A totally unimodular view of structured sparsity

Marwa El Halabi
LIONS, EPFL

Volkan Cevher
LIONS, EPFL

Abstract

This paper describes a simple framework for structured sparse recovery based on convex optimization. We show that many interesting structured sparsity models can be naturally represented by linear matrix inequalities on the support of the unknown parameters, where the constraint matrix has a totally unimodular (TU) structure. For such structured models, tight convex relaxations can be obtained in polynomial time via linear programming. Our modeling framework unifies the prevalent structured sparsity models in the literature, introduces new interesting ones, and renders their tightness and tractability arguments transparent.

1 Introduction

Many important machine learning problems reduce to extracting parameters from dimensionality reduced and potentially noisy data [25]. The most common data model in this setting takes the familiar linear form

\[ b = A(x^\natural) + w, \quad (1) \]

where \( x^\natural \in \mathbb{R}^p \) is the unknown parameter that we seek; the linear operator \( A : \mathbb{R}^p \to \mathbb{R}^n \) compresses \( x^\natural \) from \( p \) dimensions to \( n \ll p \) dimensions; and \( w \in \mathbb{R}^n \) models noise. In a typical scenario, we only know \( A \) and \( b \) in (1).

In the absence of additional assumptions, it is impossible to reliably learn \( x^\natural \) when \( n \ll p \), since \( A \) has a nontrivial null space. Hence, we must exploit application-specific knowledge on \( x^\natural \) that results in our dataset \( b \) via (1). This knowledge often imposes \( x^\natural \) to be simple, e.g., well-approximated by a sparse set of coefficients that obey domain structure. Indeed, structured sparse parameters frequently appear in machine learning, signal processing, and theoretical computer science, and have broader generalizations including structure on matrix valued \( x^\natural \) based on its rank.

This paper relies on the convex optimization perspective for structured sparse recovery, which offers a rich set of analysis tools for establishing sample complexity for recovery and algorithmic tools for obtaining numerical solutions [5]. To describe our approach, we focus on the proto-problem

Find the simplest \( x^\natural \) subject to structure and data. \quad (2)

Fortunately, we often have immediate access to convex functions that encode the information within data (e.g., \( \| Ax - b \| \leq \sigma \) for some \( \sigma \in \mathbb{R}_+ \) in (2).

However, choosing convex functions that jointly address simplicity and structure in the proto-problem requires some effort, since their natural descriptions are inherently combinatorial [4, 2, 17, 10]. For instance, sparsity (i.e., the number of nonzero coefficients) of \( x^\natural \) subject to discrete restrictions on its support (i.e., the locations of the sparse coefficients) initiates many of the structured sparsity problems. Unsurprisingly, there is a whole host of useful convex functions in the literature that induce sparsity with the desiderata in this setting (cf., [2] for a review). As a result, our choice of convex functions for (2) typically depends on the tradeoff between the computational tractability and the tightness in capturing the combinatorial models [11, 18].

To this end, this paper introduces a combinatorial sparse modeling framework that simultaneously addresses both tractability and tightness issues that arise as a result of convex relaxation. In retrospect, our key idea is quite simple and closely follows the recipe in [11, 18], but with some new twists: We first summarize the discrete constraints that encode structure as linear inequalities. We then identify whether the structural constraint matrix is totally unimodular (TU), which can be verified in polynomial-time [24]. We then investigate classical discrete notions of simplicity and derive the Fenchel biconjugate of the combinatorial descriptions to obtain the convex relaxations for (2).

We illustrate how TU descriptions of simplicity and structure make many popular norms in the literature transparent, such as (latent) group norm, hierarchical norms, and norms that promote exclusivity. Moreover, we show that TU descriptions of sparsity structures support tight convex relaxations and polynomial-time solution complexity for (2). Our tightness result is a direct corollary of the fact that TU inequalities result in an integral constraint polyhedra where we can minimize linear costs exactly by convex relaxations (cf., Lemma [2]).

Our specific contributions can be summarized as follows.
We propose a new generative framework to construct convex programs for structured sparse recovery. Our results complement the structured norms perspective via submodular modeling, and go beyond it by deriving tight convex norms for non-submodular models. We also derive novel theoretical results using our modeling framework. For instance, the latent group lasso norm is the tightest convexification of the group $l_0$-norm [3]; Hierarchical group lasso is the tightest convexification of the sparse rooted connected tree model [3]; Sparse-group lasso leads to combinatorial descriptions that are provably not totally unimodular; Exclusive lasso norm is tight even for overlapping groups.

2 Preliminaries

We denote scalars by lowercase letters, vectors by lowercase boldface letters, matrices by boldface uppercase letters, and sets by uppercase script letters.

We denote the ground set by $\mathcal{P} = \{1,\ldots,p\}$, and its power set by $2^\mathcal{P}$. The $i$-th entry of a vector $x$ is $x_i$, the projection of $x$ over a set $S \subseteq \mathcal{P}$ is $x_S$, i.e., $(x_S)_i = 0, \forall i \notin S$. The vector of positive entries of $x$ is denoted by $x_{\{\}} = \min\{x_i \neq 0\}$ (min taken element wise). The absolute value of $|x|$ is taken element wise. Similarly, the comparison $x \geq y$ is taken element wise, i.e., $x_i \geq y_i, \forall i \in \mathcal{P}$.

