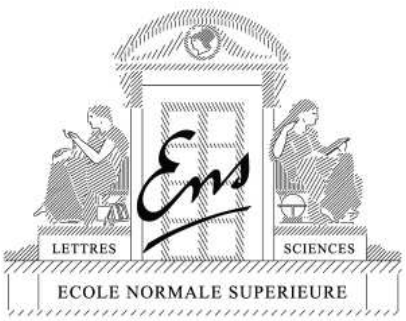
Convex relaxations for Combinatorial Penalties



Guillaume Obozinski

SIERRA team - INRIA - ENS - Paris

Joint work with Francis Bach

Journées de Statistique - Bruxelles - 21 mai 2012

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$

From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

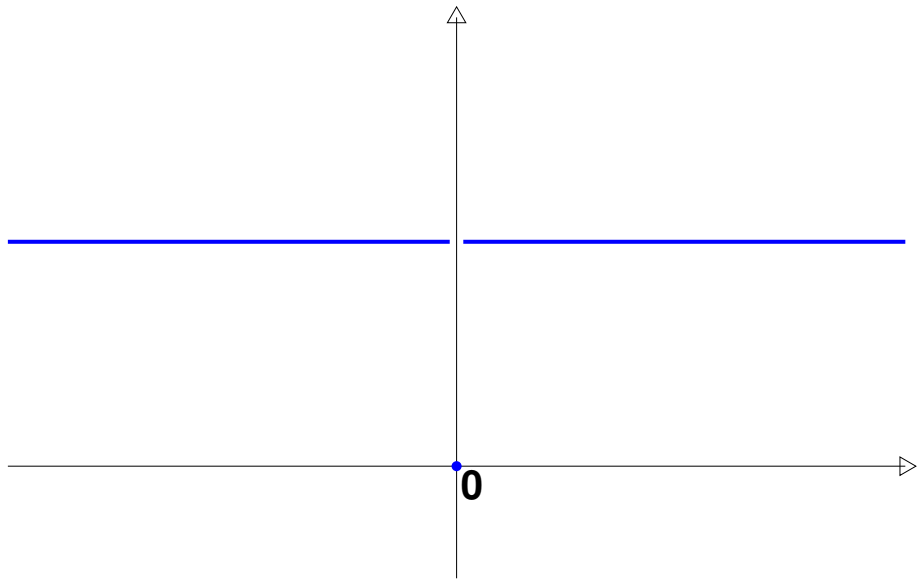
$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$



From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

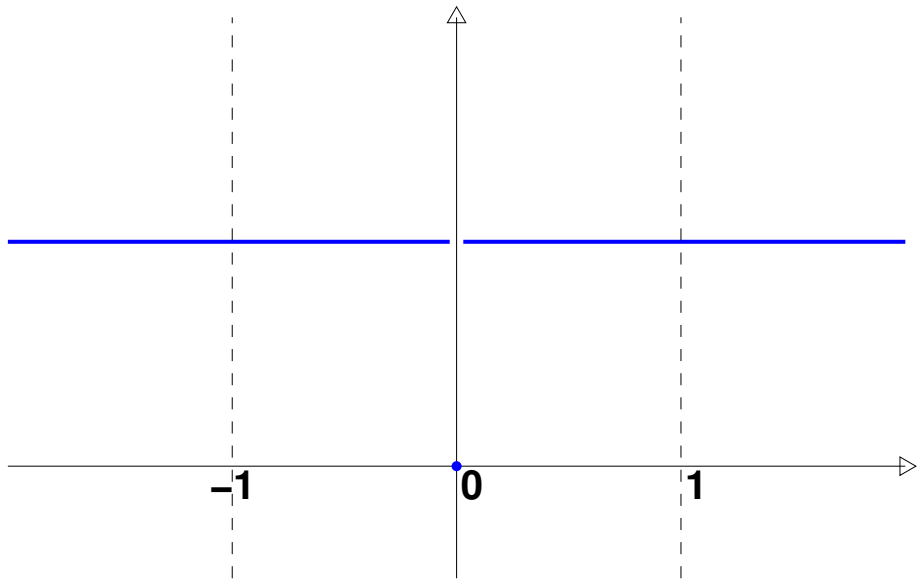
$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$



From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

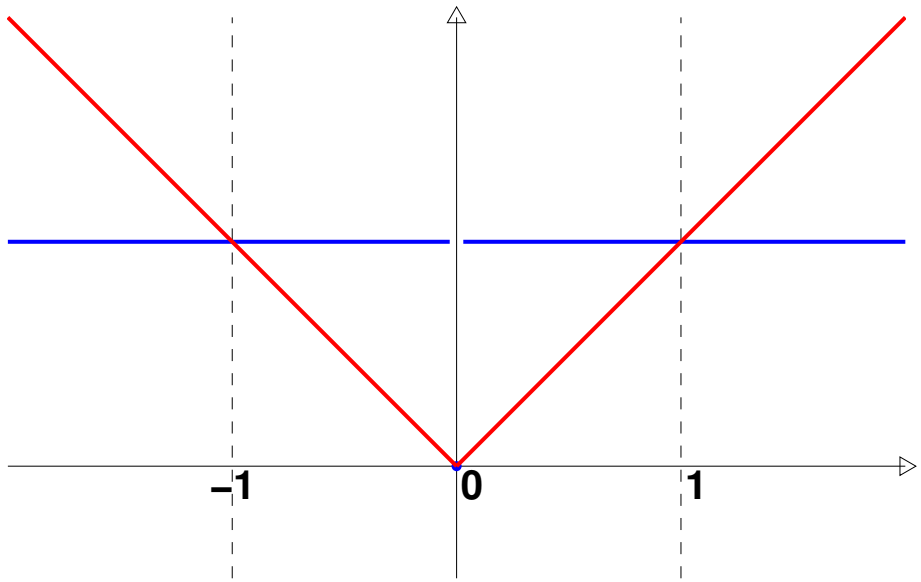
$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$



From sparsity...

- Empirical risk: for $w \in \mathbb{R}^d$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

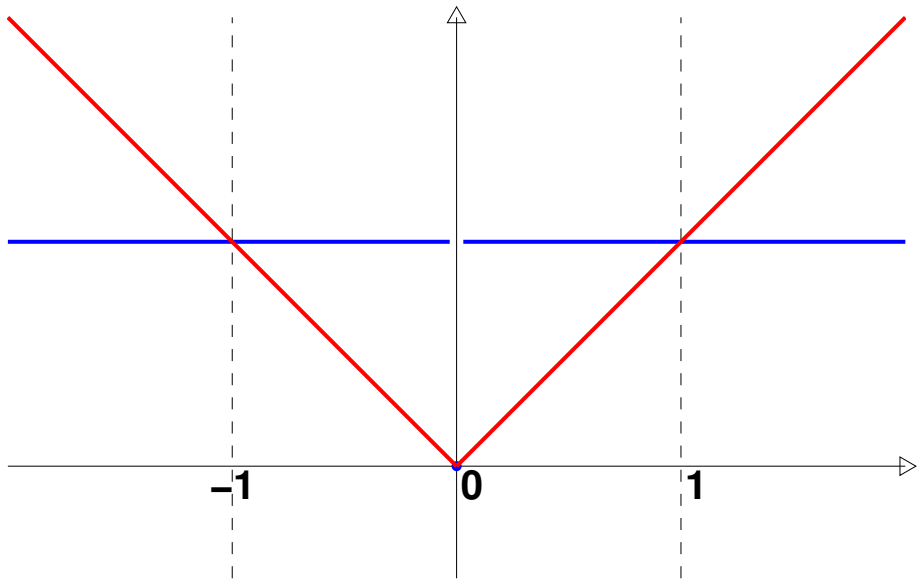
$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$

Lasso

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \|w\|_1$$



... to Structured Sparsity

The support is not only **sparse**, but, in addition,
we have prior information about its **structure**.

