Learning Preference Models with Sparse Interactions of Criteria*

Margot Herin[1] Patrice Perny[1] Nataliya Sokolovska[2]
[1]LIP6 - Sorbonne University [2]LCQB - Sorbonne University

PFIA (Plate-Forme Intelligence Artificielle), June 2023

*To appear in proceedings of IJCAI 2023

Preference Models with Interaction of Criteria

Multicriteria Decision Making:

- set of criteria $N = \{1 \dots n\}$;
- alternatives $x = (x_1, \dots, x_n)$;
- x_i utility of x w.r.t. criterion i, for $i = 1, \dots, n$.
- preference model : $x \succsim y \iff F(x) \geq F(y)$

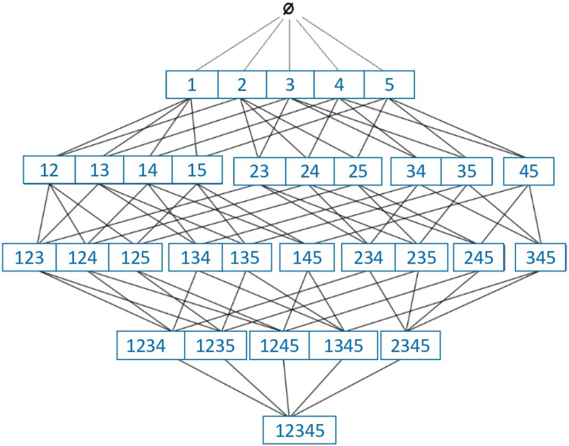
Need for interactions to model natural preferences

The preference for balanced solutions can not be modeled with a weighted sum:

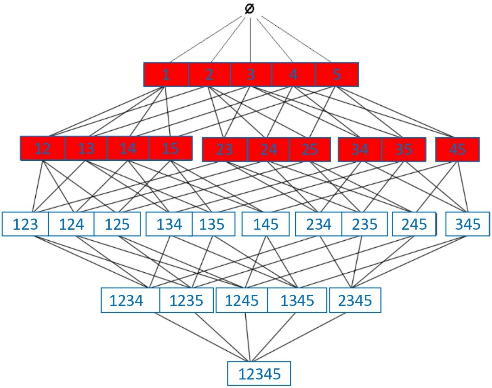
- $(0.5, 0.5) \succ (0, 1)$ and $(0.5, 0.5) \succ (1, 0)$ for $F(x) = \sum_i w_i x_i + \prod_i x_i$ or $F(x) = \sum_i w_i x_i + \min_i\{x_i\}$.

Challenge: combinatorial complexity

n criteria $\Rightarrow 2^n - 1$ possible interactions.



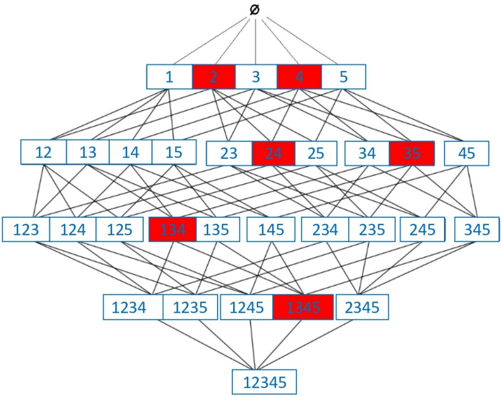
K-additivity



Descriptive limit: discards simple but n-additive models

- $F(x) = \min_i\{x_i\}$ (egalitarian criterion);
- $F(x) = \alpha \min_i\{x_i\} + (1 - \alpha) \max_i\{x_i\}$ ([Hurwicz, 1951]).

Our approach: sparse interactions



Preference Models with Interaction of Criteria

Multilinear model [Keeney et al., 1993]

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

Choquet Integral [Schmeidler, 1989]

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

(.) such that $x_{(i)} \leq x_{(i+1)}$ and $X_{(i)} = \{(i), \ldots, (n)\}$, $i = 1 \ldots n$ with $x_{(0)} = 0$, $X_{(n+1)} = \emptyset$.