We call the set of non-zero elements of a vector $x$, the support of $x$, denoted by $\text{supp}(x) = \{i : x_i \neq 0\}$. For binary vectors $s \in \{0, 1\}^\mathcal{P}$, with a slight abuse of notation, we will use $s$ and $\text{supp}(s)$ interchangeably, for example, given a set function $F$, we write $F(s) = F(\text{supp}(s))$. We let $\mathbf{1}_p$ be the vector in $\mathbb{R}^p$ of all ones, and $\mathbf{1}_p$ the $p \times p$ identity matrix. We drop subscripts whenever the dimensions are clear from the context. In particular, $\mathbf{1}_{\text{supp}(x)}$ denotes the projection of $\mathbf{1}_p$ over the set $\text{supp}(x)$.

We introduce some definitions that are used in the sequel.

Definition 1 (Submodularity). A set function $F : 2^\mathcal{P} \rightarrow \mathbb{R}$ is submodular iff it satisfies the following diminishing returns property: $\forall S \subseteq T \subseteq \mathcal{P}, \forall e \in \mathcal{P} \setminus T, F(S \cup \{e\}) - F(S) \geq F(T \cup \{e\}) - F(T)$.

Definition 2 (Fenchel conjugate). Given a function $g : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{+\infty\}$, its Fenchel conjugate, $g^* : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{+\infty\}$, is defined as:

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

where $\text{dom}(g) := \{x : g(x) < +\infty\}$. The Fenchel conjugate of the Fenchel conjugate of a function $g$ is called the biconjugate, and is denoted by $g^{**}$.

Definition 3 (Total unimodularity). A matrix $M \in \mathbb{R}^{\times m}$ is totally unimodular (TU) iff the determinant of every square submatrix of $M$ is 0 or ±1.

3 A generative view of sparsity models

3.1 Foundations

We can describe the simplicity and the structured constraints in the proto-problem by encoding them concisely into a combinatorial set function $F$ on the support of the unknown parameter $\mathbf{1}_\mathbb{R}[1,18]$. Hence, we can reduce our task of finding the best surrogate convex function to determining the convex envelope (i.e., the largest convex lower bound) of $F(\text{supp}(x))$, which is given by its biconjugate.

Let us first identify a sufficient condition for tractable computation of the convex envelope of $F(\text{supp}(x))$.

Lemma 1. Given a set function $F : 2^\mathcal{P} \rightarrow \mathbb{R} \cup \{+\infty\}$, let $g(x) = F(\text{supp}(x))$. If $F$ satisfies

A1. $F$ admits a proper $(\text{dom}(f) \neq \emptyset)$ lower semi-continuous (l.s.c.) convex extension $f$, i.e., $f(s) = F(s), \forall s \in \{0, 1\}^\mathcal{P}$;

A2. $\max_{s \in \{0, 1\}^\mathcal{P}} |y|^T s - f(s) = \max_{s \in \{0, 1\}^\mathcal{P}} |y|^T s - f(s), \forall y \in \mathbb{R}^\mathcal{P}$;

A3. $\min_{s \in \{0, 1\}^\mathcal{P}} \{f(s) : s \geq |x|\}$ can be efficiently minimized, $\forall x \in \mathbb{R}^\mathcal{P}$;

then the biconjugate of $g(x)$ over the unit $\ell_\infty$-ball can be efficiently computed.

It is also interesting to compute the biconjugate of $g(x)$ over other unit balls in the Euclidean space, which we will not discuss in this paper. The proof of Proposition 1 is elementary, and is provided below for completeness.

Proof. [Proposition 1] It holds that

$$g^*(y) = \sup_{\|x\|_\infty \leq 1} x^T y - F(\text{supp}(x))$$

$$= \sup_{s \in \{0, 1\}^\mathcal{P}} \sup_{\|x\|_\infty \leq 1} x^T y - F(s)$$

$$= \max_{s \in \{0, 1\}^\mathcal{P}} |y|^T s - F(s)$$

(by Hölder’s inequality)

$$= \max_{s \in \{0, 1\}^\mathcal{P}} |y|^T s - f(s)$$

(by A1 and A2)

Without assumption A2, the conjugate is a discrete optimization problem which, in general, is hard to solve, and the last equality will then only hold as an upper bound.

$$g^{**}(x) = \sup_{y \in \mathbb{R}^p} x^T y - g^*(y)$$

$$= \sup_{y \in \mathbb{R}^p} \min_{s \in \{0, 1\}^\mathcal{P}} |y|^T (|x| - s) + f(s)$$

$$= \max_{s \in \{0, 1\}^\mathcal{P}} \sup_{y \in \mathbb{R}^p} |y|^T (|x| - s) + f(s)$$

$$= \begin{cases} \min_{s \in \{0, 1\}^\mathcal{P}} f(s) & \text{if } x \in [-1, 1]^\mathcal{P} \cap \text{dom}(f) \\ \infty & \text{otherwise.} \end{cases}$$
Given assumption A1, (*) holds by Sion’s minimax theorem [22, Corollary 3.3]. Assumption A3 guarantees that the final convex minimization problem is tractable.

Remark 1. It is worth noting that, without assumption A2, the resulting convex function will still be a convex lower bound of \( g(x) \), albeit not necessarily the tightest one.

Remark 2. Note that we had to restrict the biconjugate over the box \([-c, c]^p\), otherwise the biconjugate would evaluate to a constant. In the sequel, unless otherwise stated, we assume \( c = 1 \) without loss of generality.

In general, computing even the conjugate is a hard problem. If the chosen combinatorial penalty has a tractable biconjugate, its envelope can be numerically approximated by a subgradient method [15].

