Learning with Submodular Functions

Francis Bach

Sierra project-team, INRIA - Ecole Normale Supérieure

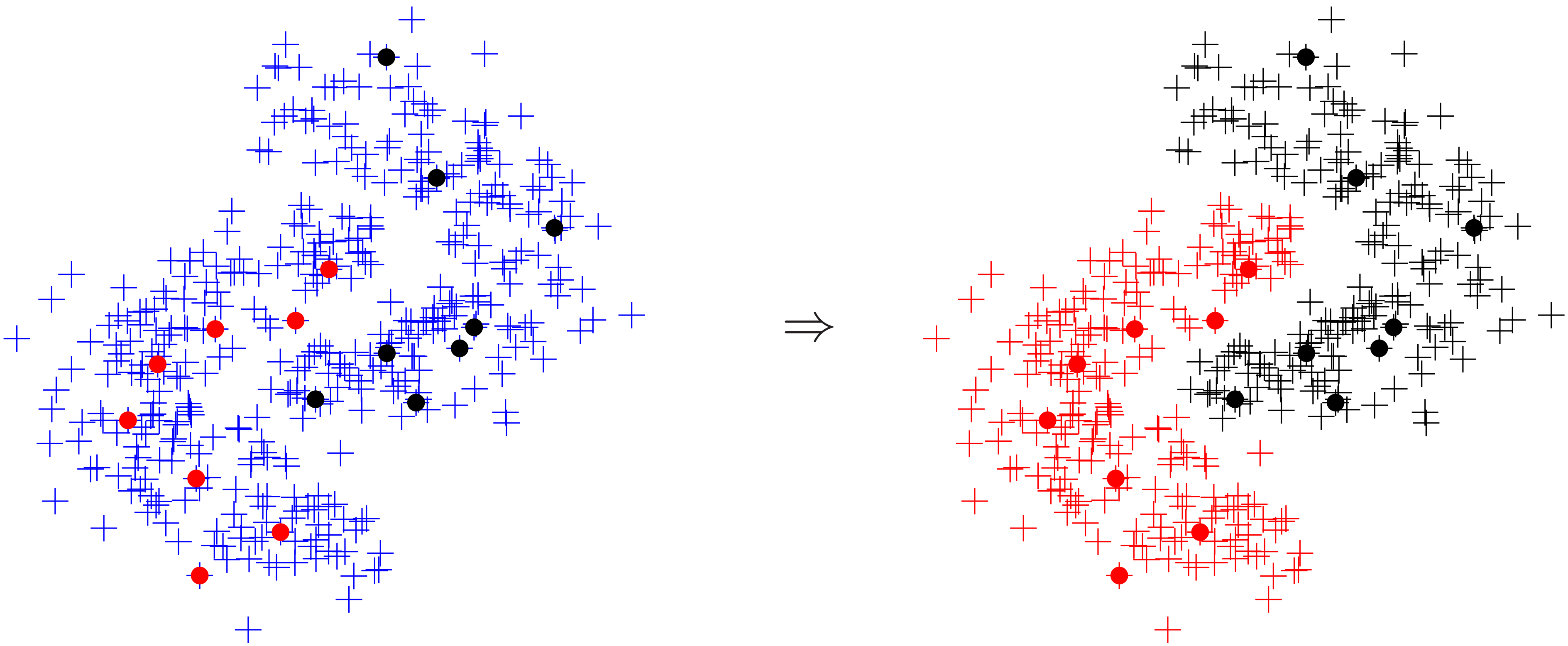
Machine Learning Summer School, Kyoto
September 2012

Submodular functions- References and Links

- **References based on from combinatorial optimization**
  - *Submodular Functions and Optimization* (Fujishige, 2005)
  - *Discrete convex analysis* (Murota, 2003)
- **Tutorial paper based on convex optimization** (Bach, 2011)
  - www.di.ens.fr/~fbach/submodular_fot.pdf
- **Slides for this class**
  - www.di.ens.fr/~fbach/submodular_fbach_mlss2012.pdf
- Other tutorial slides and code at submodularity.org/
- Lecture slides at ssli.ee.washington.edu/~bilmes/ee595a_spring_2011/

Submodularity (almost) everywhere
Clustering

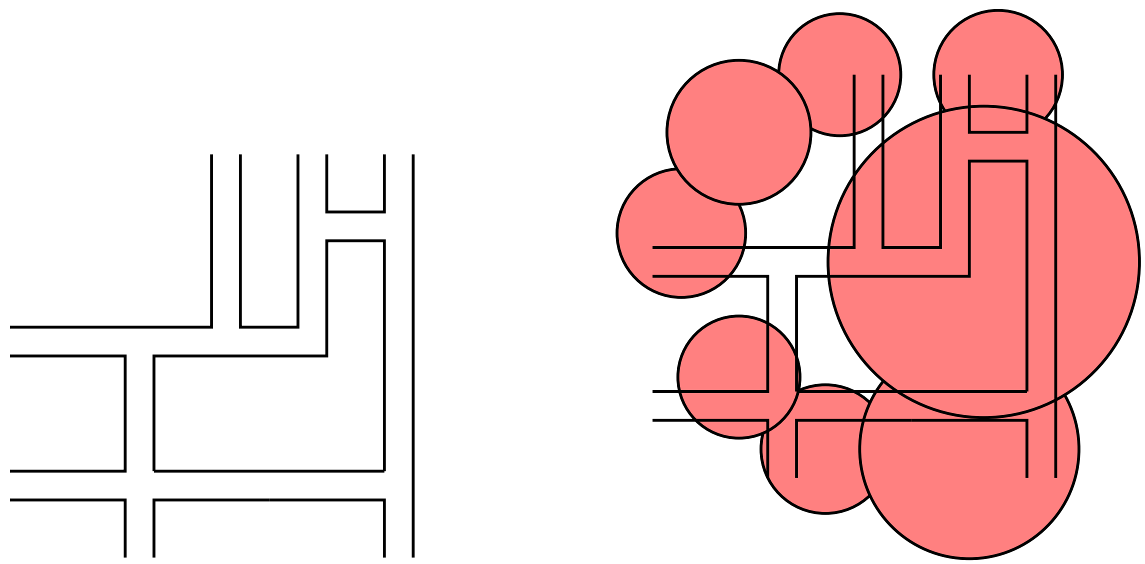
- Semi-supervised clustering



- Submodular function minimization

Submodularity (almost) everywhere
Sensor placement

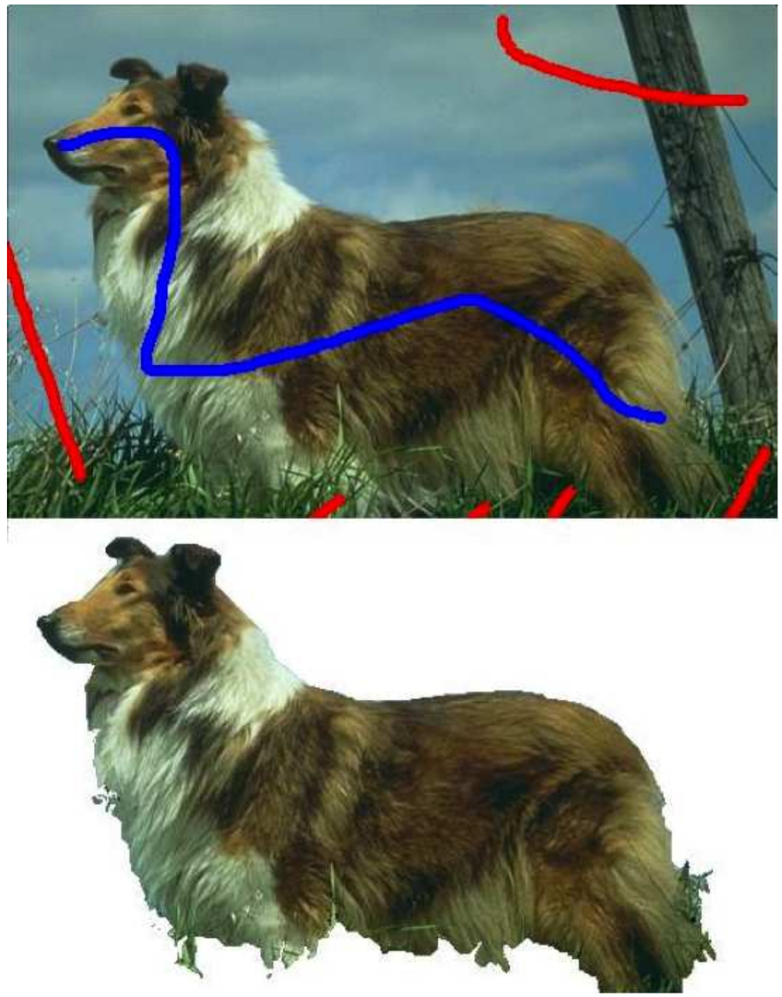
- Each sensor covers a certain area (Krause and Guestrin, 2005)
  - Goal: maximize coverage



- Submodular function maximization
- Extension to experimental design (Seeger, 2009)

Submodularity (almost) everywhere

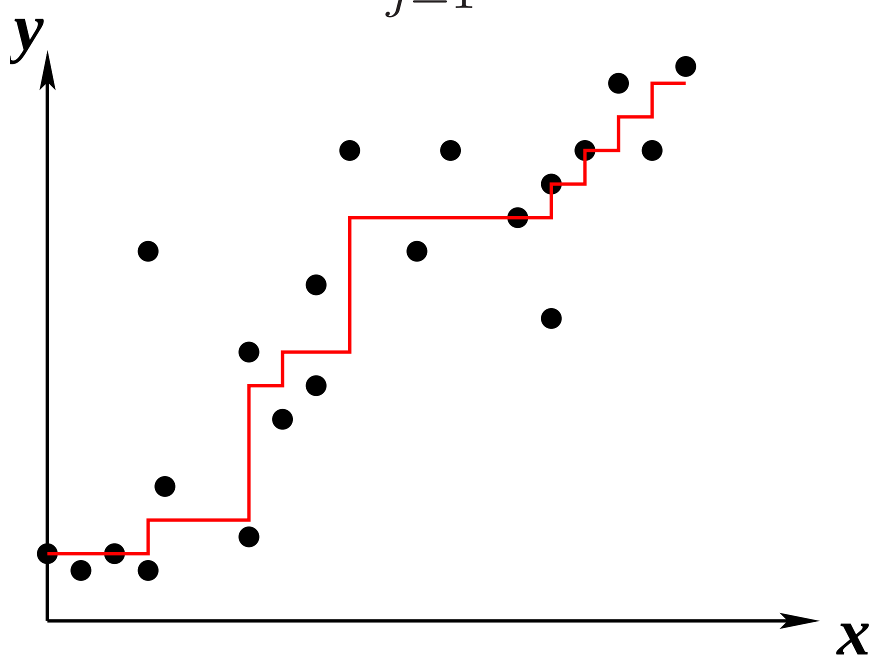
Graph cuts



- Submodular function minimization

Submodularity (almost) everywhere
Isotonic regression

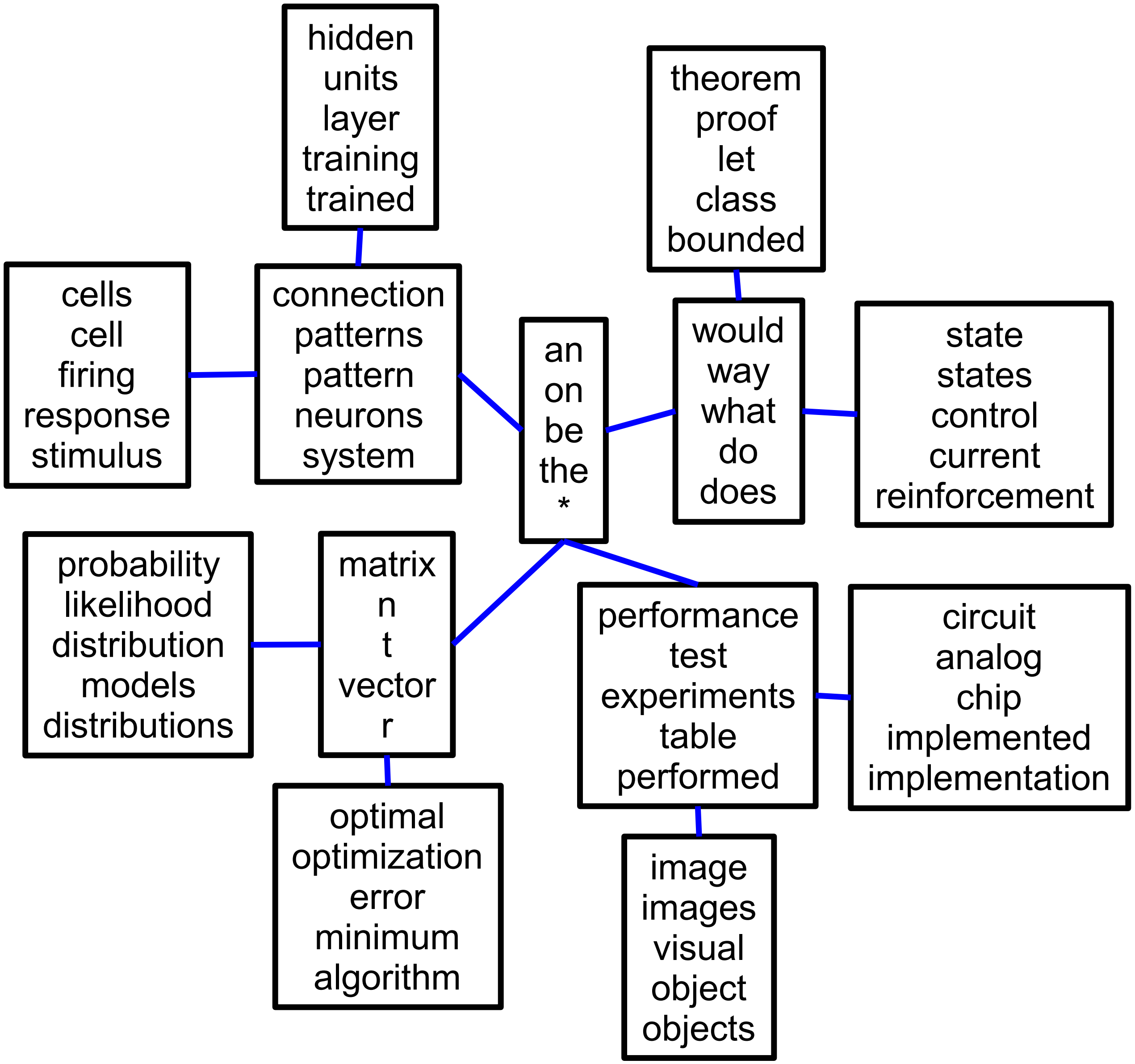
- Given real numbers x_i, $i = 1, \dots, p$
  – Find $y \in \mathbb{R}^p$ that minimizes $\frac{1}{2}\sum_{j=1}^{p}(x_i - y_i)^2$ such that $\forall i, y_i \leqslant y_{i+1}$



- Submodular convex optimization problem

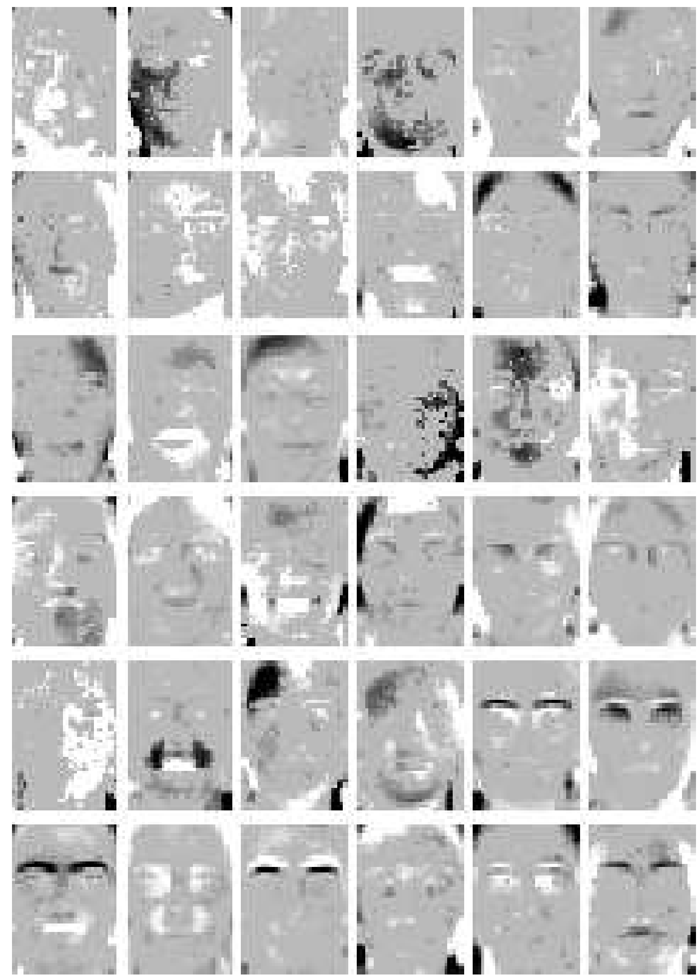
Submodularity (almost) everywhere

Structured sparsity - I



Submodularity (almost) everywhere
Structured sparsity - II

raw data

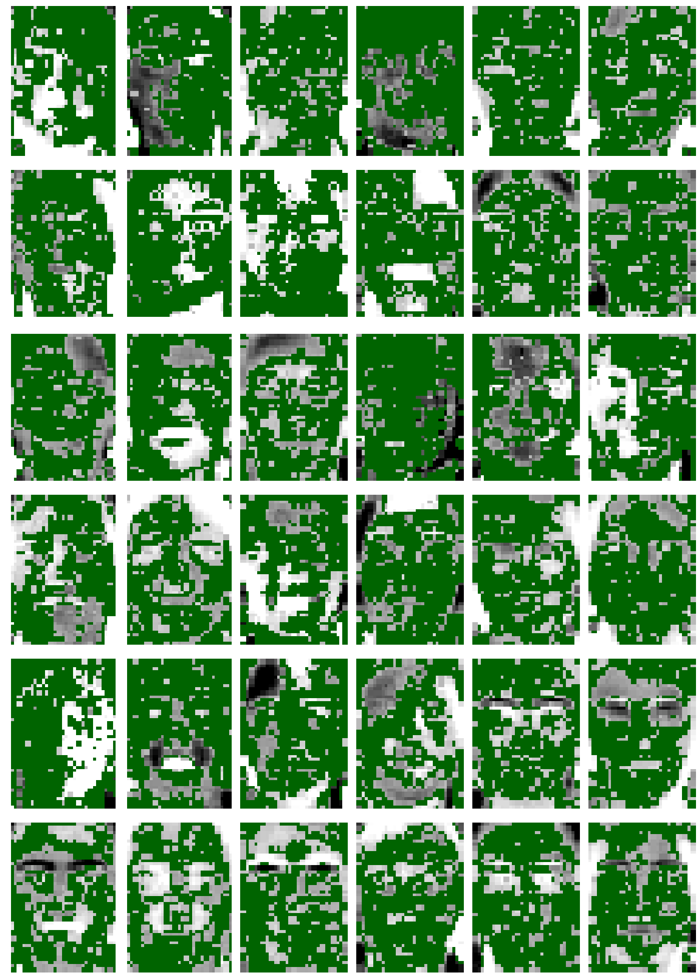


sparse PCA

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

Submodularity (almost) everywhere
Structured sparsity - II

raw data



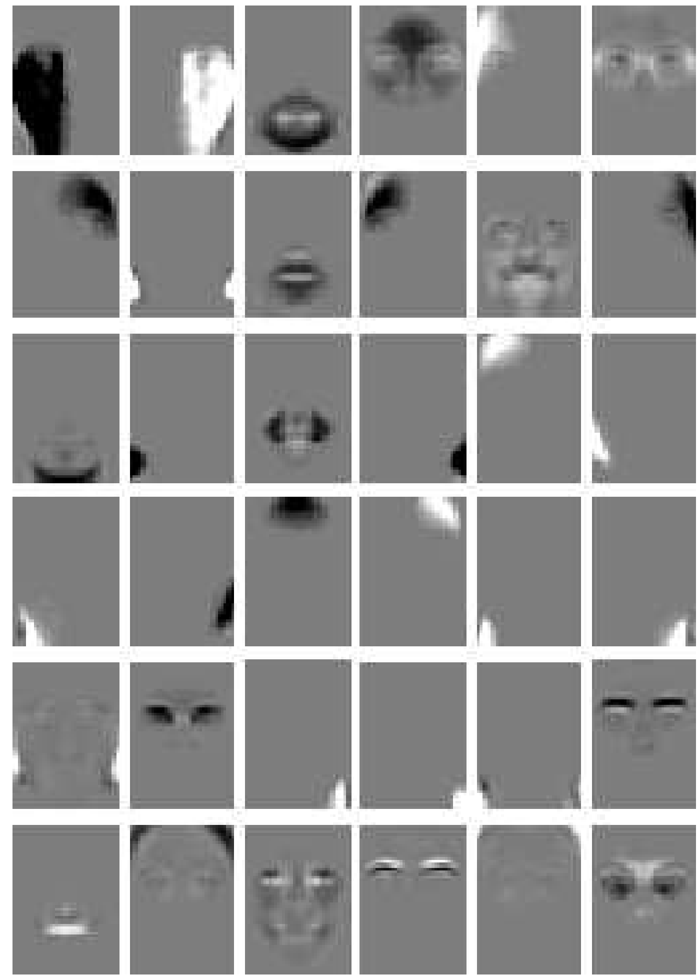
sparse PCA

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

Submodularity (almost) everywhere

Structured sparsity - II

raw data

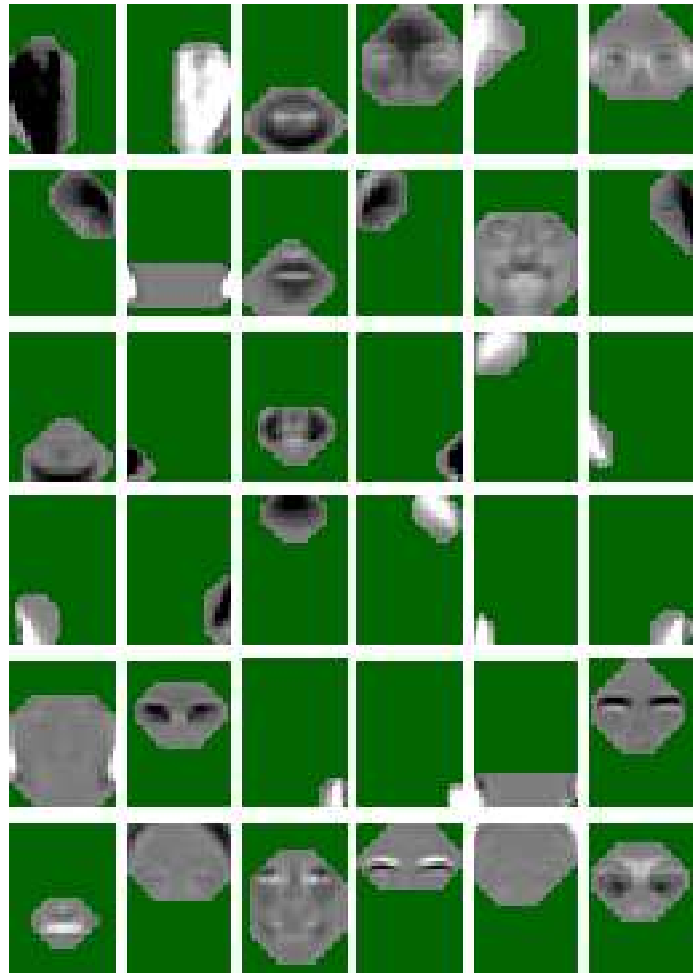


Structured sparse PCA

- Submodular convex optimization problem

Submodularity (almost) everywhere
Structured sparsity - II

raw data



Structured sparse PCA

- Submodular convex optimization problem

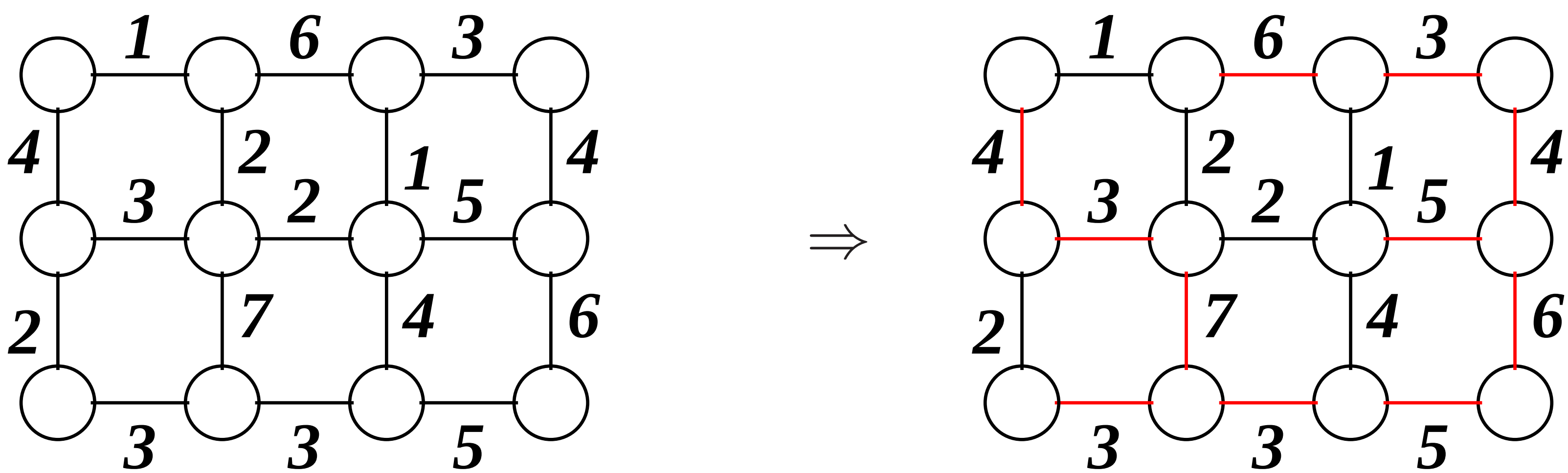
Submodularity (almost) everywhere
Image denoising

- 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

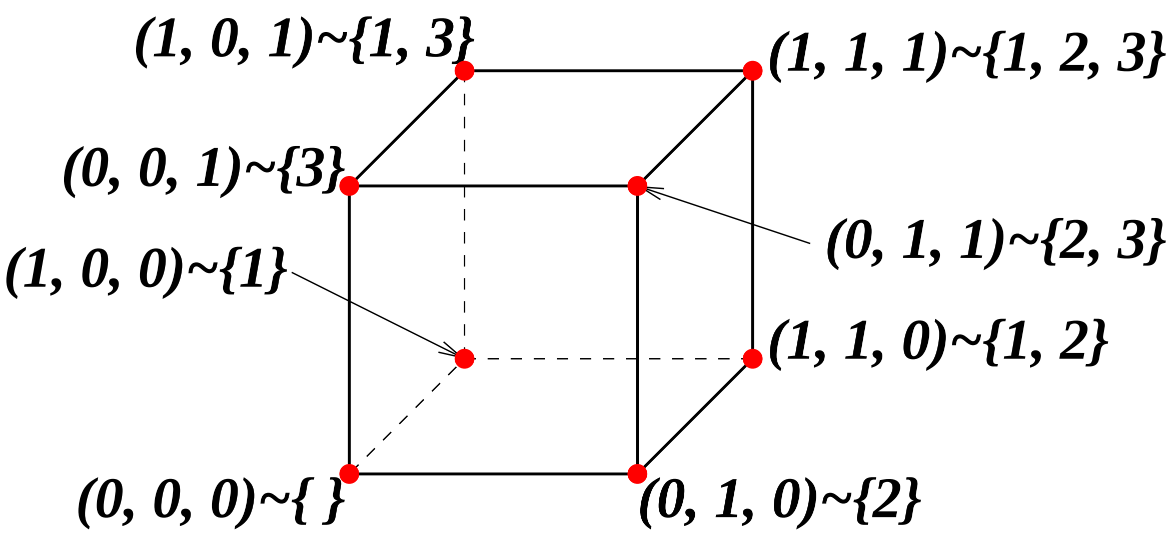
- Set $V = \{1, \dots, p\}$
- Power set 2^V = set of all subsets, of cardinality 2^p
- Minimization/maximization of a set function $F : 2^V \to \mathbb{R}$.

