Slack and Margin Rescaling as Convex Extensions of Supermodular Functions

Matthew B. Blaschko
Center for Processing Speech & Images
Departement Elektrotechniek, KU Leuven
Kasteelpark Arenberg 10
3001 Leuven, Belgium
matthew.blaschko@esat.kuleuven.be

Abstract

Slack and margin rescaling are variants of the structured output SVM. They define convex surrogates to task specific loss functions, which, when specialized to non-additive loss functions for multi-label problems, yield extensions to increasing set functions. We demonstrate in this paper that we may use these concepts to define polynomial time convex extensions of arbitrary supermodular functions. We further show that slack and margin rescaling can be interpreted as dominating convex extensions over multiplicative and additive families, and that margin rescaling is strictly dominated by slack rescaling. However, we also demonstrate that, while the function value and gradient for margin rescaling can be computed in polynomial time, the same for slack rescaling corresponds to a non-supermodular maximization problem.

1 Introduction

It is known that submodular minimization is strongly polynomial time solvable while non-submodular minimization is NP-hard even to find an $O(\sqrt{n})$ approximation. However, NP-hardness is a worst case analysis over a set of problems. We therefore take a more refined view and seek to characterize a class of piecewise-linear convex extensions of set functions that minimizes (within a certain class) the probability of a non-integral solution to the corresponding LP relaxation when problem instances are sampled from a given distribution. We do so through the notion of a dominating convex extension, i.e. one that is closer to the convex closure at all points within the unit cube.

In this work, we explore analogous strategies analogous to margin and slack rescaling in structured output support vector machines [8] for set function minimization in general, and substantially advance the theory of supermodular set function minimization. In particular, we advance the theory of convex extensions of supermodular extensions by showing an explicit form for the closest convex extension to the convex closure within a specific multiplicative family (Proposition 6). We further show that this multiplicative family dominates a related polynomial time computable additive family (i.e. that the closed form extension shown previously is closer to the convex closure than any member of this additive family) (Proposition 11). Subsequently, we prove that computation of an extension operator derived from a modular upper bound on submodular functions introduced by Jegelka & Bilmes does not correspond with a supermodular maximization (Proposition 15). This in general suggests a tradeoff between the tightness of the extension and the tractability of its computation, with only margin rescaling demonstrated to be polynomial time computable despite being dominated by slack rescaling. A summary of key theoretical results is given in Table 1. We may additionally summarize key theoretical contributions through a partial ordering over (convex) extensions of set functions.
Table 1: Key theoretical results in this paper. \( S \) denotes the extension obtained by the analogue of slack rescaling (Definition 9), \( M \) denotes the analogue of margin rescaling (Definition 10), \( G \) is the set of all supermodular functions.

| Result | Location |
|--------|----------|
| \( S \geq_g H, \forall \ell(x) = \max_{A \subseteq V} \ell(A) \cdot h(x, A) \) | Proposition 5 |
| \( S \geq_g M \geq_g H, \forall \ell(x) = \max_{A \subseteq V} \ell(A) + h(x, A) \) | Proposition 6 |

functions considered in this paper using the notion of restricted operator inequality (Definition 5):

\[
\begin{align*}
S_1 & \leq \leq g, S & \leq g & C \leq V \\
M & \leq g S
\end{align*}
\]

in which \( V \) denotes the (non-convex) multi-linear extension due to Vondrák [9] (Appendix B).

Remark 1. If a fixed finite set of poly-time convex extensions does not have a total ordering, the maximum of these extensions is a poly-time convex extension that dominates them all.

1.1 Related Work

Supermodular minimization (submodular maximization) has been widely studied both in the context of relaxations as well as for fully combinatorial approximation algorithms [5]. Notable contributions include the multilinear extension [9, 10] (mentioned above and in Appendix B), which, although non-convex, has been used in an approximate minimization framework.

A large amount of work has been done on relaxations approximate energy minimization in the context of random field models with low maximal clique size [6], including roof duality [4], or by making additional assumptions on the structure of the graph such as balancedness [11].

The connection between the structured output SVM [8] and convex extensions of set functions seems to be due to [12]. The comparative difficulty of optimizing slack rescaling vs. margin rescaling is known, and an interesting approach to the optimization of slack rescaling via multiple margin rescaling operations can be found in [1].

2 Mathematical Preliminaries

Definition 1 (Set function). A set function \( \ell \) is a mapping from the power set of a base set \( V \) to the reals:

\[
\ell: \mathcal{P}(V) \mapsto \mathbb{R}.
\]

We will assume w.l.o.g. that \( \ell(\emptyset) = 0 \) for all set functions in the sequel.