3.2 Submodular sparsity models

In the generative modeling approach, non-decreasing submodular functions provide a flexible framework that is quite popular. While the best known method for checking submodularity has sub-exponential time complexity [20], we can often identify submodular structures by inspection, or we can restrict ourselves to known submodular models that closely approximate our objectives.

In the light of Lemma[1] submodular functions (cf., Def.[1]) indeed satisfy the three assumptions, which allow tractable computation of tight convex relaxations. For instance, the convex envelope of submodular non-decreasing functions is given by their Lovász extension [1], and optimization with the resulting convex regularizers can be done efficiently. In fact, their proximity operator is equivalent to solving a submodular minimization problem using the minimum-norm point algorithm [1], which usually runs in \( O(p^2) \)-time [1]. However, recent results show that in the worst case analysis min-norm point algorithm solves submodular minimization problems in \( O(p^7) \)-time [6].

4 Totally unimodular sparsity models

Combinatorial descriptions that satisfy Lemma[1] are not limited to submodular functions. Indeed, we can intuitively model the classical sparsity penalties that encourage the simplest support subject to structure constraints via basic linear inequalities [15, 17]. When the matrix encoding the structure is TU, such models admit tractable convex relaxations that are tight, which is supported by the following

Lemma 2 ([16]). Suppose the matrix \( M \) is TU and the vector \( c \) is integral. Then, the linear program (LP) \( \min_{\omega \in \{0,1\}^M} \{ d^T \omega + e^T \omega : M \beta \leq c \} \) has integral optimal solutions.

To demonstrate the utility of this result, let us first provide a simple linear template for TU models:

**Definition 4 (TU penalties).** We define TU penalties as discrete penalties over the support of \( x \) that can be written as

\[
\|x\|_{TU} := \min_{\omega \in \{0,1\}^M} \{ d^T \omega + e^T \omega : M \beta \leq c, 1_{\text{supp}(x)} = \omega \}
\]

for all feasible \( x \), \( \|x\|_{TU} = \infty \) otherwise, where \( M \in \mathbb{R}^{1 \times (M+p)} \) is a TU matrix, \( \beta = \left( \omega \right) / s, \omega \in \{0,1\}^M \) is a vector of auxiliary binary variables, \( d \in \mathbb{R}^M \) and \( e \in \mathbb{R}^p \) are arbitrary vectors, \( c \in \mathbb{Z}^d \) is an integral vector, and \( \kappa \) is a constant.

By Lemma[2] it follows that TU penalties satisfy the sufficient conditions described in Lemma[1] where the convex extension is the function itself, and the resulting convex envelope is given below.

**Proposition 1 (Convexification of TU penalties).** The convex envelope of TU penalties is given by the following LP:

\[
\|x\|_{TU}^* = \min_{s \in \{0,1\}^p, \omega \in \{0,1\}^M} \{ d^T \omega + e^T \omega + s + \kappa : M \beta \leq c, |x| \leq s \}
\]

for all feasible \( x \). \( \|x\|_{TU}^* = \infty \) otherwise.

Note that when the matrix \( M \) in Definition[4] is not TU, the above LP is still useful since it is a convex lower bound of the penalty, despite being non-tight as noted in Remark[1].

Besides allowing tractable tight convexifications, the choice of TU penalties is motivated by their ability to capture several important structures encountered in practice. In what follows, we study several TU penalties and their convex relaxations. We hope the reader will appreciate the reinterpretation of several well-known convex norms in the literature, as well as the introduction of the new ones.

5 Group sparsity

Group sparsity is an important class of structured sparsity models that arise naturally in machine learning applications (cf., [25] and the citations therein), where prior information on \( x^G \) dictates certain groups of variables to be selected or discarded together.

A group sparsity model thus consists of a collection of potentially overlapping groups \( \mathcal{G} = \{G_1, \ldots, G_M\} \) that cover the ground set \( \mathcal{P} \), where each group \( G_i \subseteq \mathcal{P} \) comprises a subset of variables. A group structure construction immediately supports two compact graph representations (cf., Figure[1]).

First, we can represent \( \mathcal{G} \) as a bipartite graph [3], where the groups form one set, and the variables form the other. A
variable $i \in \mathcal{P}$ is connected by an edge to a group $G_j \in \mathcal{G}$ iff $i \in G_j$. We denote by $B \in \{0, 1\}^{\mathcal{P} \times \mathcal{G}}$ the biadjacency matrix of this bipartite graph: $B_{ij} = 1$ iff $i \in G_j$, and by $E \in \{0, 1\}^{\mathcal{E} \times (\mathcal{M} + \mathcal{P})}$ its edge-node incidence matrix; $E_{ij} = 1$ iff the vertex $j$ is incident to the edge $e_i \in \mathcal{E}$.

Second, we can represent $\mathcal{G}$ as an intersection graph $\mathcal{I}$, where the vertices are the groups $G_i \in \mathcal{G}$. Two groups $G_i$ and $G_j$ are connected by an edge iff $G_i \cap G_j \neq \emptyset$. This structure makes it explicit whether groups themselves have cyclic interactions via variables and identifies computational difficulties.

### 5.1 Counting group intersections

In group sparse models, we typically seek to express the support of $x^\mathcal{G}$ using only a few groups. One natural penalty to consider then is the non-decreasing submodular function that sums up the weight of the groups intersecting with the support $F_0(S) = \sum_{G_i \in \mathcal{G}, G_i \cap S \neq \emptyset} d_i$. The convexification of this submodular function results in the $\ell_\infty$-group lasso (also known as $\infty$-CAP penalties) \cite{11,26}, as shown in \cite{1}.

