Learning Preference Models with Sparse Interactions of Criteria

Margot Herin[1] , **Patrice Perny** [1] , **Nataliya Sokolovska**[2]
[1]Sorbonne University, CNRS, LIP6, Paris, France
[2] Sorbonne University, CNRS, LCQB, Paris, France
{margot.herin,patrice.perny}@lip6.fr, nataliya.sokolovska@sorbonne-universite.fr

Abstract

Multicriteria decision making requires defining the result of conflicting and possibly interacting criteria. Allowing criteria interactions in a decision model increases the complexity of the preference learning task due to the combinatorial nature of the possible interactions. In this paper, we propose an approach to learn a decision model in which the interaction pattern is revealed from preference data and kept as simple as possible. We consider weighted aggregation functions like multilinear utilities or Choquet integrals, admitting representations including non-linear terms measuring the joint benefit or penalty attached to some combinations of criteria. The weighting coefficients known as Möbius masses model positive or negative synergies among criteria. We propose an approach to learn the Möbius masses, based on iterative reweighted least square for sparse recovery, and dualization to improve scalability. This approach is applied to learn sparse representations of the multilinear utility model and conjunctive/disjunctive forms of the discrete Choquet integral from preferences examples, in aggregation problems possibly involving more than 20 criteria.

1 Introduction

One of the main challenges of preference modeling in the context of multicriteria decision making is to construct simple and explainable decision models keeping sufficient flexibility to accurately model human preferences and decision behaviors. In the field of multiattribute/multicriteria decision making, the presence of possible interactions among criteria is a source of complexity for preference modeling because it prevents representing preferences by simple linear models such as weighted arithmetic means. More sophisticated weighted evaluation models including non-linear terms measuring the joint benefit or penalty attached to some groups of criteria are needed. Interactions may be represented by product terms as in the multilinear utility model [Keeney *et al.*, 1993], or by minimum or maximum operations as in the Choquet integral [Grabisch *et al.*, 2009], or possibly by other monotonic non-linear factors involving several attributes or criteria.

For example, if the decision maker (DM) prefers solutions with balanced utility vectors then one cannot simply use a weighted arithmetic mean of type $\sum_{i=1}^{n} w_i x_i$ to define the overall utility attached to a utility vector $(x_1, \ldots, x_n)$. For instance, it would not be possible to make $(0.5, 0.5)$ better than $(1, 0)$ and $(0, 1)$ simultaneously. A simple solution in this case may be to allocate a bonus to alternatives that perform well on all criteria. It is sufficient to add a conjunctive term to the weighted mean, e.g., $\prod_{i=1}^{n} x_i$ or $\min\{x_1, \ldots, x_n\}$, weighted by a sufficiently large positive coefficient. We observe indeed that this conjunctive term takes a strictly positive value on $(0.5, 0.5)$ but remains null on (1,0) and (0,1). More generally, beyond the linear part of the aggregation function, various interaction terms could be inserted to model positive but also negative synergies within some groups of criteria.

However, allowing the possibility of interactions in a decision model is a source of complexity in preference modeling and preference learning due to the combinatorial nature of these interactions. In an aggregation model involving n criteria, interactions may appear in any of the $2^n - n - 1$ subsets of criteria including more than one element. For $n = 10$ criteria it represents slightly more than 1000 possible interactions to analyze. When $n = 20$ it already represents more than one million possible interactions. In order to preserve scalability in learning the interactions, a standard approach is to reduce the combinatorial aspect of the problem by allowing only a limited number of them. For example one can just consider pairwise interactions of criteria. More generally one could limit interactions to subsets of size k for some k significantly smaller than n. However, this prior restriction eliminates very simple and natural representations of preferences that require larger interactions. For example, in the above example, we have shown that the introduction of a conjunctive term including all criteria may be natural to promote balanced solutions. Besides, another simple example is the Hurwicz criterion [Hurwicz, 1951] which is standardly used to make a tradeoff between the worst and the best components. It is defined as a convex combination of $\min\{x_1, \ldots, x_n\}$ and $\max\{x_1, \ldots, x_n\}$ and therefore includes two interactions terms. Both of them involve the entire set of criteria and cannot be simply approximated by interactions on smaller sets.

*See http://www-desir.lip6.fr/~perny/ijcai23/appendix.pdf for the proofs not included in the paper.

Here we would like to propose another approach where no prior restriction on the possible interaction groups is made. The useful groups will emerge from preference data with the aim of constructing a model as simple as possible, that fits well the preference examples. In this perspective, we consider an aggregation function defined as a weighted sum of factors made of criteria in interaction and propose a method to learn a sparse representation of these weights.

The paper is organized as follows: In Section 2 we recall some background on multicriteria aggregation models including interacting components. Then in Section 3 we introduce an approach to obtain sparse representations of interactions by iterative reweighted least square regularization, and dualization to improve the scalability of the approach. Section 4 presents some numerical tests to evaluate the performance of the proposed approach both in terms of computation time and generalizing performances.

2 Evaluation Models with Interacting Criteria

The general framework. In a multidimensional decision problem, the alternatives are described with respect to n points of views ($n > 1$) that must be considered in the evaluation process. Depending on the decision context, these viewpoints may refer to different attributes describing the alternatives (multiattribute decision making), different criteria used to compare the alternatives (multicriteria decision making), or different individuals expressing their preferences (multiagent decision making). In any case, every alternative x is described by a vector $(c_1(x), \ldots, c_n(x))$ of consequences where $c_i(x)$ represents the value of x with respect to the i^{th} viewpoint. Let $N = \{1, \ldots, n\}$ denote the index set of viewpoints. Let X_i denote the set of possible consequences on the i^{th} viewpoint for all $i \in N$ and $X = X_1 \times \ldots \times X_n$ the set of all possible consequence vectors. Let $\succsim$ be the preference relation of the DM over X. One standard approach in preference modeling consists of representing $\succsim$ by a decomposable function u on X of the form:

$$u(x) = F(u_1(c_1(x)), \ldots, u_n(c_n(x))) \tag{1}$$

where $u_i : X_i \to [0, 1], i \in N$ are marginal utility functions representing the attractiveness of consequences $c_i(x)$ for the DM and $F : [0, 1]^n \to [0, 1]$ is an aggregation function non-decreasing in each argument. Function u is said to represent $\succsim$ when $x \succsim y$ if and only if $u(x) \geq u(y)$. Throughout the paper, $\succ$ is used to denote the asymmetric part (strict preference) of $\succsim$ whereas $\sim$ denotes its symmetric part (indifference).