Definition 2 (Extension). An extension of a set function is an operator that yields a continuous function over the \( p \)-dimensional unit cube (where \( p = |V| \)) such that the function value on each vertex \( x \) of the unit cube is equal to \( \ell(A) \) for \( A \) such that \( x_i = 1 \iff i \in A \).

Definition 3 (Submodular set function). A set function \( f \) is said to be submodular if for all \( A \subseteq B \subseteq V \) and \( x \in V \setminus B \),

\[
f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B).
\]

A set function is said to be supermodular if its negative is submodular, and a function is said to be modular if it is both submodular and supermodular. We denote the set of all submodular functions \( S \), the set of all supermodular functions \( G \) := \{\( g \mid g \in \mathbb{R}^V \) \}, and \( G_+ := \{g \mid g(S) \geq 0, \forall S \subseteq V \} \) \( \cap \mathbb{R}_+ \) denotes the set of all non-negative supermodular functions. Additional properties of submodular functions can be found in [2].

Lemma 1. Any \( g \in G \) such that \( g(\{x\}) \geq 0, \forall x \in V \) is an increasing function, i.e. \( g(A \cup \{x\}) \geq g(A), \forall A \subseteq V, x \in V \setminus A \).
Proof. We have that \( g(\emptyset) = 0 \) and \( g(\{x\}) \geq 0, \forall x \in V \) by assumption. From the supermodularity of \( g, \forall A \subset V, x \in V \setminus A, \)

\[
g(A) + g(\{x\}) \leq g(A \cup \{x\}) + g(A \cap \{x\}) \Rightarrow g(A) \leq g(A \cup \{x\}) \quad \text{(5)}
\]

\[
\Rightarrow g(A) \leq g(A \cup \{x\}) \quad \text{(6)}
\]

Corollary 1. Any \( g \in \mathcal{G}_+ \) is an increasing function.

Definition 4 (Convex closure). The convex closure of a set function \( C\ell \) is defined for all \( x \in [0, 1]^{\vert V \vert} \) as

\[
C\ell(x) := \min_{\alpha \in \mathbb{R}^{\vert V \vert}_+} \sum_{A \subseteq V} \alpha_A \ell(A), \text{ s.t. } \sum_{A \subseteq V} \alpha_A 1_A = x, \sum_{A \subseteq V} \alpha_A = 1, \alpha_A \geq 0 \quad \forall A \subseteq V. \quad \text{(7)}
\]

In general, the convex closure is not polynomial time solvable, with the notable exception of the Lovász extension [2, 7].

Definition 5 (Restricted operator inequality). For two convex extensions \( A \) and \( B \), we write \( A \prec B \) (analogously \( A \preceq B \)) if \( A \ell(x) < B \ell(x) \) \( \forall x \in [0, 1]^p \). We may subscript the inequality sign with a class of set functions (e.g. \( B \preceq_S L \)) if the inequality holds for all set functions in that class.

Proposition 1 (Transitivity of restricted operator inequality). For two function sets \( A \) and \( B \) and three operators \( F, G, \) and \( H \),

\[
F \preceq_A G \preceq_B H \implies F \preceq_{A \cup B} H. \quad \text{(8)}
\]

Proposition 2. \( B \preceq C \) for all convex extensions \( B \).

Proof. Assume that there exists some convex extension \( B \ell \) such that \( B \ell(x) > C\ell(x) \) for some \( x \) in the \( p \)-dimensional unit cube. If we map all \( A \subseteq V \) to vertices of the \( p \)-dimensional unit cube, and identify \( \ell \) with \( 2^p \) distinct points in \( \{0, 1\}^p \times \mathbb{R} \), then the convex closure is the lower hull of these points taken with respect to the dimension corresponding to the function value. As each of the \( 2^p \) points corresponding to the function are vertices of this lower hull, any \( B \ell \) that has a point greater than a corresponding point of \( C\ell \) is not a lower hull and cannot contain all of these vertices, which contradicts the definition of a set function extension.

One of the key applications of set function extensions is their application to (approximate) minimization of set functions through LP relaxations. In Proposition 3 we formalize the unsurprising but important result that dominating convex extensions result in integral solutions to LP relaxations with higher probability. For some set \( S \subseteq \mathbb{R}^d \), we denote \( \text{int}(S) := [S \cap \{0, 1\}^d \neq \emptyset] \), where we have used Iverson bracket notation on the rhs of the definition.

