Learning with Submodular Functions: A Convex Optimization Perspective

Francis Bach

Sierra project-team, INRIA - Ecole Normale Supérieure

Thanks to R. Jenatton, J. Mairal, G. Obozinski

December 2011

Convex optimization with combinatorial structure

- **Supervised learning**
  - 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

- **Submodular functions**
  - Links with convexity through Lovász extension
  - Optimization on submodular polyhedra
- **Structured sparsity-inducing norms**
  - Relaxation of the penalization of supports
  - Examples
  - Unified algorithms and analysis
- **Approximate submodular function minimization**

Outline

- **Submodular functions**
  - Links with convexity through Lovász extension
  - Optimization on submodular polyhedra
- **Structured sparsity-inducing norms**
  - Relaxation of the penalization of supports
  - Examples
  - Unified algorithms and analysis
- **Approximate submodular function minimization**
- (for more details, see tutorial / technical report on web page)

Submodular functions

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

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

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

Submodular functions

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

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

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

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

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

Submodular functions - Examples

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

Submodular functions - Lovász extension

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

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

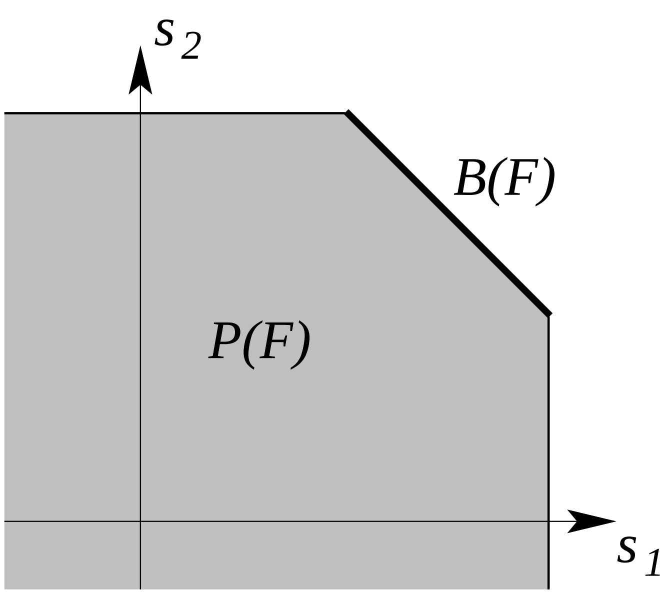
  – If $w = 1_A$, $f(w) = F(A) \Rightarrow$ extension from $\{0,1\}^p$ to $\mathbb{R}^p$
  – f is piecewise affine and positively homogeneous

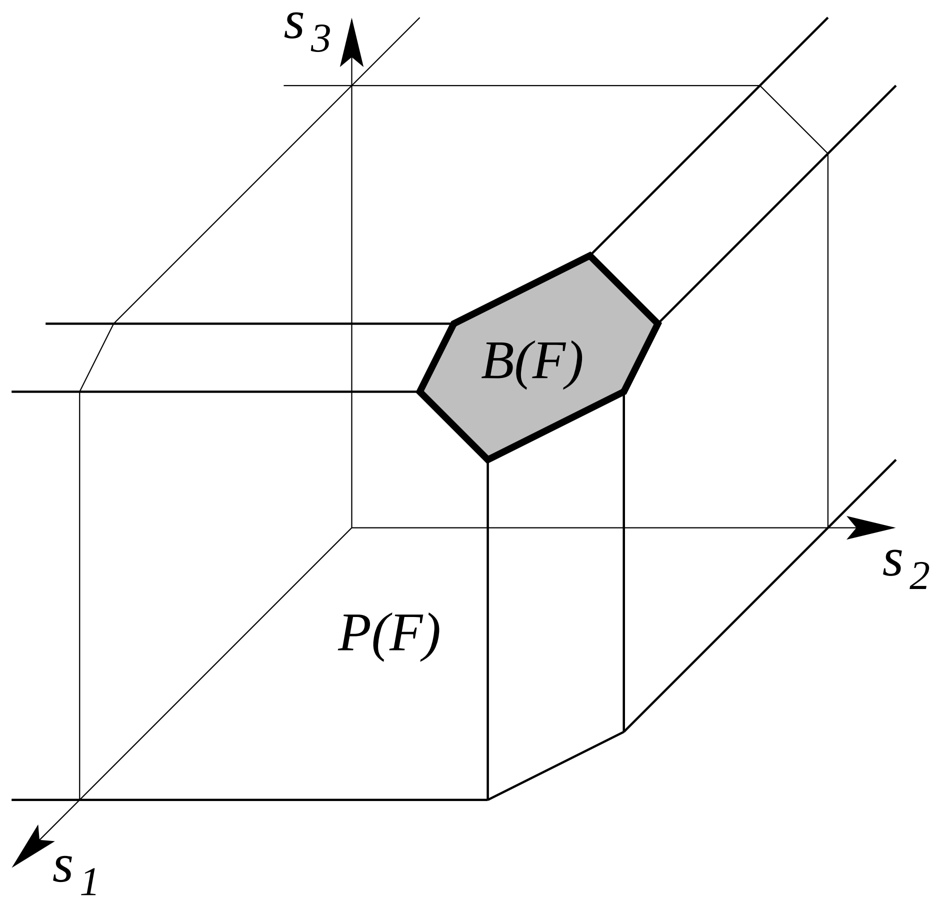
- F is submodular if and only if f is convex (Lovász, 1982)
  – Minimizing $f(w)$ on $w \in [0,1]^p$ equivalent to minimizing F on 2^V

Submodular functions - Submodular polyhedra

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

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





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 $\displaystyle\max_{s \in P(F)} w^\top s = f(w)$
  - if $w \in \mathbb{R}^p$, then $\displaystyle\max_{s \in B(F)} w^\top s = f(w)$

- Maximizer obtained by **greedy algorithm**:
  - Sort the components of w, as $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$
  - Set $s_{j_k} = F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})$

- Other operations on submodular polyhedra (see, e.g., Bach, 2011)

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**
  - Stobbe and Krause (2010)
  - Jegelka, Lin, and Bilmes (2011)

Separable optimization on base polyhedron

- Assume each ψ_k is a strictly convex function $\mathbb{R} \to \mathbb{R}$
- **Proposition**: the two following problems are dual to each other

$$\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$$

$$\max_{s \in B(F)} \sum_{k \in V} -\psi_k(-s_k)$$

Separable optimization on base polyhedron

- Assume each ψ_k is a strictly convex function $\mathbb{R} \to \mathbb{R}$

- **Proposition**: the two following problems are dual to each other

$$\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$$

$$\max_{s \in B(F)} \sum_{k \in V} -\psi_k(-s_k)$$

- **Proposition** (Chambolle and Darbon, 2009): let w^* be the solution of $\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$. Then, for $\alpha \in \mathbb{R}$,

$$\min_{A \subset V} F(A) + \sum_{j \in A} \psi_k'(\alpha)$$

has minimal minimizer $\{w^* > \alpha\}$ and maximal minimizer $\{w^* \geqslant \alpha\}$

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

Outline

- **Submodular functions**
  - Links with convexity through Lovász extension
  - Optimization on submodular polyhedra
- **Structured sparsity-inducing norms**
  - Relaxation of the penalization of supports
  - Examples
  - Unified algorithms and analysis
- **Approximate submodular function minimization**