Let us recall two standard examples of function u, widely used to represent preferences in multiattribute/multicriteria decision problems involving criteria in interaction:

Example 1. *The multilinear utility model defined by:*

$$ML_v(x) = \sum_{S \subseteq N} v(S) \prod_{i \in S} u_i(c_i(x)) \prod_{i \notin S} (1 - u_i(c_i(x))) \tag{2}$$

was introduced in the context of multiattribute decision making under risk [Keeney *et al.*, 1993] but is also used and axiomatically justified in the context of multiattribute decision making [Dyer and Sarin, 1979] and [Grabisch, 2016, Chap. 6]. In the context of this paper, v is a set function defined on the power set 2^N and valued in the unit interval, assigning a weight to any subset of viewpoints. Another well-known decision model defined from a set function v is the following:

Example 2. *The discrete Choquet integral*

$$C_v(x) = \sum_{i=1}^{n} \left[v(X_{(i)}) - v(X_{(i+1)}) \right] u_{(i)}(c_{(i)}(x)) \tag{3}$$

where (.) is any permutation of N such that $u_{(i)}(c_{(i)}(x)) \leq u_{(i+1)}(c_{(i+1)}(x))$ and $X_{(i)} = \{(i), \ldots, (n)\}, i \in N$ with $x_{(0)} = 0$ and $X_{(n+1)} = \emptyset$. For instance, if $n = 3$ and x is such that $u_2(c_2(x)) \leq u_1(c_1(x)) \leq u_3(c_3(x))$, then $C_v(x) = [v(1, 2, 3) - v(1, 3)]u_2(c_2(x)) + [v(1, 3) - v(3)]u_1(c_1(x)) + v(3)u_3(c_3(x))$.

The Choquet integral was initially introduced in the context of decision making under uncertainty [Schmeidler, 1989] but is also widely used in the context of multiattribute/multicriteria decision making [Grabisch, 1996; Grabisch and Labreuche, 2010].

In both models (multilinear and Choquet), one can assume that v is normalized, i.e., it satisfies boundary conditions $v(\emptyset) = 0$ and $v(N) = 1$. It is also generally assumed that v is *monotonic* with respect to set inclusion (which guarantees the monotonicity of u w.r.t. weak Pareto dominance). More formally, if $v(A) \leq v(B)$ for all subsets $A, B \subseteq N$ such that $A \subseteq B$ then $u_i(c_i(x)) \geq u_i(c_i(y))$ for all $i \in N$ implies $u(x) \geq u(y)$ for all pairs (x, y) whether u is defined by (2) or (3). Such monotonic set functions named *capacities* are widely used in preference aggregation [Grabisch *et al.*, 2003; Grabisch, 2016]. The capacity provides a non-necessarily additive weighting system (the weight of a set is not necessarily the sum of the weights of its elements) and super-additivity and sub-additivity are used to model positive and negative synergies between the components of the decision model.

The capacity is a preference parameter that must be elicited by questioning the DM or learned from preference examples. The other preference parameters used in these models are marginal utility functions $u_i, i \in N$ that can be obtained before the identification of the capacity. In the case of the multilinear model, marginal utilities can be elicited using standard gamble queries under *mutual utility independence*, an axiom usually assumed to justify the multilinear model under uncertainty [Keeney *et al.*, 1993]. They can alternatively be derived from comparisons of preference intensities under *weak difference independence*, an axiom usually assumed to justify the multilinear model in multicriteria/multiattribute decision making [Grabisch, 2016, Chap. 6]. For the Choquet integral, the utility functions can be obtained using standard sequences of tradeoff queries [Wakker and Deneffe, 1996], or constructed with the Macbeth method [Grabisch and Labreuche, 2010] or learned from preference examples [Herin *et al.*, 2022b]. We assume here that functions u_i have been elicited beforehand using one of the above mentioned methods and we focus on learning the capacity from preference examples. From now on, any alternative x is described by the utility vector $\mathbf{x} = (x_1, \ldots, x_n) \in [0, 1]^n$ where $x_i = u_i(c_i(x)), i \in N$.

Respresentations based on Möbius masses. A useful alternative representation of any capacity v is given by its Möbius transform m_v defined as follows:

$$m_v(S) = \sum_{T \subseteq S} (-1)^{|S \setminus T|} v(T) \text{ with } v(S) = \sum_{T \subseteq S} m_v(T)$$

The values $m_v(S)$ are called Möbius masses. We remark that we necessarily have $\sum_{T \subseteq N} m_v(T) = 1$, since $v(N) = 1$.

It is interesting to note that $ML_v(x)$ can be direclty defined from Möbius masses as follows [Owen, 1975]:

$$ML_v(x) = \sum_{S \subseteq N} m_v(S) \prod_{i \in S} x_i \quad (4)$$

Similarly, $C_v(x)$ admits several reformulations from m_v [Chateauneuf and Jaffray, 1989; Grabisch *et al.*, 2009]:

$$C_v(x) = \sum_{S \subseteq N} m_v(S) \min_{i \in S}\{x_i\} \text{ conjunctive form} \quad (5)$$

$$C_v(x) = \sum_{S \subseteq N} m_{\bar{v}}(S) \max_{i \in S}\{x_i\} \text{ disjunctive form} \quad (6)$$

where $\bar{v}$ is the conjugate of v, i.e., the capacity defined by $\bar{v}(S) = v(N) - v(N \setminus S)$ for all $S \subseteq N$.

Example 3. *Let $N = \{1, 2, 3\}$ and $v, \bar{v}$ defined on N by:*

S	1	2	3	1, 2	1, 3	2, 3	1, 2, 3
$v(S)$	0.1	0.2	0.3	0.3	0.4	0.5	1.0
$m_v(S)$	0.1	0.2	0.3	0.0	0.0	0.0	0.4
$\bar{v}(S)$	0.5	0.6	0.7	0.7	0.8	0.9	1.0
$m_{\bar{v}}(S)$	0.5	0.6	0.7	−0.4	−0.4	−0.4	0.4

Let $x = (x_1, x_2, x_3)$ with $x_2 \leq x_1 \leq x_3$. Then we have:
$ML_v(x) = 0.1\, x_1 + 0.2\, x_2 + 0.3\, x_3 + 0.4\, x_1 x_2 x_3$ *(Eq. 4)*
$C_v(x) = 0.1\, x_1 + 0.2\, x_2 + 0.3\, x_3 + 0.4\, x_2$ *(Eq. 5).*

Here the disjunctive form of C_v (Eq. 6) is less interesting because $m_{\bar{v}}$ is less sparse than m_v. The converse holds for $C_{\bar{v}}$. We have $C_{\bar{v}}(x) = 0.1x_1 + 0.2x_2 + 0.3x_3 + 0.4x_3$ by Eq. 6 whereas the conjunctive form based on $m_{\bar{v}}$ is less compact.

In order to factorize and generalize Equations 4-6, we will now consider a general decision model of the form:

$$F(x) = \sum_{S \subseteq N} m_S \phi_S(x_S) \quad (7)$$

where m_S are Möbius masses and ϕ_S aggregates the quantities $x_i, i \in S$ to define the interaction term $\phi_S(x_S)$. Thus ϕ_S is the product if F is the multilinear model and ϕ_S is the min (resp. max) operation if F is the conjunctive (resp. disjunctive) form of the Choquet integral. Note that function $F(x)$ reads as the following inner product $F(x) = \langle \mathbf{m}, \phi(\mathbf{x}) \rangle$ where $\mathbf{m} = (m_S)_{S \subseteq N}$ and $\phi : \mathbb{R}^n \to \mathbb{R}^{2^n}$ maps x into a non-linear feature space: $\phi(x) = (\phi_S(x_S))_{S \subseteq N}$. Both vectors $\mathbf{m}$ and $\phi(\mathbf{x})$ are indexed by the subsets $S \subseteq N$ numbered in lexicographic order.

