A totally unimodular view of structured sparsity

Marwa El Halabi  
marwa.elhalabi@epfl.ch

Laboratory for Information and Inference Systems (LIONS)  
École Polytechnique Fédérale de Lausanne (EPFL)  
Switzerland

EPFL-Idiap-ETH Sparsity Workshop 2015

Joint work with: Volkan Cevher
Supervised learning and inverse problems

Running example:

\[
b = Ax + w
\]

Applications: Machine learning, signal processing, theoretical computer science...
Supervised learning and inverse problems

Running example:

\[
\begin{bmatrix}
  b \\
  A \\
  x^\flat \\
  w
\end{bmatrix}
= 
\begin{bmatrix}
  n \\
  p
\end{bmatrix}
\]

A difficult estimation challenge when \( n < p \):

Nullspace (null) of \( A \): \( x^\flat + \delta \rightarrow b, \ \forall \delta \in \text{null}(A) \)

- Needle in a haystack: **We need additional information on** \( x^\flat \)!
Sparsity to the rescue!

\[ \tilde{A} \begin{bmatrix} b \\ y \end{bmatrix} = \begin{bmatrix} \tilde{A} \\ y \end{bmatrix} \]

- \( b \in \mathbb{R}^n, \tilde{A} \in \mathbb{R}^{n \times p} \), and \( n < p \)
Sparsity to the rescue!

\[ \tilde{A} \]

\[ \Psi \]

- \( b \in \mathbb{R}^n \), \( \tilde{A} \in \mathbb{R}^{n \times p} \), and \( n < p \)
- \( \Psi \in \mathbb{R}^{p \times p} \), \( x^{\dagger} \in \Sigma_s \), and \( s < n < p \)
$b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times p}$, $x^\dagger \in \Sigma_s$, and $s < n < p$
Sparsity to the rescue!

\[
\begin{bmatrix}
    \mathbf{b} \\
    \mathbf{A} \\
    \mathbf{x}^\dagger \\
\end{bmatrix}
\]

\[
\begin{array}{ccc}
    n \times 1 & n \times s & s \times 1 \\
\end{array}
\]

- \( \mathbf{b} \in \mathbb{R}^n \), \( \mathbf{A} \in \mathbb{R}^{n \times p} \), \( \mathbf{x}^\dagger \in \Sigma_s \), and \( s < n < p \)

**Impact:** Support restricted columns of \( \mathbf{A} \) leads to an overcomplete system.
Beyond sparsity towards model-based or *structured* sparsity

- The following signals can look the same from a sparsity perspective!

Sparse image  
Wavelet coefficients of a natural image  
Spike train  
Background subtracted image
Beyond sparsity towards model-based or \textit{structured} sparsity

- The following signals can look the \textit{same} from a \textit{sparsity} perspective!

- In reality, these signals have additional \textit{structures} beyond the simple sparsity
Beyond sparsity towards model-based or \textit{structured} sparsity

**Sparsity model:** Union of all \( s \)-dimensional canonical subspaces.

**Structured sparsity model:** A \textbf{particular} union of \( m_s \ s \)-dimensional canonical subspaces.

**Three upshots of structured sparsity:**

1. Reduced sample complexity
2. Better noise robustness
3. Better interpretability
A simple template for linear inverse problems

Find the “sparsest” $x$ subject to structure and data.

- **Sparsity**
  
  We can generalize this desideratum to other notions of simplicity

- **Structure**
  
  We only allow certain sparsity patterns

- **Data fidelity**
  
  We have many choices of convex constraints & losses to represent data; e.g.,

  $$\| b - Ax \|_2 \leq \kappa$$
Simple sparsity

A combinatorial approach for estimating $x^\dagger$ from $b = Ax^\dagger + w$

$$x^* \in \arg \min_{x \in \mathbb{R}^p} \left\{ \| x \|_0 : \| b - Ax \|_2 \leq \| w \|_2 \right\} \quad (P_0)$$

where $\| x \|_0 := 1^T s$, $s = 1_{\text{supp}(x)}$, $\text{supp}(x) = \{ i | x_i \neq 0 \}$
Simple sparsity