Sparsity in supervised machine learning

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

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

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

Sparsity in unsupervised machine learning

- **Multiple responses/signals** $y = (y^1, \dots, y^k) \in \mathbb{R}^{n \times k}$
- **Dictionary learning**
  - Learn $X = (x^1, \dots, x^p) \in \mathbb{R}^{n \times p}$ such that $\forall j,\ \|x^j\|_2 \leqslant 1$

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

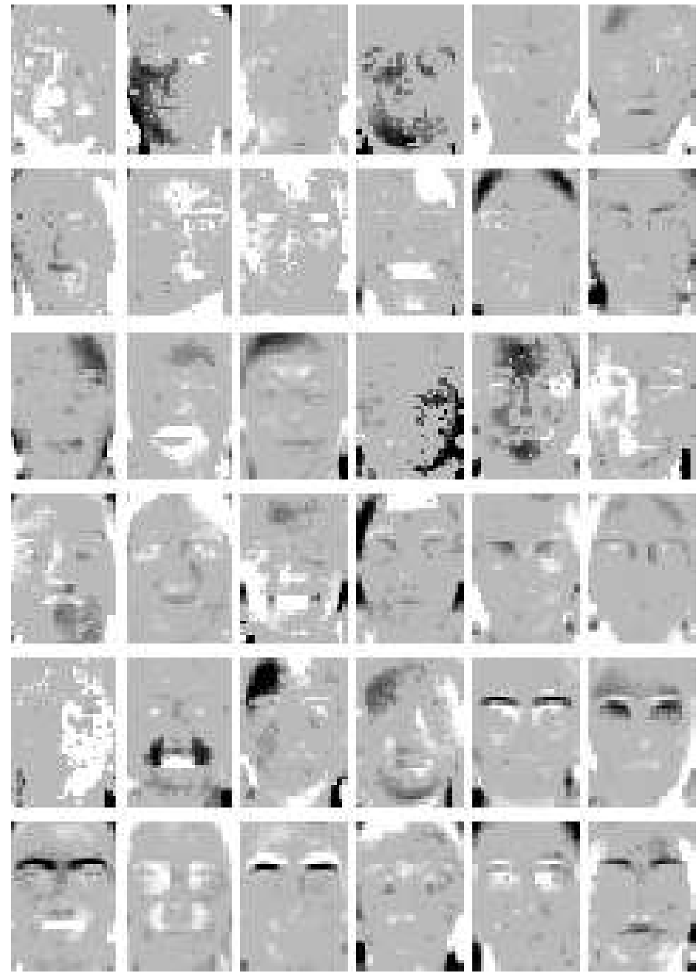
  - Olshausen and Field (1997); Elad and Aharon (2006)
- **sparse PCA**: replace $\|x^j\|_2 \leqslant 1$ by $\Theta(x^j) \leqslant 1$

