Combinatorial Penalties:
Structure preserved by convex relaxations

Marwa El Halabi[1], Francis Bach[2], and Volkan Cevher[1]

[1]Laboratory for Information and Inference Systems (LIONS), EPFL
[2]Département d'Informatique de l'Ecole Normale Supérieure, INRIA-Sierra project-team

August 31, 2017

Abstract

In this paper, we study convex relaxations of combinatorial penalty functions. Specifically, we consider models penalized by the sum of an ℓ_p-norm and a set function defined over the support of the unknown parameter vector, which encodes prior knowledge on supports. We consider both *homogeneous* and *non-homogeneous* convex relaxations, and highlight the difference in the tightness of each relaxation through the notion of the *lower combinatorial envelope* of a set-function. We characterize necessary conditions under which the support of the unknown parameter vector can be correctly identified. We then propose a general adaptive estimator for convex monotone regularizers based on majorization-minimization, and identify sufficient conditions for support recovery in the asymptotic setting.

1 Introduction

In many applications, one aims at identifying a model of small complexity, well-approximated by a sparse set of coefficients that obey domain structure. Indeed, structured sparse parameters frequently appear in machine learning, signal processing, and statistics. Several penalties have been proposed in the literature to encode a priori knowledge on the structure of the support (set of non-zero coefficients), e.g., group Lasso [34], overlapping group Lasso [19, 21], hierarchical group Lasso [20, 35], exclusive Lasso [36], latent group Lasso [30]. Other works introduced more general formulations, based on submodular functions [1], atomic norms [6], totally unimodular constraints [10], graph models [17], or general ℓ_p-regularized set functions [29]. Non-convex approaches were also proposed in [4, 18]. For an overview, we refer the reader to [2] and references within.

In this paper, given a model parametrized by a vector of coefficients $w \in \mathbb{R}^V$ where $V = \{1, \cdots, d\}$, we consider regularizers that are convex relaxations of combinatorial penalties of the form $\frac{1}{q}F(\mathrm{supp}(w)) + \frac{1}{p}\|w\|_p^p$, where the set function F controls the structure of a model in terms of favored non-zero patterns and the ℓ_p-norm controls the magnitude of their coefficients for $p \in (1, \infty]$.

Convex relaxations of general combinatorial penalties were studied in several prior works. In particular, [1] showed that computing the tightest convex relaxation over the unit ℓ_∞-ball is tractable for the ensemble of *monotone submodular functions*. Similarly, the authors in [10] showed the tractability of such a relaxation for combinatorial penalties that can be described via *totally unimodular* constraints. Considering the case where $p \in (1, \infty)$ is appealing to avoid the clustering artifacts of the values of the learned vector, induced by the ℓ_∞-norm. This was proposed in [29], where the authors consider the tightest *homogeneous* convex relaxation of $\frac{1}{q}F(\mathrm{supp}(w)) + \frac{1}{p}\|w\|_p^p$ for general set functions and draw connections to the latent group Lasso norm [30].

There is a subtle difference between the two possible approaches of convexifications; by computing the tightest *homogeneous* convex relaxation, as adopted in [29], vs. computing the tightest *non-homogeneous* convex relaxation, as adopted in [10] for the special case where $p = \infty$. It is of interest to study the difference between the two approaches, in terms of which non-zero patterns can be encoded under each relaxation. In fact, the problem of support recovery, in the context of a learning problem regularized by a structure-inducing penalty, was only investigated so far in special cases, e.g., for submodular functions [1], or for the latent group Lasso [30]. The main objective of this paper is to study sparsity-inducing properties of both homogeneous and non-homogenous convex relaxation of general ℓ_p-regularized combinatorial penalties.

To that end, this paper makes the following contributions:

- We derive the non-homogeneous tightest convex relaxation of general ℓ_p-regularized combinatorial penalties (Section 2.1).
- We show that non-homogeneous relaxation is tight for any *monotone* set function, while the homogeneous relaxation is tight only for a smaller subset of set-functions. This is characterized through the notion of *lower combinatorial envelope* (Section 2.2).
- We characterize necessary conditions for non-zero patterns to be allowed as solutions to learning problems regularized by convex monotone penalties (Section 3.1), and in particular, for regularizers that correspond to convex relaxations of combinatorial functions (Section 4).
- We propose an adaptive weight estimator based on majorization-minimization, and identify sufficient conditions for support recovery, in the asymptotic regime (Section 3.2). We illustrate numerically in Section 5 that the adaptive scheme outperforms non-adaptive ones.

Notation. Given $w \in \mathbb{R}^d$ and a matrix $Q \in \mathbb{R}^{d\times d}$, w_J and Q_{JJ} denote the corresponding subvector and submatrix of w and Q. J^c denotes the complement of J. We let $\mathbb{1}_J$ be the indicator vector of the set J and accordingly $\mathbb{1}_i$ is the i-th basis vector. We drop the subscript for $J = V$, so that $\mathbb{1}_V = \mathbb{1}$ denotes the vector of all ones. The absolute value of $|w|$ is taken element-wise. Similarly, the comparison $w \geq w'$ and the product $w \circ v$ are taken element wise. For $p > 0$, the ℓ_p-(quasi) norm is given by $\|w\|_p = (\sum_{i=1}^d |w_i|^p)^{1/p}$, and $\|w\|_\infty = \max_i |w_i|$. For $p \in [1, \infty]$, we define the conjugate $q \in [1, \infty]$ through $\frac{1}{p} + \frac{1}{q} = 1$. We call the set of non-zero elements of a vector w the support, denoted by $\mathrm{supp}(w) = \{i : w_i \neq 0\}$. We use the common notation from submodular analysis, $w(A) = \sum_{i\in A} w_i$. We write $\overline{\mathbb{R}}_+$ for $\mathbb{R}_+ \cup \{+\infty\}$. For a function $f : \mathbb{R}^d \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$, we will denote by f^* its Fenchel-Legendre conjugate. We will denote by $\iota_S(w)$ the indicator function of the set S, taking value 0 on the set S and $+\infty$ outside.

2 Combinatorial penalties and convex relaxations

Let $V = \{1, \dots, d\}$ and $2^V = \{A | A \subseteq V\}$ be its power-set. We will consider positive-valued *monotone* (i.e., such that $F(A) \leq F(B), \forall A \subseteq B$) set functions of the form $F : 2^V \to \overline{\mathbb{R}}_+$ such that $F(\varnothing) = 0, F(A) > 0$, and $\forall A \subseteq V$. The domain of F is defined as $\mathcal{D} := \{A : F(A) < +\infty\}$ and we assume that it covers V, i.e., $\cup_{A\in\mathcal{D}} A = V$ (which is equivalent to assuming that F is finite at singletons).

A set function F is *submodular* if and only if for all $A \subseteq B$ and $i \in B^c$, $F(B \cup \{i\}) - F(B) \leqslant F(A \cup \{i\}) - F(A)$ (see, e.g.,[15, 2]). In what follows, unless explicitly stated, we do not assume that F is submodular.

We consider ℓ_p-regularized combinatorial penalties of the form $F_p(w) = \frac{1}{q}F(\mathrm{supp}(w)) + \frac{1}{p}\|w\|_p^p$ for $p \in [1, \infty)$ and $F_\infty(w) = F(\mathrm{supp}(w)) + \iota_{\|w\|_\infty \leq 1}(w)$. Such combinatorial regularizers lead to computationally intractable problems. Hence, it is necessary to find tight convex surrogate penalties that capture the encoded structure as much as possible. A natural candidate for a convex surrogate of F_p is then its *convex envelope* (largest convex lower bound), given by the biconjugate (the Fenchel conjugate of the Fenchel conjugate) F_p^{**}. Two general approaches were proposed in the literature to do this; one requires the surrogate to also be positively homogeneous [29] and thus considers the convex envelope of the positively homogeneous envelope of F_p, given by $F(\mathrm{supp}(w))^{1/q}\|w\|_p$, which we denote by Ω_p, the other computes instead the convex envelope of F_p directly [10], which we denote by Θ_p. Note that from the definition of convex envelope, it holds that $\Theta_p \geq \Omega_p$. In what follows, we defer all proofs to the Appendix.

2.1 Homogeneous and non-homogeneous convex envelopes

The homogeneous convex envelope Ω_p of F_p was derived in [29]. We recall in Lemma 1 one variational form of Ω_∞ and of Ω_p which highlights the relation between the two. Other variational forms are presented in the Appendix.

Lemma 1 ([29])**.** *The homogeneous convex envelope Ω_p of F_p is given by*

$$\Omega_p(w) = \inf_{\eta \in \mathbb{R}_+^d} \frac{1}{p} \sum_{j=1}^{d} \frac{|w_j|^p}{\eta_j^{p-1}} + \frac{1}{q}\Omega_\infty(\eta), \tag{1}$$

$$\Omega_\infty(w) = \min_{\alpha} \Big\{ \sum_{S \subseteq V} \alpha_S F(S) : \sum_{S \subseteq V} \alpha_S \mathbb{1}_S \geq |w|, \alpha_S \geq 0 \Big\}. \tag{2}$$

The non-homogeneous convex envelope of a set function F, over the unit ℓ_∞-ball was derived in [10]. The following new proposition generalizes it to any $p \in [1, \infty)$. For simplicity, the variational form (3) presented below holds only for monotone functions F; the general form and other variational forms that parallel the ones known for the homogeneous envelope are presented in the Appendix.

Lemma 2. *The non-homogeneous convex envelope Θ_p of F_p, for monotone functions F, is given by*

$$\Theta_p(w) = \inf_{\eta \in [0,1]^d} \frac{1}{p} \sum_{j=1}^{d} \frac{|w_j|^p}{\eta_j^{p-1}} + \frac{1}{q}\Theta_\infty(\eta), \tag{3}$$

$$\Theta_\infty(w) = \min_{\alpha} \Big\{ \sum_{S \subseteq V} \alpha_S F(S) : \sum_{S \subseteq V} \alpha_S \mathbb{1}_S \geq |w|, \sum_{S \subseteq V} \alpha_S = 1, \alpha_S \geq 0 \Big\}. \tag{4}$$

The infima in (1) and (3) (for $w \in \mathrm{dom}(\Theta_p)$) can be replaced by a minimization, if we extend $b \to \frac{a}{b}$ by continuity in zero with $\frac{a}{0} = \infty$ if $a \neq 0$ and 0 otherwise, as suggested in [22] and [3].

