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 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 norms 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 [26]. The most common data model in this setting takes the familiar linear form

$$b = A\left(x^{\natural}\right) + w, \qquad (1)$$

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 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 nullspace. Hence, we must exploit application-specific knowledge on $x^{\natural}$. 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 anal-

ysis 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 subject to *structure* and *data*. (2)

Fortunately, we often have immediate access to convex functions that encode the information within *data* (e.g., $\|Ax - b\| \le \sigma$ 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, 18, 11]. 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). The challenge resides in finding computationally tractable convex surrogates that tightly captures the combinatorial models.

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 [1, 19], 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 [25]. 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 the (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 polyhedron where we can optimize linear costs exactly by convex relaxation (cf., Lemma 2).

Our specific contributions are summarized as follows.

We propose a new generative framework to construct structure-inducing convex programs (which are not necessarily structure-inducing norms). 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 $\mathcal{P} = \{1, \cdots, p\}$, and its power set by $2^{\mathcal{P}}$. The i-th entry of a vector $\boldsymbol{x}$ is x_i, the projection of $\boldsymbol{x}$ over a set $\mathcal{S} \subseteq \mathcal{P}$ is $\boldsymbol{x}_{\mathcal{S}}$, i.e., $(\boldsymbol{x}_{\mathcal{S}})_i = 0, \forall i \notin \mathcal{S}$. The vector containing the positive part of $\boldsymbol{x}$ is denoted by $\boldsymbol{x}_+ = \min\{\boldsymbol{x}, 0\}$ (min taken element wise). The absolute value of $|\boldsymbol{x}|$ is taken element wise. Similarly, the comparison $\boldsymbol{x} \geq \boldsymbol{y}$ is taken element wise, i.e., $x_i \geq y_i, \forall i \in \mathcal{P}$. For $q \geq 1$, the ℓ_q-norm of a vector $\boldsymbol{x} \in \mathbb{R}^p$ is given by $\|\boldsymbol{x}\|_q = (\sum_{i=1}^p x_i^q)^{1/q}$, and $\|\boldsymbol{x}\|_\infty = \max_i\{|x_i|\}$.

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

We introduce some definitions that are used in the sequel.

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

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

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

where $\mathrm{dom}(g) := \{\boldsymbol{x} : g(\boldsymbol{x}) < +\infty\}$. 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 $\boldsymbol{M} \in \mathbb{R}^{l \times m}$ is totally unimodular (TU) iff the determinant of every square submatrix of $\boldsymbol{M}$ is 0 or ± 1.*

In what follows, some proofs have been omitted due to lack of space; see the supplementary material.

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 [1, 19]. Hence, we can reduce our task of finding the tightest surrogate convex function to determining the convex envelope, i.e., the largest convex lower bound of $F(\mathrm{supp}(\boldsymbol{x}))$, which is given by its biconjugate.

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

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

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

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

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

then the biconjugate of $g(\boldsymbol{x})$ over the unit ℓ_∞-ball can be efficiently computed.

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

Proof. [Lemma 1] It holds that

$$
\begin{aligned}
g^*(\boldsymbol{y}) &= \sup_{\|\boldsymbol{x}\|_\infty \leq 1} \boldsymbol{x}^T \boldsymbol{y} - F(\mathrm{supp}(\boldsymbol{x})) \\
&= \sup_{\boldsymbol{s} \in \{0,1\}^p} \sup_{\substack{\|\boldsymbol{x}\|_\infty \leq 1 \\ \mathbb{1}_{\mathrm{supp}(\boldsymbol{x})} = \boldsymbol{s}}} \boldsymbol{x}^T \boldsymbol{y} - F(\boldsymbol{s}) \\
&= \max_{\boldsymbol{s} \in \{0,1\}^p} |\boldsymbol{y}|^T \boldsymbol{s} - F(\boldsymbol{s}) \quad \text{(by Hölder's inequality)} \\
&= \max_{\boldsymbol{s} \in [0,1]^p} |\boldsymbol{y}|^T \boldsymbol{s} - f(\boldsymbol{s}) \quad\quad \text{(by A1 and A2)}
\end{aligned}
$$

The conjugate is a discrete optimization problem which, in general, is hard to solve. Assumption $A2$ guarantees that its convex relaxation has integral optimal solutions, otherwise the last equality will only hold as an upper bound.

$$
\begin{aligned}
g^{**}(\boldsymbol{x}) &= \sup_{\boldsymbol{y} \in \mathbb{R}^p} \boldsymbol{x}^T \boldsymbol{y} - g^*(\boldsymbol{y}) \\
&= \sup_{\boldsymbol{y} \in \mathbb{R}^p} \min_{\boldsymbol{s} \in [0,1]^p} \boldsymbol{y}^T \boldsymbol{x} - |\boldsymbol{y}|^T \boldsymbol{s} + f(\boldsymbol{s}) \\
&\overset{\star}{=} \min_{\boldsymbol{s} \in [0,1]^p} \sup_{\substack{\boldsymbol{y} \in \mathbb{R}^p \\ \mathrm{sign}(\boldsymbol{y}) = \mathrm{sign}(\boldsymbol{x})}} |\boldsymbol{y}|^T (|\boldsymbol{x}| - \boldsymbol{s}) + f(\boldsymbol{s}) \\
&= \begin{cases} \min_{\substack{\boldsymbol{s} \in [0,1]^p \\ \boldsymbol{s} \geq |\boldsymbol{x}|}} f(\boldsymbol{s}) & \text{if } \boldsymbol{x} \in [-1, 1]^p \cap \mathrm{dom}(f) \\ \infty, & \text{otherwise.} \end{cases}
\end{aligned}
$$

