Granul. Comput.
DOI 10.1007/s41066-017-0045-6

ORIGINAL PAPER

Measures of the Shapley index for learning lower complexity fuzzy integrals

Anthony J. Pinar[1] · Derek T. Anderson[2] · Timothy C. Havens[3] · Alina Zare[4] · Titilope Adeyeba[2]

Received: 12 January 2017 / Accepted: 1 June 2017

Abstract The fuzzy integral (FI) is used frequently as a parametric nonlinear aggregation operator for data or information fusion. To date, numerous data-driven algorithms have been put forth to learn the FI for tasks like signal and image processing, multi-criteria decision making, logistic regression and minimization of the sum of squared error (SEE) criteria in decision-level fusion. However, existing work has focused on learning the densities (worth of just the individual inputs in the underlying fuzzy measure (FM)) relative to an imputation method (algorithm that assigns values to the remainder of the FM) or the full FM is learned relative to a single criteria (e.g., SSE). Only a handful of approaches have investigated how to learn the FI relative to some minimization criteria (logistic regression or SSE) in conjunction with a second criteria, namely model complexity. Including model complexity is important because it allows us to learn solutions that are less prone to overfitting and we can lower a solution's cost (financial, computational, etc.). Herein, we review and compare different indices of model (capacity) complexity. We show that there is no global best. Instead, applications and goals (contexts) are what drives which index is appropriate. In addition, we put forth new indices based on functions of the Shapley index. Synthetic and real-world experiments demonstrate the range and behavior of these different indices for decision-level fusion.

Keywords Fuzzy integral · Choquet integral · Fuzzy measure learning · Capacity learning · Regularization · Shapley index

1 Introduction

At the heart of challenges like machine intelligence, robotics, Big Data, geospatial intelligence, and humanitarian demining, to name a few, is the dilemma of data and information fusion, specifically aggregation. A key element of fusion is the underlying mathematics to convert multiple, potentially heterogeneous inputs into fewer outputs (typically one), with the general hope that the aggregation either summarizes the sources well or enhances a system's performance by exploiting interactions across the different sources. A well-known scenario is where "the whole can be worth more than the sum of its parts". Aggregation methods with roots in fuzzy measure theory have been shown to be powerful (Xu and Gou 2017; Das et al. 2017; Xu and Wang 2016; Tang et al. 2017). Herein, we consider the fuzzy integral (FI), specifically Sugeno's fuzzy Choquet integral (CI), for data and information aggregation (Sugeno 1974).

✉ Anthony J. Pinar
ajpinar@mtu.edu

Derek T. Anderson
anderson@ece.msstate.edu

Timothy C. Havens
thavens@mtu.edu

Alina Zare
azare@ece.ufl.edu

[1] Department of Electrical and Computer Engineering, Michigan Technological University, Houghton, USA

[2] Department of Electrical and Computer Engineering, Mississippi State University, Starkville, USA

[3] Department of Electrical and Computer Engineering and Department of Computer Science, Michigan Technological University, Houghton, USA

[4] Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA

Published online: 12 June 2017

The CI is an aggregation operator generator as it is parameterized by the fuzzy measure (FM) (Sugeno 1974), aka a monotone and normal capacity. The FM is defined over the power set of N unique inputs and assigns a "worth" to each subset. It is in this way the CI provides a granular approach to aggregation—each subset represents a unique granule of the universe (Yao 2005; Pawlak 1998), and the CI aggregates contributions over a certain set of these granules.[1] This granulation leads to an extremely flexible aggregation operator that is capable of producing numerous useful aggregation operators such as the minimum, maximum, median, mean, soft maximum (minimum), other linear order statistics, etc. (Tahani and Keller 1990), but also a wealth of other more unique and tailored aggregation operators. The point is, the CI is a powerful non-linear function that is often used for fusion.

It is important to note that the CI is not trivial in any respect. The capacity has $2^N - 2$ free parameters, for N inputs, that must be either specified or learned. This exponential number can (and often does) impact an application rather quickly. For example, 10 inputs already gives rise to 1022 values that must be specified or learned (and 5110 monotonicity constraints at that). The reader can refer to Sugeno (1974), Anderson et al. (2014a) and Grabisch et al. (1995, 2000) for reviews of analytical methods for specifying a capacity relative to just knowledge about the worth of the individual sources (called densities, $g(\{x_i\})$ for $i \in \{1, 2, \ldots, N\}$). Examples of FMs include the Sugeno λ-FM (Sugeno 1974), S-decomposable FMs (Anderson et al. 2014a; Klir and Yuan 1995), belief (and plausibility) and possibility (and necessity) FMs (Klir and Yuan 1995), and k-additive approaches (which model a limited number of capacity terms up to a k-tuple number) (Grabisch 1997).

To date, numerous methods have been proposed to learn the CI. While similar in underlying objective, these methodologies can and often do vary greatly with respect to factors like application domain, mathematics and how the CI is being used (e.g., signal and image processing, regression, decision-level fusion, etc.). In Grabisch et al. (1995, 2000), Grabisch proposed quadratic programming (QP) to acquire the full capacity based on the idea of minimizing the sum of squared error (SSE). In Keller and Osborn (1996), Keller et al. used gradient descent and then penalty and reward (Keller and Osborn 1995) to learn the densities in combination with the Sugeno λ-FM. In Beliakov (2009), Beliakov used linear programming and, in Anderson et al. (2010), a genetic algorithm was used to learn a higher-order (type-1) fuzzy set-valued capacity for

[1] Due to a sorting operation in the CI, the set of granules over which the CI aggregates is defined by a particular input. See Definition 3 for details.

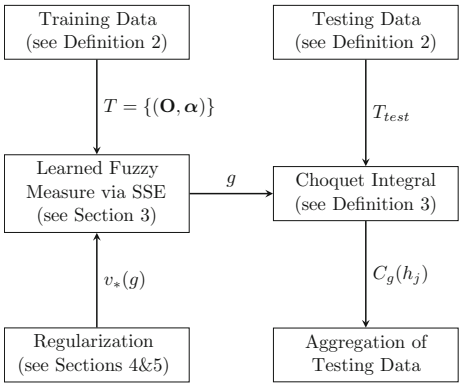


Fig. 1 Summary of the process of learning the fuzzy measure for aggregation via the Choquet integral. The novelty of this work lies in the regularization functions described in Sects. 4 and 5

the Sugeno integral (SI). Alternatively, the works Wagner and Anderson (2012), Havens et al. (2013, 2015) proposed different ways to automatically acquire, and subsequently aggregate, full capacities of specificity and agreement based on the idea of crowd sourcing when the worth of the individuals is not known but is instead extracted from the data. The reader can refer to Grabisch et al. (2008) for a detailed review of other existing work prior to 2008.

An underrepresented and unsolved challenge is learning the CI with respect to more than one criteria—a process summarized in Fig. 1. Herein, we focus on minimization of the SSE criteria, but do it in conjunction with model complexity. In Mendez-Vazquez and Gader (2007), are the first that we are aware to study the inclusion of an information-theoretic index of capacity-complexity in learning the CI. Specifically, Mendez-Vazquez and Gader explored the task of sparsity promotion in learning the CI and provided examples in decision level fusion. Their work has two parts: the Gibbs sampler and the exploration of a lexicographically encoded capacity vector as the ℓ_p-norm complexity term. The goal of their regularization term was to explicitly reduce the number of nonzero parameters in the capacity to eliminate non-informative or "useless" information sources. In Anderson et al. (2014b), the idea of learning the CI based on the use of a QP solver and a lexicographically encoded capacity vector was also explored. The novelty in that work is the study of different properties of the regularization term in an attempt to unearth what it was promoting in different scenarios (with respect to both the capacity but also the resultant aggregation operator induced by the CI). In the theme of information theoretic measures of capacities, but not regularization with respect to such indices, Kojadinovic et al. (2005) and Yager (2000, 2002) both explored the concept of the entropy of an FM. Furthermore, in Labreuche (2008), explored the identification of an FM with an ℓ_1 entropy. Labreuche proposed a linearized entropy of an

Table 1 Notation

N	Number of information sources
X	Set of information sources, $X = \{x_1, \ldots, x_N\}$
m	Number of training data elements
T	Training data, $T = \{(O_j, \alpha_j) : j = 1, \ldots, m\}$
$h_j(x_i)$	$\Re$-valued integrand with respect to source i and object O_j
g	Fuzzy measure (aka normal and monotone capacity)
$\int h_j \circ g = C_g(h_j)$	Choquet integral with respect to capacity g and O_j
$g_{i_1,\ldots,i_k}$	Lexicographic ordering of g, i.e., $g(\{x_{i_1}, \ldots, x_{i_k}\})$
$\mathbf{u}$	Lexicographic ordered capacity vector, $\mathbf{u}^t = (g_1, \ldots, g_{12\ldots N})^t$
$\|\mathbf{u}\|_p^2$	ℓ_p-norm of $\mathbf{u}$ (and thus g)
$\Phi_g(i)$	Shapley value of source i, where $\sum_{i=1}^N \Phi_g(i) = 1$
$\zeta_{X,1}(K) = \frac{(\lvert X\rvert - \lvert K\rvert - 1)!\lvert K\rvert!}{\lvert X\rvert!}$	Shapley normalization term
$\mathbf{I}_g(i,j)$	Interaction index of sources i and j, where $\mathbf{I}_g(i,j) \in [-1, 1]$
$-\sum_{i=1}^N \Phi_g(i) ln(\Phi_g(i))$	Shannon entropy of the Shapley values
$1 - \sum_{i=1}^N \Phi_g(i)^2$	Gini–Simpson index (coefficient) of the Shapley values
$-\sum_{i=1}^N \left(\sum_{K \subseteq X \setminus \{i\}} \zeta_{X,1}(K) \lvert A_i - \frac{1}{N}\rvert\right)$	Labreuche's entropy of g, where $A_i = g(K \cup \{i\}) - g(K)$
$-\sum_{i=1}^N \left(\sum_{K \subseteq X \setminus \{i\}} \zeta_{X,1}(K) A_i ln(A_i)\right)$	Marichal's entropy of g, where $A_i = g(K \cup \{i\}) - g(K)$
$\frac{1}{N}\sum_{i=1}^N \left(\sum_{K \subseteq X \setminus \{i\}} \zeta_{X,1}(K) \left(A_i - \frac{1}{N}\right)^2\right)$	Kojadinovic's variance of g, where $A_i = g(K \cup \{i\}) - g(K)$
$\sum_{A \subseteq X} f(\lvert A\rvert) \lvert \mathcal{M}(A)\rvert$	k-additive index of Mobius transform of g
Γ_i	ith lexicographically encoded Shapley coefficient vector
$1 - \sum_{i=1}^N \left(\Gamma_i^t \mathbf{u}\right)^2$	Gini–Simpson index of Shapley values
$\lVert\Phi_g\rVert_0 = \lvert\{i : \Phi_g(i) \neq 0\}\rvert$	ℓ_0-norm of Shapley vector, $\Phi_g = \left(\Phi_g(1), \ldots, \Phi_g(N)\right)^t$
$\lVert\Phi_g\rVert_1 = \sum_{i=1}^N \lvert\Phi_g(i)\rvert = \sum_{i=1}^N \Phi_g(i) = 1$	ℓ_1-norm of Shapley vector, $\Phi_g = \left(\Phi_g(1), \ldots, \Phi_g(N)\right)^t$

FM, allowing for the identification of an FM using linear programming.