Remark 1. *If F is a monotone submodular function then $\Theta_\infty(w) = \Omega_\infty(w) = f_L(|w|), \forall w \in [-1, 1]^d$, where f_L denotes the Lovász extension of F [26], i.e., homogeneous and non-homogeneous envelopes are identical, on the unit ℓ_∞-ball, for monotone submodular functions.*

Note moreover that, for $p = 1$, both relaxations reduce to $\Omega_1 = \Theta_1 = \|\cdot\|_1$. Hence, the ℓ_1-relaxations essentially lose the combinatorial structure encoded in F. Thus, we will focus on the case $p > 1$.

In general however, the two relaxations do not coincide: note the added constraints $\eta \in [0, 1]^d$ in (3) and the sum constraint on α in (4). Moreover, this is not the case for the simple example of the ℓ_2-regularized cardinality function $F_2^{card}(w) = \frac{1}{2}\|w\|_0 + \frac{1}{2}\|w\|_2^2$, illustrated in Figure 2.1, where the non-homogeneous envelope is *tighter* than the homogeneous one. Indeed, the homogeneous envelope of F_2^{card} is simply the ℓ_1-norm, while the non-homogeneous envelope of F_2^{card} is given by $[F_2^{card}(w)]_i = |w_i|$ if $|w_i| \leq 1$ and

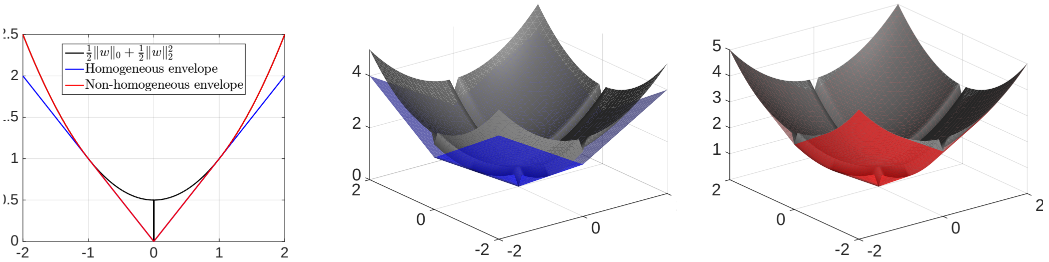

Figure 1: ℓ_2-regularized cardinality example in one dimension (left) and two dimensions (middle: homogeneous, right: non-homogeneous).

$[F_2^{card}(w)]_i = \frac{1}{2}|w_i|^2 + \frac{1}{2}$ otherwise. This penalty, called "Berhu" penalty [31], was introduced to produce a robust ridge regression estimator and was shown to be the convex envelope of F_2^{card} in [23].

In [29], it was shown that Ω_p are norms that belong to the broad family of H-norms [28, 3]. On the other hand, by virtue of being non-homogeneous, Θ_p are not norms in general.

The formulations (1) and (4) are jointly convex in (w, η), as noted in [29]. In general however, Ω_p and Θ_p can still be intractable to compute and to optimize. But for certain classes of functions, they are tractable. For example, for monotone submodular functions, $\Omega_\infty = \Theta_\infty$ is the Lovász extension of F, and hence can be computed by the usual greedy algorithm [2]. Moreover, efficient algorithms to compute Ω_p, the associated proximal operator and to solve learning problems regularized with Ω_p were proposed in [29]. Similarly, if F belongs to the class of "TU penalties", i.e., penalties that can be expressed by integer programs over totally unimodular constraints, introduced in [10], then Ω_∞, Θ_∞ and their associated Fenchel-type operators can be computed efficiently by linear programs.

2.2 Lower combinatorial envelopes

In this section, we characterize the tightness of the two convex relaxations. To that end, we generalize the notion of *lower combinatorial envelope* (LCE), which was introduced in [29]. The homogeneous LCE F_- of F is defined as the set function which agrees with the ℓ_∞-homogeneous convex relaxation of F at the vertices of the unit hypercube, i.e., $F_-(A) = \Omega_\infty(\mathbb{1}_A), \forall A \subseteq V$. For the non-homogeneous relaxation, we define the non-homogeneous LCE similarly as $\tilde{F}_-(A) = \Theta_\infty(\mathbb{1}_A)$. The ℓ_∞-relaxation reflects most directly the combinatorial structure of the function F, hence defining the LCE through it makes sense. Indeed, ℓ_p-relaxations only depend on F through the ℓ_∞-relaxation as expressed in the variational forms (1) and (3).

We say Ω_∞ is a tight relaxation of F if $F = F_-$. Similarly, Θ_∞ is a tight relaxation of F if $\tilde{F}_- = F$. Ω_∞ and Θ_∞ are then *extensions* of F from $\{0,1\}^d$ to $\mathbb{R}^d$; in this sense, the relaxation is tight for all w of the form $w = \mathbb{1}_A$. Moreover, following the definition of convex envelope, the relaxation Ω_∞ (resp. Θ_∞) is always the same for F and F_- (resp. F and $\tilde{F}_-$), and thus the LCE can be interpreted as the combinatorial function which the relaxation is actually able to capture.

For monotone submodular functions Ω_∞ is the Lovász extension [1], thus $F_-(A) = \Omega_\infty(\mathbb{1}_A) = f_L(\mathbb{1}_A) = F(A)$ and by Remark 1 $\tilde{F}_-(A) = \Theta_\infty(\mathbb{1}_A) = \Omega_\infty(\mathbb{1}_A) = F(A)$. Hence, both relaxations are tight in the case of monotone submodular functions, and the two LCEs are equal. We will see below that the LCEs are not equal in general and that the non-homogeneous is tighter.

The LCE value $F_-(A)$ can be interpreted, via the variational form (2), as the minimal fractional weighted set-cover A, a classical relaxation of the minimal weighted set-cover problem [27], as noted in [29]. It is in general not equal to $F(A)$. The following proposition shows that $\tilde{F}_-$ is equal to the *monotinization* of F, that is $\tilde{F}_{-}(A) = \inf_{S \subseteq V}\{F(S) : A \subseteq S\}$, for all set functions F, and is thus equal to the function itself if F is monotone.

Proposition 1. *The non-homogenous lower combinatorial envelope can be written as*

$$\tilde{F}_{-}(A) = \Theta_\infty(\mathbb{1}_A) = \inf\{\sum_{S \subseteq V} \alpha_S F(S) : \sum_{S \subseteq V} \alpha_S \mathbb{1}_S \geq \mathbb{1}_A, \sum_{S \subseteq V} \alpha_S = 1, \alpha_S \in \{0,1\}\}$$
$$= \inf_{S \subseteq V}\{F(S) : A \subseteq S\}.$$

Proof. To see why we can restrict α_S to be integral, let $\mathcal{E} = \{S : \alpha_S > 0\}$, then $\forall T \subseteq V$ such that $\exists e \in A, e \notin T$, then $\sum_{\alpha_S > 0, S \neq T} \alpha_S = 1$ and hence $\alpha_T = 0$. Hence $\forall S \in \mathcal{E}$ we have $A \subseteq S$ and $\sum_{\alpha_S > 0} \alpha_S F(S) = \min_{\alpha_S > 0} F(S)$. □

The non-homogeneous convex envelope is thus always tight for monotone functions and hence a "tighter" relaxation than the homogeneous one. Indeed, in certain instances, the homogeneous convex envelope loses the combinatorial structure encoded in F_p.

Example 1 (Range function)**.** *Consider the range function defined as range$(A) = \max(A) - \min(A) + 1$ where $\max(A)$ ($\min(A)$) denotes maximal (minimal) element in A, which induces the selection of interval supports. It was shown in [29] that the homogeneous LCE of the range function is the cardinality. In fact, this holds for any set function where $F(\{e\}) = 1$ for all singletons and $F(A) \geq |A|$. By Proposition 1, the non homogeneous LCE of the range function is itself.*

Example 2 (Dispersive ℓ_0-"norm")**.** *Given a set of predefined groups $\{G_1, \cdots, G_M\}$, consider the dispersive ℓ_0-"norm" defined as $F(A) = |A| + \iota_{B^T\mathbb{1}_A \leq \mathbb{1}}(A)$ where the columns of B correspond to the indicator vectors of the groups, i.e., $B_{V,i} = \mathbb{1}_{G_i}$. This penalty enforces the selection of sparse supports which are dispersive, in the sense that no two non-zeros are selected from the same group. The homogeneous LCE of the dispersive ℓ_0-"norm" is also the cardinality. By Proposition 1, the non homogeneous LCE of the dispersive ℓ_0-"norm" is itself.*

3 Sparsity inducing properties of monotone convex regularizers

The notion of LCE allowed us to characterize the combinatorial structure that can be captured by convex relaxations. We are further interested in investigating the combinatorial structure that can be enforced on solutions of learning problems regularized by convex monotone penalties in general, and by convex envelopes of ℓ_p-regularized combinatorial functions in particular.

We consider a fixed design matrix $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$ a vector of random responses. Given $\lambda_n > 0$, we define $\hat{w}$ as a minimizer of the regularized least-squares cost:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|y - Xw\|_2^2 + \lambda_n \Phi(w), \tag{5}$$

where Φ is any proper convex function which is monotonic in the absolute values of w, that is $|w| \geq |w'| \Rightarrow \Phi(w) \geq \Phi(w')$. We study the sparsity-inducing properties of solutions of (5). We determine in Section 3.1 which non-zero patterns are allowed and in Section 3.2 which sufficient conditions lead to correct estimation.

3.1 Continuous stable supports

We now introduce the notion of *continuous* stable supports, which characterizes supports with respect to the continuous penalty Φ. In Section 4.1, we will relate this to the notion of *discrete* stable supports, which characterize supports with respect to the combinatorial penalty F.

Definition 1 (Decomposability). *Given $J \subset V$ and $w \in \mathbb{R}^d$, $\operatorname{supp}(w) \subseteq J$, we say that Φ is* decomposable *at w w.r.t J if $\exists M_J > 0$ such that $\forall \Delta \in \mathbb{R}^d, \operatorname{supp}(\Delta) \subseteq J^c$,*

$$\Phi(w + \Delta) \geq \Phi(w) + M_J \|\Delta\|_\infty.$$

Definition 2 (Continuous stability). *We say that $J \subset V$ is* weakly stable *w.r.t Φ if* there exists $w \in \mathbb{R}^d$, $\operatorname{supp}(w) = J$ *such that Φ is* decomposable *at w w.r.t J. Furthermore, we say that $J \subset V$ is* strongly stable *w.r.t Φ if* for all $w \in \mathbb{R}^d$ *such that* $\operatorname{supp}(w) \subseteq J$*, Φ is* decomposable *at w w.r.t J.*

Proposition 2 considers slightly more general learning problems than (5) and shows that weak stability is a necessary condition for a non-zero pattern to be allowed as a solution.

