Convex relaxation for Combinatorial Penalties

Guillaume Obozinski

Equipe Imagine
Laboratoire d’Informatique Gaspard Monge
Ecole des Ponts - ParisTech

Joint work with Francis Bach

Fête Parisienne in Computation, Inference and Optimization
IHES - March 20th 2013
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:**

\[
\text{Supp}(w) = \{ i \mid w_i \neq 0 \}.
\]
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:**

\[
\text{Supp}(w) = \{ i \mid w_i \neq 0 \}.
\]

**Penalization for variable selection**

\[
\min_{w \in \mathbb{R}^d} L(w) + \lambda |\text{Supp}(w)|
\]
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:
  \[
  \text{Supp}(w) = \{ i \mid w_i \neq 0 \}.
  \]

Penalization for variable selection

\[
\min_{w \in \mathbb{R}^d} L(w) + \lambda |\text{Supp}(w)|
\]
From sparsity...

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

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

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

- Support of the model:

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

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda |\text{Supp}(w)|$$
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:**

\[
\text{Supp}(w) = \{ i \mid w_i \neq 0 \}.
\]

**Penalization for variable selection**

\[
\min_{w \in \mathbb{R}^d} L(w) + \lambda |\text{Supp}(w)|
\]

**Convex relaxation for Combinatorial Penalties**
From sparsity...

- **Empirical risk:** for $w \in \mathbb{R}^d$, 
  \[ L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2 \]
  \[ |\text{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}} \]

- **Support of the model:**
  \[ \text{Supp}(w) = \{ i \mid w_i \neq 0 \} \]

**Penalization for variable selection**
\[
\min_{w \in \mathbb{R}^d} L(w) + \lambda |\text{Supp}(w)|
\]
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:**

\[
\text{Supp}(w) = \{ i \mid w_i \neq 0 \}.
\]

**Penalization for variable selection**

\[
\min_{w \in \mathbb{R}^d} L(w) + \lambda |\text{Supp}(w)|
\]

**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*. 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.
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.
Difficult inverse problem in Brain Imaging

Scale 6 - Fold 9

-5.00e-02 5.00e-02

Jenatton et al. (2012)
Hierarchical Dictionary Learning

**Figure:** Hierarchical dictionary of image patches

**Figure:** Hierarchical Topic model

Mairal, Jenatton, Obozinski and Bach (2010)

Convex relaxation for Combinatorial Penalties
Ideas in structured sparsity
Group Lasso and $\ell_1/\ell_p$ 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 $\ell_1/\ell_p$ norm  (Yuan and Lin, 2006)

Group Lasso

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

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

Overlapping groups: direct extension of Jenatton et al. (2011).

Interesting induced structures

→ Induce patterns of rooted subtree
→ Induce “convex” patterns on a grid
Hierarchical Norms (Zhao et al., 2009; Bach, 2008)

(Jenatton, Mairal, Obozinski and Bach, 2010a)

- A covariate can only be selected after its ancestors
- Structure on parameters $w$
Hierarchical Norms (Zhao et al., 2009; Bach, 2008)

A covariate can only be selected after its ancestors

Structure on parameters $w$

Hierarchical penalization: $\Omega(w) = \sum_{g \in G} \|w_g\|_2$ where groups $g$ in $G$ are equal to the set of descendants of some nodes in a tree.
A new approach based on combinatorial functions
General framework

Let \( V = \{1, \ldots, d\} \).
Given a set function \( F : 2^V \rightarrow \mathbb{R}_+ \).

Examples of combinatorial functions
Use recursivity or counts of structures (e.g. tree) with DP
Block-coding (Huang et al., 2011)
\[ \tilde{G}(A) = \min \sum F(B_i) \text{ s.t. } B_1 \cup \ldots \cup B_k \supset A \]
Submodular functions
(Work on convex relaxations by Bach (2010))
Convex relaxation for Combinatorial Penalties
General framework

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

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

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

\[
\min_{w \in \mathbb{R}^d} L(w) + F(\text{Supp}(w))
\]

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?

$$\min_{w \in \mathbb{R}^d} L(w) + F(\text{Supp}(w))$$
A relaxation for $F$...

How to solve?

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

→ Greedy algorithms
→ Non-convex methods
→ Relaxation
A relaxation for $F$...?

How to solve?