For the most part, it appears that the vast majority of these works are largely unaware of each other. Therefore, in Sect. 2.2 we bridge this gap by analyzing and comparing different properties of these indices. In general, there does not appear to be a clear "winner". That is, different indices exist and are useful for various applications and goals (contexts). Therefore, it is important to understand each index and ultimately what context it supports.

In addition to our review of existing regularization indices, we put forth new indices based on the Shapley values as discussed in Sects. 4 and 5. These new indices comprise this work's novelty, and intend to promote "simpler models" that have either lower diversity in the Shapley values or fewer numbers of inputs (fewer non-zero Shapley terms). In fields such as statistics and machine learning, such a strategy can help with addressing challenges like preventing overfitting (aka improving the generalizability of a learned model). It is also the case that we are often concerned with problems having too many inputs, as more inputs are typically associated with higher cost—e.g., greater financial cost of different sensors, more computational or memory resources, time, or even physical cost in a health setting where an input is something like the result of a bone marrow biopsy.

This article is structured as follows: Sect. 2 is a review of important concepts in FM and FI theory. Section 2.1 discusses the Shapley and interaction indices and Sect. 2.2 reviews and compares existing indices for measuring different notions of complexity of a capacity. In Sect. 2.3, the ℓ_0-norm, ℓ_1-norm and Gini–Simpson index of the Shapley values are proposed. In Sect. 3 we propose the novel methods for optimizing the CI relative to the SSE criteria based on the QP. We propose Gini–Simpson index-based regularization of the Shapley values in Sects. 4 and 5 discusses ℓ_0-norm based regularization, and Sect. 6 contains experiments illustrating different scenarios encountered in practice. Table 1 is notation used in this article.

2 Fuzzy measure and integral

The aggregation of data/information using the FI has a rich history. Much of the theory and several applications can be found in Anderson et al. (2014a), Grabisch et al.

(1995, 2000), Tahani and Keller (1990), Yang et al. (2008), Cho and Kim (1995), Melin et al. (2011) and Wu et al. (2013). There are a number of (high-level) ways to describe the FI, e.g., motivated by Calculus, signal processing, pattern recognition, fuzzy set theory, etc. Herein, we set the stage by considering a finite set of N sources of information, $X = \{x_1, \ldots, x_N\}$, and a function h that maps X into some domain (initially [0, 1]) that represents the partial support of a hypothesis from the standpoint of each source of information. Depending on the problem domain, X can be a set of experts, sensors, features, pattern recognition algorithms, etc. The hypothesis is often thought of as an alternative in a decision process or a class label in pattern recognition. Both Choquet and Sugeno integrals take partial support for the hypothesis from the standpoint of each source of information and fuse it with the (perhaps subjective) worth (or reliability) of each subset of X in a nonlinear fashion. This worth is encoded in an FM (aka capacity). Initially, the function h (integrand) and FM ($g : 2^X \rightarrow [0, 1]$) took real number values in [0, 1]. Certainly, the output range for the support function and capacity can be (and have been) defined more generally, e.g., $\Re^+$, but it is convenient to think of them on [0, 1] for confidence fusion. We now review the capacity and FI.

Definition 1 (*Fuzzy Measure* Sugeno 1974) The FM is a set function, $g : 2^X \rightarrow \Re^+$, such that

P1.
(Boundary condition) $g(\phi) = 0$ (often $g(X) = 1$);
P2.
(Monotonicity) If $A, B \subseteq X$ and $A \subseteq B$, $g(A) \leq g(B)$.

Note, if X is an infinite set, a third condition guaranteeing continuity is required, but this is a moot point for finite X. As already noted, the FM has 2^N values; actually, $2^N - 2$ "free parameters" as $g(\phi) = 0$ and $g(X) = 1$. Before a definition can be given for the FI, notation must be established for the training data used to learn the capacity.

Definition 2 (*Training Data*) Let a training data set, T, be

$$T = \{(O_j, \alpha_j) | j = 1, \ldots, m\}$$

where $\mathbf{O} = \{O_1, \ldots, O_j, \ldots, O_m\}$ is a set of "objects" and α_j are their corresponding labels (specifically, $\Re$-valued numbers). For example, O_j could be the strengths in some hypothesis from N different experts, signal inputs at time j, algorithm outputs for input j, kernel inputs or kernel classifier outputs for feature vector j, etc. Subsequently, α_j could be the corresponding function output, class label, membership degree, etc. Next, we provide a definition for the FI, namely the CI with respect to T. To that end, let h_j be the jth integrand, i.e., $h_j(x_i)$ is the input for the ith source with respect to object j.

Definition 3 The discrete CI, for finite X and object O_j is

$$\begin{aligned} \int h_j \circ g &= C_g(h_j) \\ &= \sum_{i=1}^{N} [h_j(x_{\pi_j(i)}) - h_j(x_{\pi_j(i+1)})] g(A_{\pi_j(i)}), \end{aligned} \tag{1}$$

for $A_{\pi_j(i)} = \{x_{\pi_j(1)}, \ldots, x_{\pi_j(i)}\}$ and permutation π_j such that $h_j(x_{\pi_j(1)}) \geq \cdots \geq h_j(x_{\pi_j(N)})$, where $h_j(x_{\pi_j(N+1)}) = 0$ (Sugeno 1974).

2.1 Shapley and interaction indices

The CI is parameterized by the capacity. Specifically, the capacity encodes all of the rich tuple-wise interactions between the different sources and the CI utilizes this information to aggregate the inputs (the integrand, h). It is important to note that the CI operates on a weaker (and richer) premise than a great number of other aggregation operators that assume additivity (a stronger property than monotonicity). However, the capacity has a large number of values. It is not trivial to understand a capacity. For example, a commonly encountered question is what is the so-called worth of a single individual source? Information theoretic indices aid us in summarizing information such as this in the capacity. The point is, most of our questions are not about a single capacity value; we are interested in a complex question whose answer is dispersed across the capacity. For example, the Shapley index has been proposed to summarize the so-called worth of an individual source and the interaction index summarizes interactions between different sources.

Definition 4 (*Shapley Index*) The Shapley values of g are

$$\Phi_g(i) = \sum_{K \subseteq X \backslash \{i\}} \zeta_{X,1}(K)(g(K \cup \{i\}) - g(K)), \tag{2a}$$

$$\zeta_{X,1}(K) = \frac{(|X| - |K| - 1)!|K|!}{|X|!}, \tag{2b}$$

where $K \subseteq X \backslash \{i\}$ denotes all proper subsets from X that do not include source i. The Shapley value of g is the vector $\Phi_g = (\Phi_g(1), \ldots, \Phi_g(N))^t$ and $\sum_{i=1}^{N} \Phi_g(i) = 1$. The Shapley index can be interpreted as the average amount of contribution of source i across all coalitions. Equation (2a) makes its decision based on the weighted sum of (positive-valued) numeric differences between consecutive steps (layers) in the capacity.

Remark 1 It is important to note the following property. When $g(A) = 0, \forall A \subset X$, the CI is the minimum operator. The Shapley values are $\Phi_g(1) = \Phi_g(2) = \cdots = \Phi_g(N)$. This is easily verifiable as the Shapley value is a weighted sum of differences between $g(X)$ and $g(X \backslash x_i)$ (one of our

inputs). Thus, each Shapley value reduces to the same calculation, $\zeta_{X,1}(K)$, where $K \in 2^X$ and $|K| = |X| - 1$.

Definition 5 (*Interaction Index*) The interaction index (Murofushi and Soneda 1993) between i and j is

$$\mathbf{I}_g(i,j) = \sum_{K \subseteq X \setminus \{i,j\}} \zeta_{X,2}(K)(g(K \cup \{i,j\}) - g(K \cup \{i\}) - g(K \cup \{j\}) + g(K)), \quad i = 1, \ldots, N, \tag{3a}$$

$$\zeta_{X,2}(K) = \frac{(|X| - |K| - 2)!|K|!}{(|X| - 1)!}, \tag{3b}$$

where $\mathbf{I}_g(i,j) \in [-1,1], \forall i,j \in \{1,2,\ldots,N\}$. A value of 1 (respectively, -1) represents the maximum complementarity (respective redundancy) between i and j. The reader can refer to Grabisch and Roubens (2000) for further details about the interaction index, its connections to game theory and interpretations. Grabisch later extended the interaction index to the general case of any coalition (Grabisch and Roubens 2000).

Definition 6 (*Interaction Index for coalition A*) The interaction index for any coalition $A \subseteq X$ is

$$\mathbf{I}_g(A) = \sum_{K \subseteq X \setminus A} \zeta_{X,3}(K,A) \sum_{C \subseteq A} (-1)^{|A \setminus C|} g(C \cup K), \quad i = 1, \ldots, N, \tag{4a}$$

$$\zeta_{X,3}(K,A) = \frac{(|X| - |K| - |A|)!|K|!}{(|X| - |A| + 1)!}. \tag{4b}$$

Equation (4a) is a generalization of both the Shapley index and Murofushi and Soneda's interaction index as $\Phi_g(i)$ corresponds with $\mathbf{I}_g(\{i\})$ and $\mathbf{I}_g(i,j)$ with $\mathbf{I}_g(\{i,j\})$.

While the Shapley and interaction indices are extremely useful, they do not, in their current explicit form, inform us about capacity complexity. In the next subsection, we review additional information theoretic capacity indices and we discuss their interpretations.

2.2 Existing indices for capacity complexity

Excluding indices that are subsumed by others, the bottom line is various indices exist for different reasons. First, some indices are simpler computationally while others are mathematically simpler in terms of our ability to manipulate and use them for tasks like optimization. Second, and arguably the most important, complexity can and often does mean different things to different people/applications. As we discuss in this article, there is no clear winner (i.e., index). Different indices are important for different applications and knowledge of their existence and associated benefits is what is ultimately important. In this section we review existing information theoretic indices for complexity.

Definition 7 (ℓ_1-*Norm of a Lexicographically Encoded Capacity Vector*) Let $\mathbf{u} \in \mathbb{R}^{2^N-1}$ be $\mathbf{u} = (g_1, g_2, \ldots, g_{12}, g_{13} \ldots, g_{12\ldots N})^t$. Note that we define this ordering such that it is also sorted by cardinality.[2] A relatively simple index of the complexity of g is

$$\upsilon_{\ell 1}(g) = \sum_{j=1}^{2^N-1} |\mathbf{u}_j| = \sum_{j=1}^{2^N-1} \mathbf{u}_j. \tag{5}$$

As stated in Anderson et al. (2014b) and Mendez-Vazquez and Gader (2007), the intent of $\upsilon_{\ell 1}(g)$ was to help reduce the number of nonzero parameters in the capacity to eliminate non-informative or useless information sources. However, this index is not as sophisticated as desired. The index is minimized when all $\mathbf{u}_j$ are equal to zero, i.e., the FM $g(A) = 0, \forall A \subset X$, which is a minimum operator for the CI (Tahani and Keller 1990). We also note that this is an FM of "ignorance", as we assert that the answer resides in X, however we have assigned $g(A) = 0$ to all subsets outside X. The index is maximized for the FM $g(A) = 1, \forall A \subseteq X$, which is a maximum operator for the CI (Tahani and Keller 1990). There are really two problems with this index. First, it does not take advantage of any capacity summary mechanism like the Shapley index, interaction index or k-additivity. Second, it is well-known that the ℓ_1 is (geometrically) inferior to the ℓ_0 when it comes to promoting sparsity. However, the ℓ_1-norm gives rise to convex problems that we can more easily solve for while the later does not.