Given assumption $A1$, ($\star$) holds by Sion's minimax theorem [23, 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 $g(\boldsymbol{x})$, albeit not necessarily the* tightest *one.*

Remark 2. *Note that in lemma 1, we had to restrict the biconjugate over the box $[-c, c]^p$ (with $c = 1$), otherwise it 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 conjugate, its envelope can be numerically approximated by a subgradient method [14].

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 [21], 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 the tractable computation of tight convex relaxations. For instance, the convex envelope of a submodular non-decreasing function is given by its Lovász extension [1], and optimization with the resulting convex regularizer can be done efficiently. In fact, the corresponding proximity operator is equivalent to solving a submodular minimization problem (SFM) using the minimum-norm point algorithm [1], which empirically runs in $O(p^2)$-time [1]. However, recent results show that in the worst-case analysis, min-norm point algorithm solves SFM in $O(p^7)$-time [6].

3.3 ℓ_q-regularized combinatorial sparsity models

In some applications, we may want to control not only the location of the non-zero coefficients, but also their magnitude. In this setting, it makes sense to consider a combination of combinatorial penalties with continuous regularizers. In particular, functions of the form $\mu F(\text{supp}(\boldsymbol{x})) + \nu\|\boldsymbol{x}\|_q$ are studied in [19]. The positive homogeneous convex envelope of a ℓ_q-regularized set function is then given by the dual of $\Omega_q^*(\boldsymbol{y}) = \max_{\boldsymbol{s}\in\{0,1\}^p, \boldsymbol{s}\neq 0} \frac{\|\boldsymbol{y}_{\text{supp}(\boldsymbol{s})}\|_q}{F(\boldsymbol{s})^{1/q}}$, which can be computed efficiently, for example, in the special case where F is submodular. However, if we only seek to enforce the combinatorial structure, using this approach will fail, as discussed in Section 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 [16, 18]. 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 ([17]). *Given a TU matrix $\boldsymbol{M} \in \mathbb{R}^{l\times m}$, an integral vector $\boldsymbol{c} \in \mathbb{Z}^l$, and a vector $\boldsymbol{\theta} \in \mathbb{R}^m$. The linear program (LP) $\max_{\boldsymbol{\beta}\in[0,1]^m}\{\boldsymbol{\theta}^T\boldsymbol{\beta} : \boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}\}$ has integral optimal solutions.*

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 $\boldsymbol{x}$ that can be written as*

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

for all feasible $\boldsymbol{x}$, and $g_{\text{TU}}(\boldsymbol{x}) = \infty$ otherwise, where $\boldsymbol{M} \in \mathbb{R}^{l\times(M+p)}$ is a TU matrix, $\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\omega} \\ \boldsymbol{s}\end{bmatrix}$, $\boldsymbol{\omega} \in \{0,1\}^M$ is a vector of binary variables useful for modeling latent variables , $\boldsymbol{d} \in \mathbb{R}^M$ and $\boldsymbol{e} \in \mathbb{R}^p$ are arbitrary weight vectors, and $\boldsymbol{c} \in \mathbb{Z}^l$ is an integral vector.

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 a TU penalty is given by the following LP:*

$$g_{\text{TU}}^{**}(\boldsymbol{x}) = \min_{\boldsymbol{s}\in[0,1]^p, \boldsymbol{\omega}\in[0,1]^M}\{\boldsymbol{d}^T\boldsymbol{\omega} + \boldsymbol{e}^T\boldsymbol{s} : \boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}, |\boldsymbol{x}| \leq \boldsymbol{s}\}$$

for all feasible $\boldsymbol{x}$, and $g_{\text{TU}}^{**}(\boldsymbol{x}) = \infty$ 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.

Remark 3. *The simplicity description does not need to be a linear function of $\boldsymbol{\omega}$ and $\boldsymbol{s}$. We can often find TU descriptions of higher order interactions that can be "lifted" to result in the linear TU penalty framework (cf., Section 6.3).*

Remark 4. *A weaker sufficient condition for lemma 2 to hold is for the system $\boldsymbol{M}\boldsymbol{\beta} \leq \boldsymbol{c}$ to be total dual integral [9] (e.g., submodular polyhedra). Then, penalties of the form described in Definition 4 will again satisfy Lemma 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 present a reinterpretation of several well-known convex norms in the literature, as well as introduce new ones.
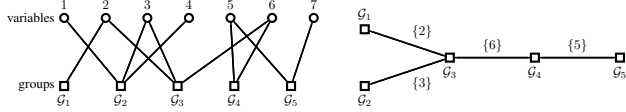
Figure 1: (Left) Bipartite graph representation, (Right) Intersection graph representation of the group structure $\mathfrak{G} = \{\mathcal{G}_1 = \{2\}, \mathcal{G}_2 = \{1,3,4\}, \mathcal{G}_3 = \{2,3,6\}, \mathcal{G}_4 = \{5,6\}, \mathcal{G}_5 = \{5,7\}\}$

5 Group sparsity

Group sparsity is an important class of structured sparsity models that arise naturally in machine learning applications (cf., [26] and the citations therein), where prior information on $x^\natural$ dictates certain groups of variables to be selected or discarded together.