Möbius masses and interactions. A capacity v is said to be k-additive if $m_v(S) \neq 0$ for some S of size k and $m_v(S) = 0$ for all S of size greater than k [Grabisch, 1997]. When v is 1-additive, all interaction terms vanish and F boils down to a weighted arithmetic mean of components x_i. When v is 2-additive, only pairwise interactions are possible in the model (terms of type $\phi_{ij}(x_i, x_j)$ in Equation 7). Their coefficients $m_v(\{i, j\}) = v(\{i, j\}) - v(\{i\}) - v(\{j\})$ can be positive or negative and their magnitude measures the importance of the interaction between i and j (the gap to additivity).

It is frequently assumed that capacities are k-additive for some $k < n$ prior to preference analysis so as to make sure that v admits a polynomial-size representation in terms of Möbius masses. However, it strongly reduces the descriptive power of the ML and Choquet models. For example, the preference for balanced solutions mentioned in the introduction requires $m_v(N) > 0$ (a bonus is given to alternatives that perform well on all criteria) which is incompatible with k-additivity for any $k < n$. In this case, a n-additive capacity is needed without any restriction on the size of possible interactions. Several measures of interaction have been proposed in the literature to measure the magnitude of interactions within sets of any size. For example, the Banzhaf and Shapley interactions indices $I_B(S)$ and $I_{Sh}(S)$ associated to a subset $S \subseteq N$ in models ML_v and C_v are related to Möbius masses m_v as follows [Grabisch *et al.*, 2009]:

$$I_B(S) = \sum_{T \supseteq S} \frac{m_v(T)}{2^{|T|-|S|}} \qquad I_{Sh}(S) = \sum_{T \supseteq S} \frac{m_v(T)}{|T| - |S| + 1}$$

We remark that these interaction indices tend to vanish as Möbius masses tend to 0. Thus, favouring the sparsity of the Möbius vector $\mathbf{m}$ in an aggregation function defined by Equation 7 tends to reduce the criteria interactions in the associated decision model.

Related work. As far as the identification of the capacity used in a decision model is concerned, several approaches based on the least squares criterion or variance minimization of the model under preference constraints have been proposed in the field of multicriteria analysis for the Choquet integral [Grabisch *et al.*, 2008; Grabisch and Labreuche, 2010]. In the field of machine learning different learning algorithms have been proposed, e.g., Choquistic regression [Tehrani and Hülermeier, 2013], support vector machines (SVM) with Choquet kernel [Tehrani, 2021], and ridge regression for Choquet regression [Kakula *et al.*, 2020]. Moreover a neural network was recently proposed to learn a hierarchical Choquet model [Bresson *et al.*, 2020]. Active learning approaches also exists based on regret minimization [Benabbou *et al.*, 2017]. Some recent contributions using regression also exist for the multilinear model [Pelegrina *et al.*, 2018; Pelegrina *et al.*, 2020].

The complexity of models involving capacities, having in general $2^n - 2$ free parameters, is prohibitive for many real-life applications. Very often, a prior complexity reduction is obtained by considering models with k-additive capacities ($k = 2$ being the most common choice) [Grabisch *et al.*, 2008; Hüllermeier and Tehrani, 2013; Galand and Mayag, 2017; Ah-Pine *et al.*, 2018; Bresson *et al.*, 2020; Pelegrina *et al.*, 2020]. Similar restrictions exist for limiting interactions with the notion of k-interactivity [Beliakov and Wu, 2019].

A less restrictive and more flexible attempt to reduce models complexity is to derive a sparse representation of the capacity from preference data using the L_1 penalty term [Anderson *et al.*, 2014; Adeyeba *et al.*, 2015; Pinar *et al.*, 2017; de Oliveira *et al.*, 2022] where the regularization was applied either to the capacity, or to the interaction index. Recently, some evidence was given that Möbius representations often

lead to more compact preference representations and an approach based on linear programming (LP) was proposed to learn a sparse Möbius transform of the capacity in the Choquet integral [Herin *et al.*, 2022a]. However the absence of prior restriction on capacities comes at the expense of computation times and scalability.

Our contribution in this paper is to propose a faster and more scalable algorithm to learn sparse Möbius representations of capacities from preference examples, for the multilinear and Choquet models and any other instance of Eq. 7. Our approach relies on iteratively re-weighted least squares and dualization as explained in the next section.

3 A Dual IRLS for Sparse Preference Learning

Our objective is to learn a sparse representation of $\mathbf{m}$ based on a training set of preferences statements $\{(x^i, y^i) \in \mathcal{X}^2 : x^i \succ y^i, i \in P\}$ and possibly of indifference statements $\{(x^i, y^i) \in \mathcal{X}^2 : x^i \sim y^i, i \in I\}$. A well-known workhorse for learning sparse models is the L_1-norm penalty. This is indeed a sparse-inducing penalty, in the sense that it promotes solutions with few non-null coefficients. A major application of this regularization is the LASSO estimate [Tibshirani, 1996] in linear regression. Then, our learning problem is to minimize both the error on the preference examples and the L_1 norm of the Möbius vector. The L_1-penalization is only applied to the terms involving at least two criteria so as to minimize interactions. The approximation problem is thus formulated as follows:

$$
\begin{aligned}
(\mathcal{P})\ \min\ & \textstyle\sum_{i\in P} \epsilon_i + \sum_{i\in I}(\epsilon_i^- + \epsilon_i^+) + \lambda \sum_{j=n+1}^{2^n} |m_j| \\
& \langle \mathbf{m}, \phi(\mathbf{x^i})\rangle - \langle \mathbf{m}, \phi(\mathbf{y^i})\rangle + \epsilon_i \geq \delta,\ i \in P && (8)\\
& \langle \mathbf{m}, \phi(\mathbf{x^i})\rangle - \langle \mathbf{m}, \phi(\mathbf{y^i})\rangle + \epsilon_i^+ - \epsilon_i^- = 0,\ i \in I && (9)\\
& \langle \mathbf{m}, \mathbf{1}\rangle = 1 && (10)\\
& \epsilon_i \geq 0,\ i \in P,\ \ \epsilon_i^+, \epsilon_i^- \geq 0,\ i \in I && (11)
\end{aligned}
$$

where variable m_j is the j^{th} component of vector $\mathbf{m}$. The hyper-parameter $\lambda > 0$ controls the level of regularization and δ is a strictly positive discrimination threshold used to separate preference from indifference situations. Variable ϵ_i models the positive error made on the preference example $\mathbf{x^i} \succ \mathbf{y^i}$, while $\epsilon_i^+ - \epsilon_i^-$ models the signed error made on the indifference $\mathbf{x^i} \sim \mathbf{y^i}$. The constraint $m(\emptyset) = 0$ is implicit.