Proposition 2. *The minimizer $\hat{w}$ of $\min_{w \in \mathbb{R}^d} L(w) - z^\top w + \lambda \Phi(w)$, where L is a strongly-convex and smooth loss function and $z \in \mathbb{R}^d$ has a continuous density, has a weakly stable support w.r.t. Φ, with probability one.*

This result extends and simplifies previous results, e.g., in [1] where this was proved for the special case of Ω_∞ and submodular functions with quadratic loss functions. The proof we present, in the Appendix, is short and simpler.

Corollary 1. *Assume $y \in \mathbb{R}^d$ has a continuous density and $X^T X$ is invertible. Then the minimizer $\hat{w}$ of Eq.* (5) *is unique and its support* $\operatorname{supp}(\hat{w})$ *is weakly stable w.r.t Φ, with probability one.*

3.2 Adaptive estimation

Restricting the choice of regularizers in (5) to convex relaxations as surrogates to the combinatorial penalties is motivated by computational tractability concerns. However, other non-convex sparsity-inducing regularization functions have been proposed in the literature. For example, ℓ_α-quasi-norms [24, 14] or more generally penalties of the form $\Phi(w) = \sum_{i=1}^d \phi(|w_i|)$, where ϕ is a monotone concave penalty [12, 8, 16] can be more advantageous than the ℓ_1-norm. Such penalties are closer to the ℓ_0-quasi-norm and penalize more aggressively small coefficients, and thus lead to a sparsity-inducing effect stronger than ℓ_1. The authors in [22] extended this to define ℓ_α/ℓ_2- quasi-norm $\Phi(w) = \sum_{i=1}^M \|w_{G_i}\|_\alpha$ for some $\alpha \in (0, 1)$, which enforce sparsity at the group level more aggressively. We generalize this to $\Phi(|w|^\alpha)$ where Φ is any structured sparsity-inducing monotone and convex regularizer.

These non-convex penalties lead to intractable estimation problems, but approximate solutions can be obtained by majorization-minimization algorithms, as suggested for e.g., in [13, 38, 5].

Lemma 3. *Let Φ be any monotone convex function, then for all $w^0 \in \mathbb{R}^d$, $\Phi(|w|^\alpha)$ admits the following majorizer $\Phi(|w|^\alpha) \leq (1-\alpha)\Phi(|w^0|^\alpha) + \alpha\Phi(|w^0|^{\alpha-1} \circ |w|)$, which is tight at w^0.*

We consider the adaptive weight estimator (6) resulting from applying a 1-step majorization-minimization to (5),

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|y - Xw\|_2^2 + \lambda_n \Phi(|w^0|^{\alpha-1} \circ |w|), \tag{6}$$

where w^0 is a $\sqrt{n}$-consistent estimator to w^*, that is converging to $\boldsymbol{w}^*$ at rate $1/\sqrt{n}$ (typically obtained from $w^0 = \mathbb{1}$ or ordinary least-squares).

We study sufficient support recovery and estimation consistency conditions for (6) for general convex monotone regularizers Φ. To that end, we assume that the linear model is well-specified, with $y = Xw^* + \epsilon$, where ϵ is a vector of i.i.d. random variables with mean 0 and variance σ^2. Sufficient support recovery and estimation consistency conditions for the (non-adaptive) estimator (5) have been established for homogeneous convex envelopes of submodular functions, for $p = \infty$ in [1] and for general p in [29], in the high dimensional setting, and for latent group Lasso norm in [30], in the classical setting.

For simplicity, we consider in this paper the classical asymptotic regime in which the model generating the data is of fixed finite dimension p while $n \to \infty$. We further assume that $Q = X^T X/n$ is positive definite and thus the minimizer of (6) is unique, we denote it by $\hat{w}$. The following Theorem extends the results from [37] for the ℓ_1-norm.

Theorem 1. *[Consistency and Support Recovery] Let $\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a convex, normalized ($\Phi(0) = 0$) and monotone function and denote by J the true support $J = \operatorname{supp}(w^*)$. If J is strongly stable with respect to Φ and λ_n satisfies $\frac{\lambda_n}{\sqrt{n}} \to 0$, $\frac{\lambda_n}{n^{\alpha/2}} \to \infty$, then the estimator* (6) *is consistent and asymptotically normal, i.e., it satisfies asymptotic normality, i.e.:*

$$\sqrt{n}(\hat{w}_J - w^*_J) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q^{-1}_{JJ}), \tag{7}$$

and

$$P(\operatorname{supp}(\hat{w}) = J) \to 1. \tag{8}$$

Hence, the adaptive estimator (6) is able to correctly identify any strongly stable support for any nomalized monotone convex regularizer. In particular, we are interested in structure-inducing regularizers that correspond to convex relaxations of combinatorial penalties.

4 Sparsity inducing properties of relaxations of combinatorial penalties

In this Section, we study the sparsity inducing properties of Ω_p and Θ_p. Both penalties are normalized monotone convex functions to which the necessary and sufficient conditions identified in Sections 3.1 and 3.2 apply. We investigate how these conditions translate to conditions with respect to combinatorial penalties. To that end we recall the concept of discrete stable sets [1], also referred to as *flat* or *closed* sets [25]. We refer to such sets as weak discrete stable sets and introduce a stronger notion of discrete stability.

4.1 Discrete stable supports

Definition 3 (Discrete stability)**.** *Given a monotone set function $F : 2^V \to \mathbb{R} \cup \{+\infty\}$, a set $J \subseteq V$ is said to be* weakly stable *w.r.t F if $\forall i \in J^c, F(J \cup \{i\}) > F(J)$.*
A set $J \subseteq V$ is said to be strongly stable *w.r.t F if $\forall A \subseteq J, \forall i \in J^c, F(A \cup \{i\}) > F(A)$.*

It is interesting to note that for monotone submodular functions, weak and strong stability are equivalent. In fact, this equivalence holds for a more general class of functions, we call ρ-submodular.

Definition 4. *A function $F : 2^V \to \mathbb{R}$ is ρ-submodular iff $\exists \rho \in (0, 1]$ s.t., $\forall B \subseteq V, A \subseteq B, i \in B^c$*

$$\rho[F(B \cup \{i\}) - F(B)] \leq F(A \cup \{i\}) - F(A)$$

The notion of ρ-submodularity is related to another notion of “approximate" submodularity, called weak submodularity (c.f., [7, 11]). We show in the appendix that ρ-submodularity is a stronger condition than weak submodularity.

Proposition 3. *If F is a monotone function, F is ρ-submodular iff weak stability is equivalent to strong stability.*

Example 3. *The range function range$(A) = \max(A) - \min(A) + 1$ is ρ-submodular with $\rho = \frac{1}{d-1}$.*

4.2 Relation between discrete and continuous stability

It is more natural to characterize which supports can be correctly estimated w.r.t the combinatorial penalty itself, without going through its relaxations. This is indeed achieved by the notion of discrete strong stability.

Proposition 4. *Given any monotone set function F, all sets $J \subseteq V$ strongly stable w.r.t to F are also strongly stable w.r.t Ω_p and Θ_p.*

It follows then by Theorem 1 that discrete strong stability is a sufficient condition for correct estimation.

Corollary 2. *If Φ is equal to Ω_p or Θ_p and $\mathrm{supp}(w^*) = J$ is strongly stable w.r.t F then the adaptive estimator* (6) *is consistent and correctly recovers the support.*

Furthermore, if F is ρ-submodular, then by Proposition 3, it is enough for $\mathrm{supp}(w^*) = J$ to be weakly stable w.r.t F for Corollary 2 to hold. Conversely, Proposition 5 below shows that discrete strong stability is also a necessary condition for continuous strong stability, in the case where $p = \infty$ and F is equal to its LCE.

Proposition 5. *If $F = F_-$ and J is strongly stable w.r.t Ω_∞, then J is strongly stable w.r.t F. Similarly, for any monotone F, if J is strongly stable w.r.t Θ_∞, then J is strongly stable w.r.t F.*

Finally, in the special case of monotone submodular function, the following Corollary 3 and Proposition 4 demonstrate that all definitions of stability become equivalent.

Corollary 3. *If F is a monotone submodular function and J is weakly stable w.r.t $\Omega_\infty = \Theta_\infty$ then J is weakly stable w.r.t F.*

Corollary 3 recovers the result in [1] showing that weakly stable supports correspond to the set of allowed sparsity patterns for monotone submodular functions.

4.3 Examples

Cardinality: The cardinality function and both its homogeneous and non-homogeneous relaxation, given by the ℓ_1-norm, are strictly monotone, hence all sets are stable (strongly and weakly) w.r.t to them.

Range function: Since the range function is $\frac{1}{d-1}$-submodular, then its stable (strongly and weakly) supports are exactly interval supports. Since the range function is monotone, then by Proposition 5, sets strongly stable w.r.t its non-homogeneous convex envelope Θ^r_∞ are interval supports too. On the other hand, its homogeneous convex envelope $\Omega^r_\infty = \|\cdot\|_1$ admits all sets as strongly stable.

Modified range function: The range function can be made to be a submodular function, if scaled by a constant as suggested in [1], yielding the monotone submodular function $F^{\mathrm{mr}}(A) = d - 1 + \mathrm{range}(A), \forall A \neq \emptyset$ and $F^{\mathrm{mr}}(\varnothing) = 0$. Since F^{mr} is submodular, both homogeneous and non-homogeneous ℓ_∞-convex envelopes are identical and correspond to the ℓ_1/ℓ_∞-group norm with groups defined as $\mathcal{G} = \{[1,k] : 1 \leq k \leq d\} \cup \{[k,d] : 1 \leq k \leq d\}$. This norm was proposed to induce interval patterns by [21] and shown to be the convex envelope of F^{mr} in [1].