Definition 8 (*Marichal's Index*) Marichal's Shannon-based entropy of g (Marichal 1998, 2000),

$$\nu_M(g) = (-1) \sum_{j=1}^{N} \left(\sum_{K \subseteq X \setminus \{j\}} \zeta_{X,1}(K)(g(K \cup \{j\}) - g(K)) \times \ln(g(K \cup \{j\}) - g(K)) \right), \tag{6a}$$

is motivated in terms of the following (Kojadinovic 2006). Consider the set of all maximal chains of the Hasse diagram $(2^N, \subseteq)$. A maximal chain in $(2^N, \subseteq)$ is a sequence

$$\phi, \{x_{\pi(1)}\}, \{x_{\pi(1)}, x_{\pi(2)}\}, \ldots, \{x_{\pi(1)}, \ldots, x_{\pi(N)}\},$$

where π is a permutation of N. On each chain, we can define a "probability distribution",

[2] As an example, consider the case of $N = 3$ where the vector $\mathbf{u}$ is $\mathbf{u} = (g_1, g_2, g_3, g_{12}, g_{13}, g_{23}, g_{123})^t$.

$$p_\pi^g(i) := g(\{x_{\pi(i)}, \ldots, x_{\pi(N)}\}) - g(\{x_{\pi(i+1)}, \ldots, x_{\pi(N)}\}),$$
$$i = 1, \ldots, N, \pi \in \Pi_N.$$

It is not entirely clear to us why this is called a probability distribution. For example, it is confusing why this is the case for a Belief measure, a Possibility measure, etc. We assume it is interpreted as such due to the properties of positivity and the distribution sums to 1. Furthermore, Kojadinovic (2006) states that "...the intuitive notion of uniformity of a capacity g on N can then be defined as an average of the uniformity values of the probability distributions" (distributions provided according to $p_\pi^g(i)$) (Kojadinovic 2006). Regardless, this account of entropy is the average of the uniformity values of the underlying probability distributions. In general, such an index can be of help with respect to the maximum entropy principle. Furthermore, maximization of index $v_M(g)$ is non-linear and not quadratic (Labreuche 2008). As stated in Labreuche (2008), we can obtain a quadratic problem under linear constraints, considering a special case of Renyi entropy.

It is trivial to prove that minimum entropy for Eq. (6a) occurs if and only if $g(K \cup \{j\}) - g(K)$ yields values in $\{0, 1\}$. Note, Eq. (6a) is defined for

$$g(K \cup \{j\}) - g(K) = 0$$

by choosing 0. While many properties of this definition of entropy are discussed in Kojadinovic (2006), a few important properties were not discussed. First, there is not a single unique "solution" (minimum). That is, an FM of all 0s (minimum operator) and an FM of all 1s (maximum operator) both satisfy this criteria. There are other FMs that satisfy this criteria as, e.g., the N different order-statistics where a single input becomes the output and all other inputs are discarded (one input has a Shapley value of 1 and all other inputs have a Shapley value of 0). Also, there is the case of the ordered weighted average (OWA) (Yager 1988). An OWA is a special case of the CI when the FM is defined such that sets of equal cardinality have equal measure. The OWA weights are simply the differences between the constant-cardinality tuple values, i.e., $w_i = g(A_i) - g(A_i \backslash \{i\})$. For N inputs, we have N such OWAs that yield the mentioned minimum—capacities with values of 1 for all sets $A \subseteq X$ with $|A| \geq k$ and 0 otherwise, for $k = 1 \cdots N$. Note, two of these N cases are the maximum and minimum aggregation operators. On the other hand, maximum entropy occurs in the case of a "uniform distribution" (all $\frac{1}{N}$ values). This only occurs in the case of a capacity in which $g(A) = |A|/|X|$, which is a CI-based average operator. This uniqueness of the maximization case was one of the motivating reasons for the proposal of Marichal's index (maximum entropy principle).

Definition 9 (*Shannon's Entropy of the Shapley Values*) In Anderson et al. (2014b), a related but different formulation of Shannon's entropy was explored in terms of the Shapley values,

$$v_S(g) = (-1) \sum_{j=1}^{N} \Phi_g(j) \ln\big(\Phi_g(j)\big). \tag{7}$$

Note, the Shapley index values sum to 1, i.e.,

$$\sum_{j=1}^{N} \Phi_g(j) = 1.$$

Furthermore, Eq. (7) is not defined for $\Phi_g(j) = 0$; it is by choosing $\ln\big(\Phi_g(j)\big) = 0$. When only one source is needed, i.e., one source is consistently superior to all others, a single Shapley value is 1 and the others are 0, i.e., Equation (7) equals 0. There are N such unique cases. There are also not any such cases in which the Shapley values are all 0s or 1s (by definition). On the other hand, the more uniformly distributed the Shapley values become, the more inputs are required (each are important relative to solving the task at hand). In the extreme case, as when all Shapley values are $\frac{1}{N}$, all sources are needed and we obtain maximum entropy. This occurs when g causes the CI to reduce to an OWA and there is an infinite set of such capacities/OWAs (for a real-valued FM).

In summary, there are fewer and more easily rationalized solutions for Eqs. (7) versus (6a) in the case of minimizing the entropy of a capacity and the latter has fewer solutions for maximizing entropy. However, while there are more solutions in the case of Eq. (7), they can easily be rationalized (all such capacities treat the inputs as equally important in terms of the CI). These two definitions of entropy are similar but not equivalent.

Definition 10 Kojadinovic's variance of g is (Kojadinovic 2006)

$$v_K(g) = \frac{1}{N} \sum_{j=1}^{N} \Bigg(\sum_{K \subseteq X \backslash \{j\}} \zeta_{X,1}(K) \big(g(K \cup \{j\}) - g(K) - \frac{1}{N}\big)^2 \Bigg). \tag{8a}$$

It is trivial to verify that this index equals 0 if and only if the differences between the capacity terms equal $\frac{1}{N}$. This is unique in the fact that it only occurs in the case of $g(A) = |A|/|X|$ (i.e., a CI that reduces to the average operator). As Kojadinovic discusses, Eq. (8a) is a simpler way (versus Marichal's index which involves logarithms) to measure the uniformity of a distribution. Also, Eq. (8a) equates to 0 if and only if we have the case of a "uniform distribution".

Kojadinovic's goal, in the theme of Marichal's notion of entropy, is that of maximum entropy—the "least specific" capacity compatible with the initial preferences of a decision maker. Kojadinovic's variance is maximized in the case that the difference of the two capacity terms equals 0 or 1. This occurs in the case of a minimum operator, maximum operator, or the other $(N-2)$ OWAs discussed in the case of Marichal's entropy. Thus, Kojadinovic's variance and Marichal's entropy are tightly coupled, while Eq. (7) is once again different in its design and set of relevant solutions.

Definition 11 Labreuche's linearized entropy of g is (Labreuche 2008)

$$v_L(g) = (-1)\sum_{j=1}^{N}\left(\sum_{K\subseteq X\setminus\{j\}}\zeta_{X,1}(K)\left|g(K\cup\{j\}) - g(K) - \frac{1}{N}\right|\right). \tag{9a}$$

The primary goal of this index is to linearize Kojadinovic's index to assist in optimization (apply linear programming). Labreuche's goal was to also satisfy, with respect to the different probability distributions, symmetry (value regardless of input permutation), maximality and minimality (probability distribution of all $\frac{1}{N}$ values and the distribution of all zeros with a single value of one). Kojadinovic's index does not satisfy the last two properties. Furthermore, index $v_L(g)$ has a single minimum, the capacity $g(A) = \frac{|A|}{|X|}$, which results in the CI becoming the mean operator. Labreuche's index also has a single maximum, one for each probability distribution. In terms of the capacity, this equates to the minimum operator, the maximum operator, and the other $(N-2)$ OWAs discussed for Kojadinovic's index.

Definition 12 The k-additive based index is (Tehrani et al. 2012; Tehrani and Hüllermeier 2013; Tehrani 2013)

$$v_T(g) = \sum_{A\subseteq X} f(|A|)|\mathcal{M}(A)|, \tag{10}$$

where f is a strictly increasing function defined on the cardinality of subsets of X and $\mathcal{M}$ is the Mobius transform of g (Grabisch et al. 2000; Grabisch and Roubens 2000). The Mobius transform of g is used here to highlight and exploit k-additivity, i.e., $\mathcal{M}(B) = 0, \forall B \subseteq X$ with $|B| > k$. This is a different approach as k-additivity allows for what could be considered a "compact" representation of g (under a set of restrictions) to combat the otherwise combinatorial explosion of g: $\sum_{i=1}^{k}\binom{N}{i}$ terms versus 2^N. In summary, $v_T(g)$ favors the restriction that capacities have a low level of nonadditivity.

It is well-known that the sum of the Mobius terms for the capacity is equal to one (Beliakov et al. 2008). However, $v_T(g)$ considers the sum of the absolute values of the Mobius terms. It is trivial to prove that $v_T(g)$ has a single maximum for the case of a capacity of all ones, $g(A) = 1, \forall A \subseteq X$, i.e., the maximum operator. Although these values sum to one, they can be any value in $[-1, 1]$. This index does not have a unique minimum. For example, a capacity of all zeros (except $g(X)$) has an index value of 1, the mean operator, $g(A) = \frac{|A|}{|X|}$, has an index of 1, and a capacity where a single input has a Shapley value of 1 has an index of 1. In general, the higher the k-additivity, the greater the $v_T(g)$.

While the above indices are all useful in their respective contexts, none truly address the desire to favor fewer numbers of inputs. In the next subsection we explore a few new indices to achieve this goal based on utilization of the Shapley values.

2.3 New indices of complexity based on the Shapley values

Definition 13 (ℓ_1-*Norm of Shapley Values*) Let $\Phi_g = \big(\Phi_g(1),\ \Phi_g(2),\ \ldots,\ \Phi_g(N)\big)^t$, a vector of size $N \times 1$, be the vector of Shapley values. The so-called ℓ_1-norm of Φ_g is

$$\|\Phi_g\|_1 = \sum_{i=1}^{N}|\Phi_g(i)| = \sum_{i=1}^{N}\Phi_g(i) = 1. \tag{11}$$

It is important to note that the constraint that the Shapley values sum to 1 renders the ℓ_1 index useless for regularization (as it yields a constant). Next, we explore the ℓ_0.

Definition 14 (ℓ_0-*Norm of Shapley Values*) Let $\Phi_g = \big(\Phi_g(1),\ \Phi_g(2),\ \ldots,\ \Phi_g(N)\big)^t$, a vector of size $N \times 1$, be the vector of Shapley values. The so-called ℓ_0-norm of Φ_g is

$$\|\Phi_g\|_0 = \big|\{i : \Phi_g(i) \neq 0\}\big|. \tag{12}$$

Technically, the ℓ_0-norm is not really a norm. It is a cardinality function that counts the number of non-zero terms. The ℓ_0-norm has been used extensively in areas like compressive sensing, where the goal is to typically find the sparsest solution for an under-determined linear system. If we define Eq. (12) on the lexicographically encoded capacity vector, versus the Shapley values vector, then we would be back in the same predicament of striving for a capacity of all 0s (except for $g(X) = 1$), viz., the minimum operator for the CI. It is clear that Eq. (12) has its minimum

for the case of one Shapley value equal to 1 (thus all other Shapley values are equal to 0). Its next smallest value is for the case of two Shapley values greater than zero and all other Shapley values are equal to zero (and so forth). It is clear to see that sparsity, in the sense of the fewest number of non-zero values, is preserved via the ℓ_0-norm. Specifically, Eq. (12) has N minima, e.g., $\Phi_g = (1, 0, \ldots, 0)^t$, $\Phi_g = (0, 1, 0, \ldots, 0)^t$, ..., $\Phi_g = (0, 0, \ldots, 0, 1)^t$. For two non-zero values, there are $\binom{N}{2}$ such solutions ($\binom{N}{k}$ in general for k non-zero inputs).