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

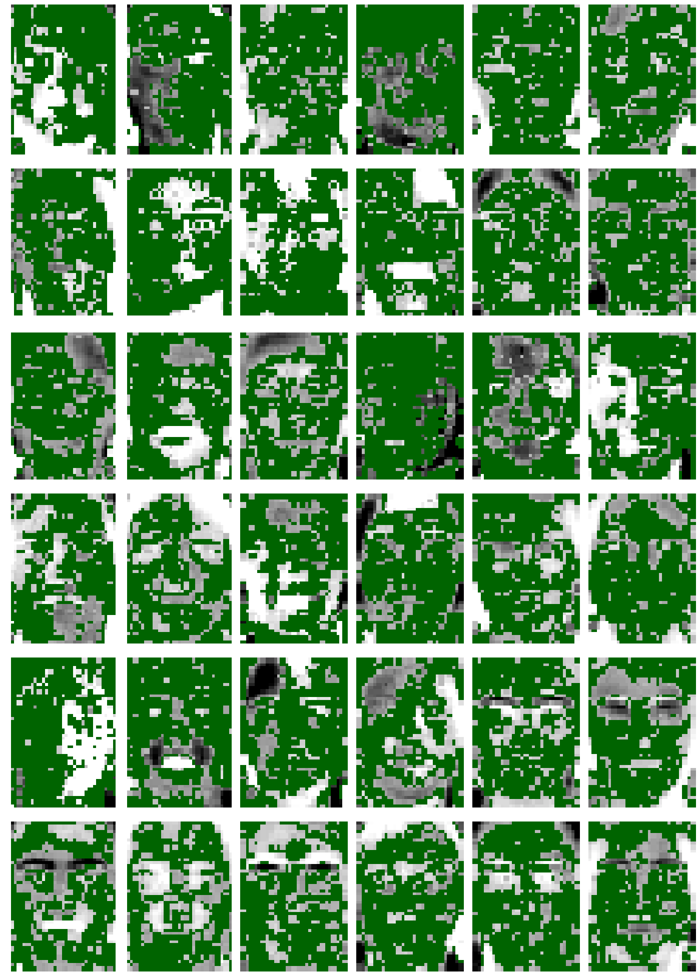


sparse PCA

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

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

raw data

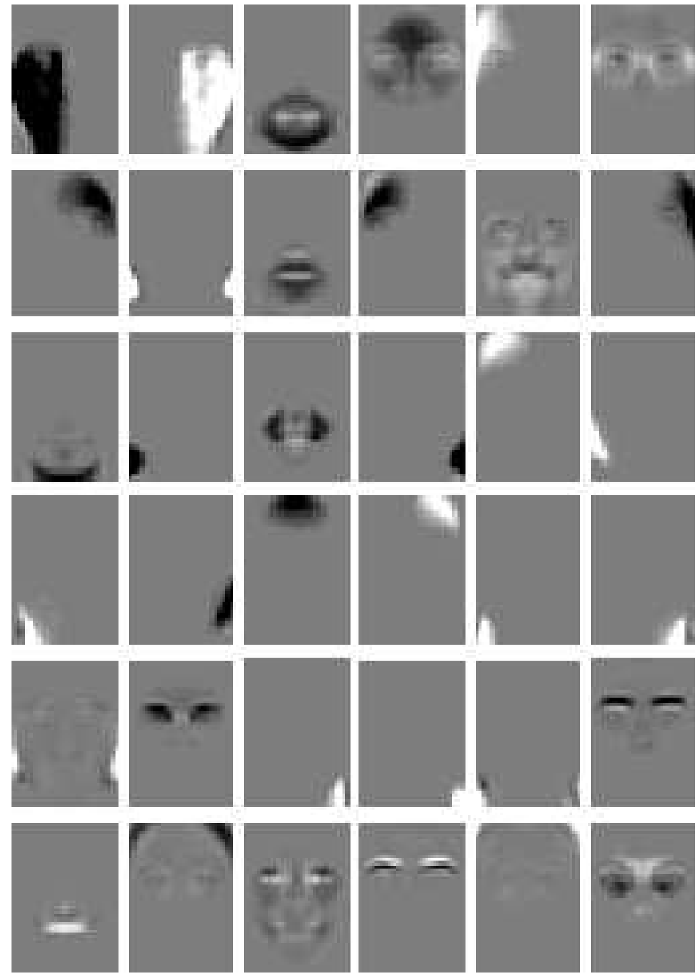


sparse PCA

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

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

raw data

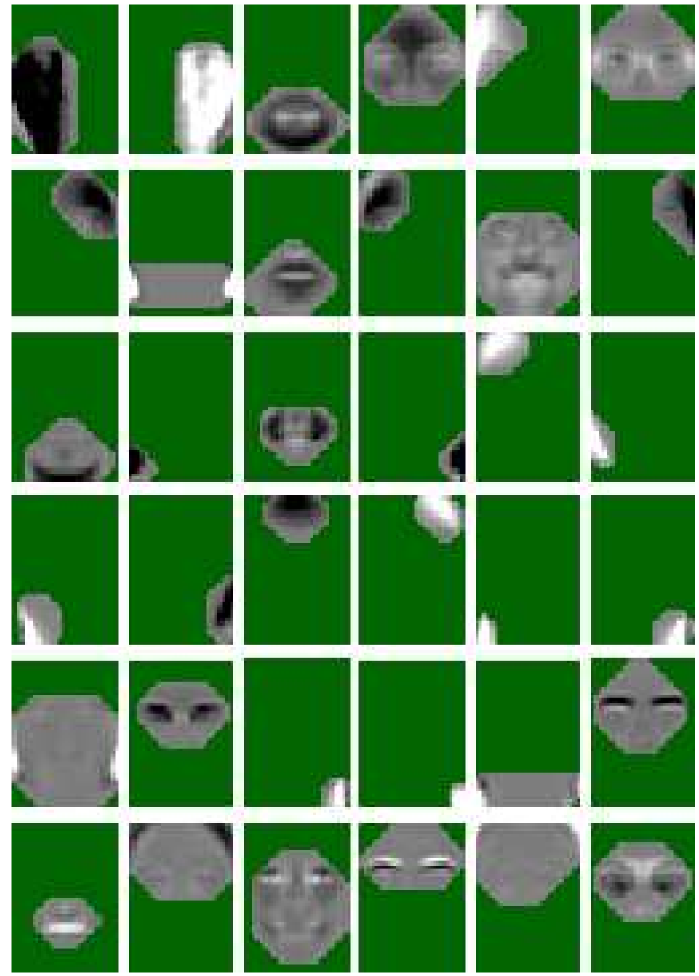


Structured sparse PCA

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

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

raw data



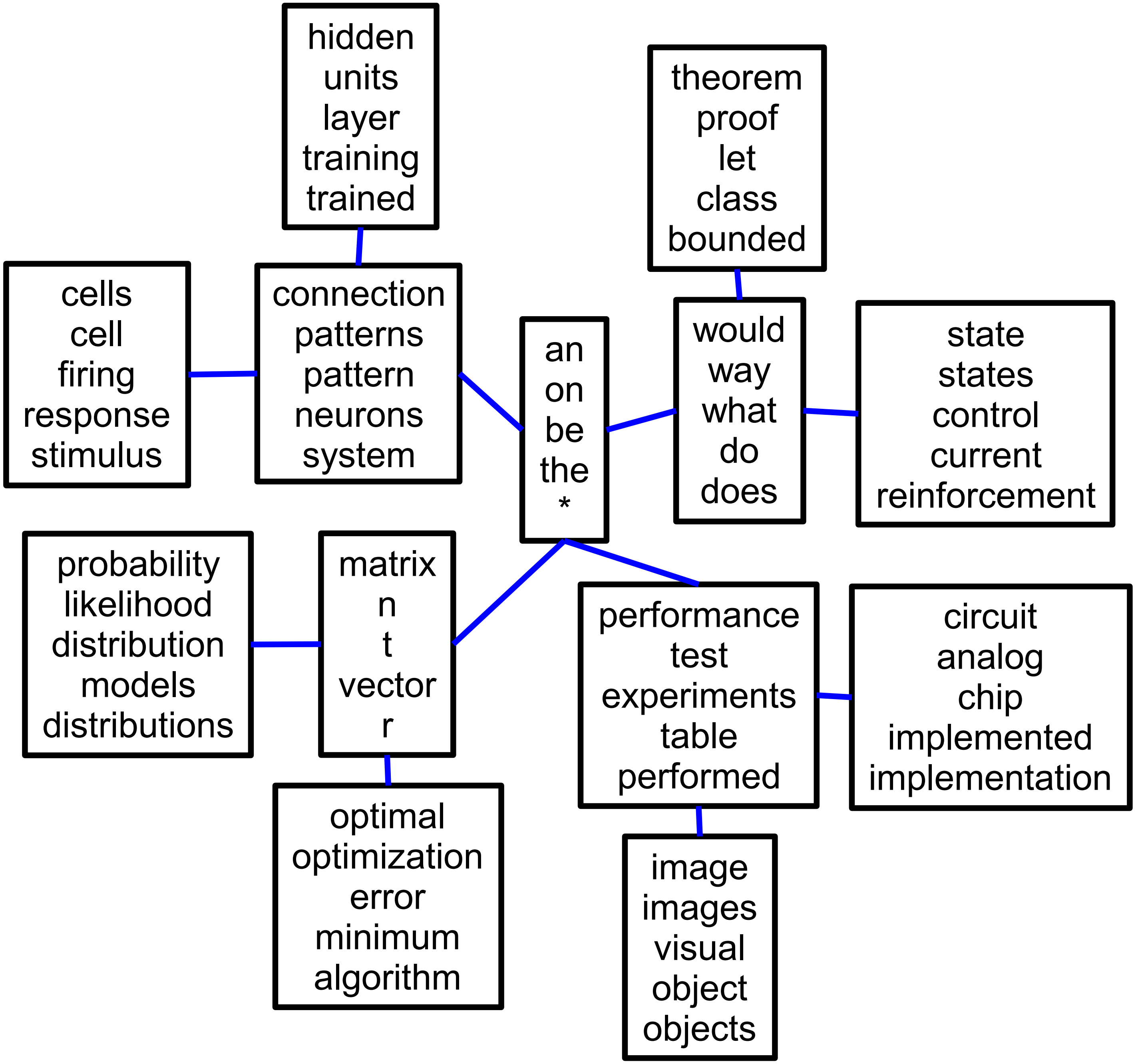
Structured sparse PCA

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

Why structured sparsity?

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

Modelling of text corpora (Jenatton et al., 2010)



Why structured sparsity?

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

Why structured sparsity?