A group sparsity model thus consists of a collection of potentially overlapping groups $\mathfrak{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_M\}$ that cover the ground set $\mathcal{P}$, where each group $\mathcal{G}_i \subseteq \mathcal{P}$ is a subset of variables. A group structure construction immediately supports two compact graph representations (c.f., Figure 1).

First, we can represent $\mathfrak{G}$ 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 $\mathcal{G}_j \in \mathfrak{G}$ iff $i \in \mathcal{G}_j$. We denote by $\boldsymbol{B} \in \{0,1\}^{p \times M}$ the biadjacency matrix of this bipartite graph; $B_{ij} = 1$ iff $i \in \mathcal{G}_j$, and by $\boldsymbol{E} \in \{0,1\}^{|\mathcal{E}| \times (M+p)}$ 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 $\mathfrak{G}$ as an intersection graph [3], where the vertices are the groups $\mathcal{G}_i \in \mathfrak{G}$. Two groups $\mathcal{G}_i$ and $\mathcal{G}_j$ are connected by an edge iff $\mathcal{G}_i \cap \mathcal{G}_j \neq \emptyset$. This structure makes it explicit whether groups themselves have cyclic interactions via variables, and identifies computational difficulties.

5.1 Group intersection sparsity

In group sparse models, we typically seek to express the support of $x^\natural$ using only 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 $F_\cap(\mathcal{S}) = \sum_{\mathcal{G}_i \in \mathfrak{G}, \mathcal{S} \cap \mathcal{G}_i \neq \emptyset} d_i$. The convexification of this function results in the ℓ_∞-group lasso norm (also known as ∞-CAP penalties) [12, 27], as shown in [1].

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

Definition 5 (Group intersection sparsity).

$$g_{\mathfrak{G},\cap}(\boldsymbol{x}) := \min_{\boldsymbol{\omega} \in \{0,1\}^M} \{\boldsymbol{d}^T \boldsymbol{\omega} : \boldsymbol{H}\boldsymbol{\beta} \leq 0, \mathbb{1}_{\text{supp}(x)} = \boldsymbol{s}\}$$

where $\boldsymbol{H}$ is the following matrix:

$$\boldsymbol{H} := \begin{bmatrix} -\boldsymbol{I}_M, \boldsymbol{H}_1 \\ -\boldsymbol{I}_M, \boldsymbol{H}_2 \\ \cdots \\ -\boldsymbol{I}_M, \boldsymbol{H}_p \end{bmatrix}, \boldsymbol{H}_k(i,j) = \begin{cases} 1 & \text{if } j = k, j \in \mathcal{G}_i \\ 0 & \text{otherwise} \end{cases}$$
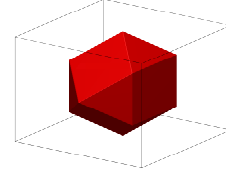

Figure 2: Unit norm ball of $g_{\mathfrak{G},\cap}^{**}$, for $\mathfrak{G} = \{\{1,2\}, \{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

and the vector $\boldsymbol{d} \in \mathbb{R}_+^M$ here corresponds to positive group weights. Recall that $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\omega} \\ \boldsymbol{s} \end{bmatrix}$, and thus $\boldsymbol{H}\boldsymbol{\beta} \leq 0$ simply corresponds to $s_j \leq w_i, \forall j \in \mathcal{G}_i$.

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

Here, $\boldsymbol{H}$ is TU, since each row of $\boldsymbol{H}$ 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 [17, Proposition 2.6].

Proposition 2 (Convexification). *The convex envelope of $g_{\mathfrak{G},\cap}(\boldsymbol{x})$ over the unit ℓ_∞-ball is*

$$g_{\mathfrak{G},\cap}^{**}(\boldsymbol{x}) = \begin{cases} \sum_{\mathcal{G}_i \in \mathfrak{G}} d_i \|\boldsymbol{x}_{\mathcal{G}_i}\|_\infty & \text{if } \boldsymbol{x} \in [-1,1]^p \\ \infty & \text{otherwise} \end{cases}$$

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 $x^\natural$ as the union of groups in $\mathfrak{G}$. In particular, we can seek the minimal set cover of $x^\natural$:

Definition 6 (Group ℓ_0-"norm", [3]). *The group ℓ_0-"norm" computes the weight of the minimal weighted set cover of $\boldsymbol{x}$ with group weights $\boldsymbol{d} \in \mathbb{R}_+^M$:*

$$g_{\mathfrak{G},0}(\boldsymbol{x}) := \min_{\boldsymbol{\omega} \in \{0,1\}^M} \{\boldsymbol{d}^T \boldsymbol{\omega} : \boldsymbol{B}\boldsymbol{\omega} \geq \mathbb{1}_{\text{supp}(\boldsymbol{x})}\},$$

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

Note that computing the group ℓ_0-"norm" is NP-Hard, since it corresponds to the minimum weight set cover problem. $g_{\mathfrak{G},0}(\boldsymbol{x})$ is a penalty that was previously considered in [3, 20, 11], and the latent group lasso was proposed in [20] as a potential convex surrogate for it, but it was not established as the tightest possible convexification.