As an index, the ℓ_0 with respect to the Shapley values is fantastic at helping promote fewer number of non-zero parameters (inputs). Problem solved, correct? Not entirely. One (big) challenge is that the ℓ_0 results in a non-convex optimization problem that is NP-hard. Before we consider ℓ_0 approximation, we investigate an alternative, but theoretically inferior (the tradeoff), index that is simpler to solve for based on the Gini-Simpson coefficient.

The Gini coefficient (aka Gini index or Gini ratio) is a summary statistic of the Lorenz curve and it was developed by the Italian statistician Corrado Gini in 1912. It is important to note that numerous mathematical formulations exist, from Corrado's original formula to Brown's Gini-style index measuring inequalities in health (Gini 1936; Brown 1994). A full review of the Gini index and its various discrete and continuous formulations is beyond the scope of this article (for a recent review, see Farris 2010). The Gini index is used extensively in areas like biological systems (for measuring species similarity Leinster and Cobbold 2012), knowledge discovery in databases (often referred to as an "impurity function"), social sciences and economics. For example, it is often used as a measure of income inequality within a population. On one extreme, the Gini index equates to perfect "equality" (everyone has the same income) and, at the other extreme, to perfect "inequality" (one person has all the wealth and everyone else has zero income). Herein, we use a mathematically simple, but pleasing nevertheless, instantiation of the Gini index—at Eq. (13)—often referred to as the Gini–Simpson index (or in ecology as the probability of interspecific encounter) with respect to a probability distribution (the Shapley values satisfy this criterion). However, the Gini–Simpson function belongs to a larger family of functions parameterized by a variable q (the sensitivity parameter) and Z (a matrix of similarity values between elements in the distribution) (Leinster and Cobbold 2012). Based on q and Z, we get different diversity measures, e.g., Shannon's entropy, Rao's quadratic entropy, "species richness" index, etc.

Definition 15 (*Gini–Simpson Index of Shapley Values*) The Gini–Simpson index of the Shapley values is

$$\upsilon_G(g) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \Phi_g(i)\Phi_g(j) = 1 - \sum_{i=1}^{N} \Phi_g(i)^2. \tag{13}$$

Note, $\upsilon_G(g) = 0$ if and only if there is a single Shapley value equal to 1 (therefore all other values are equal to 0). There are N such possible unique solutions to this criteria. If $\Phi_g(i) = 1$ and $\Phi_g(j) = 0, \forall j \neq i$, then all g subsets that contain input i are of value 1 and 0 elsewhere. Also, the maximum of $\upsilon_G(g)$ occurs only when all Shapley values are equal. When these values are all equal, all inputs are "equally important" and increasing the cardinality of any set of inputs in this case increases the FM by a constant value, regardless of which input was used to increase cardinality. This essentially reduces the CI to an OWA. It is obvious that Eq. (13) is nothing more than one minus the squared ℓ_2-norm of the Shapley values. Next, we provide simple numeric examples (Table 2) to (empirically) demonstrate some similarities and differences between the ℓ_0 and the Gini–Simpson.

Again, the ℓ_0 wants the fewest number of non-zero parameters and the Gini–Simpson index is a measure of diversity in the Shapley values, or more specifically one minus the squared ℓ_2-norm, that ultimately aims to promote, in the extreme case, a single dominant input (one Shapley value of 1 and all other values equal to 0). They both have lowest value for a single input (case g_a) and maximum value for the case of a uniform distribution (case g_g). While their trends are often similar, e.g., both prefer g_a to g_b and g_b to g_d and g_e, they do not always obviously agree. For example, consider g_c and g_d. The ℓ_0-norm prefers g_c to g_d as the prior has one zero term and the latter has no zero terms. However, the Gini–Simpson index prefers g_d to g_c. In g_c, the Shapley values indicate that one input is not important while the other two inputs are equally important. In g_d, the Shapley values indicate that one input is important and the other two inputs are equal and not that

Table 2 Numeric examples, for $N = 3$, illustrating ℓ_0 and Gini–Simpson differences

FM	Shapley values	ℓ_0	Gini–Simpson
g_a	$\Phi_{g_a} = (1, 0, 0)$	1	0
g_b	$\Phi_{g_b} = (0.8, 0.2, 0)$	2	0.320
g_c	$\Phi_{g_c} = (0.5, 0.5, 0)$	2	0.500
g_d	$\Phi_{g_d} = (0.8, 0.1, 0.1)$	3	0.340
g_e	$\Phi_{g_e} = (0.4, 0.3, 0.3)$	3	0.660
g_f	$\Phi_{g_f} = (0.999, 0.0005, 0.0005)$	3	0.002
g_g	$\Phi_{g_g} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	3	0.667

important at that. According to the Gini–Simpson index, g_d is closer to a single input vs g_c. This behavior is further emphasized by g_f.

In addition, due to the relationship between the Gini–Simpson and the ℓ_2-norm for the Shapley values, the underlying geometric interpretation and sparseness of solutions for the family of ℓ_p-norms is well-known and heavily published (e.g., see Tibshirani et al. 2005). In the following sections, we outline a way to perform regularization based on the ℓ_0-norm and the Gini-Simpson index of the Shapley values. However, we first review QP based solutions to capacity learning and ℓ_p-norm regularization.

3 Sum of squared error and quadratic programming

Definition 16 (*Sum of Squared Error of CI and T*) Let the SSE between T and the CI be (Grabisch et al. 2000; Anderson et al. 2014b)

$$E_1 = \sum_{j=1}^{m} (C_g(h_j)) - \alpha_j)^2. \tag{14}$$

Equation (14) can be expanded as follows,

$$E_1 = \sum_{j=1}^{m} (\mathbf{A}_{O_j}^t \mathbf{u} - \alpha_j)^2,$$

where $\mathbf{u}$ is the lexicographically encoded capacity vector and $\mathbf{A}_{O_j}^t = \left(\ldots, h_j(x_{\pi_j(1)}) - h_j(x_{\pi_j(2)}), \ldots, 0, \ldots, h_j(x_{\pi_j(N)}) \right)^t$ is of size $1 \times (2^N - 1)$. Note, the function differences, $h_j(x_{\pi_j(i)}) - h_j(x_{\pi_j(i+1)})$, correspond to their respective g locations in $\mathbf{u}$. Folding Eq. (14) out further, we find

$$\begin{aligned} E_1 &= \sum_{j=1}^{m} (\mathbf{u}^t \mathbf{A}_{O_j} \mathbf{A}_{O_j}^t \mathbf{u} - 2\alpha_j \mathbf{A}_{O_j}^t \mathbf{u} + \alpha_j^2) \\ &= \mathbf{u}^t \mathbf{D} \mathbf{u} + \mathbf{f}^t \mathbf{u} + \sum_{\mathbf{j=1}}^{\mathbf{m}} \alpha_{\mathbf{j}}^{\mathbf{2}}, \end{aligned} \tag{15}$$

where $\mathbf{D} = \sum_{\mathbf{j=1}}^{\mathbf{m}} \mathbf{A}_{\mathbf{O_j}} \mathbf{A}_{\mathbf{O_j}}^{\mathbf{t}}$ and $\mathbf{f} = \sum_{\mathbf{j=1}}^{\mathbf{m}} (-\mathbf{2}\alpha_{\mathbf{j}} \mathbf{A}_{\mathbf{O_j}})$. In total, the capacity has $(N(2^{N-1} - 1))$ monotonicity constraints. These constraints can be represented in a compact linear algebra (aka matrix) form. The following is the minimum number of constraints needed to represent the FM. Let $\mathbf{Cu} + \mathbf{b} \leq \mathbf{0}$, where $\mathbf{C^t} = \left(\Psi_{\mathbf{1}}^{\mathbf{t}}, \Psi_{\mathbf{2}}^{\mathbf{t}}, \ldots, \Psi_{\mathbf{N+1}}^{\mathbf{t}}, \ldots, \Psi_{N(2^{N-1}-1)}^{t} \right)^t$, and Ψ_1 is a vector representation of constraint 1, $g_1 - g_{12} \leq 0$. Specifically, $\Psi_1^t \mathbf{u}$ recovers $\mathbf{u}_1 - \mathbf{u}_{N+1}$. Thus, C is simply a matrix of $\{0, 1, -1\}$ values,

$$\mathbf{C} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \ldots & \mathbf{-1} & \mathbf{0} & \ldots & \ldots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{-1} & \ldots & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{1} & \mathbf{-1} \end{bmatrix}, \tag{16}$$

which is of size $(N(2^{N-1} - 1)) \times (2^N - 1)$. Also, $\mathbf{b} = \mathbf{0}$, a vector of all zeroes. Note, in some works, $\mathbf{u}$ is of size $(2^N - 2)$, as $g(\phi) = 0$ and $g(X) = 1$ are explicitly encoded. In such a case, $\mathbf{b}$ is a vector of 0s and the last N entries are of value -1. Herein, we use the $(2^N - 1)$ format as it simplifies the subsequent Shapley index mathematics. Given T, the search for FM g reduces to a QP of the form

$$\min_{\mathbf{u}} \frac{1}{2} \mathbf{u}^t \hat{\mathbf{D}} \mathbf{u} + \mathbf{f}^{\mathbf{t}} \mathbf{u}, \tag{17}$$

subject to $\mathbf{Cu} + \mathbf{b} \geq \mathbf{0}$ and $(\mathbf{0}, 1)^t \leq \mathbf{u} \leq \mathbf{1}$. The difference between Eqs. (17) and (15) is $\hat{\mathbf{D}} = 2\mathbf{D}$ and inequality in Eq. (16) need only be multiplied by -1.

In Mendez-Vazquez and Gader (2007), it was pointed out that the QP approach for learning the CI is not without flaw due to the exponential size of the input. While scalability is definitely of concern, many techniques have and continue to be proposed for solving QPs with respect to fairly large and sparse matrices (Cevher et al. 2014). This attention and progress is coming primarily as a response to machine learning, statistics and signal processing. A somewhat large and sparse matrix is not a "game stopper". We do agree that there is mathematically a point where the task at hand does become extremely difficult to solve and may eventually become intractable. However, most FI applications utilize a relatively small number of inputs, i.e., on the order of 3– 5, versus 50, 100. The notion that the QP has little-to-no value just because it is difficult (and may become intractable) to solve with respect to a sparse matrix for a large number of inputs is no reason to dismiss it. This challenge is akin to the current Big Data revolution, where previously intractable problems are being solved on a daily basis.