... to Structured Sparsity

The support is not only **sparse**, but, in addition, we have prior information about its **structure**.

Examples

- The variables should be selected in groups.

... to Structured Sparsity

The support is not only **sparse**, but, in addition, we have prior information about its **structure**.

Examples

- The variables should be selected in groups.
- The variables lie in a hierarchy.

... to Structured Sparsity

The support is not only **sparse**, but, in addition,
we have prior information about its **structure**.

Examples

- The variables should be selected in groups.
- The variables lie in a hierarchy.
- The variables lie on a graph or network and the support should be localized or densely connected on the graph.

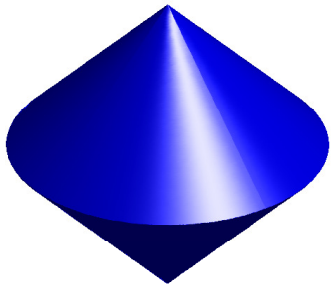
Ideas in structured sparsity

Group Lasso and ℓ_1/ℓ_p norm (Yuan and Lin, 2006)

Group Lasso

Given $\mathcal{G} = \{A_1, \ldots, A_m\}$ a partition of $V := \{1, \ldots, d\}$ consider

$$\|w\|_{\ell_1/\ell_p} = \sum_{A \in \mathcal{G}} \delta^A \|w_A\|_p$$

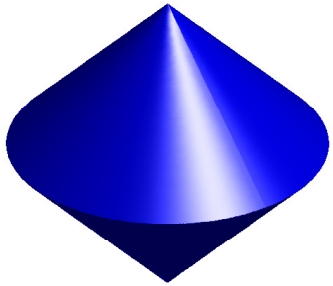


Group Lasso and ℓ_1/ℓ_p norm (Yuan and Lin, 2006)

Group Lasso

Given $\mathcal{G} = \{A_1, \dots, A_m\}$ a partition of $V := \{1, \dots, d\}$ consider

$$\|w\|_{\ell_1/\ell_p} = \sum_{A \in \mathcal{G}} \delta^A \|w_A\|_p$$

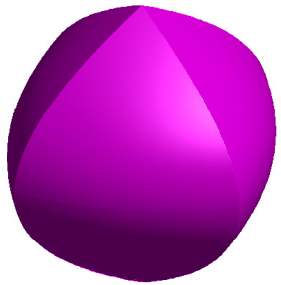


Overlapping groups: direct extension of Jenatton, Audibert and Bach (2009).

Interesting induced structures

$\rightarrow$ Induce patterns of rooted subtree

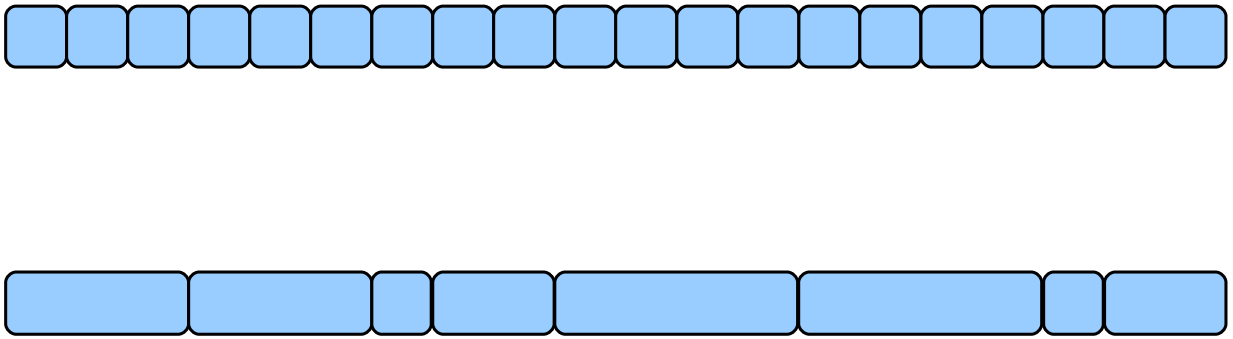
$\rightarrow$ "Convex" sets



More generally sets of patterns that are *stable by intersection*

Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$



Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

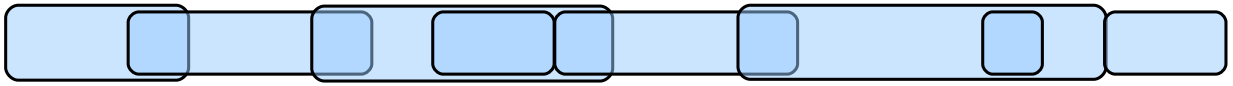
Group Lasso: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

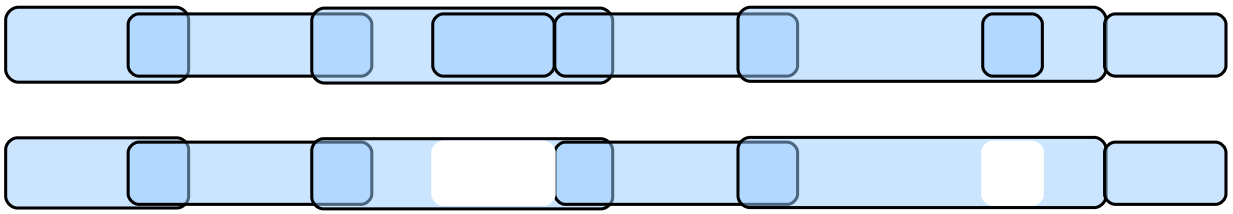


Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

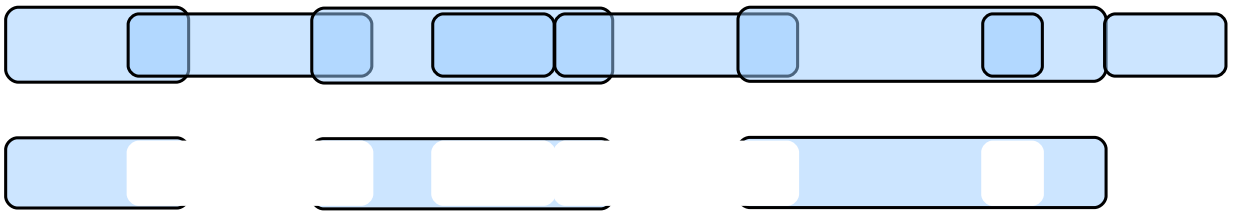


Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

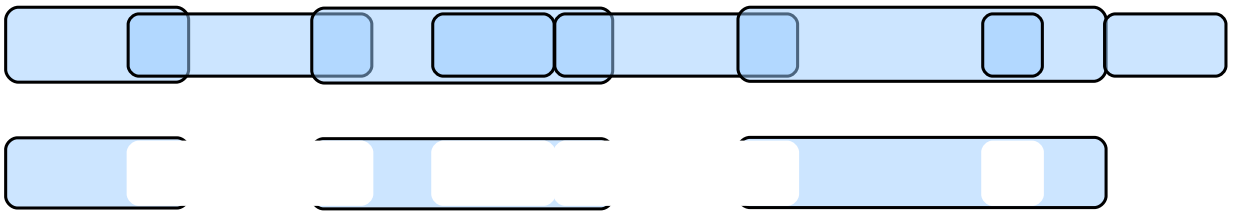


Sparsity patterns induced for $L(w) + \lambda\,\Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in \mathcal{G}} \|w_g\|$