The group ℓ_0-"norm" is *not* a submodular function, but if $\boldsymbol{B}$ is TU, it is a TU penalty, and thus is admits a tight
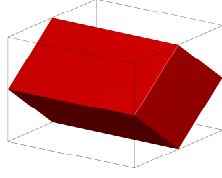
Figure 3: Unit norm ball of $g^{**}_{\mathfrak{G},0}$, for $\mathfrak{G} = \{\{1,2\}, \{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

convex relaxation. We show below that the convex envelope of the group ℓ_0-"norm" is indeed the ℓ_∞-latent group norm. It is worth noting that the ℓ_q-latent group lasso was also shown in [19] to be the positive homogeneous convex envelope of the ℓ_q-regularized group ℓ_0-"norm", i.e., of $\mu g(\boldsymbol{x})_{\mathfrak{G},0} + \nu\|\boldsymbol{x}\|_q$.

Proposition 3 (Convexification)**.** *When the group structure leads to a TU biadjacency matrix $\boldsymbol{B}$, the convex envelope of the group ℓ_0-"norm" over the unit ℓ_∞-ball is*

$$g^{**}_{\mathfrak{G},0}(\boldsymbol{x}) = \begin{cases} \min_{\boldsymbol{\omega}\in[0,1]^M}\{\boldsymbol{d}^T\boldsymbol{\omega} : \boldsymbol{B}\boldsymbol{\omega} \geq |\boldsymbol{x}|\} & \text{if } \boldsymbol{x} \in [-1,1]^p \\ \infty & \text{otherwise} \end{cases}$$

Thus, given a group structure $\mathfrak{G}$, one can check in polynomial time if it is TU [25] to guarantee that the ℓ_∞-latent group lasso will be the tightest relaxation.

Remark 5. *One important class of group structures that leads to a TU matrix $\boldsymbol{B}$ is given by acyclic groups, as shown in [3, Lemma 2]. The induced intersection graph for such groups is acyclic, as illustrated in Figure 1. In this case, the ℓ_∞-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 thus consider a natural extension of the above two penalties, where each group is weighed by the ℓ_0-norm of $\boldsymbol{x}$ restricted to the group.

Definition 7 (Group models with sparsity within groups)**.**

$$g_{\mathfrak{G},s}(\boldsymbol{x}) = \min_{\boldsymbol{\omega}\in\{0,1\}^M}\{\sum_{i=1}^M \omega_i\|\boldsymbol{x}_{\mathcal{G}_i}\|_0 : \boldsymbol{M}\boldsymbol{\beta} \leq 0, \mathbb{1}_{\mathrm{supp}(\boldsymbol{x})} = \boldsymbol{s}\}$$

where $\boldsymbol{M}$ here is either $\boldsymbol{M} = \boldsymbol{H}$ in Definition 5 or $\boldsymbol{M} = [-\boldsymbol{B}, \boldsymbol{I}_p]$ in Definition 6.

Unfortunately, this penalty leads to a non-TU penalty, and thus its corresponding convex surrogate given by Proposition 1 is not guaranteed to be tight.

Proposition 4. *Given any group structure $\mathfrak{G}$, $g_{\mathfrak{G},s}(\boldsymbol{x})$ is not a TU penalty.*

Proposition 5 (Convexification)**.** *The convex surrogate via Proposition 1 for $g_{\mathfrak{G},s}(\boldsymbol{x})$ with $\boldsymbol{M} = \boldsymbol{H}$ (i.e., the group intersection model with sparse groups) is given by*

$$\Omega_{\mathfrak{G},s}(\boldsymbol{x}) := \sum_{(i,j)\in\mathcal{E}} (\|\boldsymbol{x}_{\mathcal{G}_i}\|_\infty + |x_j| - 1)_+$$

for $\boldsymbol{x} \in [-1,1]^p$, and $\Omega_{\mathfrak{G},s}(\boldsymbol{x}) := \infty$ otherwise. Note that $\Omega_{\mathfrak{G},s}(\boldsymbol{x}) \leq g^{**}_{\mathfrak{G},s}(\boldsymbol{x})$.

By construction, the convex penalty proposed by Proposition 5 is different from the sparse group lasso in [22].

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

Proposition 6 (Convexification)**.** *The convex surrogate via Proposition 1 for $g_{\mathfrak{G},s}(\boldsymbol{x})$ with $\boldsymbol{M} = [-\boldsymbol{B}, \boldsymbol{I}_p]$ (i.e., the group ℓ_0-"norm" with sparse groups) is given by*

$$\Omega_{\mathfrak{G},s}(\boldsymbol{x}) := \min_{\boldsymbol{\omega}\in[0,1]^M}\{\sum_{(i,j)\in\mathcal{E}} (\omega_i + |x_j| - 1)_+ : \boldsymbol{B}\boldsymbol{\omega} \geq |\boldsymbol{x}|\}$$

for $\boldsymbol{x} \in [-1,1]^p$, and $g(\boldsymbol{x})_{\mathfrak{G},s} = \infty$ otherwise.