In general, the challenge of QP-based learning of the CI relative to a regularization term for tasks like decision-level fusion is the optimization of

$$E_2 = \sum_{j=1}^{m} (\mathbf{u}^t \mathbf{A}_{O_j} \mathbf{A}_{O_j}^t \mathbf{u} - 2\alpha_j \mathbf{A}_{O_j}^t \mathbf{u} + \alpha_j^2) + \lambda v_*(g), \tag{18}$$

where $v_*(g)$ is one of our indices. In order for Eq. (18) to be suitable for the QP, v_* must be linear or quadratic.

Note, in certain problems one can simply fold the ℓ_1 regularization term into the linear term of the quadratic objective. We can rewrite $\|\mathbf{u}\|_1 = \mathbf{1}^t \mathbf{u}$, where $\mathbf{1}$ is the vector of all ones. Adding the regularization term to the QP, we get

$$\min_{\mathbf{u}} \frac{1}{2}\mathbf{u}^t\hat{\mathbf{D}}\mathbf{u} + \mathbf{f}^t\mathbf{u} + \lambda\mathbf{1}^t\mathbf{u} = \min_{\mathbf{u}} \frac{\mathbf{1}}{\mathbf{2}}\mathbf{u}^t\hat{\mathbf{D}}\mathbf{u} + (\mathbf{f} + \lambda\mathbf{1})^t\mathbf{u}. \quad (19)$$

4 Gini–Simpson index-based regularization of the Shapley values

We begin by considering a vectorial encoding of the Shapley index. The Shapley value of input 1 is

$$\Phi_g(1) = \sum_{K\subseteq X\setminus\{i=1\}} \Gamma_X(K)(g(K\cup\{i=1\}) - g(K)), \quad (20a)$$

$$= \eta_1 g(\{x_1\}) + \eta_2[(g(\{x_1,x_2\}) - g(\{x_2\})) + (g(\{x_1,x_3\}) - g(\{x_3\})) + \cdots] + \cdots, \quad (20b)$$

$$= \eta_1 g(\{x_1\}) - [\eta_2 g(\{x_2\}) + \cdots + \eta_2 g(\{x_N\})] + [\eta_2 g(\{x_1,x_2\}) + \eta_2 g(\{x_1,x_N\})] + \cdots, \quad (20c)$$

where $\eta_i = \Gamma_X(K)$, and $K \in 2^X$, s.t. $|K| = i-1$ (Shapley normalization constants). What Eq. (20a) tells us is the following. The Shapley index can be represented in linear algebra/vectorial form,

$$\Gamma_i = \left(\Gamma_{i,1}, \Gamma_{i,2}, \ldots, \Gamma_{i,2^N-1}\right)^t, \quad (21)$$

where Γ_i is the same size as g and the Γ_i terms are the coefficients of Eq. (20a) arranged such that multiplication with the lexicographic FM vector yields a particular Shapley value. For example, for $N = 3$,

$$\Gamma_1 = \left(\frac{1}{3}, -\frac{1}{6}, -\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, -\frac{1}{3}, \frac{1}{3}\right)^t.$$

Thus, we can formulate a compact expression of an individual Shapley value as such,

$$\Phi_g(i) = \Gamma_i^t\mathbf{u}, \quad (22)$$

where $\Phi_g(i) \in [0,1]$. Therefore, the Gini–Simpson index in linear algebra form becomes

$$\upsilon_G(g) = 1 - \sum_{k=1}^{N} \left(\Gamma_k^t\mathbf{u}\right)^2. \quad (23)$$

Expanding Eq. (23) exposes an attractive property:

$$\begin{aligned}\upsilon_G(g) &= 1 - \sum_{k=1}^{N} \left(\Gamma_k^t\mathbf{u}\right)^2, \\ &= 1 - \sum_{k=1}^{N} \left(\mathbf{u}^t\Gamma_k\Gamma_k^t\mathbf{u}\right), \\ &= 1 - \mathbf{u}^t\mathbf{Z}\mathbf{u},\end{aligned} \quad (24)$$

where $\mathbf{Z} = \Gamma_1\Gamma_1^t + \cdots + \Gamma_N\Gamma_N^t$. First, $\Gamma_k\Gamma_k^t$ is positive semi-definite (PSD). Hence, $\mathbf{Z}$ is also PSD, as it is simply the addition of PSD matrices and addition preserves the PSD property. We propose a Gini–Simpson index-based regularization of E_1 at Eq. (14) as follows.

Definition 17 (*SSE with Gini–Simpson Index Regularization*) The Gini–Simpson index regularization is

$$E_3 = \mathbf{u}^t\mathbf{D}\mathbf{u} + \mathbf{f}^t\mathbf{u} + \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{m}} \alpha_{\mathbf{j}}^2 + \lambda - \lambda\left(\mathbf{u}^t\mathbf{Z}\mathbf{u}\right), \quad (25)$$

where the regularization term can be simply folded into the quadratic term in the SSE yielding

$$\min_{\mathbf{u}} \mathbf{u}^t\left(\mathbf{D} - \lambda\mathbf{Z}\right)\mathbf{u} + \mathbf{f}^t\mathbf{u}, \quad (26)$$

subject to $\mathbf{C}\mathbf{u} \geq \mathbf{0}$ and $(\mathbf{0}, 1)^t \leq \mathbf{u} \leq \mathbf{1}$.

This is of the form of Tikhonov regularization, where $-\lambda\mathbf{Z}$ is the Tikhonov matrix (Tikhonov 1943). As one can clearly see, the Gini–Simpson index does not result in a linear term and the constant is not part of the QP formulation. All that makes it into the quadratic term is $\mathbf{u}^t\mathbf{Z}\mathbf{u}$, our scaled (squared) ℓ_2-norm.

5 ℓ_0-norm based regularization of the Shapley values

The main difficulty behind the ℓ_0-norm of the Shapley values is how do we carry out its optimization? Our QP task with an ℓ_0-norm is a non convex problem, which makes it difficult to understand theoretically and solve computationally (NP-hard problem). There are numerous articles focused on approximation techniques for the ℓ_0-norm. Herein, we take the approach of enhancing sparsity through reweighted ℓ_1 minimization. In Candes et al. (2008), Candes proposed a simple and computationally attractive recursively reweighted formulation of ℓ_1-norm minimization designed to more democratically penalize nonzero coefficients. His approach finds a local minimum of a concave penalty function that approximates the ℓ_0-norm. Specifically, the weighted ℓ_1 minimization task can be viewed as a relaxation of a weighted ℓ_0 minimization task.

Definition 18 (*SSE with Weighted ℓ_1-Norm*) The SSE and weighted ℓ_1-norm of the Shapley index based regularization is

$$E_4 = \mathbf{u}^t\mathbf{D}\mathbf{u} + \mathbf{f}^t\mathbf{u} + \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{m}} \alpha_{\mathbf{j}}^2 - (\lambda_1\Gamma_1 + \cdots + \lambda_N\Gamma_N)^t\mathbf{u}. \quad (27)$$

Thus, our goal is

$$\min_{\mathbf{u}} \mathbf{u}^t\mathbf{D}\mathbf{u} + \left(\mathbf{f} - (\lambda_1\Gamma_1 + \cdots + \lambda_N\Gamma_N)\right)^t\mathbf{u}, \quad (28)$$

subject to $\mathbf{Cu} + \mathbf{b} \geq \mathbf{0}$ and $(\mathbf{0}, 1)^t \leq \mathbf{u} \leq \mathbf{1}$, where $\mathbf{0}$ is a vector of all zeros of length 2^{N-2}, and $\mathbf{1}$ is a vector of all ones of length 2^{N-1}.

Algorithm 1 is exactly the method of Candes et al. just with the Shapley values as the parameters. For further mathematical analysis of Candes's approximation, see Candes et al. (2008). Herein, our goal is not to advance this approximation technique. Instead, we simply apply it for learning the CI. As better approximations become available, the reader can employ those strategies. In Algorithm 1, $\epsilon > 0$ is used to provide stability and to ensure that a zero-valued component in $1 - \Gamma_k^t(t-1)\mathbf{u}(t-1)$ does not strictly prohibit a nonzero estimate at the next step (as done in Candes et al. 2008). Intuitively, the update step takes the previous $\lambda_k(t-1)$ terms and divides them by one minus their respective Shapley values. Thus, the "more important" (the larger) the Shapley value the smaller the divisor (number in [0, 1]) and therefore the larger the $\lambda_k(t)$. Different stopping criteria exist for Algorithm 1. For example, the user can provide a maximum allowable SSE. The user can also compare the difference between the weights from iteration to iteration relative to a user specified threshold. Furthermore, the user can provide a maximum number of allowable iterations.

Algorithm 1 Weighted Iterative l_1-Norm Regularization

1: Initialize the weights, e.g., $\lambda_k(t) = 1$, for $1 \leq k \leq N$
2: Initialize the counter, $t = 1$
3: **while** NOT DONE* **do**
4: Solve for $\mathbf{u}(t)$ by minimizing E_4 given $\lambda_k(t)$
5: $t = t + 1$
6: Update, $\lambda_k(t) = \frac{\lambda_k(t-1)}{\left(1-\mathbf{\Gamma}_k^t(t-1)\mathbf{u}(t-1)\right)+\epsilon}$
7: **end while**

*As described in (Candes et al., 2008), this algorithm can be terminated either upon convergence or when t reaches a specified maximum number of iterations.

6 Experiments

In this section, we explore our methods with both synthetic and real-world data sets. These experiments demonstrate the capability of the CI with respect to the learned capacity in a supervised learning task, i.e., classification or regression. The goal of the synthetic experiments (Experiments 1 through 3) is to investigate the general behavior of the proposed theories under controlled settings in which we know the "answer" (the generating capacity). The goal of the real-world experiment (Experiment 4) is to investigate classification performance on widely-available benchmark data sets.

In Sect. 2.2 we reviewed and mathematically compared different indices, however, we do not include all indices in these experiments as they do not operate on the same basis. Each index more-or-less interprets complexity differently and, thus, each has its own place (application) and rationale for existence, both in terms of capacity theory and also in terms of how the CI is applied. In the experiments that follow, we restrict analysis to the study of the Gini–Simpson and the ℓ_0-norm of the Shapley values and we compare it to the most related indices for decision-level fusion—specifically, the ℓ_1 and ℓ_2-norm of a lexicographically encoded capacity vector and the Mobius-based index.

In our synthetic experiments (Experiments 1 through 3) we elected to not report a single summarized number or statistic, e.g., classification accuracy. Instead, we show the behavior of our technique across different possible choices of the regularization parameter λ. While somewhat overwhelming at first, we believe it is important to give the reader a better (more detailed) feel for how the methods behave in general. However, it is worth briefly noting some λ selection strategies used in practice. For example, we can pick a "winner" by trying a range of values of λ in the context of cross validation (i.e., a grid search). Such an experiment emphasizes learning less complex models with respect to the idea of avoiding over fitting (one use of an information theoretic index). We employ the same strategy in our real-world benchmark data set experiment for kernel classification. If the reader desires, they can refer to one of many works in statistics or machine learning for further assistance in automatically determining or experimentally selecting λ (Candes et al. 2008).