We now show how to express this penalty as a TU penalty.

**Definition 5** (Groups intersection count).

\[
\|x\|_{\mathcal{G} \cap} := \min_{\omega \in \{0, 1\}^M} \{d^T \omega : H \beta \leq 0, 1_{\text{supp}(x)} = s \}
\]

where $H$ is the following matrix:

\[
H := \begin{pmatrix}
-I_M & H_1 \\
-I_M & H_2 \\
\vdots & \vdots \\
-I_M & H_p
\end{pmatrix},
\]

and the vector $d \in \mathbb{R}^M$ here corresponds to positive group weights. Recall that $\beta = \left( \begin{array}{c} \omega \\ s \end{array} \right)$, and thus $H \beta \leq 0$ simply corresponds to $s_j \leq w_i, \forall j \in G_i$.

$\|x\|_{\mathcal{G} \cap}$ indeed sums up the weight of the groups intersecting with the support, since for any coefficient in the support of $x$ the constraint $H \beta \leq 0$ forces all the groups that contain this coefficient to be selected.

Here, $H$ is TU, since each row of $H$ contains at most two non-zero entries, and the entries in each row with two non-zeros sum up to zero, which is a sufficient condition for total unimodularity \cite[Proposition 2.6]{15}.

**Proposition 2** (Convexification). The convex envelope of $\|x\|_{\mathcal{G} \cap}$ over the unit $\ell_\infty$-ball is

\[
\|x\|_{\mathcal{G} \cap}^{**} = \begin{cases} \sum_{G_i \in \mathcal{G}} d_i \|x_{G_i}\|_\infty & \text{if } x \in [-1, 1]^p \\
\infty & \text{otherwise} \end{cases}
\]

**Proof.** Since $\|x\|_{\mathcal{G} \cap}$ is a TU-penalty, we can use Proposition 1 to compute its convex envelope:

\[
\|x\|_{\mathcal{G} \cap}^{**} = \min_{\omega \in \{0, 1\}^M} \{d^T \omega : H \beta \leq 0, |x| \leq s \}
\]

\[
= \min_{\omega \in \{0, 1\}^M} \{d^T \omega : H \left( \begin{array}{c} \omega \\ s \end{array} \right) \leq 0 \}
\]

\[
= \sum_{G_i \in \mathcal{G}} d_i \|x_{G_i}\|_\infty \quad (\text{since } w_i^* = \|x_{G_i}\|_\infty)
\]

for $x \in [-1, 1]^p$, $\|x\|_{\mathcal{G} \cap}^{**} = \infty$ otherwise. \hfill \square

### 5.2 Minimal group cover

The groups intersections penalty induces supports corresponding to the intersection of the complements of groups, while in several applications, it is desirable to explain the support of $x^\mathcal{G}$ as the union of groups in $\mathcal{G}$. In particular, we can seek the minimal set cover of $x^\mathcal{G}$:

**Definition 6** (Group $\ell_0$-“norm”, \cite{3}). The group $\ell_0$-“norm” computes the weight of the minimal weighted set cover of $x$ with group weights $d \in \mathbb{R}_+^M$:

\[
\|x\|_{\mathcal{G},0} := \min_{\omega \in \{0, 1\}^M} \{d^T \omega : B \omega \geq 1_{\text{supp}(x)} \}
\]

where $B$ is the biadjacency matrix of the bipartite graph representation of $\mathcal{G}$.

Note that computing the group $\ell_0$-“norm” is NP-Hard, since it corresponds to the minimum weight set cover problem. $\|x\|_{\mathcal{G},0}$ is a penalty that was previously considered in \cite{3,19,10}, and the latent group lasso was proposed in \cite{19} as a potential convex surrogate for it, but it was not established as the tightest possible convexification.

The group $\ell_0$-“norm” is not a submodular function, but if $B$ is TU, it is a TU penalty, and thus is admits a tight convex relaxation. We show below that the convex envelope of the group $\ell_0$-“norm” is indeed the $\ell_\infty$-latent group norm. It is worth noting that the $\ell_0$-latent group lasso was also shown in \cite{13} to be the positive homogeneous convex envelope of the $\ell_p$-regularized group $\ell_0$-“norm”, i.e., of $\mu \|x\|_{\mathcal{G},0} + \nu \|x\|_p$.

**Proposition 3** (Convexification). When the group structure leads to a TU biadjacency matrix $B$, the convex envelope of the group $\ell_0$-“norm” over the unit $\ell_\infty$-ball is

\[
\|x\|_{\mathcal{G},0}^{**} = \begin{cases} \min_{\omega \in \{0, 1\}^M} \{d^T \omega : B \omega \geq |x| \} & \text{if } x \in [-1, 1]^p \\
\infty & \text{otherwise} \end{cases}
\]
Proof. Note that \( \|x\|_{\ell,0} \) can be written in the form given in Definition 4 with \( M = [-B, I] \) and \( c = 0 \). Thus, when \( B \) is TU, so is \( M \) (Proposition 2.1), and thus we can use Proposition 7 to compute its convex envelope:

\[
\|x\|_{\ell,0}^* = \min_{\omega \in [0,1]^M} \{ d^T \omega : B \omega \geq s, \|x\| \leq s \}
\]

\[
= \min_{\omega \in [0,1]^M} \{ d^T \omega : B \omega \geq |x| \}
\]

for \( x \in [-1,1]^p \), \( \|x\|_{\ell,0}^* = \infty \) otherwise.