5.4 Sparse G-group cover

In this section, we provide a more direct formulation to enforce sparsity both on the coefficients and the group level. If the true signal $\boldsymbol{x}^\natural$ we are seeking is a sparse signal covered by at most G groups, it would make sense to look for the sparsest signal with a G-group cover that explains the data in (2). This motivates the following natural penalty.

Definition 8 (Sparse G-group cover)**.**

$$g_{\mathfrak{G},G}(\boldsymbol{x}) := \min_{\boldsymbol{\omega}\in\{0,1\}^p}\{\mathbb{1}^T\boldsymbol{s} : \boldsymbol{B}\boldsymbol{\omega} \geq \boldsymbol{s}, \mathbb{1}^T\boldsymbol{\omega} \leq G, \boldsymbol{s} = \mathbb{1}_{\mathrm{supp}(\boldsymbol{x})}\}$$

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

If the actual number of active groups is not known, G would be a parameter to tune. Note that $g_{\mathfrak{G},G}$ is an extension of the minimal group cover penalty (c.f., Section 5.2), where instead of looking for the signal with the smallest cover, we seek the sparsest signal that admit a cover with fewer than G groups. $g_{\mathfrak{G},G}$ is a TU penalty whenever $\widetilde{\boldsymbol{B}} = \begin{bmatrix} \boldsymbol{B} \\ \mathbb{1} \end{bmatrix}$ is TU [17, Proposition 2.1], which is the case, for example, when $\boldsymbol{B}$ is an interval matrix.

Proposition 7 (Convexification)**.** *When the group structure leads to a TU constraint matrix $\widetilde{\boldsymbol{B}}$, the convex envelope of $g_{\mathfrak{G},G}$ over the unit ℓ_∞-ball is*

$$g^{**}_{\mathfrak{G},G}(\boldsymbol{x}) = \min_{\boldsymbol{\omega}\in[0,1]^M}\{\|\boldsymbol{x}\|_1 : \boldsymbol{B}\boldsymbol{\omega} \geq |\boldsymbol{x}|, \mathbb{1}^T\boldsymbol{\omega} \leq G\}$$

for $\boldsymbol{x} \in [-1,1]^p$, and $g^{**}_{\mathfrak{G},G}(\boldsymbol{x}) = \infty$ otherwise.
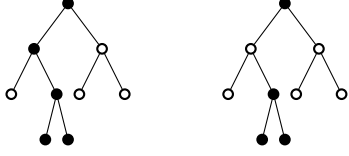
Figure 4: Valid selection (left), *Invalid* selection (right)
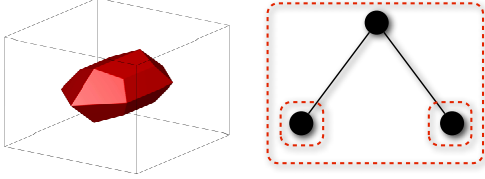
Figure 5: Unit norm ball of $g_{T,0}^{**}$, $\mathfrak{G}_H = \{\{1,2,3\},\{2\},\{3\}\}$

The resulting convex program thus combines the latent group lasso (c.f., Section 5.2) with the ℓ_1 norm and provides an alternative to the sparse group lasso in [22], for the overlapping groups case. In the supplementary material we provide a numerical illustration of its performance.

5.5 Hierarchical model

We study the hierarchical sparsity model, where the coefficients of $\boldsymbol{x}^\natural$ are organized over a tree $\mathcal{T}$, and the non-zero coefficients form a rooted connected subtree of $\mathcal{T}$ (cf., Figure 4). This model is popular in image processing due to the natural structure of wavelet coefficients [13, 7, 24]. We can describe such a hierarchical model as a TU model:

Definition 9 (Tree ℓ_0-"norm"). *We define the penalty encoding the hierarchical model on $\boldsymbol{x}$ as*

$$g_{T,0}(\boldsymbol{x}) := \begin{cases} \|\boldsymbol{x}\|_0 & \text{if } \boldsymbol{T}\mathbb{1}_{\mathrm{supp}(\boldsymbol{x})} \geq 0 \\ \infty & \text{otherwise} \end{cases}$$

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

This is indeed a TU model since each row of $\mathcal{T}$ contains at most two non-zero entries that sum up to zero [17, Proposition 2.6].

Proposition 8. *(Convexification) The convexification of the tree ℓ_0-"norm" over the unit ℓ_∞-ball is given by*

$$g_{T,0}^{**}(\boldsymbol{x}) = \begin{cases} \sum_{\mathcal{G} \in \mathfrak{G}_H} \|x_{\mathcal{G}}\|_\infty & \text{if } \boldsymbol{x} \in [-1,1]^p \\ \infty & \text{otherwise} \end{cases}$$

where the groups $\mathcal{G} \in \mathfrak{G}_H$ are defined as each node and all its descendants.