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

→ Greedy algorithms
→ Non-convex methods
→ Relaxation

| $|A|$ | $F(A)$ |
|-----|--------|
| $L(w) + \lambda |\text{Supp}(w)|$ |        |
A relaxation for $F$...

How to solve?

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

→ Greedy algorithms
→ Non-convex methods
→ Relaxation

| $|A|$ | $F(A)$ |
|------|--------|
| $L(w) + \lambda |\text{Supp}(w)|$ | $L(w) + \lambda \|w\|_1$ |
| $\downarrow$ | |

Convex relaxation for Combinatorial Penalties
A relaxation for $F$...?

How to solve?

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

→ Greedy algorithms
→ Non-convex methods
→ Relaxation

| $|A|$ | $F(A)$ |
|----------------------------------|----------------------------------|
| $L(w) + \lambda |\text{Supp}(w)|$ | $L(w) + \lambda F(\text{Supp}(w))$ |
| ↓                                | ↓                                |
| $L(w) + \lambda \|w\|_1$          | $L(w) + \lambda \|w\|_1$          |
A relaxation for $F$…?

How to solve?

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

→ Greedy algorithms
→ Non-convex methods
→ Relaxation

| $|A|$ | $F(A)$ |
|---|---|
| $L(w) + \lambda |\text{Supp}(w)|$ | $L(w) + \lambda F(\text{Supp}(w))$ |
| $\downarrow$ | $\downarrow \text{?}$ |
| $L(w) + \lambda \|w\|_1$ | $L(w) + \lambda \ldots \text{?} \ldots$ |
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 $\|w\|_\infty \leq 1$. 

Limitations and open issues:

The relaxation is defined on the unit $\ell_\infty$ ball.

Seems to implicitly assume that the $w$ to be estimated is in a fixed $\ell_\infty$ ball.

The choice of $\ell_\infty$ seems arbitrary.

The $\ell_\infty$ relaxation induces undesirable clustering artifacts of the coefficients absolute values.

What happens in the non-submodular case?
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 $\ell_\infty$ ball.

- Seems to implicitly assume that the $w$ to be estimated is in a fixed $\ell_\infty$ 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 $\ell_\infty$ ball.

- Seems to implicitly assume that the $w$ to be estimated is in a fixed $\ell_\infty$ ball
- The choice of $\ell_\infty$ 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 $\ell_\infty$ ball.

- Seems to implicitly assume that the $w$ to be estimated is in a fixed $\ell_\infty$ ball
- The choice of $\ell_\infty$ seems arbitrary
- The $\ell_\infty$ 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 $\ell_{\infty}$ ball.

- Seems to implicitly assume that the $w$ to be estimated is in a fixed $\ell_{\infty}$ ball
- The choice of $\ell_{\infty}$ seems arbitrary
- The $\ell_{\infty}$ 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 $\ell_{\infty}$ ball.

- Seems to implicitly assume that the $w$ to be estimated is in a fixed $\ell_{\infty}$ ball
- The choice of $\ell_{\infty}$ seems arbitrary
- The $\ell_{\infty}$ relaxation induces undesirable clustering artifacts of the coefficients absolute values.

What happens in the non-submodular case?
Penalizing and regularizing...

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

$$\text{pen}(w) = \mu F(\text{Supp}(w)) + \nu \|w\|_p^p.$$
Penalizing *and* regularizing...

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

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

**Motivations**

- Compromise between variable selection and smooth regularization
Penalizing and regularizing...

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

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

Motivations

- Compromise between variable selection and smooth regularization
- Required for $F$ allowing large supports such as $A \mapsto 1\{A \neq \emptyset\}$
Penalizing *and* regularizing...

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

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

**Motivations**

- Compromise between variable selection and smooth regularization
- Required for $F$ allowing large supports such as $A \mapsto 1\{A \neq \emptyset\}$
- Leads to a penalty which is *coercive* so that a convex relaxation on $\mathbb{R}^d$ will not be trivial.
A convex and *homogeneous* relaxation

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

Definition (Homogeneous extension of a function $g$)

$$g_{h}: x \mapsto \inf_{\lambda > 0} \lambda g(\lambda x).$$

Proposition

The tightest convex positively homogeneous lower bound of a function $g$ is the convex envelope of $g_{h}$. 