A combinatorial approach for estimating $x^\dagger$ from $b = Ax^\dagger + w$

\[
x^* \in \arg\min_{x \in \mathbb{R}^p} \left\{ \|x\|_0 : \|b - Ax\|_2 \leq \|w\|_2 \right\} \quad (P_0)
\]

where $\|x\|_0 := 1^T s$, $s = 1_{\text{supp}(x)}$, $\text{supp}(x) = \{i | x_i \neq 0\}$

\(P_0\) has the following characteristics:

- sample complexity: $O(s)$
- computational effort: NP-Hard
- stability: No
Simple sparsity

A combinatorial approach for estimating $\mathbf{x}^\dagger$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\dagger + \mathbf{w}$

$$\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2 \right\} \quad (\mathcal{P}_0)$$

where $\|\mathbf{x}\|_0 := 1^T s, s = 1_{\text{supp}(\mathbf{x})}, \text{supp}(\mathbf{x}) = \{ i \mid x_i \neq 0 \}$

$\mathcal{P}_0$ has the following characteristics:

- sample complexity: $\mathcal{O}(s)$
- computational effort: NP-Hard
- stability: No

Convex relaxation:

Convex envelope is the largest convex lower bound.

A technicality: Restrict $\mathbf{x}^\dagger \in [-1, 1]^p$. 

$\|\mathbf{x}\|_0$ over the unit $\ell_\infty$-ball
Simple sparsity

A combinatorial approach for estimating $\mathbf{x}^\dagger$ from $\mathbf{b} = A\mathbf{x}^\dagger + \mathbf{w}$

$$\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - A\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2 \right\} \quad (P_0)$$

where $\|\mathbf{x}\|_0 := 1^T s$, $s = 1_{\text{supp}(\mathbf{x})}$, $\text{supp}(\mathbf{x}) = \{i | x_i \neq 0\}$

$\mathcal{P}_0$ has the following characteristics:

- sample complexity: $\mathcal{O}(s)$
- computational effort: NP-Hard
- stability: No

Convex relaxation:
Convex envelope is the largest convex lower bound.

$\|\mathbf{x}\|_1$ is the convex envelope of $\|\mathbf{x}\|_0$

A technicality: Restrict $\mathbf{x}^\dagger \in [-1, 1]^p$. 
The role of convexity: Tractable & stable recovery

A combinatorial approach for estimating $x^\dagger$ from $b = Ax^\dagger + w$

$$x^* \in \arg \min_{x \in \mathbb{R}^p} \left\{ \|x\|_1 : \|b - Ax\|_2 \leq \|w\|_2, \|x\|_\infty \leq 1 \right\} \quad (BP)$$

where $\|x\|_1 := 1^T|x|$
The role of convexity: Tractable & stable recovery

A combinatorial approach for estimating $x^\dagger$ from $b = Ax^\dagger + w$

$$x^* \in \arg \min_{x \in \mathbb{R}^p} \left\{ \|x\|_1 : \|b - Ax\|_2 \leq \|w\|_2, \|x\|_\infty \leq 1 \right\} \quad (BP)$$

where $\|x\|_1 := 1^T|x|$

$\|x\|_1$ is the convex envelope of $\|x\|_0$

BP has the following characteristics [13]:

- sample complexity: $O(s \log(\frac{p}{s}))$
- computational effort: Tractable; $O(n^2 p^{1.5} \log(\frac{1}{\epsilon}))$ via IPM (for $w = 0$)
- stability: Robust to noise

A technicality: Restrict $x^\dagger \in [-1, 1]^p$. 

A TU view of structured sparsity | Marwa El Halabi, marwa.elhalabi@epfl.ch
Convex relaxations in general?

We encode the structure over the support by $g(x) = F(\text{supp}(x))$

- $\text{supp}(x) = \{i | x_i \neq 0\}$
- $F(s) : \{0, 1\}^p \rightarrow \mathbb{R} \cup \{+\infty\}$

How to compute the convex relaxation of $g$ in general?