Proposition 3. For two operators that yield convex extensions for all set functions \( \ell \in \mathcal{A} \), \( F \preceq_A G \preceq_A C \), and for all distributions \( p : \mathcal{A} \rightarrow \mathbb{R}_+ \),

\[
\mathbb{E}_{\ell \sim p} \left\{ \text{int} \left( \arg \min_{x \in [0, 1]^{\vert V \vert}} F \ell(x) \right) \right\} \leq \mathbb{E}_{\ell \sim p} \left\{ \text{int} \left( \arg \min_{x \in [0, 1]^{\vert V \vert}} G \ell(x) \right) \right\}, \quad \text{(9)}
\]

where \( \text{arg min} \) is defined as a map to the set of minimizers of an expression.

Proof. It is sufficient to demonstrate that

\[
\left( \arg \min_{x \in [0, 1]^{\vert V \vert}} F \ell(x) \right) \cap \{0, 1\}^{\vert V \vert} = \emptyset \iff \left( \arg \min_{x \in [0, 1]^{\vert V \vert}} G \ell(x) \right) \cap \{0, 1\}^{\vert V \vert} = \emptyset. \quad \text{(10)}
\]

Denote \( x^* \) an (integral) optimum of \( \arg \min_{x \in [0, 1]^{\vert V \vert}} C\ell(x) \). As \( C\ell(x) \) is total dual integral, we have that \( C\ell(x^*) = \min_{A \subseteq V} \ell(A) \). If \( \ell(x^*) = G \ell(x^*) \neq \min_{x \in [0, 1]^{\vert V \vert}} G \ell(x) \), there must be a negative directional gradient \( \nabla_v G \ell(x^*) \) for some \( v \in \mathbb{R}^{\vert V \vert} \) pointing into the unit cube from \( x^* \). From the definition of set function extensions and the fact that \( F \preceq_A G \), we have that \( 0 > \nabla_v G \ell(x^*) \geq \nabla_v F \ell(x^*) \) and therefore \( x^* \notin \arg \min_{x \in [0, 1]^{\vert V \vert}} F \ell(x) \).
3 Slack and margin rescaling

3.1 Slack rescaling

**Definition 6** (Slack rescaling for increasing set functions [8, 12, 13]). For an increasing set function $\ell$, 

$$S_1 \ell(x) = \max_{A \subseteq V} \ell(A) \left(1 - 2|A| + 2 \sum_{i \in A} x_i\right). \quad (11)$$

The derivation of this extension from the related convex surrogate is shown in Appendix A. We note that the factor 2 in the above definition is an artifact of the origin of the extension as a convex surrogate. We show below that we may instead use a tighter variant in the sense that $S_1 \leq S_+$, where $\leq$ denotes the inequality holds for all increasing functions.

**Definition 7** (A tighter slack rescaling variant for increasing set functions). For an increasing set function $\ell$, 

$$S_+ \ell(x) = \max_{A \subseteq V} \ell(A) \left(1 - |A| + \sum_{i \in A} x_i\right). \quad (12)$$

**Proposition 4.** Computation of $S_+ \ell(x)$ corresponds to a non-supermodular maximization problem for $x \in [0, 1]|V|$ and supermodular increasing $\ell$.

The proof is given in Appendix D.

**Proposition 5.** Definition 7 yields a convex extension for all increasing $\ell$.

**Proof.** Equation (12) is convex: The r.h.s. is a maximum over linear functions of $x$, and is therefore convex in $x$.

Equation (12) is an extension: We will use the notation set : $\{0, 1\}|V| \mapsto \mathcal{P}(V)$ to denote the set associated with an indicator vector. We must have that $S_+ \ell(x) = \ell(A)$ whenever $x \in \{0, 1\}|V|$ and set($x$) = $A$. For set($x$) = $A$ we have that 

$$\ell(A) \left(1 - |A| + \sum_{i \in A} x_i\right) = \ell(A). \quad (13)$$

We now must show that when set($x$) = $A$

$$\ell(A) \geq \ell(B) \left(1 - |B| + \sum_{i \in B} x_i\right) = \ell(B) \left(1 - |B| + |A \cap B|\right). \quad (14)$$

For any $B \not\subseteq A$:

$$\frac{\geq 0}{\ell(B) \left(1 - |B| + |A \cap B|\right)} \leq 0 \leq \ell(A), \quad (15)$$

and for any $B \subseteq A$

$$\frac{\leq \ell(A)}{\ell(B) \left(1 - |B| + |A \cap B|\right)} = 1 \leq \ell(A). \quad (16)$$

An example of the convex extension achieved with Definition 7 applied to the set function $\ell(\emptyset) = 0$, $\ell(\{a\}) = 0.5$, $\ell(\{b\}) = 1.5$, $\ell(\{a, b\}) = 4$ is shown in Figure 1.