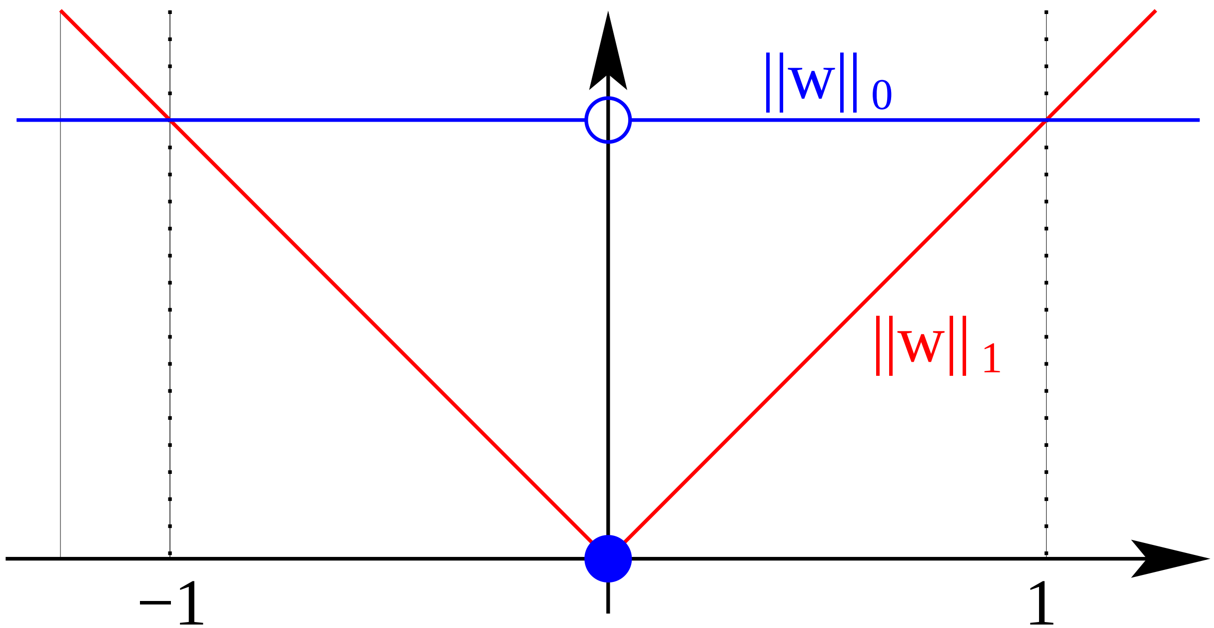
- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements "organized" in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)
- **Stability and identifiability**
  - Optimization problem $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda \|w\|_1$ is unstable
  - "Codes" w^j often used in later processing (Mairal et al., 2009)
- **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)

ℓ_1-norm = convex envelope of cardinality of support

- Let $w \in \mathbb{R}^p$. Let $V = \{1, \dots, p\}$ and $\mathrm{Supp}(w) = \{j \in V, \; w_j \neq 0\}$

- **Cardinality of support**: $\|w\|_0 = \mathrm{Card}(\mathrm{Supp}(w))$

- Convex envelope = largest convex lower bound (see, e.g., Boyd and Vandenberghe, 2004)



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

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

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

- Define $\Theta(w) = F(\text{Supp}(w))$: How to get its convex envelope?

  1. Possible if F is also **submodular**
  2. Allows **unified** theory and algorithm
  3. Provides **new** regularizers

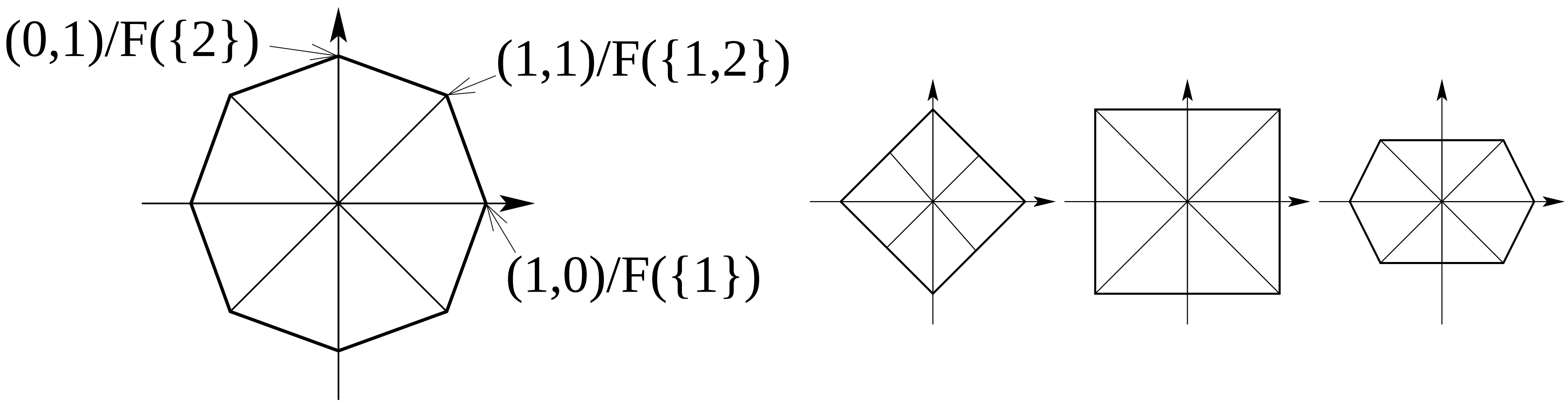
Submodular functions and structured sparsity

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

- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\mathrm{Supp}(w))$ on the ℓ_∞-ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F

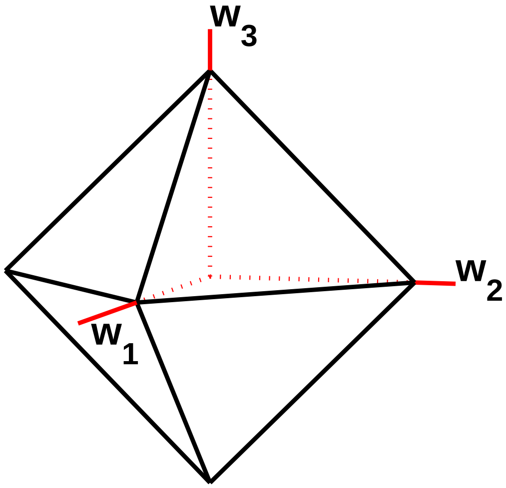
Submodular functions and structured sparsity

- 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

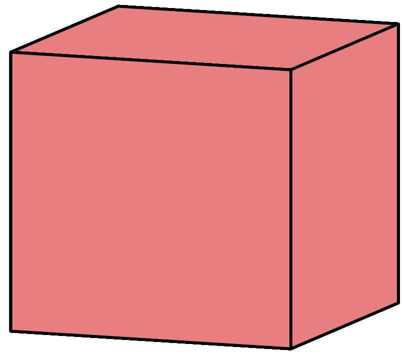


  - A if stable if for all $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$
  - With probability one, stable sets are the only allowed patterns

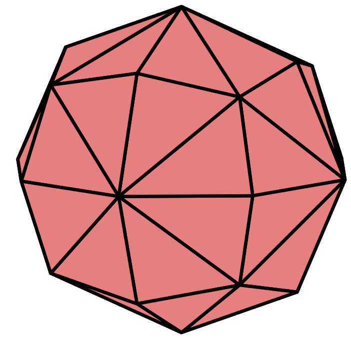
Polyhedral unit balls



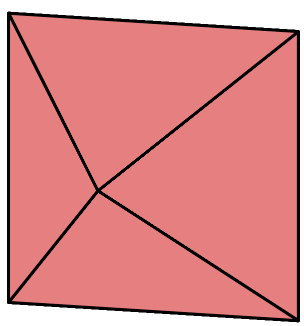
$$F(A) = |A|$$
$$\Omega(w) = \|w\|_1$$



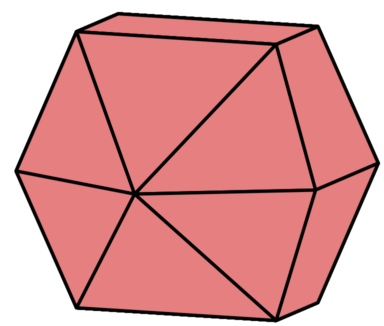
$$F(A) = \min\{|A|, 1\}$$
$$\Omega(w) = \|w\|_\infty$$



