5-9-2015

Insights and Characterization of l1-norm Based Sparsity Learning of a Lexicographically Encoded Capacity Vector for the Choquet Integral

Titilope Adeola Adeyeba

Insights and characterization of l1-norm based sparsity learning of a lexicographically encoded capacity vector for the Choquet Integral

By

Titilope Adeola Adeyeba

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Electrical and Computer Engineering
in the Department of Electrical and Computer Engineering

Mississippi State, Mississippi

May 2015

Insights and characterization of l1-norm based sparsity learning of a lexicographically encoded capacity vector for the Choquet Integral

By

Titilope Adeola Adeyeba

Approved:

_____

Derek T. Anderson
(Major Professor)

_____

Nicholas H. Younan
(Committee Member)

_____

Sherif Abdelwahed
(Committee Member)

_____

James E. Fowler
(Graduate Coordinator)

_____

Jason Keith
Interim Dean
Bagley College of Engineering

Name: Titilope Adeola Adeyeba

Date of Degree: May 8, 2015

Institution: Mississippi State University

Major Field: Electrical and Computer Engineering

Major Professor: Dr. Derek T. Anderson

Title of Study: Insights and characterization of l1-norm based sparsity learning of a lexicographically encoded capacity vector for the Choquet Integral

Pages in Study: 43

Candidate for Degree of Master of Science

This thesis aims to simultaneously minimize function error and model complexity for data fusion via the Choquet integral (CI). The CI is a generator function, i.e., it is parametric and yields a wealth of aggregation operators based on the specifics of the underlying fuzzy measure. It is often the case that we desire to learn a fusion from data and the goal is to have the smallest possible sum of squared error between the trained model and a set of labels. However, we also desire to learn as "simple'' of solutions as possible. Herein, L_1-norm regularization of a lexicographically encoded capacity vector relative to the CI is explored. The impact of regularization is explored in terms of what capacities and aggregation operators it induces under different common and extreme scenarios. Synthetic experiments are provided in order to illustrate the propositions and concepts put forth.

DEDICATION

I dedicate this thesis to God.

ACKNOWLEDGEMENTS

I acknowledge my advisor, Dr Derek T. Anderson for his relentless effort, help and guidance towards attaining my degree. I also acknowledge Olufemi A. Asafa for his continuous support and encouragement. I acknowledge my son, Olujimi J. Asafa for cooperating with me to make this a possibility. Lastly, I acknowledge Jesus Christ for His love towards me and for giving me this opportunity.

.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

h	$\Re$-valued integrand, $h: X \rightarrow [0,1]$
C_g	Choquet FI with respect to FM g
X	Set of information sources, $X = \{x_1, x_{2,\dots}\}$
T	Training data,$T = \{(O_j, \alpha_j): j = 1, \dots, m\}$
m	Number of training data elements
N	Number of information sources
g	Fuzzy measure (aka normal and monotone capacity)
$g_{i1}, g_{i2}, \dots, g_{ik}$	Lexicographic ordering of g, i.e., $g(\{x_{i1}, \dots, x_{ik}\})$
u	FM vector, $u^t = (g_1, g_2, \dots, g_{12}, \dots, g_N)^t$
$\|\|u\|\|^2_p$	Lp-norm regularization term

CHAPTER I

INTRODUCTION

Overview

In many fields, multiple sources (e.g., sensors, humans or algorithms) are needed in order to achieve some goal. The data (or information) from these sources can be large, potentially heterogeneous and fusion often changes from one application to another. Data/information aggregation is the study of intelligent ways to combine inputs to reach a single result that is hopefully more accurate or reliable than an answer obtained from just one input alone. Aggregation is not fusion, meaning fusion is something much more than just aggregation. However, fusion is an extremely hard concept to define and it is often very illusive. Aggregation on the other hand is a more specific topic and something that can be better defined and a science built around. One of the extreme challenges of fusion is discovering functions to carry out aggregation and subsequently identifying ways to tailor these functions to different problems and application domains. Therefore, it is of great interest to rigorously study different mathematics to learn and tailor aggregation based on information such as training labels and a desire to have as simple of solutions as possible for task at hand due to reasons such as financial and/or computational concerns.

Many authors have defined aggregation differently based on the specifics of their respective fields. In [1], Grabisch defines it as the fusion of several inputs values into a single output. In [2], L.Hu et al. describe it as a tool that can be used in kernel theory to

provide an elegant way to map multi-source heterogeneous data into a single combined homogeneous (implicit) space for pattern recognition (feature level fusion). In [3], Joint Directors Laboratory (JDL) define fusion as "data fusion is the process of combining data to refine state estimates and predictions". Other definitions and models exist as well, e.g., Dasarthy's functional model [4], the TRIP model [5] and the Omnibus model [6]. The point is, many have tried to define fusion and have come up short due to being overly general or overly specific. Regardless of its lack of sufficient definition, aggregation and fusion are basic concerns for all kinds of knowledge based systems like signal/image processing, decision making, pattern recognition and machine learning. This impacts a number of fields such as multi-criteria decision making [7], sensor fusion [7], decision-making [7] and data mining [7]. The mechanics of fusion can, and do change, e.g., the form of rules, neural networks and variation in terms of underlying theory such as probability, possibility and/or evidence theory.

Common examples of algorithms for fusion include Bayes-based techniques and the fuzzy integral (FI). In this work, the Choquet integral (CI) (a specific type of FI) is used [8, 9, 10, 11]. The CI is a well-known aggregation operator that is a function generator, i.e., it is a parametric function that yields a wealth of aggregation operators based on the particulars of the underlying fuzzy measure (FM), aka monotone and normal capacity [8, 9, 10, 11]. One of its major advantages is that it models and uses rich information about the various interactions across different inputs.

The uses of the FI cannot be over stated. It has been used in different domains and problems such as image processing [12], multi-criteria decision making [13], skeletal age-at-death estimation in forensic anthropology [14], multi-source (e.g., feature,

algorithm, sensor, confidence) fusion [15,16], used as a distance metric [17], classification [18], and pattern recognition [19, 20]. The FI is most often used to combine the (often objective) support in some hypothesis, e.g., algorithm outputs or confidences, from multiple inputs with the (often subjective) worth of the different subsets of sources, encoded in a FM. However, it is also of great utility for combining evidence as well as signal data/information. Most applications rely on the real-valued integrand and capacity, however numerous extension exist for higher-order uncertainty, e.g., unrestricted type-1 fuzzy set-valued integrands [21] and type-2 valued integrands [22].

In [23], a technique was put forth to learn the FI, specifically the CI from data. That work is unique because it attempts to also minimize model complexity. However, that article focused solely on the mechanics of carrying out the task in the context of quadratic programming (QP), not the true meaning and characterization under different scenarios. Herein, the goal of this thesis is to formally study and characterize the proposed methods so as to know what it is really doing in different cases. This aids in understanding what complexity means relative to the CI, when one should use such a procedure and when it breaks down and new research is needed. Figure 1 is a high-level illustration of this thesis and sub-sections are labeled relative to the different concepts.