Leads us to consider:

$$\text{pen}_{h}(w) = \inf_{\lambda > 0} \lambda (\mu F(\text{Supp}(\lambda w)) + \nu \|\lambda w\|^{p}) \propto \Theta(w) := \|w\|^{1/q} \text{ with } 1/p + 1/q = 1.$$
A convex and *homogeneous* relaxation

- Looking for a convex relaxation of $\text{pen}(w)$.
- Require as well that it is *positively homogeneous* → 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 \( \text{pen}(w) \).
- 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 $\text{pen}(w)$.
- 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:

$$\text{pen}_h(w) = \inf_{\lambda > 0} \frac{1}{\lambda} \left( \mu F(\text{Supp}(\lambda w)) + \nu \|\lambda w\|_p^p \right)$$
A convex and **homogeneous** relaxation

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

**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:

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

\[
\propto \Theta(w) := \|w\|_p F(\text{Supp}(w))^{1/q} \quad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1.
\]
Envelope of the homogeneous penalty $\Theta$

Consider $\Omega_p$ with dual norm

$$
\Omega^*_p(s) = \max_{A \subseteq V, A \neq \emptyset} \frac{\|SA\|^q}{F(A)^{1/q}}.
$$
Envelope of the homogeneous penalty $\Theta$

Consider $\Omega_p$ with dual norm

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

with unit ball: $B_{\Omega^*_p} := \{s \in \mathbb{R}^d \mid \forall A \subset V, \|s_A\|_q \leq F(A)\}$
Envelope of the homogeneous penalty $\Theta$

Consider $\Omega_p$ with dual norm

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

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

**Proposition**

*The norm $\Omega_p$ is the convex envelope (tightest convex lower bound) of the function $w \mapsto \|w\|_p F(\text{Supp}(w))^{1/q}$.*
Envelope of the homogeneous penalty $\Theta$

Consider $\Omega_p$ with dual norm

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

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

**Proposition**

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

**Proof.**

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

$$\Theta^*(s) = \max_{w \in \mathbb{R}^d} w^\top s - \|w\|_p F(\text{Supp}(w))^{1/q}$$
Envelope of the homogeneous penalty $\Theta$

Consider $\Omega_p$ with dual norm

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

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

**Proposition**

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

**Proof.**

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

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

$$= \max_{A \subset V} \max_{w_A \in \mathbb{R}^A} w_A^\top s_A - \|w_A\|_p F(A)^{1/q}$$
Envelope of the homogeneous penalty $\Theta$

Consider $\Omega_p$ with dual norm

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

with unit ball: $B_{\Omega_p^*} := \{s \in \mathbb{R}^d \mid \forall A \subseteq V, \|s_A\|_q \leq F(A)\}$

**Proposition**

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

**Proof.**

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

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

$$= \max_{A \subseteq V} \max_{w_A \in \mathbb{R}^A} w_A^\top s_A - \|w_A\|_p F(A)^{1/q}$$

$$= \max_{A \subseteq V} \ell \{\|s_A\|_q \leq F(A)^{1/q}\}$$
Envelope of the homogeneous penalty $\Theta$

Consider $\Omega_p$ with dual norm

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

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

Proposition

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

Proof.

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

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

$$= \max_{A \subset V} \max_{w_A \in \mathcal{R}^A} w_A^\top s_A - \|w_A\|_p F(A)^{1/q}$$

$$= \max_{A \subset V} l\{\|s_A\|_q \leq F(A)^{1/q}\} = l\{\Omega_p^*(s) \leq 1\}$$
Graphs of the different penalties for $w \in \mathbb{R}^2$

$$F(\text{Supp}(w))$$
Graphs of the different penalties for $w \in \mathbb{R}^2$

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

$\text{pen}(w) = \mu F(\text{Supp}(w)) + \nu \|w\|_2^2$
Graphs of the different penalties for $w \in \mathbb{R}^2$

$$
\Theta(w) = \sqrt{F(\text{Supp}(w))}\|w\|_2
$$
Graphs of the different penalties for $w \in \mathbb{R}^2$

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

$$\Omega^F(w)$$
A large latent group Lasso (Jacob et al., 2009)

\[ \mathcal{V} = \{ \mathbf{v} = (v^{A})_{A \subset \mathcal{V}} \in (\mathbb{R}^{\mathcal{V}})^{2^{\mathcal{V}}} \text{ s.t. } \text{Supp}(v^{A}) \subset A \} \]