The support obtained is

- An intersection of the complements of the groups set to 0 (cf. Jenatton et al. (2011))
- Not a union of groups

Latent group Lasso Jacob, Obozinski and Vert (2009)

- Given parameters $w \in \mathbb{R}^d$
- Consider latent parameters $v^A \in \mathbb{R}^d$, for $A \in \mathcal{G} \subset 2^V$

Latent group Lasso Jacob, Obozinski and Vert (2009)

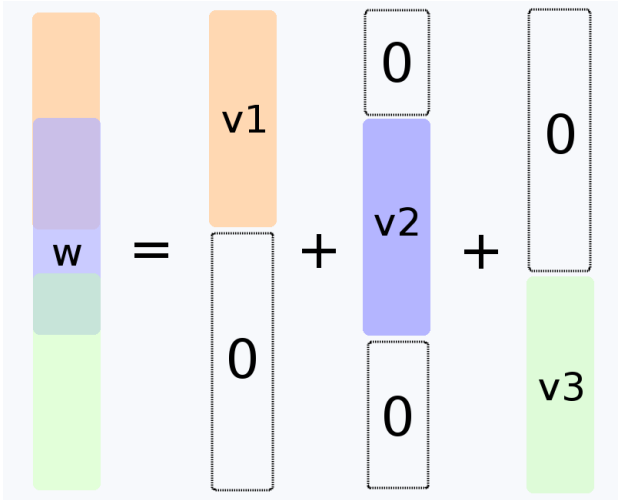
$$w = v^{A_1} + v^{A_2} + v^{A_3}$$

- Given parameters $w \in \mathbb{R}^d$
- Consider latent parameters $v^A \in \mathbb{R}^d$, for $A \in \mathcal{G} \subset 2^V$
- Such that $w = \sum_{A \in \mathcal{G}} v^A$

Latent group Lasso Jacob, Obozinski and Vert (2009)

$$w = v^{A_1} + v^{A_2} + v^{A_3}$$

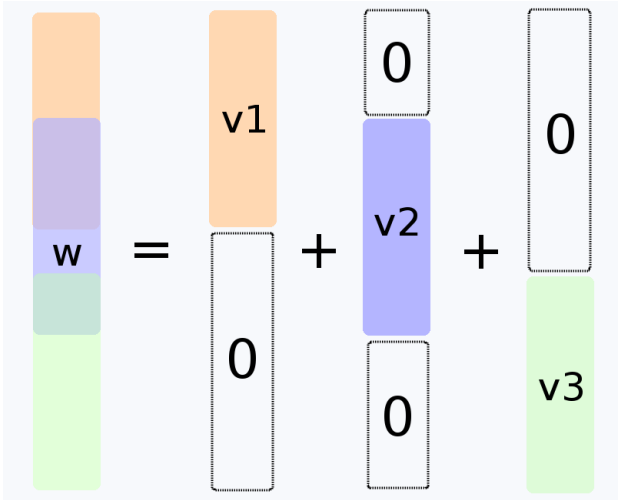
- Given parameters $w \in \mathbb{R}^d$
- Consider latent parameters $v^A \in \mathbb{R}^d$, for $A \in \mathcal{G} \subset 2^V$
- Such that $w = \sum_{A \in \mathcal{G}} v^A$
- And such that $\mathrm{Supp}(v^A) \subset A$.



Latent group Lasso Jacob, Obozinski and Vert (2009)

$$w = v^{A_1} + v^{A_2} + v^{A_3}$$

- Given parameters $w \in \mathbb{R}^d$
- Consider latent parameters $v^A \in \mathbb{R}^d$, for $A \in \mathcal{G} \subset 2^V$
- Such that $w = \sum_{A \in \mathcal{G}} v^A$
- And such that $\mathrm{Supp}(v^A) \subset A$.



Define a new regularization:

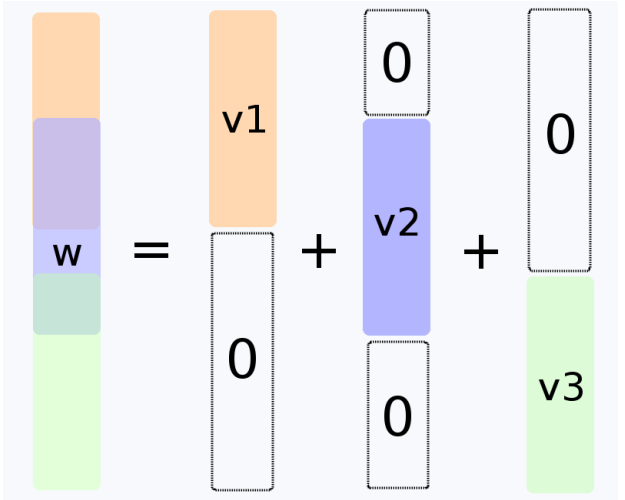
$$\Omega_{\star_{\inf}}(w) = \min_{(v^A \in \mathbb{R}^d)_{A \in \mathcal{G}}} \sum_{A \in \mathcal{G}} \delta^A \|v^A\|_p$$

$$\text{s.t.} \quad \forall A \in \mathcal{G},\ \mathrm{Supp}(v^A) \subset A, \quad \text{and} \quad w = \sum_{A \in \mathcal{G}} v^A.$$

Latent group Lasso Jacob, Obozinski and Vert (2009)

$$w = v^{A_1} + v^{A_2} + v^{A_3}$$

- Given parameters $w \in \mathbb{R}^d$
- Consider latent parameters $v^A \in \mathbb{R}^d$, for $A \in \mathcal{G} \subset 2^V$
- Such that $w = \sum_{A \in \mathcal{G}} v^A$
- And such that $\mathrm{Supp}(v^A) \subset A$.



Define a new regularization:

$$\Omega_{\star_{\inf}}(w) = \min_{(v^A \in \mathbb{R}^d)_{A \in \mathcal{G}}} \sum_{A \in \mathcal{G}} \delta^A \|v^A\|_p$$

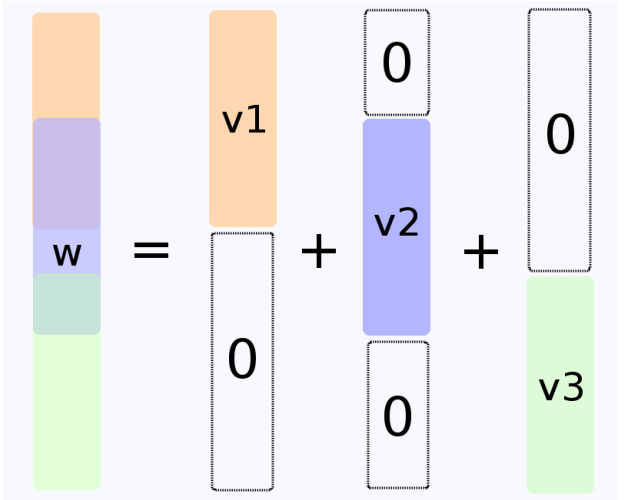
$$\text{s.t.} \quad \forall A \in \mathcal{G},\ \mathrm{Supp}(v^A) \subset A, \quad \text{and} \quad w = \sum_{A \in \mathcal{G}} v^A.$$

$\Omega_{\star_{\inf}}$ is a norm !

