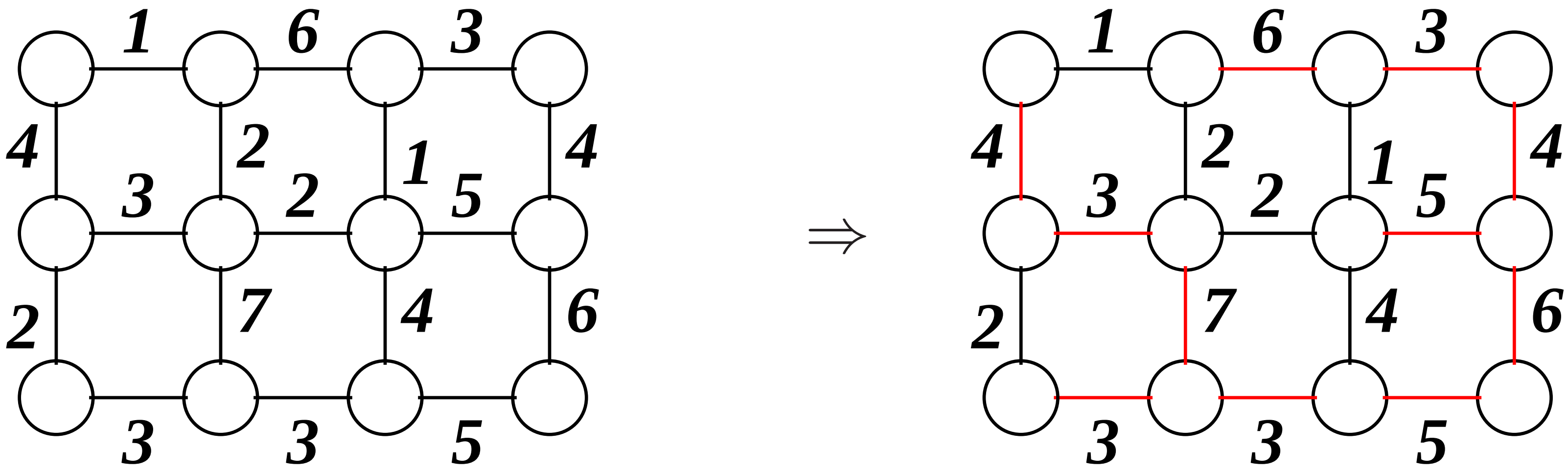
Submodularity (almost) everywhere
Image denoising

- Total variation denoising (Chambolle, 2005)

- Submodular convex optimization problem

Submodularity (almost) everywhere
Maximum weight spanning trees

- Given an undirected graph $G = (V, E)$ and weights $w : E \mapsto \mathbb{R}_+$
  - find the maximum weight spanning tree



- Greedy algorithm for submodular polyhedron - matroid

Submodularity (almost) everywhere
Combinatorial optimization problems

- Set $V = \{1, \dots, p\}$

- Power set 2^V = set of all subsets, of cardinality 2^p

- Minimization/maximization of a set function $F : 2^V \to \mathbb{R}$.

$$\min_{A \subset V} F(A) = \min_{A \in 2^V} F(A)$$

Submodularity (almost) everywhere
Combinatorial optimization problems

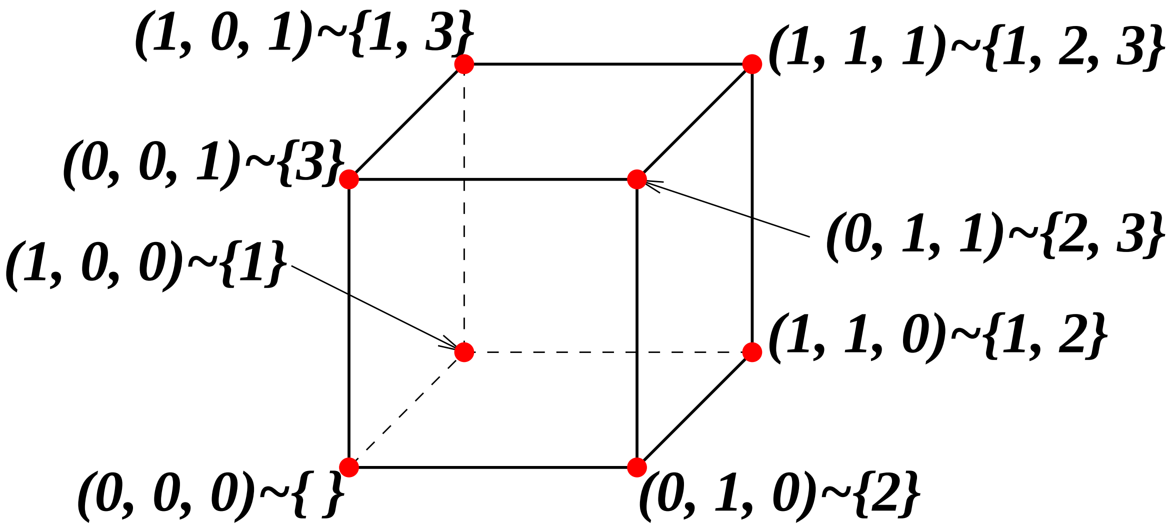
- Set $V = \{1, \dots, p\}$
- Power set 2^V = set of all subsets, of cardinality 2^p
- Minimization/maximization of a set function $F : 2^V \to \mathbb{R}$.

$$\min_{A \subset V} F(A) = \min_{A \in 2^V} F(A)$$

- Reformulation as (pseudo) Boolean function

$$\min_{w \in \{0,1\}^p} f(w)$$

with $\forall A \subset V,\ f(1_A) = F(A)$



Submodularity (almost) everywhere
Convex optimization with combinatorial structure

- **Supervised learning / signal processing**
  - Minimize regularized empirical risk from data (x_i, y_i), $i = 1, \dots, n$:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \Omega(f)$$

  - $\mathcal{F}$ is often a vector space, formulation often convex

- **Introducing discrete structures within a vector space framework**
  - Trees, graphs, etc.
  - Many different approaches (e.g., stochastic processes)

- **Submodularity allows the incorporation of discrete structures**

Outline

1. **Submodular functions**
   - Review and examples of submodular functions
   - Links with convexity through Lovász extension
2. **Submodular minimization**
   - Non-smooth convex optimization
   - Parallel algorithm for special case
3. **Structured sparsity-inducing norms**
   - Relaxation of the penalization of supports by submodular functions
   - Extensions (symmetric, ℓ_q-relaxation)

Submodular functions
Definitions

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

$$\forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B)$$

  – NB: equality for *modular* functions
  – Always assume $F(\varnothing) = 0$

Submodular functions
Definitions

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

$$\forall A, B \subset V, \quad F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B)$$

  - NB: equality for *modular* functions
  - Always assume $F(\varnothing) = 0$

- **Equivalent definition**:

$$\forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

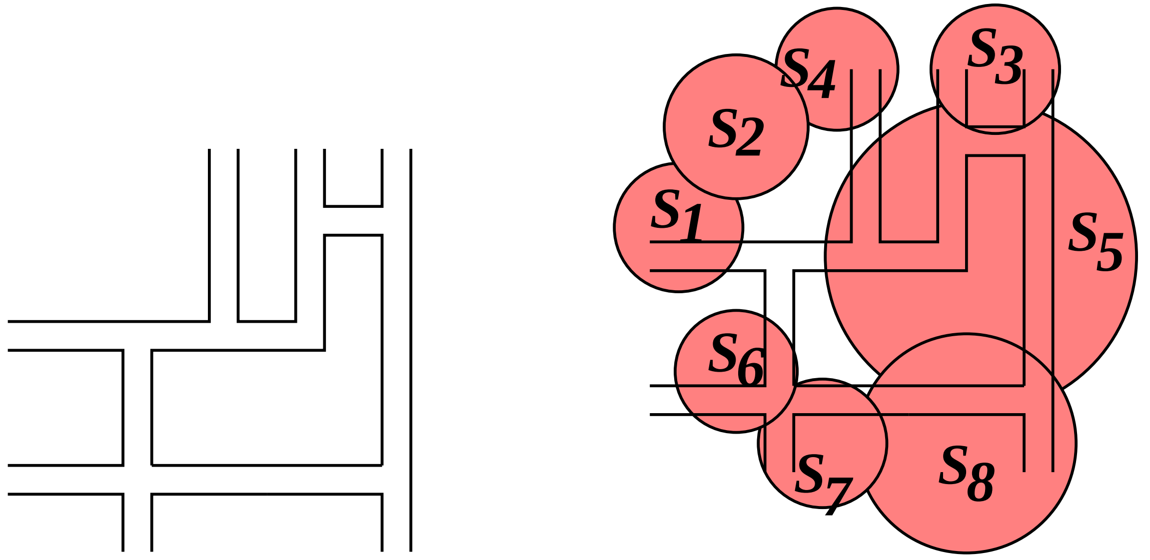
$$\Leftrightarrow \quad \forall A \subset B, \ \forall k \notin A, \ \ F(A \cup \{k\}) - F(A) \geqslant F(B \cup \{k\}) - F(B)$$

  - "Concave property": Diminishing return property

Examples of submodular functions
Cardinality-based functions

- Notation for modular function: $s(A) = \sum_{k \in A} s_k$ for $s \in \mathbb{R}^p$
  - If $s = 1_V$, then $s(A) = |A|$ (cardinality)
- **Proposition**: If $s \in \mathbb{R}_+^p$ and $g : \mathbb{R}_+ \to \mathbb{R}$ is a concave function, then $F : A \mapsto g(s(A))$ is submodular
- **Proposition 2**: If $F : A \mapsto g(s(A))$ is submodular for all $s \in \mathbb{R}_+^p$, then g is concave
- Classical example:
  - $F(A) = 1$ if $|A| > 0$ and 0 otherwise
  - May be rewritten as $F(A) = \max_{k \in V} (1_A)_k$

Examples of submodular functions
Covers

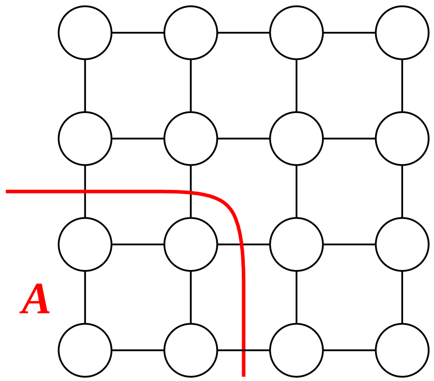


- Let W be any "base" set, and for each $k \in V$, a set $S_k \subset W$
- Set cover defined as $F(A) = \left| \bigcup_{k \in A} S_k \right|$
- *Proof of submodularity* $\Rightarrow$ *homework*

Examples of submodular functions

Cuts

- Given a (un)directed graph, with vertex set V and edge set E
  - $F(A)$ is the total number of edges going from A to $V \backslash A$.



- Generalization with $d : V \times V \to \mathbb{R}_+$

$$F(A) = \sum_{k \in A, j \in V \backslash A} d(k, j)$$

- *Proof of submodularity* $\Rightarrow$ *homework*

Examples of submodular functions
Entropies

- Given p random variables $X_1, \dots, X_p$ with finite number of values
  - Define $F(A)$ as the joint entropy of the variables $(X_k)_{k \in A}$
  - F **is submodular**
- *Proof of submodularity* using data processing inequality (Cover and Thomas, 1991): if $A \subset B$ and $k \notin B$,

$$F(A \cup \{k\}) - F(A) = H(X_A, X_k) - H(X_A) = H(X_k | X_A) \geqslant H(X_k | X_B)$$

- Symmetrized version $G(A) = F(A) + F(V \backslash A) - F(V)$ is mutual information between X_A and $X_{V \backslash A}$
- Extension to continuous random variables, e.g., Gaussian: $F(A) = \log \det \Sigma_{AA}$, for some positive definite matrix $\Sigma \in \mathbb{R}^{p \times p}$

Examples of submodular functions
Flows

- Net-flows from multi-sink multi-source networks (Megiddo, 1974)
- See details in Fujishige (2005); Bach (2011b)
- **Efficient formulation for set covers**

Examples of submodular functions
Matroids

- The pair $(V, \mathcal{I})$ is a matroid with $\mathcal{I}$ its family of independent sets, iff:

(a) $\varnothing \in \mathcal{I}$
(b) $I_1 \subset I_2 \in \mathcal{I} \Rightarrow I_1 \in \mathcal{I}$
(c) for all $I_1, I_2 \in \mathcal{I}$, $|I_1| < |I_2| \Rightarrow \exists k \in I_2 \backslash I_1,\ I_1 \cup \{k\} \in \mathcal{I}$

- Rank function of the matroid, defined as $F(A) = \max_{I \subset A,\ A \in \mathcal{I}} |I|$ is submodular (*direct proof*)

- **Graphic matroid**
  - V edge set of a certain graph $G = (U, V)$
  - $\mathcal{I}$ = set of subsets of edges which do not contain any cycle
  - $F(A) = |U|$ minus the number of connected components of the subgraph induced by A

Outline

1. **Submodular functions**
   - Review and examples of submodular functions
   - Links with convexity through Lovász extension

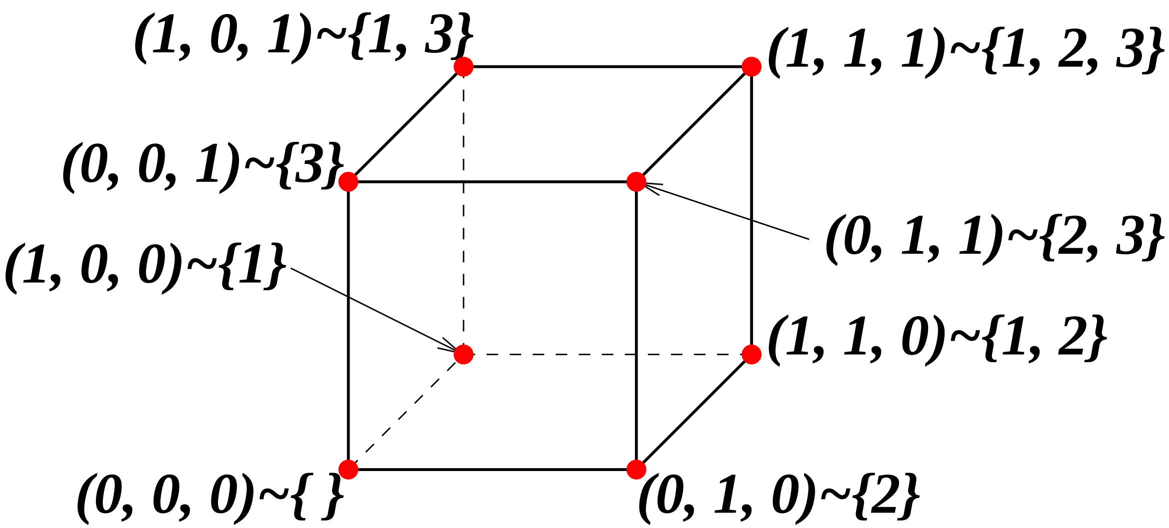
2. **Submodular minimization**
   - Non-smooth convex optimization
   - Parallel algorithm for special case

3. **Structured sparsity-inducing norms**
   - Relaxation of the penalization of supports by submodular functions
   - Extensions (symmetric, ℓ_q-relaxation)

Choquet integral (Choquet, 1954) - 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:

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



Choquet integral (Choquet, 1954) - Lovász extension

Properties

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

- For any set-function F (even not submodular)
  - f is piecewise-linear and positively homogeneous
  - If $w = 1_A$, $f(w) = F(A) \Rightarrow$ extension from $\{0,1\}^p$ to $\mathbb{R}^p$

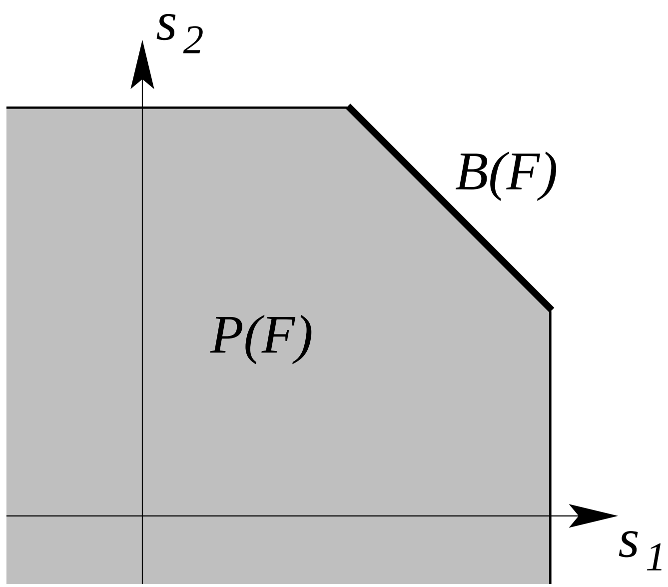
Submodular functions
Links with convexity (Edmonds, 1970; Lovász, 1982)

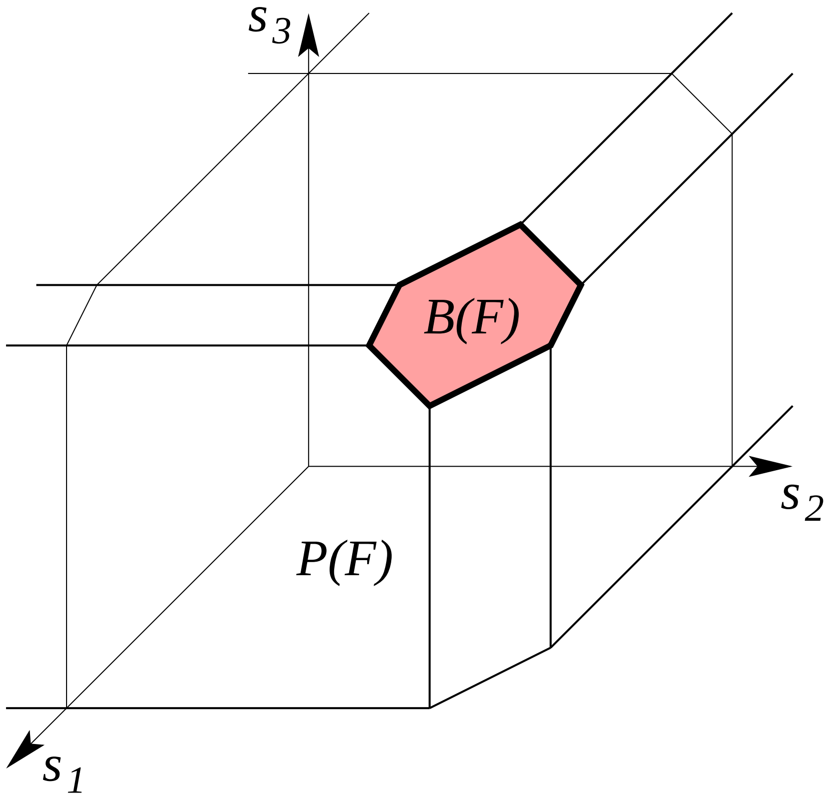
- **Theorem** (Lovász, 1982): F is submodular if and only if f is convex
- Proof requires additional notions from Edmonds (1970):
  - **Submodular and base polyhedra**

Submodular and base polyhedra - Definitions

- Submodular polyhedron: $P(F) = \{s \in \mathbb{R}^p, \ \forall A \subset V, \ s(A) \leqslant F(A)\}$

- Base polyhedron: $B(F) = P(F) \cap \{s(V) = F(V)\}$





- Property: $P(F)$ has non-empty interior

Submodular and base polyhedra - Properties

- Submodular polyhedron: $P(F) = \{s \in \mathbb{R}^p, \ \forall A \subset V, \ s(A) \leqslant F(A)\}$
- Base polyhedron: $B(F) = P(F) \cap \{s(V) = F(V)\}$
- Many facets (up to 2^p), many extreme points (up to $p!$)

Submodular and base polyhedra - Properties

- Submodular polyhedron: $P(F) = \{s \in \mathbb{R}^p, \ \forall A \subset V, \ s(A) \leqslant F(A)\}$

- Base polyhedron: $B(F) = P(F) \cap \{s(V) = F(V)\}$

- Many facets (up to 2^p), many extreme points (up to $p!$)

- **Fundamental property** (Edmonds, 1970): If F is submodular, maximizing linear functions may be done by a "greedy algorithm"
  - Let $w \in \mathbb{R}^p_+$ such that $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$
  - Let $s_{j_k} = F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})$ for $k \in \{1, \dots, p\}$
  - Then $f(w) = \max_{s \in P(F)} w^\top s = \max_{s \in B(F)} w^\top s$
  - Both problems attained at s defined above

- Simple proof by convex duality

Submodular functions
Links with convexity

- **Theorem** (Lovász, 1982): If F is submodular, then

$$\boxed{\min_{A \subset V} F(A) = \min_{w \in \{0,1\}^p} f(w) = \min_{w \in [0,1]^p} f(w)}$$

- Consequence: Submodular function minimization may be done in polynomial time (through ellipsoid algorithm)

- **Representation of $f(w)$ as a support function** (Edmonds, 1970):

$$\boxed{f(w) = \max_{s \in B(F)} s^\top w}$$

  – Maximizer s may be found efficiently through the greedy algorithm

Outline

1. **Submodular functions**
   - Review and examples of submodular functions
   - Links with convexity through Lovász extension

2. **Submodular minimization**
   - Non-smooth convex optimization
   - Parallel algorithm for special case

3. **Structured sparsity-inducing norms**
   - Relaxation of the penalization of supports by submodular functions
   - Extensions (symmetric, ℓ_q-relaxation)