Capacity

A capacity is a function $v : 2^N \to [0, 1]$ such that $v(\emptyset) = 0$ and $v(N) = 1$.
v is monotonic w.r.t. set inclusion if $\forall T \subseteq S$, $v(T) \leq v(S)$.

Möbius representation

Möbius transform

$$\forall S \subseteq N, \quad m_v(S) = \sum_{T \subseteq S} (-1)^{|S \setminus T|} v(T)$$

Multilinear model

$$ML_v(x) = \textstyle\sum_{S \subseteq N} m_v(S) \prod_{i \in S} x_i \qquad (1)$$

Choquet Integral

$$C_v(x) = \textstyle\sum_{S \subseteq N} m_v(S) \min_{i \in S}\{x_i\} \qquad \text{conjunctive form} \qquad (2)$$

$$C_v(x) = \textstyle\sum_{S \subseteq N} m_{\bar{v}}(S) \max_{i \in S}\{x_i\} \qquad \text{disjunctive form} \qquad (3)$$

with $\overline{v} : S \to v(N) - v(N \setminus S)$.

Sparse Möbius representation

For any monotonic capacity v, $\|\mathbf{m_v}\|_0 \leq \|\mathbf{v}\|_0$ ($||.||_0$: *number of non-null coefficients*)

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

$$C_v(x) = 0.1x_1 + 0.2x_2 + 0.3x_3 + 0.4\min\{x_1, x_2, x_3\}$$

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

$$C_v(x) = 0.1x_1 + 0.2x_2 + 0.3x_3 + 0.4\max\{x_1, x_2, x_3\}$$

General model

- $\mathbf{ML_v(x)} = \langle \mathbf{m_v}, \phi(\mathbf{x}) \rangle$ with $\phi(x) = (\prod_{i \in S}\{x_i\})_{S \subseteq N}$
- $\mathbf{C_v(x)} = \langle \mathbf{m_v}, \phi(\mathbf{x}) \rangle$ with $\phi(x) = (\min_{i \in S}\{x_i\})_{S \subseteq N}$
- $\mathbf{C_v(x)} = \langle \mathbf{m_{\overline{v}}}, \phi(\mathbf{x}) \rangle$ with $\phi(x) = (\max_{i \in S}\{x_i\})_{S \subseteq N}$

$$F(x) = \sum_{S \subseteq N} m_S \phi_S(x_S) = \langle \mathbf{m}, \phi(\mathbf{x}) \rangle$$

where $\mathbf{m} = (m_S)_{S \subseteq N}$ and $\phi : \mathbb{R}^n \to \mathbb{R}^{2^n}$ maps x into a nonlinear feature space such that $\phi(x) = (\phi_S(x_S))_{S \subseteq N}$.

Contribution: a faster and more scalable algorithm to learn sparse Mobius representations m from preference examples, without prior complexity reduction like k-additivity constraints.

Learning Problem

- $\{(x^i, y^i) \in \mathcal{X}^2 : x^i \succ y^i, i \in P\}$
- $\{(x^i, y^i) \in \mathcal{X}^2 : x^i \sim y^i, i \in I\}$
- $\lambda > 0$: regularization hyper-parameter;
- $\delta > 0$: discrimination threshold;

L_1 Regularization

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

Linear programming

$$(\mathcal{P}) \ \min \quad \sum_{i\in P} \epsilon_i + \sum_{i \in I}(\epsilon_i^- + \epsilon_i^+) + \lambda \sum_{j=n+1}^{2^n} |m_j|$$

$$\langle \mathbf{m}, \phi(\mathbf{x^i})\rangle - \langle \mathbf{m}, \phi(\mathbf{y^i})\rangle + \epsilon_i \geq \delta, \ i \in P \quad (4)$$