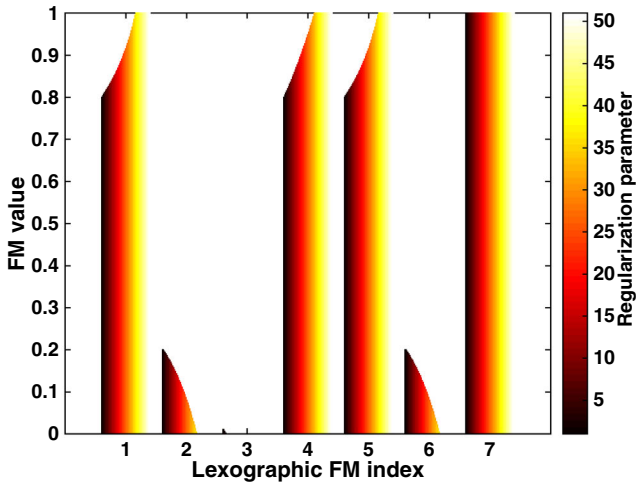
6.1 Experiment 1: important, relevant, and irrelevant inputs

For this experiment, we consider the case of learning an aggregation function for three inputs ($N = 3$)—in essence, a regression task. While this experiment is easily generalized to more than three inputs, the advantage of $N = 3$ is that we can clearly visualize the results, since it becomes difficult to view the results for more inputs as the number of elements in the capacity grows exponentially with respect to N.

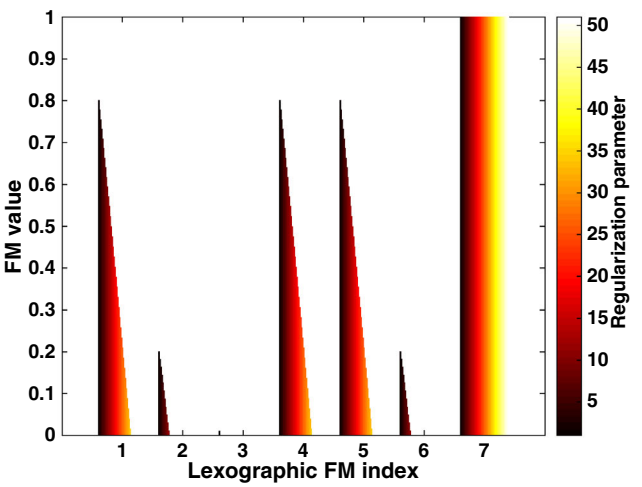
To generate the synthetic data, we define the densities of the FM as $g(x_1) = 0.8$, $g(x_2) = 0.2$, and $g(x_3) = 0.01$, such that the inputs can be considered important, relevant, and irrelevant, respectively. The capacity beyond the densities is determined as $g(A) = \max_{x_i \in A} g(x_i)$, for $A \in 2^X \backslash \{x_1, x_2, \ldots, x_N, X\}$, making the FM an S-Decomposable FM, specifically a possibility measure. This synthetic capacity was used in conjunction with the Choquet integral to produce training labels from 500 uniform (pseudo)randomly generated N-tuples, each in the range [0, 1].

We expect the FM learned from this synthetic training data to behave as follows. We would like for the third input to be ignored and the second input should be driven down to zero worth (in the Shapley sense) before the first input is driven toward zero worth. Figure 2 shows the results of this experiment. Views (a, b) show the FM values learned for values of λ between 0 and 50—the left side of each bar (the black) corresponds to the learned FM values at $\lambda = 0$ and the right side of each bar (the bright yellow) corresponds to the FM values at $\lambda = 50$. Views (c, d) show the value of the Gini–Simpson index for the learned FM and the resulting SSE versus each value of λ. The scale for the solid blue line—the Gini–Simpson index—is shown on the left of each plot and the scale for the dashed red line—the SSE—is shown on the right of each plot.

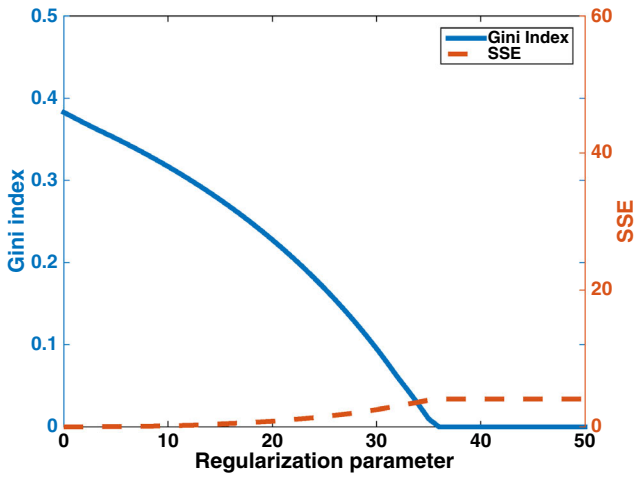
Figure 2 tells the following story. The black color bars in views (a, b) show that both methods recover the desired possibility measure (the one that minimizes just the SSE criteria) when no regularization $\lambda = 0$ is used—after all, the methods are equivalent, i.e., no regularization, when $\lambda = 0$. View (a) shows that the Gini–Simpson index regularization pushes the contribution of input 3 to zero very quickly—at $\lambda \approx 5$—and the contribution of input 2 is reduced to zero at $\lambda \approx 35$. The contribution of $\mathbf{u}_6 = g(\{x_2, x_3\})$ is also pushed to zero as λ increases. On the contrary, the contribution of input 1 and the FM values in the lattice that include input 1—i.e., $\mathbf{u}_4$ and $\mathbf{u}_5$—are gradually increased with increasing λ. Figure 2c shows that as λ is increased the Gini–Simpson index decreases, which is echoed in the FM values shown in view (a). As the



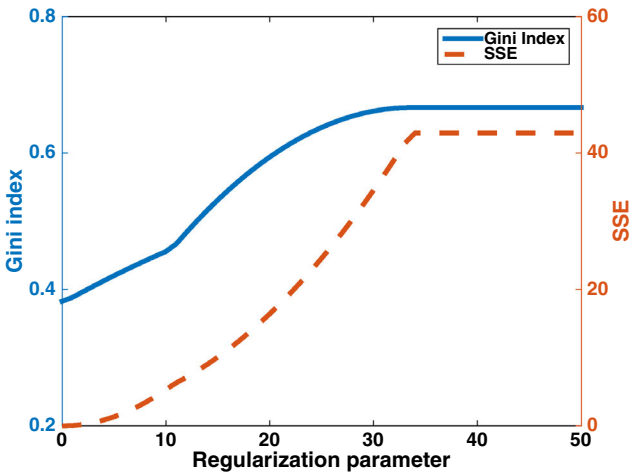
(a) Learned FM values using Gini-Simpson regularization



(b) Learned FM values using ℓ_1 regularization



(c) Gini-Simpson regularization performance plots



(d) ℓ_1 regularization performance plots

Fig. 2 Experiment 1 results. **a**, **b** Learned FM values in lexicographical order for $\lambda = 0$ to 50. Bin 1 is $\mathbf{u}(1) = g(x_1)$, bin $N+1$ is $\mathbf{u}(N+1) = g(\{x_1, x_2\})$, etc. *Height of bar* indicates FM value; *color* indicates λ value. **c**, **d** Plots showing performance of each regularization method in terms of SSE and Gini–Simpson index of the learned FM at each regularization parameter λ

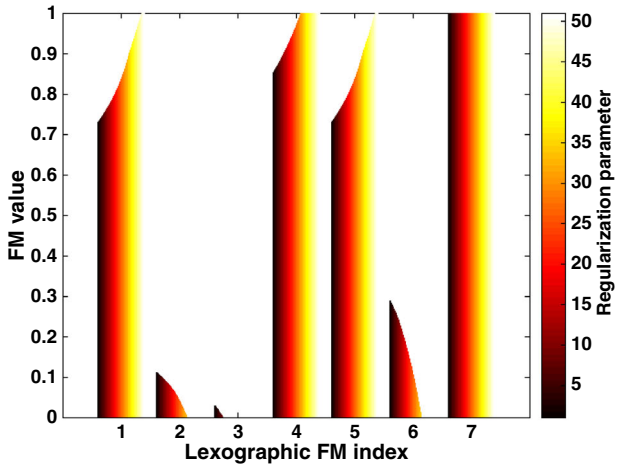
model becomes more simple, by increasing λ, the SSE increases (albeit, slightly). At $\lambda \approx 35$, the Gini-Simpson index goes to zero, indicating the model is as simple as it can get. Increasing $\lambda > 35$ has no effect on the model because it is already as simple as possible, with only one input (#1) being considered in the solution. The SSE of this minimum Gini–Simpson index model is about 4.

Figure 2b, d show visualizations of the same experiment for ℓ_1 regularization. As view (b) shows, this regularization starts decreasing all of the FM values as λ is increased. The contribution of input 3, $\mathbf{u}_3 = g(x_3)$, is quickly pushed to zero, at $\lambda \approx 2$, while the values $\mathbf{u}_2 = g(x_2)$ and $\mathbf{u}_6 = g(\{x_2, x_3\})$ go to zero at $\lambda \approx 10$. Last, $\mathbf{u}_1 = g(x_1)$, $\mathbf{u}_4 = g(\{x_1, x_2\})$, and $\mathbf{u}_5 = g(\{x_1, x_3\})$ go to zero at $\lambda \approx 32$. At $\lambda \gtrsim 32$, the ℓ_1 regularization learns, as expected, the FM of ignorance. Figure 2d shows that despite a lower complexity model, in terms of ℓ_1-norm, the Gini–Simpson index increases as λ is increased; SSE also increases, as expected. The FM of ignorance learned at $\lambda \gtrsim 32$ has an SSE of about 45. Compare this to the SSE of 4 achieved with the lowest complexity model with Gini–Simpson regularization.

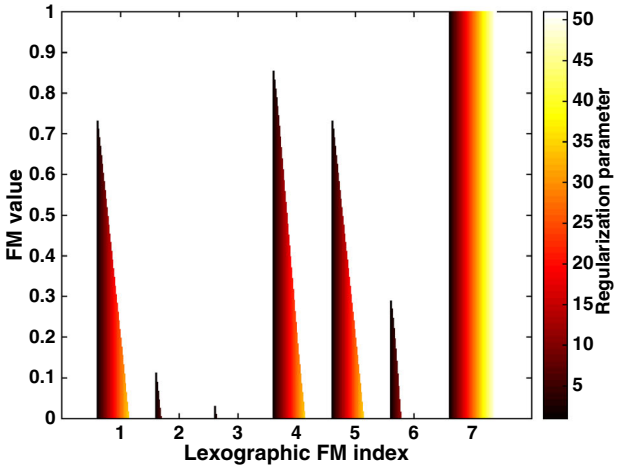
In summary, this initial experiment shows that the Gini–Simpson index regularization and ℓ_1-regularization of a lexicographically encoded capacity vector do as advertised.

6.2 Experiment 2: random AWGN noise

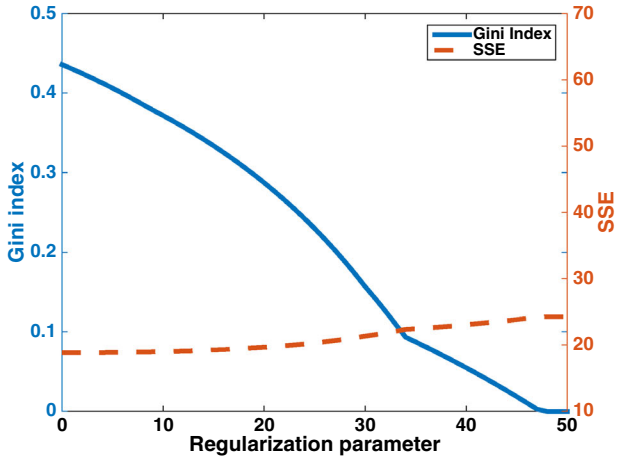
In Experiment 2, we use our setup from Experiment 1; however, (pseudo)random AWGN noise ($\sigma = 0.2$) is added to the labels. Figure 3 shows the results of Experiment 2.