1. Case by case heuristics
2. Biconjugation ($\equiv$ convex envelope): Fenchel conjugate of Fenchel conjugate.

Recall Fenchel conjugate: $g^*(y) := \sup_{x: \text{dom}(g)} x^T y - g(x)$

**Proposition (Hardness of conjugation)**

*The Fenchel conjugate of $g$ results in the following combinatorial problem*

$$g^*(y) = \sup_{s \in \{0, 1\}^p} |y|^T s - F(s).$$

*which is NP-Hard in general.*
Tractable convex relaxation

Prior work:

1. Monotone submodular penalties [1]
   - Tractable biconjugation via Lovász extension
   - Limited to certain structures

2. $\ell_q$-regularized combinatorial functions [11] ($\mu F(\text{supp}(x)) + \nu \|x\|_q$)
   - Tractable biconjugation even for some non-submodular functions
   - Not always tractable
   - May lose structure

Our work: New framework for tractable convex relaxations

- Easy to design
- Tractable biconjugation via linear programming (LP)
- Applicable to various submodular and non-submodular structures
Template for TU structures

**Sparsity and structure together [5]**
Given some weights $d \in \mathbb{R}^d$, $e \in \mathbb{R}^p$ and an integral vector $c \in \mathbb{Z}^l$, we define

$$g_{TU}(x) := \min_{\omega} \{d^T \omega + e^T s : M \begin{bmatrix} \omega \\ s \end{bmatrix} \leq c, \mathbf{1}_{\text{supp}(x)} = s, \omega \in \{0, 1\}^d\}$$

for all feasible $x$, $\infty$ otherwise. The parameter $\omega$ is useful for latent modeling.

**Total unimodular (TU):** $M \in \mathbb{R}^{l \times m}$ is TU iff the determinant of every square submatrix of $M$ is 0, or $\pm 1$.

**Relaxation of ILP to LP [10]**
When $M$ is TU and $c$ is integral, then the LP

$$\max_{\beta \in \mathbb{R}^m} \{\theta^T \beta : M \beta \leq c, \beta \geq 0\}$$

has integer optimal solutions (i.e., ILP $\equiv$ LP).

- “Exact convex relaxation” of: $g^*(y) = \sup_{s \in \{0, 1\}^p} |y|^T s - F(s)$.
- Same idea behind the tractable biconjugation of submodular functions
Convexification of TU structures

TU convex relaxation given by LP

\[ g_{TU}^{**}(x) := \min_{\omega} \{ d^T \omega + e^T s : M \begin{bmatrix} \omega \\ s \end{bmatrix} \leq c, |x| \leq s, \omega \in \{0, 1\}^d \} \]

for all feasible \( x \), \( \infty \) otherwise.

- Special cases:
  - Rederive the convex envelope of several submodular models
  - Establish the tightness of some convex regularizers for non-submodular models
- Beyond linear objectives, some quadratic objectives can also be handled
Group cover sparsity: Minimal group cover [2, 12, 8]

**Structure:** We seek the signal covered by a minimal number of groups.

**Objective:** \( \mathbf{d}^T \mathbf{\omega} \)

**Linear description:** For each non-zero coefficient, at least one group containing it is selected

\[ B \mathbf{\omega} \geq \mathbf{s} \]

where \( B \) is the biadjacency matrix of \( G \), i.e., \( B_{ij} = 1 \) iff \( i \)-th coefficient is in \( G_j \).

When \( B \) is an interval matrix, or \( G \) has a loopless group intersection graph it is TU.
Group cover sparsity: **Minimal group cover** $[2, 12, 8]$

$$\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}, \text{ unit group weights } d = 1.$$ 

**Structure:** *We seek the signal covered by a minimal number of groups.*

**Objective:** $d^T \omega$

**Linear description:** For each non-zero coefficient, at least one group containing it is selected

$$B \omega \geq s$$

where $B$ is the biadjacency matrix of $\mathcal{G}$, i.e., $B_{ij} = 1$ iff $i$-th coefficient is in $G_j$. When $B$ is an interval matrix, or $\mathcal{G}$ has a *loopless* group intersection graph it is **TU**.