Thus, given a group structure \( \mathcal{G} \), one can check in polynomial time if it is TU (Lemma 2) to guarantee that the \( \ell_\infty \)-latent group lasso will be the tightest relaxation.

Remark 4. One important class of group structures that leads to a TU matrix \( B \) is given by acyclic groups, as shown in [7] Lemma 2. The induced intersection graph for such groups is acyclic, as illustrated in Figure 7. In this case, the \( \ell_\infty \)-latent group norm is a tight relaxation.

5.3 Sparsity within groups

Both group model penalties we considered so far only induce sparsity on the group level; if a group is selected, all variables within the group are encouraged to be non-zero. In some applications, it is desirable to also enforce sparsity within groups. We consider then a natural extension of the above two penalties, where each group is weighted by the \( \ell_0 \)-norm of \( x \) restricted to the group.

Definition 7 (Group models with sparsity within groups).

\[
\|x\|_{\omega,0} = \min_{\omega \in [0,1]^M} \left\{ \sum_{i=1}^M \omega_i \|x_{\omega_i}\|_0 : M \beta \leq 0, 1_{\text{supp}(x) = s} \right\}
\]

where \( M \) here is either \( M = H \) in Definition 5 or \( M = [-B, I] \) in Definition 2.

Unfortunately, this penalty leads to a non-TU penalty, and thus its corresponding convex surrogate given by Proposition 7 is not guaranteed to be tight.

Proposition 4. Given any group structure \( \mathcal{G} \), \( \|x\|_{\omega,0} \) is not a TU penalty.

Recall that \( E \) is the edge-node incidence matrix of \( G(\mathcal{G} \cup \mathcal{P}, \mathcal{E}) \). The constraint \( E \beta \leq z + 1 \) corresponds to \( z_{ij} \geq \omega_i + s_j - 1, \forall (i,j) \in \mathcal{E} \). Although both matrices \( M \) and \( E \) are TU, their concatenation \( \tilde{M} = \left[ \begin{array}{c} E \end{array} M \right] \) is not TU. To see this, let us first focus on the case where \( M = [-B, I] \):

Given any coefficient \( i \in \mathcal{P} \) covered by at least one group \( \mathcal{G}_i \), we denote the corresponding edge in the bipartite graph by \( e_j = (i, M + i) \), which corresponds to the \( j \)-th row of \( E \). This translates into having the entries \( M_{i,i} = -1, M_{i,M+i} = 1, M_{p+j,i} = 1, \) and \( M_{p+j,M+i} = 1 \). The determinant of the submatrix resulting from these entries is \(-2\), which contradicts the definition of TU (cf., Def. 3). It follows then that \( \tilde{M} \) is TU iff \( \mathcal{G} = \{ \emptyset \} \).

A similar argument holds for \( M = H \). 

Proposition 5 (Convexification). The convex surrogate via Proposition 7 for \( \|x\|_{\omega,0} \) with \( M = H \) (i.e., the sparse group model, which counts the group intersections) is given by

\[
\Omega_{\omega,0}(x) := \sum_{(i,j) \in \mathcal{E}} \left( \|x_{\omega_i}\|_\infty + |x_j| - 1 \right)
\]

for \( x \in [-1,1]^p \), and, \( \Omega_{\omega,0}(x) := \infty \) otherwise. Note that \( \Omega_{\omega,0}(x) \leq \|x\|_{\omega,0}^* \).

Proof. For \( x \in [-1,1]^p \),

\[
\Omega_{\omega,0}(x) = \min_{\omega \in [0,1]^M} \left\{ \sum_{(i,j) \in \mathcal{E}} \omega_i \|x_{\omega_i}\|_\infty + |x_j| - 1 \right\}
\]

since \( \omega_i^* = \|x_{\omega_i}\|_\infty, s^* = |x|, \) and \( z_{ij}^* = (\omega_i^* + s_j^* - 1) \).

By construction, the convex penalty proposed by Proposition 5 is different from the \( \ell_1 \)-regularized sparse group lasso penalty in [21], which penalizes the sparsity of the parameter in addition to counting group intersections.

Analogous to the latent group norm, we can seek to convexify the sparest set cover with sparsity within groups:

Proposition 6 (Convexification). The convex surrogate given by Proposition 7 for \( \|x\|_{\omega,0} \) with \( M = [-B, I] \) (i.e., the group \( \ell_0 \)-“norm” with sparse groups) is given by

\[
\Omega_{\omega,0}(x) := \min_{\omega \in [0,1]^M} \left\{ \sum_{(i,j) \in \mathcal{E}} \omega_i \|x_{\omega_i}\|_\infty + |x_j| - 1 \right\}
\]

for \( x \in [-1,1]^p \), \( \|x\|_{\omega,0}^* = \infty \) otherwise.

Proof. For \( x \in [-1,1]^p \),

\[
\Omega_{\omega,0}(x) = \min_{\omega \in [0,1]^M} \left\{ \sum_{(i,j) \in \mathcal{E}} \omega_i \|x_{\omega_i}\|_\infty + |x_j| - 1 \right\}
\]

\[
= \min_{\omega \in [0,1]^M} \{ \sum_{(i,j) \in \mathcal{E}} \omega_i + |x_j| - 1 \}
\]

\[
= \min_{\omega \in [0,1]^M} \{ \sum_{(i,j) \in \mathcal{E}} \omega_i + |x_j| - 1 \}
\]
since $s^* = |x|$, and $z^*_i = (\omega_i + s^*_i - 1)_+$. □

5.4 Hierarchical model

We study the hierarchical sparsity model, where the coefficients of $x^s$ are organized over a directed tree structure $T$, and the sparse coefficients form a rooted connected subtree of $T$. This model is popular in image processing due to the natural structure of wavelet coefficients [12] [7] [23].