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

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

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

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

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)

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

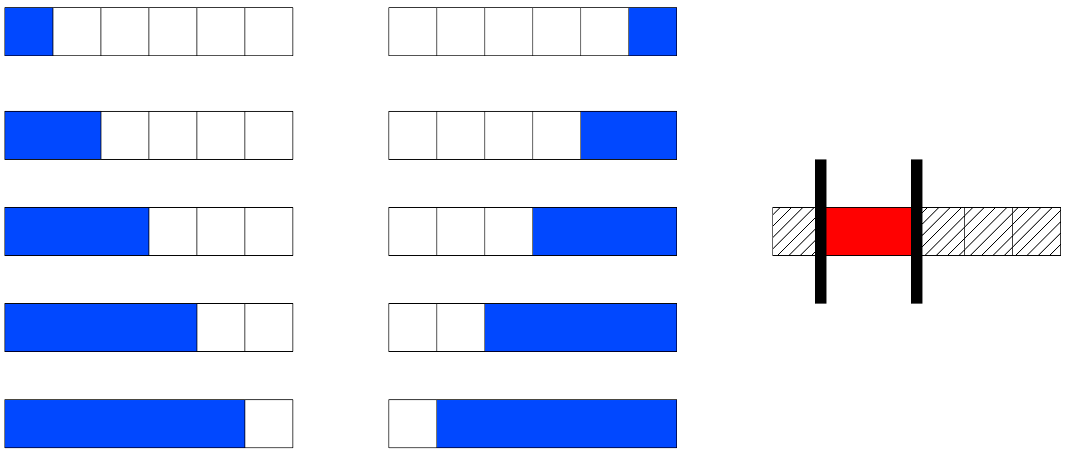
  - ℓ_1-ℓ_∞ norm $\Rightarrow$ sparsity at the group level
  - Some w_G's are set to zero for some groups G

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

  - Justification not only limited to allowed sparsity patterns

Selection of contiguous patterns in a sequence

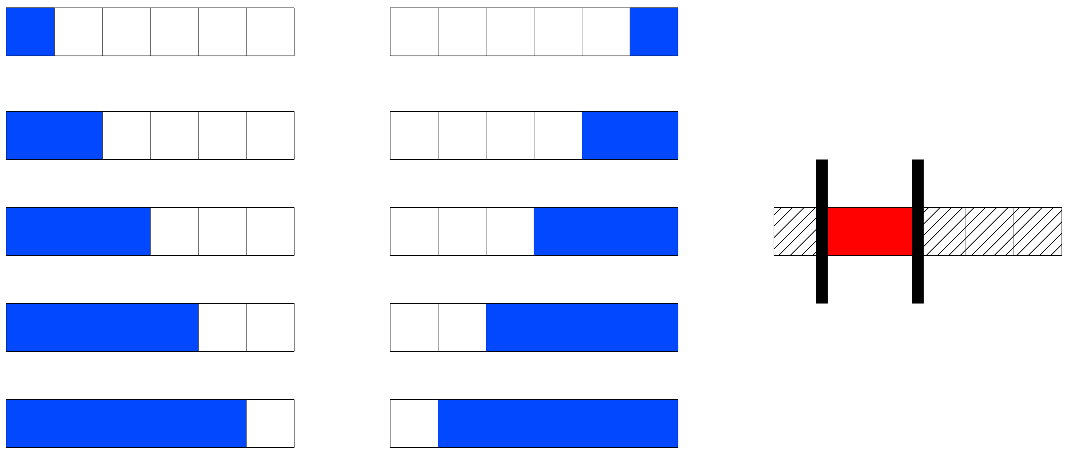
- Selection of contiguous patterns in a sequence



- $\mathcal{G}$ 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

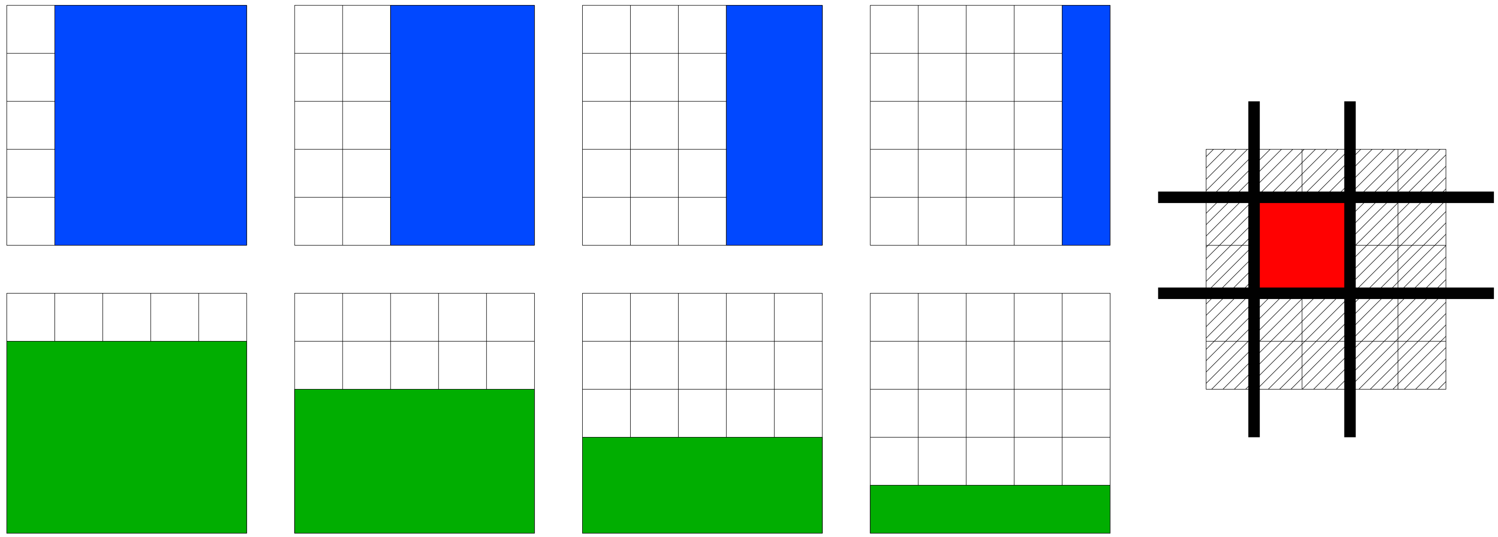


- $\mathcal{G}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a **contiguous pattern**