5 Numerical Illustration

To illustrate the results presented in this paper, we consider the problem of estimating the support of a parameter vector $w \in \mathbb{R}^d$ whose support is an interval. It is natural then to choose as combinatorial penalty the range function whose stable supports are intervals. We aim to study the effect of adaptive weights, as

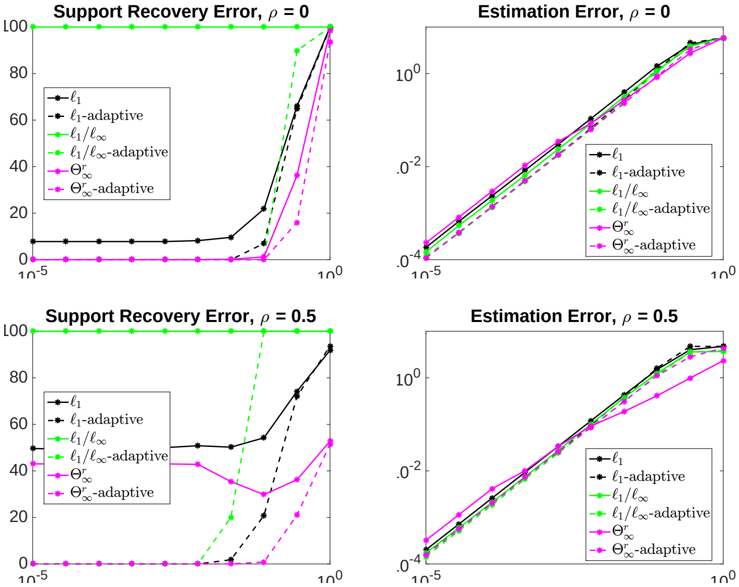

Figure 2: (Left column) Best Hamming distance and (Right column) best least square error to the true vector w^*, along the regularization path, averaged over 5 runs.

well as the effect the choice of homogeneous vs. non-homogeneous convex relaxation for regularization, on the quality of support recovery.

As discussed in Section 4.3, the ℓ_∞-homogeneous convex envelope of the range is simply the ℓ_1-norm. Its ℓ_∞-non-homogeneous convex envelope Θ^r_∞ can be computed using the formulation (3), where only interval sets need to be considered in the constraints, leading to a quadratic number of constraints. We also consider the ℓ_1/ℓ_∞-norm that corresponds to the convex relaxation of the modified range function F^{mr}.

We consider a simple regression setting in which $w^* \in \mathbb{R}^d$ is a constant signal whose support is an interval. The choice of $p = \infty$ is well suited for constant valued signals. The design matrix $X \in \mathbb{R}^{d \times n}$ is either drawn as (1) an i.i.d Gaussian matrix with normalized columns, or (2) a correlated Gaussian matrix with normalized columns, with the off-diagonal values of the covariance matrix set to a value $\rho = 0.5$. We observe noisy linear measurements $y = Xw^* + \epsilon$, where the noise vector is i.i.d. with variance σ^2, where σ is varied between 10^{-5} and 1. We solve problem (6) with and without adaptive weights $|w^0|^{\alpha-1}$, where w^0 is taken to be the least squares solution and $\alpha = 0.3$.

We assess the estimators obtained through the different regularizers both in terms of support recovery and in terms of estimation error. Figure 5 plots (in logscale) these two criteria against the noise level σ. We plot the best achieved error on the regularization path, where the regularization parameter λ was varied between 10^{-6} and 10^3. We set the parameters to $d = 250, k = 100, n = 500$.

We observe that the adaptive weight scheme helps in support recovery, especially in the correlated design setting. Indeed, Lasso is only guaranteed to recover the support under an "irrepresentability condition" [37]. This is satisfied with high probability only in the non-correlated design. On the other hand, adaptive weights allow us to recover any strongly stable support, without any additional condition, as shown in Theorem 1. The ℓ_1/ℓ_∞-norm performs poorly in this setup. In fact, the modified range function F^{mr}, introduced a gap of d between non-empty sets and the empty set. This leads to the undersirable behavior, already documented in [1, 21] of adding all the variables in one step, as opposed to gradually. Adaptive weights seem to correct for this effect, as seen by the significant improvement in performance. Finally, note that choosing the "tighter" convex relaxation leads to better support recovery. Indeed, Θ^r_∞ performs better than ℓ_1-norm in all setups.

6 Conclusion

We presented an analysis of homogeneous and non-homogeneous convex relaxations of ℓ_p-regularized combinatorial penalties. Our results show that structure encoded by submodular priors can be equally well expressed by both relaxations, while non-homogeneous relaxation is able to express the structure of general monotone set functions. We also identified necessary and sufficient stability conditions on the supports to be correctly identified. We proposed an adaptive weight scheme that is guaranteed to recover supports that satify the necessary stability conditions, with no other additional assumption.

Acknowledgments

This work was supported in part by the European Commission under ERC Future Proof, SNF 200021-146750, SNF CRSII2-147633, NCCR Marvel. Francis Bach acknowledges support from the chaire Economie des nouvelles données with the data science joint research initiative with the fonds AXA pour la recherche, and the Initiative de Recherche "Machine Learning for Large-Scale Insurance" from the Institut Louis Bachelier.

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] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[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] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted minimization. *Journal of Fourier analysis and applications*, 14(5):877–905, 2008.

[6] 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.

[7] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[8] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[9] Shaddin Dughmi. Submodular functions: Extensions, distributions, and algorithms. a survey. *arXiv preprint arXiv:0912.0322*, 2009.

[10] M. El Halabi and V. Cevher. A totally unimodular view of structured sparsity. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, pp. 223–231*, 2015.

[11] Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity. *arXiv preprint arXiv:1612.00804*, 2016.

[12] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[13] Mário AT Figueiredo, José M Bioucas-Dias, and Robert D Nowak. Majorization–minimization algorithms for wavelet-based image restoration. *IEEE Transactions on Image processing*, 16(12):2980–2991, 2007.

[14] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[15] S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

[16] Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *IEEE Transactions on Signal Processing*, 57(12):4686–4698, 2009.

[17] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. A nearly-linear time framework for graph-structured sparsity. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 928–937, 2015.

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

[19] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.

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

[21] Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12(Oct):2777–2824, 2011.

[22] Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 366–373, 2010.

[23] 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.

[24] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

[25] Andreas Krause and Carlos E Guestrin. Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*, 2012.

[26] L. Lovász. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pages 235–257. Springer, 1983.

[27] László Lovász. On the ratio of optimal integral and fractional covers. *Discrete mathematics*, 13(4):383–390, 1975.

[28] Charles A Micchelli, Jean M Morales, and Massimiliano Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, pages 1–35, 2013.

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

[30] Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

[31] Art B Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

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

[33] Jan Vondrák. Continuous extensions of submodular functions. CS 369P: Polyhedral techniques in combinatorial optimization, `http://theory.stanford.edu/~jvondrak/CS369P-files/lec17.pdf`, November 2010.

[34] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[35] 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.

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

[37] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[38] Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.

7 Appendix

7.1 Variational forms of convex envelopes (Proof of lemma 2 and Remark 1)

In this section, we recall the different variational forms of the homogeneous convex envelope derived in [29] and derive similar variational forms for the non-homogeneous convex envelope, which includes the ones stated in lemma 2). These variational forms will be needed in some of our proofs below.

Lemma 4. *The homogeneous convex envelope Ω_p of F_p admits the following variational forms.*

$$\Omega_\infty(w) = \min_\alpha \{ \sum_{S \subseteq V} \alpha_S F(S) : \sum_{S \subseteq V} \alpha_S \mathbb{1}_S \geq |w|, \alpha_S \geq 0 \}. \tag{9}$$

$$\Omega_p(w) = \min_v \{ \sum_{S \subseteq V} F(S)^{1/q} \|v^S\|_p : \sum_{S \subseteq V} v^S = |w|, \mathrm{supp}(v^S) \subseteq S \}. \tag{10}$$

$$= \max_{\kappa \in \mathbb{R}^d_+} \sum_{i=1}^d \kappa_i^{1/q} |w_i| \textit{ s.t. } \kappa(A) \leq F(A), \forall A \subseteq V. \tag{11}$$

$$= \inf_{\eta \in \mathbb{R}^d_+} \frac{1}{p} \sum_{j=1}^d \frac{|w_j|^p}{\eta_j^{p-1}} + \frac{1}{q} \Omega_\infty(\eta). \tag{12}$$

The non-homogeneous convex envelope of a set function F, over the unit ℓ_∞-ball was derived in [10]. The following proposition generalizes it to any $p \in [1, \infty)$ and derive variational forms that parallel the ones known for the homogeneous envelope.

Lemma 5. *The non-homogeneous convex envelope Θ_p of F_p admits the following variational forms.*

$$\Theta_\infty(w) = \inf \{ \sum_{S \subseteq V} \alpha_S F(S) : \sum_{S \subseteq V} \alpha_S \mathbb{1}_S \geq |w|, \sum_{S \subseteq V} \alpha_S = 1, \alpha_S \geq 0 \}. \tag{13}$$

$$\Theta_p(w) = \max_{\kappa \in \mathbb{R}^d_+} \sum_{j=1}^d \psi_j(\kappa_j, w_j) + \min_{S \subseteq V} F(S) - \kappa(S), \ \forall w \in \mathrm{dom}(\Theta_\infty(w)). \tag{14}$$

$$= \inf_{\eta \in [0,1]^d} \frac{1}{p} \sum_{j=1}^d \frac{|w_j|^p}{\eta_j^{p-1}} + \frac{1}{q} f^-(\eta), \tag{15}$$

where we define

$$\psi_j(\kappa_j, w_j) := \begin{cases} \kappa_j^{1/q} |w_j| & \textit{if } |w_j| \leq \kappa_j^{1/p} \\ \frac{1}{p}|w_j|^p + \frac{1}{q}\kappa_j & \textit{otherwise.} \end{cases}$$

In [10], it was shown that the non-homogeneous convex envelope of a set function F, over the unit ℓ_∞-ball is given by $\inf_{s \in [0,1]^d} \{ f(s) : s \geq |w| \}$ where f is any proper $(\mathrm{dom}(f) \neq \emptyset)$ lower semi-continuous (l.s.c.) convex extension of F (c.f., Lemma 1 [10]). A natural choice for f is the convex closure of F, which corresponds to the *tightest* convex extension of F on $[0, 1]^d$.