Submodular function minimization
Dual problem

- Let $F : 2^V \to \mathbb{R}$ be a submodular function (such that $F(\varnothing) = 0$)
- **Convex duality** (Edmonds, 1970):

$$\begin{aligned}
\min_{A \subset V} F(A) &= \min_{w \in [0,1]^p} f(w) \\
&= \min_{w \in [0,1]^p} \max_{s \in B(F)} w^\top s \\
&= \max_{s \in B(F)} \min_{w \in [0,1]^p} w^\top s = \max_{s \in B(F)} s_-(V)
\end{aligned}$$

Exact submodular function minimization
Combinatorial algorithms

- Algorithms based on $\min_{A \subset V} F(A) = \max_{s \in B(F)} s_-(V)$
- Output the subset A and a base $s \in B(F)$ as a certificate of optimality
- Best algorithms have polynomial complexity (Schrijver, 2000; Iwata et al., 2001; Orlin, 2009) (typically $O(p^6)$ or more)
- Update a sequence of convex combination of vertices of $B(F)$ obtained from the greedy algorithm using a specific order:
  - Based only on function evaluations
- Recent algorithms using efficient reformulations in terms of generalized graph cuts (Jegelka et al., 2011)

Approximate submodular function minimization

- **For most machine learning applications, no need to obtain exact minimum**
  - For convex optimization, see, e.g., Bottou and Bousquet (2008)

$$\boxed{\min_{A \subset V} F(A) = \min_{w \in \{0,1\}^p} f(w) = \min_{w \in [0,1]^p} f(w)}$$

Approximate submodular function minimization

- **For most machine learning applications, no need to obtain exact minimum**
  - For convex optimization, see, e.g., Bottou and Bousquet (2008)

$$\boxed{\min_{A \subset V} F(A) = \min_{w \in \{0,1\}^p} f(w) = \min_{w \in [0,1]^p} f(w)}$$

- **Important properties of f for convex optimization**
  - Polyhedral function
  - Representation as maximum of linear functions

$$f(w) = \max_{s \in B(F)} w^\top s$$

- **Stability vs. speed vs. generality vs. ease of implementation**

Projected subgradient descent (Shor et al., 1985)

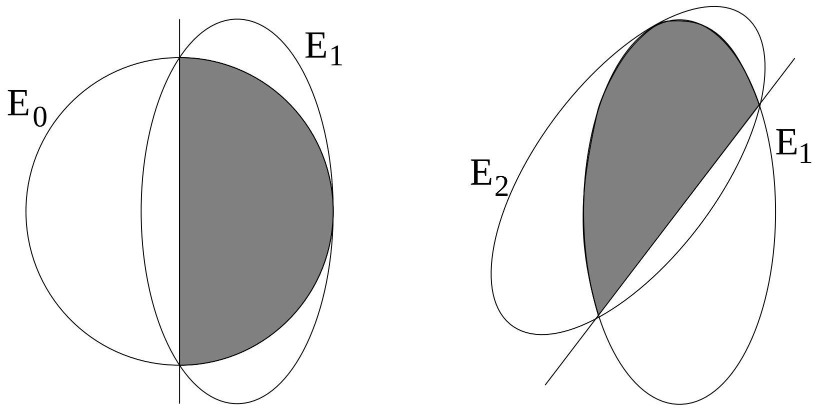
- Subgradient of $f(w) = \max\limits_{s \in B(F)} s^\top w$ through the greedy algorithm

- Using **projected subgradient descent** to minimize f on $[0,1]^p$
  - Iteration: $w_t = \Pi_{[0,1]^p}\big(w_{t-1} - \frac{C}{\sqrt{t}} s_t\big)$ where $s_t \in \partial f(w_{t-1})$
  - Convergence rate: $f(w_t) - \min_{w \in [0,1]^p} f(w) \leqslant \frac{\sqrt{p}}{\sqrt{t}}$ with primal/dual guarantees (Nesterov, 2003)

- Fast iterations but slow convergence
  - need $O(p/\varepsilon^2)$ iterations to reach precision ε
  - need $O(p^2/\varepsilon^2)$ function evaluations to reach precision ε

Ellipsoid method (Nemirovski and Yudin, 1983)

- Build a sequence of minimum volume ellipsoids that enclose the set of solutions



- Cost of a single iteration: p function evaluations and $O(p^3)$ operations

- Number of iterations: $2p^2\big(\max_{A\subset V} F(A) - \min_{A\subset V} F(A)\big) \log \frac{1}{\varepsilon}$.
  - $O(p^5)$ operations and $O(p^3)$ function evaluations

- Slow in practice (the bound is “tight”)

Analytic center cutting planes (Goffin and Vial, 1993)

- **Center of gravity method**
  - improves the convergence rate of ellipsoid method
  - cannot be computed easily

- **Analytic center** of a polytope defined by $a_i^\top w \leqslant b_i$, $i \in I$

$$\min_{w \in \mathbb{R}^p} - \sum_{i \in I} \log(b_i - a_i^\top w)$$

- **Analytic center cutting planes (ACCPM)**
  - Each iteration has complexity $O(p^2|I| + |I|^3)$ using Newton's method
  - No linear convergence rate
  - Good performance in practice

Simplex method for submodular minimization

- Mentioned by Girlich and Pisaruk (1997); McCormick (2005)
- **Formulation as linear program**: $s \in B(F) \Leftrightarrow s = S^\top \eta$, $S \in \mathbb{R}^{d \times p}$

$$\max_{s \in B(F)} s_-(V) = \max_{\eta \geqslant 0,\ \eta^\top 1_d = 1} \sum_{i=1}^{p} \min\{(S^\top \eta)_i, 0\}$$

$$= \max_{\eta \geqslant 0,\ \alpha \geqslant 0,\ \beta \geqslant 0} -\beta^\top 1_p \text{ such that } S^\top \eta - \alpha + \beta = 0,\ \eta^\top 1_d = 1.$$

- **Column generation for simplex methods**: only access the rows of S by maximizing linear functions
  - no complexity bound, may get global optimum if enough iterations

Separable optimization on base polyhedron

- **Optimization of convex functions** of the form $\boxed{\Psi(w) + f(w)}$ with f Lovász extension of F, and $\Psi(w) = \sum_{k \in V} \psi_k(w_k)$

- **Structured sparsity**
  - Total variation denoising - isotonic regression
  - Regularized risk minimization penalized by the Lovász extension

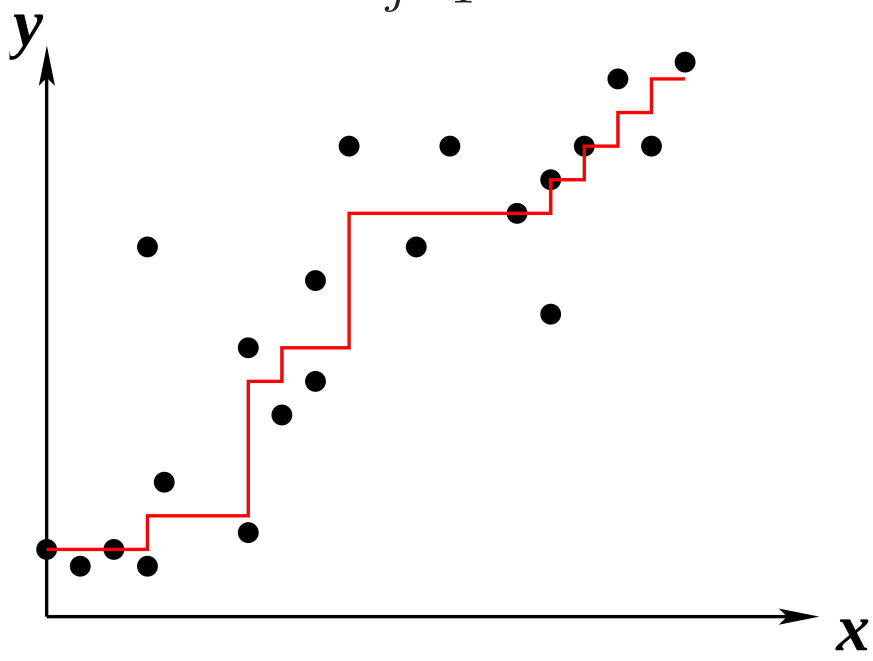
Total variation denoising (Chambolle, 2005)

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

- d symmetric $\Rightarrow f =$ total variation

Isotonic regression

- Given real numbers x_i, $i = 1, \dots, p$
  - Find $y \in \mathbb{R}^p$ that minimizes $\frac{1}{2}\sum_{j=1}^{p}(x_i - y_i)^2$ such that $\forall i, y_i \leqslant y_{i+1}$



- For a directed chain, $f(y) = 0$ if and only if $\forall i, y_i \leqslant y_{i+1}$
- Minimize $\frac{1}{2}\sum_{j=1}^{p}(x_i - y_i)^2 + \lambda f(y)$ for λ large

Separable optimization on base polyhedron

- **Optimization of convex functions** of the form $\boxed{\Psi(w) + f(w)}$ with f Lovász extension of F, and $\Psi(w) = \sum_{k \in V} \psi_k(w_k)$

- **Structured sparsity**
  - Total variation denoising - isotonic regression
  - Regularized risk minimization penalized by the Lovász extension

Separable optimization on base polyhedron

- **Optimization of convex functions** of the form $\boxed{\Psi(w) + f(w)}$ with f Lovász extension of F, and $\Psi(w) = \sum_{k \in V} \psi_k(w_k)$

- **Structured sparsity**
  - Total variation denoising - isotonic regression
  - Regularized risk minimization penalized by the Lovász extension

- **Proximal methods** (see second part)
  - Minimize $\Psi(w) + f(w)$ for smooth Ψ as soon as the following "proximal" problem may be obtained efficiently

$$\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - z\|_2^2 + f(w) = \min_{w \in \mathbb{R}^p} \sum_{k=1}^{p} \frac{1}{2}(w_k - z_k)^2 + f(w)$$

- **Submodular function minimization**

Separable optimization on base polyhedron

Convex duality

- Let $\psi_k : \mathbb{R} \to \mathbb{R}$, $k \in \{1, \dots, p\}$ be p functions. Assume
  - Each ψ_k is strictly convex
  - $\sup_{\alpha \in \mathbb{R}} \psi_j'(\alpha) = +\infty$ and $\inf_{\alpha \in \mathbb{R}} \psi_j'(\alpha) = -\infty$
  - Denote $\psi_1^*, \dots, \psi_p^*$ their Fenchel-conjugates (then with full domain)

Separable optimization on base polyhedron

Convex duality

- Let $\psi_k : \mathbb{R} \to \mathbb{R}$, $k \in \{1, \dots, p\}$ be p functions. Assume
  - Each ψ_k is strictly convex
  - $\sup_{\alpha \in \mathbb{R}} \psi_j'(\alpha) = +\infty$ and $\inf_{\alpha \in \mathbb{R}} \psi_j'(\alpha) = -\infty$
  - Denote $\psi_1^*, \dots, \psi_p^*$ their Fenchel-conjugates (then with full domain)

$$
\begin{aligned}
\min_{w \in \mathbb{R}^p} f(w) + \sum_{j=1}^p \psi_i(w_j) &= \min_{w \in \mathbb{R}^p} \max_{s \in B(F)} w^\top s + \sum_{j=1}^p \psi_j(w_j) \\
&= \max_{s \in B(F)} \min_{w \in \mathbb{R}^p} w^\top s + \sum_{j=1}^p \psi_j(w_j) \\
&= \max_{s \in B(F)} - \sum_{j=1}^p \psi_j^*(-s_j)
\end{aligned}
$$

Separable optimization on base polyhedron
Equivalence with submodular function minimization

- For $\alpha \in \mathbb{R}$, let $A^\alpha \subset V$ be a minimizer of $A \mapsto F(A) + \sum_{j \in A} \psi_j'(\alpha)$
- Let w^* be the unique minimizer of $w \mapsto f(w) + \sum_{j=1}^p \psi_j(w_j)$
- **Proposition** (Chambolle and Darbon, 2009):
  - Given A^α for all $\alpha \in \mathbb{R}$, then $\forall j,\ w_j^* = \sup(\{\alpha \in \mathbb{R},\ j \in A^\alpha\})$
  - Given w^*, then $A \mapsto F(A) + \sum_{j \in A} \psi_j'(\alpha)$ has minimal minimizer $\{w^* > \alpha\}$ and maximal minimizer $\{w^* \geqslant \alpha\}$
- Separable optimization equivalent to a sequence of submodular function minimizations
  - NB: extension of known results from parametric max-flow

Equivalence with submodular function minimization
Proof sketch (Bach, 2011b)

- Duality gap for $\displaystyle \min_{w\in\mathbb{R}^p} f(w) + \sum_{j=1}^p \psi_i(w_j) = \max_{s\in B(F)} -\sum_{j=1}^p \psi_j^*(-s_j)$

$$
\begin{aligned}
& f(w) + \sum_{j=1}^p \psi_i(w_j) - \sum_{j=1}^p \psi_j^*(-s_j) \\
= \quad & f(w) - w^\top s + \sum_{j=1}^p \Big\{ \psi_j(w_j) + \psi_j^*(-s_j) + w_j s_j \Big\} \\
= \quad & \int_{-\infty}^{+\infty} \Big\{ (F + \psi'(\alpha))(\{w \geqslant \alpha\}) - (s + \psi'(\alpha))_-(V) \Big\} d\alpha
\end{aligned}
$$

- Duality gap for convex problems = sums of duality gaps for combinatorial problems

Separable optimization on base polyhedron
Quadratic case

- Let F be a submodular function and $w \in \mathbb{R}^p$ the unique minimizer of $w \mapsto f(w) + \frac{1}{2}\|w\|_2^2$. Then:

(a) $s = -w$ is the point in $B(F)$ with minimum ℓ_2-norm
(b) For all $\lambda \in \mathbb{R}$, the maximal minimizer of $A \mapsto F(A) + \lambda|A|$ is $\{w \geqslant -\lambda\}$ and the minimal minimizer of F is $\{w > -\lambda\}$

- **Consequences**
  - Threshold at 0 the minimum norm point in $B(F)$ to minimize F (Fujishige and Isotani, 2011)
  - Minimizing submodular functions with cardinality constraints (Nagano et al., 2011)

From convex to combinatorial optimization

- Solving $\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$ to solve $\min_{A \subset V} F(A)$
  - Thresholding solutions w at zero if $\forall k \in V, \psi_k'(0) = 0$
  - For quadratic functions $\psi_k(w_k) = \frac{1}{2} w_k^2$, equivalent to projecting 0 on $B(F)$ (Fujishige, 2005)

From convex to combinatorial optimization and vice-versa...

- Solving $\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$ to solve $\min_{A \subset V} F(A)$
  - Thresholding solutions w at zero if $\forall k \in V, \psi_k'(0) = 0$
  - For quadratic functions $\psi_k(w_k) = \frac{1}{2} w_k^2$, equivalent to projecting 0 on $B(F)$ (Fujishige, 2005)

- Solving $\min_{A \subset V} F(A) - t(A)$ to solve $\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$
  - General decomposition strategy (Groenevelt, 1991)
  - Efficient only when submodular minimization is efficient

Solving $\min_{A\subset V} F(A) - t(A)$ to solve $\min_{w\in\mathbb{R}^p} \sum_{k\in V} \psi_k(w_k) + f(w)$

- General recursive divide-and-conquer algorithm (Groenevelt, 1991)

- NB: Dual version of Fujishige (2005)

1. Compute minimizer $t \in \mathbb{R}^p$ of $\sum_{j\in V} \psi_j^*(-t_j)$ s.t. $t(V) = F(V)$
2. Compute minimizer A of $F(A) - t(A)$
3. If $A = V$, then t is optimal. Exit.
4. Compute a minimizer s_A of $\sum_{j\in A} \psi_j^*(-s_j)$ over $s \in B(F_A)$ where $F_A : 2^A \to \mathbb{R}$ is the restriction of F to A, i.e., $F_A(B) = F(A)$
5. Compute a minimizer $s_{V\backslash A}$ of $\sum_{j\in V\backslash A} \psi_j^*(-s_j)$ over $s \in B(F^A)$ where $F^A(B) = F(A\cup B) - F(A)$, for $B \subset V\backslash A$
6. Concatenate s_A and $s_{V\backslash A}$. Exit.

Solving $\min_{w\in\mathbb{R}^p} \sum_{k\in V} \psi_k(w_k) + f(w)$ to solve $\min_{A\subset V} F(A)$

- Dual problem: $\max_{s\in B(F)} -\sum_{j=1}^p \psi_j^*(-s_j)$

- Constrained optimization when linear functions can be maximized
  - **Frank-Wolfe algorithms**

- Two main types for convex functions

Approximate quadratic optimization on $B(F)$

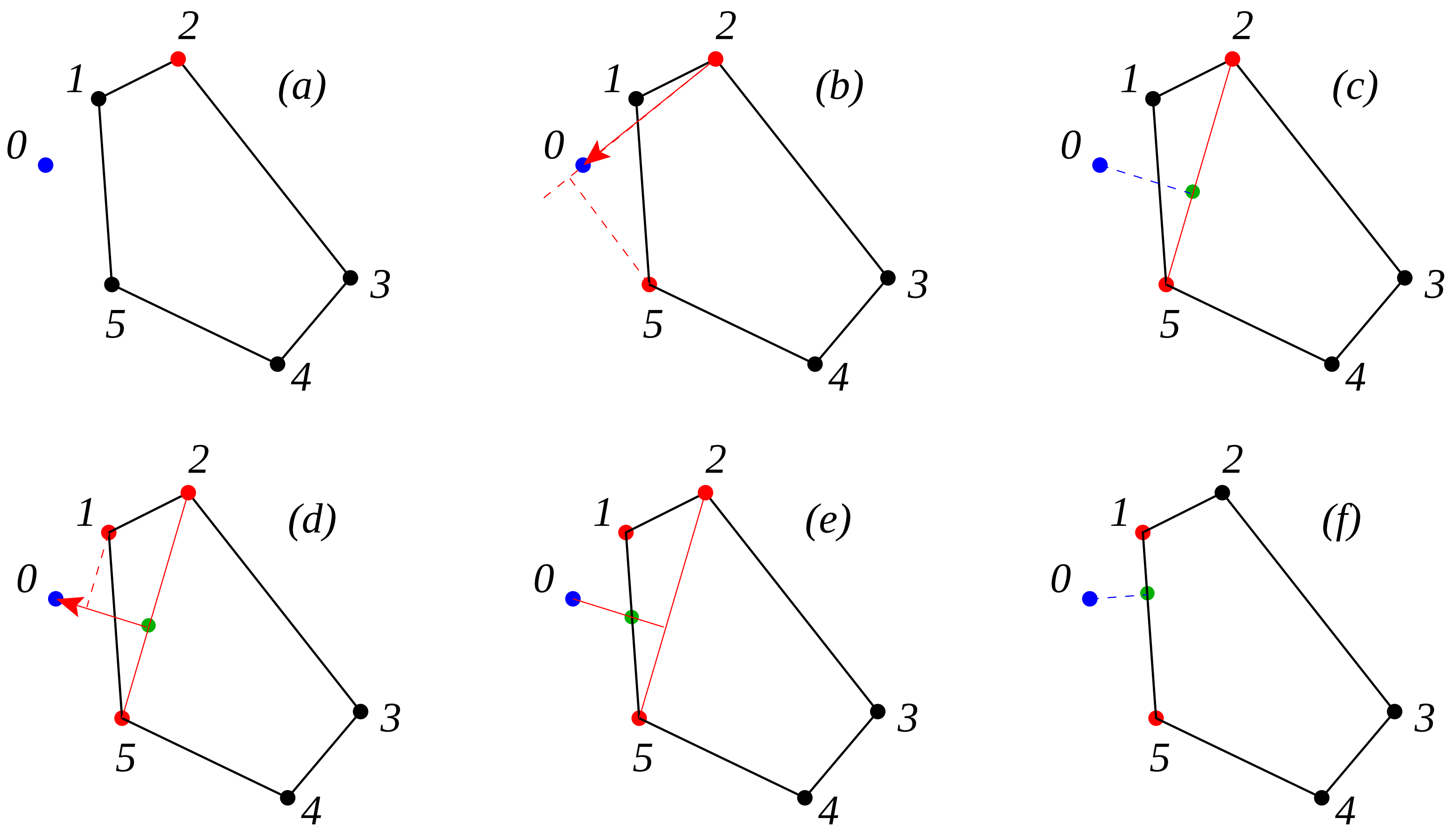
- **Goal**: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w\|_2^2 + f(w) = \max_{s \in B(F)} -\frac{1}{2}\|s\|_2^2$

- Can only maximize linear functions on $B(F)$

- **Two types of "Frank-wolfe" algorithms**

- **1. Active set algorithm ($\Leftrightarrow$ min-norm-point)**
  - Sequence of maximizations of linear functions over $B(F)$ $+$ overheads (affine projections)
  - Finite convergence, but no complexity bounds

Minimum-norm-point algorithm (Wolfe, 1976)



Approximate quadratic optimization on $B(F)$

- **Goal**: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w\|_2^2 + f(w) = \max_{s \in B(F)} -\frac{1}{2}\|s\|_2^2$