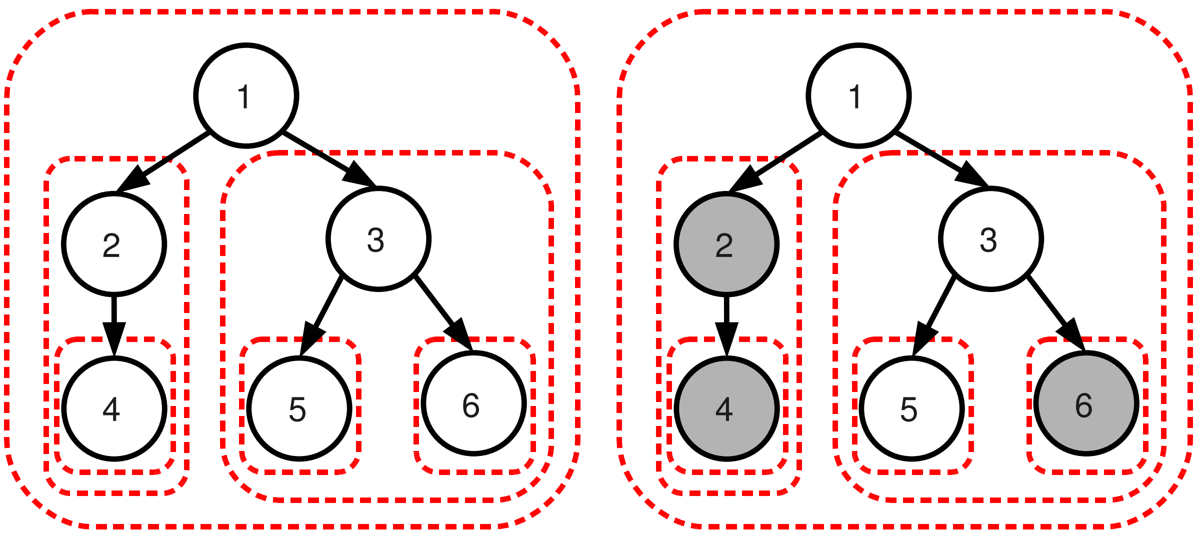
- $\sum_{G \in \mathcal{G}} \|w_G\|_\infty \;\Rightarrow\; F(A) = p - 2 + \mathrm{Range}(A)$ if $A \neq \varnothing$

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)

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

  - 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)

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

  - Justification not only limited to allowed sparsity patterns

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

Unified optimization algorithms

- **Polyhedral norm** with exponentially many faces and extreme points
  - Not suitable for 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 exponentially many faces and extreme points
  - Not suitable for linear programming toolboxes

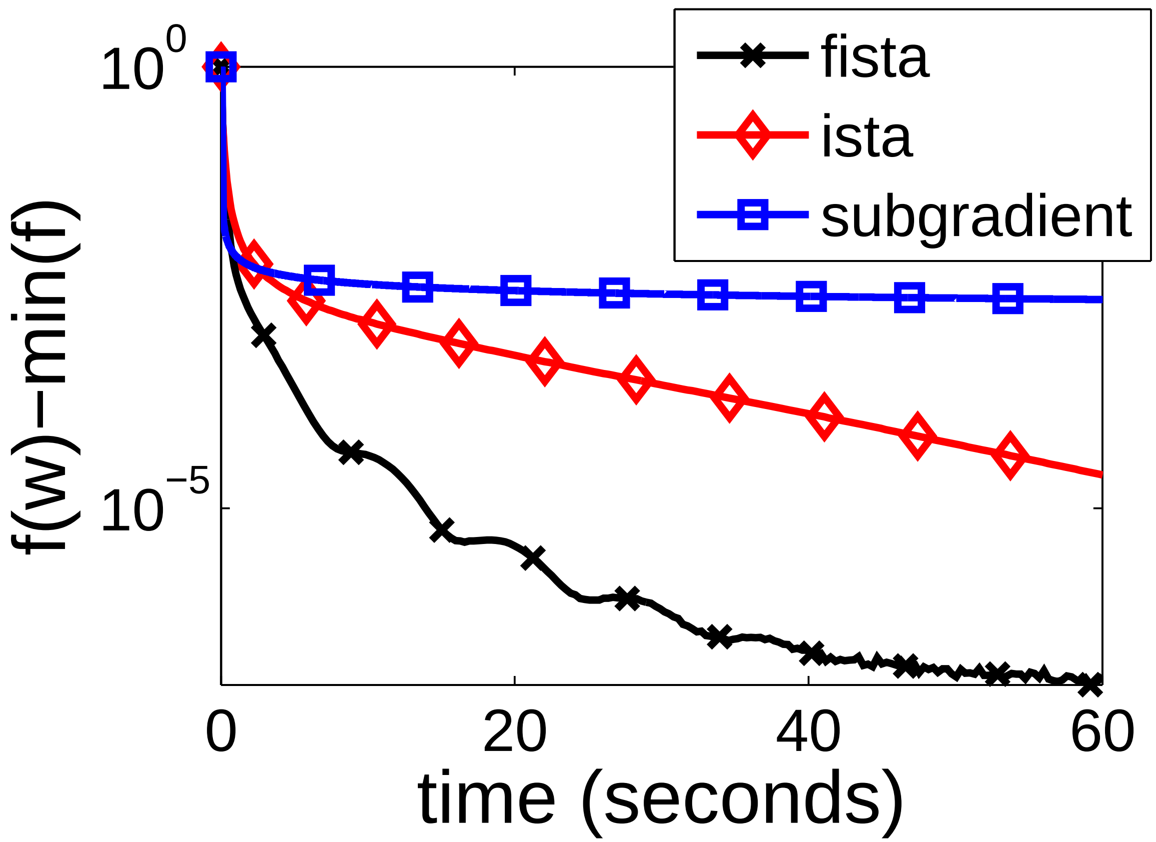
- **Subgradient** ($w \mapsto \Omega(w)$ non-differentiable)
  - subgradient may be obtained in polynomial time $\Rightarrow$ too slow

- **Proximal methods** (see, e.g., Beck and Teboulle, 2009; Bach, Jenatton, Mairal, and Obozinski, 2011)
  - $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda\Omega(w)$: differentiable + non-differentiable
  - Efficient when $(P): \ \min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - v\|_2^2 + \lambda\Omega(w)$ is "easy"

- **The proximal problem (P) is equivalent to a sequence of submodular function minimizations**
  - Decomposition strategy (Groenevelt, 1991) or min-norm-point

Comparison of optimization algorithms

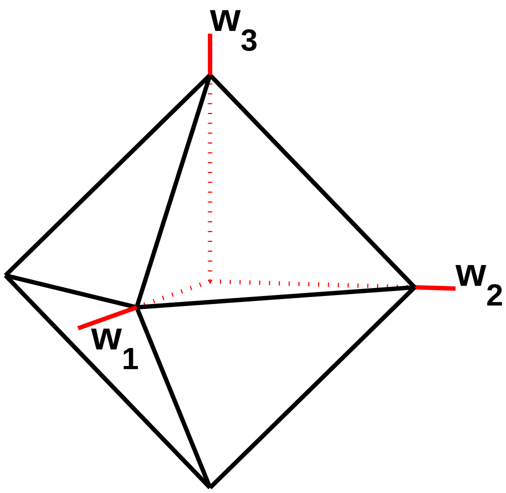
- Synthetic example with $p = 1000$ and $F(A) = |A|^{1/2}$
- ISTA: proximal method
- FISTA: accelerated variant (Beck and Teboulle, 2009)



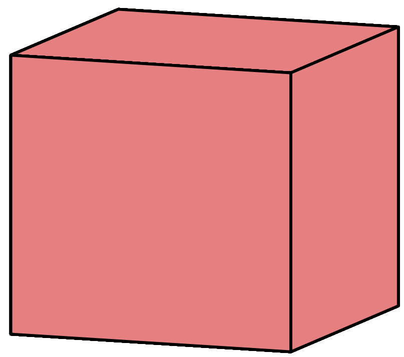
Extensions

- **Unified statistical analysis** (Bach, 2010)
  - support recovery and estimation consistency
- **Extension to symmetric submodular functions**
  - Shaping level sets (Bach, 2011)
- **Avoiding artefacts linked with ℓ_∞-norms**
  - See poster at this workshop (Obozinski and Bach, 2011)
- **Generalization to other set-functions**
  - See same poster at this workshop (Obozinski and Bach, 2011)