We can see that Definition 7 is an instance of a general strategy for multiplicative extensions of increasing functions:

$$B\ell(x) = \max_{A \subseteq V} \ell(A) \cdot h_x(x, A) \quad (17)$$
where \( h_x : [0, 1]^{|V|} \times \mathcal{P}(V) \to \mathbb{R} \) is a function convex in \( x \). We require \( h_x(x, A) \leq \frac{\ell(\text{set}(x))}{\ell(A)} \) for integral \( x \) and all \( A \) in order to ensure that \( \ell(A)h_x(x, A) \leq \ell(\text{set}(x)) \), which is the general form of the inequality in Equation (14). This strategy is quite general if \( h_x \) can have a dependency on \( \ell \) as the convex closure is recovered simply by setting \( h_x(x, A) = \frac{\ell(\text{set}(x))}{\ell(A)} \); in Proposition 6 we will limit ourselves to \( h_x \) that have no explicit dependency on \( \ell \) beyond that it is an increasing function. For increasing \( \ell \), \( \frac{\ell(\text{set}(x))}{\ell(A)} \) is bounded below by 1 for all \( A \subseteq \text{set}(x) \) and is bounded below by 0 for all \( A \nsubseteq \text{set}(x) \). If we additionally assume that \( \ell \in \mathcal{G}_+ \), these bounds remain unchanged. If we take these bounds with respect to all increasing \( \ell \), or indeed all \( g \in \mathcal{G}_+ \), the bounds are sharp.

**Proposition 6.** \( h_x(x, A) = (1 - |A| + \sum_{i \in A} x_i)_+ \) is the largest possible convex function over the unit cube satisfying the constraints that for integral \( x \), \( h_x(x, A) \leq 1 \forall A \subseteq \text{set}(x) \), \( h_x(x, A) \leq 0 \forall A \nsubseteq \text{set}(x) \), and \( h_x(x, \text{set}(x)) = 1 \), where \((\cdot)_+ = \max(0, \cdot)\).

**Proof.** Each constraint is an equality constraint or an upper bound, and the values of the function are constrained only at the vertices of the unit cube, so the maximum convex function satisfying the bounds must be the lower hull of the constraint points, that is \( h_x(x, A) = \mathcal{C}[A \subseteq \cdot](x) \), where we have employed Iverson bracket notation in our definition of the set function to which we apply \( \mathcal{C} \):

\[
\mathcal{C}[A \subseteq \cdot](x) = \min_{\alpha \in \mathbb{R}^{|V|}} \sum_{B \subseteq V} \alpha_B[A \subseteq B] = \min_{\alpha \in \mathbb{R}^{|V|}} \sum_{B \supseteq A} \alpha_B \tag{18}
\]

subject to:

\[
\sum_{B \subseteq V} \alpha_B 1_B = x, \quad \sum_{B \subseteq V} \alpha_B = 1, \quad \alpha_B \geq 0 \forall B. \tag{19}
\]

For all \(|V| \in \mathbb{Z}_+ \) and \( A \subseteq V \), we have that

\[
\mathcal{C}[A \subseteq \cdot](x) = 0, \quad \forall x \in \text{conv} \left( \{ y \in \{0, 1\}^{|V|} | \text{set}(y) \nsubseteq A \} \right). \tag{20}
\]

For \( x \in \text{conv} \left( \{ y \in \{0, 1\}^{|V|} | \text{set}(y) \supseteq A \lor \exists i \in V, \text{set}(y) \cup \{i\} \supseteq A \} \right) \), we must have a linear function that interpolates between 0 at all integral points \( x \) such that \( \text{set}(x) \nsubseteq A \) and 1 at points such that \( \text{set}(x) \supseteq A \). This linear function is exactly \( 1 - |A| + \sum_{i \in A} x_i \).

In practice, we do not need to threshold \( h_x \) at zero (i.e. we do not need to apply the \((\cdot)_+ \) operation above) as this is redundant with the maximization in Equation (17) being achieved by \( A = \emptyset \).

**Remark 2.** The constraints in Proposition 6 are necessary and sufficient conditions on \( h \) for this multiplicative family to yield a convex extension for all increasing \( \ell \). They are also necessary and sufficient to yield a convex extension for all \( g \in \mathcal{G}_+ \).
We may similarly define the following convex extension based on margin rescaling:

\[ m(\{x\}) = g(\{x\}), \quad \forall x \in V, \quad (22) \]

and denote this function \( m \).