**Biconjugate:**

$$g_{TU}^*(x) = \min_{\omega \in [0,1]^M} \{d^T \omega : B \omega \geq |x|\} \text{ for } x \in [-1,1]^p, \infty \text{ otherwise}$$
Group cover sparsity: **Minimal group cover** $[2, 12, 8]$

$\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$, unit group weights $d = 1$.

**Structure:** *We seek the signal covered by a minimal number of groups.*

**Objective:** $d^T \omega$

**Linear description:** For each non-zero coefficient, at least one group containing it is selected

$$B \omega \geq s$$

where $B$ is the biadjacency matrix of $\mathcal{G}$, i.e., $B_{ij} = 1$ iff $i$-th coefficient is in $G_j$.

When $B$ is an interval matrix, or $\mathcal{G}$ has a *loopless* group intersection graph it is **TU**.

**Biconjugate:** $g_{TU}^*(x) = \min_{\omega \in [0,1]^M} \{d^T \omega : B \omega \geq |x|\}$ for $x \in [-1, 1]^p$, $\infty$ otherwise

$$= \min_{v_i \in \mathbb{R}^p} \{\sum_{i=1}^M d_i \|v_i\|_\infty : x = \sum_{i=1}^M v_i, \forall \text{supp}(v_i) \subseteq G_i\},$$
Group intersection sparsity $[9, 14, 1]$

Structure: We seek the signal intersecting with minimal number of groups.

Objective: $d^T \omega$  (submodular: $F(S) = \sum_{G_i \in G_i, S \cap G_i \neq \emptyset} d_i$)

Linear description: All groups containing a non-zero coefficient are selected

$$H_k s \leq \omega, \forall k \in \{0, \cdots, p\}$$

where $H_k(i, j) = \begin{cases} 1 & \text{if } j = k, j \in G_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.
Group intersection sparsity [9, 14, 1]

\[ \emptyset = \{\{1, 2\}, \{2, 3\}\} \], unit group weights \( d = 1 \)

Structure: *We seek the signal intersecting with minimal number of groups.*

Objective: \( d^T \omega \)  \(\text{submodular: } F(S) = \sum_{g_i \in \emptyset, S \cap g_i \neq \emptyset} d_i \)

Linear description: All groups containing a non-zero coefficient are selected

\[
H_k s \leq \omega, \forall k \in \{0, \cdots, p\}
\]

where

\[
H_k(i, j) = \begin{cases} 
1 & \text{if } j = k, j \in G_i \\
0 & \text{otherwise}
\end{cases}
\]

which is TU.

Biconjugate: \( g^{**}_{TU}(x) = \min_{\omega \in [0,1]^M} \{d^T \omega : H_k |x| \leq \omega, \forall k \in \Psi\} \)

for \( x \in [-1, 1]^p, \infty \) otherwise.
Group intersection sparsity \([9, 14, 1]\)

\[
G = \\{\{1, 2\}, \{2, 3\}\}, \text{ unit group weights } d = 1
\]

**Structure:** *We seek the signal intersecting with minimal number of groups.*

**Objective:** \(d^T \omega\)  
\(\text{submodular: } F(S) = \sum_{G_i \in G, S \cap G_i \neq \emptyset} d_i\)

**Linear description:** All groups containing a non-zero coefficient are selected

\[
H_k s \leq \omega, \forall k \in \{0, \cdots, p\}
\]

where \(H_k(i, j) = \begin{cases} 
1 & \text{if } j = k, j \in G_i \\
0 & \text{otherwise}
\end{cases}\), which is TU.

**Biconjugate:** \(g_{TU}^{**}(x) = \min_{\omega \in [0,1]^M} \{d^T \omega : H_k |x| \leq \omega, \forall k \in \Psi\} = \sum_{G \in G} \|x_G\|_{\infty}\)

for \(x \in [-1, 1]^p\), \(\infty\) otherwise.
Group knapsack sparsity \([15, 7, 6]\)

**Structure:** We seek the sparsest signal with group allocation constraints.

**Objective:** \( \mathbf{1}^T \mathbf{s} \)

**Linear description:** A valid support obeys budget constraints over \( \mathcal{G} \)

\[
B^T \mathbf{s} \leq \mathbf{c}_u
\]

where \( B \) is the biadjacency matrix of \( \mathcal{G} \), i.e., \( B_{ij} = 1 \text{ iff } i\text{-th coefficient is in } \mathcal{G}_j \).