Latent group Lasso Jacob, Obozinski and Vert (2009)

$$w = v^{A_1} + v^{A_2} + v^{A_3}$$

- Given parameters $w \in \mathbb{R}^d$
- Consider latent parameters $v^A \in \mathbb{R}^d$, for $A \in \mathcal{G} \subset 2^V$
- Such that $w = \sum_{A \in \mathcal{G}} v^A$
- And such that $\mathrm{Supp}(v^A) \subset A$.



Define a new regularization:

$$\Omega_{\star_{\inf}}(w) = \min_{(v^A \in \mathbb{R}^d)_{A \in \mathcal{G}}} \sum_{A \in \mathcal{G}} \delta^A \|v^A\|_p$$

$$\text{s.t.} \quad \forall A \in \mathcal{G},\ \mathrm{Supp}(v^A) \subset A, \quad \text{and} \quad w = \sum_{A \in \mathcal{G}} v^A.$$

$\Omega_{\star_{\inf}}$ is a norm ! Patterns are typically union of subcollections of the sets defining the norm.

Latent group Lasso example

Consider V=$\{1, 2, 3\}$.

$\mathcal{G} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$

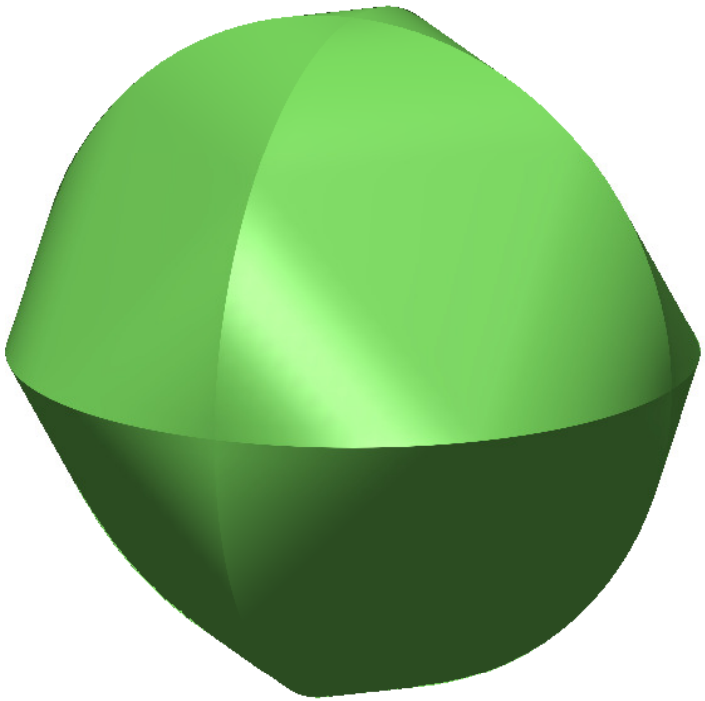
- $\delta^{\{1,2\}} = 1$,
- $\delta^{\{1,3\}} = 1$,
- $\delta^{\{2,3\}} = 1$,

Latent group Lasso example

Consider V=$\{1, 2, 3\}$.

$\mathcal{G} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$

- $\delta^{\{1,2\}} = 1$,
- $\delta^{\{1,3\}} = 1$,
- $\delta^{\{2,3\}} = 1$,



A new approach based on combinatorial functions

General framework

Let $V = \{1, \ldots, d\}$.
Given a set function $F : 2^V \mapsto \mathbb{R}_+$ consider

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

General framework

Let $V = \{1, \ldots, d\}$.
Given a set function $F : 2^V \mapsto \mathbb{R}_+$ consider

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

Examples of combinatorial functions

- Use **recursivity** or **counts** of structures (e.g. tree) with DP
- **Block-coding** (Huang et al., 2011)

$$\tilde{G}(A) = \min_{B_i} F(B_1) + \ldots + F(B_k) \quad \text{s.t.} \quad B_1 \cup \ldots \cup B_k \supset A$$

- **Submodular functions** (Work on convex relaxations by Bach (2010))

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

- $\rightarrow$ Greedy algorithms
- $\rightarrow$ Non-convex methods
- $\rightarrow$ Relaxation

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

- $\rightarrow$ Greedy algorithms
- $\rightarrow$ Non-convex methods
- $\rightarrow$ Relaxation

$\lvert A \rvert$	$F(A)$
$L(w) + \lambda \lvert \mathrm{Supp}(w) \rvert$	

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

- $\rightarrow$ Greedy algorithms
- $\rightarrow$ Non-convex methods
- $\rightarrow$ Relaxation

$\lvert A \rvert$	$F(A)$
$L(w) + \lambda \lvert \mathrm{Supp}(w) \rvert$ $\downarrow$ $L(w) + \lambda \lVert w \rVert_1$	

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

- $\rightarrow$ Greedy algorithms
- $\rightarrow$ Non-convex methods
- $\rightarrow$ Relaxation

$\lvert A \rvert$	$F(A)$
$L(w) + \lambda \lvert \mathrm{Supp}(w) \rvert$ $\downarrow$ $L(w) + \lambda \lVert w \rVert_1$	$L(w) + \lambda F(\mathrm{Supp}(w))$

A relaxation for F...?

How to solve?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\mathrm{Supp}(w))$$

- $\rightarrow$ Greedy algorithms
- $\rightarrow$ Non-convex methods
- $\rightarrow$ Relaxation

$\lvert A \rvert$	$F(A)$
$L(w) + \lambda \lvert \mathrm{Supp}(w) \rvert$	$L(w) + \lambda F(\mathrm{Supp}(w))$
$\downarrow$	$\downarrow ?$
$L(w) + \lambda \lVert w \rVert_1$	$L(w) + \lambda \ldots ? \ldots$

Previous relaxation result

(Bach, 2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Previous relaxation result

(Bach, 2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

Previous relaxation result

(Bach, 2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball

Previous relaxation result

(Bach, 2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball
- The choice of ℓ_∞ seems arbitrary

Previous relaxation result

(Bach, 2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball
- The choice of ℓ_∞ seems arbitrary
- The ℓ_∞ relaxation induces undesirable clustering artifacts of the coefficients absolute values.

Previous relaxation result

(Bach, 2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball
- The choice of ℓ_∞ seems arbitrary
- The ℓ_∞ relaxation induces undesirable clustering artifacts of the coefficients absolute values.

What happens in the non-submodular case?

Previous relaxation result

(Bach, 2010) showed that if F is a submodular function, it is possible to construct the "tightest" convex relaxation of the penalty F for vectors $w \in \mathbb{R}^d$ such that $\|w\|_\infty \leq 1$.

Limitations and open issues:

The relaxation is defined on the unit ℓ_∞ ball.