For non-increasing supermodular \( g \), we may extend the definition of slack rescaling as follows:

**Definition 9** (Slack rescaling for all supermodular \( g \)). Assume \( g \in G \). For a modular function \( m \), we will denote \( \text{vec}(m) \in \mathbb{R}^{|V|} \) the vectorization of \( m \), and we define

\[
S_g(x) = \langle \text{vec}(m g), x \rangle + \max_{A \subseteq V} \left( g(A) - m g(A) \right) \left( 1 - |A| + \sum_{i \in A} x_i \right). \quad (23)
\]

Examples of the convex extension achieved by Definition 9 are given in Figure 2. We now show that not only does \( S \) yield an extension for non-increasing supermodular functions, it yields a tighter extension for increasing supermodular functions:

**Proposition 7.** \( S_+ \leq \hat{g}_+ \). \( S \).

**Proof.** For \( g \in G_+ \),

\[
S_g(x) - S_+ g(x) = \max_{A \subseteq V} \left( \langle \text{vec}(m g), x \rangle + (g(A) - m g(A)) \left( 1 - |A| + \sum_{i \in A} x_i \right) \right) \quad (24)
\]

To show that this difference is greater than zero, we will demonstrate that the difference of each element indexed by \( A \subseteq 0 \) is greater than zero.

\[
\langle \text{vec}(m g), x \rangle + (g(A) - m g(A)) \left( 1 - |A| + \sum_{i \in A} x_i \right) - g(A) \left( 1 - |A| + \sum_{i \in A} x_i \right) = \sum_{i \in V} g(\{i\}) x_i - \sum_{i \in A} g(\{i\}) \left( 1 + \sum_{j \in A} (x_j - 1) \right) \quad (25)
\]

\[
= \sum_{i \in V} g(\{i\}) x_i - \sum_{i \in A} g(\{i\}) + \sum_{i \in A} (1 - x_i) g(\{i\}) \quad (26)
\]

\[
= \sum_{i \in V \setminus A} g(\{i\}) x_i + \xi \geq 0 \quad (27)
\]

\[ \square \]

### 3.2 Margin rescaling

We may similarly define the following convex extension based on margin rescaling:

**Definition 10** (Margin rescaling for supermodular \( g \) \[8, 12, 13\]).

\[
M_g(x) := \langle \text{vec}(m), x \rangle + \max_{A \subseteq V} g(A) - m(A) - |A| + \sum_{i \in A} x_i. \quad (28)
\]
Analogous to the development of $S$ from $S_1$, we have removed a factor of 2. It has been demonstrated that up to a strictly positive scale factor, margin rescaling yields an extension in the convex surrogate loss setting [12, Proposition 2]. We prove this next and in doing so show necessary and sufficient conditions on the scaling of the loss function for $M$ to yield an extension.

**Proposition 8 (Proposition 2 of [12]).** For every increasing function $\ell$, there exists a scalar $\gamma > 0$ such that $\gamma^{-1}M(\gamma \ell)$ is an extension of $\ell$. We will denote $M_\gamma := \gamma^{-1}M(\gamma \cdot)$ for short.

**Proof.** Wlog, we may assume that $\ell(\{i\}) = 0$, $\forall i \in V$. For $M_\gamma$ to yield an extension, we need that for all $A, B \subseteq V$, set $(x) = A$,

$$g(A) \geq g(B) \left| A \right| + \sum_{i \in B} x_i = g(B) - \left| A \right| + \left| A \cap B \right|.$$  \hfill (29)

$$\iff \gamma \leq \frac{\gamma(A) - g(B)) + \left| A \right| - \left| A \cap B \right| \geq 0}$$  \hfill (30)

$$\forall (A, B) \text{ s.t. } g(A) < g(B)$$  \hfill (31)

where the restriction to $g(A) < g(B)$ is due to the fact that $g(B) > g(A)$ reverses the direction of the inequality as we multiply both sides by $\frac{1}{g(B) - g(A)}$. As $g$ is increasing, we cannot simultaneously have that $\left| A \cap B \right| = \left| B \right|$ and $g(A) < g(B)$ so the numerator is strictly positive and the ratio is therefore strictly positive. \hfill \square 

**Proposition 9.** For $0 < \gamma_1 < \gamma_2$ both satisfying the conditions in Equation (31), $M_{\gamma_1} \preceq M_{\gamma_2}$.