Note that the resulting convex norm is the ℓ_∞-hierarchical group norm [13], which is a special case of ℓ_∞-group

norm we studied in Section 5.1 as the convex envelope of $g_{\mathfrak{G},\cap}(\boldsymbol{x})$. In this sense, $g_{\mathfrak{G},\cap}(\boldsymbol{x})$ is equivalent to $g_{T,0}(\boldsymbol{x})$, for the group structure $\mathfrak{G}_H$ (for unit weights).

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 [28, 10, 8].

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

6.1 Group knapsack model

One natural model for dispersiveness allows only a certain budget of coefficients, e.g., only one, to be selected in each group:

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

Whenever the group structure forms a *partition* of $\mathcal{P}$, [19] shows that the positive homogeneous convex envelope of the ℓ_q-regularized group knapsack model, i.e., of $\mu F_{\boldsymbol{D}}(\mathrm{supp}(\boldsymbol{x})) + \nu\|\boldsymbol{x}\|_q$, is the exclusive norm in [28].

In what follows, we prove that $F_{\boldsymbol{D}}(\mathcal{S})$ is a TU penalty whenever the group structure leads to a TU biadjacency matrix $\boldsymbol{B}$ of the bipartite graph representation, which includes partition structures. We establish that the ℓ_∞-exclusive lasso, is actually the tightest convex relaxation of a more relaxed version of $F_{\boldsymbol{D}}(\mathrm{supp}(\boldsymbol{x}))$ for any TU group structure, and not necessarily partition groups.

Definition 10 (Group knapsack penalty). *Given a group structure $\mathfrak{G}$ that leads to a TU biadjacency matrix $\boldsymbol{B}$, $F_{\boldsymbol{D}}(\mathrm{supp}(\boldsymbol{x}))$ can be written as the following TU penalty:*

$$g_{\boldsymbol{D}}(\boldsymbol{x}) := \min_{\omega \in \{0,1\}} \{\omega : \boldsymbol{B}^T \mathbb{1}_{\mathrm{supp}(\boldsymbol{x})} \leq \omega \mathbb{1}\}$$

if $\boldsymbol{B}^T \mathbb{1}_{\mathrm{supp}(\boldsymbol{x})} \leq \mathbb{1}$, and $g_{\boldsymbol{D}}(\boldsymbol{x}) = \infty$ otherwise.

Note that if $\boldsymbol{B}$ is TU, $\boldsymbol{B}^T$ is also TU [17, Proposition 2.1]. Groups that form a partition of $\mathcal{P}$ are acyclic, thus the corresponding matrix $\boldsymbol{B}$ is TU trivially (cf., Remark 5).

Another important example of a TU group structure arises from the simple one-dimensional model of the neuronal signal suggested by [10]. 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 Δ. This structure corresponds to an interval matrix $\boldsymbol{B}^T = \boldsymbol{D}$, which is TU [17, Corollary 2.10].

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

Proposition 9 (Convexification). *The convex envelope of $g_D(\boldsymbol{x})$ over the unit ℓ_∞-ball when $\boldsymbol{B}^T$ is a TU matrix is given by*

$$g_D^{**}(\boldsymbol{x}) = \begin{cases} \max_{\mathcal{G} \in \mathfrak{G}} \|\boldsymbol{x}_{\mathcal{G}}\|_1 & \text{if } \boldsymbol{x} \in [-1,1]^p, \boldsymbol{B}^T|\boldsymbol{x}| \leq \mathbb{1} \\ \infty & \text{otherwise} \end{cases}$$

Notice that the convexification of g_D is not exactly the exclusive lasso; it has an additional budget constraint $\boldsymbol{B}^T|\boldsymbol{x}| \leq \mathbb{1}$. Thus in this case, regularizing with the ℓ_q-norm before convexifying lead to the loss of part of the structure. In fact, the exclusive norm is actually the convexification of a more relaxed version of g_D, where the constraint $\boldsymbol{\omega} \in \{0,1\}$ is relaxed to $\boldsymbol{\omega} \in \mathbb{Z}, \boldsymbol{\omega} \geq 0$.

6.2 Sparse group knapsack model

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 group knapsack penalty, resulting in the following TU penalty.

Definition 11 (Dispersive ℓ_0-"norm"). *Given a group structure $\mathfrak{G}$ that leads to a TU biadjacency matrix $\boldsymbol{B}$, we define the penalty encoding the sparse group knapsack model on $\boldsymbol{x}$ as*

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

We can compute the convex envelope of $g_{D,0}(\boldsymbol{x})$ in a similar fashion to Proposition 9.