We can describe such a hierarchical model as a TU model:

**Definition 8** (Tree $\ell_0$-“norm”). We define the penalty encoding the hierarchical model on $x$ as:

$$||x||_{T,0} := \begin{cases} ||x||_0 & \text{if } T1_{\text{supp}(x)} \geq 0 \\ \infty & \text{otherwise} \end{cases}$$

where $T$ is the edge-node incidence matrix of the directed tree $T$, i.e., $T_{li} = 1$ and $T_{lj} = -1$ if $e_l = (i,j)$ is an edge in $T$. It encodes the constraint $s_{\text{parent}} \geq s_{\text{child}}$ for $s = 1_{\text{supp}(x)}$ over the tree.

This is indeed a TU model since hierarchical structures lead to totally unimodular constraint matrix [3, Proposition 7.3].

**Proposition 7.** (Convexification) The convexification of the tree $\ell_0$-“norm” over the unit $\ell_\infty$-ball is given by

$$||x||^*_{T,0} = \min_{s \in [0,1]^p} \{ T^s : Ts \geq 0, ||x||_\infty \leq s \}$$

Proof. Since this is a TU-penalty we can use Proposition 1 to compute its convex envelope:

$$||x||^*_{T,0} = \sum_{G \in \mathcal{G}_H} ||x_G||_\infty$$

for $x \in [-1,1]^p$, $\infty$ otherwise, and where the groups $G \in \mathcal{G}_H$ are defined as each node and all its descendants. (∗) holds since any feasible $s$ should satisfy $s \geq |x|$ and $s_{\text{parent}} \geq s_{\text{child}}$, so starting from the leaves, each leaf satisfies $s_i \geq |x_i|$, and since we are looking to minimize the sum of $s_i$’s, we simply set $s_i = x_i$. For a node $i$ with two children $j,k$ as leaves, it will satisfy $s_i \geq |x_i|, |s_j|, |s_k|$, thus $s_i = \max\{|x_i|, |s_j|, |s_k|\}$, and so on. Thus, $s_i = \max\{k \text{ is a descendant of } i \text{ or } i \text{ itself} \} |x_k|$ □

Note that the resulting convex norm is the $\ell_\infty$-hierarchical group norm [12], which is a special case of $\ell_\infty$-group norm we studied in Section 5.1 as the convex envelope of $||x||_{\mathcal{G},F}$. In this sense, $||x||_{\mathcal{G},F}$ is equivalent to $||x||_{T,0}$ for the group structure $\mathcal{G}_H$ (when the weights are set to one).

6 Dispersive sparsity models

The sparsity models we considered thus far encourage clustering. The implicit structure in these models is that coefficients within a group exhibit a positive, reinforcing correlation. Loosely speaking, if a coefficient within a group is important, so are the others. However, in many applications, the opposite behavior may be true. That is, sparse coefficients within a group compete against each other [27, 9, 8].

Hence, we describe models that encourage the dispersion of sparse coefficients. Here, dispersive models still inherit a known group structure $\mathcal{G}$, which underlie their interactions in the opposite manner to the group models in Section 5.

6.1 Dispersiveness

One natural model for dispersiveness allows only one coefficient to be selected in each group:

$$F_D(S) = \begin{cases} 0 & \text{if } S = \emptyset \\ 1 & \text{if } \max_{G \in \mathcal{G}} |S \cap G| \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

Whenever the group structure forms a partition of $\mathcal{P}$, [18] shows that the positive homogeneous convex envelope of the $\ell_p$-regularized dispersive model, i.e., of $\mu F_D(\text{supp}(x)) + \nu ||x||_p$, is the exclusive norm in [27].

In what follows, we prove that $F_D(S)$ is a TU penalty whenever the group structure leads to a TU biadjacency matrix $B$ of the bipartite graph representation, which includes partition structures. We establish that another convex penalty, closely related to $\ell_\infty$-exclusive lasso, is actually the tightest convex relaxation of $F_D(\text{supp}(x))$ even without $\ell_p$-regularization for any TU group structure, and not necessarily partition groups.

**Definition 9** (Dispersive penalty). Given a group structure $\mathcal{G}$ that leads to a TU biadjacency matrix $B$, $F_D(\text{supp}(x))$ can be written as the following TU penalty:

$$||x||_D := \min_{\omega \in \{0,1\}^k} \{ \omega : B^T1_{\text{supp}(x)} \leq \omega \}$$

if $B^T1_{\text{supp}(x)} \leq 1$, $||x||_D = \infty$ otherwise.

Note that if $B$ is TU, $B^T$ is also TU [16] Proposition 2.1. Groups that form a partition of $\mathcal{P}$ are acyclic, thus the corresponding matrix $B$ is TU trivially (cf., Remark 4).

Another important example of a TU group structure arises from the simple one-dimensional model of the neuronal signal suggested by [9]. In this model, neuronal signals are seen as a train of spike signals with some refractoriness period $\Delta \geq 0$, where the minimum distance between two non-zeros is $\Delta$. This structure corresponds to an interval
matrix $B^T = D$, which is TU [15 Corollary 2.10].