Problem $\mathcal{P}$ can be solved by linear programming using the following L_1-norm linearization:

$$
\begin{aligned}
(\mathcal{P})\quad \min\ & \sum_{i\in P} \epsilon_i + \sum_{i\in I}(\epsilon_i^- + \epsilon_i^+) + \lambda \sum_{j>n}(w_j^+ + w_j^-) \\
& m_j = w_j^+ - w_j^-,\ \ j = n+1, \ldots, 2^n \\
& w_j^+, w_j^- \geq 0,\ \ j = n+1, \ldots, 2^n \\
& \text{s.t. } (8), (9), (10), (11)
\end{aligned}
$$

where variables w_j^+, w_j^- are used for the linearization of $|m_j|$.

Despite a simple linearization, the obtained linear program still drags an exponential number of variables ($2^n(3-2n) + |P| + 2|I|$) and thus is hardly solvable for more than a dozen of criteria. For the sake of scalability, we propose to solve $\mathcal{P}$ by solving a sequence of sub-problems $\mathcal{P}_k$ that admit an efficient dual formulation. More precisely, we use an iteratively reweighted least square (IRLS) algorithm [Daubechies *et al.*, 2010; Bach *et al.*, 2012; Beck, 2015; Grandvalet, 1998] that consists in approximating the solution of a L_1-penalized problem with a sequence of least squares problems. Sparsity is recovered by increasingly penalizing non significant coefficients with a squared L_2 regularization. The interest of this method lies in the fact that a least squares problem is easy to solve in general. In our case, we will show that the least square problem $\mathcal{P}_k$ admits a compact dual form whose size is no longer exponential in n the number of criteria, but linear in $|P|+|I|$, the number of preference examples. More specifically, L_1-optimization is linked to least squares problems through the quadratic variational formulation of the L_1-norm [Bach *et al.*, 2012] that allows absolute values to be expressed as infimums of weighted squared values:

$$\textstyle\sum_{j>n} |m_j| = \frac{1}{2} \min_{\mathbf{z}\geq \mathbf{0}} \sum_{j>n} (\frac{m_j^2}{z_j} + z_j) \qquad (12)$$

Using Eq. 12, we now establish Proposition 1 providing an IRLS algorithm that approximatively solves $\mathcal{P}$. The proof relies on the framework introduced in [Beck, 2015] that gives conditions under which an optimization problem can be solved by alternating minimization (here on m and z) and insights on how this algorithm can lead to IRLS sequences.

Proposition 1. *Let $\eta > 0$ be a smoothing parameter. Consider the sequence $\mathbf{m}^{(k)}$ initialized with $\mathbf{m}^{(0)} = \mathbf{1}$ such that:*

$$
\begin{aligned}
\mathbf{m}^{(k+1)} \in \operatorname{argmin} & \sum_{i\in P} \epsilon_i + \sum_{i\in I}(\epsilon_i^- + \epsilon_i^+) + \sum_{j>n} \frac{\lambda m_j^2}{\sqrt{{m_j^{(k)}}^2 + \eta^2}} \\
& \textit{s.t. } (8), (9), (10), (11)
\end{aligned}
$$

Then we have: $\lim_{k\to\infty} J(\mathbf{m}^{(k+1)}) - J^* \leq (2^n - n)\eta$ where J is the objective function of $\mathcal{P}$ and J^* its optimum. $\mathcal{P}_k$ refers to the problem solved at each iteration.

Proof. Let $\eta > 0$ be a smoothing parameter and $\mathcal{P}_\eta$ the associated surrogate problem of $\mathcal{P}$ where the sum of absolute values is replaced by a differentiable term:

$$
\begin{aligned}
(\mathcal{P}_\eta)\ \ \min & \sum_{i\in P} \epsilon_i + \sum_{i\in I}(\epsilon_i^- + \epsilon_i^+) + \lambda \sum_{j>n} \sqrt{m_j^2 + \eta^2} \\
& \text{s.t. } (8), (9), (10), (11)
\end{aligned}
$$

Remarking that $\epsilon_i = (\delta - \langle \mathbf{m}, \delta^{\mathbf{i}}\rangle)_+$ and $\epsilon_i^+ - \epsilon_i^- = \langle \mathbf{m}, \boldsymbol{\delta^i}\rangle$ at the optimum, where $\boldsymbol{\delta^i} = \phi(\mathbf{x^i}) - \phi(\mathbf{y^i})$ and $(x)_+ = \max(0, x)$, $\mathcal{P}_\eta$ can be reformulated in an unconstrained form:

$$
\begin{aligned}
(\mathcal{P}_\eta)\ \ \min & \sum_{i\in P} (\delta - \langle \mathbf{m}, \boldsymbol{\delta^i}\rangle)_+ + \sum_{i\in I} |\langle \mathbf{m}, \boldsymbol{\delta^i}\rangle| \\
& + \lambda \sum_{j>n} \sqrt{m_j^2 + \eta^2} + \mathbb{1}_{\{\langle \mathbf{m}, \mathbf{1}\rangle = 1\}}
\end{aligned}
$$

with $\mathbb{1}_{\{\langle \mathbf{m}, \mathbf{1}\rangle = 1\}} = 0$ if $\langle \mathbf{m}, \mathbf{1}\rangle = 1$ and $+\infty$ otherwise. Then, introducing the smoothing parameter η in Eq.12 yields:

$$\lambda \textstyle\sum_{j>n} \sqrt{m_j^2 + \eta^2} = \min_{\mathbf{z}\geq \frac{\eta}{2}} \frac{\lambda}{2} \sum_{j>n} (\frac{m_j^2 + \eta^2}{z_j} + z_j)$$

which leads to reformulate $\mathcal{P}_\eta$ as a problem involving two blocks of variables $(\mathbf{m}, \mathbf{z})$:

$$(\mathcal{P}_\eta) \quad \min_{\mathbf{m},\mathbf{z}} \; H(\mathbf{m},\mathbf{z}) = g_1(\mathbf{m}) + g_2(\mathbf{z}) + f(\mathbf{m},\mathbf{z})$$

$$\text{with} \begin{cases} f(\mathbf{m},\mathbf{z}) = \frac{\lambda}{2}\sum_{j>n}(\frac{m_j^2+\eta^2}{z_j} + z_j) \\ g_1(\mathbf{m}) \;\; = \sum_{i\in P}(\delta - \langle \mathbf{m}, \boldsymbol{\delta^i}\rangle)_+ + \sum_{i\in I}|\langle \mathbf{m}, \boldsymbol{\delta^i}\rangle| \\ \qquad\quad + \mathbb{1}_{\{\langle \mathbf{m},\mathbf{1}\rangle = 1\}} \\ g_2(\mathbf{z}) \;\; = \mathbb{1}_{\{\mathbf{z} \geq \frac{\eta}{2}\}} \end{cases}$$