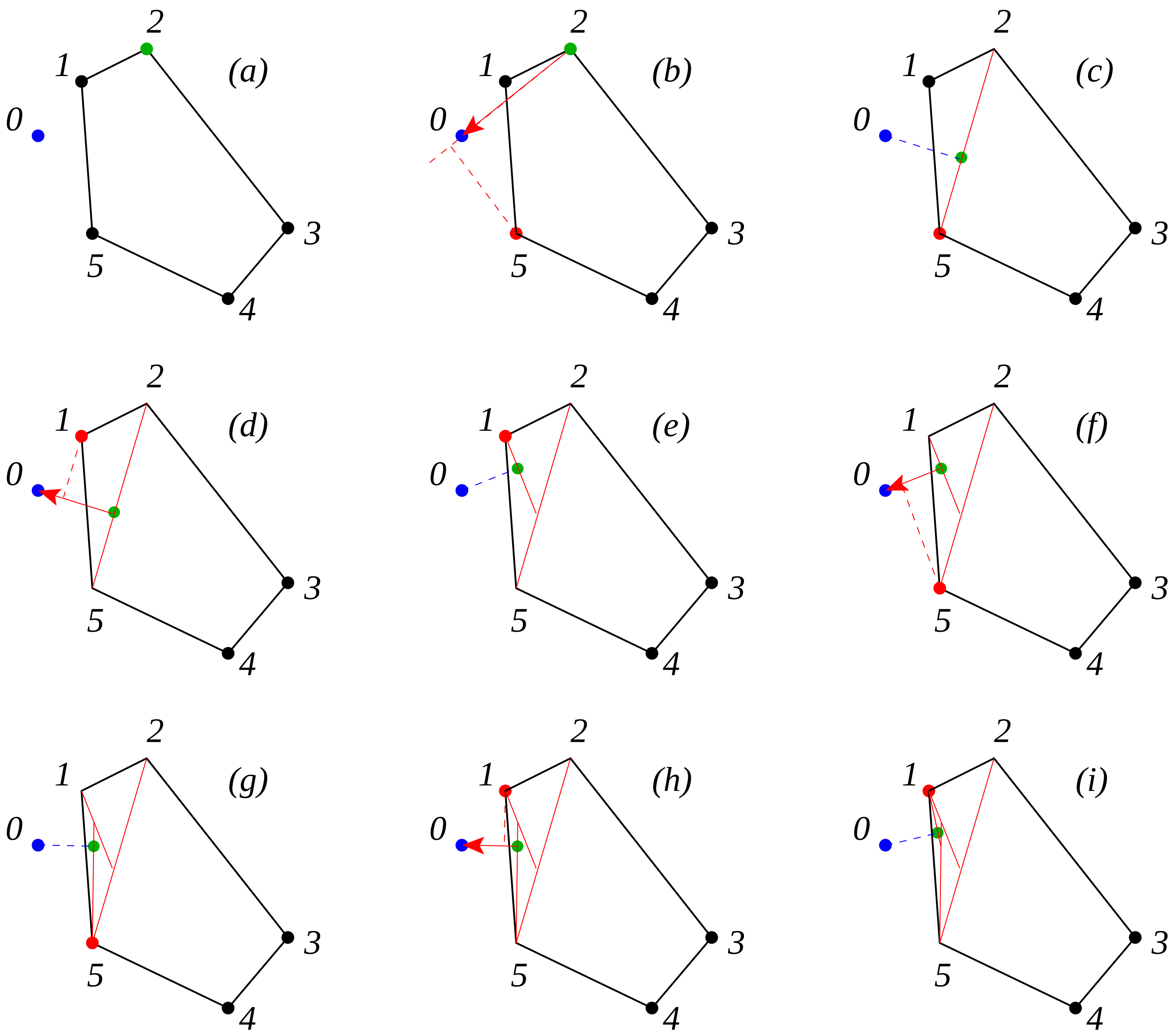
- Can only maximize linear functions on $B(F)$

- **Two types of "Frank-wolfe" algorithms**

- **1. Active set algorithm ($\Leftrightarrow$ min-norm-point)**
  - Sequence of maximizations of linear functions over $B(F)$ $+$ overheads (affine projections)
  - Finite convergence, but no complexity bounds

- **2. Conditional gradient**
  - Sequence of maximizations of linear functions over $B(F)$
  - Approximate optimality bound

Conditional gradient with line search



Approximate quadratic optimization on $B(F)$

- **Proposition**: t steps of conditional gradient (with line search) outputs $s_t \in B(F)$ and $w_t = -s_t$, such that

$$f(w_t) + \frac{1}{2}\|w_t\|_2^2 - \mathrm{OPT} \leqslant f(w_t) + \frac{1}{2}\|w_t\|_2^2 + \frac{1}{2}\|s_t\|_2^2 \leqslant \frac{2D^2}{t}$$

Approximate quadratic optimization on $B(F)$

- **Proposition**: t steps of conditional gradient (with line search) outputs $s_t \in B(F)$ and $w_t = -s_t$, such that

$$f(w_t) + \frac{1}{2}\|w_t\|_2^2 - \text{OPT} \leqslant f(w_t) + \frac{1}{2}\|w_t\|_2^2 + \frac{1}{2}\|s_t\|_2^2 \leqslant \frac{2D^2}{t}$$

- **Improved primal candidate through isotonic regression**
  - $f(w)$ is linear on any set of w with fixed ordering
  - May be optimized using isotonic regression ("pool-adjacent-violator") in $O(n)$ (see, e.g., Best and Chakravarti, 1990)
  - Given $w_t = -s_t$, keep the ordering and reoptimize

Approximate quadratic optimization on $B(F)$

- **Proposition**: t steps of conditional gradient (with line search) outputs $s_t \in B(F)$ and $w_t = -s_t$, such that

$$f(w_t) + \frac{1}{2}\|w_t\|_2^2 - \mathrm{OPT} \leqslant f(w_t) + \frac{1}{2}\|w_t\|_2^2 + \frac{1}{2}\|s_t\|_2^2 \leqslant \frac{2D^2}{t}$$

- **Improved primal candidate through isotonic regression**
  - $f(w)$ is linear on any set of w with fixed ordering
  - May be optimized using isotonic regression ("pool-adjacent-violator") in $O(n)$ (see, e.g. Best and Chakravarti, 1990)
  - Given $w_t = -s_t$, keep the ordering and reoptimize

- **Better bound for submodular function minimization?**

From quadratic optimization on $B(F)$ to submodular function minimization

- **Proposition**: If w is ε-optimal for $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w\|_2^2 + f(w)$, then at least a levet set A of w is $\left(\frac{\sqrt{\varepsilon p}}{2}\right)$-optimal for submodular function minimization

- If $\varepsilon = \dfrac{2D^2}{t}$, $\dfrac{\sqrt{\varepsilon p}}{2} = \dfrac{Dp^{1/2}}{\sqrt{2t}} \Rightarrow$ **no provable gains**, but:
  - Bound on the iterates A_t (with additional assumptions)
  - Possible thresolding for acceleration

From quadratic optimization on $B(F)$ to submodular function minimization

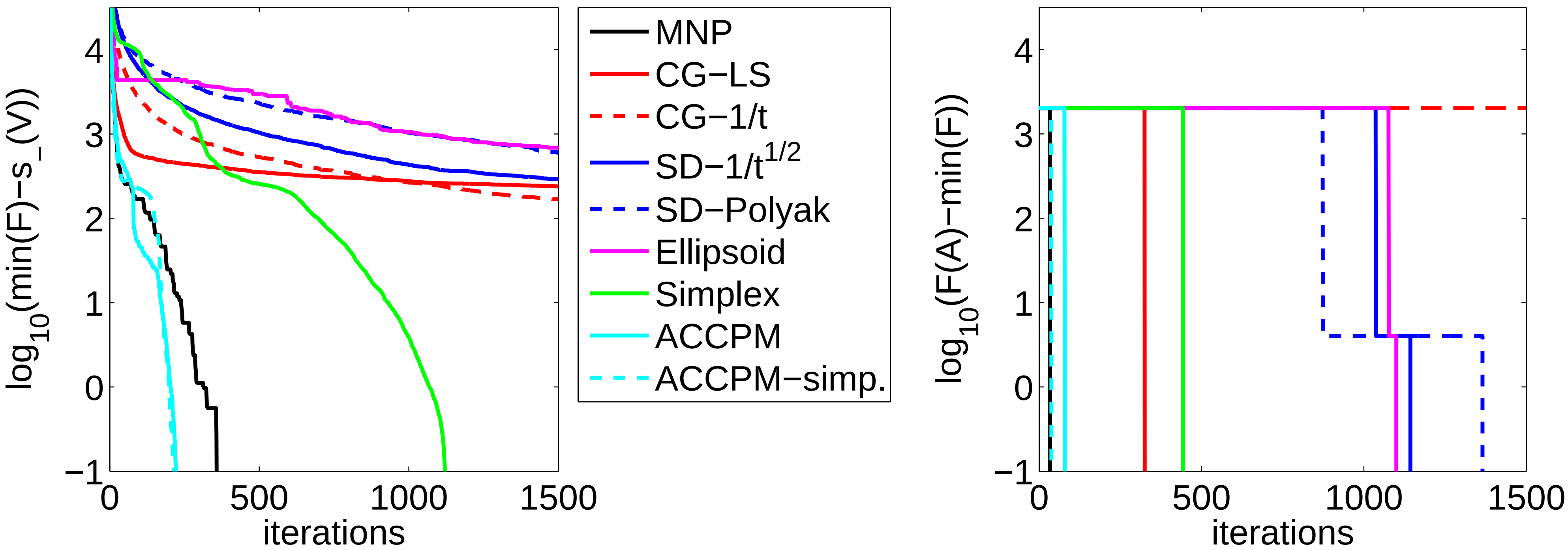
- **Proposition**: If w is ε-optimal for $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w\|_2^2 + f(w)$, then at least a levet set A of w is $\left(\frac{\sqrt{\varepsilon p}}{2}\right)$-optimal for submodular function minimization

- If $\varepsilon = \dfrac{2D^2}{t}$, $\dfrac{\sqrt{\varepsilon p}}{2} = \dfrac{Dp^{1/2}}{\sqrt{2t}} \Rightarrow$ **no provable gains**, but:
  - Bound on the iterates A_t (with additional assumptions)
  - Possible thresolding for acceleration

- **Lower complexity bound for SFM**
  - **Conjecture**: no algorithm that is based **only** on a sequence of greedy algorithms obtained from linear combinations of bases can improve on the subgradient bound (after $p/2$ iterations).

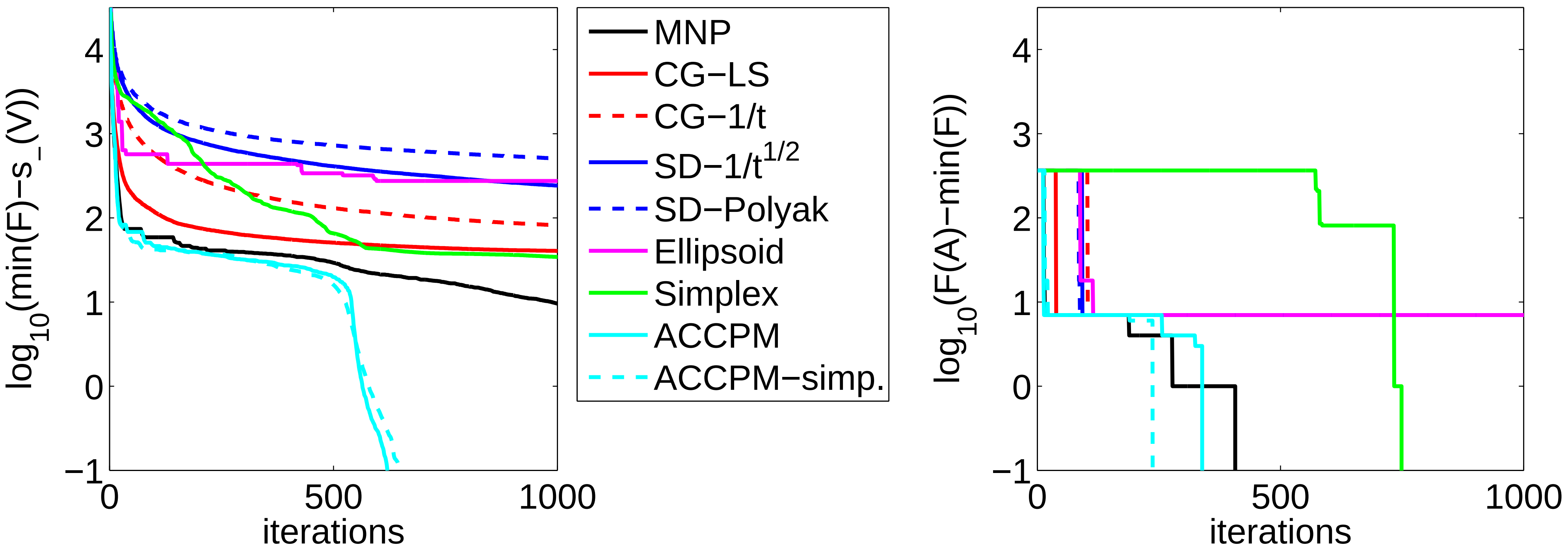
Simulations on standard benchmark "DIMACS Genrmf-wide", p = 430

- **Submodular function minimization**
  - (Left) dual suboptimality
  - (Right) primal suboptimality



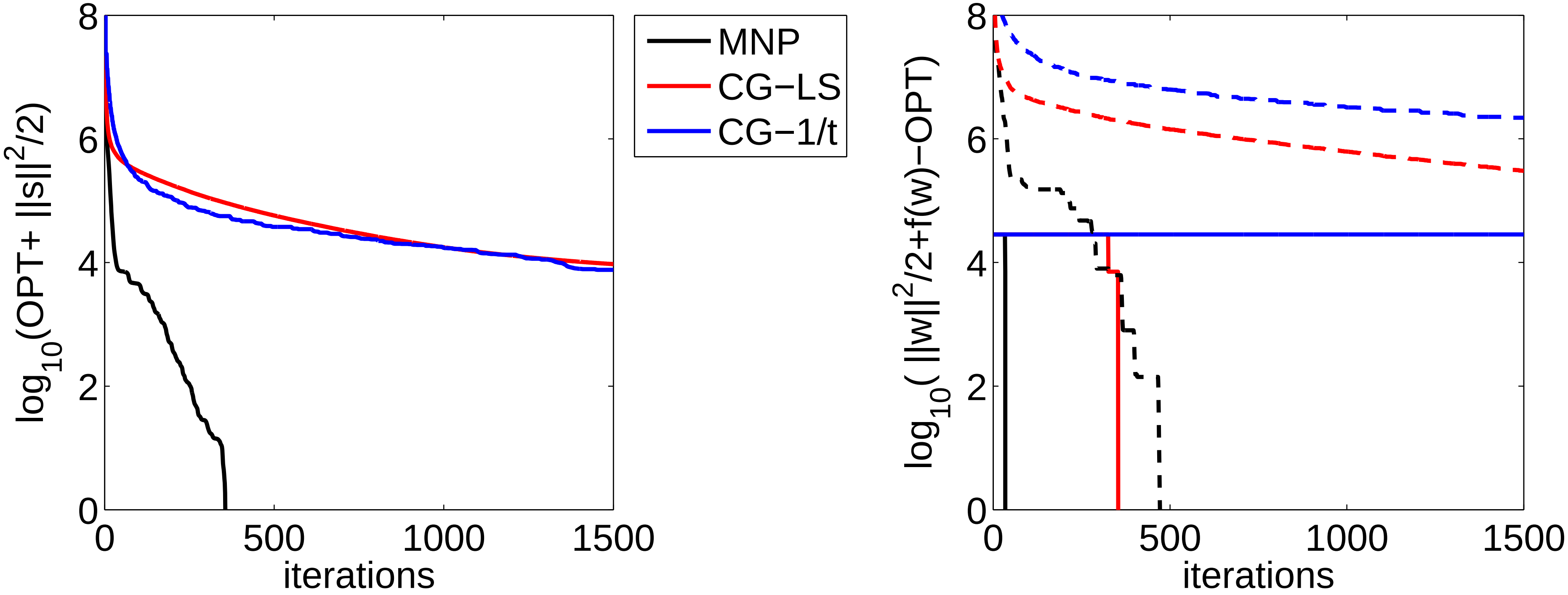
Simulations on standard benchmark "DIMACS Genrmf-long", p = 575

- **Submodular function minimization**
  - (Left) dual suboptimality
  - (Right) primal suboptimality



Simulations on standard benchmark

- **Separable quadratic optimization**
  - (Left) dual suboptimality
  - (Right) primal suboptimality
    (in dashed, before the pool-adjacent-violator correction)



Outline

1. **Submodular functions**
   - Review and examples of submodular functions
   - Links with convexity through Lovász extension

2. **Submodular minimization**
   - Non-smooth convex optimization
   - Parallel algorithm for special case

3. **Structured sparsity-inducing norms**
   - Relaxation of the penalization of supports by submodular functions
   - Extensions (symmetric, ℓ_q-relaxation)

From submodular minimization to proximal problems

- **Summary**: several optimization problems
  - Discrete problem: $\min_{A \subset V} F(A) = \min_{w \in \{0,1\}^p} f(w)$
  - Continuous problem: $\min_{w \in [0,1]^p} f(w)$
  - Proximal problem (P): $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w\|_2^2 + f(w)$

- **Solving (P) is equivalent to minimizing $F(A) + \lambda|A|$ for all λ**
  - $\arg\min_{A \subseteq V} F(A) + \lambda|A| = \{k, w_k \geqslant -\lambda\}$

- Much simpler problem but no gains in terms of (provable) complexity
  - See Bach (2011a)

Decomposable functions

- F may often be decomposed as the sum of r "simple" functions:

$$F(A) = \sum_{j=1}^{r} F_j(A)$$

  - Each F_j may be minimized efficiently
  - Example: 2D grid = vertical chains + horizontal chains

- Komodakis et al. (2011); Kolmogorov (2012); Stobbe and Krause (2010); Savchynskyy et al. (2011)
  - Dual decomposition approach but slow non-smooth problem

Decomposable functions and proximal problems (Jegelka, Bach, and Sra, 2013)

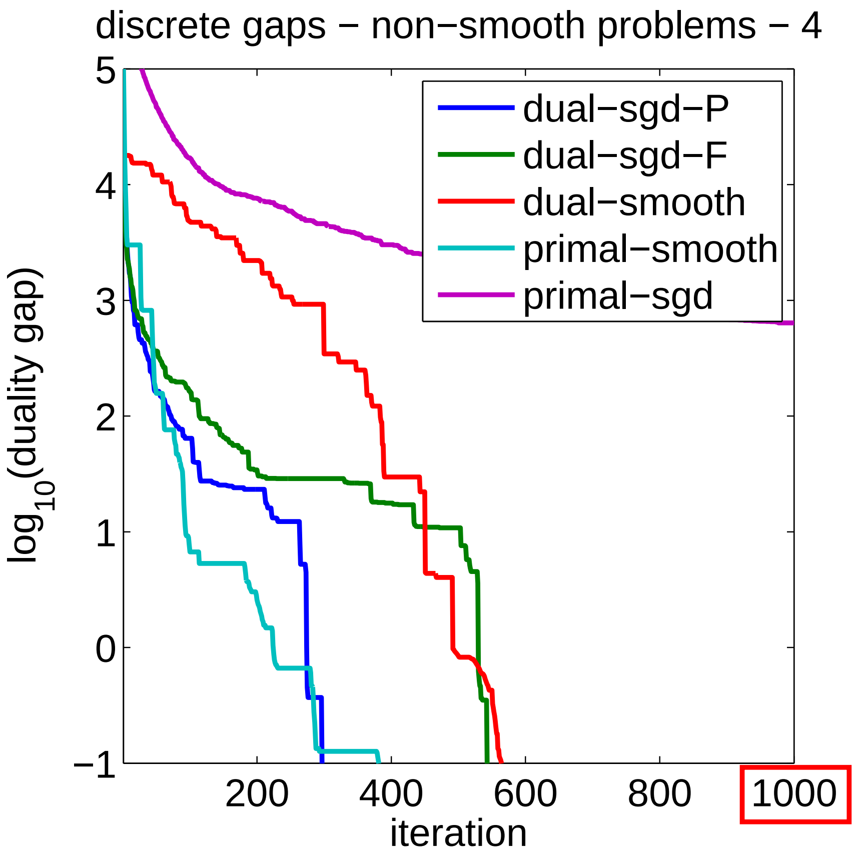
- Dual problem

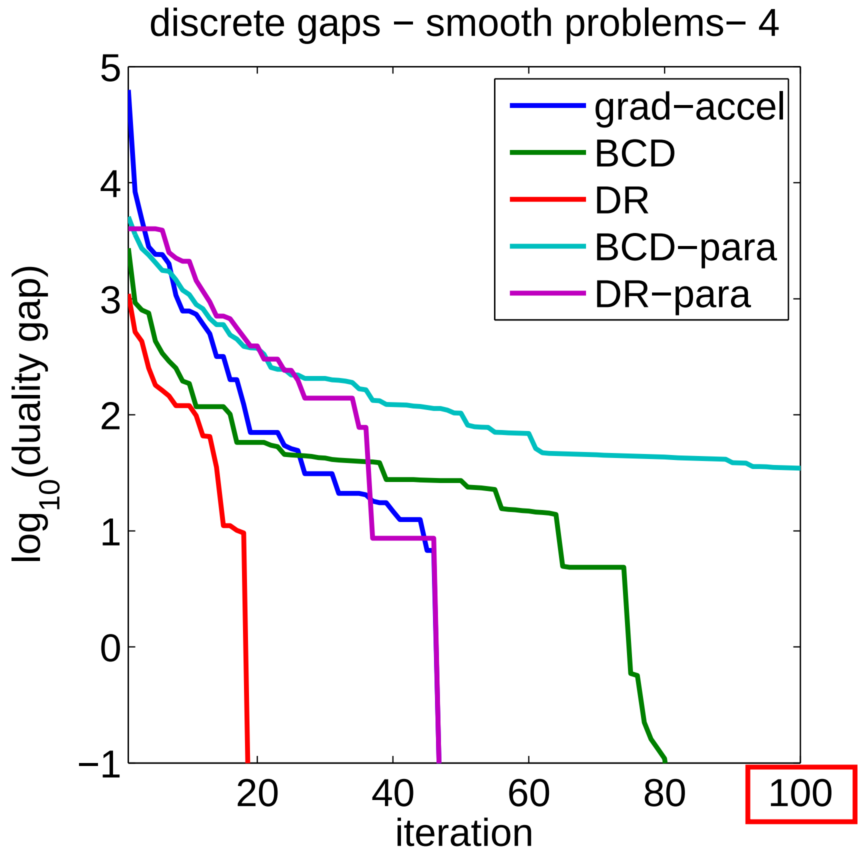
$$
\begin{aligned}
& \min_{w \in \mathbb{R}^p} f_1(w) + f_2(w) + \frac{1}{2}\|w\|_2^2 \\
= \quad & \min_{w \in \mathbb{R}^p} \max_{s_1 \in B(F_1)} s_1^\top w + \max_{s_2 \in B(F_2)} s_2^\top w + \frac{1}{2}\|w\|_2^2 \\
= \quad & \max_{s_1 \in B(F_1),\ s_2 \in B(F_2)} -\frac{1}{2}\|s_1 + s_2\|^2
\end{aligned}
$$