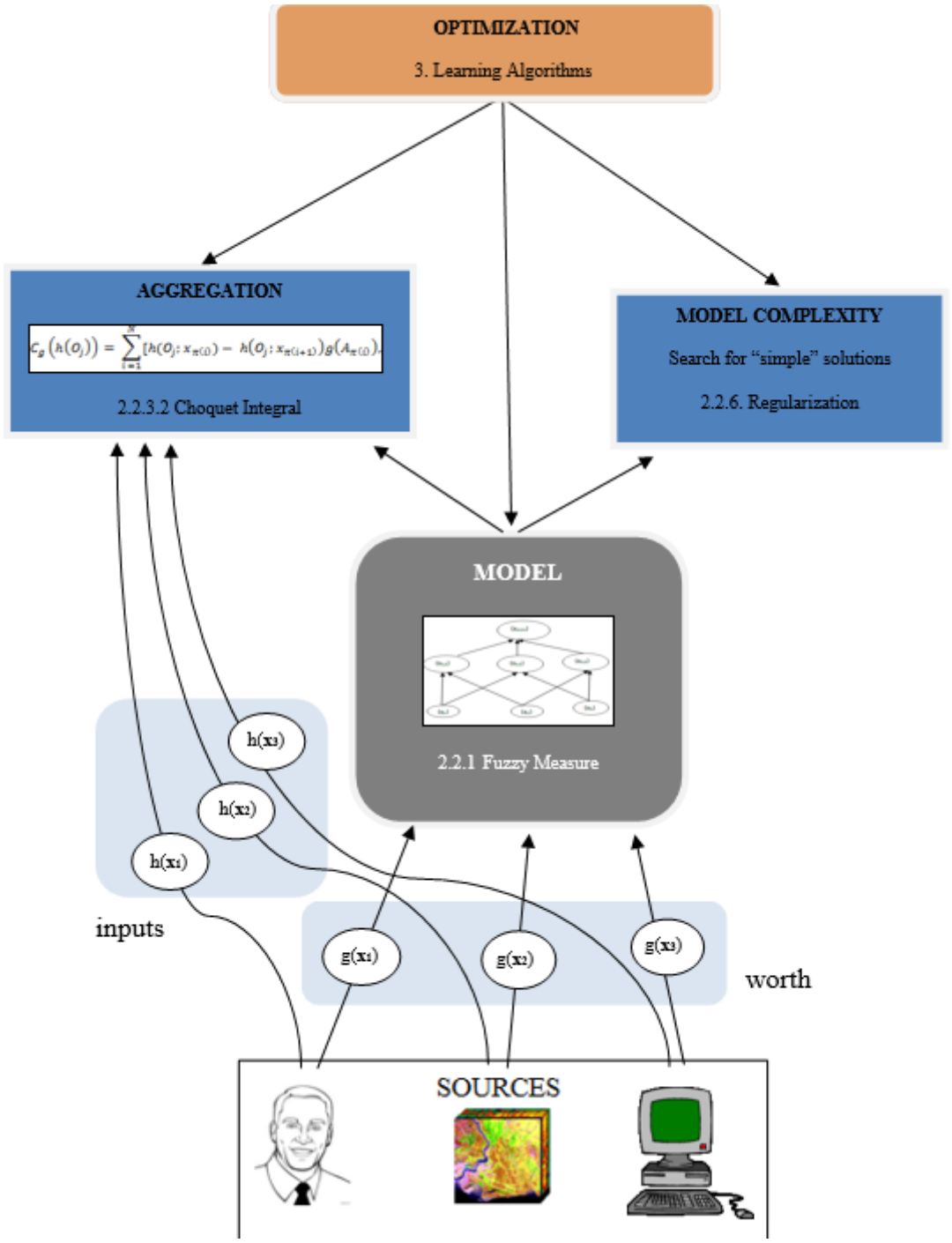

Figure 1 High level overview of thesis.

Figure 1 illustrates the major components of this thesis. First, data or information is generated from a source (human, sensor or algorithm). Next, each input provides some data/information for a task at hand. This data/information is fused using the CI (Section 2.2.3.2). However, the "worth" of different combinations is studied in terms of FM theory (Section 2.2.1). As I also seek low complexity models (FMs), Section 2.2.6 details what low complexity means and how that is measured on the FM. Last, the focus of this thesis is to learn the FM relative to the CI. Section 3 outlines a learning algorithm.

Challenges

The focus of this thesis is the use of regularization to define complexity for a FM and algorithms to use it to learn the FM from data. Specifically, my focus is the formal identification and analysis of different important conditions, in the form of remarks and propositions, regarding what happens when this procedure is used to seek low complexity solutions in conjunction with a criteria like the sum of squared error (SSE). There are a number of challenges that this thesis had to address, however the following three are the major challenges that had to be overcome.

- How to measure the complexity of a FM: A capacity is a set-valued function with 2^N values for N inputs. It is not trivial to summarize this (exponentially large) structure. Regularization is sought in order to summarize the information content of a capacity.
- Analysis and insights: If we measure the complexity of a capacity using regularization, then (at least) two questions arise. What impact does regularization under different scenarios have in terms of measure theory? Also, what is the impact of regularization under different scenarios in

terms of aggregation operators that the CI induces? This thesis studies these two questions in great depth.

- Recommendations: After analysis and insights are discovered, we must transition into understanding how the procedure operates under different conditions. This thesis tries to inform a reader about when to seek regularization and under what conditions is it not optimal (or does not produce a result that I otherwise would desire).

Contributions

Numerous works have been put forth for FM learning relative to the FI (the subject of Section 2.1.2). However, only one other recent work [24] (a conference proceeding) explored the minimization of SSE and model complexity simultaneously. They also took the approach of regularization and used a Gibbs sampler. Specifically, no rigor or analysis was provided in that work. The other related work is from Anderson et al. [23], in which a L_1-norm regularizer was introduced in conjunction with QP. In this thesis, I go beyond these two preliminary works and formally study this subject in greater depth. Specifically, my contributions to this topic include the following.

- Characterization: the formal study, in terms of both remarks and propositions, of common and extreme scenarios encountered in L_1-norm regularization of a lexicographically encoded capacity vector for the CI. I study the cases of when an exact capacity is required and when irrelevant and low quality inputs exist.

- Insights and recommendations: this thesis also sheds light on what types of measures are discovered under what conditions, what aggregation operators are unearthed, and ultimately under what scenarios does this type of approach make sense to use and how can one decide when to use it or detect undesirable conditions and avoid them.

Publications

I submitted the following conference paper to FUZZ-IEEE 2015.

- **T. A. Adeyeba**, D. T. Anderson and T. C. Havens "Insights and Characterization of L_1-Norm Based Sparsity Learning of a Lexicographically Encoded Capacity Vector for the Choquet Integral" *FUZZ-IEEE*, submitted Feb, 2015.

I am also a co-author of the following journal article (currently under review).

- D. T. Anderson, A. Zare, T. C. Havens, **T. A. Adeyeba**, "Information Theoretic Regularization of the Choquet Fuzzy Integral", *IEEE Trans. Fuzzy Systems*, submitted Jan, 2015.

Thesis Organization

In Chapter II, important concepts are defined and reviewed. In section III, the new methods put forth are discussed. Table I is the notation used throughout this thesis.

CHAPTER II

BACKGROUND

Background Work

In this background section, I review different related works. First, I discuss the basic theories and applications of the FI. Next, I review methods for learning the FM. It should be mentioned up front that the FI has been applied to numerous applications. As a result, the nature of the data/information can (and does) vary. Inputs to the FI range from low-level signal information to multi-spectral information to features and decisions. There are also many FI works focused on fusing uncertain data/information, e.g., interval-valued, set-valued (probability or possibility distributions), etc. The versatility of the FI is one of the benefits of this fusion philosophy. As a result, I have studied core topics in FM and FI theory and as applications to date have shown, their applicability of what I have found to different signal and image processing and computer vision tasks is already well established.

Theories and applications of the fuzzy integral

In addition to what was discussed in the introduction section, here are some examples of how the FI has been applied in different domains. In [8], Sugeno showed that the Sugeno integral can be applied to fuzzy inference. A fuzzy inference system (FIS) uses fuzzy set theory to map inputs (e.g., features for classification) to outputs (or

classes in classification) [25]. FIS is useful for tasks like prediction even in light of knowledge of the underlying physical process [26]. In light of the FIS adopting the use of fuzzy sets to map inputs, I proceed next to the importance of integration of information from a variety of sources. In [11], Keller et al. developed a new method of evidence fusion based on the FI that combines objective evidence (fuzzy membership functions) with the subjective worth of the sources. One of the properties of this method is its applicability to information fusion in computer vision. Decision making is a note-worthy process that has benefitted from the FI. In [27] Grabisch studied the properties of FMs and integrals inside the framework of multicriteria decision making (MCDM). He compared the FI to other aggregation operators. Other tools like the weighted sum, min, max and ordered weighted average (OWA) are too easily interpretable on the semantic point of view and also no one has been able to represent them in a way that is easily understandable. The FI is void of these drawbacks and therefore relatively works better for decision making.

Even fields outside the engineering community have not been left out in the good of the FI. In [14], Anderson et al. introduced a novel method to estimate adult skeletal age-at-death estimation using the Sugeno FI for forensic science in Anthropology. They did this by taking a multi-hypothesis testing approach to make the classical FI yield a fuzzy set-valued result based on interval-valued sources of information (aging methods). It was shown that quantitative results for summarizing the FS and comparing the single decision to a known age-at-death. They generated linguistic descriptions to establish domain standardization for assisting forensic and biological anthropologists.

Many researches today thrive on measuring distances (proximity measures and metrics) for different purposes. The FI has been a useful tool in this domain. Gader et al. proposed a method of how to apply the CI to this end. In [28], they showed that the discrete CI defines a metric if the corresponding measure satisfies certain monotonicity constraints, thereby completely characterizing the class of measures that induce a metric with the CI. In addition, the kernel-trick is a well-known way to map data from lower dimensions into higher dimensions in order to measure the similarity (inner product) of the data elements without ever explicitly performing the mapping. This is an important concept in pattern recognition (clustering and classification). The FI has also been a useful tool in this area. In [2] L.Hu et al. proposed the use of the FI for multiple kernel aggregation (FI-MK). After studying various FI formulations, they concluded that the CI for matrix wise sorting works whereas the SI does not work for per-element and matrix-wise sorting.

Recently, a number of data-driven fuzzy measure (FM) learning techniques have been introduced for the FI. Examples include, QP and evolutionary optimization. In [23], Anderson and Price explore a regularization approach to learning the FM for the Choquet FI. They put forth a L_1--norm regularization approach to reduce the complexity of a learned capacity in combination with minimization of SSE. As mentioned earlier, this thesis is a theoretical further investigation of the preliminary work put forth in [23].

Now that I have discussed a few of the different applications and theories of the FI, learning the FM for the FI will be reviewed. Specifically, I discuss past works that are related to this thesis.