**Proof.** Wlog, we will assume that $g \in G_+$ and ignore the modular transformation in Definition 10.

$$0 \leq \frac{1}{\gamma_2} \max_{A \subseteq V} (g_2 g(A) - \left| A \right| + \sum_{i \in A} x_i) - \frac{1}{\gamma_1} \max_{A \subseteq V} (g_1 g(A) - \left| A \right| + \sum_{i \in A} x_i)$$  \hfill (32)

$$\iff 0 \leq \frac{1}{\gamma_2} \left( g_2 g(A) - \left| A \right| + \sum_{i \in A} x_i \right) - \frac{1}{\gamma_1} \left( g_1 g(A) - \left| A \right| + \sum_{i \in A} x_i \right)$$  \hfill (33)

$$= \left( \frac{1}{\gamma_2} - \frac{1}{\gamma_1} \right) \left( -\left| A \right| + \sum_{i \in A} x_i \right)$$  \hfill (34)
We next demonstrate that margin rescaling is dominated by slack rescaling, and the latter is therefore closer to the convex closure.

**Proposition 10.** $M_\gamma \preceq G_S$, for all $\gamma$ satisfying Equation (31).

**Proof.** For $g \in G_+$ and $\gamma$ satisfying Equation (31),

$$0 \preceq S g(x) - M_\gamma g(x)$$

$$\iff 0 \preceq (g(A) - g(A)|A| + g(A) \sum_{x \in A} x_i) - \left( g(A) - \frac{1}{\gamma} |A| + \frac{1}{\gamma} \sum_{x \in A} x_i \right)$$

$$= \left( \frac{1}{\gamma} - g(A) \right) \left( |A| - \sum_{x \in A} x_i \right).$$

From Equation (31),

$$\frac{1}{\gamma} \geq \frac{g(A) - g(B)}{|A| - |A \cap B|} \geq g(A)$$

for some $B \subseteq A$. The second inequality follows as the numerator is less than or equal to $g(A)$ as $g$ is increasing and the denominator is at least 1. The desired result follows.

As in Equation (17), we may view margin rescaling as a special case of an additive strategy for computing convex extensions in which

$$B \ell(x) = \frac{1}{\gamma} \max_{A \subseteq V} \gamma \ell(A) + h_+(x, A)$$

where as in Equation (17), $h_+$ is some convex function. We have seen from Proposition 8 that $\gamma$ is necessary to guarantee that the operator yields an extension for members of this family. We again restrict ourselves to $h_+$ that do not have access to $\ell$ except the assumption that it is an increasing function satisfying some scaling constraints that we may enforce by setting $\gamma$ as a function of $\ell$. Allowing $\gamma$ to depend on $\ell$ in a way that may require an exponential number of oracle accesses to $\ell$ to compute is extremely generous, but we will see in Proposition 11 that even with this degree of freedom the optimal multiplicative extension dominates all additive extensions.

**Proposition 11.** Slack rescaling, which is the dominating member of the multiplicative family of extensions following Equation (17) (Corollary 2), dominates the family of additive extensions following Equation (39), where $h_+(x, A)$ has no oracle access to $\ell$.

**Proof.** For Equation (39) to yield an extension, we require that $\gamma \ell(A) + h_+(x, A) \geq \gamma \ell(set(x))$ for integral $x \in [0, 1]^{V}$, which is satisfied for $h_+(x, A) = \gamma (\ell(set(x)) - \ell(A))$. For increasing $\ell$, $\gamma (\ell(set(x)) - \ell(A))$ is bounded below by 0 for all $A \subseteq set(x)$, and for $A \not\subseteq set(x)$ we have from Equation (29) that $\gamma (\ell(set(x)) - \ell(A)) \geq |A| + |set(x) \cap A|$. These yield a set of constraints similar to those in Proposition 6 but shifted downwards by one, and without thresholding to zero. In Proposition 6 the maximal convex function satisfying the constraints had the form $(1 - |A| + \sum_{x \in A} x_i) +$, and we may conclude that the optimal function for the additive family of extensions is achieved by $h_+(x, A) = -|A| + \sum_{x \in A} x_i$.

This indicates that $M_\gamma \ell$ with maximal $\gamma$ satisfying Equation (31) (Proposition 9) is the additive convex extension closest to the convex closure of $\ell$. From Proposition 10, $M_\gamma$ is dominated by $S_+$ for all $\gamma$ such that $M_\gamma$ yields an extension.

### 4 Conclusions

In this work, we have formally analyzed slack and margin rescaling as convex extensions of supermodular set functions. Although they were originally developed for empirical risk minimization of (increasing) functions, we show that supermodularity can be exploited to first transform a set function into an increasing function with a linear number of accesses to the loss function. We have
shown that, while slack rescaling dominates margin rescaling in the sense that it is strictly closer to the convex closure, computation of slack rescaling for a supermodular function corresponds to a non-supermodular maximization problem. Margin rescaling, by contrast, remains tractable. We have further shown that margin and slack rescaling correspond to optimal additive and multiplicative extensions, respectively, given a computational budget of one oracle access to the loss function.