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

- 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

– Definitions
– Examples of submodular functions
– Links with convexity through Lovász extension

2. Submodular optimization

– Minimization
– Links with convex optimization
– Maximization

3. Structured sparsity-inducing norms

– Norms with overlapping groups
– Relaxation of the penalization of supports by submodular functions

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

Submodular functions
Definitions

- **Equivalent definition (easiest to show in practice)**:
  F is submodular if and only if $\forall A \subset V, \ \forall j,k \in V \backslash A$:

$$\boxed{F(A \cup \{k\}) - F(A) \geqslant F(A \cup \{j,k\}) - F(A \cup \{j\})}$$

Submodular functions
Definitions

- **Equivalent definition (easiest to show in practice)**:
  F is submodular if and only if $\forall A \subset V,\ \forall j,k \in V \backslash A$:

$$\boxed{F(A \cup \{k\}) - F(A) \geqslant F(A \cup \{j,k\}) - F(A \cup \{j\})}$$

- **Checking submodularity**

1. Through the definition directly
2. Closedness properties
3. Through the Lovász extension

Submodular functions
Closedness properties

- **Positive linear combinations**: if F_i's are all submodular : $2^V \to \mathbb{R}$ and $\alpha_i \geqslant 0$ for all $i \in \{1, \dots, m\}$, then

$$A \mapsto \sum_{i=1}^{n} \alpha_i F_i(A) \text{ is submodular}$$

Submodular functions
Closedness properties

- **Positive linear combinations**: if F_i's are all submodular : $2^V \to \mathbb{R}$ and $\alpha_i \geqslant 0$ for all $i \in \{1, \ldots, m\}$, then

$$A \mapsto \sum_{i=1}^{n} \alpha_i F_i(A) \text{ is submodular}$$

- **Restriction/marginalization**: if $B \subset V$ and $F : 2^V \to \mathbb{R}$ is submodular, then

$$A \mapsto F(A \cap B) \text{ is submodular on } V \text{ and on } B$$

Submodular functions
Closedness properties

- **Positive linear combinations**: if F_i's are all submodular : $2^V \to \mathbb{R}$ and $\alpha_i \geqslant 0$ for all $i \in \{1, \dots, m\}$, then
$$A \mapsto \sum_{i=1}^{n} \alpha_i F_i(A) \text{ is submodular}$$

- **Restriction/marginalization**: if $B \subset V$ and $F : 2^V \to \mathbb{R}$ is submodular, then
$$A \mapsto F(A \cap B) \text{ is submodular on } V \text{ and on } B$$

- **Contraction/conditioning**: if $B \subset V$ and $F : 2^V \to \mathbb{R}$ is submodular, then
$$A \mapsto F(A \cup B) - F(B) \text{ is submodular on } V \text{ and on } V \backslash B$$

Submodular functions
Partial minimization

- Let G be a submodular function on $V \cup W$, where $V \cap W = \varnothing$
- For $A \subset V$, define $F(A) = \min_{B \subset W} G(A \cup B) - \min_{B \subset W} G(B)$
- **Property**: the function F is submodular and $F(\varnothing) = 0$

Submodular functions
Partial minimization

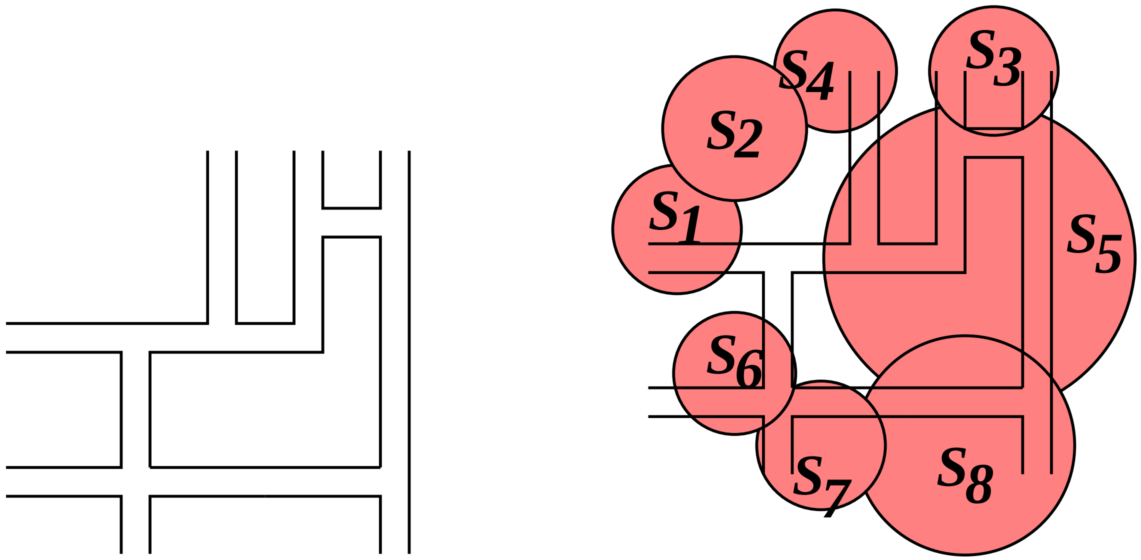
- Let G be a submodular function on $V \cup W$, where $V \cap W = \varnothing$
- For $A \subset V$, define $F(A) = \min_{B \subset W} G(A \cup B) - \min_{B \subset W} G(B)$
- **Property**: the function F is submodular and $F(\varnothing) = 0$
- NB: partial minimization also preserves convexity
- NB: $A \mapsto \max\{F(A), G(A)\}$ and $A \mapsto \min\{F(A), G(A)\}$ might not be submodular

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 1**: 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

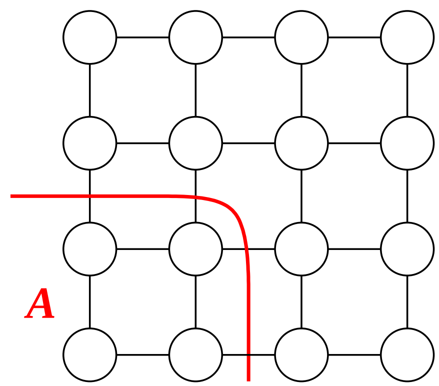


- Let W be any "base" set, and for each $k \in V$, a set $S_k \subset W$
- Set cover defined as $F(A) = \left| \bigcup_{k \in A} S_k \right|$
- *Proof of submodularity*

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*

Examples of submodular functions
Entropies

- Given p random variables $X_1, \ldots, X_p$ 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 $X_{V \backslash A}$

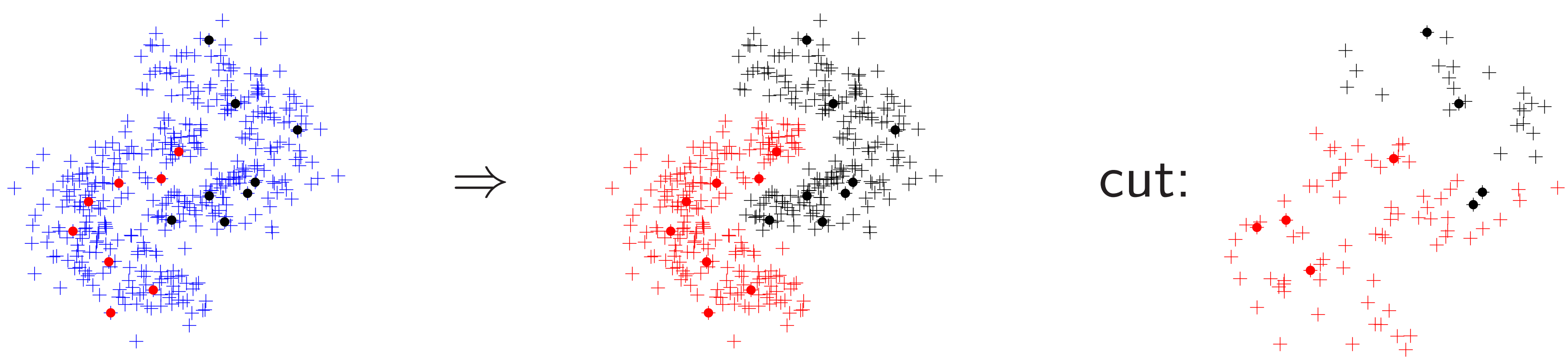
- Extension to continuous random variables, e.g., Gaussian: $F(A) = \log \det \Sigma_{AA}$, for some positive definite matrix $\Sigma \in \mathbb{R}^{p \times p}$

Entropies, Gaussian processes and clustering

- Assume a joint Gaussian process with covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$
- Prior distribution on subsets $p(A) = \prod_{k \in A} \eta_k \prod_{k \notin A} (1 - \eta_k)$
- Modeling with independent Gaussian processes on A and $V \backslash A$
- Maximum a posteriori: minimize

$$I(f_A, f_{V \backslash A}) - \sum_{k \in A} \log \eta_k - \sum_{k \in V \backslash A} \log(1 - \eta_k)$$

- Similar to independent component analysis (Hyvärinen et al., 2001)



Examples of submodular functions
Flows

- Net-flows from multi-sink multi-source networks (Megiddo, 1974)
- See details in `www.di.ens.fr/~fbach/submodular_fot.pdf`
- **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** (More later!)
  - 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**
   - Definitions
   - Examples of submodular functions
   - Links with convexity through Lovász extension

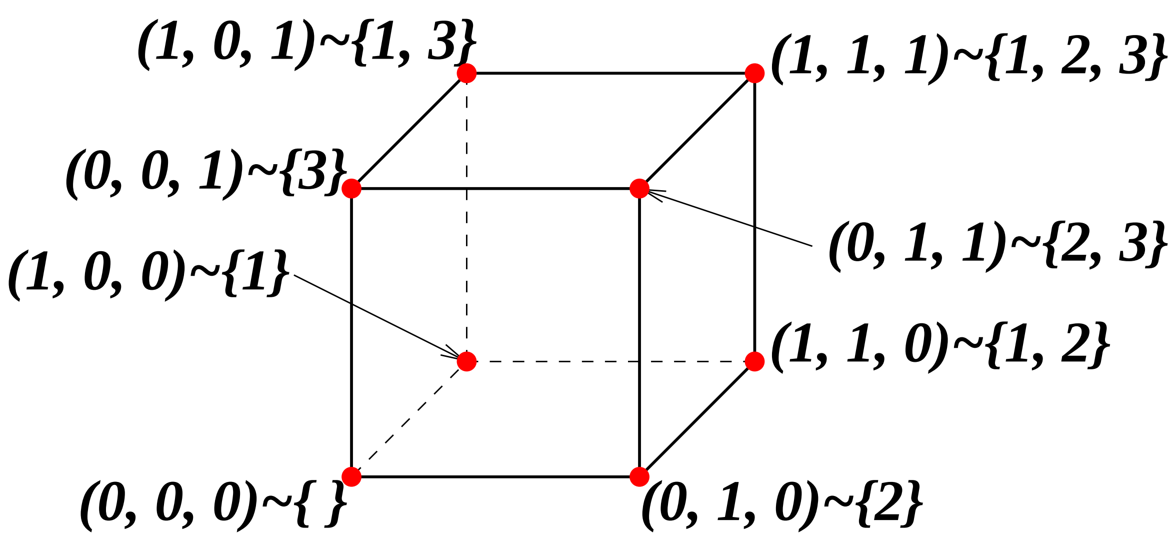
2. **Submodular optimization**
   - Minimization
   - Links with convex optimization
   - Maximization

3. **Structured sparsity-inducing norms**
   - Norms with overlapping groups
   - Relaxation of the penalization of supports by submodular functions

Choquet integral - 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:

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



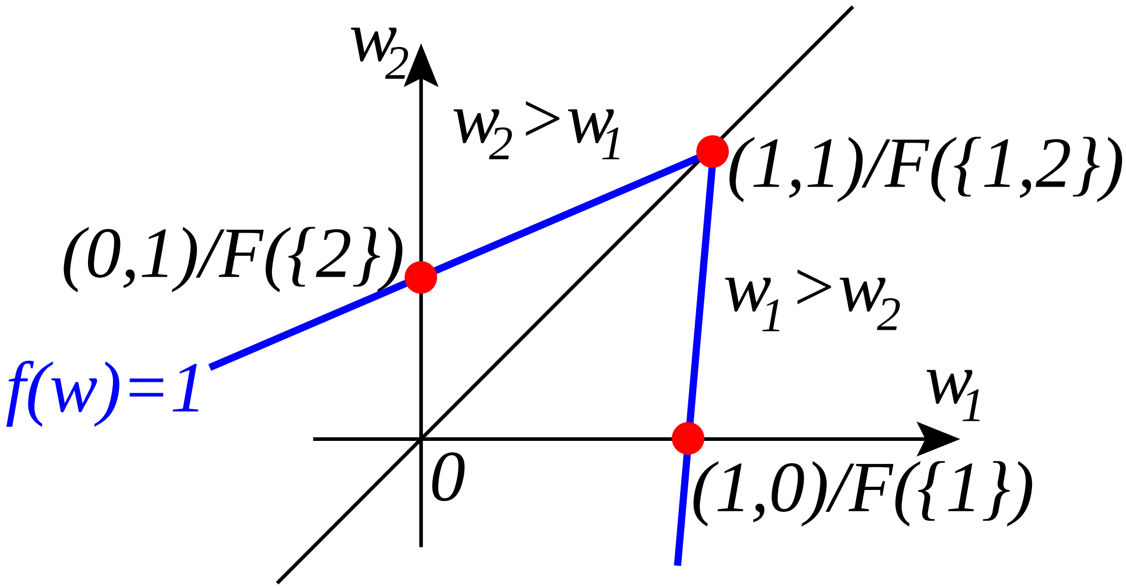
Choquet integral - Lovász extension
Properties

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

- 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 $\mathbb{R}^p$

Choquet integral - Lovász extension
Example with $p = 2$

- If $w_1 \geqslant w_2$, $f(w) = F(\{1\})w_1 + [F(\{1,2\}) - F(\{1\})]w_2$
- If $w_1 \leqslant w_2$, $f(w) = F(\{2\})w_2 + [F(\{1,2\}) - F(\{2\})]w_1$



(level set $\{w \in \mathbb{R}^2, f(w) = 1\}$ *is displayed in blue)*

- NB: Compact formulation $f(w) = -[F(\{1\})+F(\{2\})-F(\{1,2\})] \min\{w_1, w_2\} + F(\{1\})w_1 + F(\{2\})w_2$

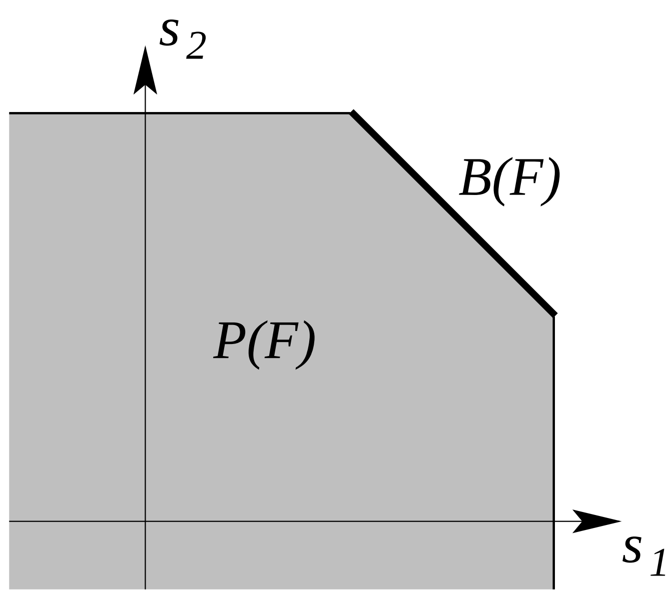
Submodular functions
Links with convexity

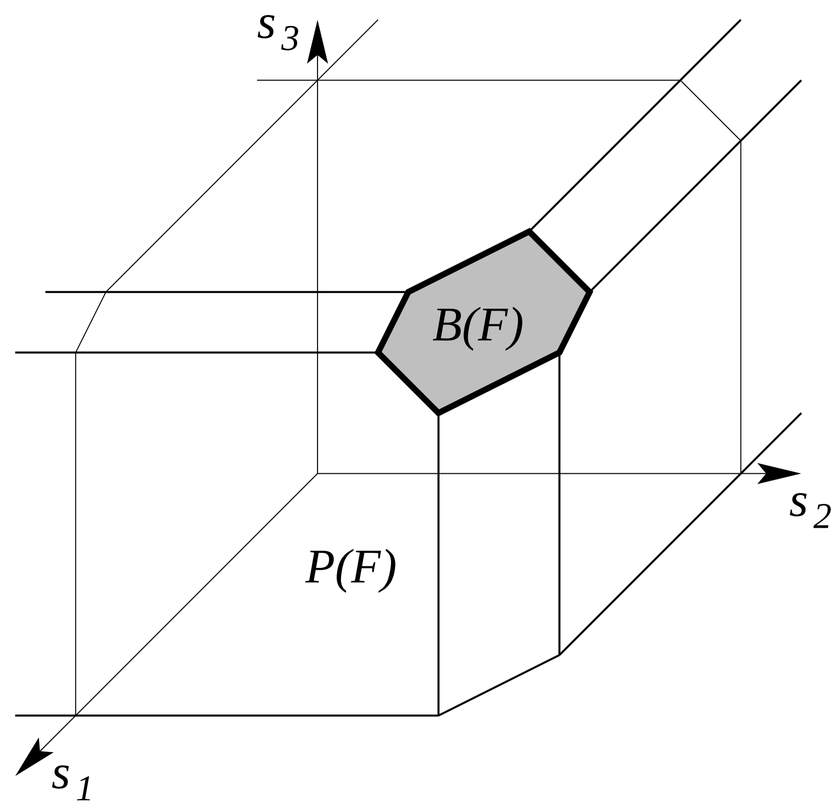
- **Theorem** (Lovász, 1982): F is submodular if and only if f is convex
- Proof requires additional notions:
  - **Submodular and base polyhedra**

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

Greedy algorithms - Proof

- Lagrange multiplier $\lambda_A \in \mathbb{R}_+$ for $s^\top 1_A = s(A) \leqslant F(A)$

$$\max_{s \in P(F)} w^\top s = \min_{\lambda_A \geqslant 0, A \subset V} \max_{s \in \mathbb{R}^p} \left\{ w^\top s - \sum_{A \subset V} \lambda_A [s(A) - F(A)] \right\}$$

$$= \min_{\lambda_A \geqslant 0, A \subset V} \max_{s \in \mathbb{R}^p} \left\{ \sum_{A \subset V} \lambda_A F(A) + \sum_{k=1}^{p} s_k \Big(w_k - \sum_{A \ni k} \lambda_A \Big) \right\}$$

$$= \min_{\lambda_A \geqslant 0, A \subset V} \sum_{A \subset V} \lambda_A F(A) \text{ such that } \forall k \in V, \ w_k = \sum_{A \ni k} \lambda_A$$

- Define $\lambda_{\{j_1, \dots, j_k\}} = w_{j_k} - w_{j_{k-1}}$ for $k \in \{1, \dots, p-1\}$, $\lambda_V = w_{j_p}$, and zero otherwise
  - λ is dual feasible and primal/dual costs are equal to $f(w)$

Proof of greedy algorithm - Showing primal feasibility

- Assume (wlog) $j_k = k$, and $A = (u_1, v_1] \cup \cdots \cup (u_m, v_m]$

$$
\begin{aligned}
s(A) &= \textstyle\sum_{k=1}^m s((u_k, v_k]) \text{ by modularity} \\
&= \textstyle\sum_{k=1}^m \big\{F((0, v_k]) - F((0, u_k])\big\} \text{ by definition of } s \\
&\leqslant \textstyle\sum_{k=1}^m \big\{F((u_1, v_k]) - F((u_1, u_k])\big\} \text{ by submodularity} \\
&= F((u_1, v_1]) + \sum_{k=2}^m \big\{F((u_1, v_k]) - F((u_1, u_k])\big\} \\
&\leqslant F((u_1, v_1]) + \textstyle\sum_{k=2}^m \big\{F((u_1, v_1] \cup (u_2, v_k]) - F((u_1, v_1] \cup (u_2, u_k])\big\} \\
&\qquad \text{by submodularity} \\
&= F((u_1, v_1] \cup (u_2, v_2]) \\
&\quad + \textstyle\sum_{k=3}^m \big\{F((u_1, v_1] \cup (u_2, v_k]) - F((u_1, v_1] \cup (u_2, u_k])\big\}
\end{aligned}
$$

- By pursuing applying submodularity, we get:
$s(A) \leqslant F((u_1, v_1] \cup \cdots \cup (u_m, v_m]) = F(A)$, i.e., $s \in P(F)$

Greedy algorithm for 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, defined as $F(A) = \max_{I \subset A,\ A \in \mathcal{I}} |I|$ is submodular

- **Greedy algorithm**:
  - Since $F(A \cup \{k\}) - F(A) \in \{0,1\}^p$, $s \in \{0,1\}^p$
  $$\Rightarrow w^\top s = \sum_{k,\ s_k=1} w_k$$
  - Start with $A = \varnothing$, orders weights w_k in decreasing order and sequentially add element k to A if set A remains independent

- Graphic matroid: Kruskal's algorithm for max. weight spanning tree!

Submodular functions
Links with convexity

- **Theorem** (Lovász, 1982): F is submodular if and only if f is convex
- **Proof**
  1. If F is submodular, f is the maximum of linear functions $\Rightarrow f$ convex
  2. If f is convex, let $A, B \subset V$.
     – $1_{A\cup B} + 1_{A\cap B} = 1_A + 1_B$ has components equal to 0 (on $V \backslash (A \cup B)$), 2 (on $A \cap B$) and 1 (on $A \Delta B = (A \backslash B) \cup (B \backslash A)$)
     – Thus $f(1_{A\cup B} + 1_{A\cap B}) = F(A \cup B) + F(A \cap B)$.
     – By homogeneity and convexity, $f(1_A + 1_B) \leqslant f(1_A) + f(1_B)$, which is equal to $F(A) + F(B)$, and thus F is submodular.