Different Methods for Learning the Fuzzy Measure

The FM determines the behavior (what specific aggregation operator is induced) of the FI. There are different types of measures e.g., the S-Decomposable FM, the Sugeno λ-FM [8], and Grabisch's k-additive FM [29]. Also, there are different approaches that exist to learn the FM from data. For example, Grabisch introduced the use of the QP in [9]. QP involves optimizing (minimizing or maximizing) an objective function subject to bounds, linear equality, and inequality constraints. Grabisch has shown that the measure for each class can be learned using the QP. This approach is useful but it requires a least squares objective function to derive a QP. The problem with the QP is that it is computationally prohibitive with large data sets and non-robust in light of noisy data. Other approaches include the use of gradient descent [30] and penalty and reward [31]. Keller et al. talked about the chain of uncertainties that develop from using the FI in a decision making environment. To overcome such uncertainties, they presented a neuron model for using the FI in multiclass decision making. They also created a way to train the fuzzy densities (measure on just the individual inputs, not any combinations of inputs) from labeled data. This training algorithm uses a reward and punishment scheme to increase the reliability of the decision making process. Furthermore, Gader used a Gibbs sampler to learn the entire FM [24]. He presented a novel algorithm for learning FMs for the CI. His method uses a hierarchical model that implements a sparsity promotion algorithm through a Gibbs sampler. In [21], Anderson et al. introduced a genetic algorithm (GA) for higher-order (type-1) fuzzy set-valued FMs relative to (type-1 valued) integrands. In [15], Anderson et al. put forth a new method to automatically acquire, and subsequently aggregate, measures of specificity and agreement based on the

notion of crowd sourcing. That specific approach is of benefit when the worth of the individuals is not known but has to be extracted from data based on agreement (conflict). In the next section I review basic concepts and mathematics in FM and FI theory.

Review of Fuzzy Measures and Fuzzy Integrals

Fuzzy Measure (aka Monotone and Normal Capacity)

Measures are a fundamental concept in mathematics, especially as it relates to integrals with respect to a measure. A key property of FMs is that they require the property of monotonicity with respect to set inclusion, a far weaker property than the additive property of a probability measure. Specifically, the FM, a normal and monotone capacity, is a set-valued function, $g: 2^X \rightarrow [0,1]$, where $X = \{x_1, x_2 \ldots, x_N\}$ is our various data or information source, that has the following properties.

P1. (Boundary condition) $g\ (\emptyset) = 0$ (and often $g\ (X) = 1$);

P2. (Monotonicity) If $A, B\ \subseteq\ X$ and $A\ \subseteq B, g(A) \leq g(B)$.

Note, there is a third condition in the case of infinite sets but it is a moot point for finite sets (which are of interest here because I always work with a finite set of inputs in real-world applications). As already stated, the capacity has 2^N values, actually $2^N - 2$ due to the two boundary conditions, that must be specified or learned. As the number of values is exponential in N, it is not typically the case that one specifies the FM. Figure 2 is an illustration of the lattice induced by the FM.

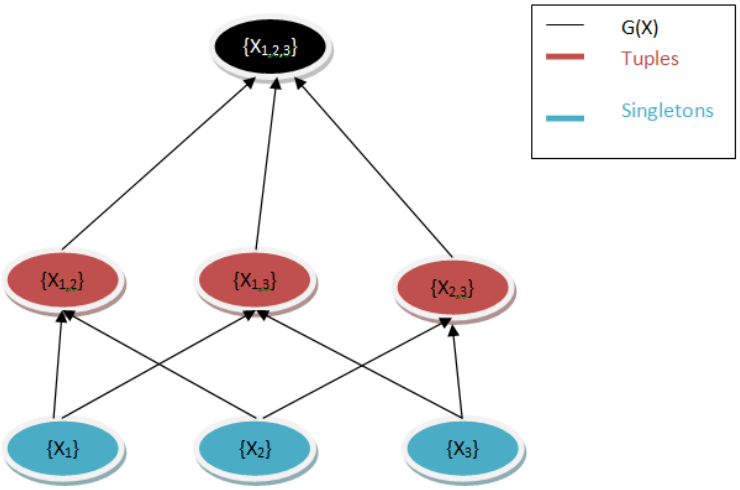

Figure 2 Lattice depicting a FM for N=3.

Notes: The first layer consists of the singletons. The second layer is the tuples and the third layer is all inputs.

A common way of acquiring the FM is to provide the value of the different singletons (which are often called the densities), i.e., $g\ (x_i)$, then use a method like the Sugeno λ-fuzzy measure [1]; which has the following additional property,

$$\text{If } A, B \subseteq\ 2^X \text{ and } A \cap B = \emptyset, \tag{1}$$

$$g(A\ \cup\ B)\ =\ g(A)\ +\ g(B)\ +\lambda_g g(A)g(B), \tag{2}$$

where λ_g is found from just the densities by solving Sugeno's famous polynomial [1].

Other note-worthy examples of measures are S-Decomposable FMs, Belief and Plausibility theory, Grabisch's k-additive FM [29]. For example, Let S be a t-conorm (generalization of an union operator). An FM g is called an S-decomposable measure if

$$g(\emptyset) = 0, g(X) = 1, \tag{3}$$

and for all A, B such that $A \cap B =\ \emptyset$,

$$g(A \cup B) = S\big(g(A), g(B)\big) \tag{4}$$

One famous example is the possibility measure, a W^* decomposable measure, where W^* is the Lukasiewicz *t*-conorm.

Aggregation Operators

Before diving into the CI, I first review some basic concepts related to aggregation. First, there are numerous different aggregation operators in the literature. A selective set has been identified and are discussed below.

The aggregation of N numbers is typically a function $F: [0,1]^n \rightarrow [0,1]$. Note, I have restricted my analysis to the interval [0,1] here for notation simplicity and convention (as many decision makers provide numbers between 0 and 1 in support of a hypothesis or I am concerned with probabilities in [0,1]). However, without loss of generality, F can (and has been) used to combine data/information in many ranges, e.g., $[-\infty, \infty]$. Applied to values $a_1, a_2, \dots, a_N$, function F produces a new number y, i.e.,

$$y = F(a_1, a_2, \dots, a_N). \tag{5}$$

Some important properties include the following.

- Continuity: i.e., F is a continuous function.
- Boundedness: i.e., $\beta_1 \leq F(a_1, a_2, \dots, a_N) \leq \beta_2$. For example, a common case is $\min(a_1, a_2, \dots, a_N) \leq F(a_1, a_2, \dots, a_N) \leq \max(a_1, a_2, \dots, a_N)$.
- Idempotency: i.e., $a_1 = a_2 = \cdots = a_N = a$, then $F(a_1, a_2, \dots, a_N) = a$.
- Monotonicy: i.e., for $(a_1, a_2, \dots, a_N)$ and $(b_1, b_2, \dots, b_N)$, if $a_i \geq b_i$ for all i, then $F(a_1, a_2, \dots, a_N) \geq F(b_1, b_2, \dots, b_N)$.

The following sub-sections are examples, not a comprehensive list, of different commonly encountered aggregator operators.

Averaging operators:

Commonly encountered, but important nonetheless, averaging operators include the following.

- Generalized means:

$$F_\alpha(a_1, a_2, \dots, a_N) = \left(\frac{a_1^\alpha + a_2^\alpha + \cdots + a_N^\alpha}{N}\right)^{\frac{1}{\alpha}} \tag{6}$$

for $\alpha \in \Re$, and $\alpha \neq 0$, and for $\alpha < 0$ $a_i \neq 0$

- Geometric mean:

$$\lim_{\alpha \to 0} F_\alpha(a_1, a_2, \dots, a_N) = (a_1, a_2, \dots, a_N)^{\frac{1}{N}}, \tag{7}$$

- Harmonic mean:

$$F_{-1}(a_1, a_2, \dots, a_N) = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2}, \dots, \frac{1}{a_N}} \tag{8}$$

- Arithmetic mean:

$$F_1(a_1, a_2, \dots, a_N) = \frac{1}{n}(a_1 + a_2 + \cdots + a_N) \tag{9}$$

Ordered Weighted Averaging Operations (OWA)

For a (weighting) vector $W = (w_1, \dots, w_N)$, where $w_i \in [0,1]$ and $\sum_{i=1}^{N} w_i = 1$, consider a permutation on $(a_1, \dots, a_N)$, $(b_1, \dots, b_N)$, such that $b_i \geq b_j$, $i \geq j$. An OWA is

$$F_w(a_1, a_2, \dots, a_N) = w_1 b_1 + w_2 b_2 + \cdots + w_N b_N. \tag{10}$$

Common OWAs include the maximum, i.e., $W = (1, 0, \dots, 0)^t$, the minimum, i.e.,

$W = (0, \dots, 0,1)^t$, the mean, $W = (1/N, 1/N, \dots, 1/N)^t$, trimmed mean, median, soft maximum and minimum, etc.

Now that I have briefly described different types of aggregation operators, I move onto the FI. The FI is a parametric function, with respect to the FM, which often reproduces common aggregator operators such as the ones listed above.

Fuzzy Integral

Choquet Integral with Respect to Training Data

Let a set of training data, T, be

$$T = \{(O_j, \alpha_j) : j = 1, \dots, m\}, \tag{11}$$

where $O = \{O_1, \dots, O_m)$ is a set of objects and α_j are the labels. The discrete CI for a finite X and object O_j is

$$C_g\left(h(O_j)\right) = \sum_{i=1}^{n} [h(O_j; x_{\pi(i)}) - h(O_j; x_{\pi(i+1)})] g(A_{\pi(i)}) \tag{12}$$

for $A_{\pi(i)} = \{x_{\pi(1)}, \dots, x_{\pi(i)}\}$, and permutation π such that

$$h(O_j; x_{\pi(1)}) \geq \cdots \geq h(O_j; x_{\pi(N)}). \tag{13}$$

The FM and CI are not trivial to understand. For example, I am interested in determining what the "worth'' is of a single input or what the "interaction'' strength is between two (or more) inputs. In order to summarize complex capacity behaviors, information theoretic indices have been put forth. In the next sub-section, I explore one such index that helps us ultimately better understand the impact of ℓ_1-norm regularization of lexicographically encoded measure vectors.