$$\langle \mathbf{m}, \phi(\mathbf{x^i})\rangle - \langle \mathbf{m}, \phi(\mathbf{y^i})\rangle + \epsilon_i^+ - \epsilon_i^- = 0, \ i \in I \quad (5)$$

$$\langle \mathbf{m}, \mathbf{1}\rangle = 1 \quad (6)$$

$$\epsilon_i \geq 0, \ i \in P, \quad \epsilon_i^+, \epsilon_i^- \geq 0, \ i \in I \quad (7)$$

$$\Leftrightarrow \quad \min \quad \sum_{i\in P} \epsilon_i + \sum_{i\in I}(\epsilon_i^- + \epsilon_i^+) + \lambda \sum_{j>n}(w_j^+ + w_j^-)$$

$$m_j = w_j^+ - w_j^-, \quad j = n+1, \dots, 2^n$$

$$w_j^+, w_j^- \geq 0, \quad j = n+1, \dots, 2^n$$

$$\text{s.t. } (4), (5), (6), (7)$$

Iterative Reweighted Least Square (IRLS)

Consider the IRLS sequence $\mathbf{m}^{(k)}$ initialized with $\mathbf{m}^{(0)} = \mathbf{1}$ such that:

$$\mathbf{m}^{(k+1)} \in \text{argmin} \sum_{i \in P} \epsilon_i + \sum_{i \in I} (\epsilon_i^- + \epsilon_i^+) + \sum_{j > n} \lambda_j^{(k)} m_j^2$$
$$\text{s.t. } (4), (5), (6), (7)$$

$\mathcal{P}_k$ refers to the problem solved at each iteration.

Then $\mathbf{m}^{(k+1)}$ converges towards the solution of $\mathcal{P}$ in the sense that:
$\lim_{k \to \infty} J(\mathbf{m}^{(k+1)}) - J^* \leq (2^n - n)\eta$ where J is the objective function of $\mathcal{P}$ and J^* its optimum, and η is a smoothing parameter.

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

How does IRLS work?

Variational formulation of the L_1-norm [Bach et al., 2012]

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

$$
\begin{aligned}
(\mathcal{P})\ \ \min_{\boldsymbol{z}\geq \mathbf{0}, \boldsymbol{m}, \boldsymbol{\epsilon}} \quad & \sum_{i\in P} \epsilon_i + \sum_{i\in I} (\epsilon_i^- + \epsilon_i^+) + \frac{\lambda}{2} \sum_{j>n} \left(\frac{m_j^2}{z_j} + z_j\right) \\
& \langle \mathbf{m}, \phi(\mathbf{x^i})\rangle - \langle \mathbf{m}, \phi(\mathbf{y^i})\rangle + \epsilon_i \geq \delta,\ \ i\in P \\
& \langle \mathbf{m}, \phi(\mathbf{x^i})\rangle - \langle \mathbf{m}, \phi(\mathbf{y^i})\rangle + \epsilon_i^+ - \epsilon_i^- = 0,\ \ i\in I \\
& \langle \mathbf{m}, \mathbf{1}\rangle = 1 \\
& \epsilon_i \geq 0,\ \ i\in P, \quad \epsilon_i^+, \epsilon_i^- \geq 0,\ \ i\in I
\end{aligned}
$$

$$\Leftrightarrow \qquad \min_{\mathbf{m},\mathbf{z}} \quad H(\mathbf{m},\mathbf{z}) = g_1(\mathbf{m}) + g_2(\mathbf{z}) + f(\mathbf{m},\mathbf{z})$$

$$
\text{with} \begin{cases}
f(\mathbf{m},\mathbf{z}) = \frac{\lambda}{2} \sum_{j>n} \left(\frac{m_j^2}{z_j} + z_j\right) \\
g_1(\mathbf{m}) \quad = \sum_{i\in P} (\delta - \langle \mathbf{m}, \boldsymbol{\delta^i}\rangle)_+ + \sum_{i\in I} |\langle \mathbf{m}, \boldsymbol{\delta^i}\rangle| + 1_{\{\langle \mathbf{m},\mathbf{1}\rangle = 1\}} \\
g_2(\mathbf{z}) \quad\ = 1_{\{\mathbf{z}\geq \mathbf{0}\}}
\end{cases}
$$