Submodular functions
Links with convexity

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

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

- **Proof**

1. Since f is an extension of F,
   $\min_{A\subset V} F(A) = \min_{w\in\{0,1\}^p} f(w) \geqslant \min_{w\in[0,1]^p} f(w)$
2. Any $w \in [0,1]^p$ may be decomposed as $w = \sum_{i=1}^m \lambda_i 1_{B_i}$ where $B_1 \subset \cdots \subset B_m = V$, where $\lambda \geqslant 0$ and $\lambda(V) \leqslant 1$:
   - Then $f(w) = \sum_{i=1}^m \lambda_i F(B_i) \geqslant \sum_{i=1}^m \lambda_i \min_{A\subset V} F(A) \geqslant \min_{A\subset V} F(A)$ (because $\min_{A\subset V} F(A) \leqslant 0$).
   - Thus $\min_{w\in[0,1]^p} f(w) \geqslant \min_{A\subset V} F(A)$

Submodular functions
Links with convexity

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

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

- **Consequence**: Submodular function minimization may be done in polynomial time
  - Ellipsoid algorithm: polynomial time but slow in practice

Submodular functions - Optimization

- **Submodular function minimization in $O(p^6)$**
  - Schrijver (2000); Iwata et al. (2001); Orlin (2009)
- **Efficient active set algorithm with no complexity bound**
  - Based on the efficient computability of the support function
  - Fujishige and Isotani (2011); Wolfe (1976)
- **Special cases with faster algorithms**: cuts, flows
- **Active area of research**
  - Machine learning: Stobbe and Krause (2010), Jegelka, Lin, and Bilmes (2011)
  - Combinatorial optimization: see Satoru Iwata's talk
  - Convex optimization: See next part of tutorial

Submodular functions - Summary

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

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

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

Submodular functions - Summary

- $F : 2^V \to \mathbb{R}$ 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 - Summary

- $F : 2^V \to \mathbb{R}$ 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: $F : A \mapsto g(\text{Card}(A))$ is submodular if g is concave

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

Submodular functions - Examples

- Concave functions of the cardinality: $g(|A|)$
- Cuts
- Entropies
  - $H((X_k)_{k \in A})$ from p random variables $X_1, \ldots, X_p$
  - 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

- Given any set-function F and w such that $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$, define:

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

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

- F is submodular if and only if f is convex

  - Minimizing $f(w)$ on $w \in [0,1]^p$ equivalent to minimizing F on 2^V

Submodular functions - Submodular polyhedra

- Submodular polyhedron: $P(F) = \{s \in \mathbb{R}^p,\ \forall A \subset V,\ s(A) \leqslant F(A)\}$
- Base polyhedron: $B(F) = P(F) \cap \{s(V) = F(V)\}$
- **Link with Lovász extension** (Edmonds, 1970; Lovász, 1982):
  - if $w \in \mathbb{R}^p_+$, then $\max_{s \in P(F)} w^\top s = f(w)$
  - if $w \in \mathbb{R}^p$, then $\max_{s \in B(F)} w^\top s = f(w)$
- Maximizer obtained by **greedy algorithm**:
  - Sort the components of w, as $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$
  - Set $s_{j_k} = F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\})$
- Other operations on submodular polyhedra (see, e.g., Bach, 2011)

Outline

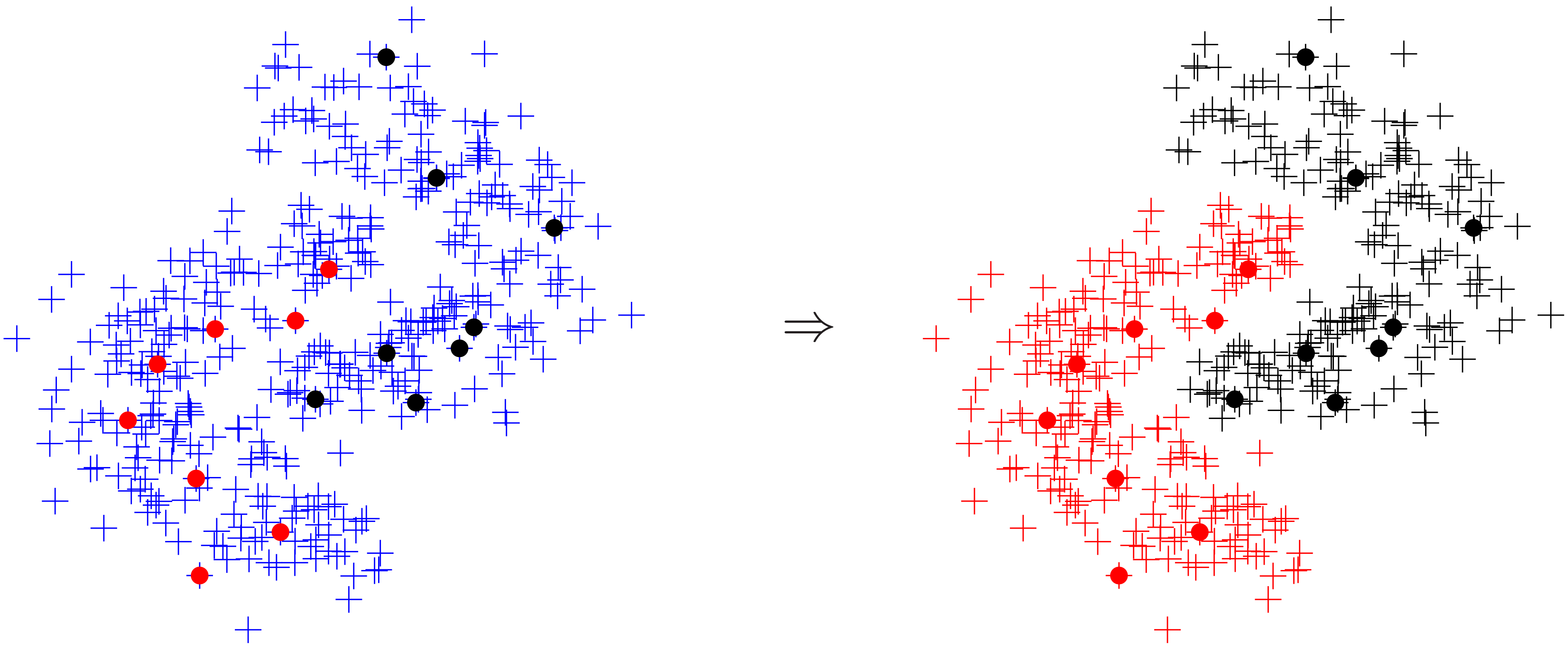
1. **Submodular functions**
   - Definitions
   - Examples of submodular functions
   - Links with convexity through Lovász extension
2. **Submodular optimization**
   - Minimization
   - Links with convex optimization
   - Maximization
3. **Structured sparsity-inducing norms**
   - Norms with overlapping groups
   - Relaxation of the penalization of supports by submodular functions

Submodular optimization problems
Outline

- **Submodular function minimization**
  - Properties of minimizers
  - Combinatorial algorithms
  - Approximate minimization of the Lovász extension
- **Convex optimization with the Lovász extension**
  - Separable optimization problems
  - Application to submodular function minimization
- **Submodular function maximization**
  - Simple algorithms with approximate optimality guarantees

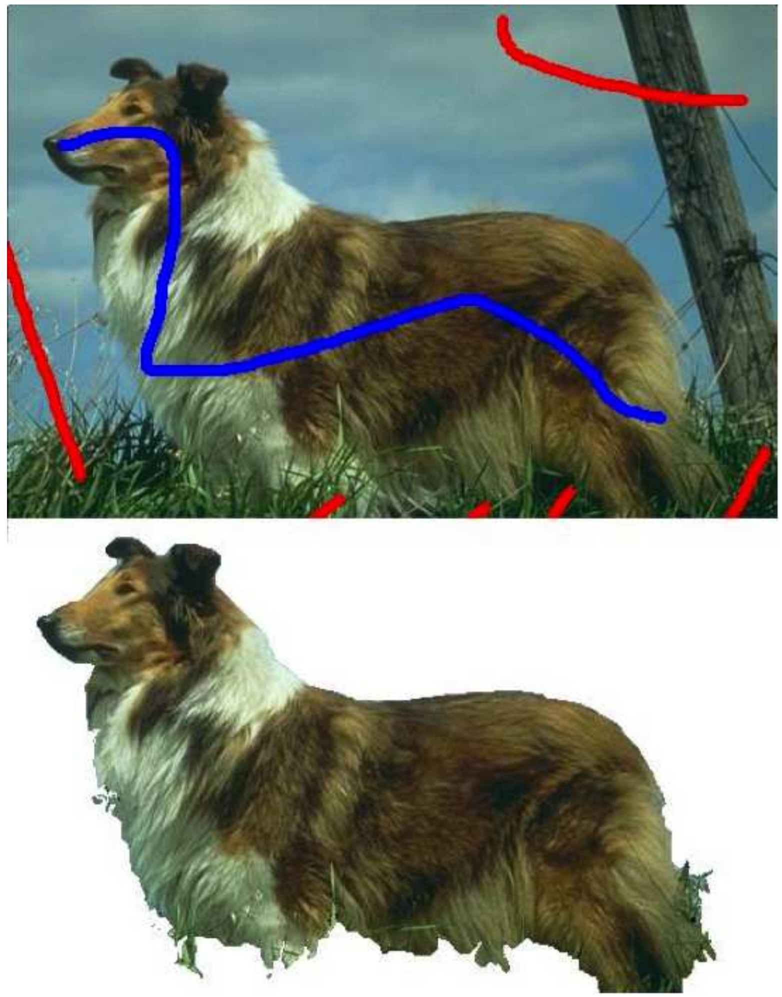
Submodularity (almost) everywhere
Clustering

- Semi-supervised clustering



- Submodular function minimization

Submodularity (almost) everywhere
Graph cuts



- Submodular function minimization

Submodular function minimization
Properties

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

- **Optimality conditions**: $A \subset V$ is a minimizer of F if and only if A is a minimizer of F over all subsets of A and all supersets of A
  - *Proof*: $F(A) + F(B) \geqslant F(A \cup B) + F(A \cap B)$

- **Lattice of minimizers**: if A and B are minimizers, so are $A \cup B$ and $A \cap B$

Submodular function minimization
Dual problem

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

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

- **Optimality conditions**: The pair (A, s) is optimal if and only if $s \in B(F)$ and $\{s < 0\} \subset A \subset \{s \leqslant 0\}$ and $s(A) = F(A)$
  - *Proof*: $F(A) \geqslant s(A) = s(A \cap \{s < 0\}) + s(A \cap \{s > 0\})$
    $\geqslant s(A \cap \{s < 0\}) \geqslant s_-(V)$

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 $s \in B(F)$ such that A is tight for s and $\{s < 0\} \subset A \subset \{s \leqslant 0\}$, 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)

Exact submodular function minimization
Symmetric submodular functions

- A submodular function F is said symmetric if for all $B \subset V$, $F(V \backslash B) = F(B)$
  - Then, by applying submodularity, $\forall A \subset V$, $F(A) \geqslant 0$
- Example: undirected cuts, mutual information
- Minimization in $O(p^3)$ over all *non-trivial* subsets of V (Queyranne, 1998)
- NB: extension to minimization of posimodular functions (Nagamochi and Ibaraki, 1998), i.e., of functions that satisfies

$$\forall A, B \subset V, \ F(A) + F(B) \geqslant F(A \backslash B) + F(B \backslash A).$$

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)

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

- Subgradient of $f(w) = \max_{s \in B(F)} s^\top w$ 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: $f(w_t) - \min_{w \in [0,1]^p} f(w) \leqslant \frac{C}{\sqrt{t}}$ with primal/dual guarantees (Nesterov, 2003; Bach, 2011)

Approximate submodular function minimization
Projected subgradient descent

- Assume (wlog.) that $\forall k \in V$, $F(\{k\}) \geqslant 0$ and $F(V\backslash\{k\}) \geqslant F(V)$

- Denote $D^2 = \sum_{k\in V} \big\{ F(\{k\}) + F(V\backslash\{k\}) - F(V) \big\}$

- Iteration: $w_t = \Pi_{[0,1]^p}\big(w_{t-1} - \frac{D}{\sqrt{pt}} s_t\big)$ with $s_t \in \operatorname*{argmin}_{s\in B(F)} w_{t-1}^\top s$

- **Proposition**: t iterations of subgradient descent outputs a set A_t (and a certificate of optimality s_t) such that

$$F(A_t) - \min_{B\subset V} F(B) \leqslant F(A_t) - (s_t)_-(V) \leqslant \frac{Dp^{1/2}}{\sqrt{t}}$$

Submodular optimization problems
Outline

- **Submodular function minimization**
  - Properties of minimizers
  - Combinatorial algorithms
  - Approximate minimization of the Lovász extension
- **Convex optimization with the Lovász extension**
  - Separable optimization problems
  - Application to submodular function minimization
- **Submodular function maximization**
  - Simple algorithms with approximate optimality guarantees

Separable optimization on base polyhedron

- **Optimization of convex functions** of the form $\boxed{\Psi(w) + f(w)}$ with f Lovász extension of F

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

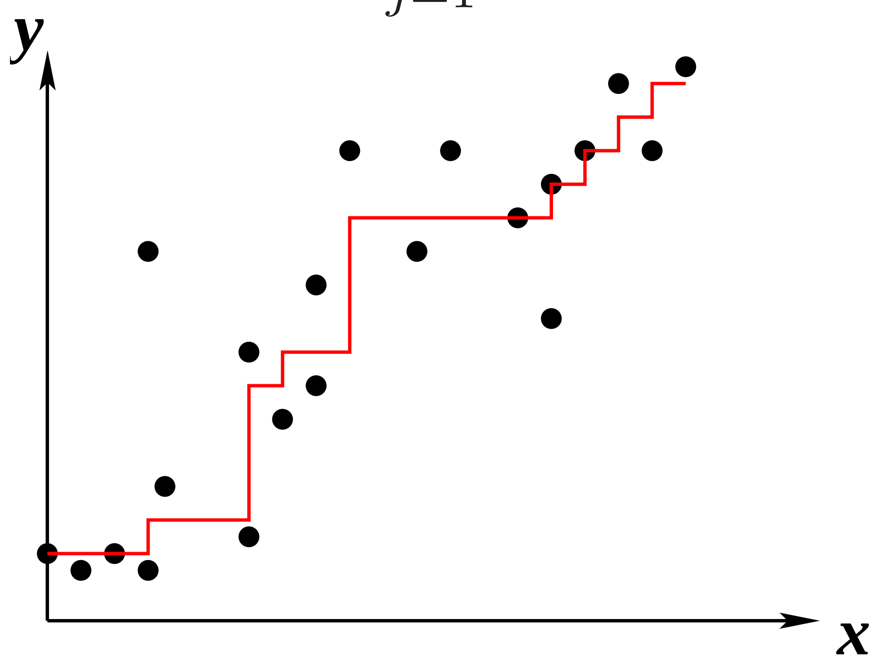
Total variation denoising (Chambolle, 2005)

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

- d symmetric $\Rightarrow f =$ total variation

Isotonic regression

- Given real numbers x_i, $i = 1, \ldots, p$
  - Find $y \in \mathbb{R}^p$ that minimizes $\displaystyle\frac{1}{2}\sum_{j=1}^{p}(x_i - y_i)^2$ 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

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

Separable optimization on base polyhedron

- **Optimization of convex functions** of the form $\boxed{\Psi(w) + f(w)}$ with f Lovász extension of F
- **Structured sparsity**
  - Regularized risk minimization penalized by the Lovász extension
  - Total variation denoising - isotonic regression
- **Proximal methods** (see next part of the tutorial)
  - 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 $\psi_k : \mathbb{R} \to \mathbb{R}$, $k \in \{1, \ldots, p\}$ 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 $\psi_1^*, \ldots, \psi_p^*$ 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 u 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 $\forall j,\ u_j = \sup(\{\alpha \in \mathbb{R},\ j \in A^\alpha\})$
  - Given u, 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

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

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

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

- 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)
  - minimum-norm-point algorithm (Fujishige and Isotani, 2011)

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)
  - minimum-norm-point algorithm (Fujishige and Isotani, 2011)

- 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 function can be maximized
  - **Frank-Wolfe algorithms**

- Two main types for convex functions

Approximate quadratic optimization on $B(F)$

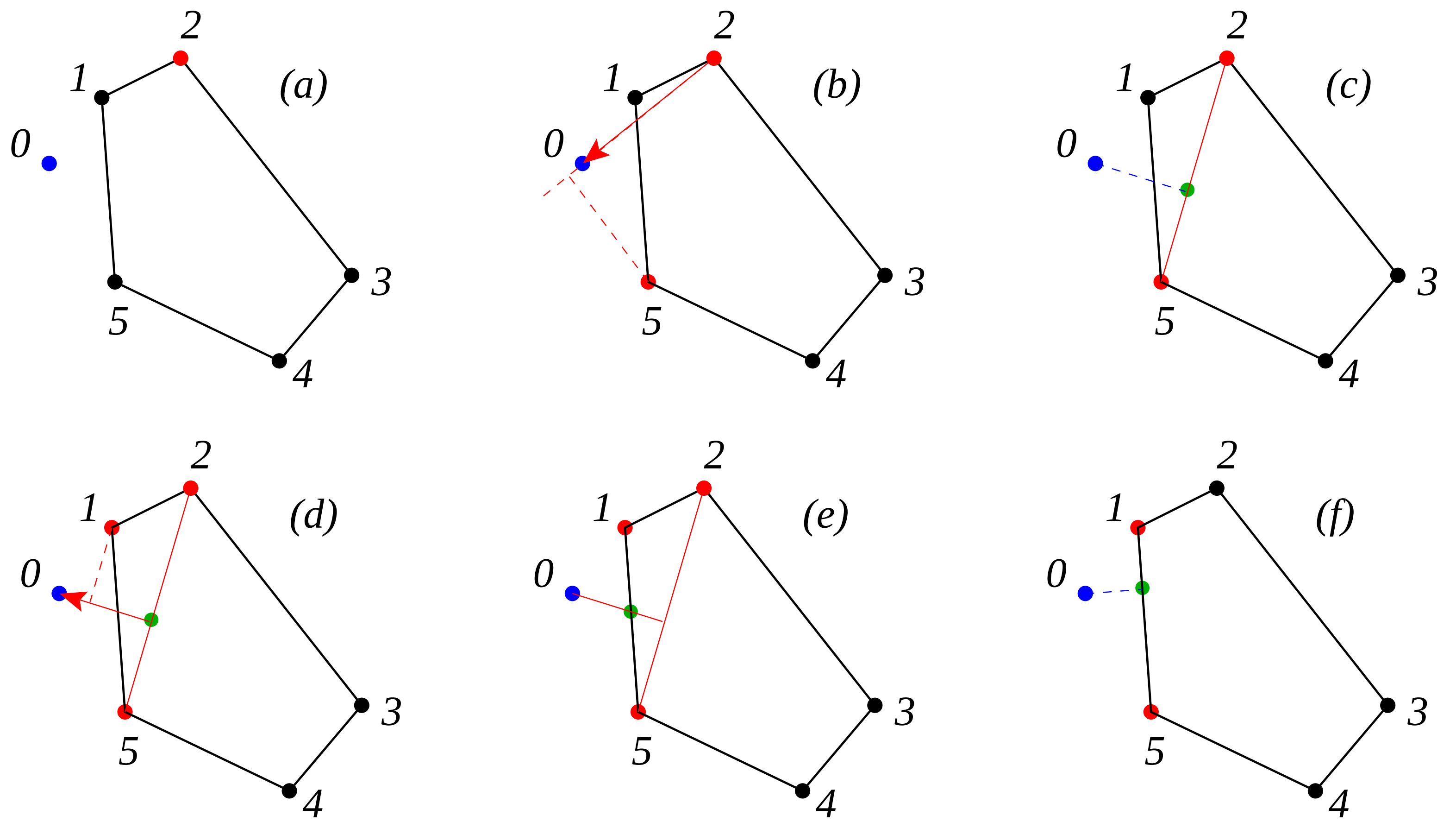
- **Goal**: $\displaystyle \min_{w \in \mathbb{R}^p} \frac{1}{2}\|w\|_2^2 + f(w) = \max_{s \in B(F)} -\frac{1}{2}\|s\|_2^2$

- 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 algorithms



Approximate quadratic optimization on $B(F)$

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

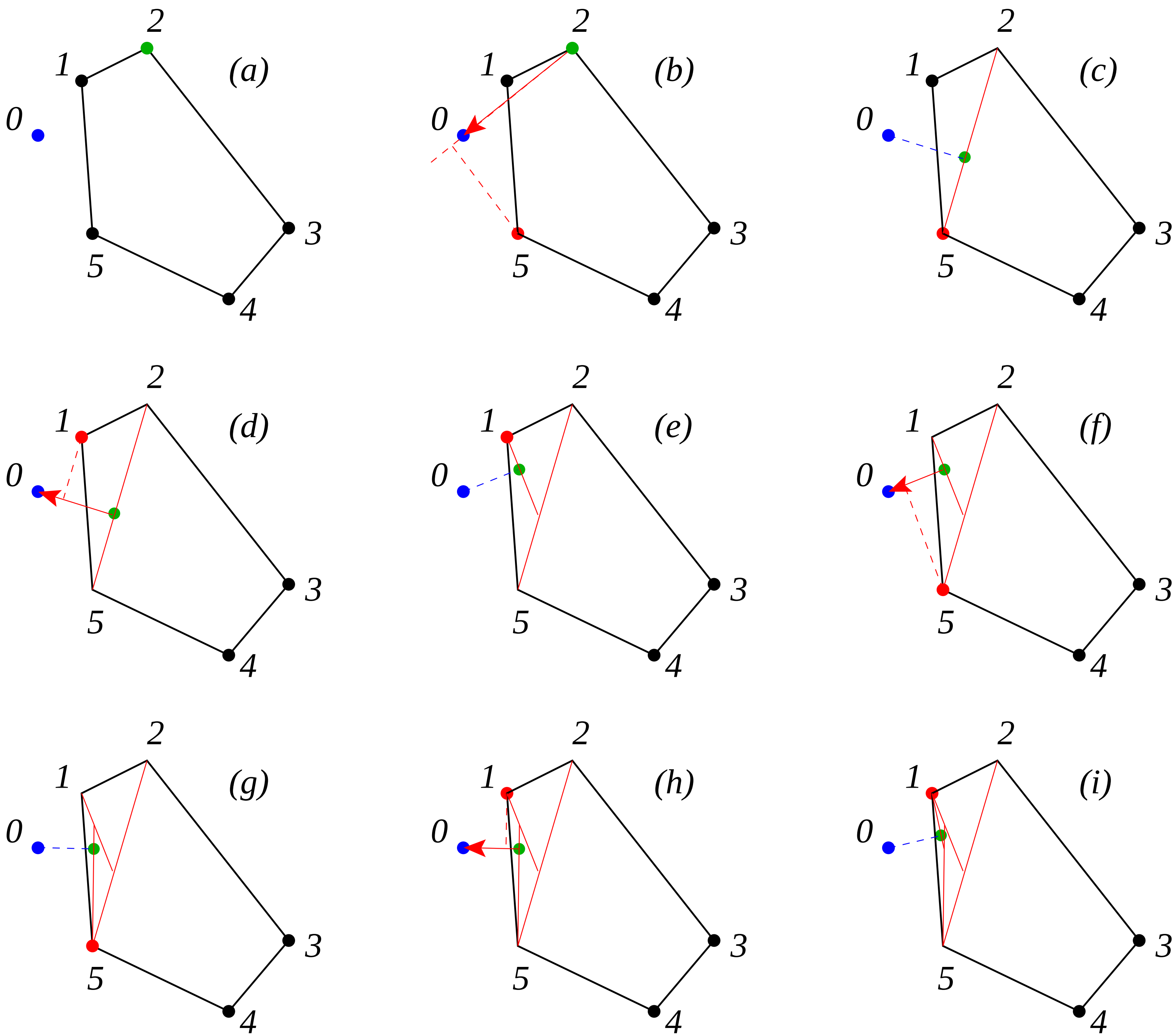
- Can only maximize linear functions on $B(F)$

- **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

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

- **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

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

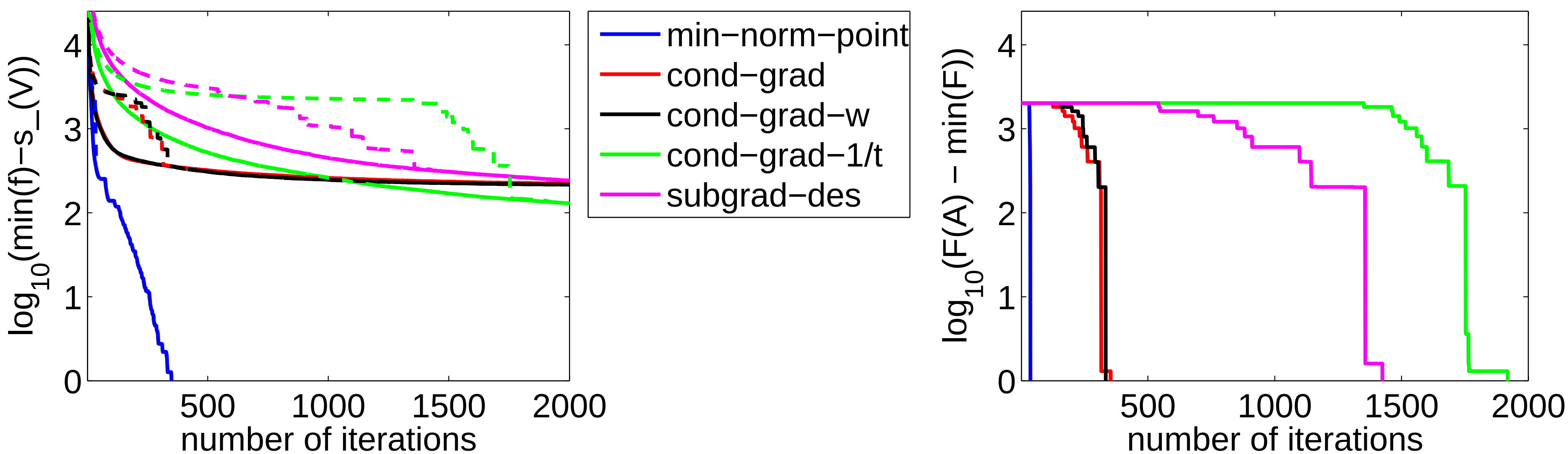
- **Proposition**: If w is ε-optimal for $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w\|_2^2 + f(w)$, then at least a levet set A of w is $\left(\frac{\sqrt{\varepsilon p}}{2}\right)$-optimal for submodular function minimization

- 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**
  - **Proposition**: 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).