Fuzzy Measure Information Theoretic Index

The Shapley index of g are

$$\phi_g(i) = \sum_{K \subseteq X \backslash \{i\}} \eta_X(K)(g(K \cup i) - g(K)), \tag{14}$$

where

$$\eta_X(K) = \frac{(|X|-|K|-1)!|K|!}{|X|!}. \tag{15}$$

Note $X\backslash\{i\}$ denotes all subsets from X that do not include input i. The Shapley value of g is a vector $\phi_g = (\phi_g(1), \ldots, \phi_g(N))^t$ such that $\sum_{i=1}^{N} \phi_g(i)$=1. The Shapley values can be interpreted as the average amount of "contribution'' of source i across all coalitions. Basically, Equation 14 is the weighted sum (positive-valued) of the numeric differences between consecutive steps (layers) in the measure (a lattice).

In many cases, our goal is to seek and eliminate *irrelevant* or *low quality* inputs to find less complex solutions. The Shapley values give us a notion of the worth of each input. However, I really need an index that provides a scalar number that is 0 when there is no complexity and a 1 when we have the most complex model (FM). I introduce the following index as a measure of model complexity,

$$s(g) = (-1) \sum_{j=1}^{n} \phi_g(j) \ln\left(\phi_g(j)\right). \tag{16}$$

Note, this function, Shannon's entropy of the Shapley values, is 0 for the case of all inputs required, i.e., $\phi_g(j) = 1/N$, and 1 when only a single input is of value 1 and all other values are 0.

Now that we know what the Shapley does, i.e. identify the worth of each input, we are pointed towards what inputs can be "safely" eliminated. In the next subsection, I

review the un-regularized way to learn the CI based on QP. This gives us an idea of what is going on when all inputs are included in the measure.

Un-Regularized Learning of Choquet Integral

Let the SSE between the CI and T defined with respect to capacity g, be

$$E_{1,g} = \sum_{i=1}^{m} \left(C_g \left(h(O_j) \right) - \alpha_j \right)^2 . \tag{17}$$

Equation (17) can be expanded as follows;

$$E_{1,g} = \sum_{i=1}^{m} \left(\mathrm{A}_{O_j}^{t} \mathrm{u} - \alpha_j \right)^2 \tag{18}$$

where

$$A_{O_j} = \begin{pmatrix} \dots \\ h(O_j; x_{\pi(1)}) - h(O_j; x_{\pi(2)}) \\ \dots \\ 0 \\ \dots \\ h(O_j; x_{\pi(N)}) - 0 \\ \dots \end{pmatrix}, \tag{19}$$

which is of size $(2^n - 1) \times 1$. The function differences, i.e., $h(O_j; x_{\pi(i)}) - h(O_j; x_{\pi(i+1)})$, corresponds to their respective g locations in u, the lexicographically encoded measure vector,

$$u = (g_1, g_2, \dots g_{12}, .. g_{12\dots \mathrm{N}})^t, \tag{20}$$

which is of size $(2^n - 1) \times 1$. Expanding Equation 17 further,

$$E_{1,g} = \sum_{i=1}^{m} (\mathrm{u}^t \mathrm{A}_{O_j} \mathrm{A}_{O_j}^{t} \mathrm{u} - 2\alpha_j \mathrm{A}_{O_j}^{t} \mathrm{u} + {\alpha_j}^2), \tag{21}$$

$$= \mathrm{u}^t \mathrm{D} \mathrm{u} + \mathrm{f}^t \mathrm{u} + \sum_{j=1}^{m} \alpha_j^2,$$

$$\mathrm{D} = \sum_{j=1}^{m} \mathrm{A}_{O_j} \mathrm{A}_{O_j}^{t}, \quad \mathrm{f} = \sum_{j=1}^{m} (-2\alpha_j \mathrm{A}_{O_j}). \tag{22}$$

In addition, the capacity has $N(2^{N-1}-1)$ monotonicity constraints, which can be represented in a compact linear form.

$$\mathrm{Cu}+\mathrm{b}\leq\ 0, \tag{23}$$

where

$$\mathrm{C}=\begin{pmatrix}\Psi_1^t\\ \Psi_2^t\\ \cdots\\ \Psi_{N+1}^t\\ \cdots\\ \Psi_{N(2^{N-1}-1)}^t\end{pmatrix}, \tag{24}$$

where Ψ_1 is a vector representation of constraint 1, $g_1\ -g_{12}\ \leq\ 0$. For $\Psi_1^t\mathrm{u},$ one recovers $\mathrm{u}_1-\mathrm{u}_{N+1}$. Thus, C is nothing more than a matrix of $\{0,1,-1\}$ values,

$$\mathrm{C}=\begin{bmatrix}1&0&\ldots&-1&0&\ldots&\ldots&0\\ 1&0&\cdots&0&-1&\cdots&\cdots&0\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\ 0&0&\cdots&0&0&\ldots&1&-1\end{bmatrix}, \tag{25}$$

which is of size $(N(2^{N-1}-1))$x(2^N-1). In addition, b is a vector of all 0s. Note, in some works, u is of size $2^{N-1}-2$, as $g\ (\emptyset)=0$ and $g\ (X)=1$. Therefore, vector b is typically a vector of 0s and the last N entries are of value -1. I used the $2^{N-1}-1$ notation as it simplifies (notationally) the subsequent Shapley mathematics. Given T, the search for g reduces to a QP of the form

$$\min_{\mathrm{u}}\frac{1}{2}\mathrm{u}^T\widehat{\mathrm{D}}\mathrm{u}+\mathrm{f}^T\mathrm{u} \tag{26}$$

subject to

$$\mathrm{Cu}+\mathrm{b}\geq\ 0,\quad\ (\mathbf{0},1)^t\ \leq\mathrm{u}\leq\mathbf{1}. \tag{27}$$

Note, Equation 21 and 26 d*i*ffer only in the fact that $\widehat{\mathrm{D}} = 2\mathrm{D}$ and our inequality need only be multiplied by -1. Now what happens when regularization term is included? The next subsection answers this question.

Optimization for L_p-Norm Regularization Term

There has been a lot of work on solving the problem of convex unconstrained optimization in areas of machine learning, statistics and signal processing. In general, the problem of interest is one of

$$\min_{\mathrm{u}} \frac{1}{2}\|\mathrm{Gu} - \mathrm{h}\|_2^2 + \lambda\|\mathrm{u}\|_p^2, \tag{28}$$

where $\mathrm{u} \in \Re^l$, $\mathrm{h} \in \Re^k$, G is a $k \times l$ matrix, λ is a non-negative parameter and $\|u\|_p^2$ is the L_p-norm of u. The inclusion of the regularizer term works to produce solutions of u that also have a small $\|u\|_p^2$. When $p = 1$, this drives the elements of u to 0 (promoting sparsity in the solution). Another common choice is the case of $p = 0$, which counts the number of non-zero values. The basic idea behind regularization is to seek solutions that have the fewest number of parameters as possible, it is often used for parameter selection, but it can also be used to help seek simpler solutions and address overfitting. It has been shown that the ℓ_1-norm versus ℓ_2-norm leads to sparser models that can often be (more) easily interpreted [28]. In general, the ℓ_2-norm does not promote sparsity. Also, higher λ values for the ℓ_2-norm tend to force the coefficients to actually be more similar to each other (to jointly minimize the 2-norm). In [33], it was shown that a weighted iterative approach to ℓ_1-norm regularization can be taken to find even more sparse solutions (in which a different λ_k is used for each regularization term). In [23], Anderson et al. used

LASSO to solve measure learning relative to the Choquet integral and the ℓ_1-norm of a lexicographically encoded measure vector. We discussed the Tibshirani Method [34] and the Non-Negative Variable Method (NNVM) [23]. They elected to use NNVM as it is a more efficient method. In summary, in [23] we put forth a procedure to optimize Equation 17 for regularization-based measure learning,

$$E_{2,g} = \sum_{i=1}^{m} \left(u^t A_{O_j} A_{O_j}^t u - 2\alpha_j A_{O_j}^t u + {\alpha_j}^2 \right) + \lambda \|u\|_1^2, \tag{29}$$

subject to

$$\mathrm{Cu} + \mathrm{b} \geq 0, \quad (\mathbf{0}, 1)^t \leq \mathrm{u} \leq \mathbf{1}. \tag{30}$$

Specifically, the objective function in this minimization is convex and the constraints define a convex set (giving rise to a convex optimization task). Two simple, but not necessarily scalable, optimization solutions were proposed by Tibshirani [35]. Numerous solutions exist to solve this problem, e.g., active set method and local linearization [36, 37], iterated ridge regression [38], grafting [39], shooting [40], etc.