Since g_1, g_2 are closed proper convex functions subdifferentiable over their domains $dom\ g_1$ and $dom\ g_2$, and f is convex and continuously differentiable over $dom\ g_1 \times dom\ g_2$ (and $\nabla_z f$ Lispschitz continuous), this problem fits in the class of problem solvable by alternating minimization [Beck, 2015]. Applied on problem $\mathcal{P}_\eta$ and initialized at $\mathbf{m^{(0)}}$ this algorithm is:

$$\mathbf{z}^{(k+1)} \in \operatorname{argmin} g_2(\mathbf{z}) + f(\mathbf{m}^{(k)}, \mathbf{z}) \tag{13}$$

$$\mathbf{m}^{(k+1)} \in \operatorname{argmin} g_1(\mathbf{m}) + f(\mathbf{m}, \mathbf{z}^{(k+1)}) \tag{14}$$

A proof of a non asymptotic sublinear convergence rate of the alternating minimization method in this case is given in [Beck, 2015] and guarantees:

$$\lim_{k\to\infty} H(\mathbf{m}^{(k+1)}, \mathbf{z}^{(k+1)}) - H^* = 0 \tag{15}$$

where H^* is the minimal value of H. Since the optimization problem in Eq. 13 has the closed form solution $z_j^{(k+1)} = \sqrt{m_j^{(k)2} + \eta^2}$, this expression can be inserted in Eq. 14 which yields the following IRLS sequence:

$$\mathbf{m}^{(k+1)} \in \operatorname{argmin} g_1(\mathbf{m}) + \sum_{j>n} \frac{\lambda m_j^2}{\sqrt{m_j^{(k)2} + \eta^2}}$$

$$\in \operatorname{argmin} \sum_{i\in P}\epsilon_i + \sum_{i\in I}(\epsilon_i^- + \epsilon_i^+) + \sum_{j>n} \frac{\lambda m_j^2}{\sqrt{m_j^{(k)2} + \eta^2}}$$
$$\text{s.t. } (8), (9), (10), (11)$$

Let J_η denote the objective function of $\mathcal{P}_\eta$. Using $|x| \leq \sqrt{x^2+\eta^2} \leq |x| + \eta$ as in [Beck, 2015], we obtain:

$$\begin{aligned} J(\mathbf{m}^{(k+1)}) - J^* &= J(\mathbf{m}^{(k+1)}) - J_\eta(\mathbf{m}^{(k+1)}) \\ &+ J_\eta(\mathbf{m}^{(k+1)}) - J_\eta^* + J_\eta^* - J^* \\ &\leq H(\mathbf{m}^{(k+1)}, \mathbf{z}^{(k+1)}) - H^* + (2^n - n)\eta \end{aligned}$$

We conclude by passing to the limit and using Eq.15. □

Proposition 1 ensures that solving problems $\mathcal{P}_k$ for a sufficient number of iterations and a sufficiently small η provides a near-optimal solution to $\mathcal{P}$. The special interest of the IRLS method in our case is revealed when considering the dual formulation of each problem $\mathcal{P}_k$. Indeed, as in kernel-based machine learning methods such as support vector machines [Shawe-Taylor *et al.*, 2004; Waegeman *et al.*, 2009; Tehrani, 2021], one can use Lagrangian duality to obtain a more compact mathematical programming formulation. More precisely, since $\mathcal{P}_k$ is a convex problem with linear constraints, strong duality holds and there is no duality gap. Then solving $\mathcal{P}_k$ or its dual form is equivalent. The efficiency of the dual form of $\mathcal{P}_k$ is detailed in the following proposition:

Proposition 2. *Problem $\mathcal{P}_k$ admits a dual formulation $\mathcal{D}_k$ which has $|P|+|I|+1$ variables and $2(|P|+|I|)$ constraints:*

$$(\mathcal{D}_k) \quad \max_{\boldsymbol{\Gamma}=(\boldsymbol{\alpha},\boldsymbol{\beta},\mu)\in\mathbb{R}^{p+q+1}} -\frac{1}{4\lambda}\boldsymbol{\Gamma}^\intercal \boldsymbol{Q}^\intercal \boldsymbol{D_k}^{-1}\boldsymbol{Q}\boldsymbol{\Gamma} + \boldsymbol{\Gamma}^\intercal \boldsymbol{L}$$
$$\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1}$$
$$-\mathbf{1} \leq \boldsymbol{\beta} \leq \mathbf{1}$$

where $p = |P|$, $q = |I|$ and D_k is a square diagonal matrix of size 2^n whose diagonal contains the current weighting coefficients $1/\sqrt{m_j^{(k)2} + \eta^2}$ (and 0 for the singletons). Also, $\boldsymbol{Q}$ (respectively $\boldsymbol{L}$) is a data dependent matrix of size $2^n \times (p+q+1)$ (respectively $p+q+1$) such that $\begin{cases} \boldsymbol{Q} = (\delta_{\mathbf{P}}, \delta_{\mathbf{I}}, \mathbf{1}) \\ \mathbf{L} = (\boldsymbol{\delta}, \mathbf{0}, 1) \end{cases}$ *where $\delta_{\mathbf{P}} = (\boldsymbol{\delta^i})_{i\in P}$ and $\delta_{\mathbf{I}} = (\boldsymbol{\delta^i})_{i\in I}$ are matrices of size $2^n \times p$ and $2^n \times q$ respectively and where $\boldsymbol{\delta} = \delta(1,\ldots,1) \in \mathbb{R}^p$ and $\mathbf{0} = (0,\ldots,0) \in \mathbb{R}^q$.*

Towards higher dimensions. For a high number of criteria n, the computation of the matrix $\boldsymbol{Q}^\intercal \mathbf{D_k}^{-1}\mathbf{Q}$ raises an issue since $\boldsymbol{Q}$ and $\boldsymbol{D_k}$ have 2^n columns. However, at the first iteration of the IRLS sequence, $\boldsymbol{D_k}$ is the identity matrix and the matrix $\boldsymbol{Q}^\intercal \boldsymbol{D_k}^{-1}\mathbf{Q} = \boldsymbol{Q}^\intercal\mathbf{Q}$ can be computed in polynomial time. In kernel-based machine learning, this property is referred to as the 'kernel trick' [Shawe-Taylor *et al.*, 2004].

As computing the matrix $\boldsymbol{Q}^\intercal\mathbf{Q}$ requires the computation of inner products of the form $\langle \boldsymbol{\phi(x)}, \boldsymbol{\phi(x')}\rangle$, the 'kernel trick' refers to a direct computation that does not require the calculation of vectors $\boldsymbol{\phi(x)}$ (which are of size 2^n here). A computation in $O(n^2)$ is provided for the case of the Choquet integral ($\phi_S(x_S) = \min(x_S)$) in [Tehrani *et al.*, 2014]:

$$\langle \boldsymbol{\phi(x)}, \boldsymbol{\phi(x')}\rangle = \langle \boldsymbol{x}, \boldsymbol{x'}\rangle + \sum_{i=1}^{n-1} x_{(i)} \left\{ \sum_{j=1}^{n-i} 2^{n-i-j} \cdot \min\left\{ x'_{(i)}, x'_{[j+1]_i} \right\} \right\}$$

where $(.)$ is a permutation of N such that $x_{(i)} \leq x_{(i+1)}$ and $[.]_i$ are permutations sorting each vector $(x'_{(i+1)}, \ldots, x'_{(n)})$ by increasing order. This formula can also be used to obtain a polynomial computation of $\langle \boldsymbol{\phi(x)}, \boldsymbol{\phi(x')}\rangle$ when $\phi_S(x_S) = \max(x_S)$ since $\max(x_S) = -\min(-x_S)$. In addition, we provide a polynomial computation of the multilinear kernel ($\phi(x_S) = \prod_{i\in S} x_i$) in the following proposition:

Proposition 3 (See also [Shawe-Taylor *et al.*, 2004])**.** *When $\phi_S(x_S) = \prod_{i\in S} x_i$, we have: $\langle \boldsymbol{\phi(x)}, \boldsymbol{\phi(x')}\rangle = \sum_{S\subseteq N}\prod_{i\in S} x_i \prod_{i\in S} x'_i = \prod_{i=1}^n (x_i x'_i + 1) - 1$.*
Then $\langle \boldsymbol{\phi(x)}, \boldsymbol{\phi(x')}\rangle$ can be computed in $O(n)$.