- Seems to implicitly assume that the w to be estimated is in a fixed ℓ_∞ ball
- The choice of ℓ_∞ seems arbitrary
- The ℓ_∞ relaxation induces undesirable clustering artifacts of the coefficients absolute values.

What happens in the non-submodular case?

Penalizing *and* regularizing...

Given a function $F : 2^V \to \bar{\mathbb{R}}_+$, consider for $\nu, \mu > 0$ the combined penalty:

$$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p.$$

Penalizing *and* regularizing...

Given a function $F : 2^V \to \bar{\mathbb{R}}_+$, consider for $\nu, \mu > 0$ the combined penalty:

$$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p.$$

Motivations

- Compromise between variable selection and smooth regularization

Penalizing *and* regularizing...

Given a function $F : 2^V \to \bar{\mathbb{R}}_+$, consider for $\nu, \mu > 0$ the combined penalty:

$$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p.$$

Motivations

- Compromise between variable selection and smooth regularization
- Required for functions F allowing large supports such as $A \mapsto 1_{\{A \neq \varnothing\}}$

Penalizing *and* regularizing...

Given a function $F : 2^V \to \bar{\mathbb{R}}_+$, consider for $\nu, \mu > 0$ the combined penalty:

$$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p.$$

Motivations

- Compromise between variable selection and smooth regularization
- Required for functions F allowing large supports such as $A \mapsto 1_{\{A \neq \varnothing\}}$

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\rightarrow$ scale invariance.

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\rightarrow$ scale invariance.

Definition (Homogeneous extension of a function g)

$$g_h : x \mapsto \inf_{\lambda > 0} \frac{1}{\lambda} g(\lambda x).$$

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\rightarrow$ scale invariance.

Definition (Homogeneous extension of a function g)

$$g_h : x \mapsto \inf_{\lambda > 0} \frac{1}{\lambda} g(\lambda x).$$

Proposition

The tightest convex positively homogeneous lower bound of a function g is the convex envelope of g_h.

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\rightarrow$ scale invariance.

Definition (Homogeneous extension of a function g)

$$g_h : x \mapsto \inf_{\lambda > 0} \frac{1}{\lambda} g(\lambda x).$$

Proposition

The tightest convex positively homogeneous lower bound of a function g is the convex envelope of g_h.

Leads us to consider:

$$\mathrm{pen}_h(w) \quad = \quad \inf_{\lambda > 0} \frac{1}{\lambda} \left(\mu F(\mathrm{Supp}(\lambda w)) + \nu \|\lambda w\|_p^p \right)$$

A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\mathrm{pen}(w)$.
- Require as well that it is *positively homogeneous* $\rightarrow$ scale invariance.

Definition (Homogeneous extension of a function g)

$$g_h : x \mapsto \inf_{\lambda > 0} \frac{1}{\lambda} g(\lambda x).$$

Proposition

The tightest convex positively homogeneous lower bound of a function g is the convex envelope of g_h.

Leads us to consider:

$$\begin{aligned} \mathrm{pen}_h(w) &= \inf_{\lambda > 0} \frac{1}{\lambda} \left(\mu F(\mathrm{Supp}(\lambda w)) + \nu \|\lambda w\|_p^p \right) \\ &\propto \Theta(w) := \|w\|_p F(\mathrm{Supp}(w))^{1/q} \qquad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1. \end{aligned}$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

with unit ball: $\mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

with unit ball: $\mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$

Proposition

The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p F(\mathrm{Supp}(w))^{1/q}$.

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

with unit ball: $\mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$

Proposition

The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p F(\mathrm{Supp}(w))^{1/q}$.

Proof.

Denote $\Theta(w) = \|w\|_p F(\mathrm{Supp}(w))^{1/q}$:

$$\Theta^*(s) \quad = \quad \max_{w \in \mathbb{R}^d} w^\top s - \|w\|_p F(\mathrm{Supp}(w))^{1/q}$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

with unit ball: $\mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$

Proposition

The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p F(\mathrm{Supp}(w))^{1/q}$.

Proof.

Denote $\Theta(w) = \|w\|_p F(\mathrm{Supp}(w))^{1/q}$:

$$\begin{aligned} \Theta^*(s) &= \max_{w \in \mathbb{R}^d} w^\top s - \|w\|_p F(\mathrm{Supp}(w))^{1/q} \\ &= \max_{A \subset V} \max_{w_A \in \mathbb{R}^A} w_A^\top s_A - \|w_A\|_p F(A)^{1/q} \end{aligned}$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

with unit ball: $\mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$

Proposition

The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p F(\mathrm{Supp}(w))^{1/q}$.

Proof.

Denote $\Theta(w) = \|w\|_p F(\mathrm{Supp}(w))^{1/q}$:

$$\begin{aligned}
\Theta^*(s) &= \max_{w \in \mathbb{R}^d} w^\top s - \|w\|_p F(\mathrm{Supp}(w))^{1/q} \\
&= \max_{A \subset V} \max_{w_A \in \mathbb{R}^A} w_A^\top s_A - \|w_A\|_p F(A)^{1/q} \\
&= \max_{A \subset V} \iota_{\{\|s_A\|_q \leqslant F(A)^{1/q}\}}
\end{aligned}$$

Envelope of the homogeneous penalty Θ

Consider Ω_p with dual norm

$$\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}.$$

with unit ball: $\mathcal{B}_{\Omega_p^*} := \left\{ s \in \mathbb{R}^d \mid \forall A \subset V,\ \|s_A\|_q^q \leq F(A) \right\}$

Proposition

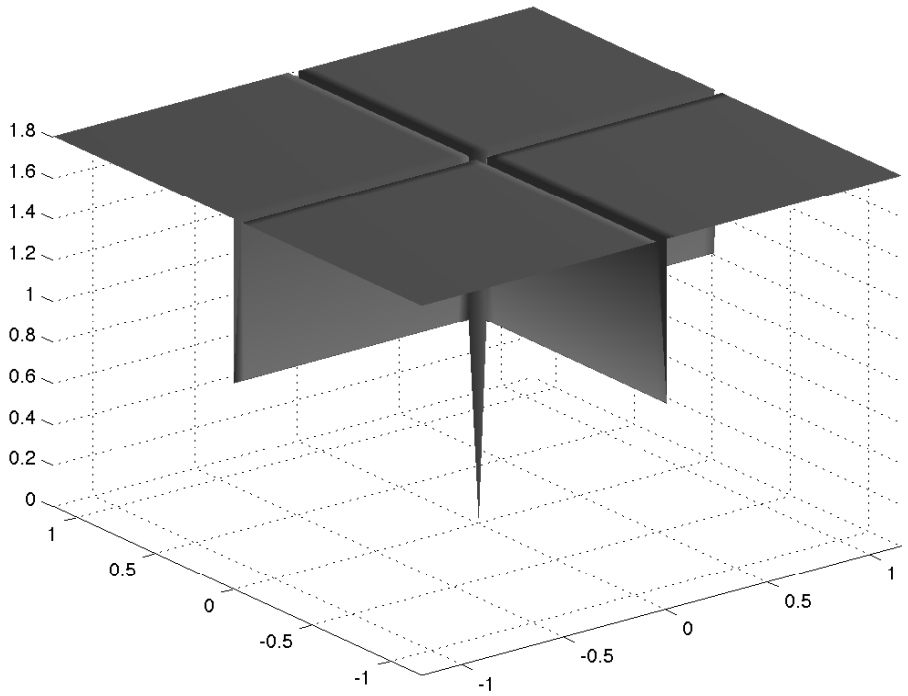
The norm Ω_p is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p F(\mathrm{Supp}(w))^{1/q}$.