Definition 5 (Convex Closure; c.f., [9, Def. 3.1])**.** *Given a set function $F : 2^V \to \overline{\mathbb{R}}$, the convex closure $f^- : [0, 1]^d \to \overline{\mathbb{R}}$ is the point-wise largest convex function from $[0, 1]^d$ to $\overline{\mathbb{R}}$ that always lowerbounds f.*

Definition 6 (Equivalent definition of Convex Closure; c.f., [33, Def. 1] and [9, Def. 3.2])**.** *Given any set function $f : \{0, 1\}^n \to \mathbb{R}$, the convex closure of f can also be defined $\forall w \in [0, 1]^n$ as:*

$$f^-(w) = \min \{ \sum_{S \subseteq V} \alpha_S F(S) : w = \sum_{S \subseteq V} \alpha_S \mathbb{1}_S, \sum_{S \subseteq V} \alpha_S = 1, \alpha_S \geq 0 \}$$

It is interesting to note that $f^-(w) = f_L(w)$ where f_L is Lovász extension iff F is a submodular function [33]. For set functions that take ∞ value, we extend definition 6 by replacing the $\min$ by $\inf$, to have $f^-(\mathbb{1}_S) = +\infty$ iff $F(S) = +\infty$.

Proposition 6 (c.f., [9, Prop. 3.23])**.** *The minimum values of a proper set function F and its convex closure f^- are equal, i.e.,*

$$\min_{w \in [0,1]^d} f^-(w) = \min_{S \subseteq V} F(S)$$

If S is a minimizer of $f(S)$, then $\mathbb{1}_S$ is a minimizer of f^-. Moreover, if w is a minimizer of f^-, then every set in the support of α, where $f^-(w) = \sum_{S \subseteq V} \alpha_S F(S)$, is a minimizer of F.

Proof. First note that, $\{0,1\}^d \subseteq [0,1]^d$ implies that $f^-(w^*) \leq F(S^*)$. On the other hand, $f^-(w^*) = \sum_{S \subseteq V} \alpha^*_S F(S) \geq \sum_{S \subseteq V} \alpha^*_S F(S^*) = F(S^*)$. The rest of the proposition follows directly. □

Given the choice of the extension $f = f^-$, the variational form (13) of Θ_∞ given in lemma 5 follows directly from definition 6 and proposition 6, as shown in the following corollary.

Corollary 4. *Given any set function $F : 2^V \to \mathbb{R} \cup \{+\infty\}$ and its corresponding convex closure f^-, the convex envelope of $F(\mathrm{supp}(w))$ over the unit ℓ_∞-ball is given by*

$$\begin{aligned} \Theta_\infty(w) &= \inf\{\sum_{S \subseteq V} \alpha_S F(S) : \sum_{S \subseteq V} \alpha_S \mathbb{1}_S \geq |w|, \sum_{S \subseteq V} \alpha_S = 1, \alpha_S \geq 0\}. \\ &= \inf_v \{\sum_{S \subseteq V} F(S) \|v^S\|_\infty : \sum_{S \subseteq V} v^S = |w|, \sum_{S \subseteq V} \|v^S\|_\infty = 1, \mathrm{supp}(v^S) \subseteq S\}. \end{aligned}$$

Proof. f^- satisfies the first 2 assumptions required in Lemma 1 of [10], namely, f^- is a lower semi-continuous convex extension of F which satisfies

$$\max_{S \subseteq V} m(S) - F(S) = \max_{w \in [0,1]^d} m^T w - f^-(w), \forall m \in \mathbb{R}^d_+$$

To see this note that $m^T w^* - f^-(w^*) = \sum_{S \subseteq V} \alpha^*_S (m^T \mathbb{1}_S - F(S)) \geq \sum_{S \subseteq V} \alpha^*_S (m^T \mathbb{1}_{S^*} - F(S^*)) = m(S^*) - F(S^*)$. The other inequality is trivial. The corollary then follows directly from Lemma 1 in [10] and (extended) definition 6. □

Note that $\mathrm{dom}(\Theta_\infty) = \{w : \exists s \in [0,1]^d \cap \mathrm{dom}(f^-), s \geq |w|\}$. Note also that Θ_∞ is monotone even if F is not. On the other hand, if F is monotone, then f^- is monotone on $[0,1]^d$ and $\Theta_\infty(w) = f^-(|w|)$. Then the proof of remark 1 follows, since if F is a monotone submodular function and f_L is its Lovász extension, then $\Theta_\infty(w) = f^-(|w|) = f_L(|w|) = \Omega_\infty(w), \forall w \in [-1,1]^d$, where the last equality was shown in [1].

Next, we derive the convex relaxation of F_p for a general $p \in [1, \infty)$.

Proposition 7. *Given any set function $F : 2^V \to \mathbb{R} \cup \{+\infty\}$ and its corresponding convex closure f^-, the convex envelope of $F_{\mu\lambda}(w) = \mu F(\mathrm{supp}(w)) + \lambda \|w\|_p^p$ is given by*

$$\Theta_p(w) = \inf_{\eta \in [0,1]^d} \lambda \sum_{j=1}^d \frac{|w_j|^p}{\eta_j^{p-1}} + \mu f^-(\eta).$$

Note that $\mathrm{dom}(\Theta_p) = \{w | \exists \eta \in [0,1]^d$ *s.t* $\mathrm{supp}(w) \subseteq \mathrm{supp}(\eta), \eta \in \mathrm{dom}(f^-)\} \supseteq \mathrm{dom}(\Theta_\infty)$.

Proof. Given any proper l.s.c. convex extension f of F, we have: First for the case where $p = 1$:

$$\begin{aligned}
F^*_{\mu\lambda}(s) &= \sup_{w \in \mathbb{R}^n} w^T s - \mu F(\mathrm{supp}(w)) - \lambda \|w\|_1 \\
&= \sup_{\eta \in \{0,1\}^d} \sup_{\substack{\mathbb{1}_{\mathrm{supp}(w)} = \eta \\ \mathrm{sign}(w) = \mathrm{sign}(s)}} |w|^T (|s| - \lambda \mathbb{1}) - \mu F(\eta) \\
&= \iota_{\{|s| \leq \lambda \mathbb{1}\}}(s) - \inf_{\eta \in \{0,1\}^d} \mu F(\eta).
\end{aligned}$$

Hence $F^{**}_{\mu\lambda}(w) = \lambda \|w\|_1 + \inf_{\eta \in \{0,1\}^d} \lambda F(\eta)$. For the case $p \in (1, \infty)$.

$$\begin{aligned}
F^*_{\mu\lambda}(s) &= \sup_{w \in \mathbb{R}^d} w^T s - \mu F(\mathrm{supp}(w)) - \lambda \|w\|_p^p \\
&= \sup_{\eta \in \{0,1\}^d} \sup_{\substack{\mathbb{1}_{\mathrm{supp}(w)} = \eta \\ \mathrm{sign}(w) = \mathrm{sign}(s)}} |w|^T |s| - \lambda \|w\|_p^p - \mu F(\eta) \\
&= \sup_{\eta \in \{0,1\}^d} \frac{\lambda(p-1)}{(\lambda p)^q} \eta^T |s|^q - \mu F(\eta) \qquad (|s_i| = \lambda p |x_i^*|^{p-1}, \forall \eta_i \neq 0) \\
&= \sup_{\eta \in [0,1]^d} \frac{\lambda(p-1)}{(\lambda p)^q} \eta^T |s|^q - \mu f^-(\eta).
\end{aligned}$$

We denote $\hat{\lambda} = \frac{\lambda(p-1)}{(\lambda p)^q}$.

$$\begin{aligned}
F^{**}_{\mu\lambda}(w) &= \sup_{s \in \mathbb{R}^d} w^T s - F^*_{\mu\lambda}(s) \\
&= \sup_{s \in \mathbb{R}^d} \min_{\eta \in [0,1]^d} s^T w - \hat{\lambda} \eta^T |s|^q + \mu f^-(\eta) \\
&\overset{\star}{=} \inf_{\eta \in [0,1]^d} \sup_{\substack{s \in \mathbb{R}^p \\ \mathrm{sign}(s) = \mathrm{sign}(w)}} |s|^T |w| - \hat{\lambda} \eta^T |s|^q + \mu f^-(\eta) \\
&= \inf_{\eta \in [0,1]^d} \lambda (|w|^p)^T \eta^{1-p} + \mu f^-(\eta),
\end{aligned}$$

where the last equality holds since $|w_i| = \hat{\lambda} \eta_i q |s_i^*|^{q-1}, \forall \eta_i \neq 0$, otherwise $s_i^* = 0$ if $w_i = 0$ and ∞ otherwise. $(\star)$ holds by Sion's minimax theorem [32, Corollary 3.3]. Note then that the minimizer η^* (if it exists) satisfies $\mathrm{supp}(w) \subseteq \mathrm{supp}(\eta^*)$. Finally, note that if we take the limit as $p \to \infty$, we recover $\Theta_\infty = \inf_{s \in [0,1]^d} \{f^-(s) : s \geq |x|\}$. □

The variational form (15) given in lemma 5 follows from proposition 7 for the choice $\mu = \frac{1}{q}, \lambda = \frac{1}{p}$.

The following proposition derives the variational form (14) for $p = \infty$.