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

Convex relaxation for Combinatorial Penalties
Some simple examples

<table>
<thead>
<tr>
<th>$F$</th>
<th>$\Omega_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>A</td>
</tr>
<tr>
<td>$1{A\neq\emptyset}$</td>
<td>$|w|_p$</td>
</tr>
</tbody>
</table>

If $\mathcal{G}$ is a partition of $\{1, \ldots, d\}$:

$$\sum_{B \in \mathcal{G}} 1\{A \cap B \neq \emptyset\} \quad \sum_{B \in \mathcal{G}} \|w_B\|_p$$

When $p = \infty$ and $F$ is submodular, our relaxation coincides with that of Bach (2010). However, when $\mathcal{G}$ is not a partition and $p < \infty$, $\Omega_p$ is not in general $\ell_1/\ell_p$-norms!
Some simple examples

<table>
<thead>
<tr>
<th>$F$</th>
<th>$\Omega_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>A</td>
</tr>
<tr>
<td>$1{A \neq \emptyset}$</td>
<td>$|w|_p$</td>
</tr>
</tbody>
</table>

If $G$ is a partition of $\{1, \ldots, d\}$: $\sum_{B \in G} 1\{A \cap B \neq \emptyset\} = \sum_{B \in G} \|w_B\|_p$

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

<table>
<thead>
<tr>
<th>$F$</th>
<th>$\Omega_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>A</td>
</tr>
<tr>
<td>$1{A\neq\emptyset}$</td>
<td>$|w|_p$</td>
</tr>
</tbody>
</table>

If $\mathcal{G}$ is a partition of $\{1, \ldots, d\}$: \[
\sum_{B \in \mathcal{G}} 1\{A \cap B \neq \emptyset\} \quad \sum_{B \in \mathcal{G}} \|w_B\|_p
\]

- When $p = \infty$ and $F$ is submodular, our relaxation coincides with that of Bach (2010).
- However, when $\mathcal{G}$ is not a partition and $p < \infty$, $\Omega_p$ is not in general an $\ell_1/\ell_p$-norms!

→ New norms... e.g. the $k$-support norm of Argyriou et al. (2012).
Example

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

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

- $F(\{1, 2\}) = 1$,
- $F(\{1, 3\}) = 1$,
- $F(\{2, 3\}) = 1$,
- $F(A) = \infty$ or defined by block-coding.
Example

Consider \( V = \{1, 2, 3\} \).

\[ G = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\} \]

- \( F(\{1, 2\}) = 1 \),
- \( F(\{1, 3\}) = 1 \),
- \( F(\{2, 3\}) = 1 \),
- \( F(A) = \infty \) or defined by block-coding.
How tight is the relaxation? Example: the range function

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

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

→ Leads to the selection of interval patterns.

⇒ $\Omega_F(p(w)) = \|w\|_1$
How tight is the relaxation? Example: the range function

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

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

→ Leads to the selection of interval patterns.

What is its convex relaxation?
How tight is the relaxation? Example: the range function

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

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

→ Leads to the selection of interval patterns.

What is its convex relaxation?

⇒ $\Omega^F_p(w) = \|w\|_1$
How tight is the relaxation? Example: the range function

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

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

→ Leads to the selection of interval patterns.

What is its convex relaxation?

\[
\Omega_F^p(w) = \|w\|_1
\]

The relaxation fails
How tight is the relaxation? Example: the range function

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

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

→ Leads to the selection of interval patterns.

What is its convex relaxation?

⇒ $\Omega_p^F(w) = \|w\|_1$

The relaxation fails

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

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

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

(1)

For these functions $\Omega^F_{\infty}(w) = f(|w|)$ 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

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

For these functions $\Omega^F_\infty(w) = f(|w|)$ 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.
Consistency for the Lasso (Bickel et al., 2009)

- Assume that $y = Xw^* + \sigma \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \text{Id}_n)$
- Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{d \times d}$.
- Denote $J = \text{Supp}(w^*)$.
- Assume the $\ell_1$-Restricted Eigenvalue condition:
  \[ \forall \Delta \text{ s.t. } \|\Delta_{J^c}\|_1 \leq 3 \|\Delta_J\|_1, \quad \Delta^\top Q \Delta \geq \kappa \|\Delta_J\|_1^2. \]

Then we have
\[
\frac{1}{n} \|X\hat{w} - Xw^*\|_2^2 \leq \frac{72|J|\sigma^2}{\kappa} \left( \frac{2 \log p + t^2}{n} \right),
\]
with probability larger than $1 - \exp(-t^2)$. 
Support Recovery for the Lasso (Wainwright, 2009)