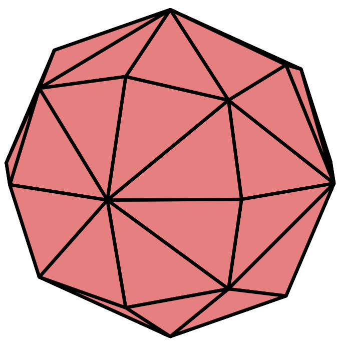
Polyhedral unit balls



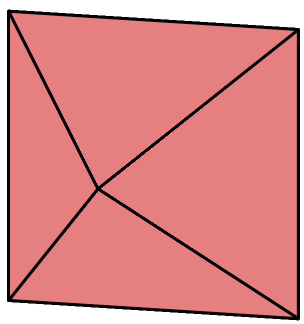
$$F(A) = |A|$$
$$\Omega(w) = \|w\|_1$$



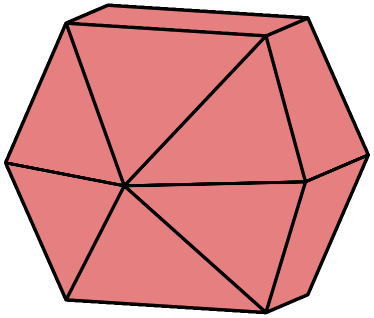
$$F(A) = \min\{|A|, 1\}$$
$$\Omega(w) = \|w\|_\infty$$



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



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



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

Outline

- **Submodular functions**
  - Links with convexity through Lovász extension
  - Optimization on submodular polyhedra
- **Structured sparsity-inducing norms**
  - Relaxation of the penalization of supports
  - Examples
  - Unified algorithms and analysis
- **Approximate submodular function minimization**

Approximate submodular function minimization

- **For most machine learning applications, no need to obtain exact minimum**

Approximate submodular function minimization

- **For most machine learning applications, no need to obtain exact minimum**
- Assume (wlog.) that $\forall k \in V$, $F(\{k\}) \geqslant 0$ and $F(V\backslash\{k\}) \geqslant F(V)$
- Denote $D^2 = \sum_{k \in V} \big\{ F(\{k\}) + F(V\backslash\{k\}) - F(V) \big\}$
- **Proposition**: t iterations of subgradient descent outputs a set A_t (and a certificate of optimality s_t) such that

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

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

Approximate quadratic optimization on $B(F)$

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

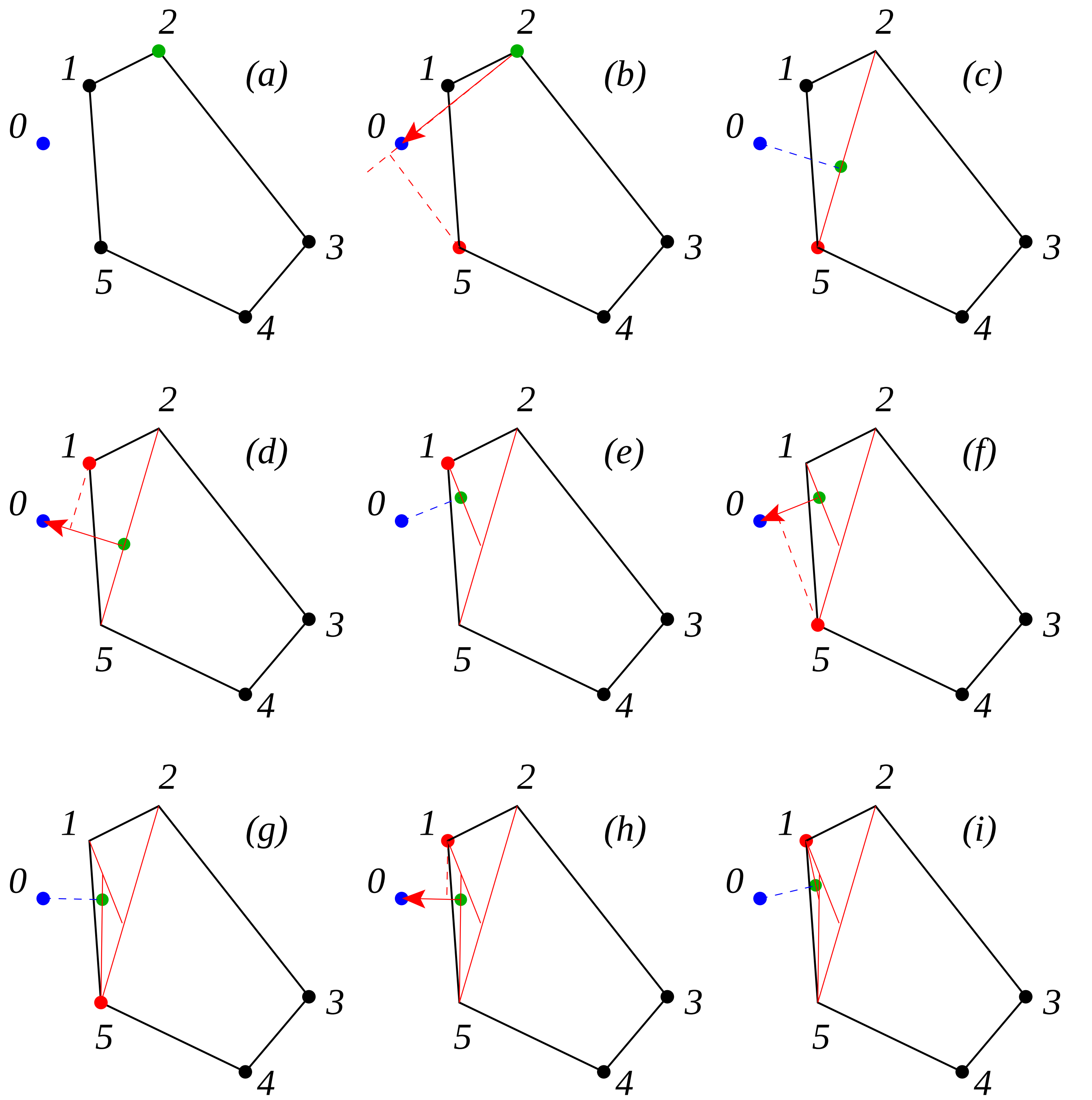
- Can only maximize linear functions on $B(F)$