- **Finding the closest point between two polytopes**
  - Several alternatives: Block coordinate ascent, Douglas Rachford splitting (Bauschke et al., 2004)
  - (a) no parameters, (b) parallelizable

Experiments

- Graph cuts on a 500×500 image

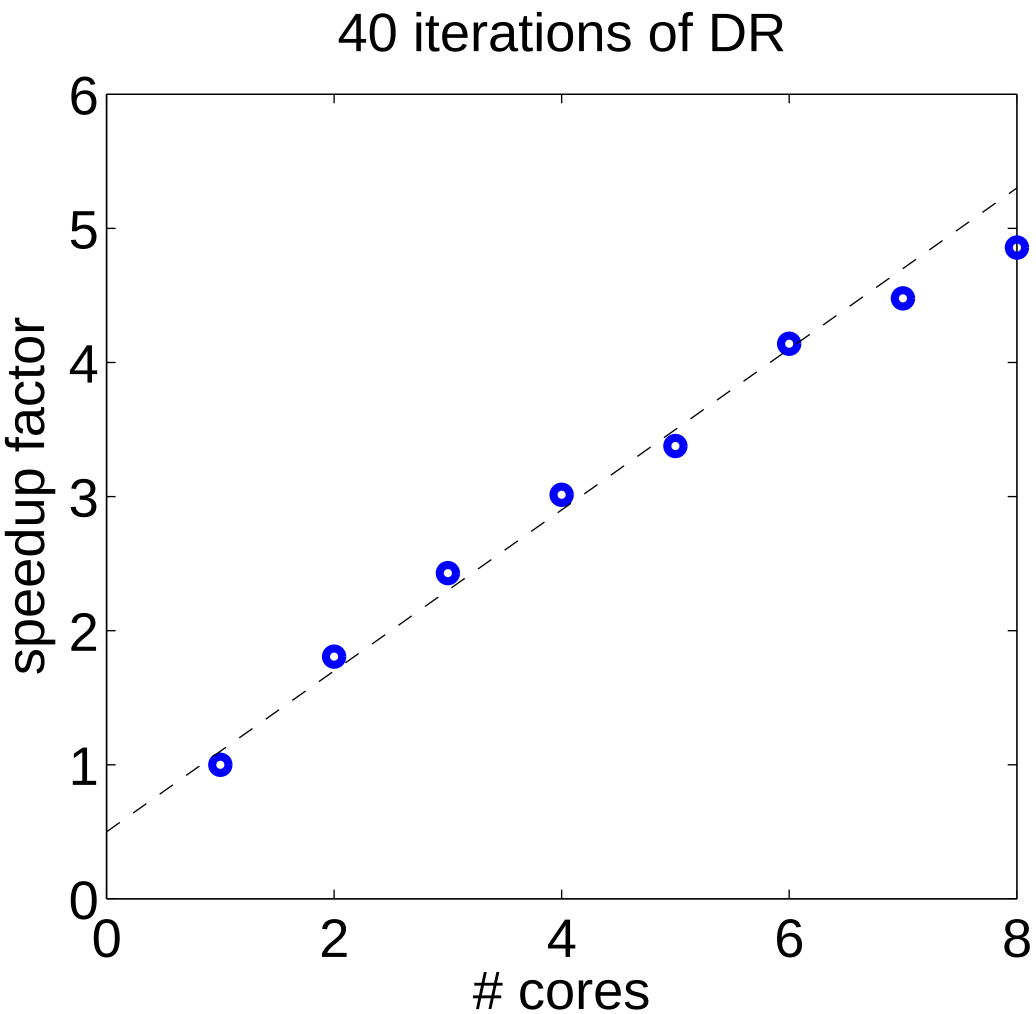




- Matlab/C implementation 10 times slower than C-code for graph cut
  - Easy to code and parallelizable

Parallelization

- **Multiple cores**



Outline

1. **Submodular functions**
   - Review and examples of submodular functions
   - Links with convexity through Lovász extension

2. **Submodular minimization**
   - Non-smooth convex optimization
   - Parallel algorithm for special case

3. **Structured sparsity-inducing norms**
   - Relaxation of the penalization of supports by submodular functions
   - Extensions (symmetric, ℓ_q-relaxation)

Structured sparsity through submodular functions
References and Links

- **References on submodular functions**
  - *Submodular Functions and Optimization* (Fujishige, 2005)
  - Tutorial paper based on convex optimization (Bach, 2011b)
    `www.di.ens.fr/~fbach/submodular_fot.pdf`
- **Structured sparsity through convex optimization**
  - Algorithms (Bach, Jenatton, Mairal, and Obozinski, 2011)
    `www.di.ens.fr/~fbach/bach_jenatton_mairal_obozinski_FOT.pdf`
  - Theory/applications (Bach, Jenatton, Mairal, and Obozinski, 2012)
    `www.di.ens.fr/~fbach/stat_science_structured_sparsity.pdf`
  - Matlab/R/Python codes: `http://www.di.ens.fr/willow/SPAMS/`
- **Slides**: `www.di.ens.fr/~fbach/fbach_cargese_2013.pdf`

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 $y = (y^1, \ldots, y^k) \in \mathbb{R}^{n \times k}$

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

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

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

  - 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)

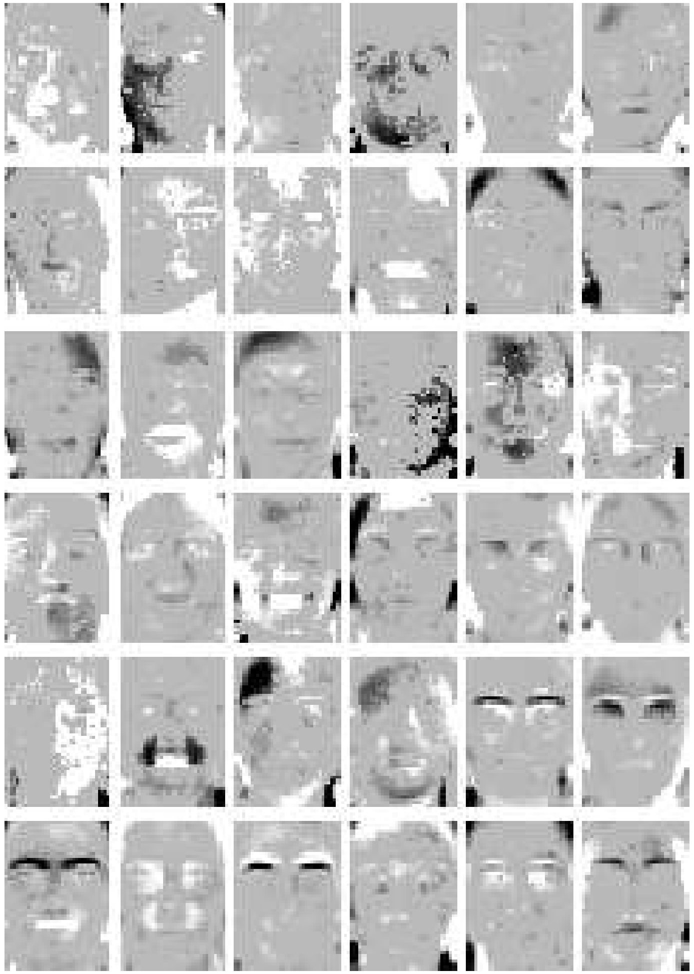
- **sparse PCA**: replace $\|d^j\|_2 \leqslant 1$ by $\Theta(d^j) \leqslant 1$

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

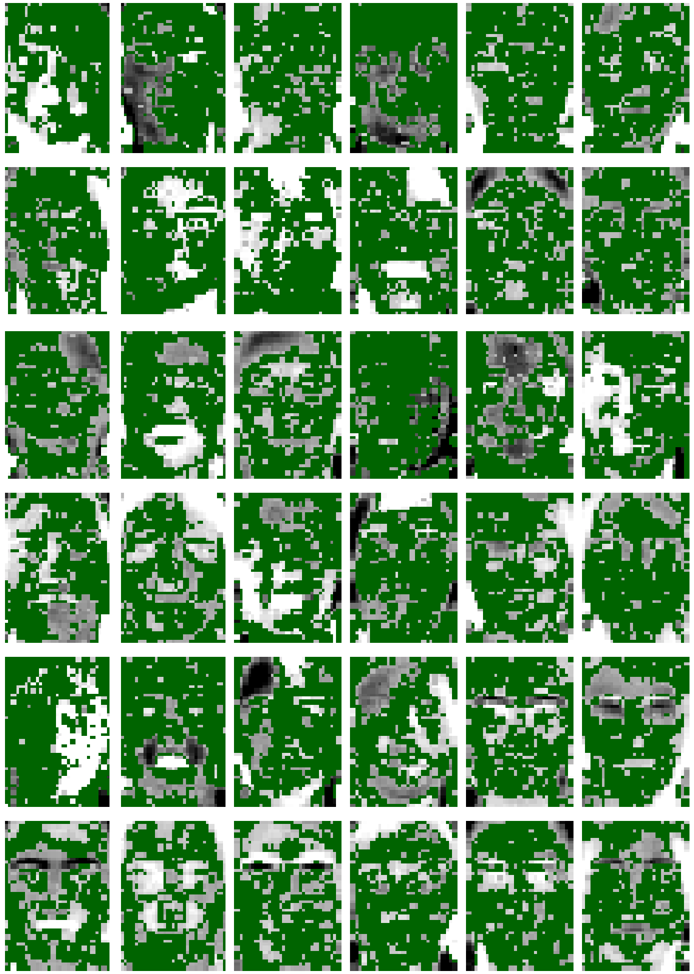


sparse PCA

- Unstructed 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 ⇒ many zeros do not lead to better interpretability

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

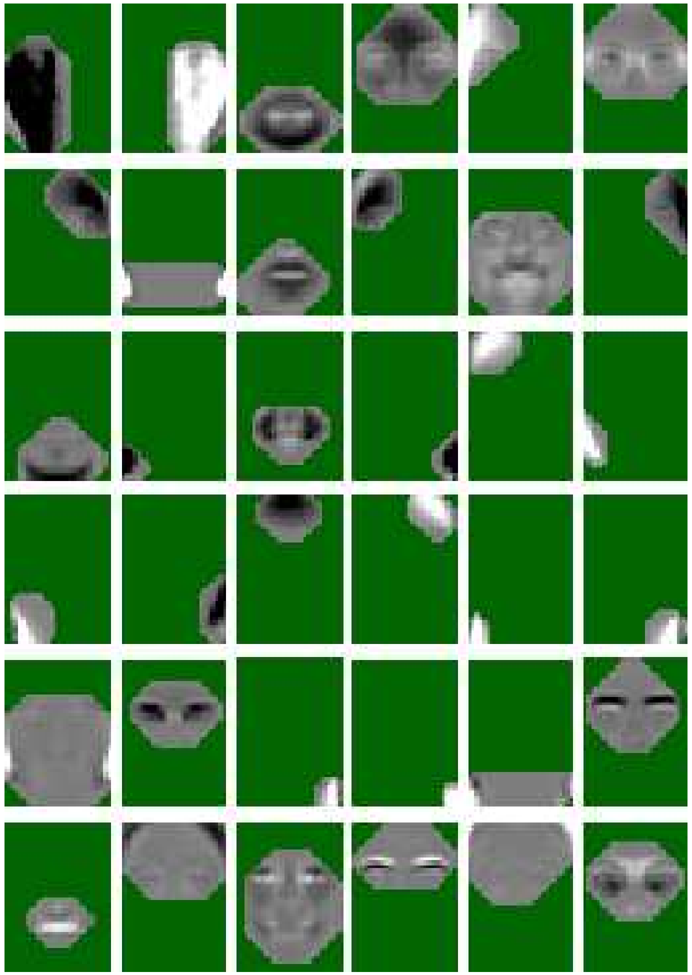
raw data

Structured sparse PCA

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

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

raw data



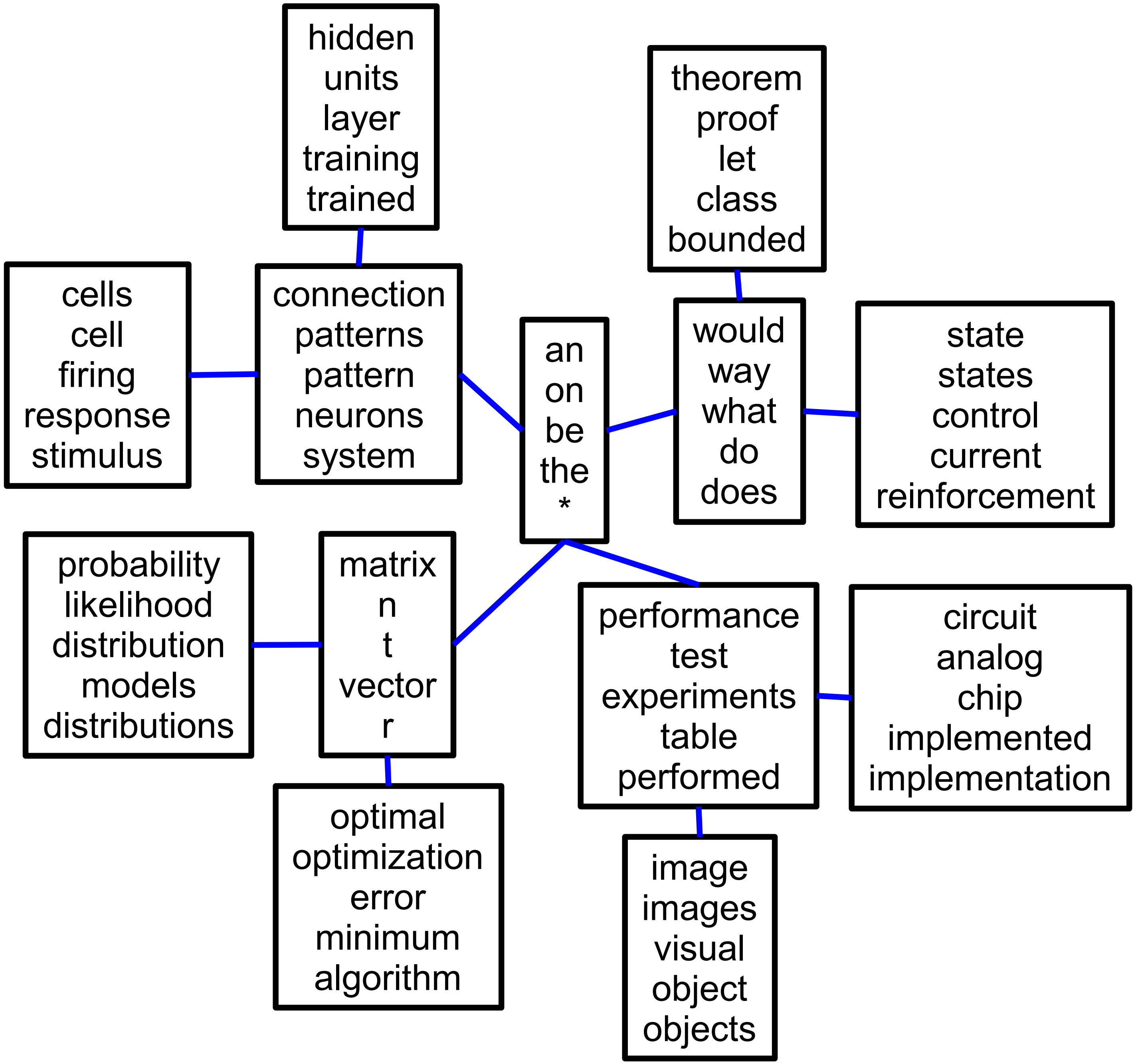
Structured sparse PCA

- Enforce selection of convex nonzero patterns $\Rightarrow$ 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)

Modelling of text corpora (Jenatton 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)

Why structured sparsity?

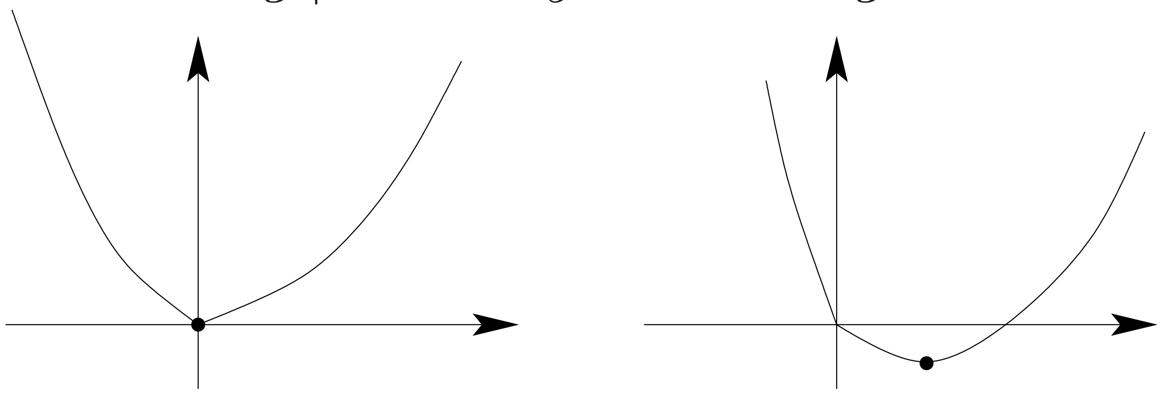
- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements "organized" in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)
- **Stability and identifiability**
- **Prediction or estimation performance**
  - When prior 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)

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

Why ℓ_1-norms lead to sparsity?

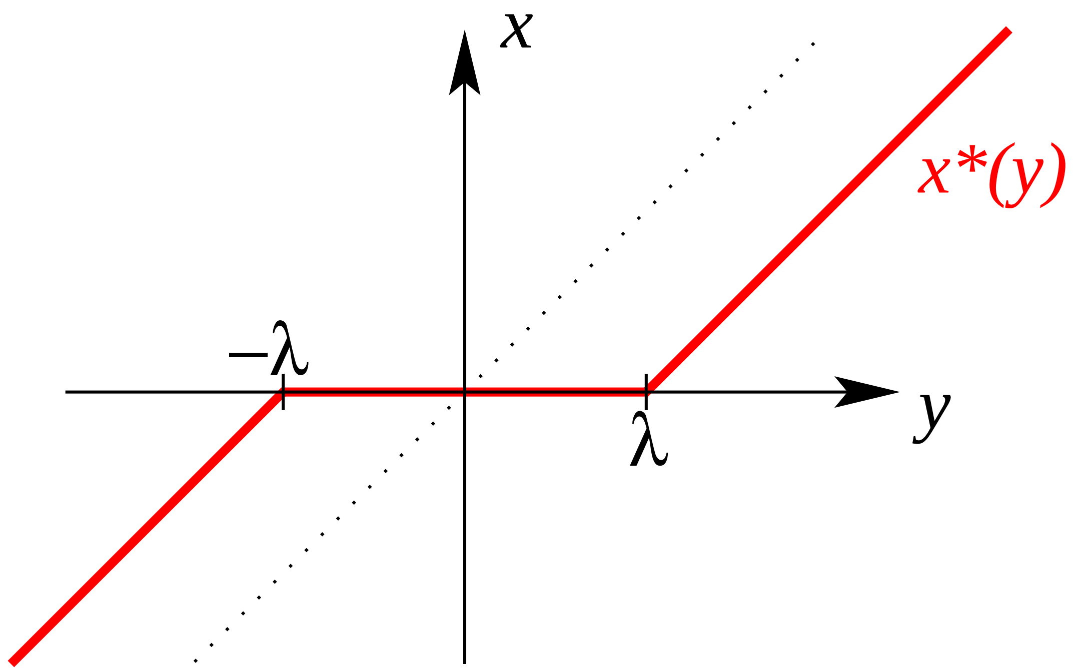
- **Example 1**: quadratic problem in 1D, i.e., $\boxed{\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
  - Derivative at $0+$: $g_+ = \lambda - y$ and $0-$: $g_- = -\lambda - y$



  - $x = 0$ is the solution iff $g_+ \geqslant 0$ and $g_- \leqslant 0$ (i.e., $|y| \leqslant \lambda$)
  - $x \geqslant 0$ is the solution iff $g_+ \leqslant 0$ (i.e., $y \geqslant \lambda$) $\Rightarrow x^* = y - \lambda$
  - $x \leqslant 0$ is the solution iff $g_- \leqslant 0$ (i.e., $y \leqslant -\lambda$) $\Rightarrow x^* = y + \lambda$
- Solution $\boxed{x^* = \mathrm{sign}(y)(|y| - \lambda)_+}$ = soft thresholding

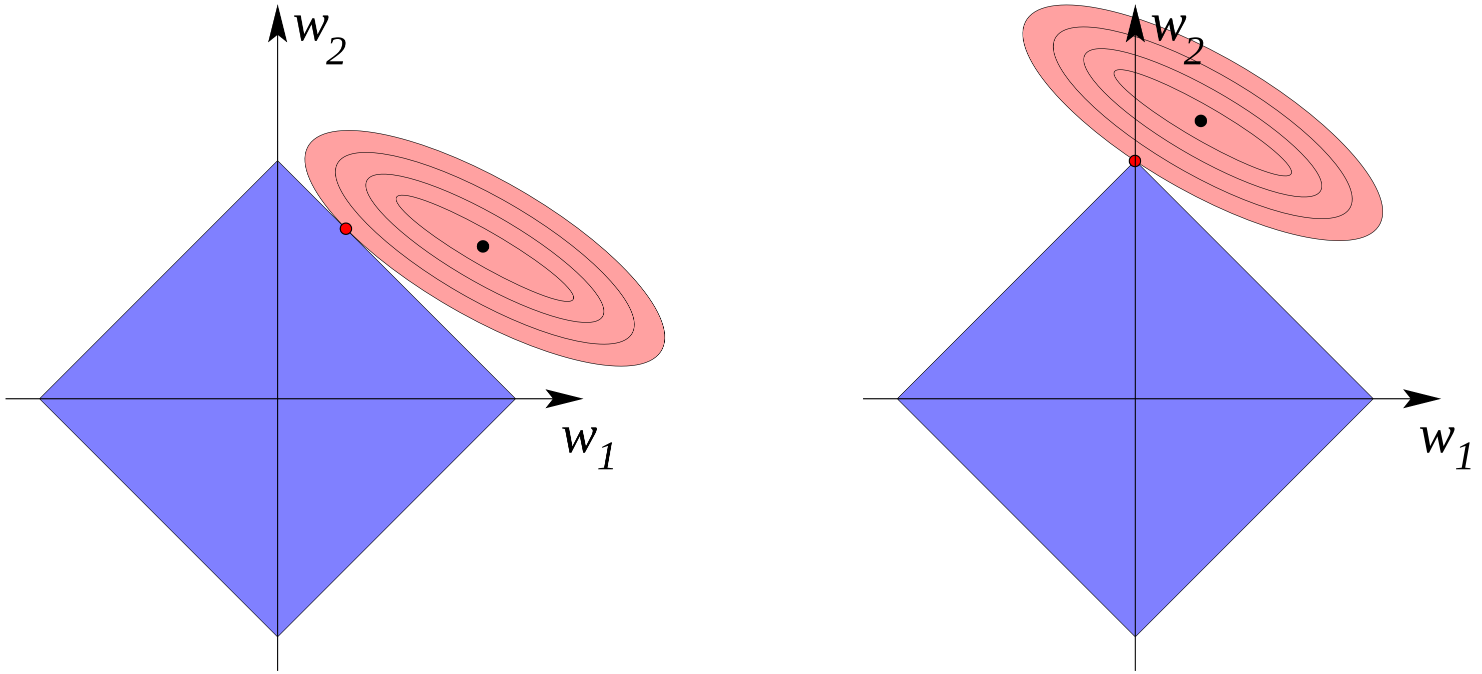
Why ℓ_1-norms lead to sparsity?