- Assume $y = Xw^* + \sigma \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \text{Id}_n)$
- Let $Q = \frac{1}{n} X^\top X \in \mathbb{R}^{d \times d}$.
- Denote by $J = \text{Supp}(w^*)$.
- Define $\nu = \min_{j, w_j^* \neq 0} |w_j^*| > 0$
- Assume $\kappa = \lambda_{\min}(Q_{JJ}) > 0$
- Assume the Irrepresentability Condition, i.e., that for $\eta > 0$,
  \[ \| Q_{JJ}^{-1} Q_{JJ^c} \|_{\infty, \infty} \leq 1 - \eta. \]

Then, if $\frac{2}{\eta} \sqrt{\frac{2\sigma^2 \log(p)}{n}} < \lambda < \frac{\kappa \nu}{|J|}$, the minimizer $\hat{w}$ is unique and has support equal to $J$, with probability larger than $1 - 4 \exp(-c_1 n \lambda^2)$. 
An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of \( \{1, \ldots, p\} \)
An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of \( \{1, \ldots, p\} \)

Natural associated penalization:

\[
F(A) = \text{range}(A) = i_{\text{max}}(A) - i_{\text{min}}(A) + 1.
\]
An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of \{1, \ldots, p\}

Natural associated penalization:

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

\[ \rightarrow F \text{ is not submodular...} \]
An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of \( \{1, \ldots, p\} \)

Natural associated penalization:

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

\( \rightarrow \) \( F \) is not submodular...

\( \rightarrow \) \( G(A) = |A| \)
An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of \( \{1, \ldots, p\} \)

**Natural associated penalization:**

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

\( \rightarrow \) \( F \) is not submodular...

\( \rightarrow \) \( G(A) = |A| \)

But \( F(A) := d - 1 + \text{range}(A) \) is submodular!
An example: penalizing the range

Structured prior on support (Jenatton et al., 2011):

- the support is an interval of \( \{1, \ldots, p\} \)

Natural associated penalization:

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

\[ \rightarrow F \text{ is not submodular...} \]

\[ \rightarrow G(A) = |A| \]

But \( F(A) := d - 1 + \text{range}(A) \) is submodular!

In fact \( F(A) = \sum_{B \in G} 1_{\{A \cap B \neq \emptyset\}} \) for \( B \) of the form:

Jenatton et al. (2011) considered \( \Omega(w) = \sum_{B \in B} \|w_B \circ d_B\|_2 \).
Experiments

Figure: Signals

- $S_1$ constant
- $S_2$ triangular shape
- $S_3$ $\mapsto |\sin(x)\sin(5x)|$
- $S_4$ a slope pattern
- $S_5$ i.i.d. Gaussian pattern

Compare:
- Lasso
- Elastic Net
- Naive $\ell_2$ group-Lasso
- $\Omega_2$ for $F(A) = d - 1 + \text{range}(A)$
- $\Omega_\infty$ for $F(A) = d - 1 + \text{range}(A)$
- The weighted $\ell_2$ group-Lasso of (Jenatton et al., 2011).
Constant signal

\[ d = 256, \ k = 160, \ \sigma = 0.5 \]

Convex relaxation for Combinatorial Penalties
Triangular signal

Best Hamming $d=256$, $k=160$, $\sigma=0.5$, $S_2$, $\text{cov}=\text{id}$

Convex relaxation for Combinatorial Penalties
\((x_1, x_2) \mapsto |\sin(x_1) \sin(5x_1) \sin(x_2) \sin(5x_2)|\) signal in 2D

**Convex relaxation for Combinatorial Penalties**

- Best Hamming
  - EN
  - GL+w
  - GL
  - L1
  - Sub \(p=\infty\)
  - Sub \(p=2\)

- Parameters:
  - \(d = 256\)
  - \(k = 160\)
  - \(\sigma = 1.0\)
  - \(\sigma = 0.5\)
i.i.d Random signal in 2D

$d=256$, $k=160$, $\sigma=1.0$

Convex relaxation for Combinatorial Penalties
A convex relaxation for functions penalizing
(a) the support via a general set function
(b) the $\ell_p$ norm of the parameter vector $w$.

Principled construction of:
- known norms like the group Lasso or $\ell_1/\ell_p$-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 II