One solution (aka Tibshirani's Method) is to convert the regularization term into a set of inequalities. One linear inequality is created for each combination of the signs of elements in x, i.e.,

$$\begin{aligned} +X_1 + X_2 + \cdots + X_l &\leq t \\ +X_1 + X_2 + \cdots - X_l &\leq t \\ -X_1 - X_2 + \cdots - X_l &\leq t \end{aligned} \tag{31}$$

where t is inversely proportional to λ. For a vector of length l, there is therefore 2^l linear inequalities. Again, the above is simple to understand, but not that scalable. A second,

and more efficient, solution (aka the Non- Negative Variable Method [26]) involves doubling the number of variables in x, i.e., $\{{X_1}^+, {X_2}^+, \dots, {X_1}^-, {X_2}^-, \dots\}$,where $X_i = {X_i}^+ - {X_i}^-$. There are $2l + 1$ constraints.

$$\begin{aligned} {X_i}^+ \geq 0, {X_i}^- \geq 0, \\ \sum_{i=1}^{n} \left({X_i}^+ + {X_i}^-\right) \leq t. \end{aligned} \tag{32}$$

Another well-known formulation (basis pursuit criterion) is

$$\min_X \|X\|_p^2 \tag{33}$$

subject to

$$\|GX - h\|_2^2 \leq \sigma, \tag{34}$$

a linear program subject to quadratic inequalities. In some applications σ can often be easier to specify (versus t).

CHAPTER III

INSIGHTS AND CHARACTERIZATION

In this section, I dig deeper and investigate the theoretical impact of ℓ_1-norm regularization of lexicographically encoded capacity vectors. Specifically, I ask a number of questions to help gauge what is going on with respect to measure theory and the CI (in terms of what aggregations are induced by a learned capacity). A range of different scenarios encountered in practice are explored to help the reader better understand when and why to apply such a technique. These insights and characterizations are important and unique to the CI. That is, they differ from regularization of support vector machines, sparsity learning for machine learning, signal processing and statistics.

CASE 1: Exact Capacity Required

The idea behind Case 1 is that the solution at hand requires a specific capacity, and therefore specific aggregation operator with respect to the CI, and any other answer leads to an increase in SSE. This scenario is addressed on two fronts: (Case 1.A) the general case of any capacity and (Case 1.B) the specific case of an OWA [13]. I am interested in studying how regularization responds to such a scenario. Ideally, regularization would be *kind* in such a condition and it would not make us deviate away from the desired solution. I would like for regularization to help with factors such as removing *low quality* and/or *irrelevant* inputs and with overfitting, but I do not want

regularization to otherwise hinder other commonly encountered and natural scenarios (such as the desire to learn an OWA, an extremely common aggregation operator encountered in practice).

Before I dive into the following two sub-sections, I must first review the OWA. An OWA is defined as

$$f_W^{OWA}(\mathrm{a}_1, \ldots, \mathrm{a}_\mathrm{N}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{W_j a}_{\pi(\mathrm{j})} \tag{35}$$

where $\mathrm{a}_{(\mathrm{j})}$ is a permutation on the inputs a such that $\mathrm{a}_{(1)} \geq \cdots \geq \mathrm{a}_{(\mathrm{N})}$ and $W=(w_1, \ldots, w_N)^{\mathrm{t}}$ is a vector of (positive valued) weights that sum to 1. In terms of the Choquet integral [43], an OWA is simply a capacity with the following property:

$$g(A) = g(B) \text{ for } A, B \subseteq 2^X \text{ when } |A| = |B|. \tag{36}$$

Common OWAs include; maximum, $W = (1, 0, \ldots, 0)^t$, minimum, $W = (0, \ldots, 0, 1)^t$, mean, $W = (1/N, 1/N, \ldots, 1/N)^t$, trimmed mean, median, soft maximum and minimum, etc. The point is, the OWA is an extremely common set of operators used in practice and valid operators that may be learned for a given task in the context of CI learning. Next, I review the general case of regularization for a specific capacity.

CASE 1.A: Regularization Impact on Capacities

We start our analysis by considering Proposition 1.

Proposition 1.

Let g^* be the minimum SSE solution for the task at hand. If any ℓ_1-norm regularization of a lexicographically encoded capacity vector is used, i.e., $\xi > 0$, then the result is an increase in the SSE.

Proof.

Trivial. Any ℓ_1-norm regularization, i.e., $\xi > 0$, drives one or more of the $g(A)$ (for $A \subseteq 2^X \backslash X$) terms to 0, meaning it lessens one or more capacity values moving us away from the minimum SSE solution, g^*. ■

While simple, Proposition 1 tells us that any use of regularization works to promote sparsity and it is not selective in the respect that if a specific capacity is required it will try to keep driving capacity values towards 0 regardless. The next few remarks give us insight into what regularization is actually doing.

Remark 1.

As $\xi \to 0$, $E_{2,g}$ reduces to $E_{1,g}$, i.e., I am minimizing SSE (when ξ=0, Equation 29 is Equation 21).

Remark 2.

As $\xi \to \infty$, the regularizer term in $E_{2,g}$ dominates the objective value, resulting in a capacity vector of 0s (except $g(X) = 1$). As $\xi \to \infty$, the regularizer term dwarfs the SSE term. The result therefore has a unique minimum with respect to the regularizer: $g(A) = 0, A \in X$ s.t. $A \neq X$. What is interesting (but well-known by some) is this informs us that optimization is driven by ξ and ultimately the regularizer and the SSE term are not complementary but competing.

Remark 3.

The result of ℓ_1-norm regularization as $\xi \to \infty$ is the minimum operator. As shown in [41], a CI for a capacity of all 0s, except $g(X) = 1$, is a minimum operator with weights $(0, 0, \ldots, 1)^t$. Remark 2 shows us that as $\xi \to \infty$ this is what the regularizer

promotes. However, I note that it is not just at $\xi = \infty$ that I get this behavior. As Experiment 1 will show, this is the case when the regularizer term is relatively large in comparison to the SSE term. Figure 1(f) shows that we get essentially get all 0s at the simpler case of $\xi = 1000$.

Remark 4.

Remark 1 described what aggregation operator is being promoted as ξ becomes relatively large (minimum operator). In a measure theoretic respect, this is a state of *total ignorance*, as we have $g(X) = 1$ yet $g(A) = 0, A \in X$ s.t. $A \neq X$. While this seems extreme, it is rationalized as such. In lue of knowledge about the SSE, we have no truly helpful information to exploit. Therefore, the solution is to take a pessimistic route. It is interesting to note that the extreme case is an OWA (a minimum operator).

CASE 1.B: Ordered Weighted Average

As already discussed, the aim of regularization is to seek less complex, but still accurate, models. However, if a problem truly requires all inputs and if the required aggregation operator is an OWA, which means that all inputs are equally important, then by definition I have the highest possible model complexity (in terms of the Shannon entropy of the Shapley values). The problem I am faced with is this: I want to acquire minimum SSE, but we cannot simultaneously obtain it and minimum model complexity. The result is that any ℓ_1-norm regularization gives sub-optimal performance. However, if I am learning the capacity from data and do not know that the answer requires an OWA, then the take away is that the use of any ℓ_1-norm regularization negatively impacts performance and I am not privileged to know this ahead of time. This is a downfall of ℓ_1-

norm regularization of a lexicographically encoded capacity vector. Proposition 1 already informed us about this behavior (in the general case). It told us that any use of regularization has the impact of working to promote sparsity and it is not selective in the respect that if a specific capacity is required then it will try to keep driving those values towards zero regardless. While I am discussing the familiar scenario of OWAs in Case 1.B, other well-known fuzzy measures exhibit this property on occasion, e.g., those derived from the densities in which the densities have equal value, including the Sugeno λ-FM and the S-Decomposable measure.

Experiment 1

In this first experiment, I explore the case of three inputs with 500 randomly selected data points. I use an OWA with weights $(0.5, 0.5, 0)^t$ to generate the labels. I vary the ℓ_1-norm regularizer from 0 to 10 in step sizes of 0.001 (and an extreme case of $\xi = 1{,}000$). Some of these values are selected for visualization in Figure 2. Figure 3 shows plots of SSE against the regularizer value and Shannon entropy of the Shapley.

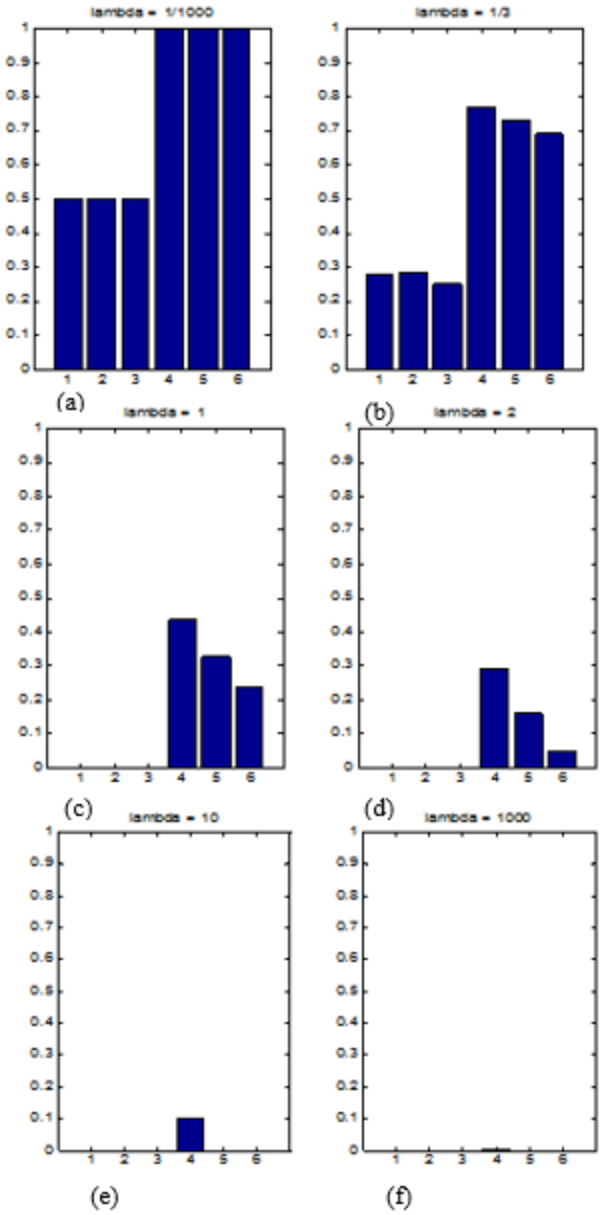

Figure 3 The effect of different ξ values

Notes: (a) $\xi = 0.001$, (b) $\xi = 0.33$, (c) $\xi = 1$, (d) $\xi = 2$, (e) $\xi = 10$, (f) $\xi = 1000$. The x-axis is lexicographically ordered capacity terms and y-axis is capacity values.