Taking into consideration these polynomial computations, we propose to proceed to a kernelized computation of matrix $\boldsymbol{Q}^\intercal\mathbf{Q}$ for the first iteration of the dual IRLS method. This provides a way to perform dimension reduction since non significant coefficients obtained after this first iteration can be discarded before going on.

Enforcing monotonicity. Even if monotonicity constraints on the capacity are omitted, it is likely that the learning algorithm captures the monotonicity of the preference examples. It has been observed in practice with the Choquet kernel SVM [Tehrani, 2021] where the learned models achieve low monotonicity violation rates. However, if for normative reasons, we must guarantee that monotonicity w.r.t weak Pareto-dominance holds for all possible alternatives, hard monotonicity constraints must be put on the capacity. For n criteria, there are $C(n) = \sum_{k=1}^{n} k\binom{n}{k}$ monotonicity constraints that read in terms of Möbius masses, as follows: $\sum_{T\subseteq S, T\ni i} m_T \geq 0, \quad \forall i \in S, \forall S \subseteq N$. Including in $\mathcal{P}$ this set of constraints induces a dual problem $\mathcal{D}_k$ with $|P| + |I| + 1 + C(n)$ variables. Thus the dualization benefit is lost and one may prefer a direct solving of $\mathcal{P}$ with LP. Still, the exponential number of variables is an obstacle to scalability and it gets even worse when an exponential number of constraints is added. Hence we propose to handle the monotonicity constraints throughout a generation constraint algorithm that allows an optimal solution to be reached while incorporating only a small portion of the entire set of constraints [Jünger *et al.*, 1993]. The algorithm is initialized with a solution of $\mathcal{P}$ found without monotonicity constraints. Then we iteratively insert the constraints violated in the current solution. The next section presents numerical evidences of the benefits of the proposed method.

4 Numerical Tests

In this section we present the results of numerical tests performed on synthetic preference data. We test the ability of our algorithm (denoted D-IRLS for dual IRLS) to learn a multilinear model or a Choquet integral for a growing number of criteria. We compare it to an exact solving of $\mathcal{P}$ with LP (denoted ES). Preference data are generated through randomly drawn sparse Möbius vectors $\mathbf{m}$ (verifying monotonicity constraints) and utilities vectors x, y are uniformly drawn within $[0, 1]^n$. The overall values $u(x)$ and $u(y)$ are computed and perturbated with a Gaussian noise ($\sigma = 0.03$) before being classified as preference or indifference training examples. We set the size of the training sets to $|P| + |I| = 500$ and of the test sets to $|P| = 1000$. The regularization parameter λ is set to $\lambda = 1$. All tests are conducted on a 2.8 GHz Intel Core i7 processor with 16GB RAM and we used the mathematical programming Gurobi solver (version 9.1.2). For the D-IRLS method, the smoothing parameter is set to $\eta = 10^{-50}$ and the algorithm terminates when $\|\mathbf{m}^{(k+1)} - \mathbf{m}^{(k)}\|_2 \leq 10^{-3}$. Also, coefficients with absolute values smaller than 10^{-5} are discarded at each iteration.

Training time and generalizing performance. In the first experiment, we generate 10 training/test sets and evaluate the average training time of both algorithms as well as the generalizing performances of the learned models (average preference inversion on a test set). In order to evaluate the scalability of our method we vary the number of criteria from $n = 7$ to $n = 22$. Figure 1 shows the results for the learning of the multilinear model (1a,1b) and for the Choquet integral in its conjunctive form (1c,1d). We observe that for both decision models ES does not provide any solution after $n = 17$.

n	$\tilde{C}(n)$	$C(n)$	Time ESG	Time ESC
6	$\mathbf{3.2 \pm 6.4}$	192	0.6 ± 0.2	$\mathbf{0.6 \pm 0.1}$
9	$\mathbf{2.4 \pm 7.2}$	2304	$\mathbf{4.2 \pm 1.9}$	18.0 ± 4.6
12	$\mathbf{151.9 \pm 222.2}$	24576	$\mathbf{61.0 \pm 30.4}$	1212.6 ± 247.6
15	$\mathbf{2777.6 \pm 4326.5}$	245760	$\mathbf{3448.6 \pm 5613.1}$	-

Table 1: $C(n)$,$\tilde{C}(n)$ and training times for ESG and ESC.

However D-IRLS allows more than 4 millions of coefficients ($n = 22$) to be learned in less than 450 seconds. In contrast we observe that the generalizing performances of the learned decision models obtained with D-IRLS and ES are comparable. Since the number of training preference examples is constant, the test error globally increases with the number of criteria for both methods. Finally, we can notice that the test errors obtained for the learning of ML_v are higher than the one obtained for the learning of C_v.

Enforcing monotonicity. In a second experiment, we assess the computational efficiency of the constraint generation algorithm used to guarantee monotonicity. We use the same experimental setting as above and let n vary from 6 to 15. We compare the exact solving of $\mathcal{P}$ under all monotonicity constraints (denoted ESC) with the exact solving of $\mathcal{P}$ with constraint generation (denoted ESG). Both are solved using LP. In Table 1 we compare $C(n)$ the total number of monotonicity constraints in ESC, and $\tilde{C}(n)$ the average number of constraints generated in ESG. We observe that ESC (including all contraints) is slower than ES and limited to $n = 12$. ESG performs significantly better (up to 15 criteria) due to the progressive introduction of monotonicity constraints. We observe that only a small fraction of the entire set of monotonicity constraints are inserted before reaching an optimal and fully monotonic capacity.

Comparison with k-additive models. The advantage of using sparse models with possible large interactions instead of k-additive models is illustrated in Table 2 where we compare our method (D-IRLS) to an exact solving of $\mathcal{P}$ with k-additivity constraints for $k = 2$ (2-add) and $k = 3$ (3-add), still under the same experimental setting. The generalizing performance is significantly improved while computation times remain admissible (and even better for a large n).

	Test Error			Training time		
n	D-IRLS	2-add	3-add	D-IRLS	2-add	3-add
8	**0.04**	0.10	0.06	28.22	**1.33**	1.40
12	**0.04**	0.17	0.23	122.53	**20.96**	21.34
16	**0.06**	0.22	0.49	**187.79**	345.13	346.78

Table 2: Average test error and training time over 10 simulations.