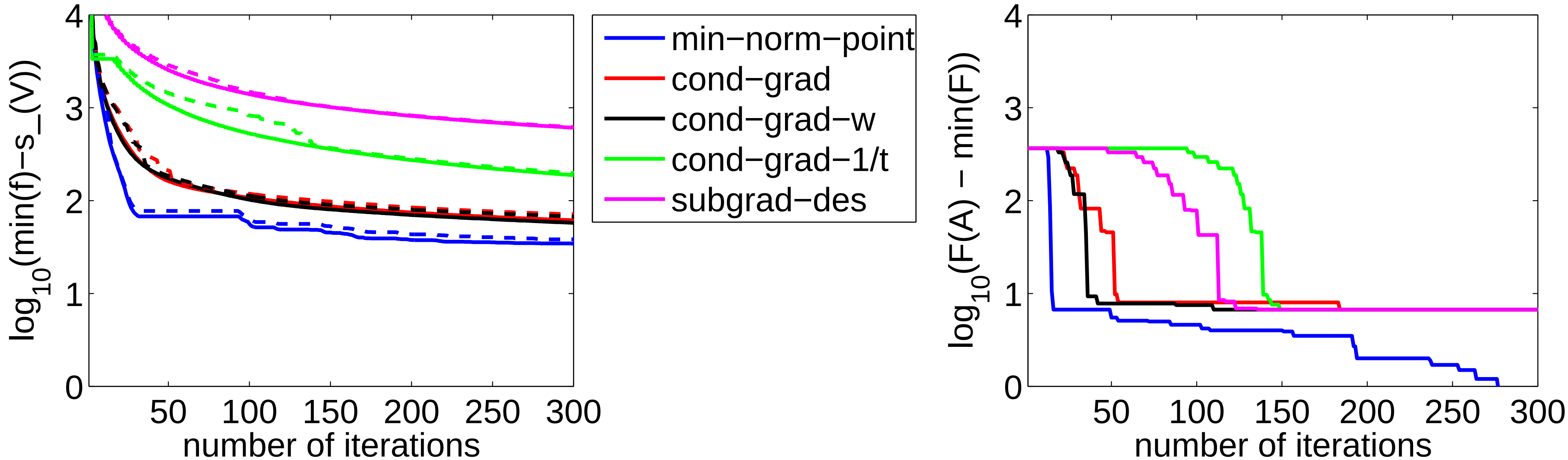
Simulations on standard benchmark "DIMACS Genrmf-wide", p = 430

- **Submodular function minimization**
  - (Left) optimal value minus dual function values $(s_t)_-(V)$ (in dashed, certified duality gap)
  - (Right) Primal function values $F(A_t)$ minus optimal value



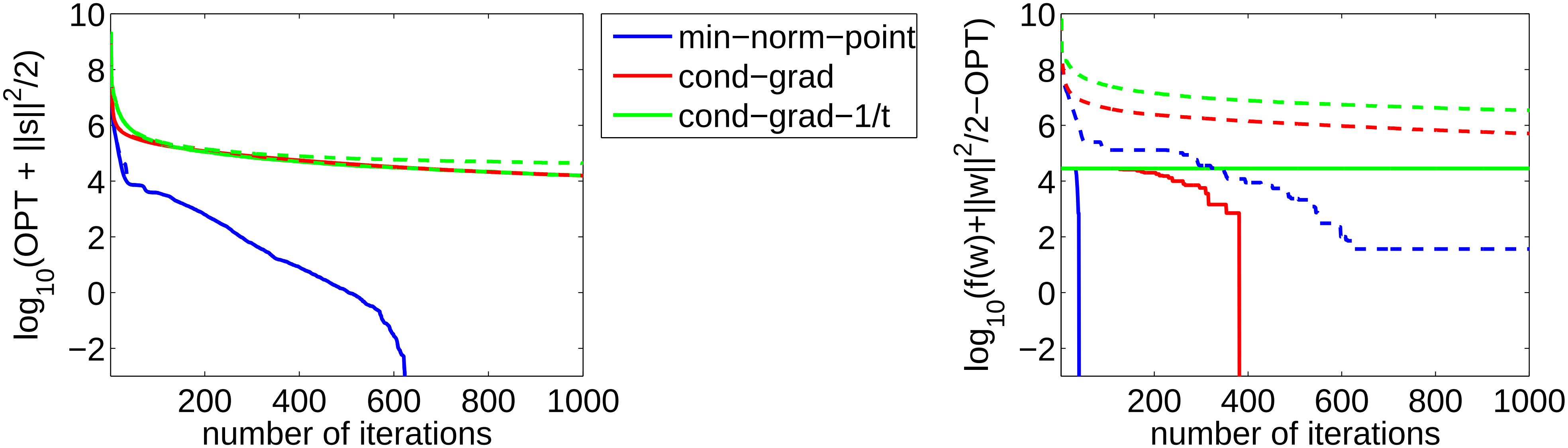
Simulations on standard benchmark "DIMACS Genrmf-long", p = 575

- **Submodular function minimization**
  - (Left) optimal value minus dual function values $(s_t)_-(V)$ (in dashed, certified duality gap)
  - (Right) Primal function values $F(A_t)$ minus optimal value



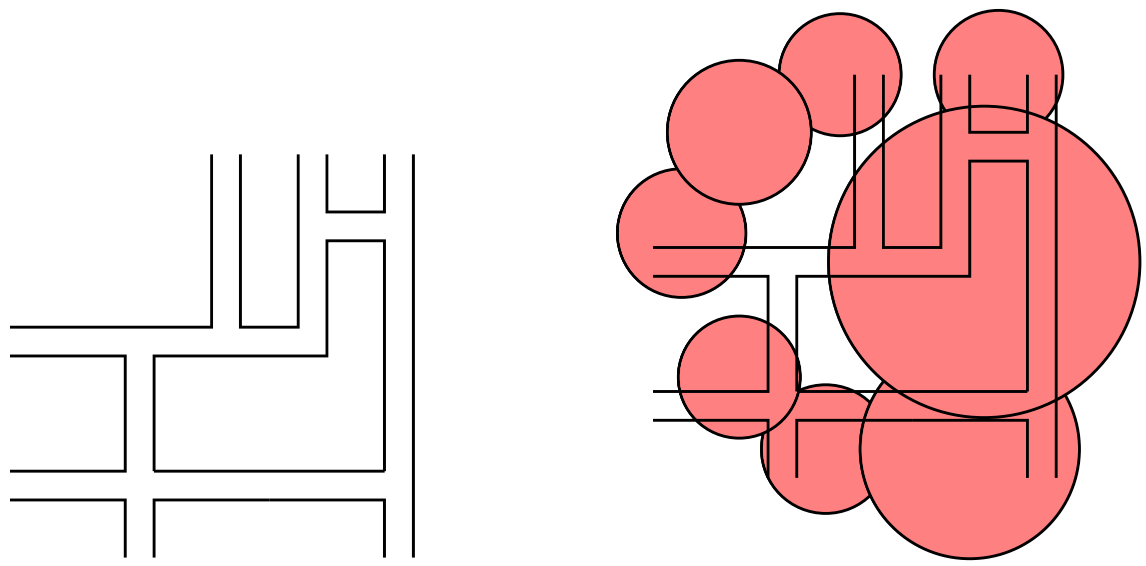
Simulations on standard benchmark

- **Separable quadratic optimization**
  - (Left) optimal value minus dual function values $-\frac{1}{2}\|s_t\|_2^2$ (in dashed, certified duality gap)
  - (Right) Primal function values $f(w_t) + \frac{1}{2}\|w_t\|_2^2$ minus optimal value (in dashed, before the pool-adjacent-violator correction)



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)

Submodular function maximization

- Occurs in various form in applications but is NP-hard
- **Unconstrained maximization**: Feige et al. (2007) shows that that for non-negative functions, a random subset already achieves at least $1/4$ of the optimal value, while local search techniques achieve at least $1/2$
- **Maximizing non-decreasing submodular functions with cardinality constraint**
  - Greedy algorithm achieves $(1 - 1/e)$ of the optimal value
  - *Proof* (Nemhauser et al., 1978)

Maximization with cardinality constraint

- Let $A^* = \{b_1, \ldots, b_k\}$ be a maximizer of F with k elements, and a_j the j-th selected element. Let $\rho_j = F(\{a_1, \ldots, a_j\}) - F(\{a_1, \ldots, a_{j-1}\})$

$$
\begin{aligned}
F(A^*) &\leqslant F(A^* \cup A_{j-1}) \text{ because } F \text{ is non-decreasing,} \\
&= F(A_{j-1}) + \sum_{i=1}^{k} \big[F(A_{j-1} \cup \{b_1, \ldots, b_i\}) - F(A_{j-1} \cup \{b_1, \ldots, b_{i-1}\}) \big] \\
&\leqslant F(A_{j-1}) + \sum_{i=1}^{k} \big[F(A_{j-1} \cup \{b_i\}) - F(A_{j-1}) \big] \text{ by submodularity,} \\
&\leqslant F(A_{j-1}) + k\rho_j \text{ by definition of the greedy algorithm,} \\
&= \sum_{i=1}^{j-1} \rho_i + k\rho_j .
\end{aligned}
$$

- Minimize $\sum_{i=1}^{k} \rho_i$: $\rho_j = (k-1)^{j-1} k^{-j} F(A^*)$

Submodular optimization problems
Summary

- **Submodular function minimization**
  - Properties of minimizers
  - Combinatorial algorithms
  - Approximate minimization of the Lovász extension

- **Convex optimization with the Lovász extension**
  - Separable optimization problems
  - Application to submodular function minimization

- **Submodular function maximization**
  - Simple algorithms with approximate optimality guarantees

Outline

1. **Submodular functions**
   - Definitions
   - Examples of submodular functions
   - Links with convexity through Lovász extension
2. **Submodular optimization**
   - Minimization
   - Links with convex optimization
   - Maximization
3. **Structured sparsity-inducing norms**
   - Norms with overlapping groups
   - Relaxation of the penalization of supports by submodular functions

Sparsity in supervised machine learning

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

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

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

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

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

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

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

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

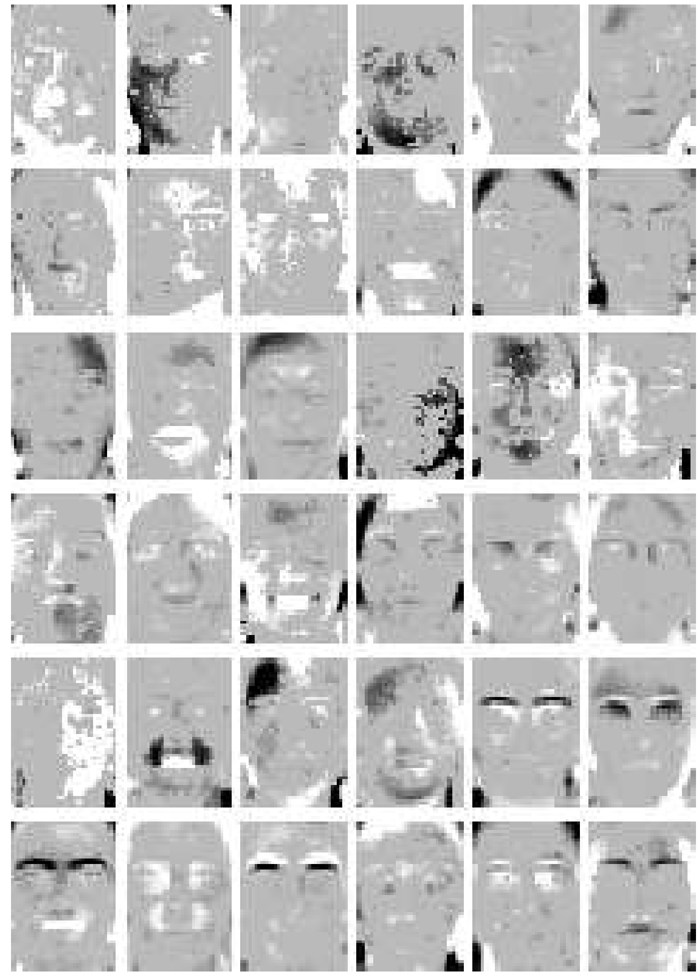
- **sparse PCA**: replace $\|d^j\|_2 \leqslant 1$ by $\Theta(d^j) \leqslant 1$

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

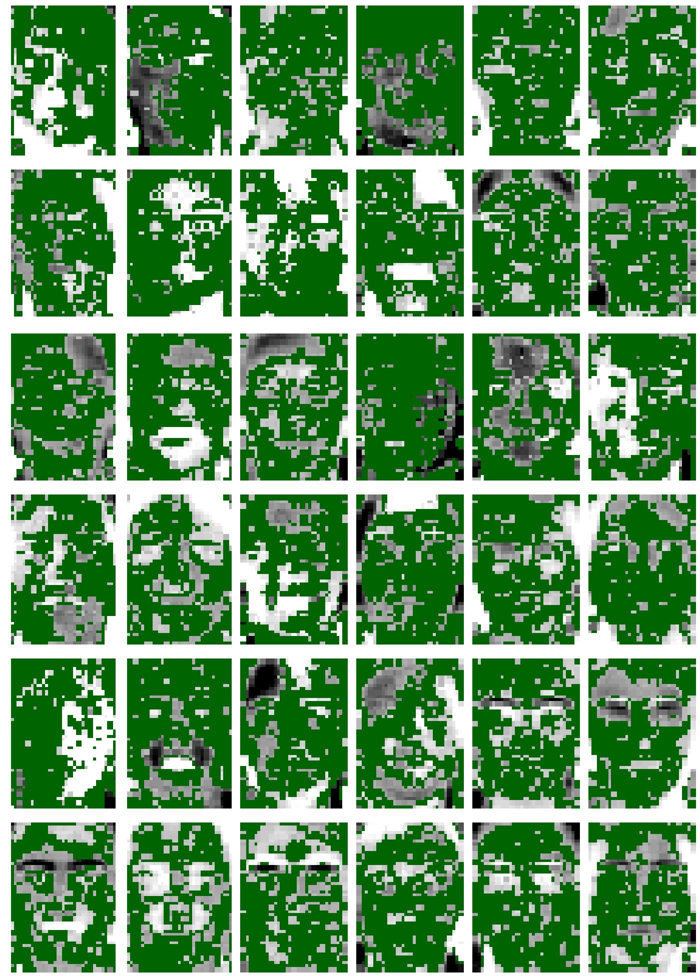


sparse PCA

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

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

raw data

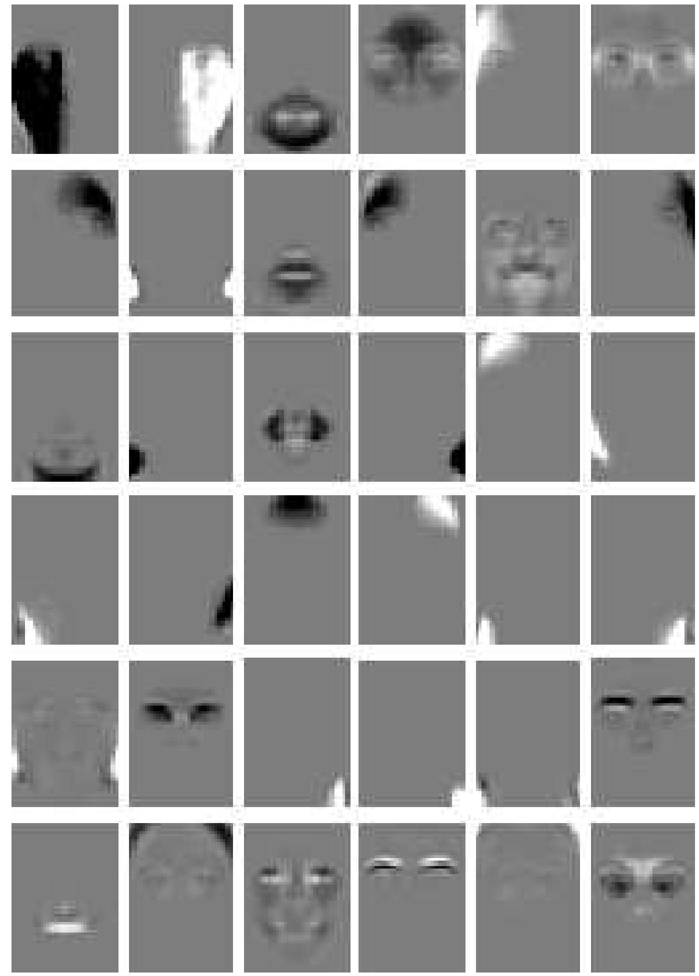


sparse PCA

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

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

raw data

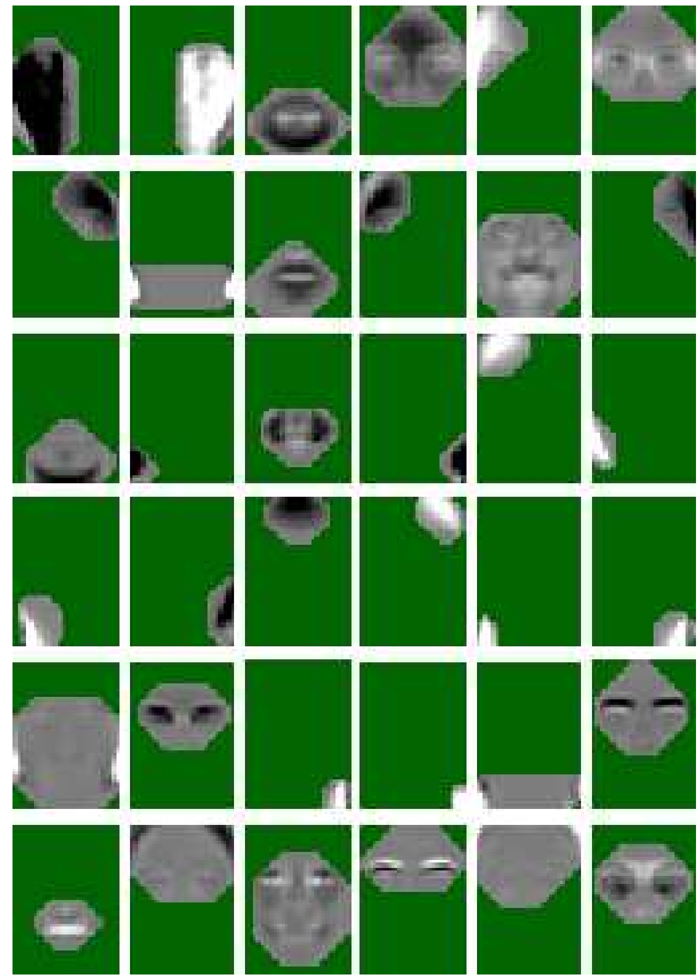


Structured sparse PCA

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

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

raw data



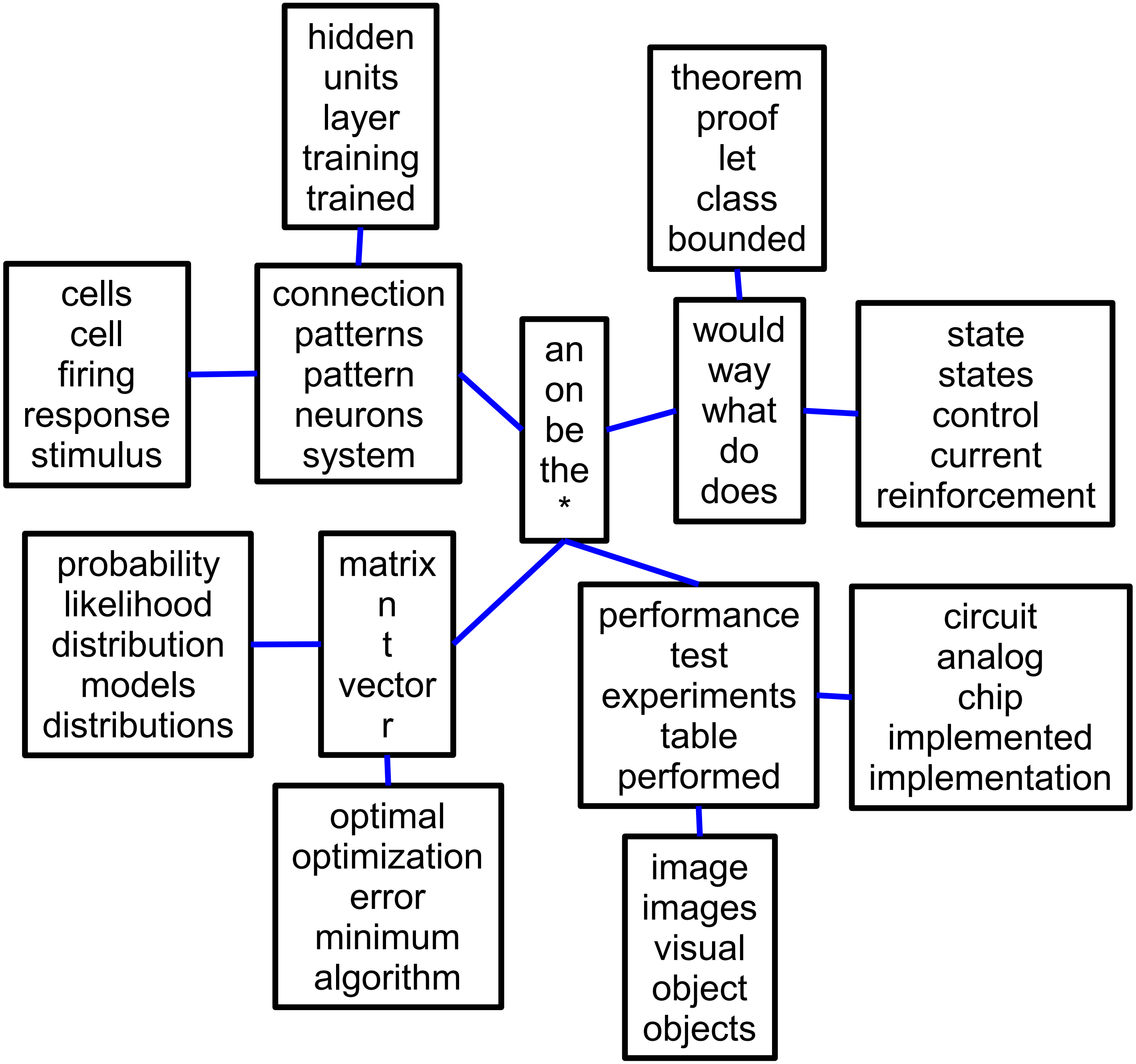
Structured sparse PCA

- 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)

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**
  - 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)

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)

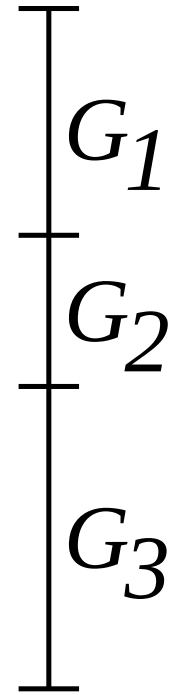
- **Convex approaches**
  - Design of sparsity-inducing norms

Sparsity-inducing norms

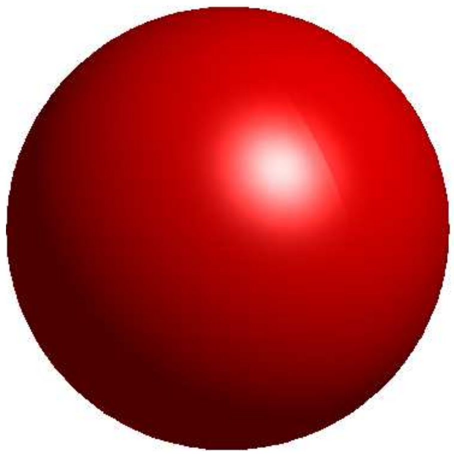
- **Popular choice for Ω**
  - The ℓ_1-ℓ_2 norm,

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

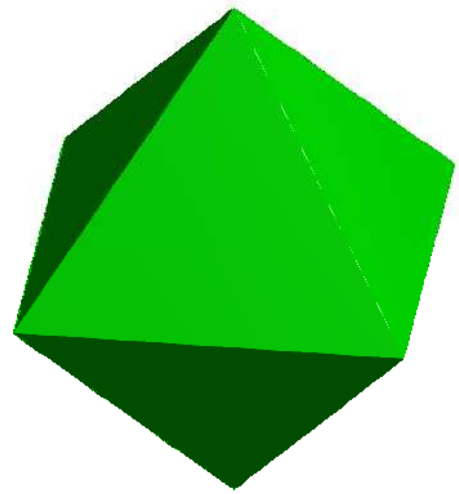
  - with $\mathbf{H}$ a partition of $\{1, \ldots, p\}$
  - The ℓ_1-ℓ_2 norm sets to zero groups of non-overlapping variables (as opposed to single variables for the ℓ_1-norm)
  - For the square loss, group Lasso (Yuan and Lin, 2006)



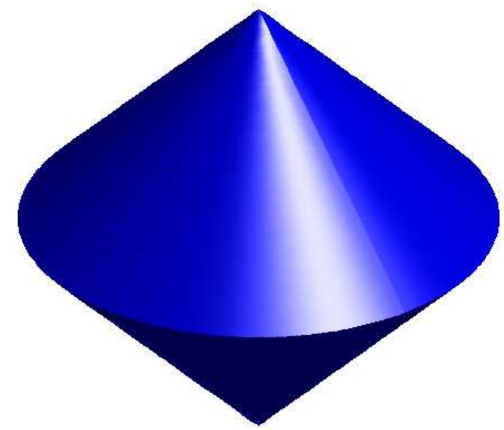
Unit norm balls
Geometric interpretation