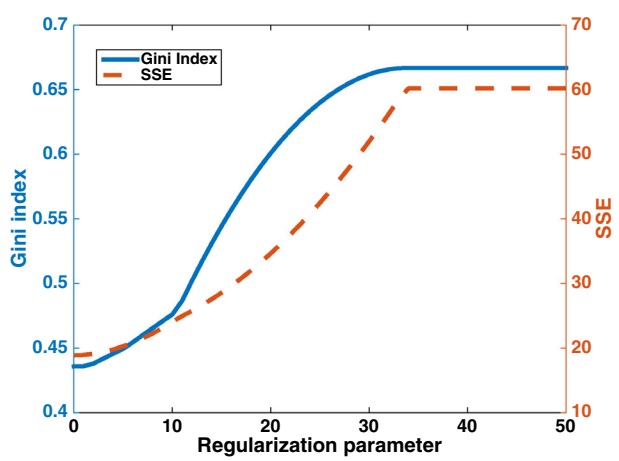
(a) Learned FM values using Gini-Simpson regularization



(b) Learned FM values using ℓ_1 regularization

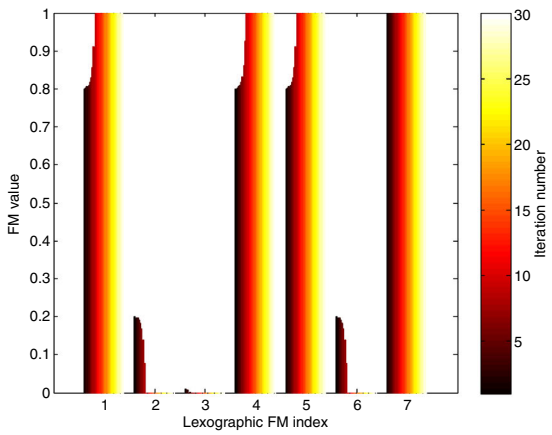


(c) Gini-Simpson regularization performance plots

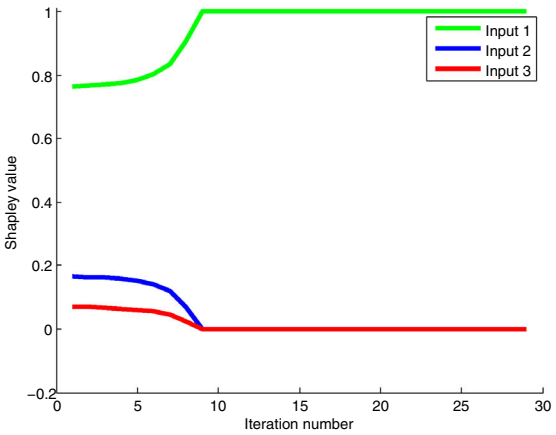


(d) ℓ_1 regularization performance plots

Fig. 3 Experiment 2 results. **a**, **b** Learned FM values in lexicographical order. Bin 1 is $\mathbf{u}(1) = g(x_1)$, bin $N+1$ is $\mathbf{u}(N+1) = g(\{x_1, x_2\})$, etc. *Height of the bar* indicates FM value; *color* indicates λ value. **c**, **d** Plots showing performance of each regularization method in terms of SSE and Gini–Simpson index values of the learned FM at each regularization parameter λ



(a) Learned FM values using iteratively reweighted ℓ_1 regularization for Experiment 1



(b) Shapley values for (a)

Fig. 4 Experiment 3 results. Learned FM values in lexicographical order for Experiment 1. Bin 1 is $\mathbf{u}(1) = g(x_1)$, bin $N+1$ is $\mathbf{u}(N+1) = g(\{x_1, x_2\})$, etc. *Height of the bar* indicates FM value; *color* indicates iteration number. Plot of the Shapley values of the learned FM for Experiment 1 at each iteration.eps

As views (c, d) show, neither procedure perfectly fits the data now due to the noise in the training labels. The Gini–Simpson procedure, shown in views (a, c), can find a solution close to our noise-free goal at small values of λ. If regularization is increased, $\lambda \gtrsim 45$, we eventually identify a single input, which interestingly still fits the data well (only a small increase in SSE). Again, the ℓ_1 procedure is only able to achieve low SSE at low λ values. As λ is increased the SSE is significantly increased (beyond that achieved by the Gini–Simpson).

6.3 Experiment 3: iteratively reweighted ℓ_1-norm

In Experiment 3, we use our setup from Experiment 1 to demonstrate the recursively reweighted ℓ_1 minimization procedure. The result (Fig. 4a, b) for the possibility FM with densities $g(x_1) = 0.8$, $g(x_2) = 0.2$, $g(x_3) = 0.01$, is as expected. After a few iterations we see the Shapley value increasing for input x_1 and decreasing for inputs x_2 and x_3. This is the same trend and final answer that we saw in Experiment 1 with respect to the Gini–Simpson index and we obviously get a different final solution than the ℓ_1 with respect to the lexographically encoded capacity vector (eventual solution of 0s and corresponding CI minimum operator).

6.4 Experiment 4: multiple kernel learning

In this final experiment we consider a problem from pattern recognition. Kernel methods for classification is a well-studied field in which data are implicitly mapped from a lower-dimensional space to a higher-dimensional space, called the reproducing kernel Hilbert space (RKHS), to improve classification accuracy. The ultimate challenge is that we are not privileged to know what transform (kernel) solves a particular task at hand—we only have an existence theorem. Multiple kernel learning (MKL) is a way to learn the fusion of multiple known Mercer kernels (the building blocks) to identify a superior kernel. In Pinar et al. (2015, 2016), Hu et al. (2013, 2014), a genetic algorithm (GA) based ℓ_p-norm linear convex sum of kernels called GAMKLp for feature-level fusion was proposed. In Pinar et al. (2015), the nonlinear fusion of kernels was also explored. Specifically, kernel classifiers were fused at the decision-level based on the fuzzy integral, a procedure called decision-level fuzzy integral MKL (DeFIMKL). In this experiment, we explore the use of QP learning and regularization for CI-based MKL in the context of support vector machine (SVM) classification with respect to well-known community benchmark data sets. In Pinar et al. (2015), the benefit of DeFIMKL and GAMKL was demonstrated versus other state-of-the-art MKL algorithms from machine learning, e.g., MKL group lasso (MKLGL). Herein, the goal is not to reestablish DeFIMKL but to explore the proposed indices and their relative performances. Note, in the other experiments we knew the answer, i.e., the "generating capacity". However, while SVMs are supervised learners and our data has labels, we do not know the true capacity in the case of MKL. Herein, like often in machine learning, success is instead evaluated in terms of ones ability to improve classification performance. The fusion of classifiers via DeFIMKL results in a classifier and this experiment demonstrates the ability of regularization to help learn an improved classifier that does not succumb to overfitting and generalizes better.

Each learner, i.e., input to fusion, is a kernel classifier trained on a separate kernel. The kth $(1 \leq k \leq N)$ SVM classifier decision value is

$$\eta_k(\mathbf{x}) = \sum_{\mathbf{i=1}}^{\mathbf{D}} \alpha_{\mathbf{ik}} \mathbf{y_i} \kappa_{\mathbf{k}}(\mathbf{x_i}, \mathbf{x}) - \mathbf{b_k},$$

which is the distance of $\mathbf{x}$ (an object from T) from the hyperplane defined by the data labels, y, the kth kernel, $\kappa_k(\mathbf{x_i}, \mathbf{x})$, and the learned SVM model parameters, α_{ik} and b_k. For the two-class SVM binary decision problem, the class label is typically computed as $\text{sgn}\{\eta_k(\mathbf{x})\}$. One could use $\text{sgn}\{\eta_k(\mathbf{x})\}$ as the training input to the capacity learning, however this eliminates information about which kernel produces the largest class separation—essentially, the difference between $\eta_k(\mathbf{x})$ for classes labeled $y = +1$ and $y = -1$. In Pinar et al. (2015) $\eta_k(\mathbf{x})$ is remapped onto the interval $[-1, +1]$, creating the inputs for learning by the sigmoid function $\frac{\eta_k(\mathbf{x})}{\sqrt{1+\eta_k^2(\mathbf{x})}}$. For training, we use our labeled data and cast the learning process as a SSE problem and the CI is learned using QP and regularization (see Pinar et al. 2015 for a full mathematical description).

The well-known LIBSVM library was used to implement the classifiers (Chang and Lin 2011). The machine learning UCI benchmark data sets used are sonar, dermatology, wine, ecoli and glass. Each experiment consists of 100 randomly sampled trials in which 80% of the data is used for training and the remaining 20% is preserved for testing. Each index was applied to the same random sample to guarantee a fair comparison. Note that in some cases multiple classes are joined together such that the classification decision is binary. Five radial basis function (RBF) kernels are used in each algorithm with respective RBF width σ linearly spaced on the interval defined in Table 3; the same RBF parameters are used for each algorithm. For the ℓ_1, ℓ_2, Gini–Simpson and k-additive indices, a dense grid search (of λ) was used and the "winner" was picked according to the highest classification accuracy on the test data. For the iteratively reweighted ℓ_1 approximation, we used Algorithm 1. Table 4 is the result of regularization on DeFIMKL.

Table 4 tells the following story. First, in each instance regularization helps. In many instances, e.g., ecoli, glass and wine, the regularization results are extremely close. However, in other cases, e.g., sonar and dermatology, the regularization results vary more (both in terms of means and standard deviations). Note, we ran the k-additive index with different levels of forced k-additivity. This was done to explore the impact of assuming and working with subsets of the capacity. The results show for the various k-additive experiments that as k is increased classification accuracy generally increases, though for the ecoli and glass datasets the performance remains stable and decreases slightly, respectively. In our other experiments we were able to analyze specific conditions and properties relating to the fusion process. While this experiment is encouraging, i.e., better classification performance, we are sadly unable to connect a story to the results. The regularization results are what they are—with these datasets some form of

Table 3 RBF kernel parameter ranges and data set properties

	Data set				
	Sonar	Dermatology	Wine	Ecoli	Glass
Parameter ranges	$[-2.2, -0.2]$	$[-2.3, 0]$	$[-6, -3]$	$[-3, 3]$	$[-2, 2]$
No. of objects	208	366	178	336	214
No. of features	60	33	13	7	9
Binary classes	$\{1\}$ vs. $\{2\}$	$\{1,2,3\}$ vs. $\{4,5,6\}$	$\{1\}$ vs. $\{2,3\}$	$\{1,2,3,4\}$ vs. $\{5,6,7,8\}$	$\{1,2,3\}$ vs. $\{4,5,6\}$

Table 4 Classifier performances—means and standard deviations

	Sonar	Dermatology	Wine	Ecoli	Glass
No Regularization	80.43 (9.25)	94.51 (3.89)	93.00 (9.02)	96.71 (2.90)	94.33 (5.23)
Lexicographic ℓ_1	86.52 (7.55)	94.57 (3.91)	93.44 (8.52)	96.71 (2.90)	**96.33 (4.73)**
Lexicographic ℓ_2	86.43 (7.42)	94.57 (3.91)	94.00 (8.27)	96.71 (2.90)	96.00 (4.82)
Gini–Simpson	87.14 (6.98)	**98.22 (2.15)**	94.22 (7.97)	**97.15 (2.63)**	96.14 (4.67)
Shapley ℓ_0 approximation	**87.38 (6.98)**	97.76 (2.40)	**94.56 (7.65)**	97.12 (2.71)	96.05 (4.59)
$k = 2$ additive	84.90 (7.63)	94.57 (3.91)	93.78 (8.25)	96.71 (2.90)	96.19 (4.69)
$k = 3$ additive	85.67 (7.49)	94.57 (3.91)	93.89 (8.26)	96.71 (2.90)	95.52 (4.78)
$k = 4$ additive	86.48 (7.37)	94.92 (3.81)	94.00 (8.27)	96.71 (2.90)	95.38 (4.90)
$k = 5$ additive	86.48 (7.37)	94.92 (3.81)	94.00 (8.27)	96.71 (2.90)	95.38 (4.90)

Bold values indicate which algorithm achieves superior results for each dataset

regularization gives superior performance in all cases with respect to the *k*-additive algorithms. We cannot go to the next step and inform the reader why a Gini–Simpson or *k*-additive index is better suited given our limited knowledge about the machine learning classification task.