Proof.

Denote $\Theta(w) = \|w\|_p F(\mathrm{Supp}(w))^{1/q}$:

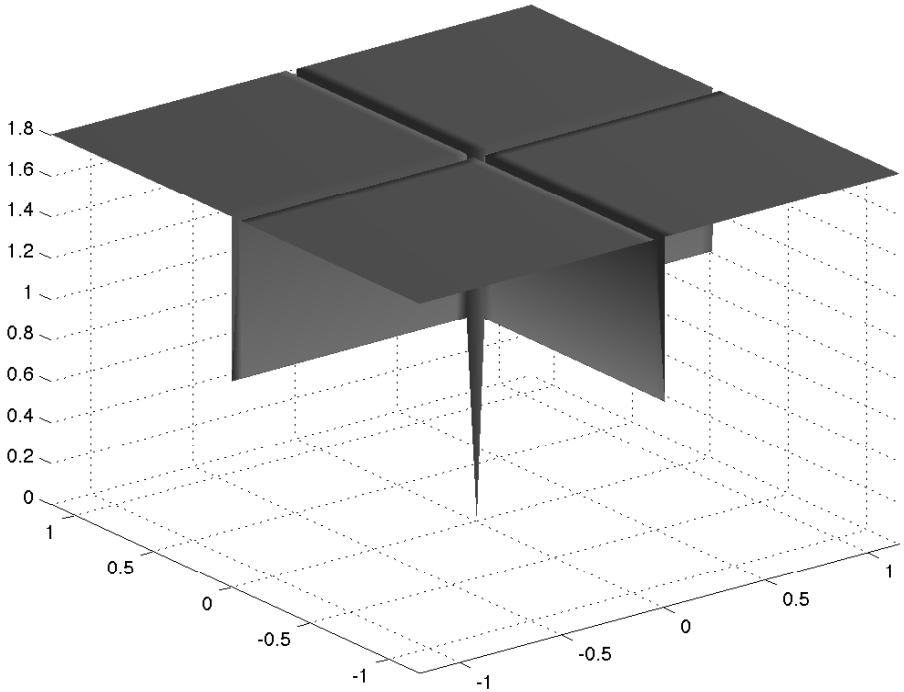
$$\begin{aligned}
\Theta^*(s) &= \max_{w \in \mathbb{R}^d} w^\top s - \|w\|_p F(\mathrm{Supp}(w))^{1/q} \\
&= \max_{A \subset V} \max_{w_A \in \mathbb{R}^A} w_A^\top s_A - \|w_A\|_p F(A)^{1/q} \\
&= \max_{A \subset V} \iota_{\{\|s_A\|_q \leqslant F(A)^{1/q}\}} = \iota_{\{\Omega_p^*(s) \leqslant 1\}}
\end{aligned}$$