Proposition 10. *(Convexification) The convexification of the dispersive ℓ_0-"norm" over the unit ℓ_∞-ball is given by*

$$g_{D,0}^{**}(\boldsymbol{x}) = \begin{cases} \|\boldsymbol{x}\|_1 & \text{if } \boldsymbol{x} \in [-1,1]^p, \boldsymbol{B}^T|\boldsymbol{x}| \leq \mathbb{1} \\ \infty & \text{otherwise} \end{cases}$$

It is worth mentioning that regularizing with the ℓ_q-norm here loses the underlying dispersive structure. In fact, the positively homogeneous convex envelope of $\mu g_{D,0}(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_q$ is given by the dual (cf., Section 3.3) of

$$\Omega_q^*(\boldsymbol{y}) = \max_{\boldsymbol{s} \in \{0,1\}^p, \boldsymbol{s} \neq 0, \boldsymbol{B}^T\boldsymbol{s} \leq \mathbb{1}} \frac{\|\boldsymbol{y}_{\text{supp}(\boldsymbol{s})}\|_q}{(\mathbb{1}^T\boldsymbol{s})^{1/q}}$$
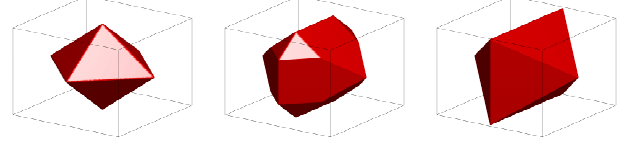
Figure 6: $g_{D,0}^{**}(\boldsymbol{x}) \leq 1$ (left) $g_{D,0}^{**}(\boldsymbol{x}) \leq 1.5$ (middle) $g_{D,0}^{**}(\boldsymbol{x}) \leq 2$ (right) for $\mathfrak{G} = \{\{1,2\},\{2,3\}\}$

which is simply the ℓ_1-norm. To see this, note that $\Omega_a^*(\boldsymbol{y})^q = \|\boldsymbol{y}\|_\infty$, since $\frac{\sum_{i \in \mathcal{S}}|y_i|^q}{|\mathcal{S}|} \leq \frac{|\mathcal{S}|\|\boldsymbol{y}\|_\infty^q}{|\mathcal{S}|}, \forall \mathcal{S} \subseteq \mathcal{P}$ which is achieved with equality by choosing the vector $\boldsymbol{s}$ having ones where $\boldsymbol{y}$ is maximal, and zeros elsewhere. Note that this vector satisfies $\boldsymbol{B}^T\boldsymbol{s} \leq \mathbb{1}$. As a result, the regularized convexification boils down to the ℓ_1-norm, since $\Omega_q(\boldsymbol{x}) = \sup_{\Omega_q^*(\boldsymbol{y}) \leq 1} \boldsymbol{x}^T\boldsymbol{y} = \|\boldsymbol{x}\|_1$, while the direct convexification is not even a norm (cf., Figure 6).

We illustrate the effect of this loss of structure via 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 12 (Pairwise dispersive penalty). *Given a graph $G(\mathcal{P}, \mathcal{E})$ with a TU edge-node incidence matrix $\boldsymbol{E}_G$ (e.g., bipartite graph), we define the penalty encoding the pairwise dispersive model as*

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

Note that this function is not submodular; in fact, $g_{\mathcal{G},\mathcal{D}}(\boldsymbol{x})$ is a *supermodular* function.

Proposition 11 (Convexification). *The convex envelope of $g_{\mathcal{G},\mathcal{D}}(\boldsymbol{x})$ over the unit ℓ_∞-ball is*

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

Proof. We use the linearization trick employed in [15] to reduce $g_{\mathcal{G},\mathcal{D}}(\boldsymbol{x})$ to a TU penalty. Let $\boldsymbol{s} = \mathbb{1}_{\text{supp}(\boldsymbol{x})}$,

$$\begin{aligned} g_{\mathcal{G},\mathcal{D}}(\boldsymbol{x}) &= \sum_{(i,j)\in\mathcal{E}} s_i s_j \\ &= \min_{\boldsymbol{z}\in\{0,1\}^{|\mathcal{E}|}} \Big\{ \sum_{(i,j)\in\mathcal{E}} z_{ij} : z_{ij} \geq s_i + s_j - 1 \Big\} \\ &= \min_{\boldsymbol{z}\in\{0,1\}^{|\mathcal{E}|}} \Big\{ \sum_{(i,j)\in\mathcal{E}} z_{ij} : \boldsymbol{E}_G\boldsymbol{s} \leq \boldsymbol{z} - \mathbb{1} \Big\} \end{aligned}$$
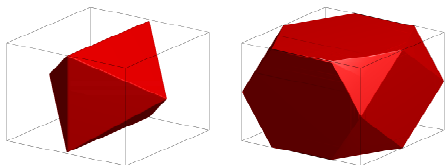
Figure 7: $g_{\mathcal{G},\mathcal{D}}^{**}(\boldsymbol{x}) = 0$ (left) $g_{\mathcal{G},\mathcal{D}}^{**}(\boldsymbol{x}) \leq 1$ (right) for $\mathcal{E} = \{\{1,2\},\{2,3\}\}$ (chain graph)