Proposition 8. *Given any set function $F : 2^V \to \mathbb{R} \cup \{+\infty\}$, and its corresponding convex closure f^-, Θ_∞ can be written $\forall w \in \mathrm{dom}(\Theta_\infty)$ as*

$$\begin{aligned}
\Theta_\infty(w) &= \max_{\kappa \in \mathbb{R}^d_+} \{\kappa^T |w| + \min_{S \subseteq V} F(S) - \kappa(S)\} \\
&= \max_{\kappa \in \mathbb{R}^d_+} \{\kappa^T |w| + \min_{S \subseteq \mathrm{supp}(w)} F(S) - \kappa(S)\} \qquad \text{(if } F \text{ is monotone)}
\end{aligned}$$

Similarly $\forall w \in \mathrm{dom}(f^-)$ we can write

$$\begin{aligned}
f^-(w) &= \max_{\kappa \in \mathbb{R}^d} \{\kappa^T |w| + \min_{S \subseteq V} F(S) - \kappa(S)\} \\
&= \Theta_\infty(w) = \max_{\kappa \in \mathbb{R}^d_+} \{\kappa^T w + \min_{S \subseteq \mathrm{supp}(x)} F(S) - \kappa(S)\} \qquad \text{(if } F \text{ is monotone)}
\end{aligned}$$

Proof. $\forall w \in \mathrm{dom}(\Theta_\infty)$, strong duality holds by Slater's condition, hence

$$\begin{aligned}
\Theta_\infty(w) &= \min_{\alpha}\{\sum_{S\subseteq V}\alpha_S F(S) : \sum_{S\subseteq V}\alpha_S \mathbb{1}_S \geq |w|, \sum_{S\subseteq V}\alpha_S = 1, \alpha_S \geq 0\}.\\
&= \min_{\alpha\geq 0}\max_{\rho\in\mathbb{R},\kappa\in\mathbb{R}^d_+}\{\sum_{S\subseteq V}\alpha_S F(S) + \kappa^T(|w| - \sum_{S\subseteq V}\alpha_S\mathbb{1}_S) + \rho(1 - \sum_{S\subseteq V}\alpha_S)\}.\\
&= \max_{\rho\in\mathbb{R},\kappa\in\mathbb{R}^d_+}\min_{\alpha\geq 0}\{\kappa^T|w| + \sum_{S\subseteq V}\alpha_S(F(S) - \kappa^T\mathbb{1}_S - \rho) + \rho\}.\\
&= \max_{\rho\in\mathbb{R},\kappa\in\mathbb{R}^d_+}\{\kappa^T|w| + \rho : F(S) \geq \kappa^T\mathbb{1}_S + \rho)\}.\\
&= \max_{\kappa\in\mathbb{R}^d_+}\{\kappa^T|w| + \min_{S\subseteq V} F(S) - \kappa(S)\}.
\end{aligned}$$

Let $J = \mathrm{supp}(|w|)$ then $\kappa^*_{J^c} = 0$. Then for monotone functions $F(S) - \kappa^*(S) \geq F(S\cap J) - \kappa^*(S)$, so we can restrict the minimum to $S \subseteq J$. The same proof holds for f^-, with the Lagrange multiplier $\kappa \in \mathbb{R}^d$ not constrained to be positive. □

The following Corollary derives the variational form (14) for $p \in [1, \infty]$.

Corollary 5. *Given any* monotone *set function* $F : 2^V \to \mathbb{R}\cup\{+\infty\}$*,* Θ_p *can be written* $\forall w \in \mathrm{dom}(\Theta_p)$ *as*

$$\Theta_p(w) = \max_{\kappa\in\mathbb{R}^d_+}\sum_{j=1}^d \psi_j(\kappa_j, x_j) + \min_{S\subseteq V} F(S) - \kappa(S).$$

where

$$\psi_j(\kappa_j, w_j) := \begin{cases}\kappa_j^{1/q}|w_j| & \textit{if } |w_j| \leq \kappa_j^{1/p}\\ \frac{1}{p}|w_j|^p + \frac{1}{q}\kappa_j & \textit{otherwise}\end{cases}$$

Proof. By Propositions 7 and 8, we have $\forall w \in \mathrm{dom}(\Theta_p)$, i.e., $\exists\eta \in [0,1]^d$, s.t $\mathrm{supp}(w) \subseteq \mathrm{supp}(\eta), \eta \in \mathrm{dom}(\Theta_\infty)$,

$$\begin{aligned}
\Theta_p(w) &= \inf_{\eta\in[0,1]^d}\frac{1}{p}\sum_{j=1}^d\frac{|w_j|^p}{\eta_j^{p-1}} + \frac{1}{q}\Theta_\infty(\eta)\\
&= \inf_{\eta\in[0,1]^d}\frac{1}{p}\sum_{j=1}^d\frac{|w_j|^p}{\eta_j^{p-1}} + \frac{1}{q}\max_{\rho\in\mathbb{R},\kappa\in\mathbb{R}^d_+}\{\kappa^T\eta + \rho : F(S) \geq \kappa^T\mathbb{1}_S + \rho\}.\\
&\overset{\star}{=} \max_{\rho\in\mathbb{R},\kappa\in\mathbb{R}^d_+}\inf_{\eta\in[0,1]^d}\{\frac{1}{p}\sum_{j=1}^d\frac{|w_j|^p}{\eta_j^{p-1}} + \frac{1}{q}\kappa^T\eta + \rho : F(S) \geq \kappa^T\mathbb{1}_S + \rho\}.
\end{aligned}$$

$(\star)$ holds by Sion's minimax theorem [32, Corollary 3.3]. Note also that for $\kappa_i \geq 0$,

$$\inf_{\eta_j\in[0,1]}\frac{1}{p}\frac{|w_j|^p}{\eta_j^{p-1}} + \frac{1}{q}\kappa_j\eta_j = \begin{cases}\kappa_j^{1/q}|w_j| & \text{if } |w_j| \leq \kappa_j^{1/p}\\ \frac{1}{p}|w_j|^p + \frac{1}{q}\kappa_j & \text{otherwise}\end{cases} := \psi_j(\kappa_j, w_j)$$

where the infimum is zero if $w_j = 0$. Otherwise, the minimum is achieved at $\eta_j^* = \min\{\frac{|w_j|}{\kappa_j^{1/p}}, 1\}$ (if $\kappa_j = 0, \eta_j^* = 1$). Hence,

$$\Theta_p(w) = \max_{\kappa\in\mathbb{R}^d_+}\sum_{j=1}^d\psi_j(\kappa_j, w_j) + \min_{S\subseteq V} F(S) - \kappa(S).$$

□

7.2 Necessary conditions for support recovery (Proof of Proposition 2)

Before proving Proposition 2, we need the following technical Lemma.
Lemma 6. *Given $J \subset V$ and a vector w s.t $\mathrm{supp}(w) \subseteq J$, if Φ is not decomposable at w w.r.t J, then $\exists i \in J^c$ such that the i-th component of all subgradients at w is zero; $0 = [\partial\Phi(w)]_i$.*

Proof. If Φ is not decomposable at w and $0 \neq [\partial\Phi(w)]_i, \forall i \in J^c$, then $\forall M_J > 0, \exists \Delta \neq 0, \mathrm{supp}(\Delta) \subseteq J^c$ s.t., $\Phi(w + \Delta) < \Phi(w) + M_J\|\Delta\|_\infty$. In particular, we can choose $M_J = \inf_{i \in J^c, v \in \partial\Phi(w_J), v_i \neq 0} |v_i| > 0$, if the inequality holds for some $\Delta \neq 0$, then let $i_{\max}$ denote the index where $|\Delta_{i_{\max}}| = \|\Delta\|_\infty$. Then given any $v \in \partial\Phi(w)$, we have

$$\begin{aligned}\Phi(w + \|\Delta\|_\infty \mathbb{1}_{i_{\max}}) \leq \Phi(w + \Delta) &< \Phi(w) + M_J\|\Delta\|_\infty \\ &\leq \Phi(w) + \langle v, \|\Delta\|_\infty \mathbb{1}_{i_{\max}} \mathrm{sign}(v_{i_{\max}})\rangle \\ &\leq \Phi(w + \|\Delta\|_\infty \mathbb{1}_{i_{\max}})\end{aligned}$$

which leads to a contradiction. □

Proposition 2. *The minimizer $\hat{w}$ of $\min_{w \in \mathbb{R}^d} L(w) - z^\top w + \lambda\Phi(w)$, where L is a strongly-convex and smooth loss function and $z \in \mathbb{R}^d$ has a continuous density, has a weakly stable support w.r.t. Φ, with probability one.*

Proof. Given any weakly unstable J, we show that the set of z such that $\mathrm{supp}(\hat{w}) = J$ has measure zero. By optimality conditions $z - \nabla L(\hat{w}) \in \partial\Phi(\hat{w})$. Hence, given z, z' and the corresponding solutions $\hat{w}, \hat{w}'$ such that $\mathrm{supp}(\hat{w}) \subseteq J, \mathrm{supp}(\hat{w}') \subseteq J$, denote by $\mu > 0$ the strong convexity constant of L. We have by convexity of Φ:

$$\begin{aligned}\big((z - \nabla L(\hat{w}) - (z' - \nabla L(\hat{w}'))\big)^\top (\hat{w} - \hat{w}') \geq 0& \\ (z - z')^\top (\hat{w} - \hat{w}'_J) &\geq (\nabla L(\hat{w}) - \nabla L(\hat{w}'))^\top (\hat{w} - \hat{w}') \\ (z - z')^\top (\hat{w} - \hat{w}') &\geq \mu\|\hat{w} - \hat{w}'\|_2^2 \\ \frac{1}{\mu}\|z - z'\|_2 &\geq \|\hat{w} - \hat{w}'\|_2\end{aligned}$$

Thus $\hat{w}$ is a deterministic Lipschitz-continuous function of z. If Φ is not decomposable at $\hat{w}$ with respect to J, we know by lemma 6 that there exists an $i \in J^c$ such that $0 = [\partial\Phi(\hat{w})]_i$, this implies that $z_i - \nabla L(\hat{w})_i = 0$ and thus z_i is a Lipschitz-continuous function of z, which can only happen with zero measure. □

7.3 Sufficient conditions for support recovery (Proof of Lemma 3 and Theorem 1)

Lemma 3. *Let Φ be any monotone convex function, then for all $w^0 \in \mathbb{R}^d$, $\Phi(|w|^\alpha)$ admits the following majorizer $\Phi(|w|^\alpha) \leq (1 - \alpha)\Phi(|w^0|^\alpha) + \alpha\Phi(|w^0|^{\alpha-1} \circ |w|)$, which is tight at w^0.*

Proof. The function $w \rightarrow w^\alpha$ is concave on $\mathbb{R}_+ \setminus \{0\}$, hence

$$\begin{aligned}|w_j|^\alpha &\leq |w_j^0|^\alpha + \alpha|w_j^0|^{\alpha-1}(|w_j| - |w_j|^0) \\ |w_j|^\alpha &\leq (1 - \alpha)|w_j^0|^\alpha + \alpha|w_j^0|^{\alpha-1}|w_j| \\ \Phi(|w|^\alpha) &\leq \Phi((1 - \alpha)|w^0|^\alpha + \alpha|w^0|^{\alpha-1} \circ |w_j|) && \text{(by monotonicity)} \\ \Phi(|w|^\alpha) &\leq (1 - \alpha)\Phi(|w^0|^\alpha) + \alpha\Phi(|w^0|^{\alpha-1} \circ |w|) && \text{(by convexity)}\end{aligned}$$

If $w_j = 0$ for any j, the upper bound goes to infinity and hence it still holds. □