Graphs of the different penalties for $w \in \mathbb{R}^2$

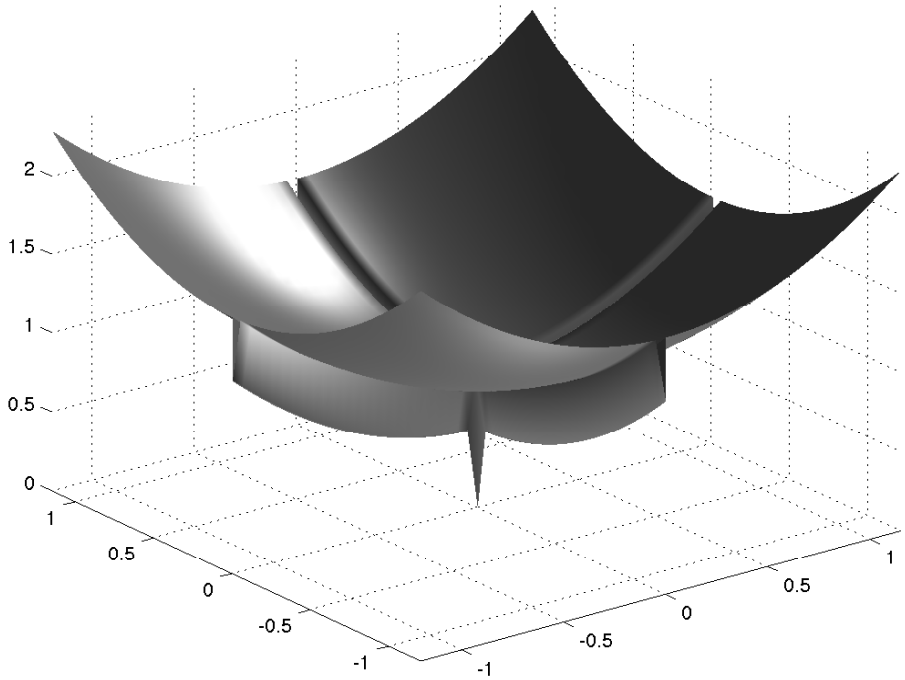


$F(\mathrm{Supp}(w))$

Graphs of the different penalties for $w \in \mathbb{R}^2$

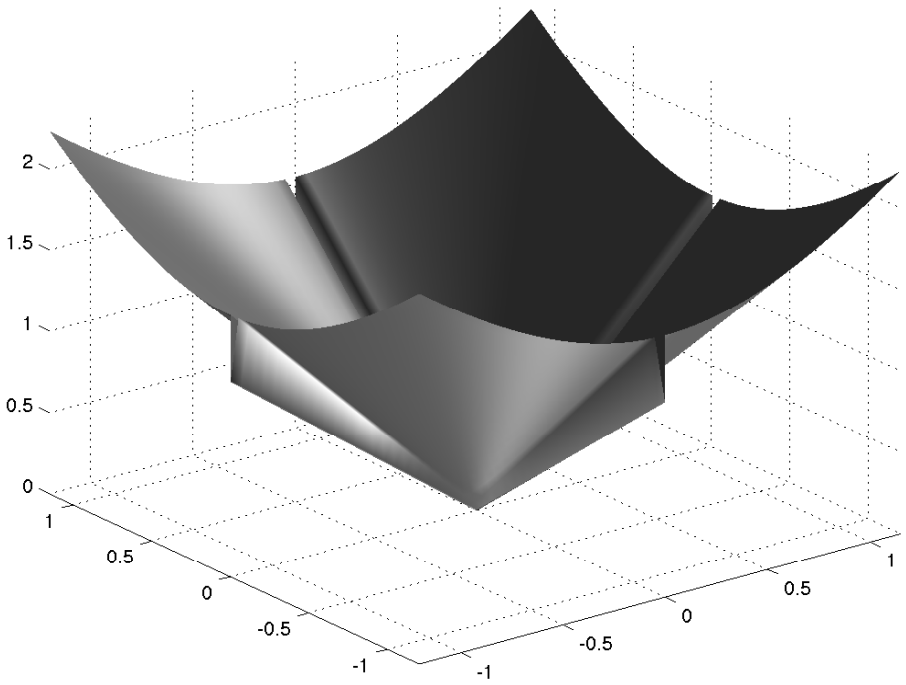


$F(\mathrm{Supp}(w))$



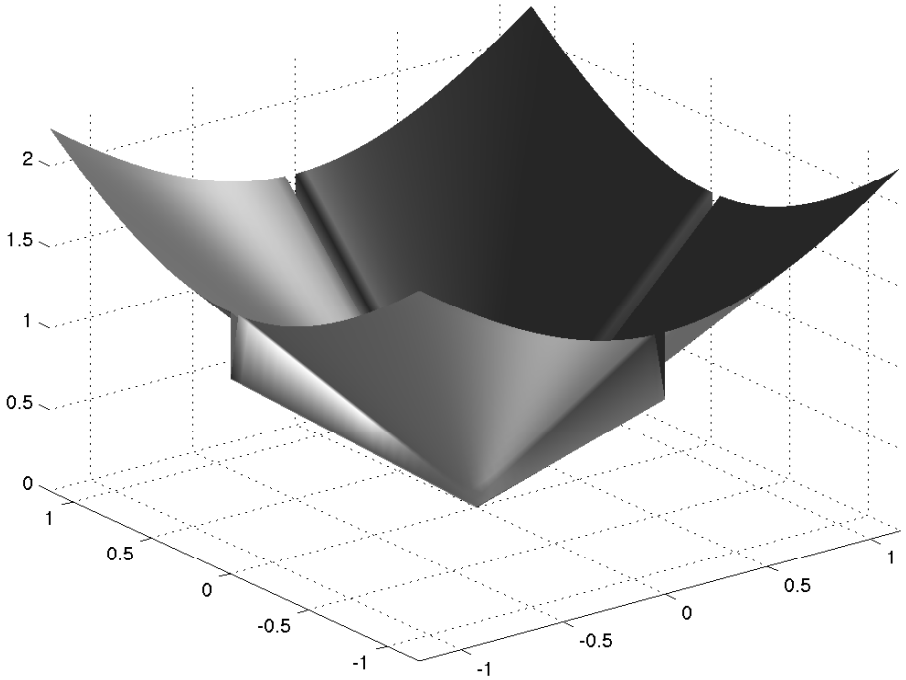
$\mathrm{pen}(w) = \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_2^2$

Graphs of the different penalties for $w \in \mathbb{R}^2$

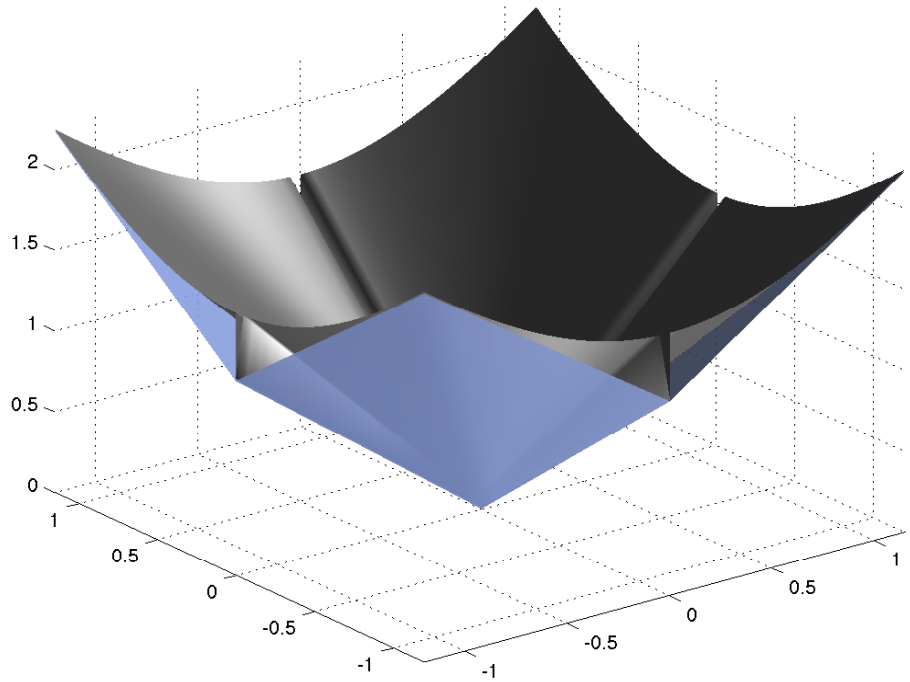


$$\Theta(w) = \sqrt{F(\mathrm{Supp}(w))}\|w\|_2$$

Graphs of the different penalties for $w \in \mathbb{R}^2$



$$\Theta(w) = \sqrt{F(\mathrm{Supp}(w))}\|w\|_2$$

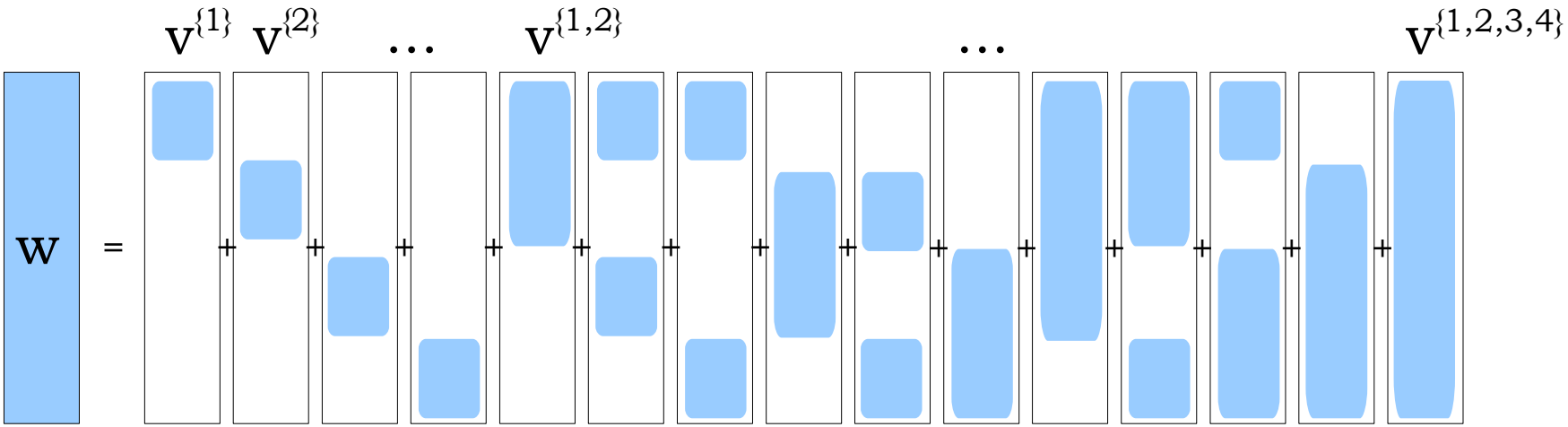


$\Omega^F(w)$

A large latent group Lasso (Jacob et al., 2009)

$$\mathcal{V} = \{v = (v^A)_{A \subset V} \in \left(\mathbb{R}^V\right)^{2^V} \text{ s.t. } \operatorname{Supp}(v^A) \subset A\}$$

$$\Omega_p(w) \quad = \quad \min_{v \in \mathcal{V}} \sum_{A \subset V} F(A)^{\frac{1}{q}} \|v^A\|_p \quad \text{s.t.} \quad w = \sum_{A \subset V} v^A,$$



Some simple examples

	F	Ω_p
	$\lvert A\rvert$	$\lVert w\rVert_1$
	$1_{\{A\neq\varnothing\}}$	$\lVert w\rVert_p$
If $\mathcal{G}$ is a partition of $\{1,\dots,d\}$:	$\sum_{B\in\mathcal{G}} 1_{\{A\cap B\neq\varnothing\}}$	$\sum_{B\in\mathcal{G}} \lVert w_B\rVert_p$

Some simple examples

	F	Ω_p
	$\lvert A\rvert$	$\lVert w\rVert_1$
	$1_{\{A\neq\varnothing\}}$	$\lVert w\rVert_p$
If $\mathcal{G}$ is a partition of $\{1,\dots,d\}$:	$\sum_{B\in\mathcal{G}} 1_{\{A\cap B\neq\varnothing\}}$	$\sum_{B\in\mathcal{G}} \lVert w_B\rVert_p$

- When $p=\infty$ and F is submodular, our relaxation coincides with that of Bach (2010).