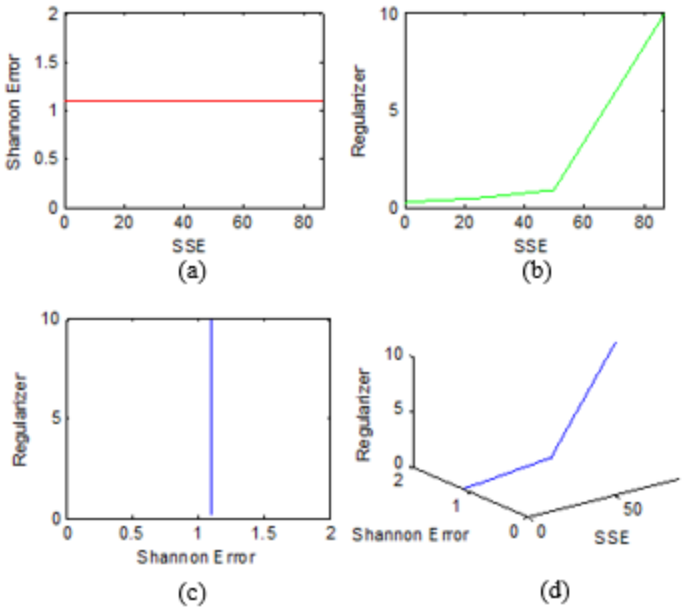

Figure 4 Plot showing relationship between the SSE, the Shapley entropy & ξ.

Notes: (a) Plot showing relationship between ξ and the Shapley entropy. (b) Plot showing relationship between ξ and the SSE. (c) Plot showing relationship between the SSE and Shapley entropy. Last, (d) plot showing relationship between the SSE, the Shapley entropy and ξ.

In Figure 3, we see that with little-to-no regularization we obtain the target OWA, $(0.5, 0.5, 0)^t$. However, as ξ grows the capacity drives towards all zeros (which is still an OWA). Figure 4 shows that as the regularizer increases, the SSE also rises. Furthermore, we see that as the regularizer increases, the Shapley entropy remains constant (as at each step I effectively have an OWA which is by definition the most complex model in the Shannon error of the Shapley index). Next, we review an index of similarity to an OWA.

Definition 1

[42]**.** The distance of g to an OWA is

$$D_{OWA} = \frac{1}{N-1}\sum_{k=1}^{N-1}\sqrt{T_1}, \tag{37}$$

$$T_1 = \frac{\sum_{I \in L(k)} (g(I) - T_2)^2}{|L(k)| - 1}, \tag{38}$$

$$T_2 = \frac{\sum_{I \in L(k)} g(I)}{|L(k)|} - \frac{\sum_{J \in L(k-1)} g(J)}{|L(k-1)|}, \tag{39}$$

where layer k in the measure is given by $L(k)$, i.e.,

$$\begin{aligned} L(1) &= \{x_1,\ x_2,\ x_3\}, \\ L(2) &= \{\ \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}\ \}, \\ L(3) &= \{\{x_1, x_2, x_3\}\}, \end{aligned} \tag{40}$$

for $N = 3$. The value $D_{OWA} \in [0,1]$. It is up to the user to determine what value of τ to use or how to determine τ automatically (as OWA-like is a fuzzy concept).

Remark 5.

If the answer to some problem is an OWA, then any ℓ_1-norm regularization pushes us away from such a goal. In order to address this challenge, I propose the following. First, learn the capacity without regularization. After this, measure the degree to which the resultant capacity is an OWA. If the degree is below a threshold, τ, then declare the capacity too much like an OWA and do not use regularization. However, if the degree is above τ and the user wants to still seek a simpler model, then regularization can still be used.

CASE 2: Irrelevant and Low Quality Inputs

First, I introduce notation to help us compactly express the following remarks. Let R be the set of relevant inputs (specifically a set of indices), let I be the set of irrelevant inputs, and let L be the set of low quality inputs. Input k is referred to as low quality if $\phi_g(k) \ll \phi_g(j)$, where $j = \text{argmin}_{z \in R}\, \phi_g(z)$ and $\phi_g(k) \neq 0$. Thus, low quality inputs

have Shapley values that are relatively small. Input k is irrelevant if $\phi_g(k) = 0$; thus, the input has no benefit towards answering the question.

In this sub-section, I discuss the effect of ℓ_1-norm regularization on capacities that represent sets of sources that contain irrelevant and low quality inputs. The focus here, versus [23], is not experimentation but rigorous analysis (characterization and insights).

Remark 6.

When there are $|I| > 0$ irrelevant inputs, then $g(k) = 0$, where $k \in I$, and $g(B) = g(B \backslash k)$; proof follows directly from Equation 14. This remark is relatively simple to understand but it needs stating. It informs us about the conditions that must occur for $\phi_g(k) = 0$. Furthermore, it tells us that if I have any irrelevant inputs, then ℓ_1-norm regularization is once again not intelligent enough to identify such a condition and respond kindly. It instead continues to drive terms toward zero, which may not be the intended goal but it is what that technique is mathematically designed to do.

Remark 7.

I use a procedure similar to that in Remark 5. A QP can be run without regularization to identify inputs that have a Shapley value below a threshold. I remove these inputs and go back seeking a regularization solution.

Next, I explore the impact and behavior of regularization in the case of low quality inputs. These are inputs that provide relatively little benefit towards solving a task. They have some contribution toward achieving minimum SSE, however, if they are removed (excluded as an input), then SSE changes only slightly. Hence, we can often achieve a "good enough" SSE and a lower model complexity by removing these low

quality inputs. The point is this, lower model complexity can give rise to a solution that requires less memory storage, less computational resources, less financial cost (e.g., fewer sensors), etc. In many situations we are willing to sacrifice some SSE for lower model complexity. I start this exploration with Proposition 2, which enables us to better understand how low quality inputs can be addressed.

Proposition 2.

As $\xi \to \infty$, $\xi \|u\|_1^2$ dominates Equation 29 and forces the capacity to $\vec{\mathbf{0}}$, except $g(X) = 1$, resulting in a Shapley value of $\phi_g(k) = \frac{1}{N}$, $\forall k \in \{1, \dots, N\}$.

Proof.

From Equation 14, the Shapley values, $\phi_g(k)$, are simply the sum of differences in the capacity. Specifically, it is a weighted sum of differences between all sets in which k is an element, $g(K \cup k)$, and the sets excluding k, $g(K)$. The regularization term is minimized when all capacity terms are 0, except for $g(X) = 1$ (Remark 2). All Shapley value differences are 0 except one term,

$$\eta_X(X\backslash\{k\})(g(X) - g(X\backslash\{k\})) = \eta_X(X\backslash\{k\})(1 - 0), \quad \forall k \in \{1, \dots, N\}. \tag{41}$$

Thus, each Shapley value is

$$\frac{(|X| - |K| - 1)!\,|K|!}{|X|!} = \frac{(N - (N-1) - 1)!\,(N-1)!}{N!} = \frac{(N-1)!}{N!} = \frac{1}{N}, \tag{42}$$

which concludes the proof. ■

Proposition 2 tells us the following story. The ℓ_1-norm regularizer gives rise to a model with all inputs of equal worth, i.e., Shapley values of $1/N$. This is confusing as one would likely assume that a simpler model would be one such that $\phi_g(k) = 1, j \neq k, \phi_g(j) = 0$. The ℓ_1-norm regularization on the lexicographically coded capacity vector does not produce the intuitive low-complexity model that we often desire.

Empirical results tell a different story; the use of this regularization scheme appears to result in lower complexity models. This is confusing, i.e., the ability to identify results with fewer number of inputs when the regularizer is actually striving for the most complex model. It turns out that it sort of does this, but it is a difficult behavior to characterize. When one uses an *adequately* valued ξ, it gives rise to interesting results due to the interplay between the SSE and regularizer term. Meaning, when ξ=0, we do not perform any regularization; we just minimize SSE. However, as ξ starts to grow in value the optimization procedure begins to *attack* the lower quality inputs first, as they contribute less to the task. It drives their values down first, resulting in a lower complexity model with respect to the Shapley. However, as ξ continues to grow, the regularizer term becomes relatively large and drives the capacity towards a measure of ignorance—the minimum operator—and uniformly equal Shapley values. Thus, particular ξ selections seem to result in the desired behavior of reducing model complexity. However, there is a point of diminishing return. As ξ is increased to seek even simpler models—again with respect to the entropy of the Shapley values—the method starts to prefer the minimum. This is a unique behavior specific to CI learning.