Before proceeding to the proof of theorem 1, we will need the following lemma.
Lemma 7. *Given any normalized ($\Phi(0) = 0$) convex function, we have for all $w \in \mathbb{R}^d$:*

$$t\Phi(w) \geq \Phi(tw), \forall t \leq 1$$
$$t\Phi(w) \leq \Phi(tw), \forall t \geq 1.$$

Proof. The first inequality holds since $\forall t \leq 1, \Phi(tw + (1-t)0) \leq t\Phi(w) + (1-t)\Phi(0) = t\Phi(w)$. The second inequality follows directly from the first one, since $\forall t \geq 1, \Phi(w) = \Phi(\frac{1}{t}(tw)) \leq \frac{1}{t}\Phi(tw)$. □

Theorem 1. *[Consistency and Support Recovery] Let $\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a convex, normalized ($\Phi(0) = 0$) and monotone function and denote by J the true support $J = \operatorname{supp}(w^*)$. If J is strongly stable with respect to Φ and λ_n satisfies $\frac{\lambda_n}{\sqrt{n}} \to 0$, $\frac{\lambda_n}{n^{\alpha/2}} \to \infty$, then the estimator* (6) *is consistent and asymptotically normal, i.e., it satisfies asymptotic normality, i.e.:*

$$\sqrt{n}(\hat{w}_J - w^*_J) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q_{JJ}^{-1}), \tag{7}$$

and

$$P(\operatorname{supp}(\hat{w}) = J) \to 1. \tag{8}$$

Proof. We will follow the proof in [37]. We write $\hat{w} = w^* + \frac{\hat{u}}{\sqrt{n}}$ and $\Phi_n(u) = \frac{1}{2n}\|y - X(w^* + \frac{u}{\sqrt{n}})\|_2^2 + \lambda_n \Phi(c \circ |w^* + \frac{u}{\sqrt{n}}|)$, where $c = |w^0|^{\alpha - 1}$. Then $\hat{u} = \arg\min_{u \in \mathbb{R}^d} \Phi_n(u)$. Let $V_n(u) = \Phi_n(u) - \Phi_n(0)$, then

$$V_n(u) = \frac{1}{2}u^T Q u - \epsilon^T \frac{Xu}{\sqrt{n}} + \lambda_n \big(\Phi(c \circ |w^* + \frac{u}{\sqrt{n}}|) - \Phi(c \circ |w^*|)\big)$$

Since w^0 is a $\sqrt{n}$-consistent estimator to w^*, then $\sqrt{n} w^0_{J^c} = O_p(1)$ and $n^{\frac{1-\alpha}{2}} c_{J^c}^{-1} = O_p(1)$. Since $\frac{\lambda_n}{n^{\alpha/2}} \to \infty$, by stability of J, we have

$$\begin{aligned}
\lambda_n \big(\Phi(c \circ |w^* + \frac{u}{\sqrt{n}}|) - \Phi(c \circ |w^*|)\big) &= \lambda_n \big(\Phi(c_J \circ |w^*_J + \frac{u_J}{\sqrt{n}}| + c_{J^c} \circ \frac{|u_{J^c}|}{\sqrt{n}}) - \Phi(c_J \circ |w^*_J|)\big) \\
&\geq \lambda_n \big(\Phi(c_J \circ |w^*_J + \frac{u_J}{\sqrt{n}}|) + M_J \|c_{J^c} \circ \frac{|u_{J^c}|}{\sqrt{n}}\|_\infty - \Phi(c_J \circ |w^*_J|)\big) \\
&= \lambda_n \big(\Phi(c_J \circ |w^*_J + \frac{u_J}{\sqrt{n}}|) - \Phi(c_J \circ |w^*_J|)\big) + M_J \|\lambda_n n^{-\alpha/2} n^{\frac{\alpha-1}{2}} c_{J^c} \circ |u_{J^c}|\|_\infty \\
&\begin{cases} \xrightarrow{p} \infty & \text{if } u_{J^c} \neq 0 \\ = \lambda_n \big(\Phi(c_J \circ |w^*_J + \frac{u_J}{\sqrt{n}}|) - \Phi(c_J \circ |w^*_J|)\big) & \text{otherwise.} \end{cases}
\end{aligned} \tag{16}$$

Since Φ is convex and normalized, then by lemma 7, it follows that:

$$\lambda_n \big(\Phi(c_J \circ |w^*_J + \frac{u_J}{\sqrt{n}}|) - \Phi(c_J \circ |w^*_J|)\big) \leq \frac{\lambda_n}{2} \big(\Phi(2c_J \circ |w^*_J|) + \Phi(2c_J \circ \frac{|u_J|}{\sqrt{n}})\big) - \lambda_n \Phi(c_J \circ |w^*_J|) \tag{17}$$

$$\leq \Phi(c_J \circ \frac{\lambda_n |u_J|}{\sqrt{n}}) \qquad (\text{if } \lambda_n \geq 2)$$

Since w^0 is a $\sqrt{n}$-consistent estimator to w^*, then $c_J = |w^0_J|^{\alpha-1} \xrightarrow{p} |w^*_J|^{\alpha-1}$. Since $\frac{\lambda_n}{\sqrt{n}} \to 0$, we have by Slutsky's theorem, $\Phi_J(c_J \circ \frac{\lambda_n |u_J|}{\sqrt{n}}) \xrightarrow{p} 0$. Hence by (17),

$$\lambda_n(\Phi_J(c_J \circ |w^*_J + \frac{u_J}{\sqrt{n}}|) - \Phi_J(c_J \circ |w^*_J|)) \xrightarrow{p} 0. \tag{18}$$

Hence by (16) and (18),

$$\lambda_n\big(\Phi(c \circ |w^* + \frac{u}{\sqrt{n}}|) - \Phi(c \circ |w^*|)\big) \xrightarrow{p} \begin{cases} 0 & \text{if } u_{J^c} = 0 \\ \infty & \text{Otherwise} \end{cases}. \tag{19}$$

By CLT, $\frac{X}{\sqrt{n}}\epsilon \xrightarrow{d} W \sim \mathcal{N}(0, \sigma^2 Q)$, it follows then that $V_n(u) \xrightarrow{d} V(u)$, where

$$V(u) = \begin{cases} \frac{1}{2}u_J^T Q_{JJ} u_J - W_J^T u_J & \text{if } u_{J^c} = 0 \\ \infty & \text{Otherwise} \end{cases}.$$

V_n is convex and the unique minimum of V is $u_J = Q_{JJ}^{-1} W_J, u_{J^c} = 0$, hence by epi-convergence results [c.f., [37]]

$$\hat{u}_J \xrightarrow{d} Q_{JJ}^{-1} W_J \sim \mathcal{N}(0, \sigma^2 Q_{JJ}^{-1}), \quad \hat{u}_{J^c} \xrightarrow{d} 0. \tag{20}$$

Since $\hat{u} = \sqrt{n}(\hat{w} - w^*)$, then it follows from (21) that

$$\hat{w}_J \xrightarrow{p} w_J^*, \quad \hat{w}_{J^c} \xrightarrow{p} 0 \tag{21}$$

Hence, $P(\text{supp}(\hat{w}) \supseteq J) \to 1$ and it is sufficient to show that $P(\text{supp}(\hat{w}) \subseteq J) \to 1$ to complete the proof.

For that denote $\hat{J} = \text{supp}(\hat{w})$ and let's consider the event $\hat{J} \setminus J \neq \emptyset$. By optimality conditions, we know that

$$-X_{\hat{J}\setminus J}^T(X\hat{w} - y) \in \lambda_n[\partial\Phi(c \circ \cdot)(\hat{w})]_{\hat{J}\setminus J}$$

Note, that $-\frac{X_{\hat{J}\setminus J}^T(X\hat{w}-y)}{\sqrt{n}} = \frac{X_{\hat{J}\setminus J}^T X(\hat{w}-w^*)}{\sqrt{n}} - \frac{X_{\hat{J}\setminus J}^T \epsilon}{\sqrt{n}}$. By CLT, $\frac{X_{\hat{J}\setminus J}^T \epsilon}{\sqrt{n}} \xrightarrow{d} W \sim \mathcal{N}(0, \sigma^2 Q_{\hat{J}\setminus J, \hat{J}\setminus J})$ and by (21) $\hat{w} - w^* \xrightarrow{p} 0$ then $-\frac{X_{\hat{J}\setminus J}^T(X\hat{w}-y)}{\sqrt{n}} = O_p(1)$.

On the other hand, $\frac{\lambda_n c_{\hat{J}\setminus J}}{\sqrt{n}} = \lambda_n n^{\frac{1-\alpha}{2}} n^{\frac{\alpha-1}{2}} c_{\hat{J}\setminus J} \to \infty$, hence $\frac{\lambda_n c_{\hat{J}\setminus J}}{\sqrt{n}} c_{\hat{J}\setminus J}^{-1} v_{\hat{J}\setminus J} \to \infty, \forall v \in \partial\Phi(c \circ \cdot)(\hat{w})$, since $c_{\hat{J}\setminus J}^{-1} v_{\hat{J}\setminus J} = O_p(1)^{-1}$. To see this, let $w'_J = \hat{w}_J$ and 0 elsewhere. Note that by definition of the subdifferential and the stability assumption on J, there must exists $M_J > 0$ s.t

$$\begin{aligned} &\Phi(c \circ w') \geq \Phi(c \circ \hat{w}) + \langle v_{\hat{J}\setminus J}, -\hat{w}_{\hat{J}\setminus J} \rangle \\ &\Phi(c \circ w') \geq \Phi(c \circ w') + M_J \|c_{\hat{J}\setminus J} \circ \hat{w}_{\hat{J}\setminus J}\|_\infty - \|c_{\hat{J}\setminus J}^{-1} \circ v_{\hat{J}\setminus J}\|_1 \|c_{\hat{J}\setminus J} \circ \hat{w}_{\hat{J}\setminus J}\|_\infty \\ \|c_{\hat{J}\setminus J}^{-1} \circ v_{\hat{J}\setminus J}\|_1 \geq M_J \end{aligned}$$

We deduce then $P(\text{supp}(\hat{w}) \subseteq J) = 1 - P(\hat{J} \setminus J \neq \emptyset) = 1 - P(\text{optimality condition holds}) \to 1$. □

7.4 Discrete stability (Proof of Proposition 3 and relation to weak submodularity)

Proposition 3. *If F is a monotone function, F is ρ-submodular iff weak stability is equivalent to strong stability.*

Proof. If F is ρ-submodular and J is weakly stable, then $\forall A \subseteq J, \forall i \in J^c, 0 < \rho[F(J \cup \{i\}) - F(J)] \leq F(J \cup \{i\}) - F(J)$, i.e., J is strongly stable w.r.t. F. If F is such that any weakly stable set is also strongly stable, then if F is not ρ-submodular, then $\forall \rho \in (0,1]$ there must exists a set $B \subseteq V$, s.t., $\exists A \subseteq B, i \in B^c$, s.t., $\rho[F(B \cup \{i\}) - F(B)] > F(A \cup \{i\}) - F(A) \geq 0$. Hence, $F(B \cup \{i\}) - F(B) > 0$, i.e., B is weakly stable and thus it is also strongly stable and we must have $F(A \cup \{i\}) - F(A) > 0$. Choosing then in particular, $\rho = \min_{B \subseteq V} \min_{A \subseteq B, i \in B^c} \frac{F(A \cup \{i\}) - F(A)}{F(B \cup \{i\}) - F(B)} \in (0,1]$, leads to a contradiction; $\min_{A \subseteq B, i \in B^c} F(A \cup \{i\}) - F(A) \geq \rho[F(B \cup \{i\}) - F(B)] > F(A \cup \{i\}) - F(A)$. □

We show that ρ-submodularity is a stronger condition than weak submodularity. First, we recall the definition of weak submodular functions.