Some simple examples

	F	Ω_p
	$\lvert A\rvert$	$\lVert w\rVert_1$
	$1_{\{A\neq\varnothing\}}$	$\lVert w\rVert_p$
If $\mathcal{G}$ is a partition of $\{1,\dots,d\}$:	$\sum_{B\in\mathcal{G}} 1_{\{A\cap B\neq\varnothing\}}$	$\sum_{B\in\mathcal{G}} \lVert w_B\rVert_p$

- When $p=\infty$ and F is submodular, our relaxation coincides with that of Bach (2010).
- However, when $\mathcal{G}$ is not a partition and $p<\infty$, Ω_p is not in general an ℓ_1/ℓ_p-norms !

$\rightarrow$ New norms...

How tight is the relaxation? Example: the range function

Consider $V = \{1, \ldots, p\}$ and the function

$$F(A) = \text{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

How tight is the relaxation? Example: the range function

Consider $V = \{1, \ldots, p\}$ and the function

$$F(A) = \mathrm{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

What is its convex relaxation?

How tight is the relaxation? Example: the range function

Consider $V = \{1, \dots, p\}$ and the function

$$F(A) = \text{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

What is its convex relaxation?

$\Rightarrow$ $\Omega_p^F(w) = \|w\|_1$

How tight is the relaxation? Example: the range function

Consider $V = \{1, \dots, p\}$ and the function

$$F(A) = \text{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

What is its convex relaxation?

$\Rightarrow$ $\Omega_p^F(w) = \|w\|_1$

The relaxation fails

How tight is the relaxation? Example: the range function

Consider $V = \{1, \ldots, p\}$ and the function

$$F(A) = \text{range}(A) = max(A) - min(A) + 1.$$

$\rightarrow$ Leads to the selection of interval patterns.

What is its convex relaxation?

$\Rightarrow$ $\Omega_p^F(w) = \|w\|_1$

The relaxation fails

- New concept of Lower Combinatorial envelope provides a tool to analyze the tightness of the relaxation.

Submodular penalties

A function $F : 2^V \mapsto \mathbb{R}$ is *submodular* if

$$\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cup B) + F(A \cap B) \tag{1}$$

For these functions $\Omega_\infty^F(w) = f(|w|)$ for f the Lovász extension of F.

Properties of submodular function

- f is computed efficiently (via the so-called "greedy" algorithm)
- decomposition ("weak" separability) properties
- F and f can be minimized in polynomial time.

Submodular penalties

A function $F : 2^V \mapsto \mathbb{R}$ is *submodular* if

$$\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cup B) + F(A \cap B) \tag{1}$$

For these functions $\Omega_\infty^F(w) = f(|w|)$ for f the Lovász extension of F.

Properties of submodular function

- f is computed efficiently (via the so-called "greedy" algorithm)
- decomposition ("weak" separability) properties
- F and f can be minimized in polynomial time.

... leads to properties of the corresponding submodular norms

- Regularized empirical risk minimization problems solved efficiently
- Statistical guarantees in terms of consistency and support recovery.

Theoretical results: consistency and fast rates

Result of consistency for the Lasso (Bickel et al., 2009)

- Assume that $y = Xw^* + \sigma\varepsilon$, with $X \in \mathbb{R}^{n \times d}$, $\varepsilon \sim \mathcal{N}(0, Id_n)$ and $\mathrm{Supp}(w^*) = s$
- For an appropriate choice of $\lambda = A\sigma\sqrt{\frac{\log d}{n}}$,
- Under the so-called ℓ_1-Restricted Eigenvalue condition on the design X

$$\frac{1}{n}\|X\hat{w} - Xw^*\|_2^2 = O\Big(\frac{s \log d}{n}\Big).$$

Theoretical results: consistency and fast rates

Result of consistency for the Lasso (Bickel et al., 2009)

- Assume that $y = Xw^* + \sigma\varepsilon$, with $X \in \mathbb{R}^{n \times d}$, $\varepsilon \sim \mathcal{N}(0, Id_n)$ and $\mathrm{Supp}(w^*) = s$
- For an appropriate choice of $\lambda = A\sigma\sqrt{\frac{\log d}{n}}$,
- Under the so-called ℓ_1-Restricted Eigenvalue condition on the design X

$$\frac{1}{n}\|X\hat{w} - Xw^*\|_2^2 = O\Big(\frac{s \log d}{n}\Big).$$

$\rightarrow$ The result can be generalized to all "submodular norms" via a generalization of the ℓ_1-Restricted-Eigenvalue condition with rates depending on the concentration of $\Omega_p^*(\varepsilon)$.

Theoretical results: support recovery

Support Recovery for the Lasso (Zhao and Yu, 2006)

- Assume that $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, Id_n)$.
- For an appropriate choice of λ
- If the non-zero coefficients of w are sufficiently large w.r.t. the noise level
- Under the so-called Irrepresentability Condition on the design X,

$$\mathrm{Supp}(\hat{w}) = \mathrm{Supp}(w^*) \qquad \text{with high probability.}$$

Theoretical results: support recovery

Support Recovery for the Lasso (Zhao and Yu, 2006)

- Assume that $y = Xw^* + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, Id_n)$.
- For an appropriate choice of λ
- If the non-zero coefficients of w are sufficiently large w.r.t. the noise level
- Under the so-called Irrepresentability Condition on the design X,

$$\mathrm{Supp}(\hat{w}) = \mathrm{Supp}(w^*) \qquad \text{with high probability.}$$

$\rightarrow$ The result can be generalized to all "submodular-norms" via a generalization of the Irrepresentability Condition.

Summary

- A convex relaxation for functions penalizing
  (a) the support via a general set function
  (b) the ℓ_p norm of the parameter vector w.
- Principled construction of:
  - known norms like the group Lasso or ℓ_1/ℓ_p-norm
  - many new sparsity inducing norms
- Caveat: the relaxation can fail to capture the structure (e.g. range function)
- For submodular functions we can obtain efficient algorithms, and theoretical results such as consistency and support recovery guarantees.

References I

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

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

Huang, J., Zhang, T., and Metaxas, D. (2011). Learning with structured sparsity. *The JMLR*, 12:3371–3412.

Jacob, L., Obozinski, G., and Vert, J. (2009). Group lasso with overlap and graph lasso. In *ICML*.

Jenatton, R., Audibert, J., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *JMLR*, 12:2777–2824.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B*, 68:49–67.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *JMLR*, 7:2541–2563.