$\|w\|_2$



$\|w\|_1$

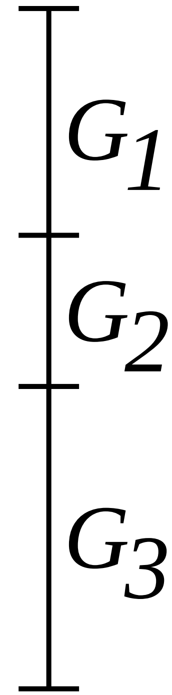


$\sqrt{w_1^2 + w_2^2} + |w_3|$

Sparsity-inducing norms

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

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

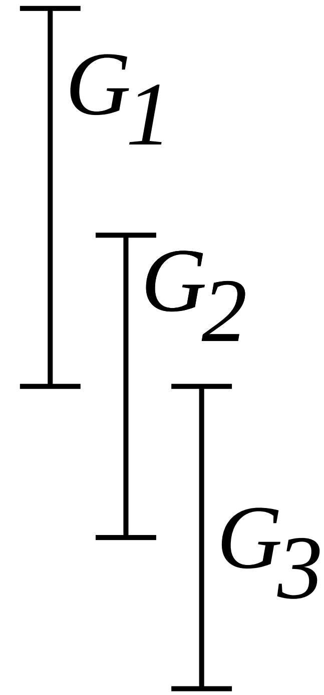


  - with $\mathbf{H}$ a partition of $\{1, \ldots, p\}$
  - The ℓ_1-ℓ_2 norm sets to zero groups of non-overlapping variables (as opposed to single variables for the ℓ_1-norm)
  - For the square loss, group Lasso (Yuan and Lin, 2006)

- However, the ℓ_1-ℓ_2 norm encodes **fixed/static prior information**, requires to know in advance how to group the variables

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

Structured sparsity with overlapping groups (Jenatton, Audibert, and Bach, 2009a)

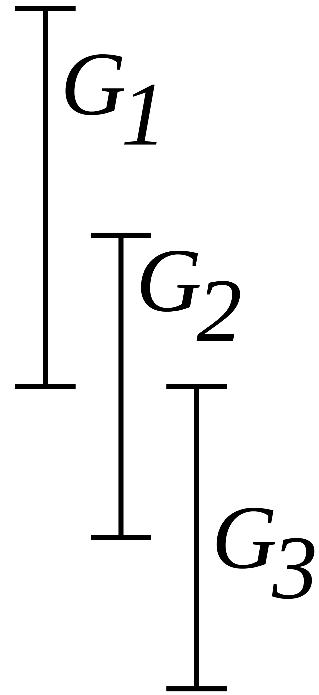


- When penalizing by the ℓ_1-ℓ_2 norm,

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

  - The ℓ_1 norm induces sparsity at the group level:
    - Some w_G's are set to zero
  - Inside the groups, the ℓ_2 norm does not promote sparsity

Structured sparsity with overlapping groups (Jenatton, Audibert, and Bach, 2009a)



- When penalizing by the ℓ_1-ℓ_2 norm,

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

  - The ℓ_1 norm induces sparsity at the group level:
    - Some w_G's are set to zero
  - Inside the groups, the ℓ_2 norm does not promote sparsity

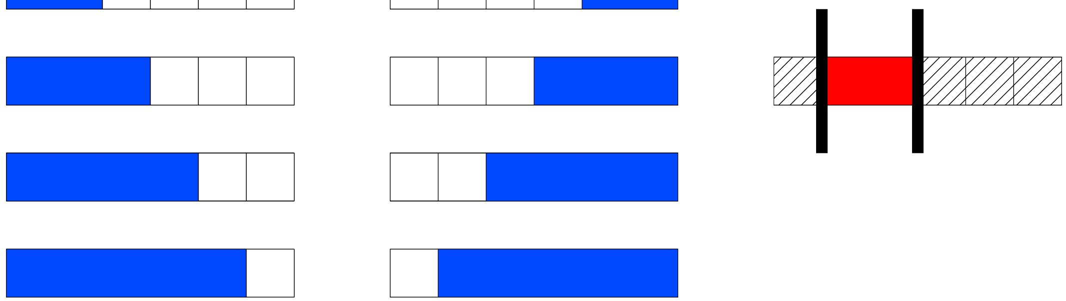
- The zero pattern of w is given by

$$\{j,\ w_j = 0\} = \bigcup_{G \in \mathbf{H}'} G \quad \text{for some } \mathbf{H}' \subseteq \mathbf{H}$$

- **Zero patterns are unions of groups**

Examples of set of groups $\mathbb{H}$

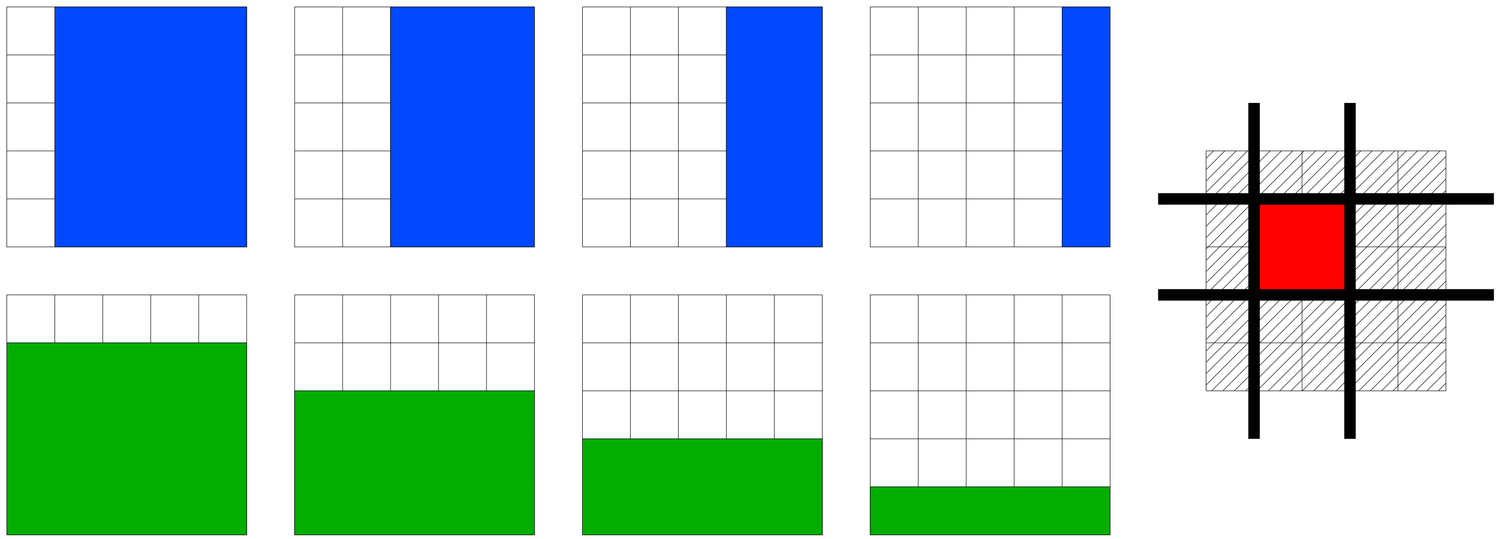
- Selection of contiguous patterns on a sequence, $p = 6$



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

Examples of set of groups $\mathbf{H}$

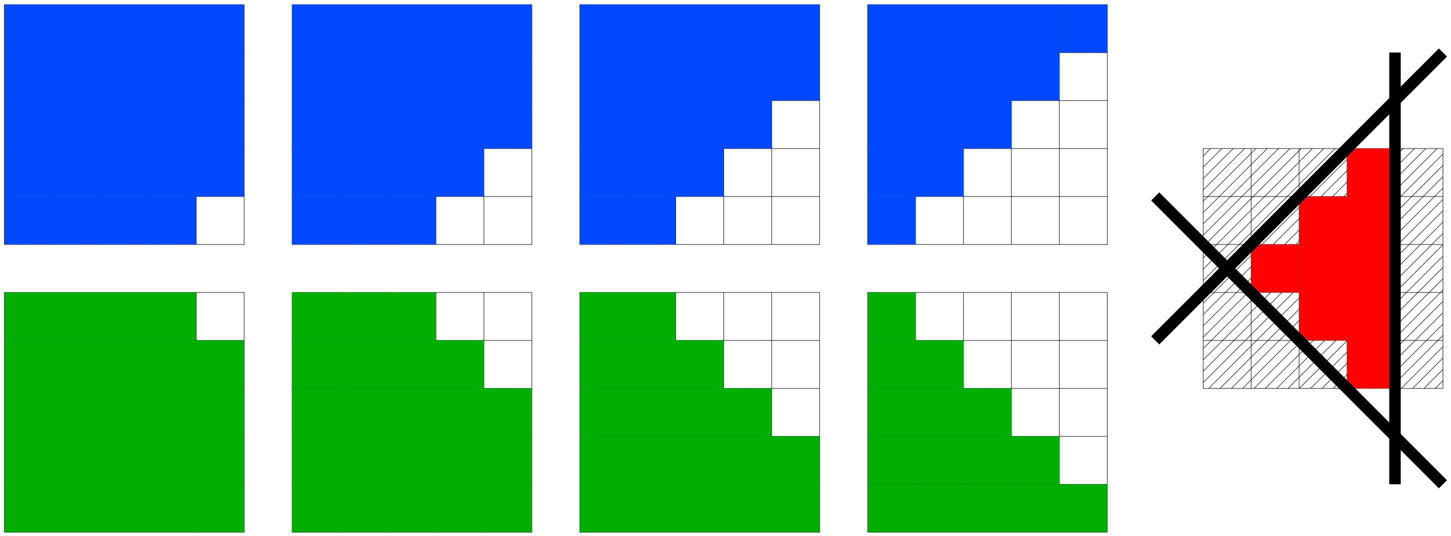
- Selection of rectangles on a 2-D grids, $p = 25$



  - $\mathbf{H}$ 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

Examples of set of groups $\mathbb{H}$

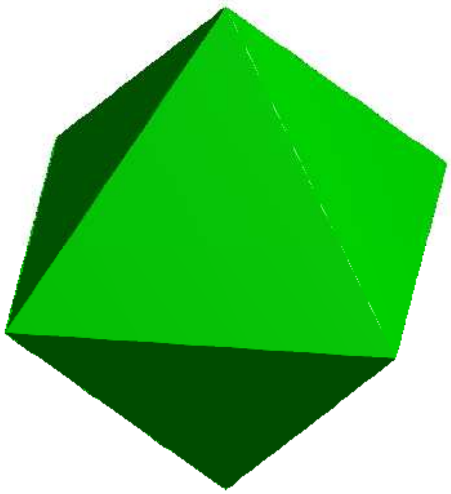
- Selection of diamond-shaped patterns on a 2-D grids, $p = 25$.



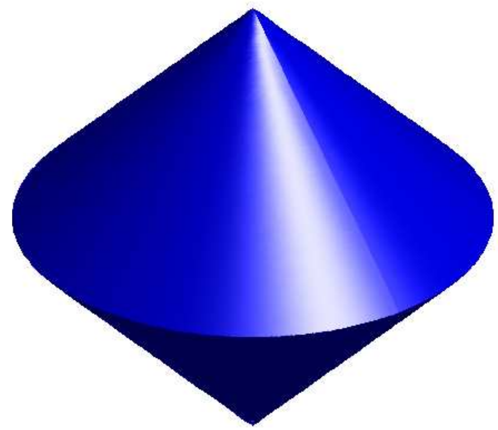
  – It is possible to extend such settings to 3-D space, or more complex topologies

Unit norm balls

Geometric interpretation



$\|w\|_1$



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

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

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

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

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

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

- 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)

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

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

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

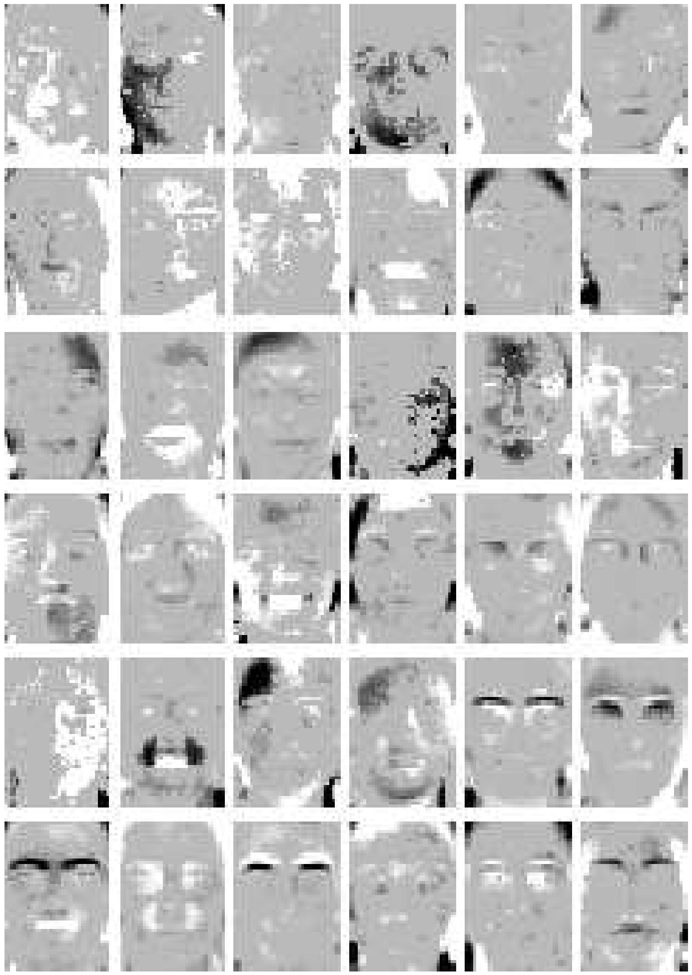
Application to face databases (1/3)

raw data

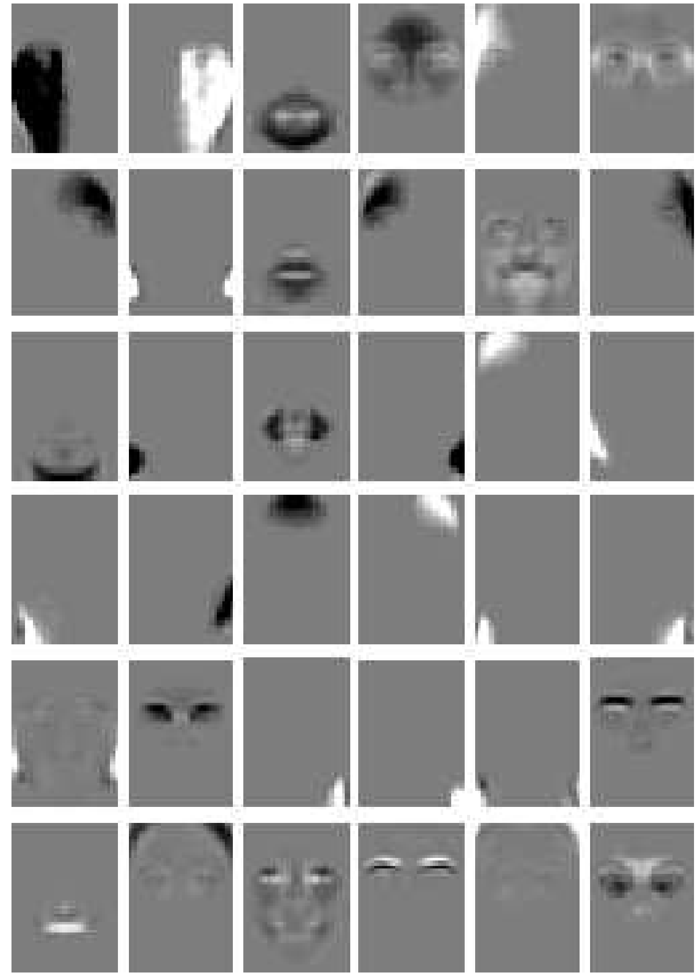
(unstructured) NMF

- NMF obtains partially local features

Application to face databases (2/3)



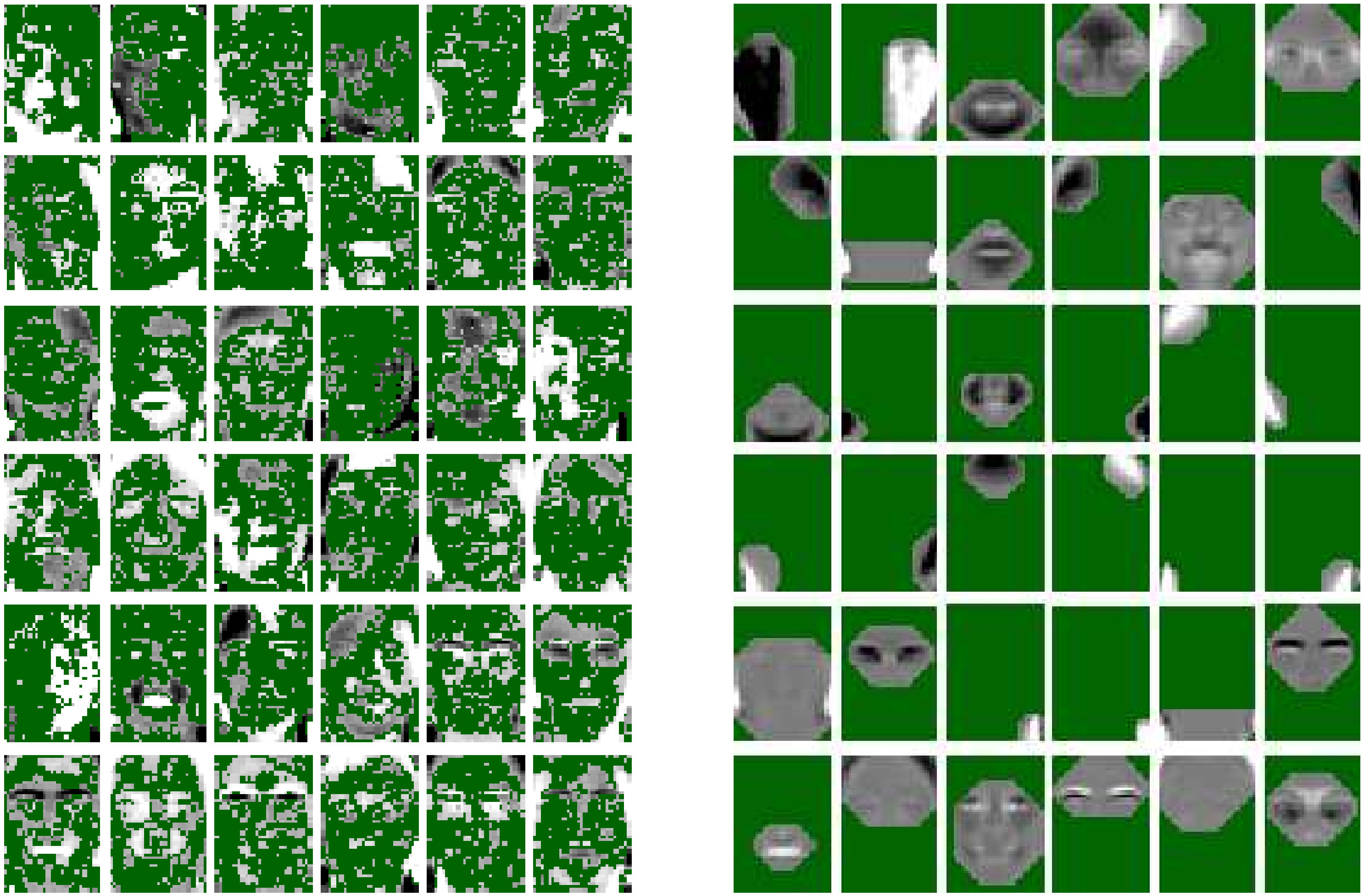
(unstructured) sparse PCA



Structured sparse PCA

- Enforce selection of convex nonzero patterns ⇒ robustness to occlusion

Application to face databases (2/3)

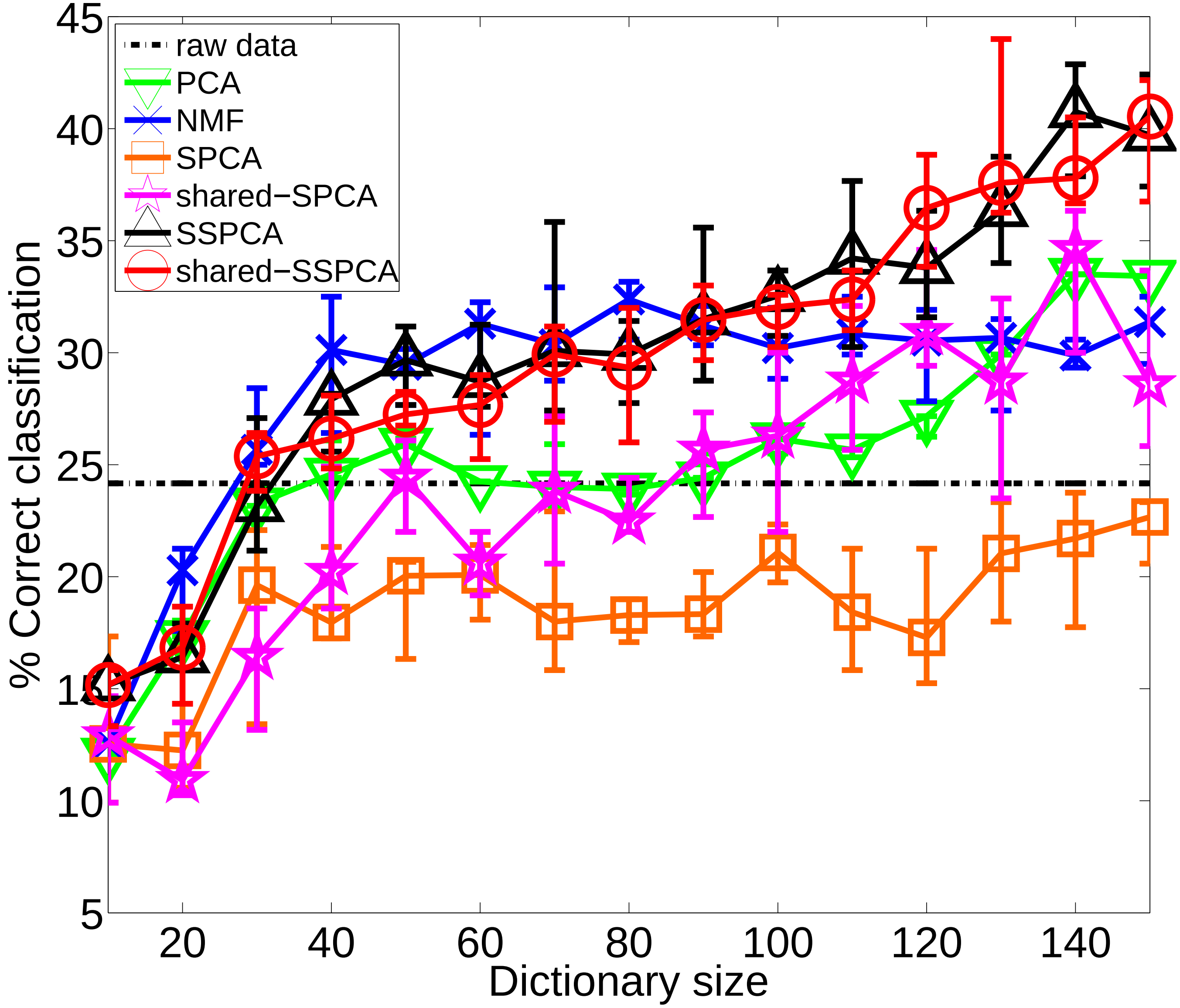


(unstructured) sparse PCA   Structured sparse PCA

- 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:

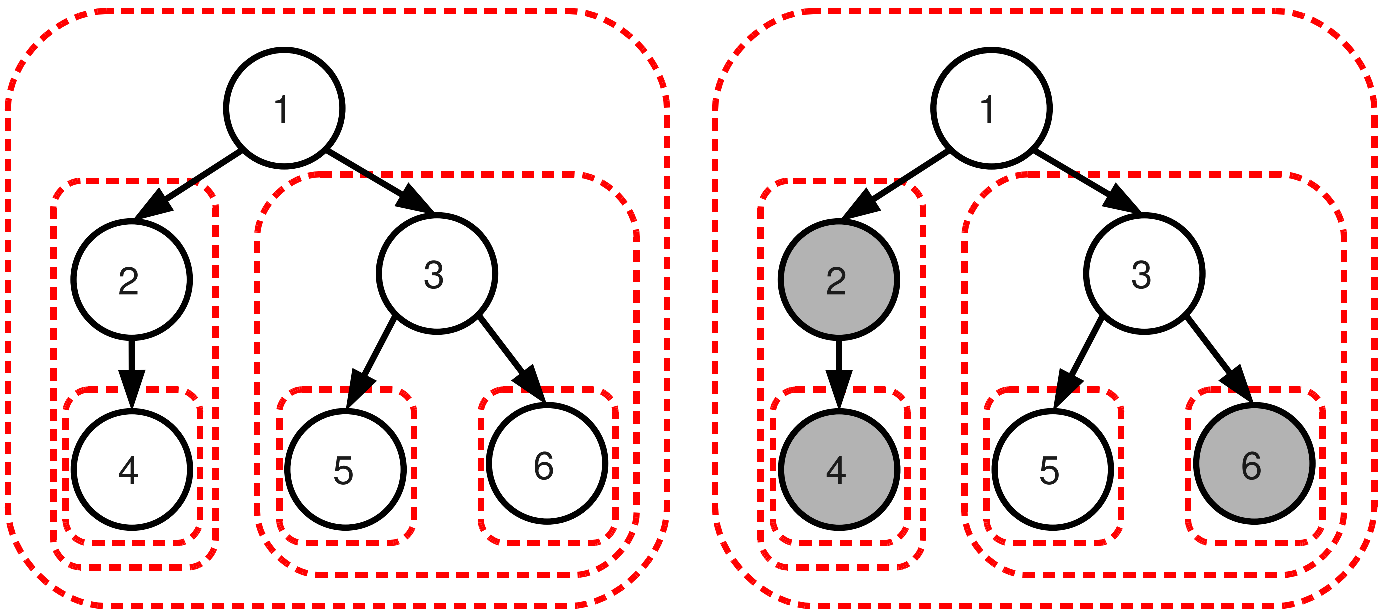
$$\min_{W\in\mathbb{R}^{k\times n},X\in\mathbb{R}^{p\times k}} \frac{1}{n}\sum_{i=1}^{n}\|y^i - Xw^i\|_2^2 + \lambda\Omega(w^i) \text{ s.t. } \forall j,\ \|x^j\|_2 \leq 1.$$