Alternating minimization algorithm

$$(\mathcal{P}_\eta) \qquad \min_{\mathbf{m},\mathbf{z}} \quad H_\eta(\mathbf{m},\mathbf{z}) = g_1(\mathbf{m}) + g_{2\eta}(\mathbf{z}) + f_\eta(\mathbf{m},\mathbf{z}) \qquad \text{(Surrogate problem)}$$

Algorithm (Convergence in $O(1/k)$ [Beck, 2015])

$$\mathbf{m}^{(0)} = \mathbf{1}$$

$$\mathbf{z}^{(k+1)} \in \operatorname{argmin} g_{2\eta}(\mathbf{z}) + f_\eta(\mathbf{m}^{(k)},\mathbf{z}) \tag{8}$$

$$\mathbf{m}^{(k+1)} \in \operatorname{argmin} g_1(\mathbf{m}) + f_\eta(\mathbf{m},\mathbf{z}^{(k+1)}) \tag{9}$$

First step (8) $\Leftrightarrow z_j^{(k+1)} = \sqrt{m_j^{(k)2} + \eta^2}$

$$\Rightarrow \mathbf{m}^{(k+1)} \in \operatorname{argmin} \sum_{i \in P} \epsilon_i + \sum_{i \in I} (\epsilon_i^- + \epsilon_i^+) + \sum_{j > n} \lambda_j^{(k)} m_j^2$$

$$\text{s.t. } (4),(5),(6),(7) \quad \text{with } \lambda_j^{(k)} = \frac{\lambda}{\sqrt{m_j^{(k)2} + \eta^2}}$$

$\lim_{k\to\infty} J(\mathbf{m}^{(k+1)}) - J^* \leq (2^n - n)\eta$ where J is the objective function of $\mathcal{P}$ and J^* its optimum, and η is the smoothing parameter

Efficient dual formulation for $|P| + |I| << 2^n$

Kernel trick

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

Toy example ($|P| = 3, n = 5$)

$$(\mathcal{P}_k) \min_{\mathbf{m}\in\mathbb{R}^{32}} \sum_{i=1}^{3} \epsilon_i + \sum_{j=5}^{32} \lambda_j^{(k)} m_j^2$$

$$\langle \mathbf{m}, \phi(\mathbf{x}^1)\rangle - \langle \mathbf{m}, \phi(\mathbf{y}^1)\rangle + \epsilon_1 \geq \delta$$

$$\langle \mathbf{m}, \phi(\mathbf{x}^2)\rangle - \langle \mathbf{m}, \phi(\mathbf{y}^2)\rangle + \epsilon_2 \geq \delta$$

$$\langle \mathbf{m}, \phi(\mathbf{x}^3)\rangle - \langle \mathbf{m}, \phi(\mathbf{y}^3)\rangle + \epsilon_3 \geq \delta$$

$$\epsilon_1, \epsilon_2, \epsilon_3 \geq 0$$

$$(\mathcal{D}_k) \max_{\boldsymbol{\alpha}\in[0,1]^3} \sum_{i,j=1}^{3} \alpha_i \alpha_j \boldsymbol{\delta}^{i\top} \boldsymbol{D_k}^{-1} \boldsymbol{\delta}^j$$

with $\boldsymbol{\delta}^i = \phi(\mathbf{x}^i) - \phi(\mathbf{y}^i)$

and $\boldsymbol{D_k} = diag((\lambda_j^{(k)})_{j=n}^{2^n})$

Preference Kernel

First iteration:

$$\delta^{i\top} D_k^{-1} \delta^j = \delta^{i\top} \delta^j$$
$$= \langle \phi(\mathbf{x}^i), \phi(\mathbf{x}^j) \rangle + \langle \phi(\mathbf{y}^i), \phi(\mathbf{y}^j) \rangle - \langle \phi(\mathbf{x}^i), \phi(\mathbf{y}^j) \rangle - \langle \phi(\mathbf{y}^i), \phi(\mathbf{x}^j) \rangle$$