- **Example 1**: quadratic problem in 1D, i.e., $\boxed{\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
- Solution $\boxed{x^* = \text{sign}(y)(|y| - \lambda)_+}$ = soft thresholding



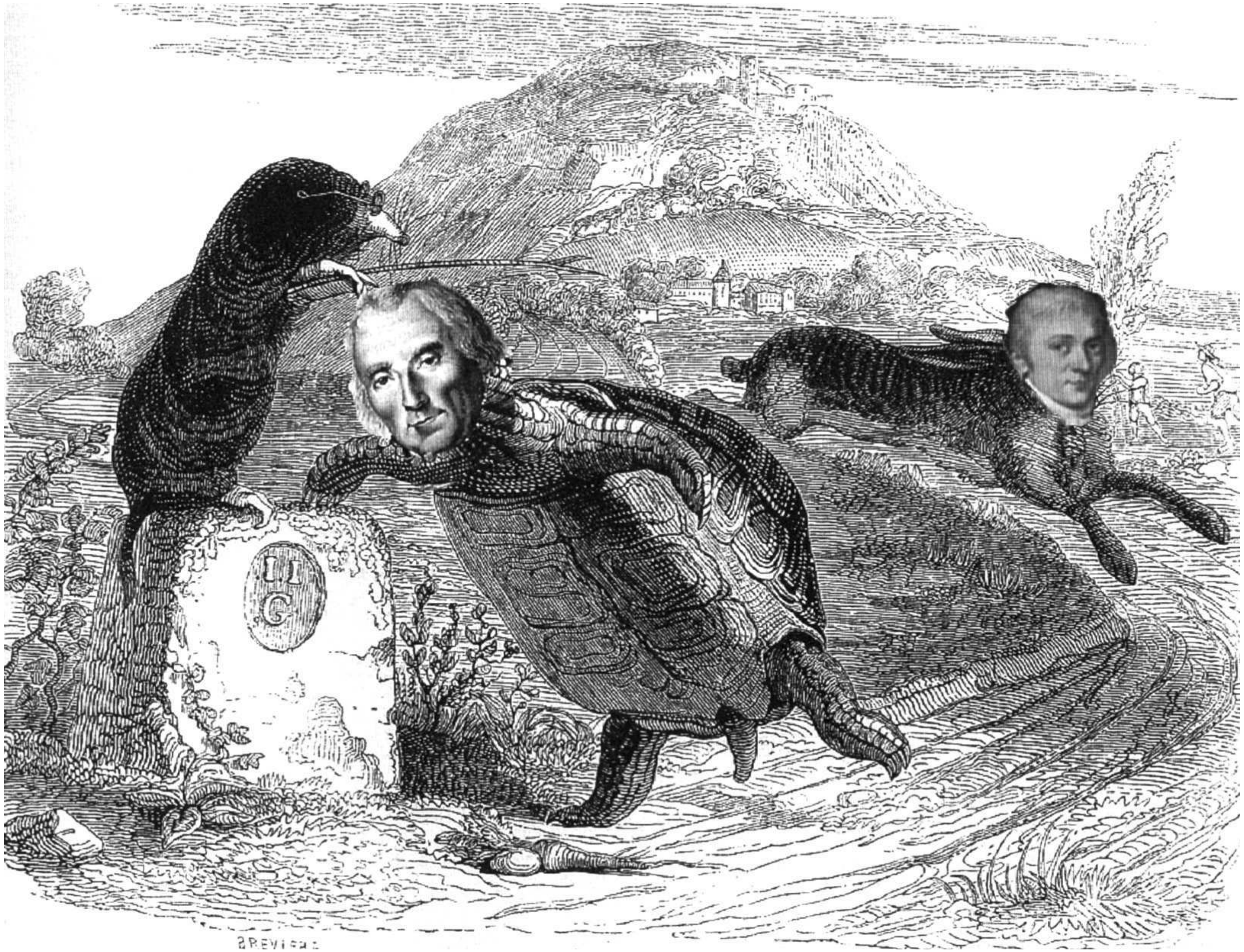
Why ℓ_1-norms lead to sparsity?

- **Example 2**: minimize quadratic function $Q(w)$ subject to $\|w\|_1 \leqslant T$.
  - coupled soft thresholding
- Geometric interpretation
  - NB : penalizing is "equivalent" to constraining



- **Non-smooth optimization!**

Gaussian hare (ℓ_2) vs. Laplacian tortoise (ℓ_1)



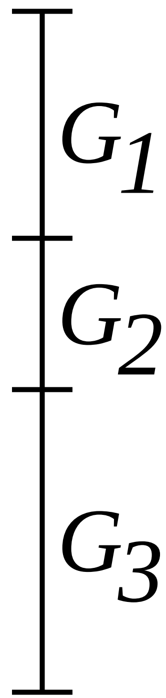
- Smooth vs. non-smooth optimization
- See Bach, Jenatton, Mairal, and Obozinski (2011)

Sparsity-inducing norms

- **Popular choice for Ω**
  - The ℓ_1-ℓ_2 norm,

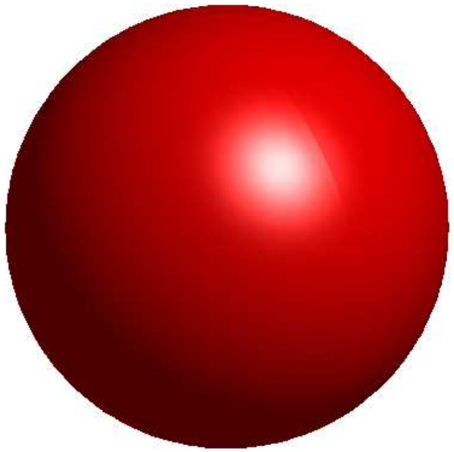
$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \Big(\sum_{j \in G} w_j^2 \Big)^{1/2}$$

  - with $\mathbf{H}$ a partition 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)

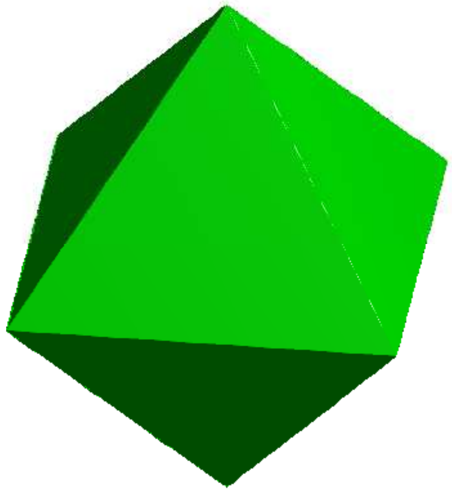


Unit norm balls

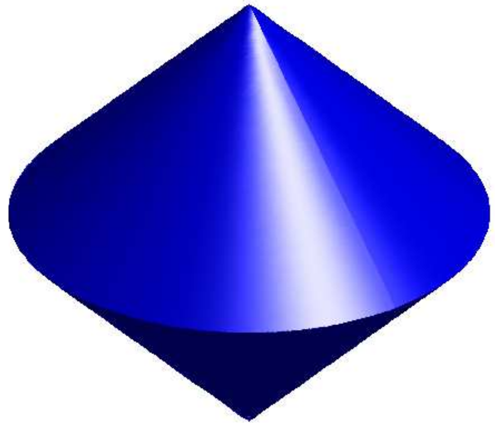
Geometric interpretation



$\|w\|_2$



$\|w\|_1$

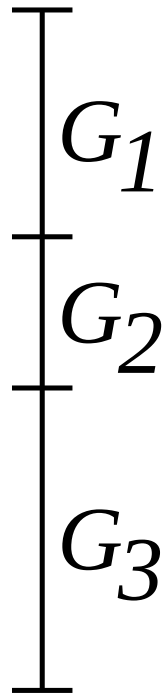


$\sqrt{w_1^2 + w_2^2} + |w_3|$

Sparsity-inducing norms

- **Popular choice for Ω**
  - The ℓ_1-ℓ_2 norm,

$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \Big(\sum_{j \in G} w_j^2 \Big)^{1/2}$$



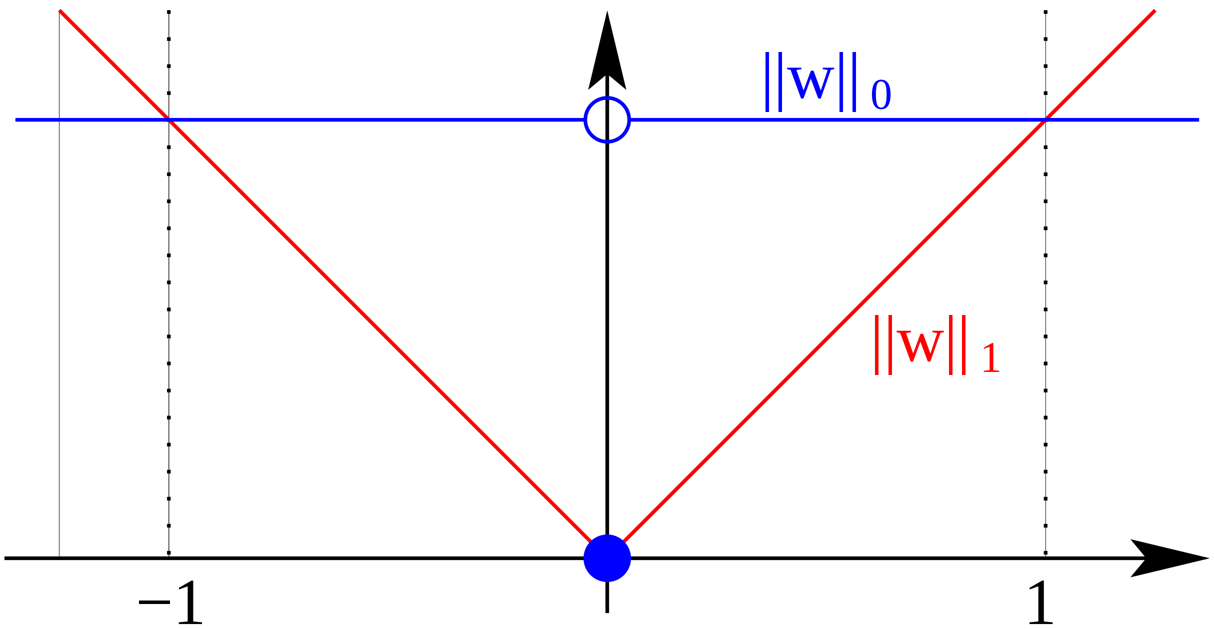
  - with $\mathbf{H}$ a partition of $\{1, \dots, p\}$
  - The ℓ_1-ℓ_2 norm sets to zero groups of non-overlapping variables (as opposed to single variables for the ℓ_1-norm)
  - For the square loss, group Lasso (Yuan and Lin, 2006)

- What if the set of groups $\mathbf{H}$ is not a partition anymore?

- **Is there any systematic way?**

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 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

Proof - I

- Notation: $g : w \mapsto F(\text{supp}(w))$ defined on $[-1, 1]^p$
- Computation of the Fenchel dual

$$
\begin{aligned}
g^*(s) &= \max_{\|w\|_\infty \leqslant 1} w^\top s - g(w) \\
&= \max_{\delta \in \{0,1\}^p} \max_{\|w\|_\infty \leqslant 1} (\delta \circ w)^\top s - f(\delta) \text{ by definition of } g \\
&= \max_{\delta \in \{0,1\}^p} \delta^\top |s| - f(\delta) \text{ by maximizing out } w \\
&= \max_{\delta \in [0,1]^p} \delta^\top |s| - f(\delta) \text{ because } F - |s| \text{ is submodular}
\end{aligned}
$$

Proof - II

- Notation: $g : w \mapsto F(\mathrm{supp}(w))$ defined on $[-1,1]^p$
- Fenchel dual: $g^*(s) = \max_{\delta \in [0,1]^p} \delta^\top |s| - f(\delta)$

Proof - II

- Notation: $g : w \mapsto F(\mathrm{supp}(w))$ defined on $[-1,1]^p$
- Fenchel dual: $g^*(s) = \max_{\delta \in [0,1]^p} \delta^\top |s| - f(\delta)$
- Computation of the Fenchel bi-dual, for all w such that $\|w\|_\infty \leqslant 1$:

$$
\begin{aligned}
g^{**}(w) &= \max_{s \in \mathbb{R}^p} s^\top w - g^*(s) \\
&= \max_{s \in \mathbb{R}^p} \min_{\delta \in [0,1]^p} s^\top w - \delta^\top |s| + f(\delta) \\
&= \min_{\delta \in [0,1]^p} \max_{s \in \mathbb{R}^p} s^\top w - \delta^\top |s| + f(\delta) \text{ by strong duality} \\
&= \min_{\delta \in [0,1]^p, \delta \geqslant |w|} f(\delta) = f(|w|) \text{ because } F \text{ is nonincreasing}
\end{aligned}
$$

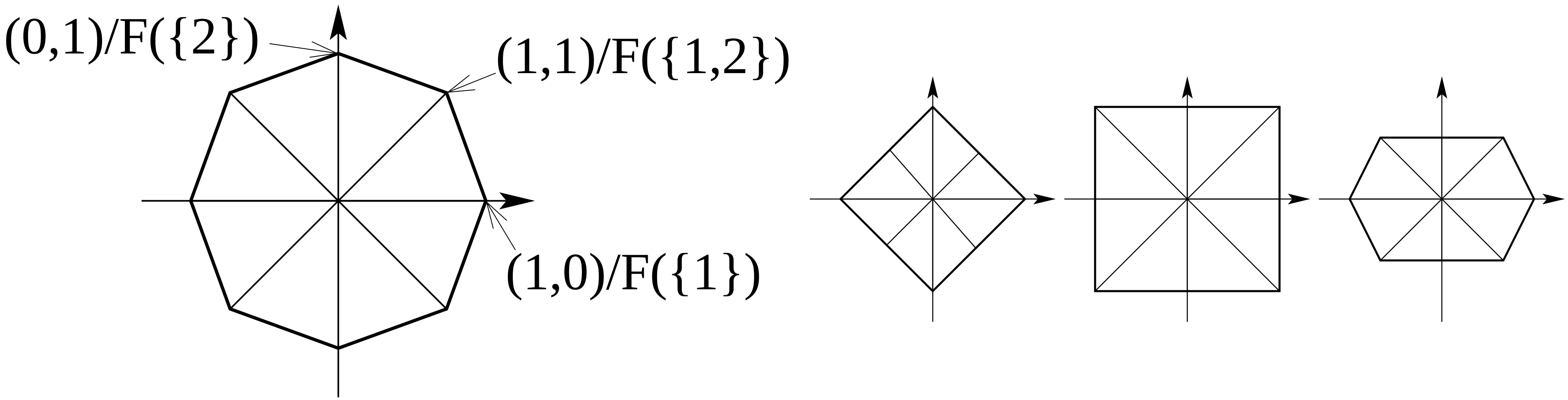
Submodular functions and structured sparsity

- 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

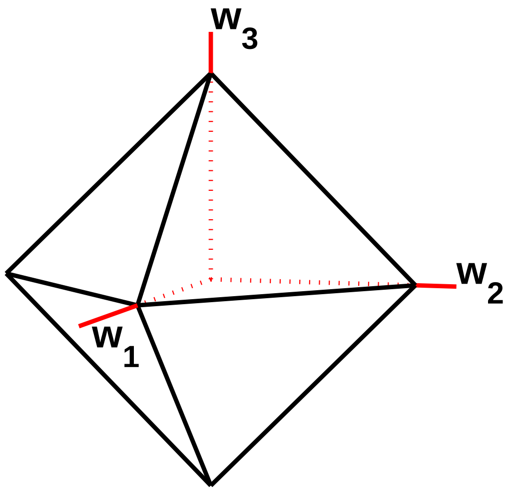
Submodular functions and structured sparsity

- Let $F : 2^V \to \mathbb{R}$ be a **non-decreasing submodular set-function**
- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\mathrm{Supp}(w))$ 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

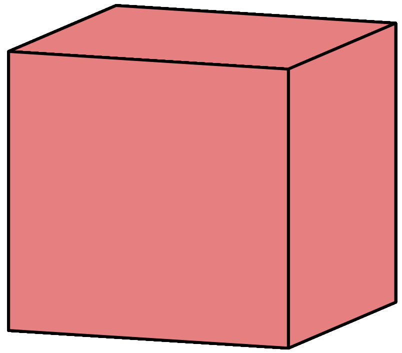


  - A if stable if for all $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$
  - With probability one, stable sets are the only allowed active sets

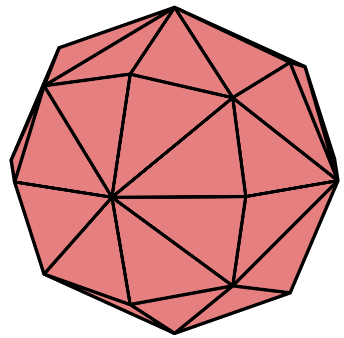
Polyhedral unit balls



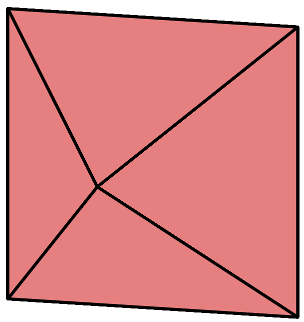
$$F(A) = |A|$$
$$\Omega(w) = \|w\|_1$$



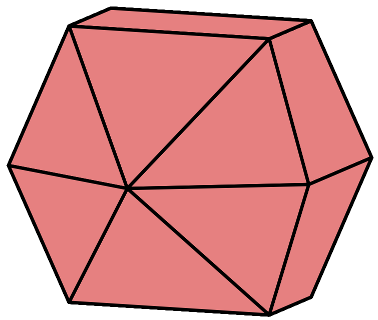
$$F(A) = \min\{|A|, 1\}$$
$$\Omega(w) = \|w\|_\infty$$