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

**Proposition 8** (Convexification). The convex envelope of \([x]^\ast_D\) over the unit $\ell_\infty$-ball when $B^T$ is a TU matrix is given by

$$\|x\|^\ast_D = \begin{cases} \max_{\omega \in [0,1]^p} \|x\|^\omega_1 & \text{if } x \in [-1,1]^p \text{, } B^T|x| \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

**Proof.** Since this is a TU penalty we can use Proposition 1 to compute its convex envelope:

\[
\|x\|^\ast_D = \begin{cases} \min_{\omega \in [0,1]^p} \{ \omega : B^T x \leq \omega 1 \} & \text{if } x \text{ feasible} \\ \infty & \text{otherwise} \end{cases}
\]

\[
= \begin{cases} \|B^T x\|_\infty & \text{if } x \in [-1,1]^p \text{, } B^T|x| \leq 1 \\ \infty & \text{otherwise} \end{cases}
\]

Notice that in this case, regularizing with the $\ell_p$ norm before convexifying lead to the loss of part of the structure.

### 6.2 Sparsest dispersiveness

In some applications, it may be desirable to seek the sparsest signal satisfying the dispersive structure. This can be achieved by incorporating sparsity into the dispersive penalty, resulting in the following TU penalty.

**Definition 10** (Dispersive $\ell_0$-norm). Given a group structure $\mathcal{G}$ that leads to a TU biadjacency matrix $B$, we define the penalty encoding the sparse dispersive model on $x$ as

$$\|x\|_{D,0} = \begin{cases} \|x\|_0 & \text{if } B^T \mathbb{1}_{\text{supp}(x)} \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

We can compute the convex envelope of $\|x\|_{D,0}$ in a similar fashion to Proposition 8.

**Proposition 9.** (Convexification) The convexification of the dispersive $\ell_0$-norm over the unit $\ell_\infty$-ball is given by

$$\|x\|^{\ast}_{D,0} = \begin{cases} \|x\|_1 & \text{if } x \in [-1,1]^p \text{, } B^T|x| \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

It is worth mentioning that regularizing with the $\ell_p$-norm here loses the underlying dispersive structure. In fact, [18] proves that the positively homogeneous convex envelope of $\mu\|x\|_{D,0} + \lambda\|x\|_p$ is given by the dual of

\[
\Omega^*_p(y) = \max_{s \in \{0,1\}^p, s \neq 0} \frac{\|y^{\supp(s)}\|_q}{(1^T s)^{1/q}}
\]

which is simply the $\ell_1$-norm. To see this, note that $\Omega^*_p(y) \equiv \|y\|_\infty$ since $\sum_{s \in \{0,1\}^p} |y^s| \leq \|s\|_\infty \|y\|_\infty \forall S \subseteq \mathcal{P}$ which is achieved with equality by choosing the vector $s$ having ones where $y$ is maximal, and zeros everywhere else. Note that this vector satisfies $B^T s \leq 1$. As a result, the convexification boils down to the $\ell_1$-norm, since $\Omega^*_p(x) = \sup_{\omega \in [0,1]} \{ |x^\omega| : \omega \mathbb{1} \} \leq \|x\|_\infty$.

We illustrate the effect of this loss of structure by a numerical example in Section 7.

### 6.3 Graph dispersiveness

In this section, we illustrate that our framework is not limited to linear costs, by considering a pairwise dispersive model. We assume that the parameter structure is encoded on a known graph $G(\mathcal{P}, \mathcal{E})$, where coefficients connected by an edge are discouraged from being on simultaneously.

**Definition 11** (Pairwise dispersive penalty). Given a graph $G(\mathcal{P}, \mathcal{E})$ with a TU edge-node incidence matrix $E_G$ (e.g., bipartite graph), we define the penalty encoding pairwise dispersive model as

$$\|x\|_{g,D} = \sum_{(i,j) \in \mathcal{E}} s_i s_j \text{ where } s = \mathbb{1}_{\text{supp}(x)}$$

Note that this function is not submodular; in fact, $\|s\|_{g,D}$ is a supermodular function.

**Proposition 10** (Convexification). The convex envelope of $\|x\|_{g,D}$ over the unit $\ell_\infty$-ball is

$$\|x\|^{\ast}_{g,D} = \begin{cases} \sum_{(i,j) \in \mathcal{E}} (|x_i| + |x_j| - 1) & \text{if } x \in [-1,1]^p \\ \infty & \text{otherwise} \end{cases}$$

**Proof.** We use the linearization trick employed in [14] to reduce $\|x\|_{g,D}$ to a TU penalty. Let $s = \mathbb{1}_{\text{supp}(x)}$,

\[
\|s\|_{g,D} = \sum_{(i,j) \in \mathcal{E}} s_i s_j = \min_{z \in \{0,1\}^{|\mathcal{E}|}} \{ \sum_{(i,j) \in \mathcal{E}} z_{ij} : z_{ij} \geq s_i + s_j - 1 \}
\]

\[
= \min_{z \in \{0,1\}^{|\mathcal{E}|}} \{ \sum_{(i,j) \in \mathcal{E}} z_{ij} : E_G s \leq z + 1 \}
\]