Multilinear kernel

$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \sum_{S \subseteq N} \prod_{i \in S} x_i \prod_{i \in S} x'_i = \prod_{i=1}^{n} (x_i x'_i + 1)$$

Choquet kernel [Tehrani et al., 2014]

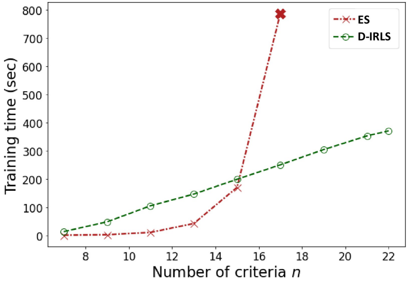
:

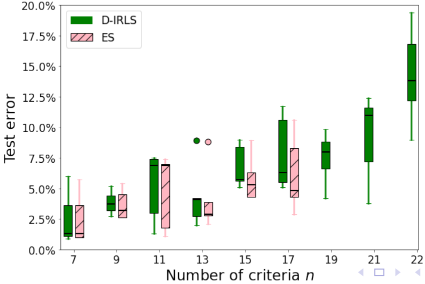
$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \langle \mathbf{x}, \mathbf{x}' \rangle + \sum_{i=1}^{n-1} x_{(i)} \left\{ \sum_{j=1}^{n-i} 2^{n-i-j} \cdot \min \left\{ x'_{(i)}, x'_{[j+1]_i} \right\} \right\}$$

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.

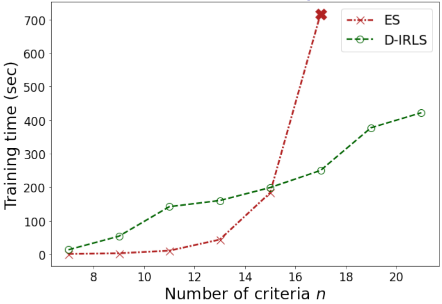
Numerical tests on synthetic data (Choquet Integral)

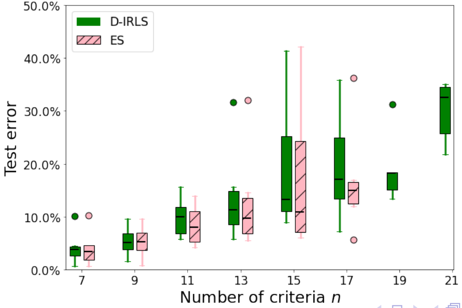
- $n = 4, \ldots, 22$, $|P| + |I| = 500$, 10 simulations





Numerical tests on synthetic data (Multilinear model)





Enforcing monotonicity

Monotonicity constraints

$\sum_{T \subseteq S, T \ni i} m_T \geq 0, \quad \forall i \in S, \forall S \subseteq N$

- Exponential number of constraints ($C(n) = \sum_{k=1}^{n} k\binom{n}{k}$) $\Rightarrow$ dual problem with an exponential number of parameters.
- Direct solving of the problem using constraint generation: in practice a small fraction $\tilde{C}(n)$ of the monotonicity constraints are needed to reach the optimal solution of the fully constrained problem.

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

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

ESG: Exact solving with constraint generation, ESC: Exact solving fully constrained.