- **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: $\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_2$ where groups G in $\mathbf{H}$ are equal to set of descendants of some nodes in a tree

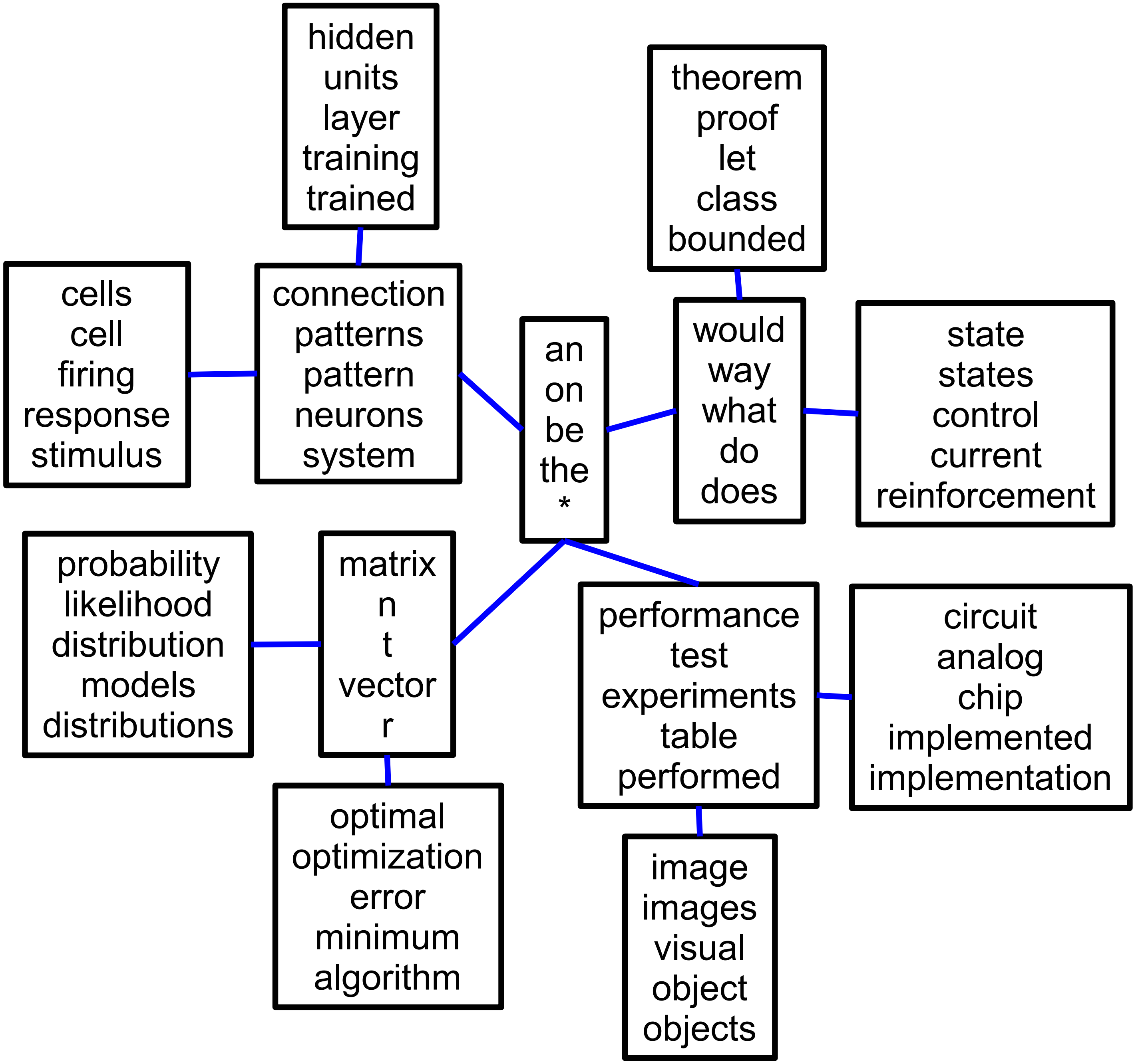


- Variable selected after its ancestors (Zhao et al., 2009; Bach, 2008)

Hierarchical dictionary learning
Modelling of text corpora

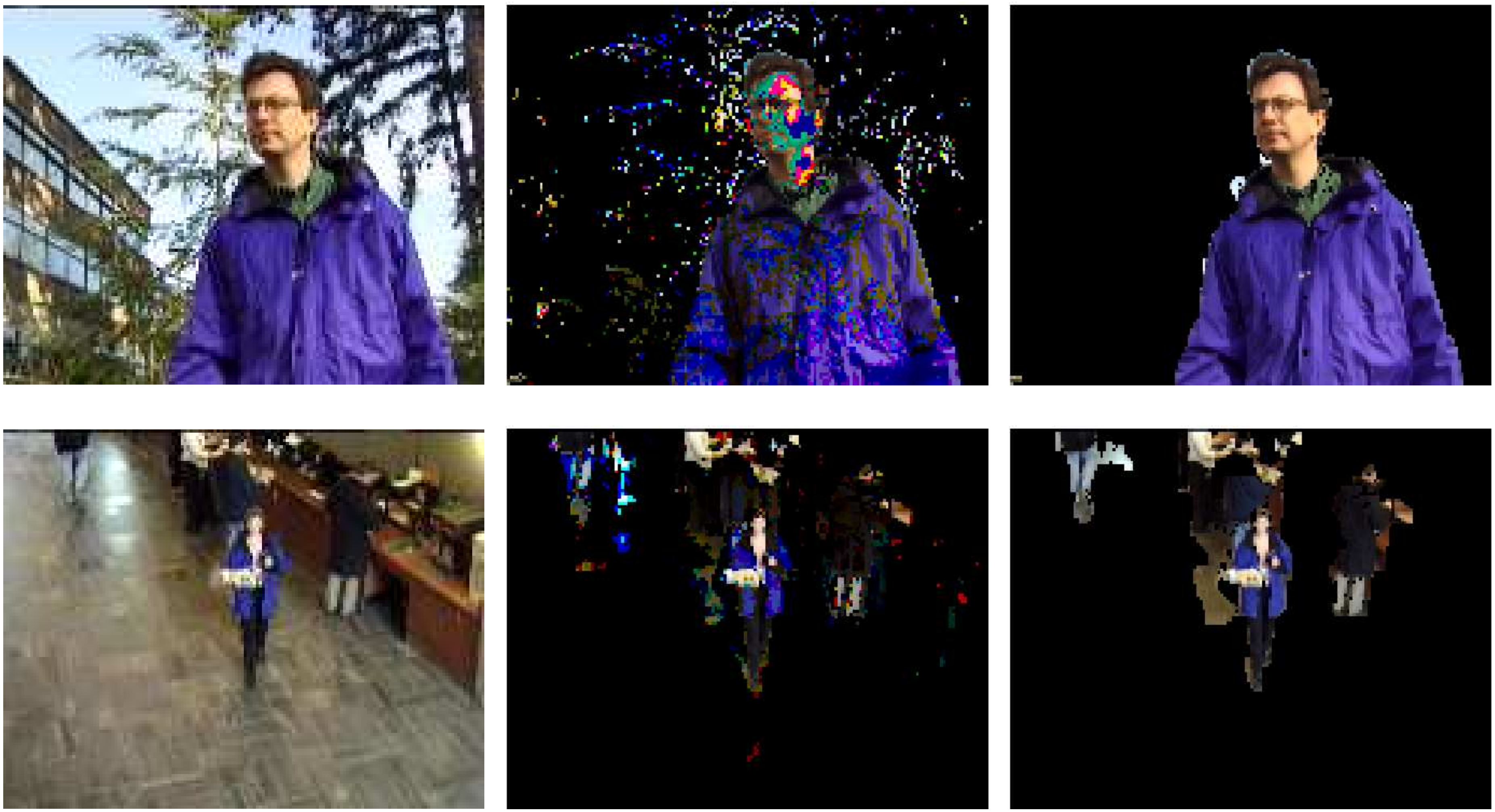
- Each document is modelled through word counts
- Low-rank matrix factorization of word-document matrix
- Probabilistic topic models (Blei et al., 2003)
  - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
  - **Can we achieve similar performance with simple matrix factorization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

Modelling of text corpora - Dictionary tree



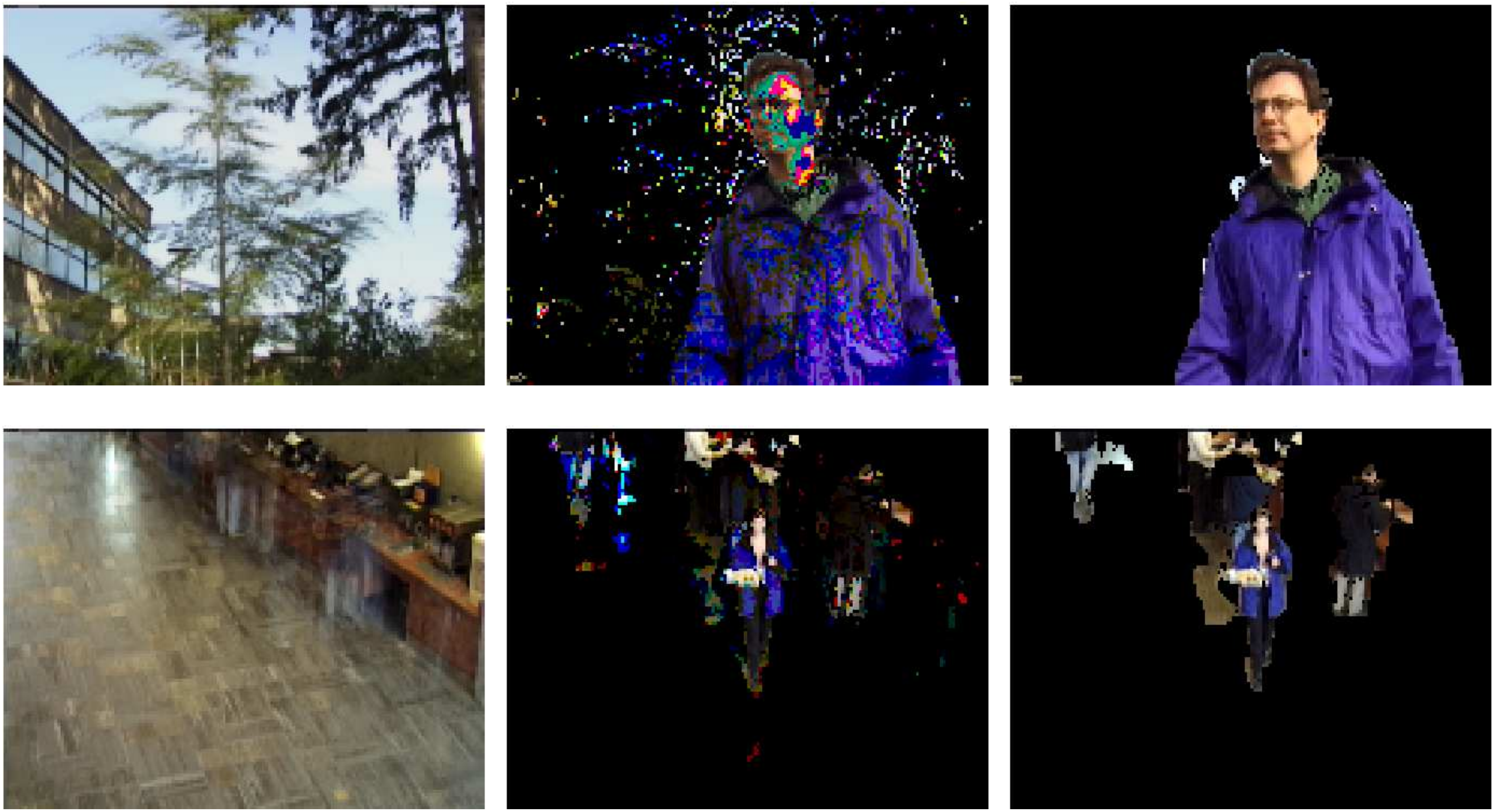
Application to background subtraction (Mairal, Jenatton, Obozinski, and Bach, 2010)

Input | ℓ_1-norm | Structured norm



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

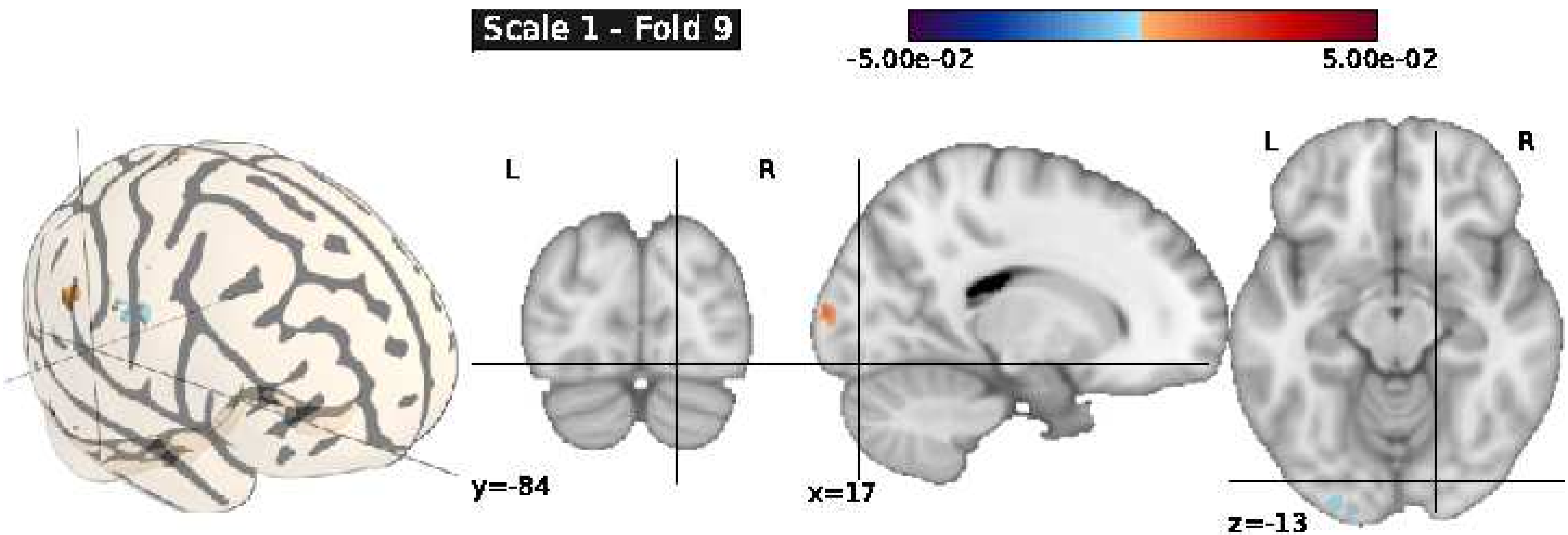
Background | ℓ_1-norm | Structured norm



Application to neuro-imaging

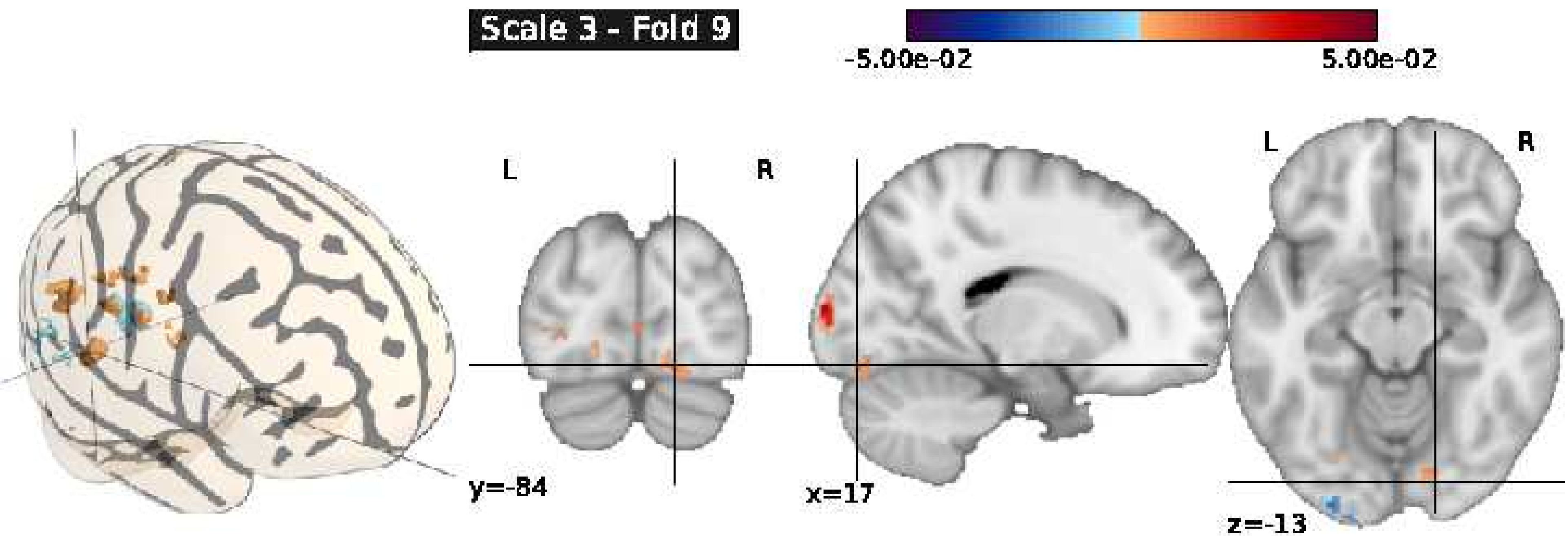
Structured sparsity for fMRI (Jenatton et al., 2011)

- "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)

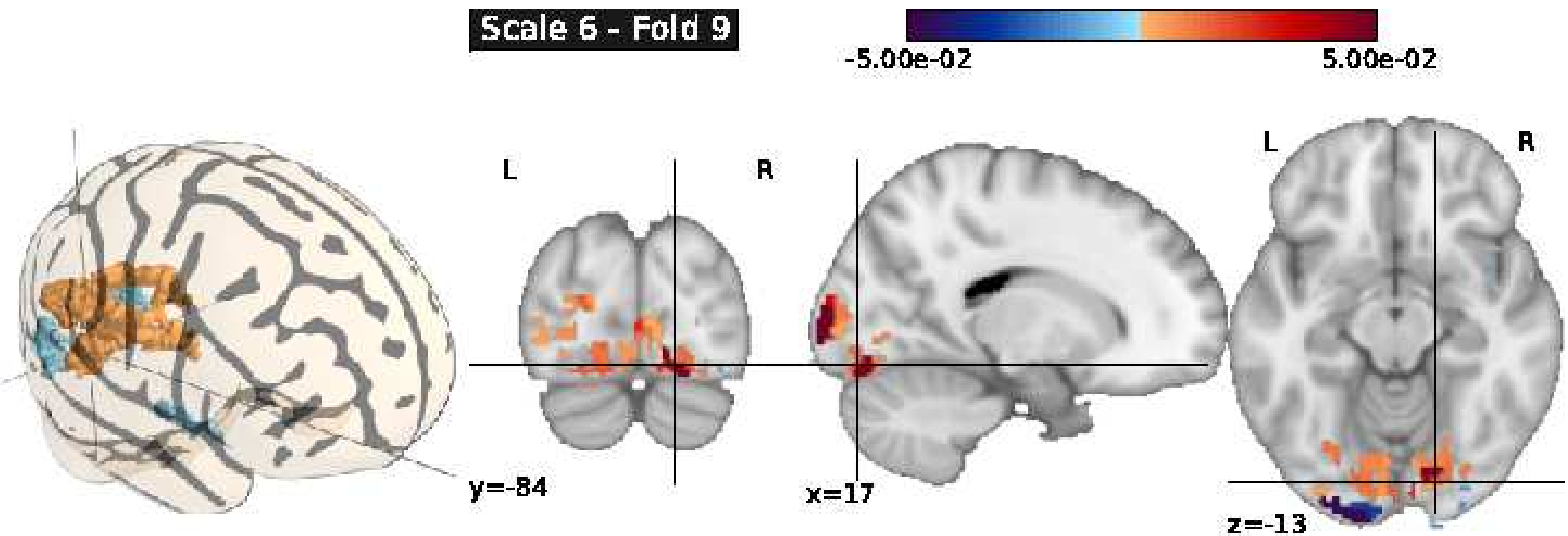
- "Brain reading": prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



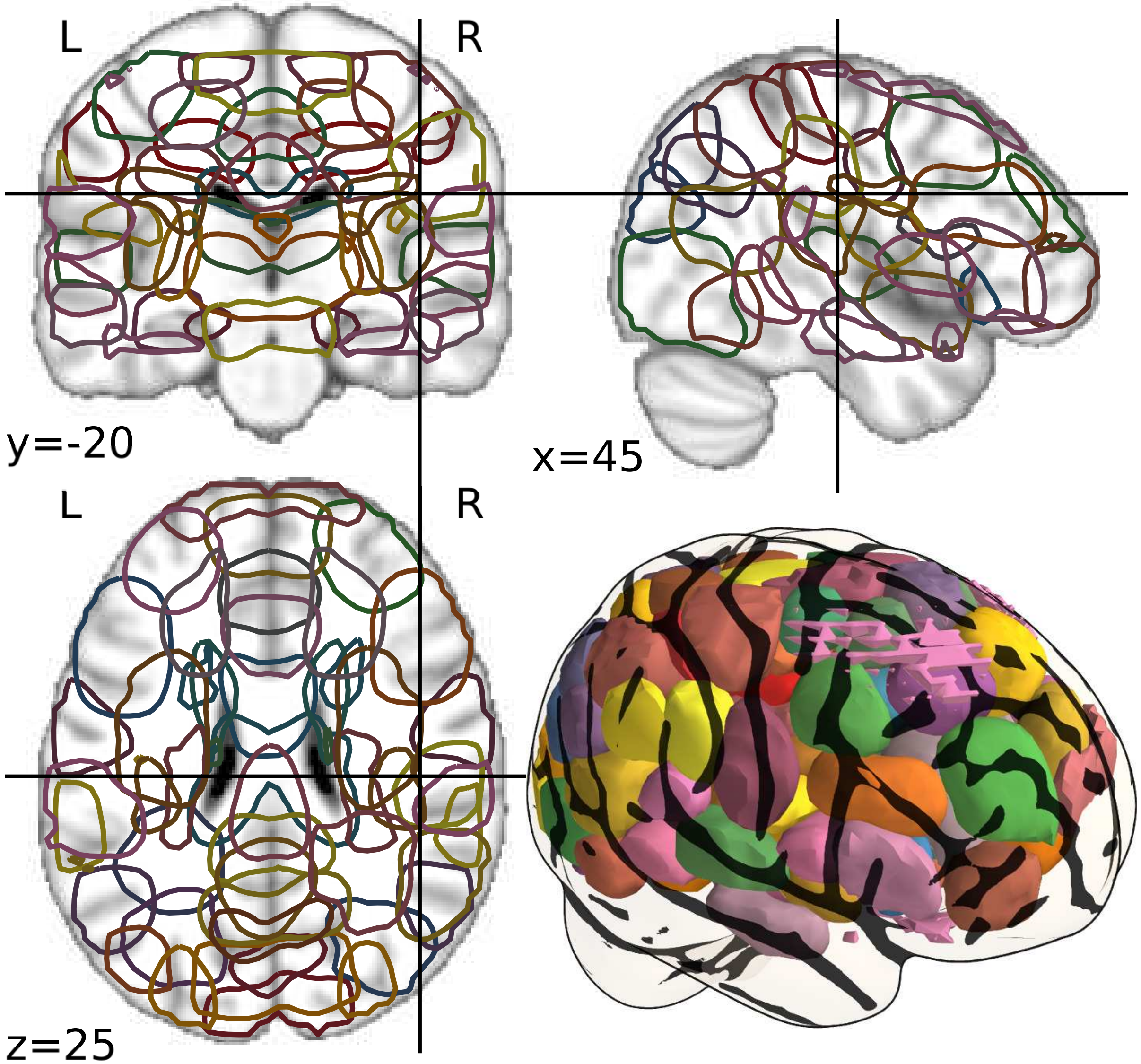
Application to neuro-imaging

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

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



Structured sparse PCA on resting state activity (Varoquaux, Jenatton, Gramfort, Obozinski, Thirion, and Bach, 2010)