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

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

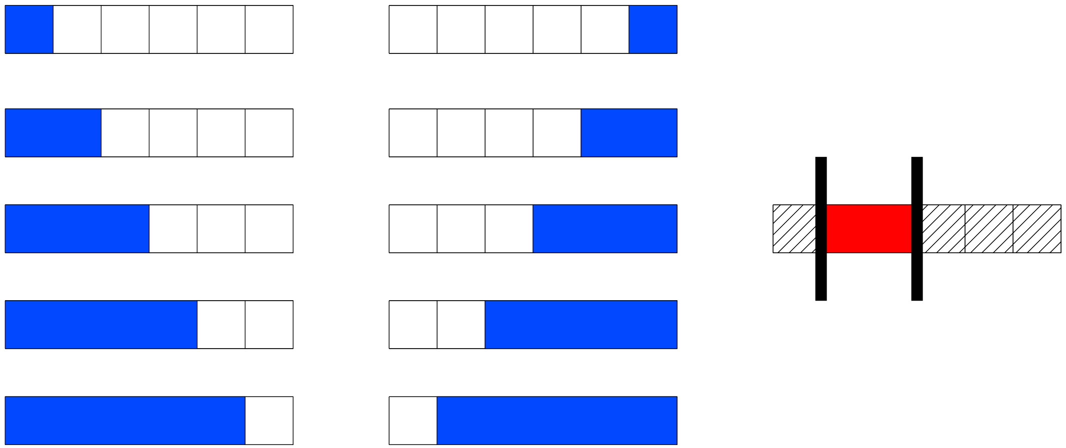
  - ℓ_1-ℓ_∞ norm $\Rightarrow$ sparsity at the group level
  - Some w_G's are set to zero for some groups G

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

  - Justification not only limited to allowed sparsity patterns

Selection of contiguous patterns in a sequence

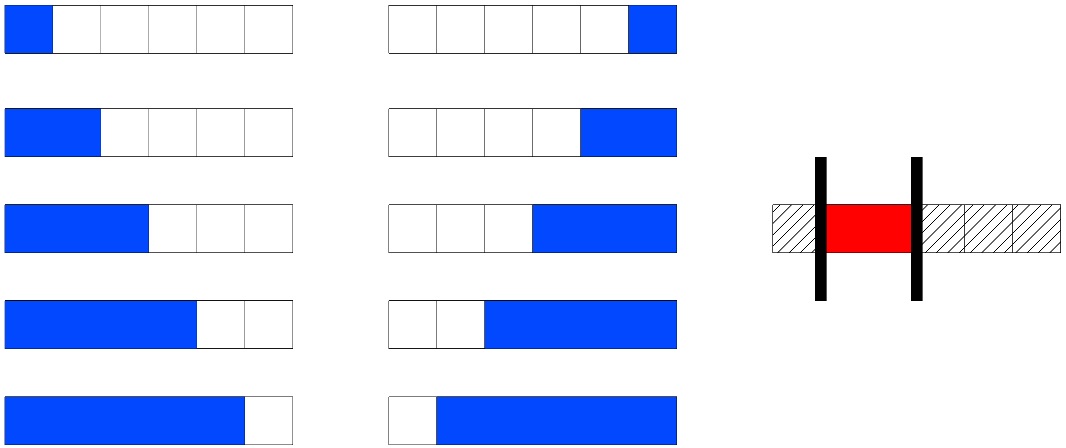
- Selection of contiguous patterns in a sequence



- $\mathbf{H}$ 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

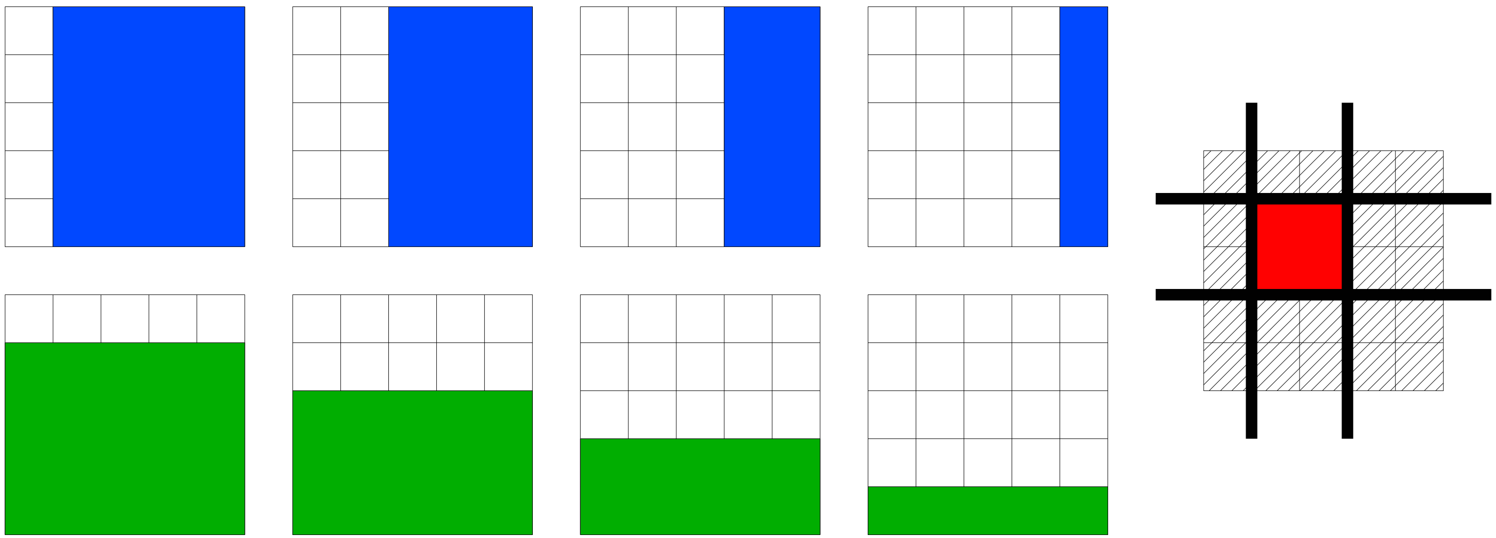


- $\mathbf{H}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a **contiguous pattern**

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

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

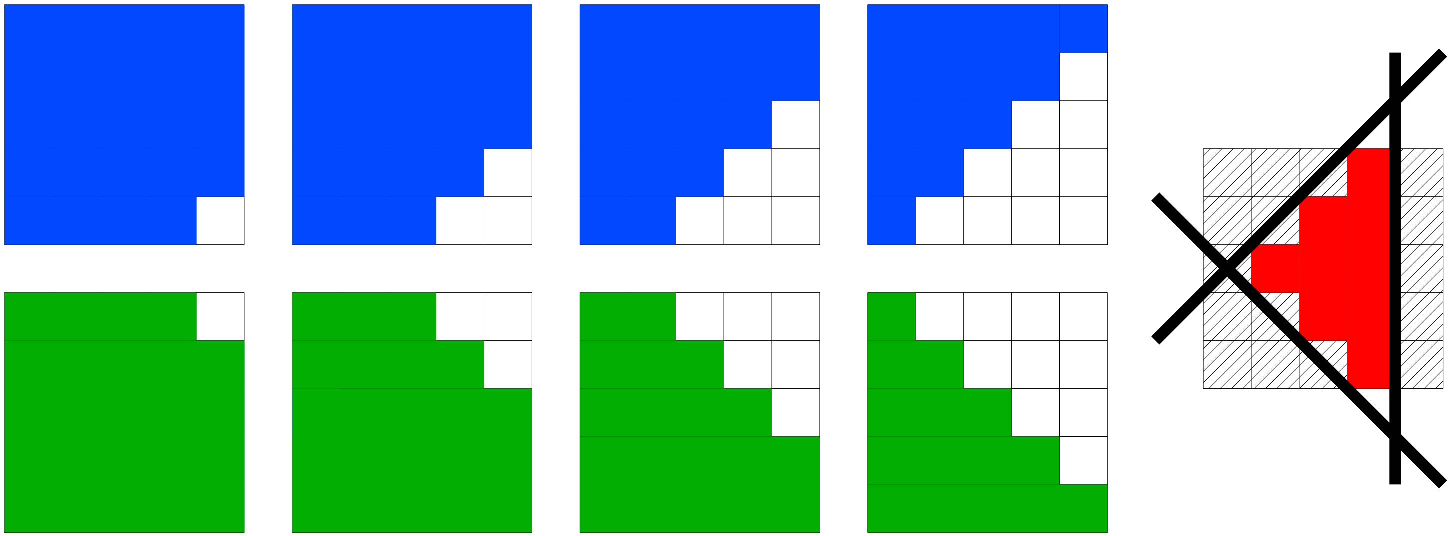
- Selection of rectangles on a 2-D grids, $p = 25$



  - $\mathbf{H}$ 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 $\mathbb{H}$

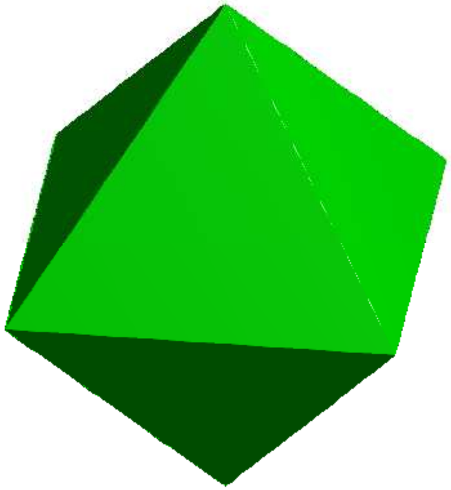
- Selection of diamond-shaped patterns on a 2-D grids, $p = 25$.



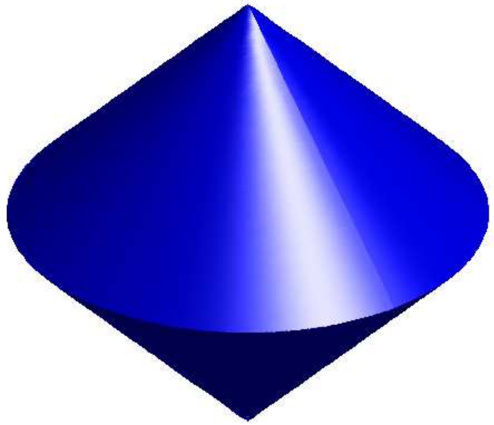
  - It is possible to extend such settings to 3-D space, or more complex topologies

Unit norm balls

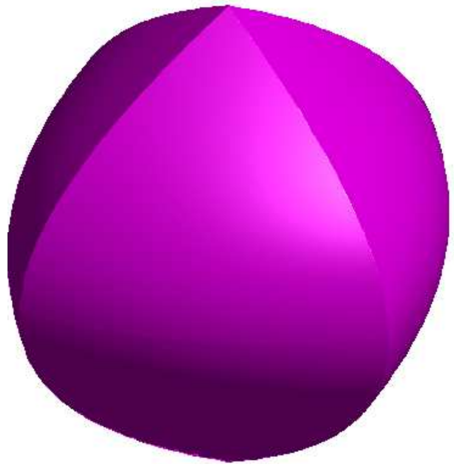
Geometric interpretation



$\|w\|_1$



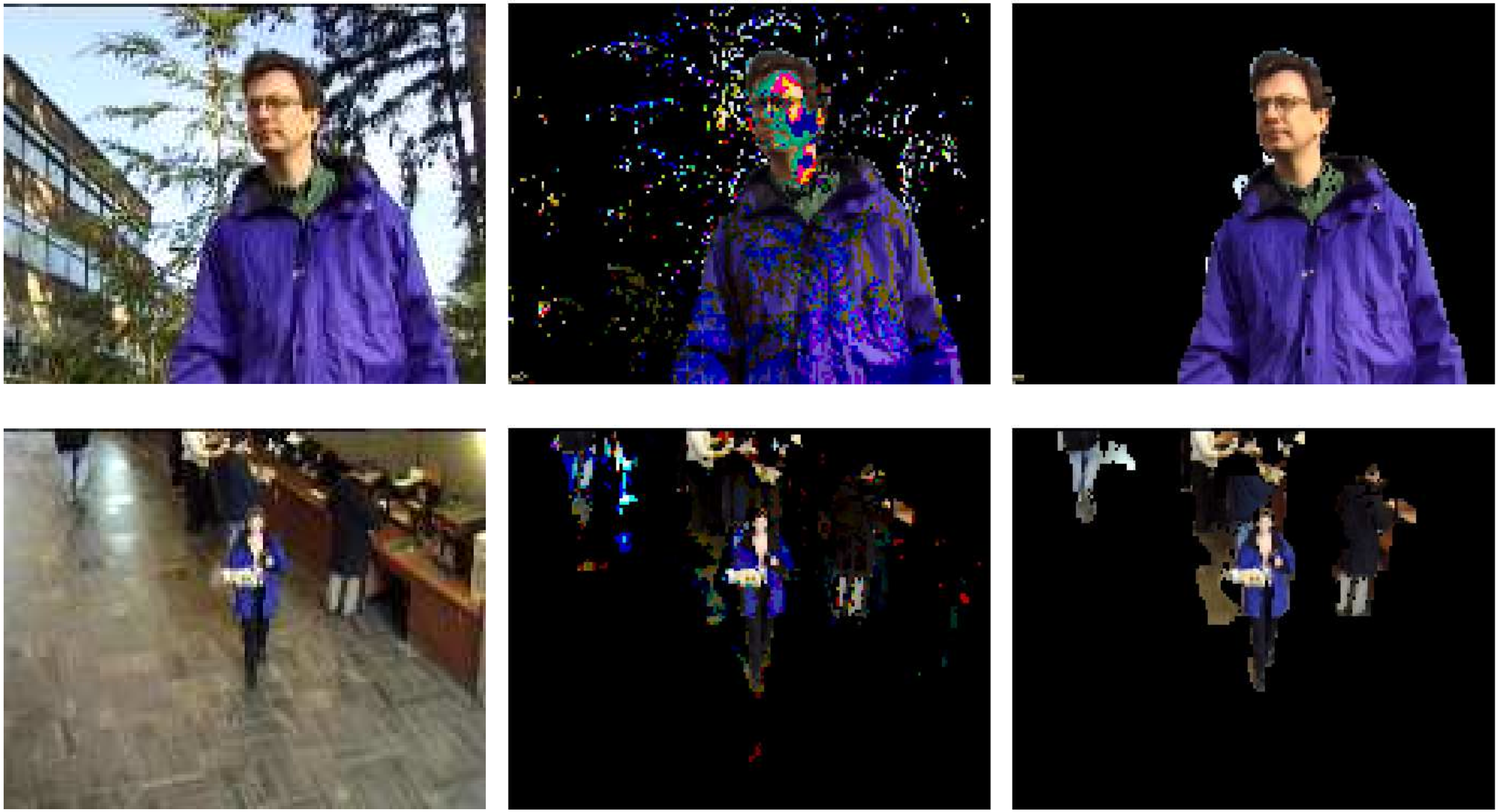
$\sqrt{w_1^2 + w_2^2} + |w_3|$



$\|w\|_2 + |w_1| + |w_2|$

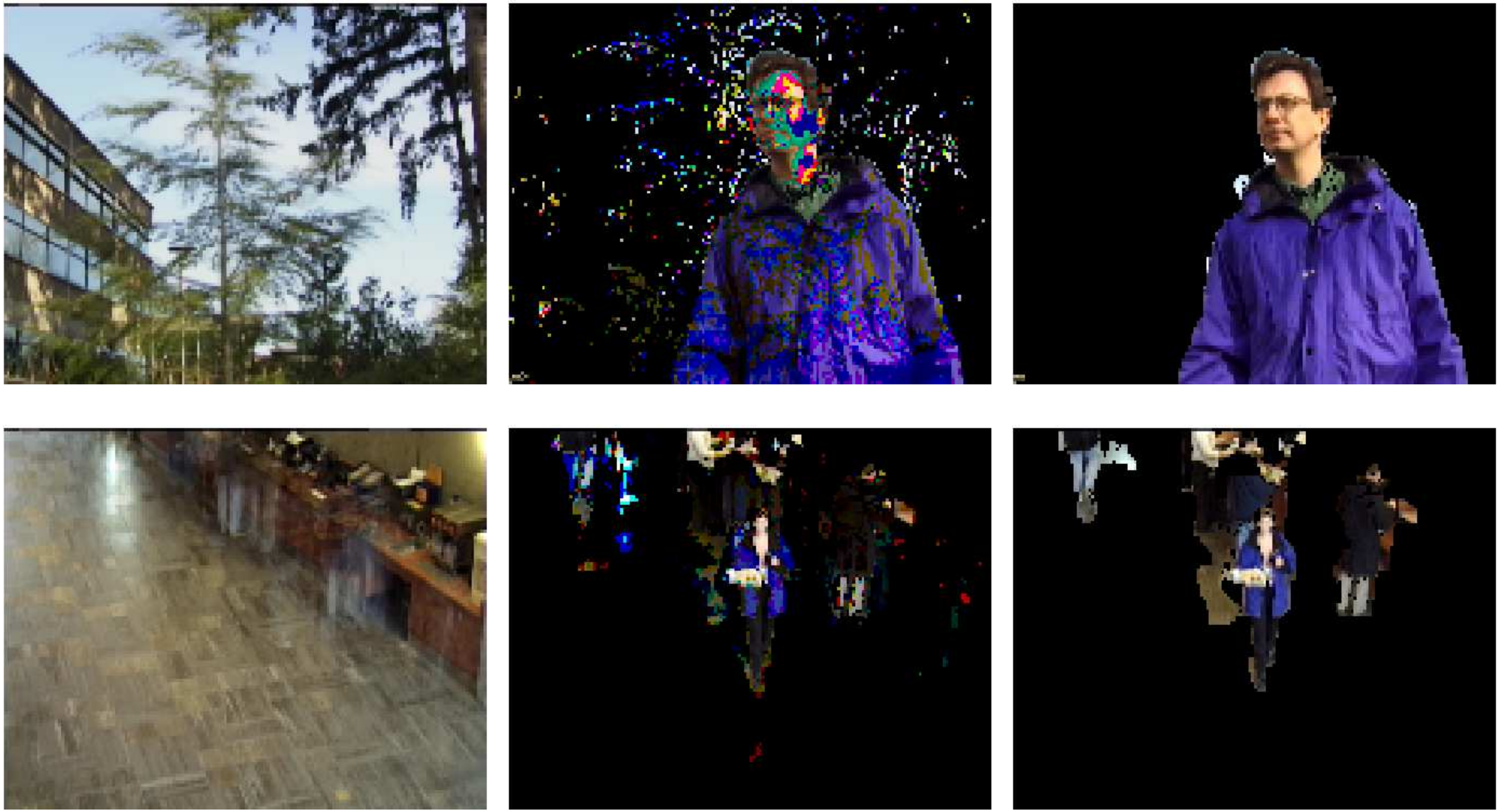
Application to background subtraction (Mairal, Jenatton, Obozinski, and Bach, 2010)

Input | ℓ_1-norm | Structured norm



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

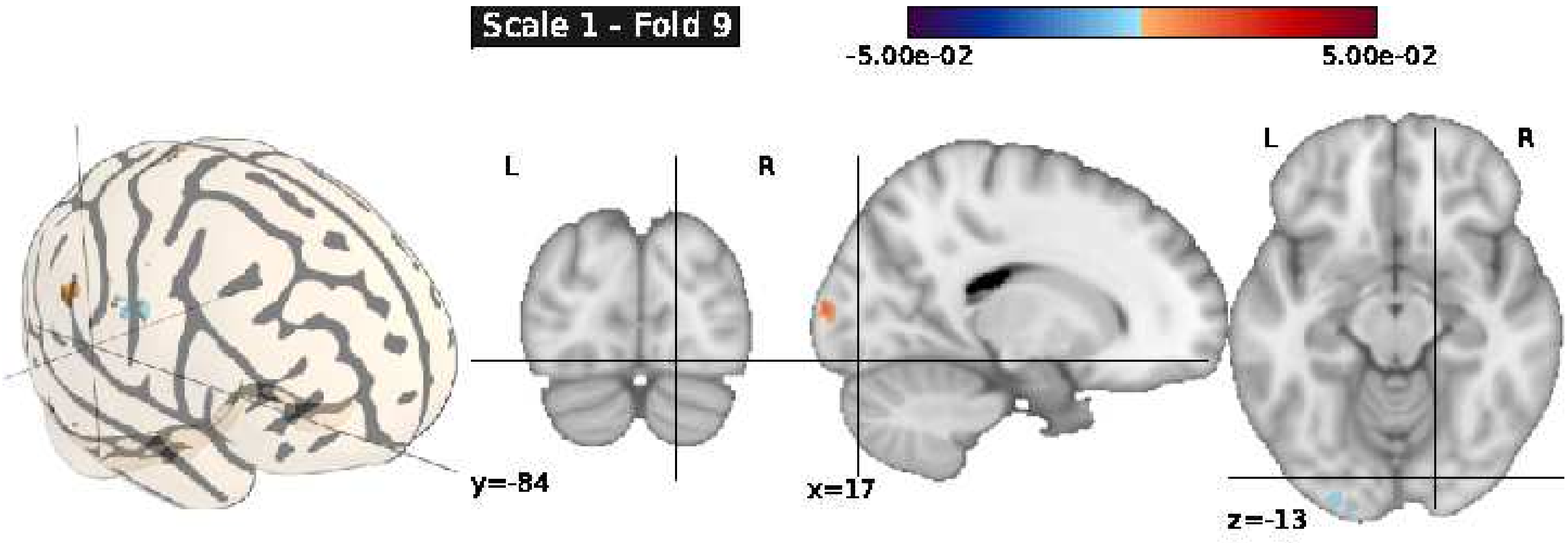
Background ℓ_1-norm Structured norm



Application to neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

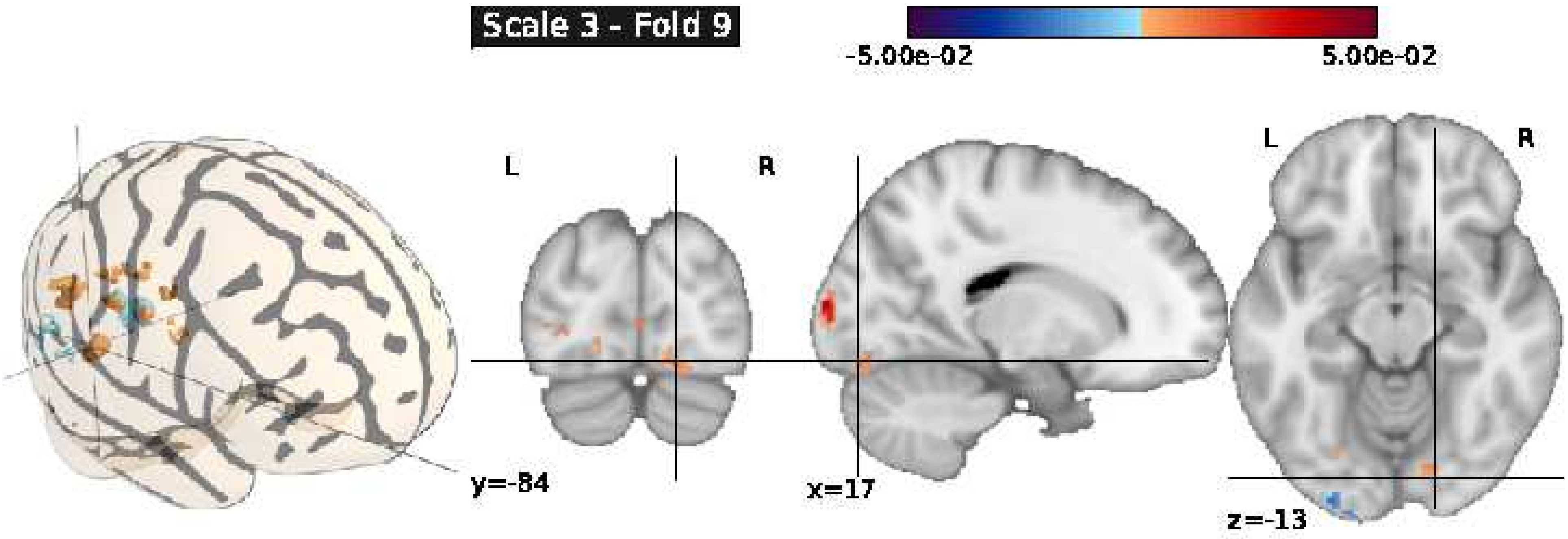
- "Brain reading": prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