Acknowledgments

This work is funded by Internal Funds KU Leuven, FP7-MC-CIG 334380, and the Research Foundation - Flanders (FWO) through project number G0A2716N.

References

[1] H. Choi, O. Meshi, and N. Srebro. Fast and scalable structural SVM with slack rescaling. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, pages 667–675, 2016.
[2] S. Fujishige. Submodular Functions and Optimization. Elsevier, 2005.
[3] S. Jegelka and J. A. Bilmes. Submodularity beyond submodular energies: Coupling edges in graph cuts. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1897–1904, 2011.
[4] F. Kahl and P. Strandmark. Generalized roof duality. Discrete Applied Mathematics, 160(16-17):2419–2434, 2012.
[5] A. Krause and D. Golovin. Submodular function maximization. In L. Bordeaux, Y. Hamadi, and P. Kohli, editors, Tractability: Practical Approaches to Hard Problems. Cambridge University Press, 2014.
[6] M. P. Kumar, V. Kolmogorov, and P. H. S. Torr. An analysis of convex relaxations for MAP estimation of discrete MRFs. Journal of Machine Learning Research, 10:71–106, 2009.
[7] L. Lovász. Submodular functions and convexity. In Mathematical Programming The State of the Art, pages 235–257. Springer, 1983.
[8] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, 6:1453–1484, 2005.
[9] J. Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing, pages 67–74, 2008.
[10] J. Vondrák. Submodularity and curvature: The optimal algorithm. RIMS Kôkyûroku Bessatsu, 23:253–266, 2010.
[11] A. Weller, D. Sontag, and M. Rowland. Tightness of LP relaxations for almost balanced models. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, pages 47–55, 2016.
[12] J. Yu and M. B. Blaschko. Learning submodular losses with the Lovász hinge. In F. Bach and D. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Journal of Machine Learning Research: W&CP, pages 1623–1631, 2015.
[13] J. Yu and M. B. Blaschko. A convex surrogate operator for general non-modular loss functions. In A. Gretton and C. Robert, editors, International Conference on Artificial Intelligence and Statistics, volume 51 of Journal of Machine Learning Research: W&CP, pages 1032–1041, 2016.

A Slack and Margin Rescaling

In this section, we show how we arrive at Definition 6 and Definition 10 from slack and margin rescaling [12, Equations (13) and (12)]. Slack and margin rescaling were introduced in the context of convex surrogate loss functions, and so the notation is defined with respect to a loss function $\Delta(\cdot, \cdot)$ that compares a correct ground truth labeling $y \in \{-1, 1\}^p$ with a (potentially incorrect) prediction.
\( y \in \{-1, 1\}^P \). We identify this function with a set function, where the input is the set of entries for which \( y \) and \( \tilde{y} \) differ [12, Equation (6)]. The convex surrogate is defined with respect to axes \( g^j \cdot y^j \), where \( g \in \mathbb{R}^p \) is a vector of predictions. To translate from coordinates for a convex surrogate loss function to those for an extension of a set function, we must take the mapping \( x_j = 1 - g^j \cdot y^j \) (where \( x \) is this paper’s notation). From the definition of the slack rescaling operator [12, Equation (13)]:

\[
\max_{\tilde{y} \in Y} \Delta(y, \tilde{y}) (1 + \langle g, \tilde{y} \rangle - \langle g, y \rangle) = \max_{A \subseteq V} \ell(A) \left(1 + \sum_{i \in A} g^i \tilde{y}^i + \sum_{j \in V \setminus A} g^j \tilde{y}^j - \sum_{k \in V} g^k y^k\right)
\]

(40)

\[
= \max_{A \subseteq V} \ell(A) \left(1 - \sum_{i \in A} (1 - x_i) + \sum_{j \in V \setminus A} (1 - x_j) - \sum_{k \in V} (1 - x_k)\right)
\]

(41)

\[
= \max_{A \subseteq V} \ell(A) \left(1 - 2 \sum_{i \in A} (1 - x_i)\right) = \max_{A \subseteq V} \ell(A) \left(1 - 2|A| + 2 \sum_{i \in A} x_i\right).
\]

(42)

The derivation for margin rescaling is analogous.

**B Multi-linear Extension**

The multi-linear extension is due to Vondrák [9].