Overall, we can conclude the following with respect to low quality inputs. ℓ_1-norm regularization helps remove the influence of these low quality inputs; however,

there is no guarantee that the procedure will kill them before reducing the influence of relevant inputs. However, the regularizer eventually results in the learning of a minimum operator and ignorance measure, which is a complex measure in the entropic respect. Experiment 2 illustrates the stated behavior for the case of irrelevant and low quality inputs.

Experiment 2

In Experiment 2, I use three inputs and 500 randomly selected training points. The reason for once again picking only three inputs is so we can easily visualize the algorithm output (as the number of capacity terms grows exponentially). Input 1 is given a worth of 0.85 and is therefore required to solve the task at hand. Furthermore, input 2 is a low quality input and has a worth of 0.15. Last, we let input 3 have a worth of 0; it is irrelevant to the task at hand. A possibility measure is used, thus the value at each 2- and 3-tuple is the max of the densities with respect to the elements in that set. We expect a quality learner to ignore the third input and we would like to see regularization drive the worth of the second input before attacking the first. In addition, we expect to observe a rise in SSE as we force out input two. Figures 3 and 4 illustrate the experiment.

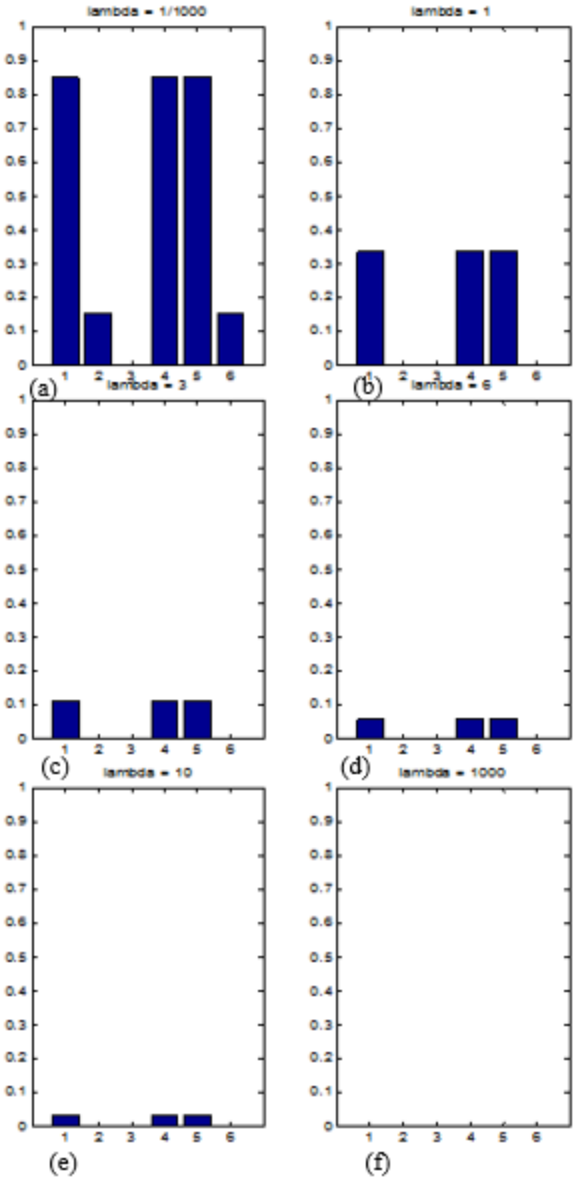

Figure 5 The lexicographically ordered FM variables learned by the QP subject to ℓ_1-norm regularization.

Notes: (a) $\xi = 0.001$, (b) $\xi = 1$,(c) $\xi = 3$, (d) $\xi = 6$, (e) $\xi = 10$, (f) $\xi = 1000$. The x-axis is lexicographically ordered capacity variables and the y-axis is the capacity value.

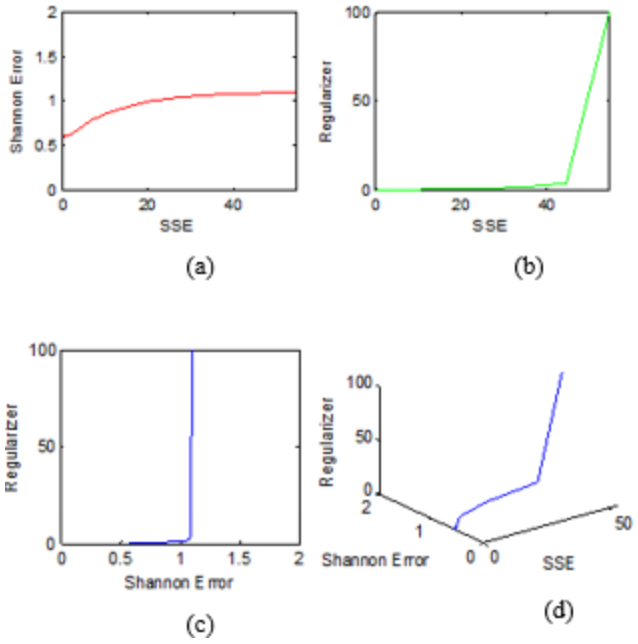

Figure 6 Plot showing relationship between the SSE, the Shapley entropy & ξ

Notes: (a) Plot showing relationship between ξ and the Shapley entropy. (b) Plot showing relationship between ξ and the SSE. (c) Plot showing relationship between the SSE and Shapley entropy. Last, (d) plot showing relationship between SSE, the Shapley entropy and ξ.

Figures 6 and 7 tell the following story. First, we find the target possibility measure for a value ξ=0.001. However, inputs two and three are still included and make for a more complex model. As we increase ξ, we eliminate the third then second input. We also see that eventually we obtain a result of all zeros (the regularizer seeks a minimum operator). This experiment reinforces the propositions and remarks made earlier.

CHAPTER IV

CONCLUSION AND FUTURE WORK

In this thesis, I focused on the simultaneous minimization of function error and model complexity for the CI. I explored the impact of ℓ_1-norm regularization with respect to a lexicographically encoded capacity vector in terms of what specific measures and aggregation operators it strives to induce. I put forth a number of propositions and remarks that showed what happens and methods to help address and remedy problems with such an approach. Overall, it is shown that such a method tries to achieve a measure of ignorance, a minimum operator and equal Shapley values. Furthermore, the true benefit of such an approach appears to be the removal of low quality inputs, which occurs at particular values of a regularizer, but it is not entirely the case as the regularizer term is increased.

This thesis helped to motivate an exploration of a more intelligent way to use an improved regularizer, versus the ℓ_1-norm regularization with respect to a lexicographically encoded capacity vector [44]. This new research is aimed at intentionally forcing out low quality inputs and it is well-suited for relevant and irrelevant inputs. This work has been documented in a paper titled “Information Theoretic Regularization of the Choquet Fuzzy Integral” that is currently under review now in the IEEE Transactions on Fuzzy Systems (submitted in Jan, 2015). We put forth two

algorithms, one based on minimization of the entropy of the Shapley values through the Gini index and another method based on direct minimization of the ℓ_1-norm of the Shapley values. The Gini index is a measure of entropy and we consider it on the Shapley index values,

$$v_G(g) = 1 - \sum_{i=1}^{N} \Phi_g\,(i)^2. \tag{43}$$

Note that $v_G(g) = 0$ iff there is a single Shapley value equal to 1 (thus all other values are 0). Also, the maximum of $v_G(g)$ occurs when all Shapley values are equal. We also show that this formula can be further simplified (conceptually) and one can achieve enhanced sparsity through reweighted ℓ_1-norm regularization of the Shapley values themselves (which gives rise to a useful iterative regularization procedure). The second approach is based on minimizing the SSE with weighted ℓ_1-norm of Shapley values and different regularization weights [44]. The SSE weighted ℓ_1-norm of Shapley values and different regularization weights is

$$E_4 = \boldsymbol{u}^t \boldsymbol{D}\boldsymbol{u} + \boldsymbol{f}^t \boldsymbol{u} + \sum_{j=1}^{m} \alpha_j^2 - (\lambda_1 \boldsymbol{\Gamma}_1 + \ldots + \lambda_N \boldsymbol{\Gamma}_N)^t\ \mathbf{u}, \tag{44}$$

The goal is

$$\min_{\boldsymbol{u}} \boldsymbol{u}^t \boldsymbol{D}\boldsymbol{u} + \left(\boldsymbol{f} - (\lambda_1 \boldsymbol{\Gamma}_1 + \ldots + \lambda_N \boldsymbol{\Gamma}_N)\right)^t \boldsymbol{u}, \tag{45}$$

subject to

$$\boldsymbol{C}\boldsymbol{u} + \boldsymbol{b} \geq \mathbf{0}, (\mathbf{0}, \mathbf{1})^t \leq \boldsymbol{u} \leq \mathbf{1}. \tag{46}$$

In [44] we showed outstanding progress towards arguably more reasonably low complexity models that a human/expert might prefer.