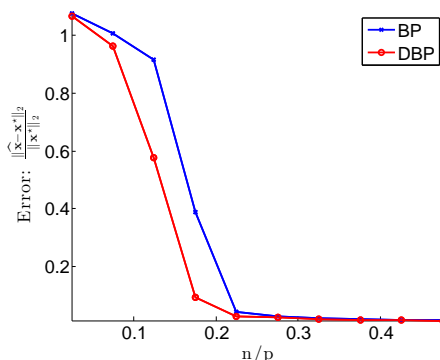Figure 8: *Recovery error of BP and DBP*

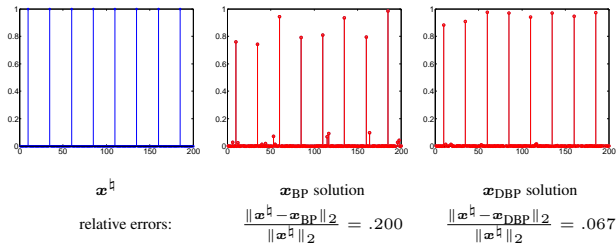| $\boldsymbol{x}^\natural$	$\boldsymbol{x}_{\mathrm{BP}}$ solution	$\boldsymbol{x}_{\mathrm{DBP}}$ solution
relative errors:	$\frac{\|\boldsymbol{x}^\natural - \boldsymbol{x}_{\mathrm{BP}}\|_2}{\|\boldsymbol{x}^\natural\|_2} = .200$	$\frac{\|\boldsymbol{x}^\natural - \boldsymbol{x}_{\mathrm{DBP}}\|_2}{\|\boldsymbol{x}^\natural\|_2} = .067$

Figure 9: *Example spike train recovery when $n = 0.18p$. The DBP formulation (right) shrinks the competing sparse coefficients within the Δ intervals, resulting in a better reconstruction overall sampling regimes than BP (middle).*

Now we can apply Proposition 1 to compute the convex envelope. The resulting convexification is again *not a norm* (c.f., Figure 7). ☐

7   Numerical illustration

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 ℓ_q-regularized dispersive ℓ_0-"norm" [19] versus its convex envelope without regularization over the unit ℓ_∞-ball in Section 6.2. To produce the recovery results in Figure 8, we generate a train of spikes of equal value for $\boldsymbol{x}^\natural$ in dimensions $p = 200$ with a refractoriness of $\Delta = 25$ (cf., Figure 9). We then recover $\boldsymbol{x}^\natural$ from its compressive measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\natural + \boldsymbol{w}$, where the noise $\boldsymbol{w}$ is also a sparse vector, with 15 non-zero Gaussian values of variance $\sigma = 0.01$ and $\boldsymbol{A}$ is a random column normalized Gaussian matrix. Since the noise is sparse, we encode the data via $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1 \leq \|\boldsymbol{w}\|_1$ using the true ℓ_1-norm of the noise. We produce the data randomly 20 times and report the averaged results.

Figure 8 measures the relative recovery error with $\frac{\|\boldsymbol{x}^\natural - \hat{\boldsymbol{x}}\|_2}{\|\boldsymbol{x}^\natural\|_2}$, as we vary the number of compressive measurements. The regularized convexification simply leads to Basis Pursuit formulation (BP), while the TU convexification results in the addition of a budget constraint $\boldsymbol{B}^T|\boldsymbol{x}| \leq \mathbb{1}$ to the BP formulation, as described in Section 6.2. We refer to the resulting criteria as Dispersive Basis Pursuit (DBP). Since the DBP criteria uses the fact that $\boldsymbol{x}^\natural$ lies in the unit ℓ_∞-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 8 shows that DBP outperforms 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 [10]. Hence, by changing the objective in the convexification, Figure 9 reinforces the message that we can lose the tightness in capturing certain structured sparsity models.

8   Conclusions

We have provided a principled recipe for designing convex formulations that jointly express models of simplicity and structure for sparse recovery, that promotes clustering or dispersiveness. 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 compute the biconjugation of the corresponding combinatorial simplicity objective subject to structure, and perform tractable recovery using standard optimization techniques.

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] FR Giles and William R Pulleyblank. Total dual integrality and integer polyhedra. *Linear algebra and its applications*, 25:191–196, 1979.

[10] 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.

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

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

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

[14] 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.

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

[16] 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.

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

[18] B. Nirav, L. Baldassarre, and V. Cevher. Tractability of interpretability via selection of group-sparse models. In *Information Theory (ISIT), 2013 IEEE International Symposium on*, 2013.

[19] G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties. *arXiv preprint arXiv:1205.1240*, 2012.

[20] G. Obozinski, L. Jacob, and J.P. Vert. Group lasso with overlaps: The latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

[21] C Seshadhri and Jan Vondrák. Is submodularity testable? *Algorithmica*, 69(1):1–25, 2014.

[22] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[23] Maurice Sion et al. On general minimax theorems. *Pacific J. Math*, 8(1):171–176, 1958.

[24] Akshay Soni and Jarvis Haupt. Efficient adaptive compressive sensing using sparse hierarchical learned dictionaries. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 1250–1254. IEEE, 2011.

[25] Klaus Truemper. Alpha-balanced graphs and matrices and GF(3)-representability of matroids. *Journal of Combinatorial Theory, Series B*, 32(2):112–139, 1982.

[26] Martin J Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253, 2014.

[27] Peng Zhao, Guilherme Rocha, and Bin Yu. Grouped and hierarchical model selection through composite absolute penalties. *Department of Statistics, UC Berkeley, Tech. Rep*, 703, 2006.

[28] Y. Zhou, R. Jin, and S. Hoi. Exclusive lasso for multitask feature selection. In *International Conference on Artificial Intelligence and Statistics*, pages 988–995, 2010.