- **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

- **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

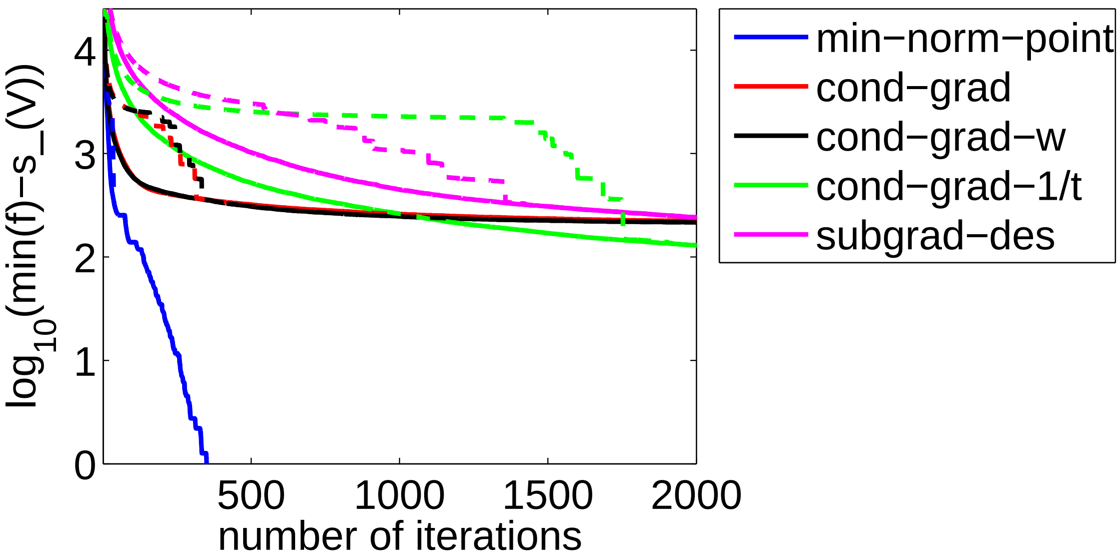
- **Proposition**: If w is ε-optimal for $\min_{w\in\mathbb{R}^p} \frac{1}{2}\|w\|_2^2 + f(w)$, then at least a levet set A of w is $\left(\frac{\sqrt{\varepsilon p}}{2}\right)$-optimal for submodular function minimization

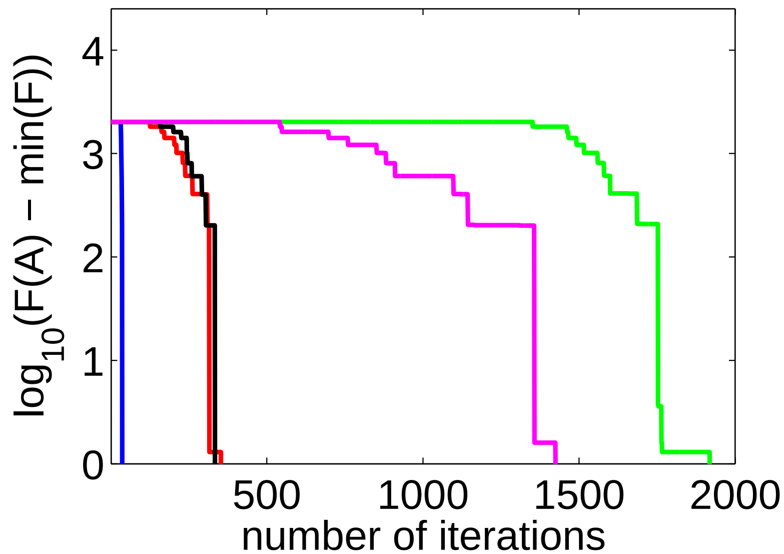
- If $\varepsilon = \dfrac{2D^2}{t}$, $\dfrac{\sqrt{\varepsilon p}}{2} = \dfrac{Dp^{1/2}}{\sqrt{2t}} \Rightarrow$ **no provable gains**, but:
  - Bound on the iterates A_t (with additional assumptions)
  - Possible thresolding for acceleration

- **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 = 575

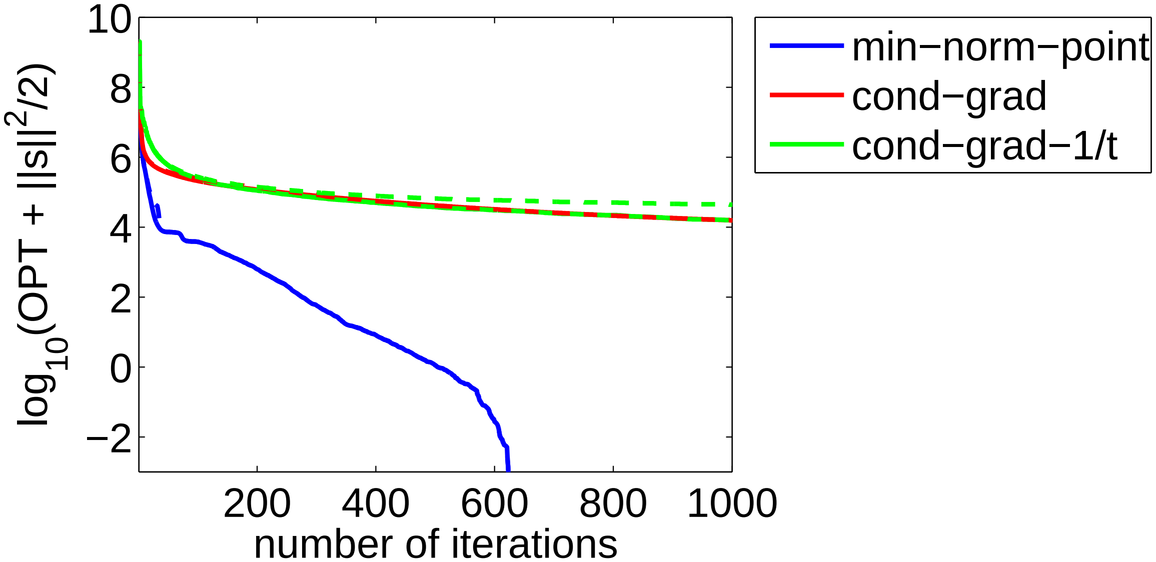
- **Submodular function minimization**
  - (Left) optimal value minus dual function values $(s_t)_-(V)$ (in dashed, certified duality gap)
  - (Right) Primal function values $F(A_t)$ minus optimal value

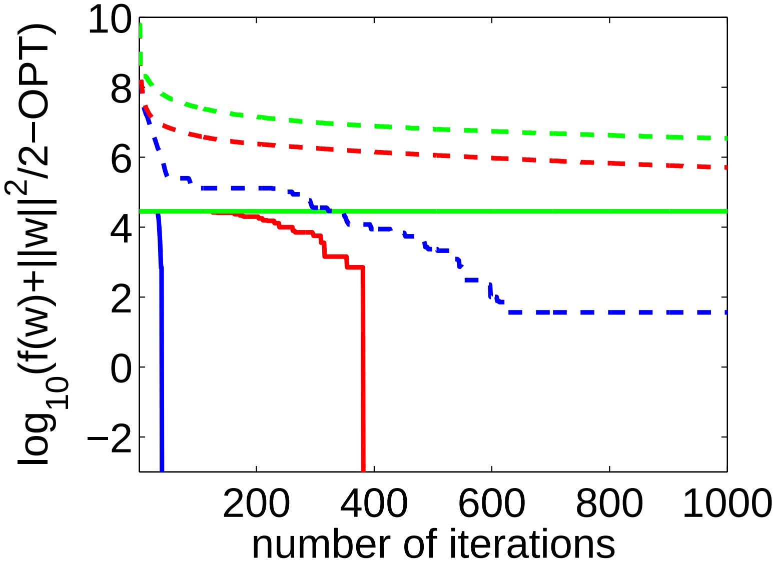




Simulations on standard benchmark

- **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)





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. 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.

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.

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

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

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

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

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

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

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

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.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.

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

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.

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

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, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *Advances in Neural Information Processing Systems (NIPS)*, 21, 2009.

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

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.

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

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.

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