Now we can apply Proposition 1 to compute the convex envelope:

\[
\|x\|^{\ast}_{g,D} = \min_{s \in \{0,1\}^p, x \in [-1,1]^p} \{ \sum_{(i,j) \in \mathcal{E}} z_{ij} : E_G s \leq z + 1, |x| \leq s \}
\]

\[
= \sum_{(i,j) \in \mathcal{E}} (|x_i| + |x_j| - 1) \quad (s^* = x, z^*_{ij} = (s^*_i + s^*_j - 1)_+)
\]

for $x \in [-1,1]^p$, $\|x\|^{\ast}_{g,D} = \infty$ otherwise.
7 Numerical illustration of the loss of structure in the sparsest dispersive model

In this section, we show the impact of convexifying two different simplicity objectives under the same dispersive structural assumptions. Specifically, we consider minimizing the convex envelope of the \( \ell_p \)-regularized dispersive model \([13]\) versus the dispersive \( \ell_1 \)-“norm” over the unit \( \ell_\infty \)-ball in Section 6.2. To produce the recovery results in Figure 2 we generate a train of spikes of equal value for \( x^\dagger \) in dimensions \( p = 200 \) with a refractoriness of \( \Delta = 25 \) (cf., Figure 3). We then recover \( x^\dagger \) from its compressive measurements \( y = Ax^\dagger + w \), where the noise \( w \) is also a sparse vector, with 15 non-zero Gaussian values of variance \( \sigma = 0.01 \) and \( A \) is a random column normalized Gaussian matrix. Since the noise is sparse, we encode the data via \( \|y - Ax\|_1 \leq \|w\|_1 \) using the true \( \ell_1 \)-norm of the noise. We produce the data randomly 20 times and report the averaged results.

Figure 2 measures the relative recovery error with \( \frac{\|\hat{x} - x^\dagger\|_2}{\|x^\dagger\|_2} \), as we vary the number of compressive measurements. The regularized convexification, as derived in \([13]\), simply leads to Basis Pursuit formulation (BP), while the TU convexification, as described in Section 6.2, results in the addition of a dispersion constraint \( B^T|z| \leq 1 \) to the Basis Pursuit formulation. We refer to the resulting criteria as Dispersive Basis Pursuit (DBP). Since the DBP criteria uses the fact that \( x^\dagger \) lies in the unit \( \ell_\infty \)-ball, we include this constraint in the BP formulation for fairness. We use an interior point method to obtain high accuracy solutions to each formulation.

Figure 2 shows that the DBP outperforms the BP as we vary the number of measurements. Note that the number of measurements needed to achieve a certain error is expected to be lower for DBP than BP, as theoretically characterized in \([9]\). Hence, by changing the objective in the convexification, Figure 3 reinforces the message that we can lose the tightness in capturing certain structured sparsity models.

8 Conclusions

We outline a principled recipe for designing convex formulations that jointly express models of simplicity and structure for sparse recovery. The main hallmark of our approach is its pithiness in generating the prevalent convex structured sparse formulations and in explaining their tightness. Our key idea relies on expressing sparsity structures via simple linear inequalities over the support of the unknown parameters and their corresponding latent group indicators. By recognizing the totally unimodularity of the underlying constraint matrices, we can tractably leverage biconjugation of the corresponding combinatorial simplicity objective subject to structure, whether it promotes clustering or dispersiveness, and perform tractable recovery using standard optimization techniques.

Acknowledgements

This work was supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof, SNF 200021- 132548, SNF 200021-146750 and SNF CRSII2-147633.

References

[1] F. Bach. Structured sparsity-inducing norms through submodular functions. In NIPS, pages 118–126, 2010.

[2] F. Bach. Learning with submodular functions: A convex optimization perspective. arXiv preprint arXiv:1111.6453, 2011.

[3] L. Baldassarre, N. Bhan, V. Cevher, and A. Kyrillidis. Group-sparse model selection: Hardness and relaxations. arXiv preprint arXiv:1303.3207, 2013.
[4] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde. Model-based compressive sensing. Information Theory, IEEE Transactions on, 56(4):1982–2001, 2010.

[5] V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. Foundations of Computational Mathematics, 12:805–849, 2012.

[6] P. Kothari D. Chakrabarty, P. Jain. Provable submodular minimization using wolfe's algorithm. NIPS, 2014.

[7] Marco F Duarte, Michael B Wakin, and Richard G Baraniuk. Wavelet-domain compressive signal reconstruction using a hidden markov tree model. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, pages 5137–5140. IEEE, 2008.

[8] W Gerstner and W. Kistler. Spiking neuron models: Single neurons, populations, plasticity. Cambridge university press, 2002.

[9] C. Hegde, M. Duarte, and V. Cevher. Compressive sensing recovery of spike trains using a structured sparsity model. In SPARS’09-Signal Processing with Adaptive Sparse Structured Representations, 2009.

[10] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. The Journal of Machine Learning Research, 12:3371–3412, 2011.

[11] R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Journal of Machine Learning Research, 12:2777–2824, 2011.

[12] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. Journal of Machine Learning Research, 12:2297–2334, 2011.

[13] Vladimir Jojic, Suchi Saria, and Daphne Koller. Convex envelopes of complexity controlling penalties: the case against premature envelopment. In International Conference on Artificial Intelligence and Statistics, pages 399–406, 2011.

[14] Marcin Kamiński. Quadratic programming on graphs without long odd cycles. 2008.

[15] A. Kyrillidis and V. Cevher. Combinatorial selection and least absolute shrinkage via the clash algorithm. In Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on, pages 2216–2220. Ieee, 2012.

[16] George L Nemhauser and Laurence A Wolsey. Integer and combinatorial optimization, volume 18. Wiley New York, 1999.