Application to neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

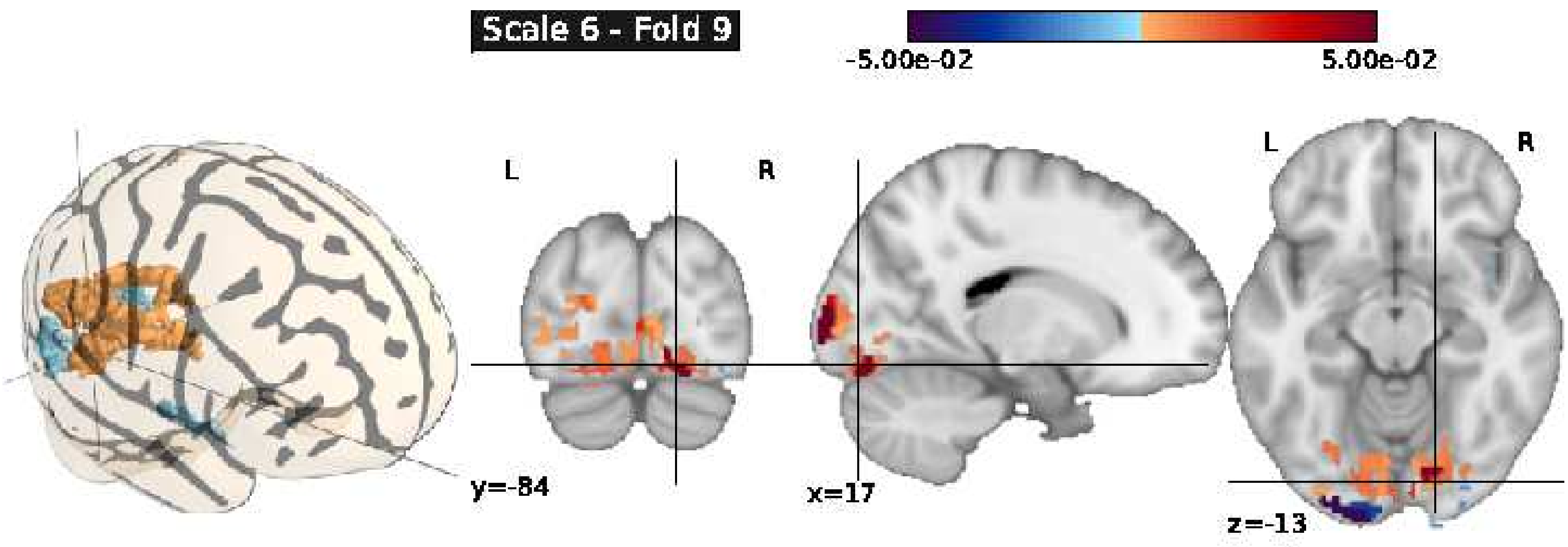
- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



Application to neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization

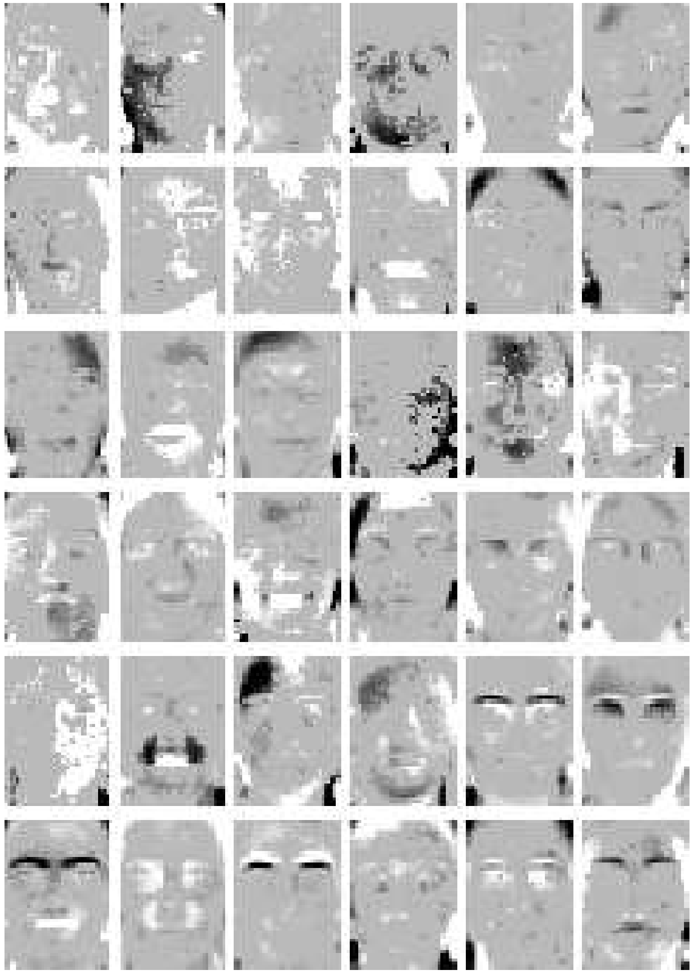


Sparse Structured PCA (Jenatton, Obozinski, and Bach, 2009b)

- 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 (2/3)

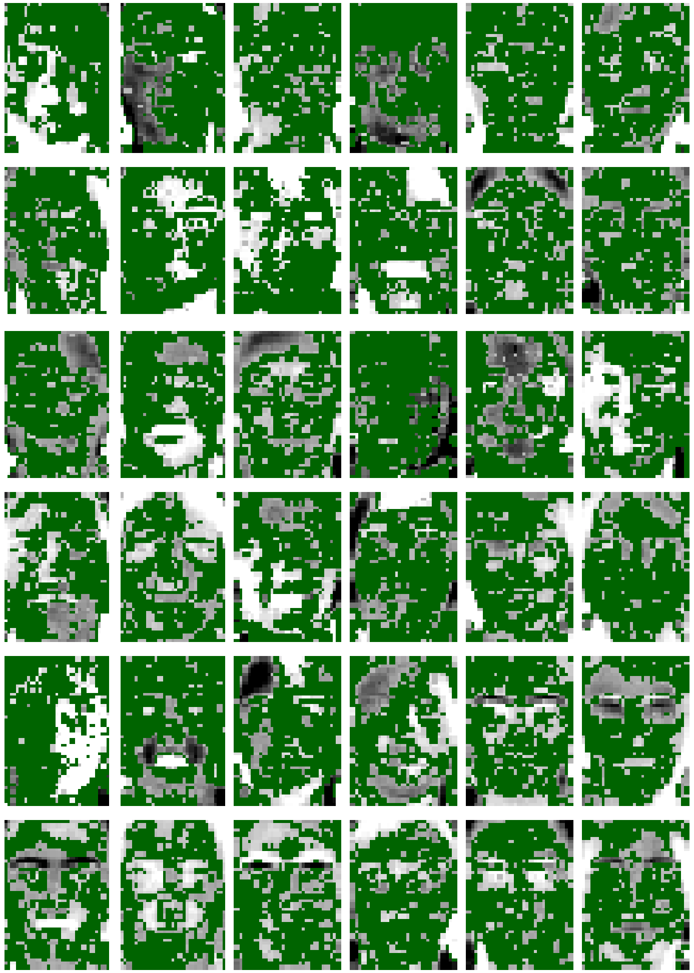


(unstructured) sparse PCA

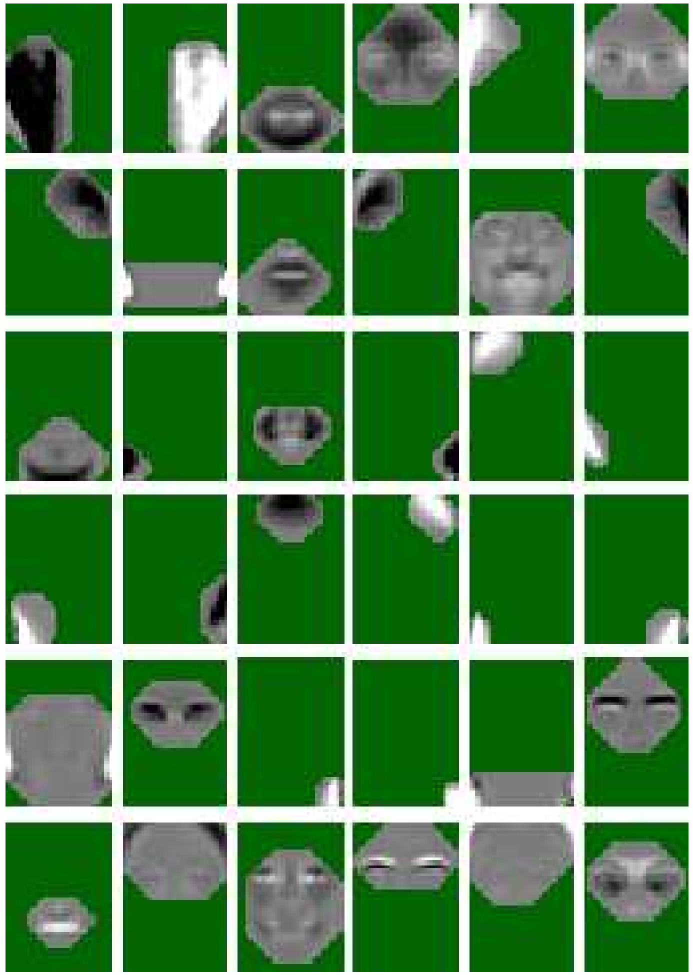
Structured sparse PCA

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

Application to face databases (2/3)



(unstructured) sparse PCA

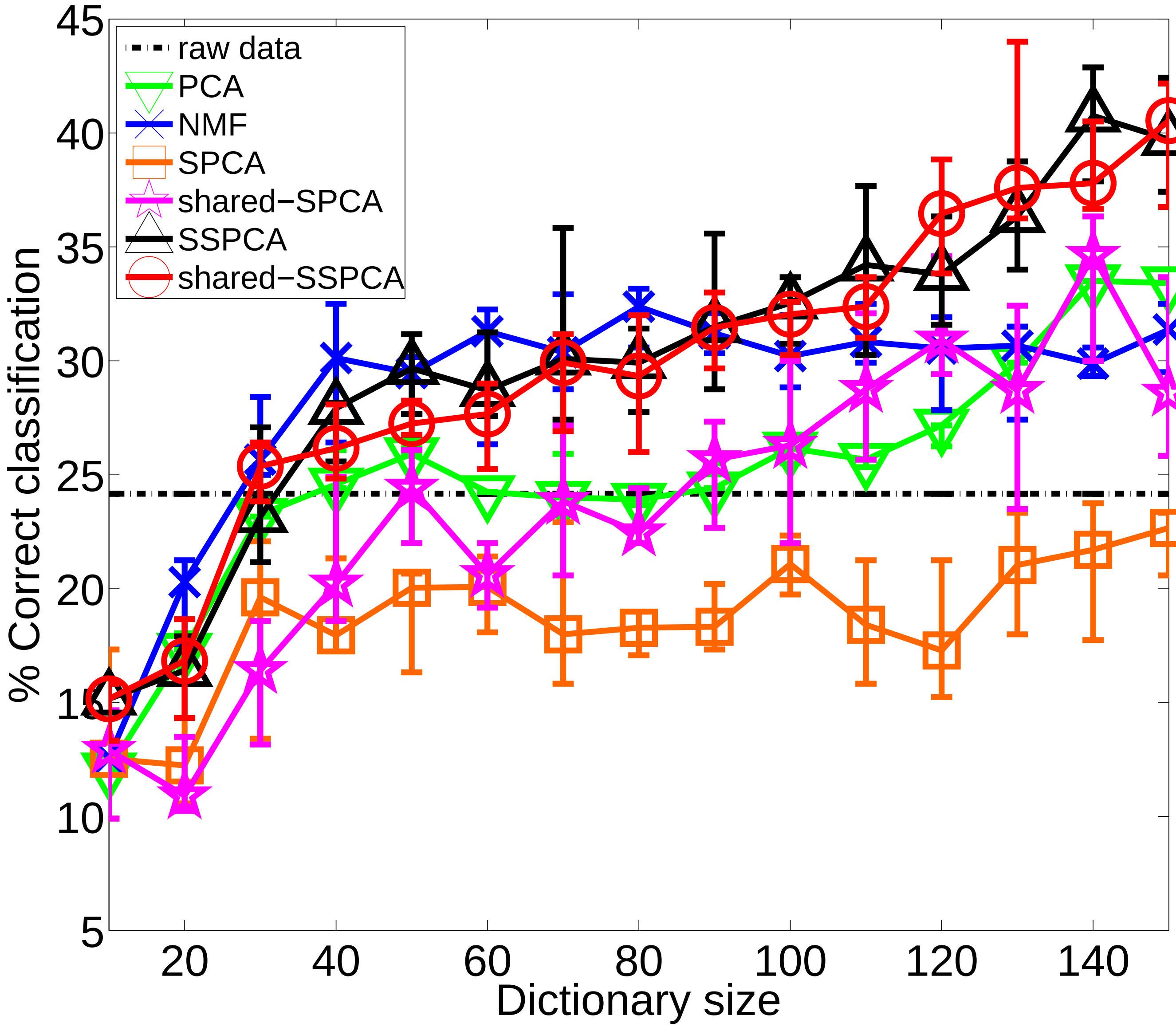


Structured sparse PCA

- 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**):

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

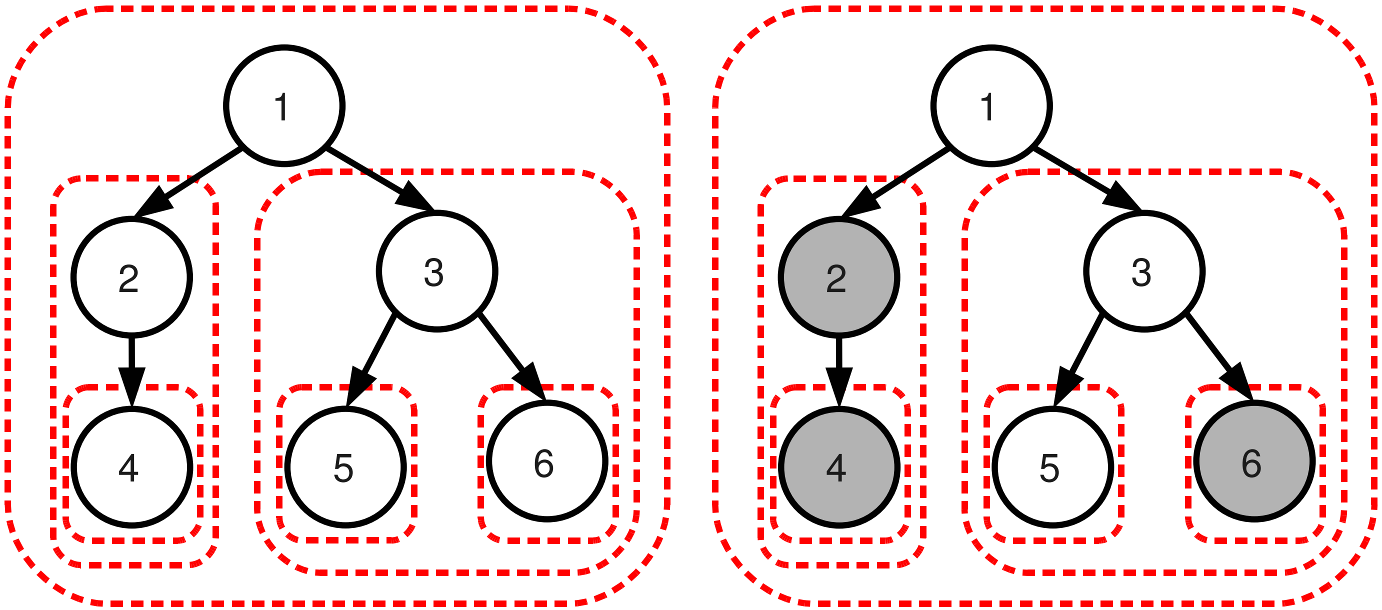
- Dictionary learning with **structured sparsity for codes** w:

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

- **Optimization**: proximal methods
  - Requires solving many times $\min_{w\in\mathbb{R}^p}\frac{1}{2}\|y-w\|_2^2 + \lambda\Omega(w)$
  - **Modularity of implementation** if proximal step is efficient (Jenatton et al., 2010; Mairal et al., 2010)

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

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

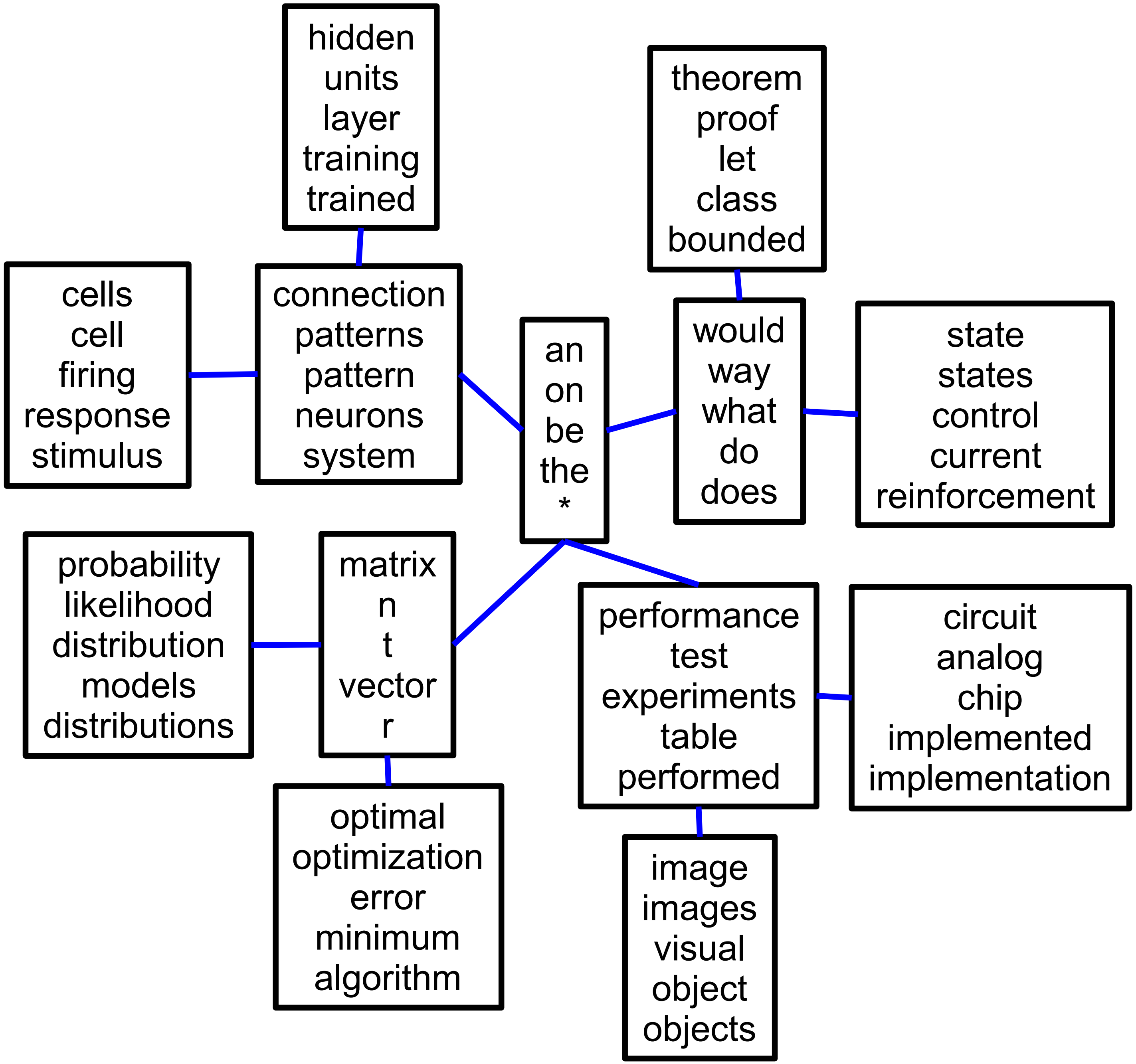


- Variable selected after its ancestors (Zhao et al., 2009; Bach, 2008)