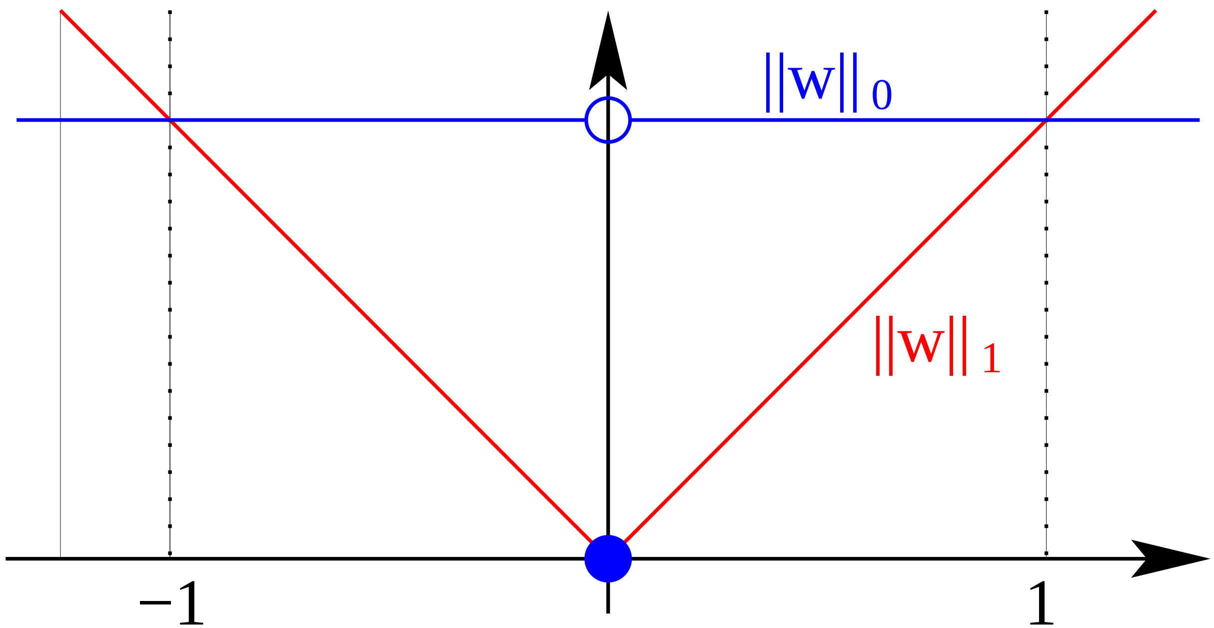


ℓ_1-norm = convex envelope of cardinality of support

- Let $w \in \mathbb{R}^p$. Let $V = \{1, \ldots, p\}$ 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, 2010)

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

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

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: $F : A \mapsto g(\text{Card}(A))$ is submodular if g is concave

Submodular functions (Fujishige, 2005; Bach, 2010)

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

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

- **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

Submodular functions (Fujishige, 2005; Bach, 2010)

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

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

- **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

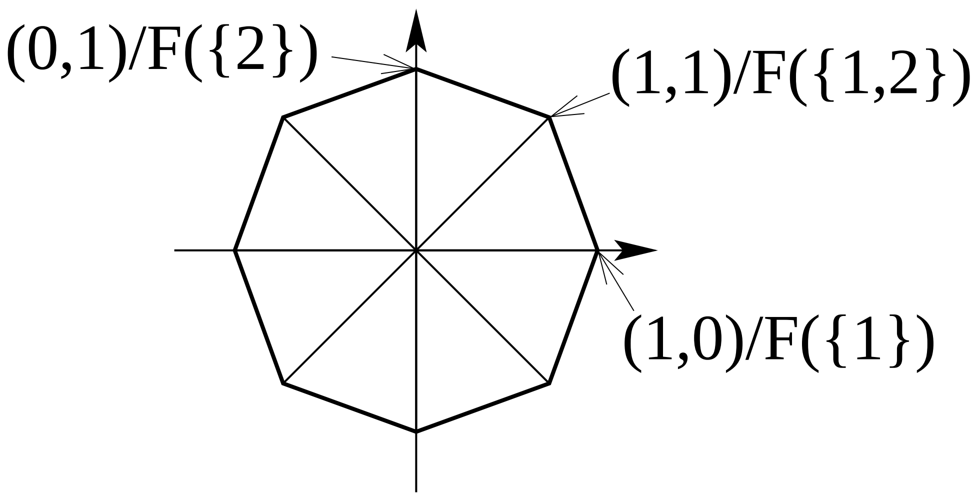
- F is submodular if and only if f is convex (Lovász, 1982)

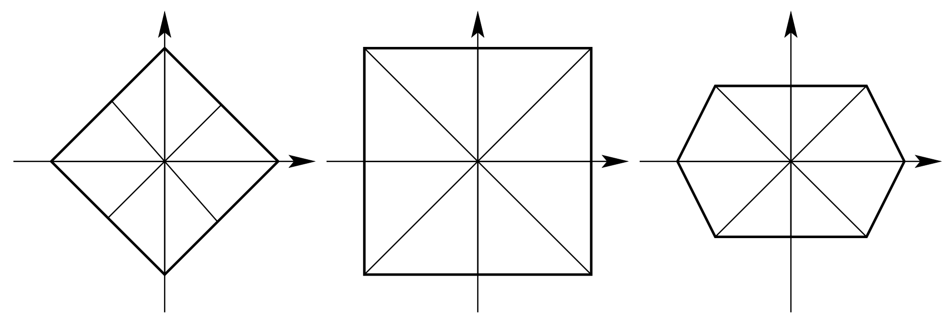
Submodular functions and structured sparsity

- 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

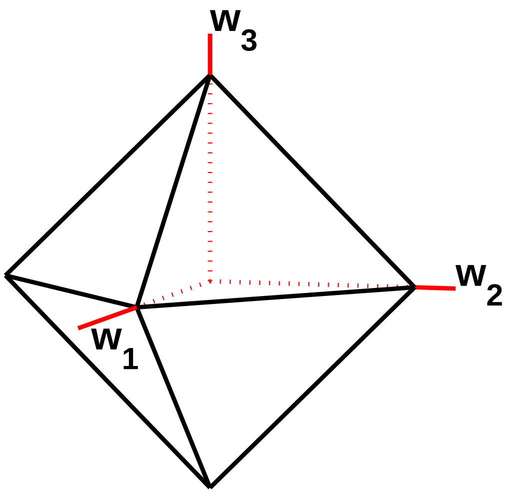
- Let $F : 2^V \to \mathbb{R}$ be a **non-decreasing submodular set-function**
- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\mathrm{Supp}(w))$ on the ℓ_∞-ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F
- **Sparsity-inducing properties**: Ω is a polyhedral norm



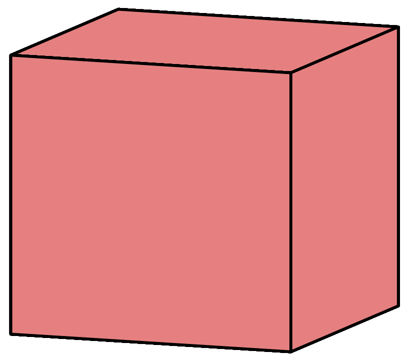


  – A if stable if for all $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$
  – With probability one, stable sets are the only allowed active sets

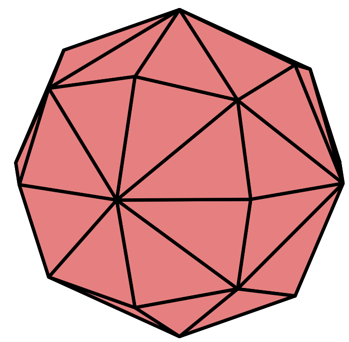
Polyhedral unit balls



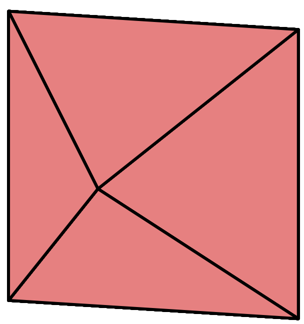
$F(A) = |A|$
$\Omega(w) = \|w\|_1$



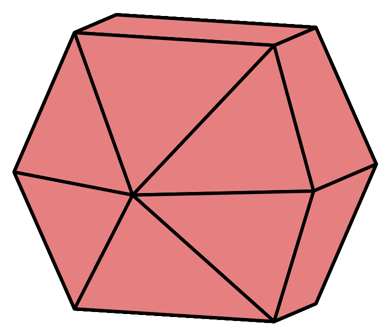
$F(A) = \min\{|A|, 1\}$
$\Omega(w) = \|w\|_\infty$



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

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

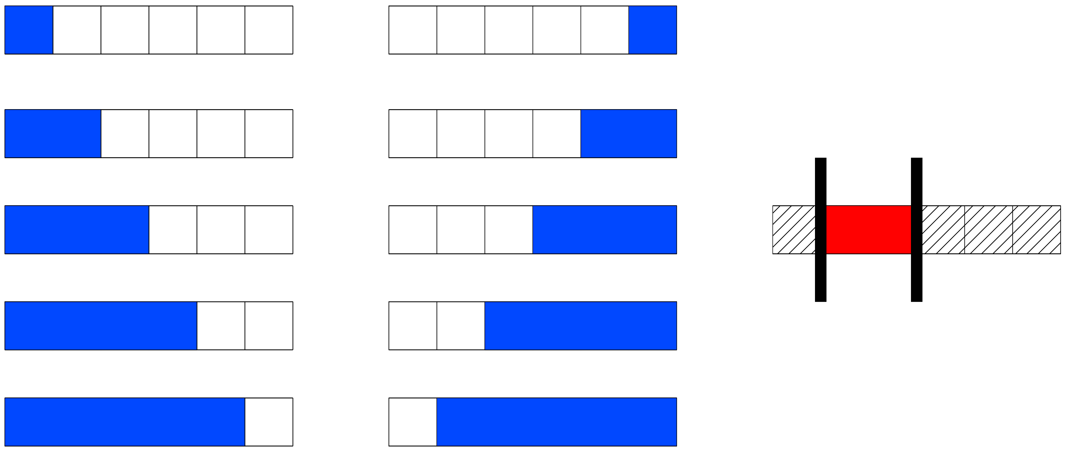
  - ℓ_1-ℓ_∞ norm $\Rightarrow$ sparsity at the group level
  - Some w_G's are set to zero for some groups G

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

  - Justification not only limited to allowed sparsity patterns

Selection of contiguous patterns in a sequence

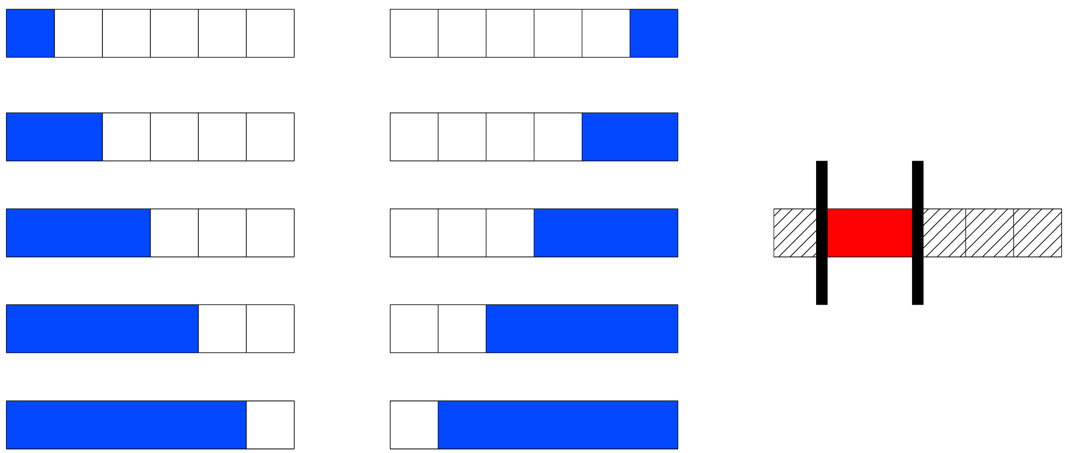
- Selection of contiguous patterns in a sequence



- $\mathbf{H}$ 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

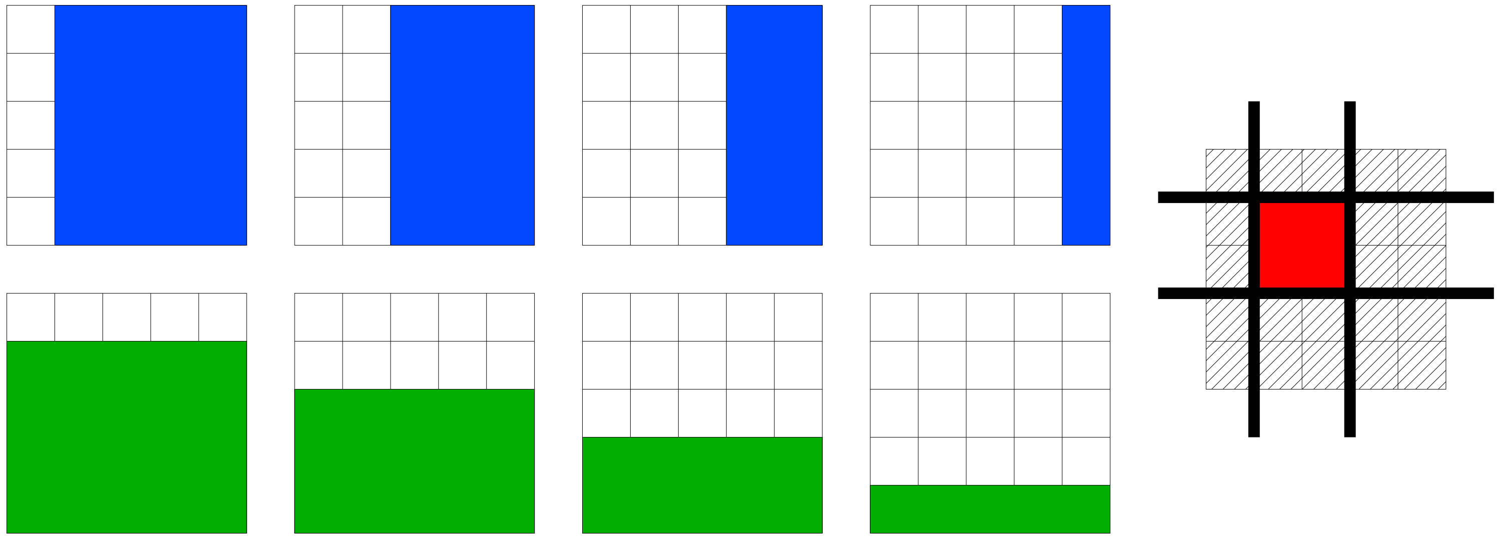


- $\mathbf{H}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a **contiguous pattern**

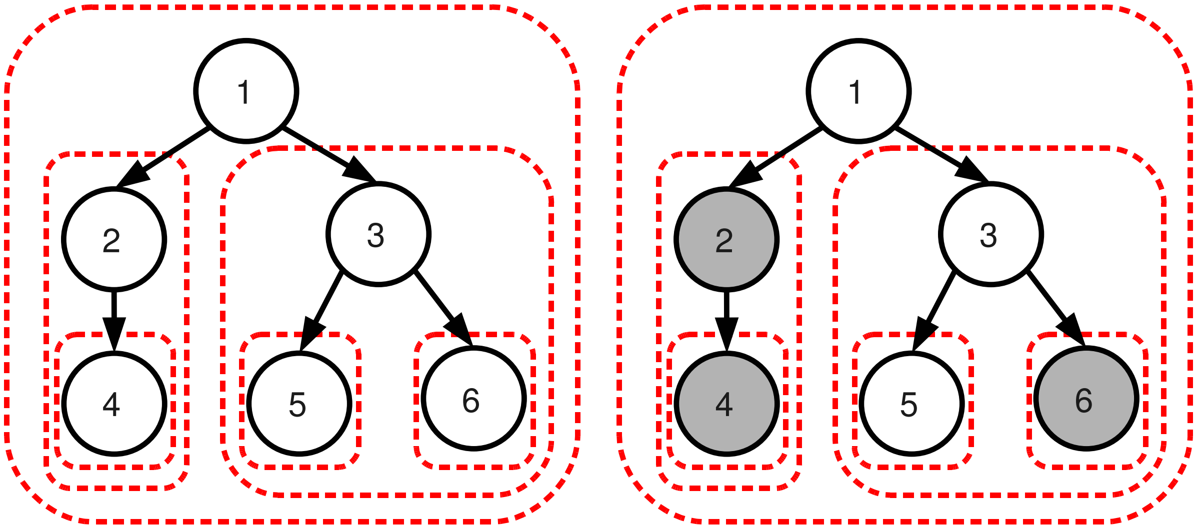
- $\sum_{G \in \mathbf{H}} \|w_G\|_\infty \;\Rightarrow\; F(A) = p - 2 + \text{Range}(A)$ if $A \neq \varnothing$
  - Jump from 0 to $p-1$: tends to include all variables simultaneously
  - Add $\nu|A|$ to smooth the kink: all sparsity patterns are possible
  - **Contiguous patterns are favored (and not forced)**

Extensions of norms with overlapping groups

- Selection of **rectangles** (at any position) in a 2-D grids



- **Hierarchies**



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** $F(A)$

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

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

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 $\pm$ 0.8	-2.6 $\pm$ 0.5	**0.6** $\pm$ **0.0**	**21.8** $\pm$ **0.9**
120	120	40	35.9 $\pm$ 0.8	**2.4** $\pm$ **0.4**	**0.3** $\pm$ **0.0**	**15.8** $\pm$ **1.0**
120	120	20	29.0 $\pm$ 1.0	**9.4** $\pm$ **0.5**	-0.1 $\pm$ 0.0	**6.7** $\pm$ **0.9**
120	120	10	20.4 $\pm$ 1.0	**17.5** $\pm$ **0.5**	-0.2 $\pm$ 0.0	-2.8 $\pm$ 0.8
120	20	20	49.4 $\pm$ 2.0	0.4 $\pm$ 0.5	**2.2** $\pm$ **0.8**	**23.5** $\pm$ **2.1**
120	20	10	49.2 $\pm$ 2.0	0.0 $\pm$ 0.6	1.0 $\pm$ 0.8	**20.3** $\pm$ **2.6**
120	20	6	43.5 $\pm$ 2.0	**3.5** $\pm$ **0.8**	**0.9** $\pm$ **0.6**	**24.4** $\pm$ **3.0**
120	20	4	41.0 $\pm$ 2.1	**4.8** $\pm$ **0.7**	-1.3 $\pm$ 0.5	**25.1** $\pm$ **3.5**

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"
- **Proposition**: (P) is equivalent to $\min_{A \subset V} \lambda F(A) - \sum_{j \in A} |v_j|$ with minimum-norm-point algorithm
  - Possible complexity bound $O(p^6)$, but empirically $O(p^2)$ (or more)
  - Faster algorithm for special case (Mairal et al., 2010)

Proximal methods for Lovász extensions

- **Proposition** (Chambolle and Darbon, 2009): let w^* be the solution of $\min_{w\in\mathbb{R}^p} \frac{1}{2}\|w-v\|_2^2 + \lambda f(w)$. Then the solutions of

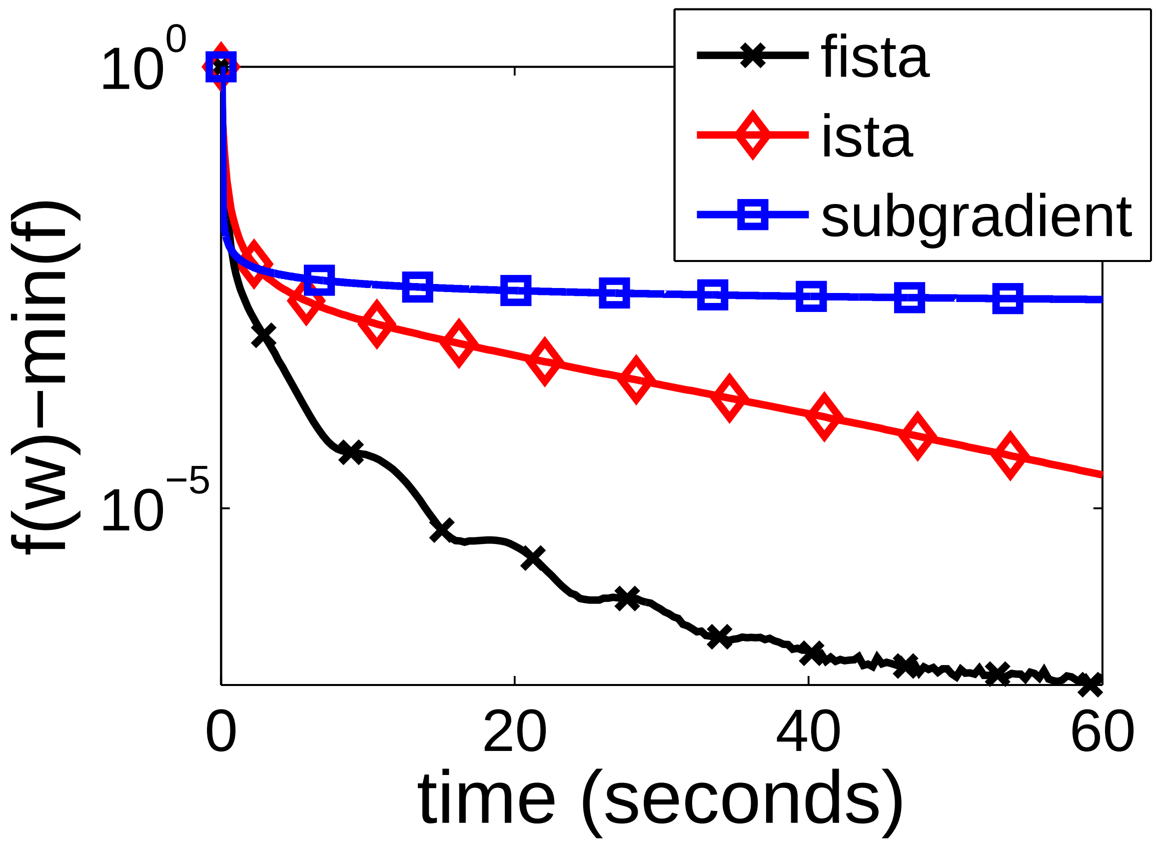
$$\min_{A\subset V} \lambda F(A) + \sum_{j\in A}(\alpha - v_j)$$

are the sets A^α such that $\{w^* > \alpha\} \subset A^\alpha \subset \{w^* \geqslant \alpha\}$

- **Parametric submodular function optimization**
  - General decomposition strategy for $f(|w|)$ and $f(w)$ (Groenevelt, 1991)
  - Efficient only when submodular minimization is efficient
  - Otherwise, minimum-norm-point algorithm (a.k.a. Frank Wolfe) is preferable

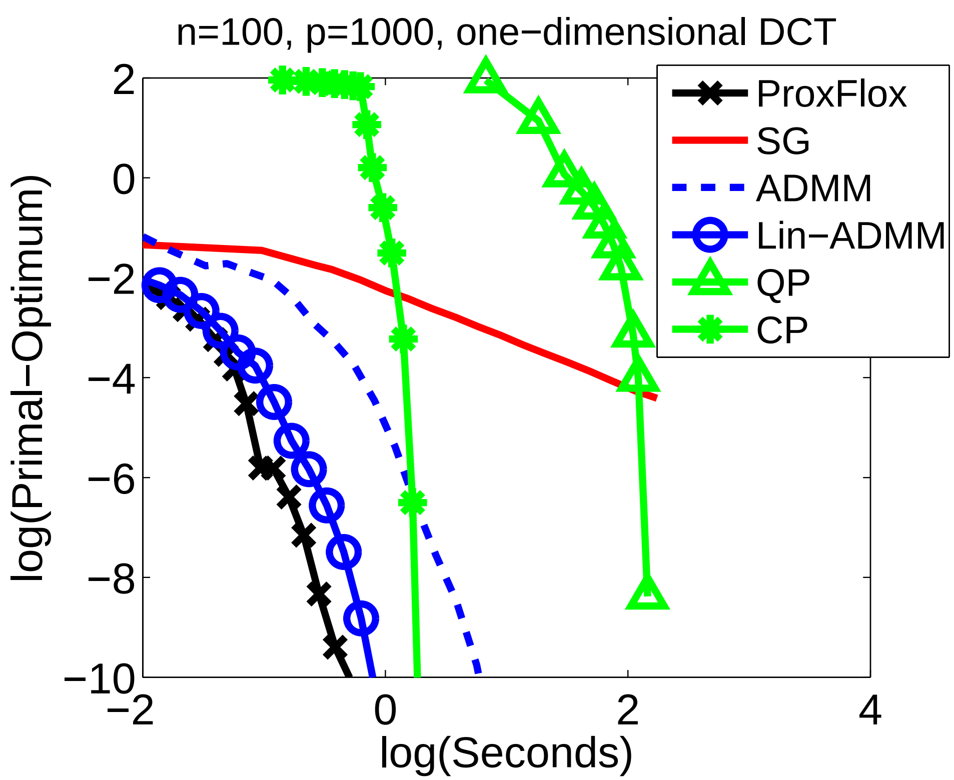
Comparison of optimization algorithms

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



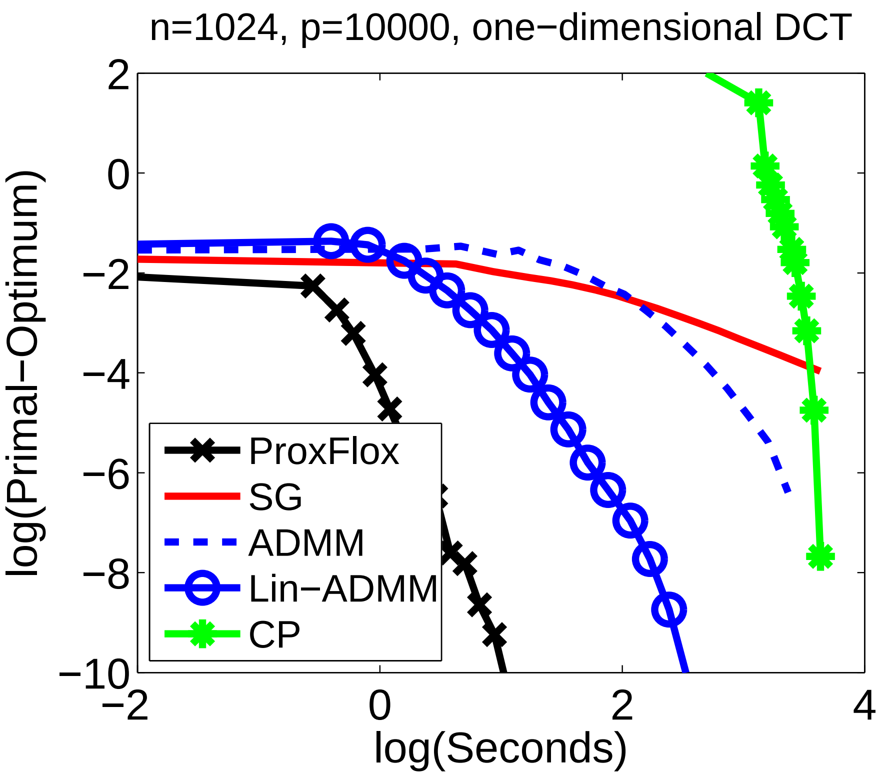
Comparison of optimization algorithms (Mairal, Jenatton, Obozinski, and Bach, 2010)
Small scale