When \( B \) is an interval matrix or \( \mathcal{G} \) has a loopless group intersection graph, it is TU.
Group knapsack sparsity $[15, 7, 6]$

$\begin{bmatrix}
1 & 1 & \cdots & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & \cdots & 1 & 1 & 0 & \cdots & 0 \\
0 & \cdots & 0 & 0 & 1 & 1 & \cdots & 1 & 1 \\
\end{bmatrix} (p-\Delta+1) \times p$

**Structure:** We seek the sparsest signal with group allocation constraints.

**Objective:** $1^T s$

**Linear description:** A valid support obeys budget constraints over $G$

$B^T s \leq c_u$

where $B$ is the biadjacency matrix of $G$, i.e., $B_{ij} = 1$ iff $i$-th coefficient is in $G_j$.

When $B$ is an interval matrix or $G$ has a *loopless* group intersection graph, it is TU.

**Biconjugate:** $g_{TU}^\ast\ast(x) = \begin{cases} 
\|x\|_1 & \text{if } x \in [-1, 1]^p, B^T|x| \leq c_u, \\
\infty & \text{otherwise}
\end{cases}$

For the neuronal spike example, we have $c_u = 1$. 
Group knapsack sparsity [15, 7, 6]

(left) \( g_{TU}^* (x) \leq 1 \) (middle) \( g_{TU}^* (x) \leq 1.5 \) (right) \( g_{TU}^* (x) \leq 2 \) for \( G = \{\{1, 2\}, \{2, 3\}\} \)

Structure: We seek the sparsest signal with group allocation constraints.

Objective: \( 1^T s \)

Linear description: A valid support obeys budget constraints over \( G \)

\[ B^T s \leq c_u \]

where \( B \) is the biadjacency matrix of \( G \), i.e., \( B_{ij} = 1 \) iff \( i \)-th coefficient is in \( G_j \).

When \( B \) is an interval matrix or \( G \) has a loopless group intersection graph, it is TU.

Biconjugate: \( g_{TU}^{**} (x) = \begin{cases} \|x\|_1 & \text{if } x \in [-1, 1]^p, B^T|x| \leq c_u, \\ \infty & \text{otherwise} \end{cases} \)

For the neuronal spike example, we have \( c_u = 1 \).
Group knapsack sparsity example: A stylized spike train

- Basis pursuit (BP): $\|x\|_1$
- TU-relax (TU):

$$g_{TU}^*(x) = \begin{cases} 
\|x\|_1 & \text{if } x \in [-1, 1]^p, B^T|x| \leq c_u, \\
\infty & \text{otherwise}
\end{cases}$$

Figure: Recovery for $n = 0.18p$. 

Relative errors:

$$\frac{\|x^h - x_{BP}\|_2}{\|x^h\|_2} = 0.200$$
$$\frac{\|x^h - x_{TU}\|_2}{\|x^h\|_2} = 0.067$$
Conclusions

Our work: TU modeling framework
- Complement previous approaches
- Convex programs (not necessarily norms)
- Tight convexifications, non-submodular examples
- **Easy to design** and “usually” efficient via an LP
References


References II

Compressive sensing recovery of spike trains using a structured sparsity model.

Learning with structured sparsity.

Multi-scale mining of fmri data with hierarchical structured sparsity.
In *Pattern Recognition in NeuroImaging (PRNI)*, 2011.

*Integer and combinatorial optimization*, volume 18.

Convex relaxation for combinatorial penalties.

[12] G. Obozinski, L. Jacob, and J.P. Vert.
Group lasso with overlaps: The latent group lasso approach.
Simple bounds for noisy linear inverse problems with exact side information.  
2013.  
arXiv:1312.0641v2 [cs.IT].

[14] Peng Zhao and Bin Yu.  
On model selection consistency of Lasso.  

Association screening of common and rare genetic variants by penalized regression.  