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



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

Unified optimization algorithms

- **Polyhedral norm** with $O(3^p)$ faces and 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 $O(3^p)$ faces and 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** (e.g., Beck and Teboulle, 2009)
  - $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda\Omega(w)$: differentiable + non-differentiable
  - Efficient when (P) : $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - v\|_2^2 + \lambda\Omega(w)$ is "easy"
  - **Fact**: (P) is equivalent to submodular function minimization

Optimization for sparsity-inducing norms
(see Bach, Jenatton, Mairal, and Obozinski, 2011)

- Gradient descent as a **proximal method** (differentiable functions)
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{B}{2}\|w - w_t\|_2^2$
  - $w_{t+1} = w_t - \frac{1}{B}\nabla L(w_t)$

Optimization for sparsity-inducing norms (see Bach, Jenatton, Mairal, and Obozinski, 2011)

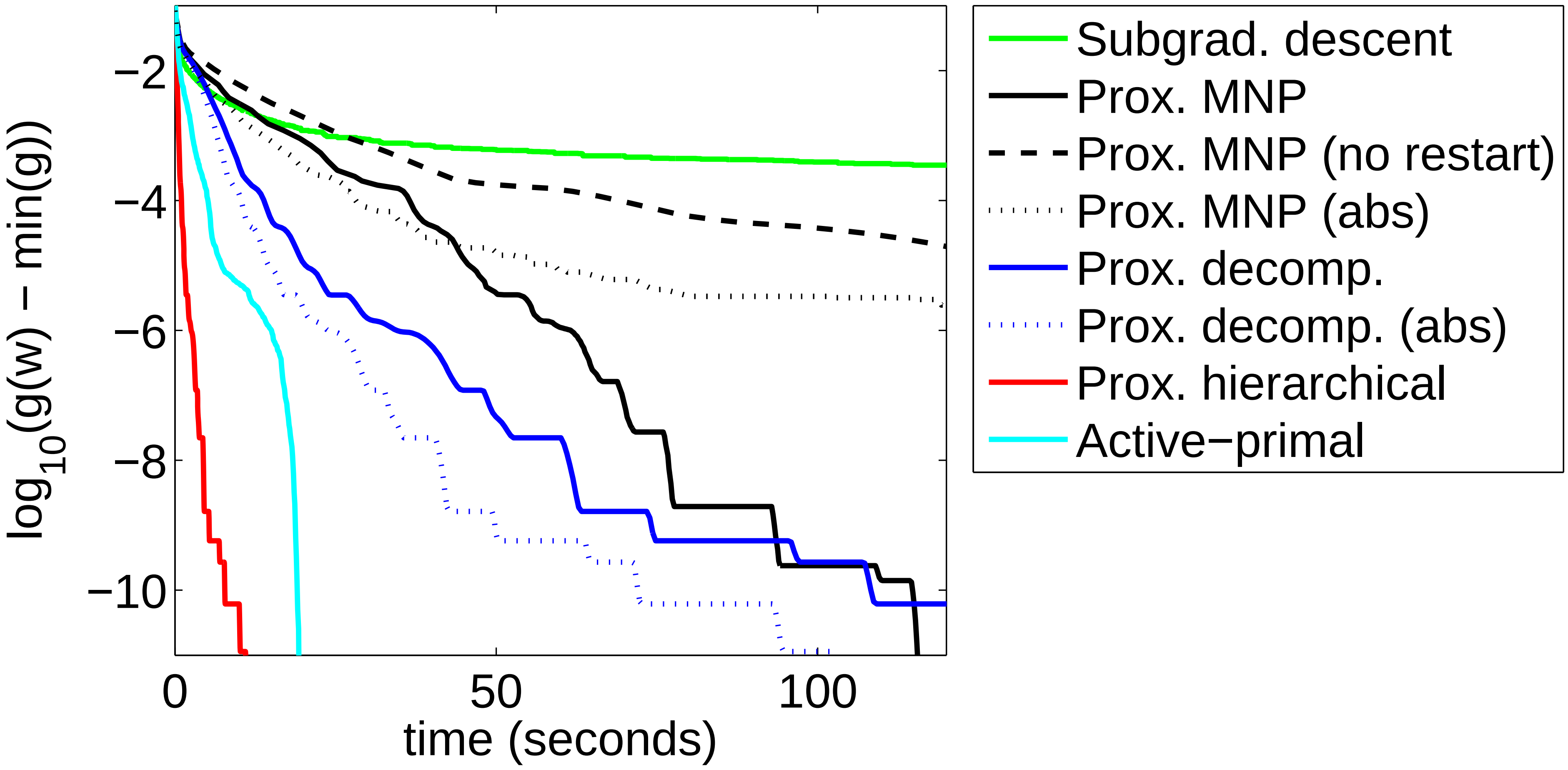
- Gradient descent as a **proximal method** (differentiable functions)
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{B}{2} \|w - w_t\|_2^2$
  - $w_{t+1} = w_t - \frac{1}{B} \nabla L(w_t)$
- Problems of the form: $\boxed{\min_{w \in \mathbb{R}^p} L(w) + \lambda \Omega(w)}$
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \lambda \Omega(w) + \frac{B}{2} \|w - w_t\|_2^2$
  - $\Omega(w) = \|w\|_1 \Rightarrow$ **Thresholded gradient descent**
- Similar convergence rates than smooth optimization
  - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)

Unified optimization algorithms

- **Polyhedral norm** with $O(3^p)$ faces and 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** (e.g., Beck and Teboulle, 2009)
  - $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda\Omega(w)$: differentiable + non-differentiable
  - Efficient when (P) : $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - v\|_2^2 + \lambda\Omega(w)$ is "easy"
  - **Fact**: (P) is equivalent to submodular function minimization
- **Active-set methods**

Comparison of optimization algorithms

- Tree-based regularization ($p = 511$)
- See Bach et al. (2011) for larger-scale problems



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 $P(\Omega^*(z) > t) \leqslant \sum_{A \in \mathcal{T}} 2^{|A|} \exp\big(- \frac{t^2 F(A)^2/2}{1^\top Q_{AA} 1}\big)$

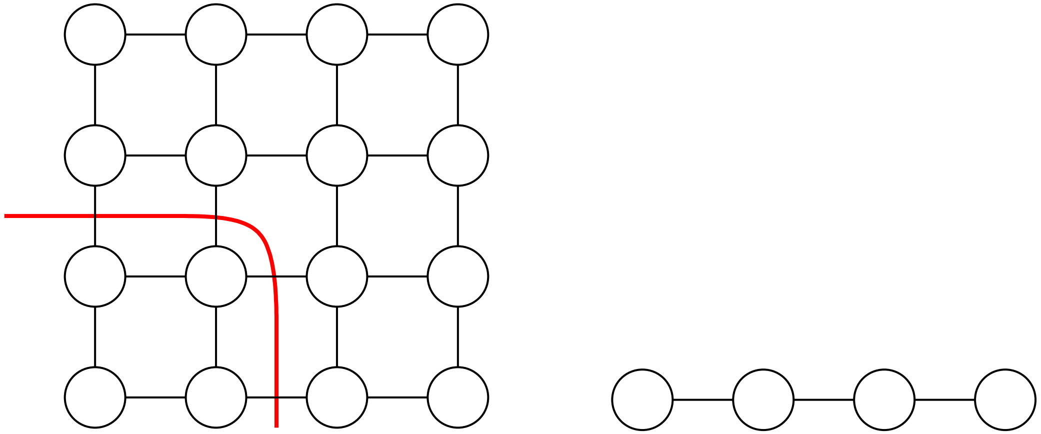
Symmetric submodular functions (Bach, 2011)

- 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 $[0,1]^p + \mathbb{R}1_V = \{w \in \mathbb{R}^p,\ \max_{k \in V} w_k - \min_{k \in V} w_k \leqslant 1\}$.
- **Shaping all level sets**

Symmetric submodular functions - Examples

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

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

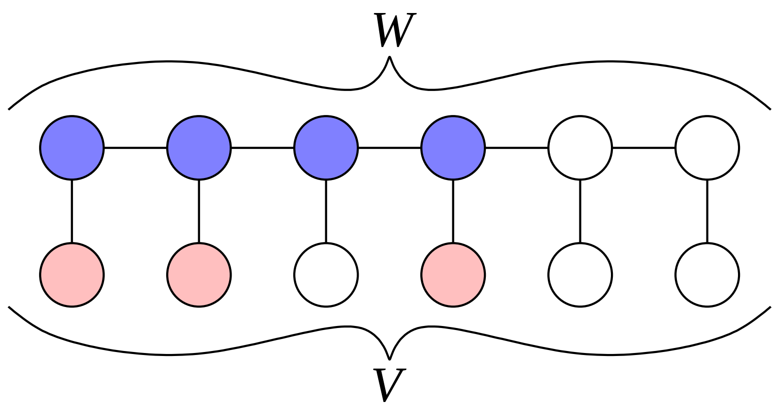


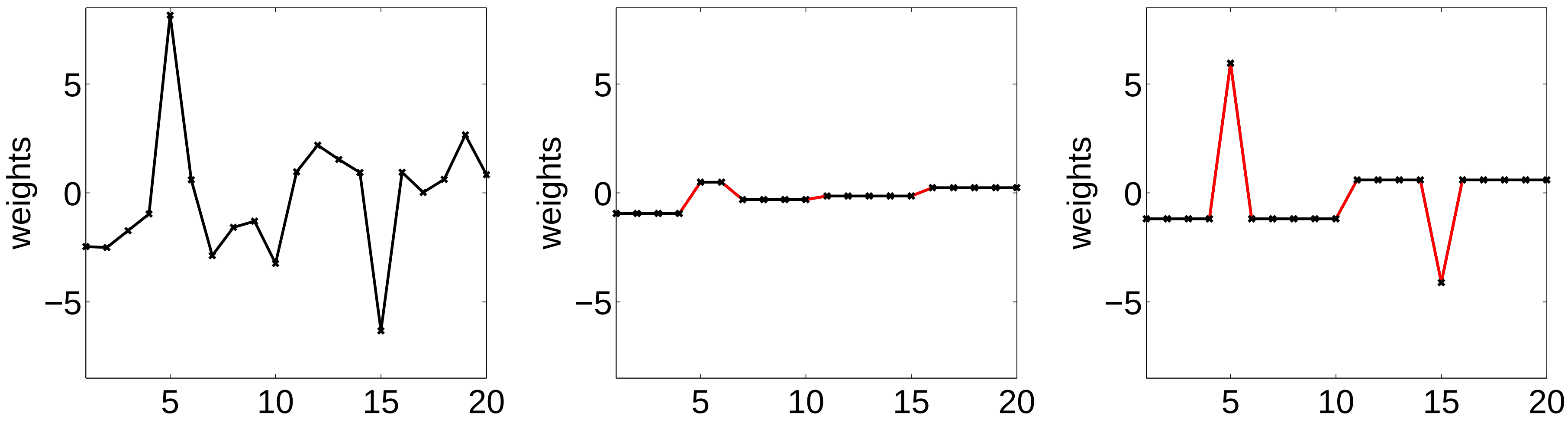
  - NB: graph may be directed
  - Application to change-point detection (Tibshirani et al., 2005; Harchaoui and Lévy-Leduc, 2008)

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)

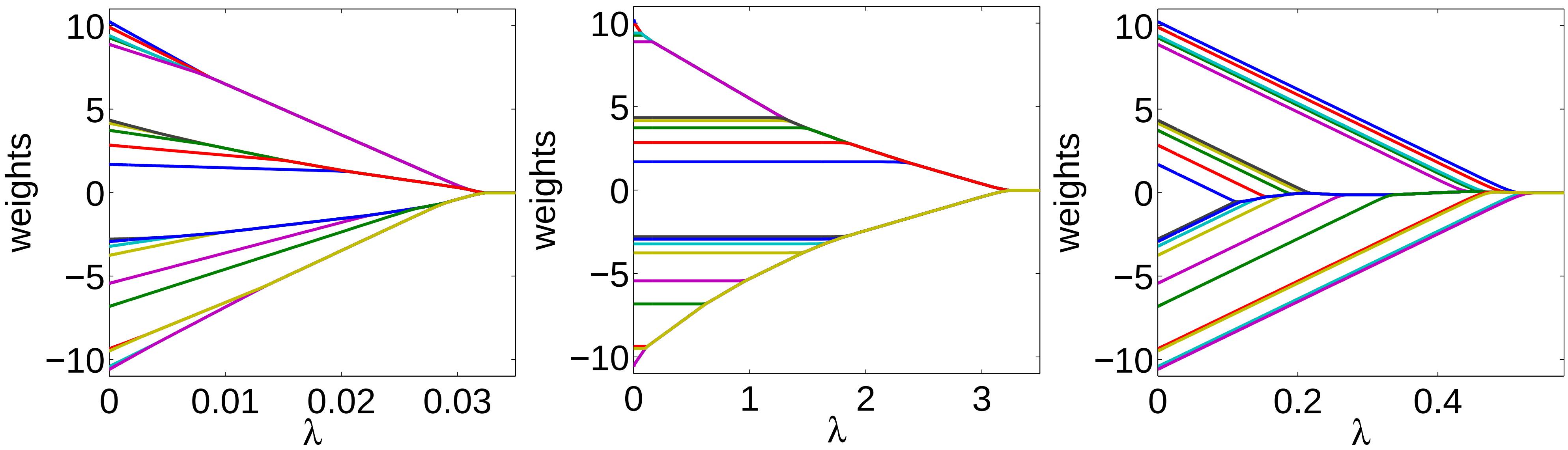
$$F(A) = \min_{B \subset W} \sum_{k \in B,\ j \in W \backslash B} d(k,j) + \lambda |A \Delta B|$$





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



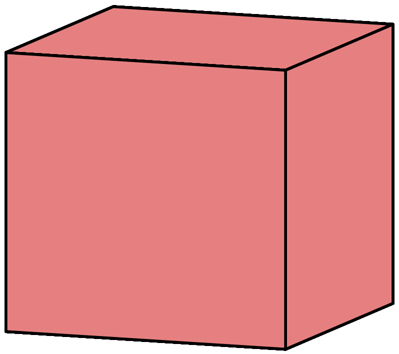
$|A|(p - |A|)$ $\quad$ $1_{|A| \in (0,p)}$ $\quad$ $\max\{|A|, p - |A|\}$

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

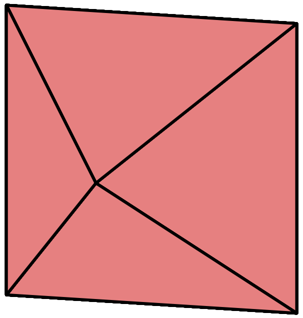
ℓ_2-relaxation of combinatorial penalties (Obozinski and Bach, 2012)

- **Main result** of Bach (2010):
  - $f(|w|)$ is the convex envelope of $F(\text{Supp}(w))$ 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)

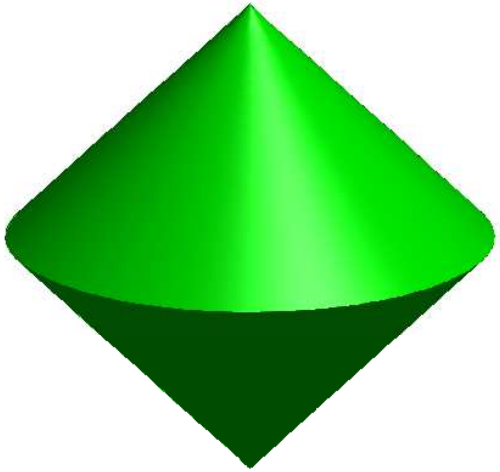
- F a nondecreasing submodular function with Lovász extension f
- Define $\Omega_2(w) = \min_{\eta \in \mathbb{R}_+^p} \dfrac{1}{2} \sum_{i \in V} \dfrac{|w_i|^2}{\eta_i} + \dfrac{1}{2} f(\eta)$
  - NB: general formulation (Micchelli et al., 2011; Bach et al., 2011)
- **Proposition 1**: Ω_2 is the convex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

$$w \mapsto \tfrac{1}{2} F(\mathrm{Supp}(w)) + \tfrac{1}{2} \|w\|_2^2$$

- **Jointly penalizing and regularizing**
  - Extension possible to ℓ_q, $q > 1$

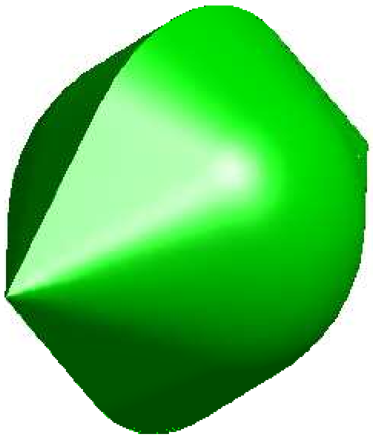
From ℓ_∞ to ℓ_2

Removal of undesired artefacts



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

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



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

- Extension to non-submodular functions + tightness study: see Obozinski and Bach (2012)

Beyond submodular functions?

- Let F be **any** set-function
- **"Edmonds extension"**: the convex envelope of $w \mapsto F(\text{Supp}(w))$ on $[0,1]^p$ is equal to

$$f(w) = \sup_{\forall A \subseteq V,\ s(A) \leqslant F(A)} w^\top s = \sup_{s \in P(F)} w^\top s$$

  – When is it an extension of F?

- **Lower combinatorial envelope**: $G(B) = f(1_B) = \sup_{s \in P(F)} s(B)$
  – $G \leqslant F$
  – Property: idempotent operation
- **A new class of set-functions: functions for which $G = F$**

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 machine learning and submodularity**
  - Improved complexity bounds for submodular function minimization
  - Submodular function maximization
  - Importing concepts from machine learning (e.g., graphical models)
  - Multi-way partitions for computer vision
  - Online learning
  - Going beyond linear programming duality?

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. Learning with submodular functions: A convex optimization perspective. *Arxiv preprint arXiv:1111.6453*, 2011a.

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

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

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.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 2012. To appear.

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

H. H. Bauschke, P. L. Combettes, and D. R. Luke. Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. *J. Approx. Theory*, 127(2):178–192, 2004.

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.

M. J. Best and N. Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1):425–439, 1990.

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.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2008.

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.

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. 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.

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

G. Choquet. Theory of capacities. *Ann. Inst. Fourier*, 5:131–295, 1954.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

J. Edmonds. Submodular functions, matroids, and certain polyhedra. In *Combinatorial optimization - Eureka, you shrink!*, pages 11–26. Springer, 1970.

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.

S. Fujishige and S. Isotani. A submodular function minimization algorithm based on the minimum-norm base. *Pacific Journal of Optimization*, 7:3–17, 2011.

E. Girlich and N. N. Pisaruk. The simplex method for submodular function minimization. Technical Report 97-42, University of Magdeburg, 1997.

J.-L. Goffin and J.-P. Vial. On the computation of weighted analytic centers and dual ellipsoids with the projective algorithm. *Mathematical Programming*, 60(1-3):81–92, 1993.

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.

Z. Harchaoui and C. Lévy-Leduc. Catching change-points with Lasso. *Adv. NIPS*, 20, 2008.

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.

S. Iwata, L. Fleischer, and S. Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.

S. Jegelka, F. Bach, and S. Sra. Reflection methods for user-friendly submodular optimization. Technical report, HAL, 2013.

Stefanie Jegelka, Hui Lin, and Jeff A. Bilmes. Fast approximate submodular minimization. In *Neural Information Processing Society (NIPS)*, Granada, Spain, December 2011.

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.

V. Kolmogorov. Minimizing a sum of submodular functions. *Disc. Appl. Math.*, 160(15), 2012.

N. Komodakis, N. Paragios, and G. Tziritas. Mrf energy minimization and beyond via dual decomposition. *IEEE TPAMI*, 33(3):531–552, 2011.

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

S. T. McCormick. Submodular function minimization. *Discrete Optimization*, 12:321–391, 2005.

N. Megiddo. Optimal flows in networks with multiple sources and sinks. *Mathematical Programming*, 7(1):97–107, 1974.

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

K. Murota. *Discrete convex analysis*. Number 10. Society for Industrial Mathematics, 2003.

K. Nagano, Y. Kawahara, and K. Aihara. Size-constrained submodular minimization through minimum norm base. In *Proc. ICML*, 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.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Pub, 2003.

Y. Nesterov. Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Tech. Rep*, 76, 2007.

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.

J.B. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. *Mathematical Programming*, 118(2):237–251, 2009.

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

B. Savchynskyy, S. Schmidt, J. Kappes, and C. Schnörr. A study of Nesterovs scheme for Lagrangian decomposition and map labeling. In *CVPR*, 2011.

A. Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.

M. Seeger. On the submodularity of linear experimental design, 2009. `http://lapmal.epfl.ch/papers/subm_lindesign.pdf`.

Naum Zuselevich Shor, Krzysztof C. Kiwiel, and Andrzej Ruszcay?ski. *Minimization methods for non-differentiable functions*. Springer-Verlag New York, Inc., 1985.

P. Stobbe and A. Krause. Efficient minimization of decomposable submodular functions. In *NIPS*, 2010.

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

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society. Series B*, 67(1):91–108, 2005.

P. Wolfe. Finding the nearest point in a polytope. *Math. Progr.*, 11(1):128–149, 1976.

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.