7 Conclusion and future work

In this article, we explored a new data-driven way to learn the CI in the context of decision-level fusion relative to the joint minimization of the SSE criteria and desire to obtain minimum model complexity. We brought together and analyzed a number of existing indices, put forth new indices based on the Shapley values, and explored their role in regularization-based learning of the CI. Our first proposed index promotes sparsity (specifically, fewer number of non-zero parameters), however it is complicated to optimize (NP-hard). Our second index is a tradeoff with respect to modeling accuracy relative to solution simplicity. The proposed indices and regularization approach are compared theoretically. We showed that there is no "winning index", as these indices strive for different goals and are therefore valid in different contexts. Synthetic and real-world data set experiments are shown that demonstrate the benefits of the proposed indices and CI learning technique.

In future work, we will seek more efficient and scalable ways to solve the problem investigated here as the number of inputs (N) grows—since the number of capacity terms and subsequently associated monotonicity constraints increases at an exponential rate. We will also explore if there are other information theoretic measures that have additional benefit towards learning lower complexity and useful CIs.

Acknowledgements This work was partially funded by the Army Research Office Grants W911NF-14-1-0114, 57940-EV, W909MY-13-C-0013, W909MY-13-C-0029, and W911NF-13-1-002 to support the U.S. Army RDECOM CERDEC NVESD and the Pacific Northwest National Laboratory, under U.S. Department of Energy Contract DE-AC05-76RL01830. Superior, a high-performance computing infrastructure at Michigan Technological University, was used in obtaining results presented in this publication.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Anderson DT, Havens TC, Wagner C, Keller JM, Anderson MF, Wescott D (2014a) Extension of the fuzzy integral for general fuzzy set-valued information. IEEE Trans Fuzzy Syst 22(6):1625–1639

Anderson DT, Keller JM, Havens TC (2010) Learning fuzzy-valued fuzzy measures for the fuzzy-valued Sugeno fuzzy integral. In: International conference on information processing and management of uncertainty, pp 502–511

Anderson DT, Price S, Havens TC (2014b) Regularization-based learning of the Choquet integral. In: 2014 IEEE international conference on fuzzy systems, pp 2519–2526

Beliakov G (2009) Construction of aggregation functions from data using linear programming. Fuzzy Sets Syst 160:65–75

Beliakov G, Pradera A, Calvo T (2008) aggregation functions: a guide for practitioners, 1st edn. Springer Publishing Company, Incorporated, Heidelberg

Brown M (1994) Using Gini-style indices to evaluate the spatial patterns of health practitioners: theoretical considerations and an application based on Alberta data. Soc Sci Med 38(9):1243–1256

Candes E, Wakin M, Boyd S (2008) Enhancing sparsity by reweighted l1 minimization. J Fourier Anal Appl 14:877–905

Cevher V, Becker S, Schmidt M (2014) Convex optimization for Big Data: scalable, randomized, and parallel algorithms for big data analytics. IEEE Signal Process Mag 31(5):32–43

Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:27. http://www.csie.ntu.edu.tw/∼cjlin/libsvm

Cho S-B, Kim JH (1995) Combining multiple neural networks by fuzzy integral for robust classification. IEEE Trans Syst Man Cybern 25(2):380–384

Das S, Kar S, Pal T (2017) Robust decision making using intuitionistic fuzzy numbers. Granul Comput 2(1):41–54

Farris FA (2010) The gini index and measures of inequality. Am Math Mon 117:851–864

Gini C (1936) On the measure of concentration with special reference to income and statistics. Colo Coll Publ 208:73–79

Grabisch M (1997) k-order additive discrete fuzzy measures and their representation. Fuzzy Sets Syst 92(2):167–189 **(Fuzzy Measures and Integrals)**

Grabisch M, Kojadinovic I, Meyer P (2008) A review of methods for capacity identification in Choquet integral based multi-attribute utility theory: Applications of the kappalab r package. Eur J Oper Res 186(2):766–785

Grabisch M, Murofushi T, Sugeno M (2000) Fuzzy measures and integrals: theory and applications. Physica-Verlag, Heidelberg

Grabisch M, Nguyen H, Walker E (1995) Fundamentals of uncertainty calculi, with applications to fuzzy inference. Kluwer Academic, Dordrecht

Grabisch M, Roubens M (2000) Application of the Choquet integral in multicriteria decision making. In: Grabish M, Murofushi T, Sugeno M (eds) Fuzzy measures and integrals: Theory and applications. Physica Verlag, Berlin, pp 348–374

Havens TC, Anderson DT, Wagner C (2015) Constructing meta-measures from data-informed fuzzy measures for fuzzy integration of interval inputs and fuzzy number inputs. IEEE Trans Fuzzy Systems 23(5):1861–1875

Havens TC, Anderson DT, Wagner C, Deilamsalehy H, Wonnacott D (2013) Fuzzy integrals of crowd-sourced intervals using a measure of generalized accord. In: IEEE International Conference on Fuzzy Systems

Hu L, Anderson DT, Havens TC (2013) Multiple kernel aggregation using fuzzy integrals. In: IEEE international conference on fuzzy systems (FUZZ-IEEE), pp 1–7

Hu L, Anderson DT, Havens TC, Keller JM (2014) Efficient and scalable nonlinear multiple kernel aggregation using the Choquet integral. In: Information processing and management of uncertainty in knowledge-based systems, vol 442, pp 206–215

Keller JM, Osborn J (1995) A reward/punishment scheme to learn fuzzy densities for the fuzzy integral. In: International fuzzy systems association world congress, pp 97–100
Keller JM, Osborn J (1996) Training the fuzzy integral. Int J Approx Reason 15(1):1–24
Klir GJ, Yuan B (1995) Fuzzy sets and fuzzy logic: theory and applications. Prentice-Hall Inc, Upper Saddle River
Kojadinovic I (2006) Minimum variance capacity identification. Q J Oper Res (4OR) 12:23–36
Kojadinovic I, Marichal J-L, Roubens M (2005) An axiomatic approach to the definition of the entropy of a discrete Choquet capacity. Inf Sci 172(1–2):131–153
Labreuche C (2008) Identification of a fuzzy measure with an l1 entropy. In: Proc. of IPMU, pp 1476–1483
Leinster T, Cobbold CA (2012) Measuring diversity: the importance of species similarity, vol 93, pp 477–489
Marichal JL (1998) Aggregation operators for multicriteria decision aid. Ph.D. thesis, University of Liege, Liege, Belgium
Marichal JL (2000) Entropy of discrete Choquet capacities. Eur J Oper Res 3(137):612–624
Melin P, Mendoza O, Castillo O (2011) Face recognition with an improved interval type-2 fuzzy logic Sugeno integral and modular neural networks. IEEE Trans Syst Man Cybern Part A Syst Hum 41(5):1001–1012
Mendez-Vazquez A, Gader P (2007) Sparsity promotion models for the Choquet integral. In: IEEE symposium on foundations of computational intelligence, pp 454–459
Murofushi T, Soneda S (1993) Techniques for reading fuzzy measures (iii): interaction index. In: Proceedings of the 9th fuzzy systems symposium, Sapporo, Japan, pp 693–696
Pawlak Z (1998) Granularity of knowledge, indiscernibility and rough sets. In: Proceedings of the IEEE international conference on computational intelligence, vol 1. IEEE, pp 106–110
Pinar A, Havens TC, Anderson DT, Hu L (2015) Feature and decision level fusion using multiple kernel learning and fuzzy integrals. In: IEEE international conference on fuzzy systems (FUZZ-IEEE), pp 1–7
Pinar AJ, Rice J, Hu L, Anderson DT, Havens TC (2016) Efficient multiple kernel classification using feature and decision level fusion. IEEE Trans Fuzzy Systems. doi:10.1109/TFUZZ.2016.2633372
Sugeno M (1974) Theory of fuzzy integrals and its applications. Ph.D. thesis, Tokyo Institute of Technology
Tahani H, Keller J (1990) Information fusion in computer vision using the fuzzy integral. IEEE Trans Syst Man Cybern 20:733–741
Tang X, Fu C, Xu D-L, Yang S (2017) Analysis of fuzzy hamacher aggregation functions for uncertain multiple attribute decision making. Inf Sci 387:19–33
Tehrani AF (2013) Learning nonlinear monotone classifiers using the Choquet integral. PhD dissertation
Tehrani AF, Cheng W, Dembczyński K, Hüllermeier E (2012) Learning monotone nonlinear models using the Choquet integral. Mach Learn 89(1–2):183–211
Tehrani AF, Hüllermeier E (2013) Ordinal choquistic regression. In: EUSFLAT conference
Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B 67:91–108
Tikhonov AN (1943) on the stability of inverse problems. Doklady Akademii Nauk SSSR 39(5):195–198
Wagner C, Anderson DT (2012) Extracting meta-measures from data for fuzzy aggregation of crowd sourced information. In: IEEE Int. Conf. Fuzzy Systems, pp 1 –8
Wu Q, Wang Z, Deng F, Chi Z, Feng DD (2013) Realistic human action recognition with multimodal feature selection and fusion. IEEE Trans Syst Man Cybern Syst 43(4):875–885
Xu Z, Gou X (2017) An overview of interval-valued intuitionistic fuzzy information aggregations and applications. Granul Comput 2(1):13–39
Xu Z, Wang H (2016) Managing multi-granularity linguistic information in qualitative group decision making: an overview. Granul Comput 1(1):21–35
Yager R (2000) On the entropy of fuzzy measures. Fuzzy Syst IEEE Trans 8(4):453–461
Yager RR (1988) On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans Syst Man Cybern 18(1):183–190
Yager RR (2002) Uncertainty representation using fuzzy measures. IEEE Trans Syst Man Cybern Part B (Cybernetics) 32(1):13–20
Yang R, Wang Z, Heng PA, Leung KS (2008) Fuzzified Choquet integral with a fuzzy-valued integrand and its application on temperature prediction. IEEE Trans Syst Man Cybern Part B (Cybernetics) 38(2):367–380
Yao J (2005) Information granulation and granular relationships. In: Proceedigns of the IEEE international conference on granular computing, vol 1. IEEE, pp 326–329