**Definition 11** (Multi-linear extension [9]).

\[
\mathcal{V} \ell(x) := \sum_{A \subseteq V} \ell(A) \prod_{i \in A} x_i \prod_{j \in V \setminus A} (1 - x_j)
\]

(43)

We note that \( \mathcal{V} \) does not yield a convex extension in general, which is implied by Proposition 2, the fact that \( \mathcal{V} \neq C \), and the following proposition:

**Proposition 12.**

\[
C \leq \mathcal{V}.
\]

(44)

**Proof.** We can see that the definition of \( \mathcal{V} \) is an expectation over Boolean random variables of the function value of \( \ell \) for the sets defined by those random variables. Consequently the multi-linear extension lies entirely within the convex hull of the points defined in the proof of Proposition 2, and the values of the multi-linear extension must therefore be pointwise greater than the values of the lower hull, which is equivalent to the convex closure. \( \square \)

**C Extensions through modular upper bounds on submodular functions**

In this section, we show that a known and frequently employed modular upper bound on submodular functions [3] does in fact define a convex extension to a supermodular function, but that computation of the extension does not correspond to a supermodular maximization in general.

**Definition 12** (Modular upper bound to a submodular function (Lemma 1 [3])). For a submodular function \( f \) and an arbitrary \( B \subseteq V \), define the modular function

\[
j_{B, f}(A) := f(B) + \sum_{i \in A \setminus B} (f(B \cup \{i\}) - f(B)) - \sum_{i \in B \setminus A} (f(V) - f(V \setminus \{i\})�).
\]

(45)

**Lemma 2.** For all \( f \in S \) and all \( A \subseteq V \)

\[
j_{A, f}(A) = f(A).
\]

(46)
where we make use of the fact that $g$ is neither submodular nor supermodular. For any modular $m$, $j_{B,f+m}(A) = j_{B,f}(A) + m(A)$.

\textbf{Proof.} Convexity is immediate as $Cm$ is linear for any modular $m$.

\begin{equation}
\mathbf{J}_g = \max_{B \subseteq V} \mathbf{C}(-j_{B,-g}(\cdot))
\end{equation}

is a convex extension.

\textbf{Proof.} The convex closure of a modular function is linear time computable. As we are showing the result for arbitrary $g \in \mathcal{G}$, the only relevant fact about $-g$ is that it is submodular. The maximization in Equation (50) is supermodular if for all $f \in \mathcal{S}$, $\forall A \subseteq B \subset V$, $k \in V \setminus B$,

\begin{equation}
-\mathbf{C}(j_{f,B\cup\{k\}}(\cdot)) \cdot (x) + \mathbf{C}(j_{f,B}(\cdot))(x) \geq -\mathbf{C}(j_{f,A\cup\{k\}}(\cdot))(x) + \mathbf{C}(j_{f,A}(\cdot))(x),
\end{equation}

where we make use of the fact that $\mathbf{C}(-m) = -\mathbf{C}m$ for modular $m$. For any $B$.

\begin{equation}
\mathbf{C}(j_{B,f}(\cdot))(x) = f(B) + \sum_{i \in V \setminus B} x_i (f(B \cup \{i\}) - f(B)) - \sum_{i \in B} (1 - x_i) (f(V) - f(V \setminus \{i\})).
\end{equation}

A numerical counterexample to this corresponding to a supermodular maximization is $g(A) = 0, \forall A \in \mathcal{P}(V) \setminus \{V,V \setminus \{a\}\}$, $g(V) = 1$, $g(V \setminus \{a\}) = \frac{1}{3}$, and $x_i = \frac{2}{3}, \forall i$. In this setting, maximizing Equation (52) w.r.t. $B$ is neither submodular nor supermodular.
Proof of Proposition 4

Proof. To show that the optimization is not in general a supermodular maximization, we may consider the following counterexample: \( V = \{a, b\} \), \( \ell(\emptyset) = \ell(\{a\}) = \ell(\{b\}) = 0 \), \( \ell(\{a, b\}) = \varepsilon \). For \( \varepsilon > 0 \) this is strictly supermodular. Denote

\[
\hat{\ell}(A) := \ell(A) \left( 1 - |A| + \sum_{i \in A} x_i \right) \tag{53}
\]

For \( x = 0 \), we have that \( \hat{\ell}(\emptyset) = \hat{\ell}(\{a\}) = \hat{\ell}(\{b\}) = 0 \) and \( \hat{\ell}(\{a, b\}) = \varepsilon (1 - 2 + 0) = -\varepsilon \), but this indicates that \( \ell = -\hat{\ell} \) and they cannot both be supermodular. \( \square \)