Hybrid models

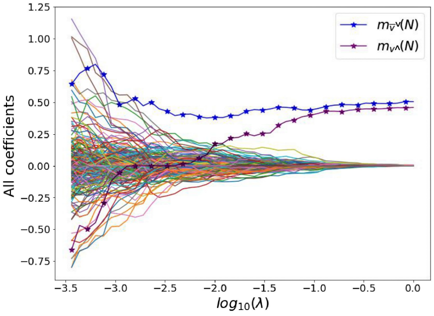
$$F(x) = \sum_{S \subseteq N} m_S \phi_S(x_S)$$

$$F(x) = \sum_{S \subseteq N} m_S^1 \phi_S^1(x_S) + m_S^2 \phi_S^2(x_S) \quad ?$$

Hybrid form of the Choquet Integral

$$v = v^{\wedge} + v^{\vee} \Rightarrow C_v(x) = C_{v^{\wedge}}(x) + C_{v^{\vee}}(x)$$
$$\Rightarrow C_v(x) = \sum_{S \subseteq N} (m_{v^{\wedge}}(S) \min_{i \in S}\{x_i\} + m_{\bar{v}^{\vee}}(S) \max_{i \in S}\{x_i\})$$

Recovering the Hurwicz model ($F(x) = \frac{1}{2}(\min_{i \in N}\{x_i\} + \max_{i \in N}\{x_i\})$)



Conclusion

- Approach to learn a large class of capacity-based decision models;
- Sparsity pattern learned from data (no cardinality-based restriction)
- Applies to instances possibly involving more than 20 criteria (millions of possible interactions)

Perspectives

- Learning the interaction shape in a predefinite set $\{\prod, \min \max\}$;
- Learning general shape: application to GAI models.

References

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.

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.

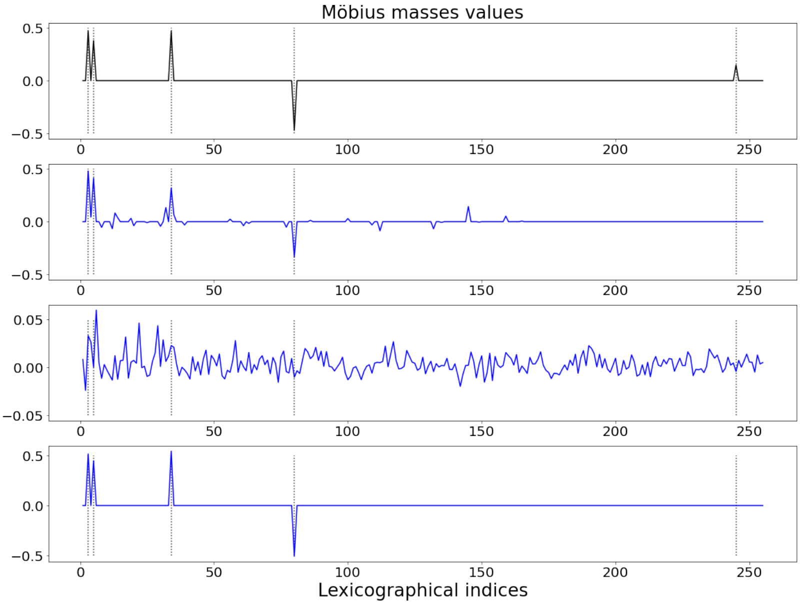
Leonid Hurwicz. The generalized bayes minimax principle: a criterion for decision making under uncertainty. *Cowles Comm. Discuss. Paper Stat*, 335:1950, 1951.

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

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

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

References


Möbius masses values
0.5
0.0
−0.5
0.05
0.00
−0.05
0
50
100
150
200
250
Lexicographical indices

References

- g_1, g_2 are closed proper convex (lower-semi continous) functions sub-differentiable over their domains *dom* g_1 and *dom* g_2
- f is convex and continuously differentiable over *dom* $g_1 \times$ *dom* g_2
- $\nabla_z f$ Lispschitz continuous

Theorem 3.2. *Let* $\{\mathbf{x}_k\}_{k \geq 0}$ *be the sequence generated by the alternating minimization method. Then for any* $k \geq 1$

$$H(\mathbf{x}_k) - H^* \leq \frac{3\max\{H(\mathbf{x}_0) - H^*, \min\{L_1, L_2\}R^2\}}{k}. \quad (3.14)$$