In the future, I plan to investigate the impact of noise and over fitting on both of the methodologies outlined in this thesis. Specifically, I would like to characterize these types of phenomena and see how the ideas put forth react (theoretically versus experimentally). Furthermore, my next step will be to take the theory developed in this thesis and apply it to different signal/image processing problems and data sets. However, we now know how these tools behave in general, so application is just a demonstration for a problem domain. Nevertheless, it will be interesting to explore different creative ways of applying it to different tasks such as signal, spectrum, algorithm and decision level fusion.

REFERENCES

[1] Grabisch, M., Murofushi, T., Sugeno, M., fuzzy measures and integrals: theory and applications, Studies in fuzziness and soft computing, Physica-Verlag, (2000)

[2] L. Hu, D. T. Anderson, and T. C. Havens, “Fuzzy integral for multiple kernel aggregation,” IEEE International Conference on Fuzzy Systems, 2013.

[3] A.N. Steinberg, C.L.Bowman, F.E. White, "Revisions to the JDL Data Fusion Model" A1AA Missile Sciences Conference, Naval Postgraduate School, Monterey, CA 17-19 November 1998

[4] Real-Time Systems: Constructs for Expressing Them, Methods of Validating Them", IEEE Transactions on Software Engineering 11(1):80-86, January, 1985.

[5] O. Kessler and B. Fabian, Estimation and ISR Process Integration, Washington D.C., Defense Advanced Research Projects Agency, 2001.

[6] M. Bedworth , J. O'Brien"The Omnibus Model:A New Model of Data Fusion?", Jemity, P.O. Box 113, Malvern,Worcestershire, WR14 3YJ, UK

[7] M. Detyniecki,” Fundamentals on Aggregation Operators”, Berkeley initiative in Soft Computing Computer Science Division University of California, Berkeley United Sates of America

[8] M. Sugeno, “Theory of fuzzy integrals and its applications,” Ph.D. thesis, vol. Tokyo Institute of Technology, 1974

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

[10] Grabisch, M., Murofushi, T., Sugeno, M., Fuzzy measures and integrals: theory and applications, Studies in fuzziness and soft computing, Physica-Verlag, (2000)

[11] Tahani, H., Keller, J., Information fusion in computer vision using the fuzzy integral, IEEE Transactions on Systems, Man, and Cybernetics, vol. 20, pp. 733-741, (1990)

[12] M. Grabisch,” Fuzzy Integral for Classification and Feature Extraction”Thomson-CSF, Corporate Research Laboratory Domaine de Corbeville 91404 Orsay Cedex, France

[13] R.R. Yager “an ordered weighted averaging aggregation operators in multi-criteria decision making” IEEE Transactions on Systems, Man and Cybernetics, 18 (1988), pp. 83–190

[14] M. Anderson, D. T. Anderson, D. Wescott, “Estimation of Adult Skeletal Age-at-Death Using the Sugeno Fuzzy Integral,” American Journal of Physical Anthropology, vol. 142 (1), pp. 30-41, 2009.

[15] C. Wagner and D. T. Anderson, “Extracting meta-measures from data for fuzzy aggregation of crowd sourced information,” in Proc. IEEE Int. Conf. Fuzzy Syst., 2012, pp. 1–8, doi: 10.1109/FUZZ-IEEE.2012.6251281.

[16] T. C. Havens, D. T. Anderson, C. Wagner, "Fuzzy Integrals of Crowd-Sourced Intervals Using A Measure of Generalized Accord," IEEE International Conference on Fuzzy Systems, 2013.

[17] J. Bolton, P. Gader, and J. N. Wilson, “Discrete choquet integral as a distance metric,” IEEE Transactions on Fuzzy Systems, vol. 16, no. 4, pp. 1107–1110, 2008.

[18] A. Mendez-Vazquez, P. Gader, J. Keller, and K. Chamberlin, “Minimum classification error training for choquet integrals with applications to landmine detection,” IEEE Transactions on Fuzzy Systems, vol. 16, no. 1, pp. 225 –238, feb. 2008

[19] M. Grabisch and J.-M. Nicolas, “Classification by fuzzy integral: Performance and tests,” Fuzzy Sets and Systems, vol. 65, no. 23, pp. 255 – 271, 1994, fuzzy Methods for Computer Vision and Pattern Recognition. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016501149490023X

[20] M. Grabisch,” A New Algorithm for Identifying Fuzzy Measures and Its Application to Pattern Recognition” Thomson-CSF, Corporate Research Laboratory Domaine de Corbeville 91404 Orsay Cedex, France

[21] D. T. Anderson, J. M. Keller, and T. C. Havens, “Learning fuzzy valued fuzzy measures for the fuzzy-valued sugeno fuzzy integral,” in International conference on information processing and management of uncertainty, 2010, pp. 502–511.

[22] T. C. Havens, D. T. Anderson, C. Wagner, "Fuzzy Integrals of Crowd-Sourced Intervals Using A Measure of Generalized Accord," IEEE International Conference on Fuzzy Systems, 2013.

[23] D. T. Anderson, S. Price, T. C. Havens, "Regularization-Based Learning of the Choquet Integral," accepted for publication in IEEE Int. Conf. Fuzzy Systems, 2014.

[24] Mendez-Vazquez, A., Gader, P., Sparsity Promotion Models for the Choquet Integral, IEEE Symposium on Foundations of Computational Intelligence, pp. 454-459, (2007)

[25] http://www.cs.princeton.edu/courses/archive/fall07/cos436/HIDDEN/Knapp/fuzzy004.htm

[26] M.H. Kazeminezhad, A. Etemad-Shahidi, S.J. Mousavi," Application of fuzzy inference system in the prediction of wave parameters" Ocean Engineering 32 (2005) 1709–1725

[27] M. Grabisch, "Fuzzy integral in multicriteria decision making," Fuzzy Sets Syst., vol. 69, no. 3, pp. 279–298, 1995. [Online]. Available:http://dx.doi.org/10.1016/0165-0114(94)00174-6

[28] J. Bolton, P. Gader, and J. N. Wilson, "Discrete choquet integral as a distance metric," IEEE Transactions on Fuzzy Systems, vol. 16, no. 4, pp. 1107–1110, 2008.

[29] P. Miranda, M. Grabisch, Characterizing k-additive measures. Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 1063-1070, (2000)

[30] Keller, J., Osborn, J., Training the Fuzzy Integral, International Journal of Approximate Reasoning, vol. 15 (1), pp. 1-24, (1996)

[31] Keller, J., Osborn, J., A Reward/Punishment Scheme to Learn Fuzzy Densities for the Fuzzy Integral, International Fuzzy Systems Association World Congress, pp. 97-100, (1995)

[32] Mendez-Vazquez, A., Gader, P., Sparsity Promotion Models for the Choquet Integral, IEEE Symposium on Foundations of Computational Intelligence, pp. 454-459, (2007)

[33] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," in Journal of the Royal Statistical Society Series B, pp. 91-108, 2005.

[34] M. Candes and S. Boyd, "Enhancing sparsity by reweighted L1 minimization," in Journal of Fourier Analysis and Applications, vol. 14, pp. 877-905, 2008.

[35] R. Tibshirani, "Regression shrinkage and selection via the lasso," in Journal of the Royal Statistical Society, Series B, vol. 58. Pp. 267-288, 1994.

[36] Osborne, M., Presnell, B., Turlach, B., On the lasso and its dual. Journal of Computational and Graphical Statistics, pp. 319337, (2000)

[37] Osborne, M., Presnell, B., Turlach, B., A new approach to variable selection in least squares problems, IMA Journal of Numerical Analysis, vol. 20 (3), pp. 389-403, (2000)

[38] Fan, J., Li, R., Variable selection via non-concave penalized likelihood and its oracle properties, pp. 1348, (2001)

[39] Perkins, S., Lacker, K., Theiler, J., Grafting: Fast, incremental feature selection by gradient descent in function space. Journal of Machine Learning Research, vol. 3, pp. 13331356, (2003).

[40] Fu, W., Penalized regressions: The bridge versus the lasso, Journal of Computational and Graphical Statistics, vol. 7(3), pp. 397416, (1998)

[41] D. T. Anderson, T.C. Havens, C. Wagner, J.M. Keller, M. Anderson, D. Wescott, "Extension of the Fuzzy Integral for General Fuzzy Set-Valued Information," in IEEE Trans. on Fuzzy Systems, vol. 22, no. 6, pp. 1625-1639, 2014.

[42] S. R. Price, D. T. Anderson, C. Wagner, T. C. Havens, J. M. Keller, "Indices for Introspection of the Choquet Integral," in 3rd Annual World Conference on Soft Computing, vol. 312, pp. 261-271, 2013.

[43] T. C. Havens, D. T. Anderson, C. Wagner, "Constructing Meta-Measures from Data-Informed Fuzzy Measures for Fuzzy Integration of Interval Inputs and Fuzzy Number Inputs," in IEEE Transactions on Fuzzy Systems, 2014. doi: 10.1109/TFUZZ.2014.2382133

[44] D. T. Anderson, A. Zare, T. C. Havens, T. A. Adeyeba, "Information Theoretic Regularization of the Choquet Fuzzy Integral", IEEE Trans. Fuzzy Systems, submitted Jan, 2015.