Mixing different models of interactions. Preference learning produces an instance of model $F(x)$ defined by Eq. 7. Möbius coefficients m_S are learned for any given monotonic function ϕ_S defining the nature of interaction terms.

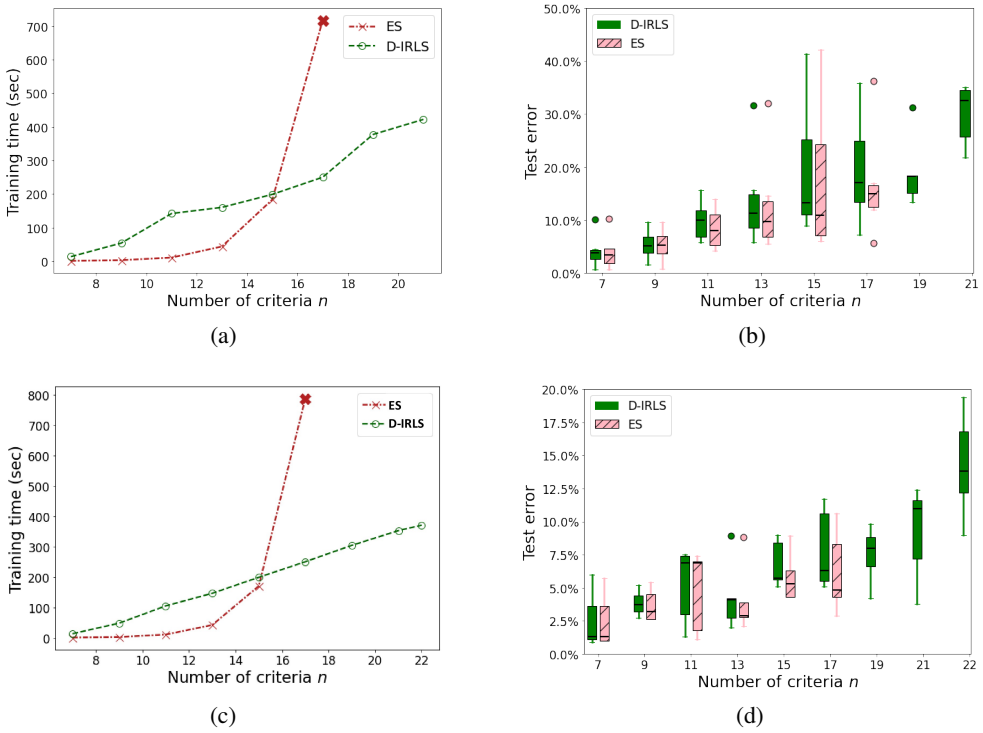

Figure 1: Mean training time and test error boxplot for D-IRLS and ES with multilinear (top) and Choquet Integral (bottom) model.

Several interaction functions may coexist in the same model with a possible benefit in terms of sparsity. As an illustration, we randomly draw preference examples (noised) based on the Hurwicz criterion: $h(x) = \frac{1}{2}(\min_{i\in N}\{x_i\} + \max_{i\in N}\{x_i\})$. Although the min (resp. max) term admits a sparse representation in the conjunctive form (Eq. 5) of the Choquet integral (resp. disjunctive form, Eq. 6) this is not the case of $h(x)$ that includes both terms. This suggests extending the model defined in Eq. 7 to include simultaneously several instances of ϕ_S (like min and max). If we write $v = v^\wedge + v^\vee$ with $(v^\wedge, v^\vee)$ two sub-normalized capacities, we have $C_v(x) = C_{v^\wedge}(x) + C_{v^\vee}(x)$. Then using the conjunctive form for $C_{v^\wedge}(x)$ and the disjunctive form for $C_{v^\vee}(x)$ we obtain: $C_v(x) = \sum_{S\subseteq N}(m_{v^\wedge}(S)\min_{i\in S}\{x_i\} + m_{\bar{v}^\vee}(S)\max_{i\in S}\{x_i\})$. Then we compute a sparse representation of C_v possibly including both conjunctive and disjunctive terms by solving a variant of problem $\mathcal{P}$ using the double penalization term $\lambda_\wedge \sum_{j>n}|m_{v^\wedge}(j)| + \lambda_\vee \sum_{j>n}|m_{\bar{v}^\vee(j)}|$ under the normalization constraint $\sum_{S\subseteq N}(m_{v^\wedge}(S) + m_{\bar{v}^\vee}(S)) = 1$. On Figure 2 we provide the regularization path obtained for increasing values of $\lambda_\wedge = \lambda_\vee$. As expected, a model including only two factors (min and max) is progressively emerging.

5 Conclusion

We have addressed the problem of preference learning with interacting criteria by considering a large class of capacity-based decision models including the multilinear utility and the Choquet integral, known for their expressiveness. We proposed a unified approach to learn the models of this class based on the search of sparse Möbius representations of capacities, leading to simple models with sparse interaction patterns. This approach applies to instances possibly involving more than 20 criteria and allows the most significant interaction factors to be identified within millions of possibilities. This represents a significant improvement compared to previous approaches limited to a dozen of criteria. Moreover, the sparsity pattern is revealed from preference examples instead of resulting from a prior cardinality-based simplification of interactions, which greatly enhances the descriptive possibilities. The main directions to extend this work are 1) going further in scalability (a larger use of kernels is worth investigating) and 2) extending the approach to learn interaction functions ϕ_S from preference data (model selection problem).

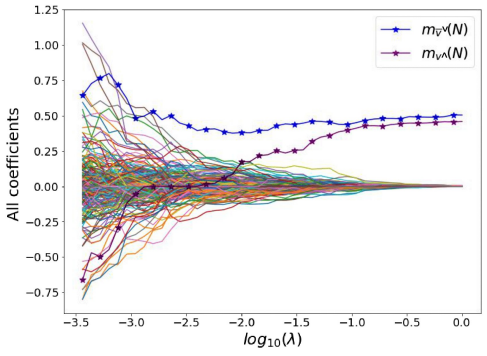

Figure 2: Selection path for the learning of the Hurwicz model.

Acknowledgments

This work is supported by the ANR project ANR-20-CE23-0018 THEMIS of the French National Research Agency.

References

[Adeyeba *et al.*, 2015] Titilope A. Adeyeba, Derek T. Anderson, and Timothy C. Havens. Insights and characterization of l1-norm based sparsity learning of a lexicographically encoded capacity vector for the Choquet integral. In *FUZZ-IEEE*, pages 1 – 7, 2015.

[Ah-Pine *et al.*, 2018] Julien Ah-Pine, Brice Mayag, and Antoine Rolland. Additive bi-capacity by using mathematical programming. In *Third International Conference on Algorithmic Decision Theory (ADT)*, pages 15–29, 2018.

[Anderson *et al.*, 2014] Derek T. Anderson, Stanton R. Price, and Timothy C. Havens. Regularization-based learning of the Choquet integral. In *FUZZ-IEEE*, pages 2519 – 2526, 2014.