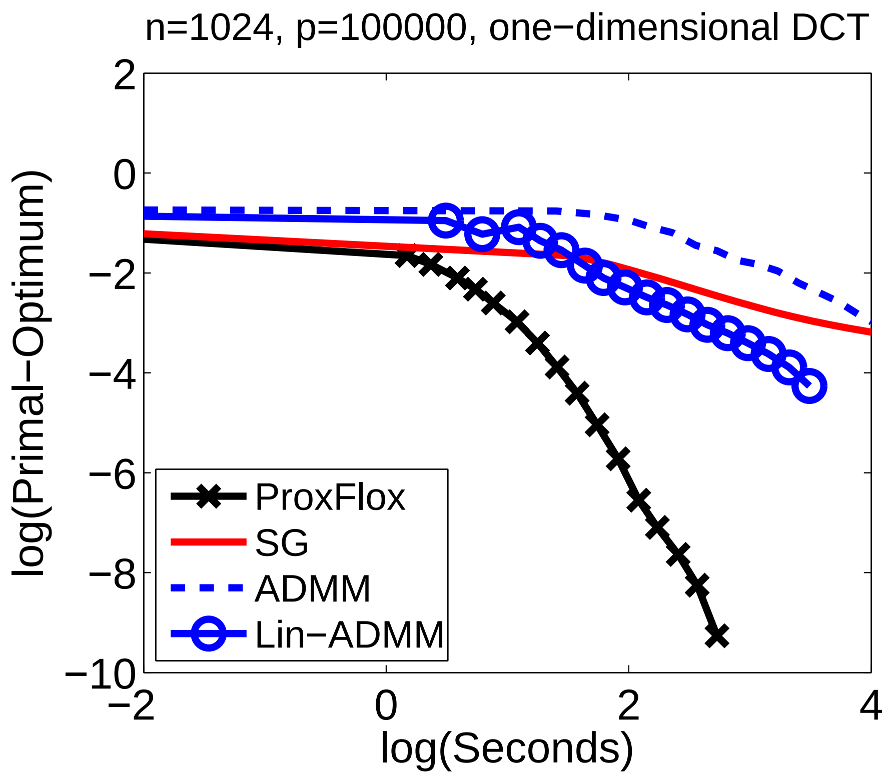
- Specific norms which can be implemented through network flows



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

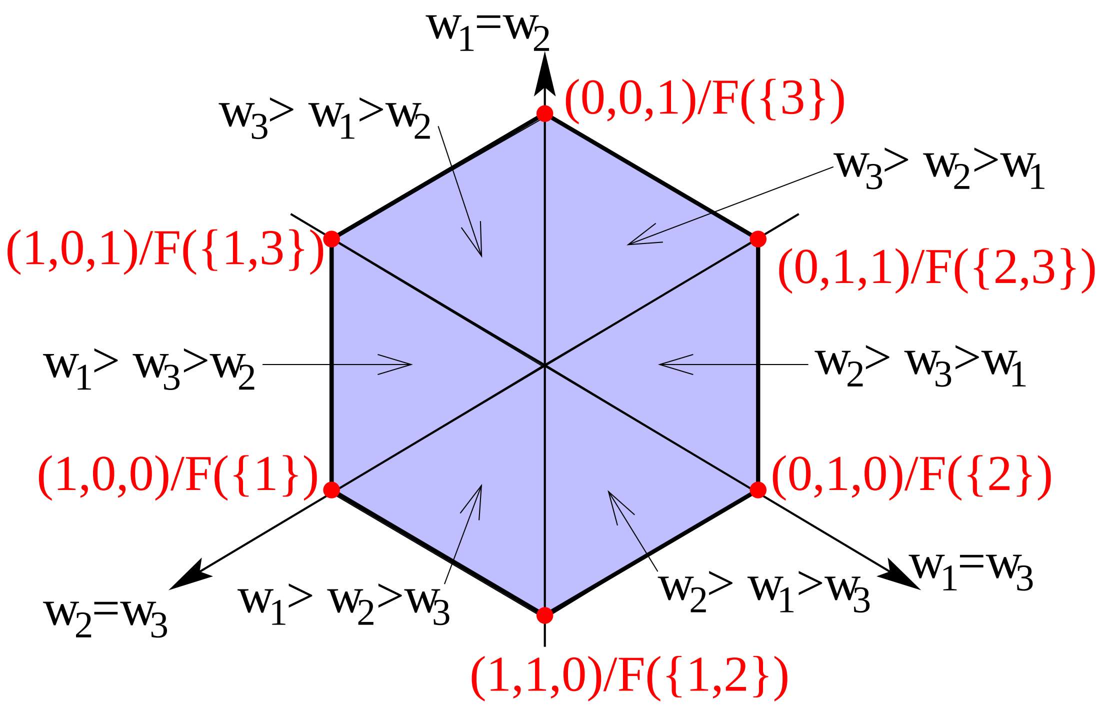
Symmetric submodular functions (Bach, 2011)

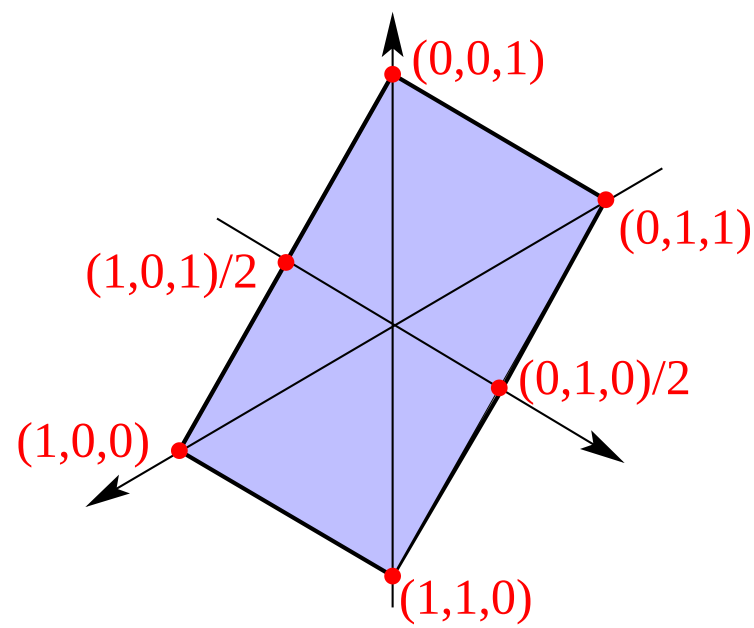
- Let $F : 2^V \to \mathbb{R}$ be a **symmetric submodular set-function**
- **Proposition**: The Lovász extension $f(w)$ is the convex envelope of the function $w \mapsto \max_{\alpha \in \mathbb{R}} F(\{w \geqslant \alpha\})$ on the set $[0,1]^p + \mathbb{R}1_V = \{w \in \mathbb{R}^p,\ \max_{k \in V} w_k - \min_{k \in V} w_k \leqslant 1\}$.

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

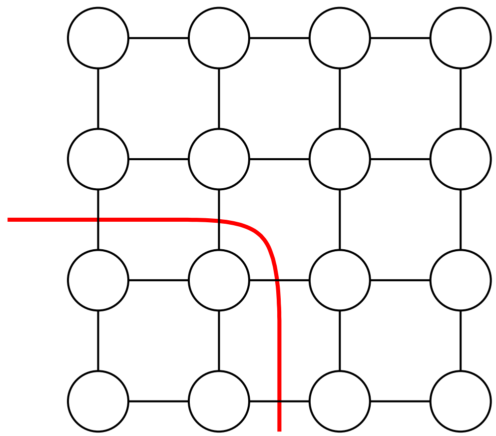


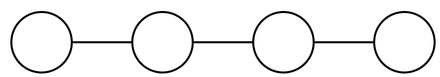


Symmetric submodular functions - Examples

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

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

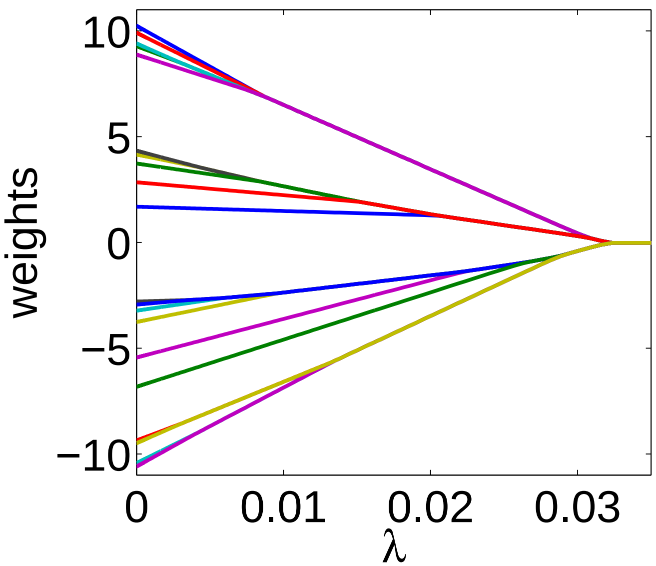




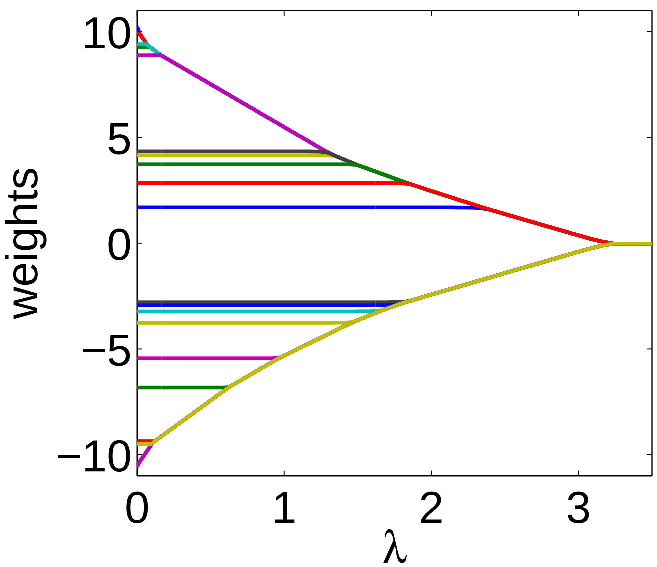
  - NB: graph may be directed

Symmetric submodular functions - Examples

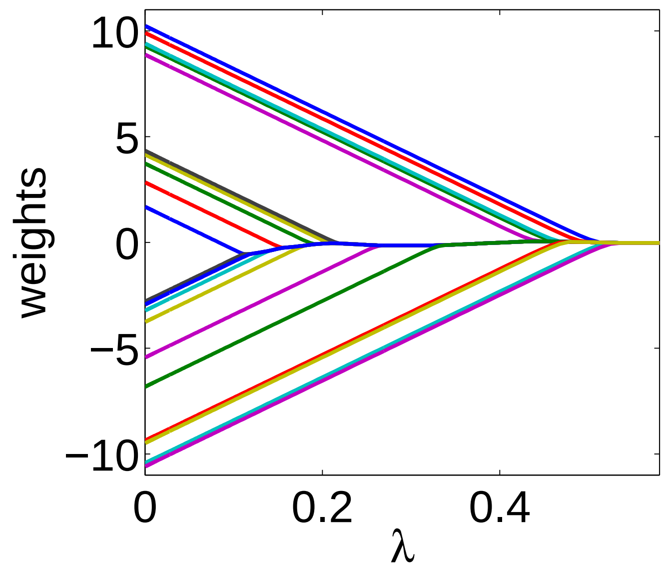
- **From** $F(A)$ **to** $\Omega(w)$: provides new sparsity-inducing norms
  - $F(A) = g(\mathrm{Card}(A)) \Rightarrow$ priors on the size and numbers of clusters



$|A|(p - |A|)$



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



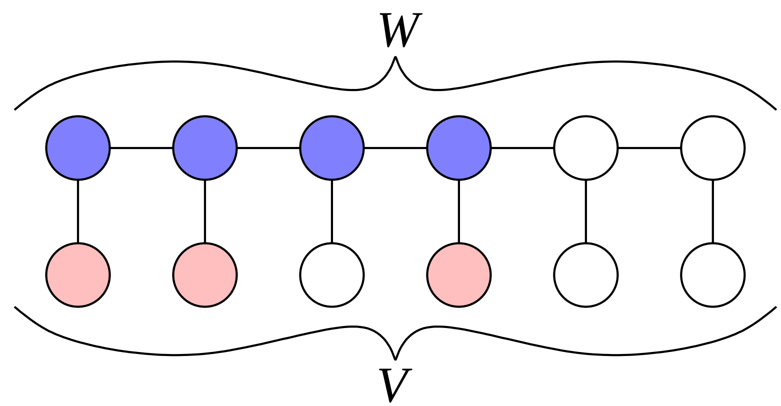
$\max\{|A|, p - |A|\}$

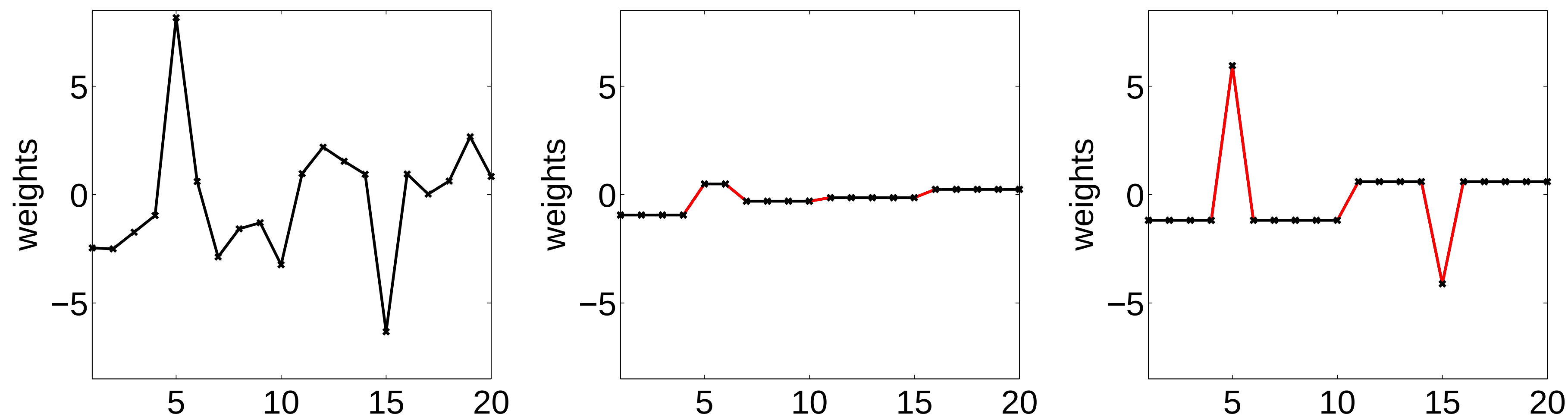
  - Convex formulations for clustering (Hocking, Joulin, Bach, and Vert, 2011)

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)

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

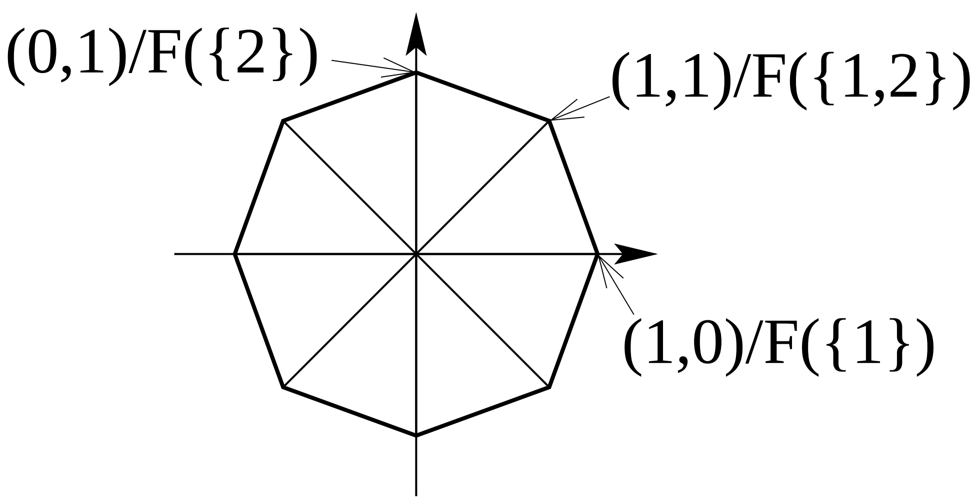


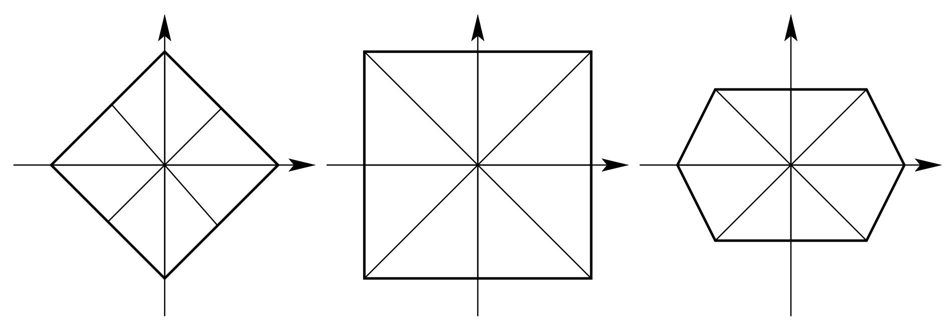


ℓ_q-relaxation of combinatorial penalties (Obozinski and Bach, 2011)

- **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





From ℓ_∞ to ℓ_2

- Variational formulations for subquadratic norms (Bach et al., 2011)

$$\Omega(w) = \min_{\eta \in \mathbb{R}_+^p} \frac{1}{2} \sum_{j=1}^{p} \frac{w_j^2}{\eta_j} + \frac{1}{2} g(\eta) = \min_{\eta \in H} \sqrt{\sum_{j=1}^{p} \frac{w_j^2}{\eta_j}}$$

where g is a convex homogeneous and $H = \{\eta, g(\eta) \leqslant 1\}$

– Often used for computational reasons (Lasso, group Lasso)
– May also be used to define a norm (Micchelli et al., 2011)

From ℓ_∞ to ℓ_2

- Variational formulations for subquadratic norms (Bach et al., 2011)

$$\Omega(w) = \min_{\eta \in \mathbb{R}^p_+} \frac{1}{2} \sum_{j=1}^p \frac{w_j^2}{\eta_j} + \frac{1}{2} g(\eta) = \min_{\eta \in H} \sqrt{\sum_{j=1}^p \frac{w_j^2}{\eta_j}}$$

  where g is a convex homogeneous and $H = \{\eta, g(\eta) \leqslant 1\}$

  – Often used for computational reasons (Lasso, group Lasso)
  – May also be used to define a norm (Micchelli et al., 2011)

- If F is a nondecreasing submodular function with Lovász extension f

  – Define $\Omega_2(w) = \min_{\eta \in \mathbb{R}^p_+} \frac{1}{2} \sum_{j=1}^p \frac{w_j^2}{\eta_j} + \frac{1}{2} f(\eta)$
  – Is it the convex relaxation of some natural function?

ℓ_q-relaxation of submodular penalties (Obozinski and Bach, 2011)

- F a nondecreasing submodular function with Lovász extension f
- Define $\Omega_q(w) = \min_{\eta \in \mathbb{R}_+^p} \frac{1}{q} \sum_{i \in V} \frac{|w_i|^q}{\eta_i^{q-1}} + \frac{1}{r} f(\eta)$ with $\frac{1}{q} + \frac{1}{r} = 1$.
- **Proposition 1**: Ω_q is the convex envelope of $w \mapsto F(\mathrm{Supp}(w)) \|w\|_q$
- **Proposition 2**: Ω_q is the homogeneous convex envelope of $w \mapsto \frac{1}{r} F(\mathrm{Supp}(w)) + \frac{1}{q} \|w\|_q^q$
- **Jointly penalizing and regularizing**
  - Special cases $q = 1$, $q = 2$ and $q = \infty$

Some simple examples

	F	Ω_q
	$\|A\|$	$\|w\|_1$
	$1_{\{A \neq \varnothing\}}$	$\|w\|_q$
If $\mathbf{H}$ is a partition of V:	$\sum_{B \in \mathbf{H}} 1_{\{A \cap B \neq \varnothing\}}$	$\sum_{B \in \mathbf{H}} \|w_B\|_q$

- Recover results of Bach (2010) when $q = \infty$ and F submodular
- However
  - when $\mathbf{H}$ is not a partition and $q < \infty$, Ω_q is not in general an ℓ_1/ℓ_q-norm !
  - F does not need to be submodular

$\Rightarrow$ **New norms**

ℓ_q-relaxation of combinatorial penalties (Obozinski and Bach, 2011)

- F any strictly positive set-function (with potentially infinite values)
- **Jointly penalizing and regularizing**. Two formulations:
  - homogeneous convex envelope of $w \mapsto F(\mathrm{Supp}(w)) + \|w\|_q^q$
  - convex envelope of $w \mapsto F(\mathrm{Supp}(w))\|w\|_q$
- **Proposition**: These envelopes are equal to a constant times a norm $\Omega_q^F = \Omega_q$ defined through its dual norm
  - its dual norm is equal to $\boxed{(\Omega_q)^*(s) = \max_{A \subset V} \frac{\|s_A\|_r}{F(A)^{1/r}}}$, with $\frac{1}{q}+\frac{1}{r} = 1$
- Three-line proof

ℓ_q-relaxation of combinatorial penalties
Proof

- Denote $\Theta(w) = \|w\|_q F(\text{Supp}(w))^{1/r}$, and compute its Fenchel conjugate:

$$\begin{aligned}
\Theta^*(s) &= \max_{w\in\mathbb{R}^p} w^\top s - \|w\|_q F(\text{Supp}(w))^{1/r} \\
&= \max_{A\subset V} \max_{w_A\in(\mathbb{R}^*)^A} w_A^\top s_A - \|w_A\|_q F(A)^{1/r} \\
&= \max_{A\subset V} \iota_{\{\|s_A\|_r\leqslant F(A)^{1/r}\}} = \iota_{\{\Omega_q^*(s)\leqslant 1\}},
\end{aligned}$$

  where $\iota_{\{s\in S\}}$ is the indicator of the set S

- Consequence: If F is submodular and $q = +\infty$, $\Omega(w) = f(|w|)$

How tight is the relaxation?
What information of F is kept after the relaxation?

- When F is submodular and $q = \infty$
  - the Lovász extension $f = \Omega_\infty$ is said to "extend" F because $\Omega_\infty^F(1_A) = f(1_A) = F(A)$
- In general we can still consider the function : $G(A) \triangleq \Omega_\infty^F(1_A)$
  - Do we have $G(A) = F(A)$?
  - How is G related to F?
  - What is the norm Ω_∞^G which is associated with G?

Lower combinatorial envelope

- Given a function $F : 2^V \to \mathbb{R}$, define its *lower combinatorial envelope* as the function G given by

$$G(A) = \max_{s \in P(F)} s(A)$$

with $P(F) = \{s \in \mathbb{R}^p, \ \forall A \subset V, \ s(A) \leq F(A)\}$.

- **Lemma 1** (Idempotence)
  - $P(F) = P(G)$
  - G is its own lower combinatorial envelope
  - For all $q \geq 1$, $\Omega_q^F = \Omega_q^G$

- **Lemma 2** (Extension property)

$$\Omega_\infty^F(1_A) = \max_{(\Omega_\infty^F)^*(s) \leq 1} 1_A^\top s = \max_{s \in P(F)} s^\top 1_A = G(A)$$

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

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

- **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)
  - Links between norm Ω, support $\mathrm{Supp}(w)$, and design X (see, e.g., Grave, Obozinski, and Bach, 2011)
  - Achieving $\log p = O(n)$ algorithmically (Bach, 2008)

Conclusion

- **Submodular functions to encode discrete structures**
  - Structured sparsity-inducing norms
- **Convex optimization for submodular function optimization**
  - Approximate optimization using classical iterative algorithms
- **Future work**
  - Primal-dual optimization
  - Going beyond linear programming

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, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. Technical Report 00613125, HAL, 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.

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.

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.

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.

U. Feige, V.S. Mirrokni, and J. Vondrak. Maximizing non-monotone submodular functions. In *Proc. Symposium on Foundations of Computer Science*, pages 461–471. IEEE Computer Society, 2007.

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.

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

E. Grave, G. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. *Arxiv preprint arXiv:1109.1990*, 2011.

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.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

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.

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.

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.

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.

H. Nagamochi and T. Ibaraki. A note on minimizing submodular functions. *Information Processing Letters*, 67(5):239–244, 1998.

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.

G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions–i. *Mathematical Programming*, 14(1):265–294, 1978.

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, 2011.

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.

M. Queyranne. Minimizing symmetric submodular functions. *Mathematical Programming*, 82(1):3–12, 1998.

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

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`.

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

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

G. Varoquaux, R. Jenatton, A. Gramfort, G. Obozinski, B. Thirion, and F. Bach. Sparse structured dictionary learning for brain resting-state activity modeling. In *NIPS Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*, 2010.

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.