Definition 7 (Weak Submodularity (c.f., [7, 11])). *A function F is weakly submodular if $\forall S, L, S \cap L = \emptyset, F(L \cup S) - F(L) > 0$,*

$$\gamma_{S,L} = \frac{\sum_{i \in S} F(L \cup \{i\}) - F(L)}{F(L \cup S) - F(L)} > 0$$

Proposition 9. *If F is ρ-submodular then F is weakly submodular. But the converse is not true.*

Proof. If F is ρ-submodular then $\forall S, L, S \cap L = \emptyset, F(L \cup S) - F(L) > 0$, let $S = \{i_1, i_2, \cdots, i_r\}$

$$\begin{aligned} F(L \cup S) - F(L) &= \sum_{k=1}^{r} F(L \cup \{i_1, \cdots, i_k\}) - F(L \cup \{i_1, \cdots, i_{k-1}\}) \\ &\leq \sum_{k=1}^{r} \frac{1}{\rho}(F(L \cup \{i_k\}) - F(L)) \\ \Rightarrow \gamma_{S,T} &= \rho > 0. \end{aligned}$$

We show that the converse is not true by giving a counter-example. Consider the set cover function, with the ground set $V = \{1,2,3\}$ and the groups $G_1 = \{1,2\}, G_2 = \{2,3\}$, then it's easy to see that $\gamma_{S,L} > 0, \forall S, L, S \cap L = \emptyset$, but F is not ρ-submodular for any $\rho \in (0,1]$ since $0 = F(\{2,3\}) - F(\{2\}) < \rho(F(\{1,2,3\}) - F(\{2,3\}))$. □

7.5 Relation between discrete and continuous stability (Proof of Propositions 4 and 5, and Corollary 3)

First, we present a useful simple lemma, which provides an equivalent definition of decomposability for monotone function.

Lemma 8. *Given $w \in \mathbb{R}^d$, $J \subseteq J$, $\operatorname{supp}(w) = J$, if Φ is a monotone function, then Φ is decomposable at w w.r.t J iff $\exists M_J > 0, \forall \delta > 0, i \in J^c$, s.t,*

$$\Phi(w + \delta \mathbb{1}_i) \geq \Phi(w) + M_J \delta.$$

Proof. By definition 2, $\exists M_J > 0, \forall \Delta \in \mathbb{R}^d, \operatorname{supp}(\Delta) \subseteq J^c$,

$$\Phi(w + \Delta) \geq \Phi(w) + M_J \|\Delta\|_\infty.$$

in particular this must hold for $\Delta = \delta \mathbb{1}_i$. On the other hand, if the inequality hold for all $\delta \mathbb{1}_i$, then given any Δ s.t. $\operatorname{supp}(\Delta) \subseteq J^c$ let $i_{\max}$ be the index where $\Delta_{i_{\max}} = \|\Delta\|_\infty$ and let $\delta = \|\Delta\|_\infty$, then

$$\Phi(w + \Delta) \geq \Phi(w + \delta_{i_{\max}}) \geq \Phi(w) + M_J \delta = \Phi(w) + M_J \|\Delta\|_\infty.$$

□

Proposition 4. *Given any monotone set function F, all sets $J \subseteq V$ strongly stable w.r.t to F are also strongly stable w.r.t Ω_p and Θ_p.*

Proof. We make use of the variational form (11). Given a set J stable w.r.t to F and $\operatorname{supp}(w) \subseteq J$, let $\kappa^* \in \arg\max_{\kappa \in \mathbb{R}^d_+} \{\sum_{i \in J} \kappa_i^{1/q} |w_i| : \kappa(A) \leq F(A), \forall A \subseteq V\}$, then $\Omega(w) = |w_J|^T (\kappa^*_J)^{1/q}$. Note that $\forall A \subseteq J, F(A \cup i) > F(A)$, by definition 3. Hence, $\forall i \in J^c$, we can define $\kappa' \in \mathbb{R}^d_+$ s.t., $\kappa'_J = \kappa^*_J$, $\kappa'_{(J \cup i)^c} = 0$ and $\kappa'_i = \min_{A \subseteq J} F(A \cup i) - F(A) > 0$. Note that κ' is feasible, since $\forall A \subseteq J, \kappa'(A) = \kappa^*(A) \leq F(A)$ and $\kappa'(A + i) = \kappa^*(A) + \kappa'_i \leq F(A) + F(A \cup i) - F(A) = F(A \cup i)$. For any other set $\kappa'(A) = \kappa'(A \cap (J + i)) \leq F(A \cap (J + i)) \leq F(A)$, by monotonicity. It follows then that $\Omega(w + \delta \mathbb{1}_i) = \max_{\kappa \in \mathbb{R}^d_+} \{\sum_{i \in J \cup i}^d \kappa_i^{1/q} |w_i| : \kappa(A) \leq F(A), \forall A \subseteq V\} \geq |w_J|^T (\kappa^*_J)^{1/q} + \delta (\kappa'_i)^{1/q} \geq \Omega(w) + \delta M$, with $M = (\kappa'_i)^{1/q} > 0$. The proposition then follows by lemma 8.

Similarly, the proof for Θ_p follows in a similar fashion. We make use of the variational form (14). Given a set J stable w.r.t to F and $\operatorname{supp}(w) \subseteq J$, first note that this implicity implies that $F(J) < +\infty$ and hence $\Theta_p(w) < +\infty$. Let $\kappa^* \in \arg\max_{\kappa \in \mathbb{R}^d_+} \sum_{j=1}^d \psi_j(\kappa_j, w_j) + \min_{S \subseteq V} F(S) - \kappa(S)$ and $S^* \in \arg\min_{S \subseteq J} F(S) - \kappa^*(S)$. Note that $\forall S \subseteq J, \forall i \in J^c, F(S \cup i) > F(S)$, by definition 3. Hence, $\forall i \in J^c$, we can define $\kappa' \in \mathbb{R}^d_+$ s.t., $\kappa'_J = \kappa^*_J$, $\kappa'_{(J \cup i)^c} = 0$ and $\kappa'_i = \min_{S \subseteq J} F(S \cup i) - F(S) > 0$. Note that $\forall S \subseteq J, F(S) - \kappa'(S) = F(S) - \kappa^*(S) \geq F(S^*) - \kappa^*(S^*)$ and $F(S + i) - \kappa'(S + i) = F(S + i) - \kappa^*(S) - \kappa'_i \geq F(S + i) - \kappa^*(S) - F(S + i) + F(S) \geq F(S^*) - \kappa^*(S^*)$. Note also that $\psi_i(\kappa'_i, \delta) = (\kappa'_i)^{1/q} \delta$ if $\delta \leq (\kappa'_i)^{1/p}$, and $\psi_i(\kappa'_i, \delta) = \frac{1}{p} \delta^p + \frac{1}{q} \kappa'_i = \delta(\frac{1}{p} \delta^{p-1} + \frac{1}{q} \kappa'_i \delta^{-1}) \geq \delta (\kappa'_i)^{1/q}$ otherwise. It follows then that $\Theta_p(w + \delta \mathbb{1}_i) \geq \sum_{j \in J} \psi_j(\kappa_j, w_j) + (\kappa'_i)^{1/q} \delta + \min_{S \subseteq J \cup i} F(S) - \kappa'(S) \geq \sum_{j \in J} \psi_j(\kappa_j, w_j) + (\kappa'_i)^{1/q} \delta + \min_{S \subseteq J} F(S) - \kappa^*(S) = \Theta_p(w) + \delta M$ with $M = (\kappa'_i)^{1/q} > 0$. The proposition then follows by lemma 8. □

Proposition 5. *If $F = F_-$ and J is strongly stable w.r.t Ω_∞, then J is strongly stable w.r.t F. Similarly, for any monotone F, if J is strongly stable w.r.t Θ_∞, then J is strongly stable w.r.t F.*

Proof. $F(A + i) = \Omega_\infty(\mathbb{1}_A + \mathbb{1}_i) = \Theta_\infty(\mathbb{1}_A + \mathbb{1}_i) > \Omega_\infty(\mathbb{1}_A) = \Theta_\infty(\mathbb{1}_A) = F(A), \forall A \subseteq J.$ □

Corollary 3. *If F is a monotone submodular function and J is weakly stable w.r.t $\Omega_\infty = \Theta_\infty$ then J is weakly stable w.r.t F.*

Proof. If F is a monotone submodular function, then $\Omega_\infty(w) = \Theta_\infty(w) = f_L(|w|)$. If J is not weakly stable w.r.t F, then $\exists i \in J^c$ s.t., $F(J \cup \{i\}) = F(J)$. Thus, given any $w, \operatorname{supp}(w) = J$, choosing $0 < \delta < \min_{i \in J} |w_i|$, result in $f_L(|w| + \delta \mathbb{1}_i) = f_L(|w|)$, which contradicts the weak stability of J w.r.t to $\Omega_\infty = \Theta_\infty$. □