[Bach *et al.*, 2012] 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.

[Beck, 2015] Amir Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1):185–209, 2015.

[Beliakov and Wu, 2019] Gleb Beliakov and Jian-Zhang Wu. Learning fuzzy measures from data: simplifications and optimisation strategies. *Information Sciences*, 494:100–113, 2019.

[Benabbou *et al.*, 2017] Nawal Benabbou, Patrice Perny, and Paolo Viappiani. Incremental elicitation of Choquet capacities for multicriteria choice, ranking and sorting problems. *Artificial Intelligence*, 246:152–180, 2017.

[Bresson *et al.*, 2020] Roman Bresson, Johanne Cohen, Eyke Hüllermeier, Christophe Labreuche, and Michèle Sebag. Neural representation and learning of hierarchical 2-additive Choquet integrals. In *IJCAI*, pages 1984–1991, 2020.

[Chateauneuf and Jaffray, 1989] Alain Chateauneuf and Jean-Yves Jaffray. Some characterizations of lower probabilities and other monotone capacities through the use of möbius inversion. *Mathematical social sciences*, 17(3):263–283, 1989.

[Daubechies *et al.*, 2010] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 63(1):1–38, 2010.

[de Oliveira *et al.*, 2022] Henrique Evangelista de Oliveira, Leonardo Tomazeli Duarte, and João Marcos Travassos Romano. Identification of the Choquet integral parameters in the interaction index domain by means of sparse modeling. *Expert Systems with Applications*, 187, 2022.

[Dyer and Sarin, 1979] James S Dyer and Rakesh K Sarin. Measurable multiattribute value functions. *Operations research*, 27(4):810–822, 1979.

[Galand and Mayag, 2017] Lucie Galand and Brice Mayag. A heuristic approach to test the compatibility of a preference information with a choquet integral model. In *Algorithmic Decision Theory - 5th International Conference, ADT*, pages 65–80, 2017.

[Grabisch and Labreuche, 2010] Michel Grabisch and Christophe Labreuche. A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *Annals of Operations Research*, 175(1):247–286, 2010.

[Grabisch *et al.*, 2003] Michel Grabisch, Christophe Labreuche, and Jean-Claude Vansnick. On the extension of pseudo-boolean functions for the aggregation of interacting criteria. *European Journal of Operational Research*, 148(1):28–47, 2003.

[Grabisch *et al.*, 2008] Michel Grabisch, Ivan Kojadinovic, and Patrick Meyer. A review of methods for capacity identification in Choquet integral based multi-attribute utility theory: Applications of the Kappalab R package. *European journal of operational research*, 186(2):766–785, 2008.

[Grabisch *et al.*, 2009] Michel Grabisch, Jean-Luc Marichal, Radko Mesiar, and Endre Pap. *Aggregation functions*, volume 127. Cambridge University Press, 2009.

[Grabisch, 1996] Michel Grabisch. The application of fuzzy integrals in multicriteria decision making. *European journal of operational research*, 89(3):445–456, 1996.

[Grabisch, 1997] Michel Grabisch. K-order additive discrete fuzzy measures and their representation. *Fuzzy sets and systems*, 92(2):167–189, 1997.

[Grabisch, 2016] Michel Grabisch. *Set functions, games and capacities in decision making*. Springer, 2016.

[Grandvalet, 1998] Yves Grandvalet. Least absolute shrinkage is equivalent to quadratic penalization. In *International Conference on Artificial Neural Networks*, pages 201–206. Springer, 1998.

[Herin *et al.*, 2022a] Margot Herin, Patrice Perny, and Nataliya Sokolovska. Learning sparse representations of preferences within Choquet expected utility theory. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

[Herin *et al.*, 2022b] Margot Herin, Patrice Perny, and Nataliya Sokolovska. Learning utilities and sparse representations of capacities for multicriteria decision making with the bipolar Choquet integral. In *Multidisciplinary Workshop on Advances in Preference Handling, IJCAI*, 2022.

[Hüllermeier and Tehrani, 2013] Eyke Hüllermeier and Ali Fallah Tehrani. Efficient learning of classifiers based

on the 2-additive choquet integral. In *Computational Intelligence in Intelligent Data Analysis*, pages 17–29, 2013.

[Hurwicz, 1951] Leonid Hurwicz. The generalized bayes minimax principle: a criterion for decision making under uncertainty. *Cowles Comm. Discuss. Paper Stat*, 335:1950, 1951.

[Jünger *et al.*, 1993] Michael Jünger, Gerhard Reinelt, and Stefan Thienel. Practical problem solving with cutting plane algorithms in combinatorialoptimization. In *Combinatorial Optimization*, 1993.

[Kakula *et al.*, 2020] Siva K Kakula, Anthony J Pinar, Timothy C Havens, and Derek T Anderson. Choquet integral ridge regression. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2020.

[Keeney *et al.*, 1993] Ralph L Keeney, Howard Raiffa, and Richard F Meyer. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.

[Owen, 1975] Guillermo Owen. Multilinear extensions and the banzhaf value. *Naval research logistics quarterly*, 22(4):741–750, 1975.

[Pelegrina *et al.*, 2018] Guilherme D Pelegrina, Michel Grabisch, Leonardo T Duarte, and MT Romano. Multilinear model: New issues in capacity identification. In *From Multiple Criteria Decision Aid to Preference Learning (DA2PL'2018)*, 2018.

[Pelegrina *et al.*, 2020] Guilherme Dean Pelegrina, Leonardo Tomazeli Duarte, Michel Grabisch, and João Marcos Travassos Romano. The multilinear model in multicriteria decision making: The case of 2-additive capacities and contributions to parameter identification. *European Journal of Operational Research*, 282(3):945–956, 2020.

[Pinar *et al.*, 2017] Anthony J. Pinar, Derek T. Anderson, Timothy C. Havens, Alina Zare, and Titilope Adeyeba. Measures of the Shapley index for learning lower complexity fuzzy integrals. *Granul. Comput.*, 2:303 – 319, 2017.

[Schmeidler, 1989] David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571–587, 1989.

[Shawe-Taylor *et al.*, 2004] John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[Tehrani and Hülermeier, 2013] Ali Fallah Tehrani and Eyke Hülermeier. Ordinal Choquistic regression. In *EUSFLAT*, 2013.

[Tehrani *et al.*, 2014] Ali Fallah Tehrani, Marc Strickert, and Eyke Hüllermeier. The choquet kernel for monotone data. In *Esann*, 2014.

[Tehrani, 2021] Ali Fallah Tehrani. The choquet kernel on the use of regression problem. *Information Sciences*, 556:256–272, 2021.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological)*, 58(1):267 – 88, 1996.

[Waegeman *et al.*, 2009] Willem Waegeman, Bernard De Baets, and Luc Boullart. Kernel-based learning methods for preference aggregation. *4OR*, 7(2):169–189, 2009.

[Wakker and Deneffe, 1996] Peter Wakker and Daniel Deneffe. Eliciting von neumann-morgenstern utilities when probabilities are distorted or unknown. *Management science*, 42(8):1131–1150, 1996.