Regularization-Based Learning of the Choquet Integral

Derek T. Anderson, Stanton R. Price, and Timothy C. Havens

***Abstract*— A number of data-driven *fuzzy measure* (FM) learning techniques have been put forth for the *fuzzy integral* (FI). Examples include quadratic programming, Gibbs sampling, gradient descent, reward and punishment and evolutionary optimization. However, most approaches focus solely on the minimization of the *sum of squared error* (SSE). Limited attention has been placed on characterizing and subsequently minimizing *model* (i.e., FM) complexity. Furthermore, the vast majority of learning techniques are highly susceptible to overfitting and noise. Herein, we explore a regularization approach to learning the FM for the Choquet FI. We investigate the mathematical motivation for such an approach, its applicability and impact on different types of FMs, and its desirable properties for *quadratic programming* (QP) based optimization. We show that L_1 regularization has a distinct meaning for measure learning and aggregation operators. Experiments are performed and validated with respect to the Shapley index. Specifically, we show that it is possible to reduce the effect of overfitting, we can identify higher quality measures and, if desired, force the learning of fewer numbers of sources.**

I. Introduction

THE *fuzzy integral* (FI), both Sugeno and Choquet, is a well-known nonlinear flexible aggregation operator. The FI is defined with respect to a *fuzzy measure* (FM), which drives the *behavior* of the FI. While the FM can be specified by an expert, it is most often not the case because the number of free parameters is exponential ($2^N - 2$) in the number of inputs, N. A number of measures, e.g., the S-Decomposable FM, the Sugeno λ-FM [1], and Grabisch's k-additive FM [2] have been proposed that require specification of just the N densities (the measure on just the singletons) or constraints are placed on the FM to produce drastically fewer number of required parameters to work with the FM. In addition, approaches also exist to learn the FM from data. For example, Grabisch introduced *quadratic programming* (QP) [3]; Keller et al. introduced methods based on gradient descent [4] and penalty and reward [5]; Mendez-Vazquez et al. used a Gibbs sampler [6]; we introduced a *genetic algorithm* (GA) for higher-order (type-1) fuzzy set-valued FMs [7]; and linear programming [8], [9]. In addition, we introduced a way to automatically acquire, and aggregate, measures of specificity and agreement based on the notion of crowd sourcing when the worth of the individuals is not known but has to be extracted from data [10].

Most data-driven FI learning methods are focused on minimizing the *sum of squared error* (SSE). However, these approaches are severely limited as they are often highly susceptible to factors such as overfitting and noise. Furthermore, their results can often be difficult to interpret (by an expert). For example, a frequent concern is the complexity of a learned model. That is, do different FMs exist to approximately solve the task at hand? If so, which solution is preferable? While one given solution might have a slightly higher SSE, if the model (FM) is less complex, we might prefer it. In terms of a real world example, it is often the case that an expert or system desires the fewest number of inputs possible. This could be due to the cost of the resultant solution (in terms of dollars or computational complexity). While indices such as the Shapley values and the interaction index [12] exist to measure such desirable properties for a candidate solution, we are interested in discovering if it is possible to incorporate information indices directly into the learning process or seek procedures that indirectly accomplish the same goal in a different way.

Recently, a surge of interest has sparked in the fields of machine learning, statistics, and signal processing. Specifically, a flood of papers has appeared on the topic of L_p-norm based regularization and so-called sparsity promotion. In general, the objective is to reduce model complexity by searching for "sparse" solutions; those that have the fewest number of non-zero parameters possible. This approach has helped in learning models with fewer numbers of parameters and solutions that are less susceptible to overfitting and noise. However, many machine learning works use such an approach "blindly." Meaning, the goal is to reduce the number of parameters at whatever cost and many do not consider highly constrained tasks. Herein, we begin by exploring the use of L_1-norm regularization for FM learning in the context of the QP. We define a procedure, investigate its implications in terms of optimization, and we experimentally demonstrate different scenarios. However, we also explore what it means in terms of FM learning, what type of models it strives to promote, and we discuss when and where we might want to use such an approach. Furthermore, we touch on a computational method to automate some of this procedure.

In [6], Mendez-Vazquez and Gader investigated the only work we are aware of related to sparsity promotion and the *Choquet integral* (CI). In general, their goal is similar to ours; reduce the number of non-zero parameters in the FM. However, we are also concerned with what the resultant model *means* and what it is truly doing (in terms of both a measure and also aggregation). Mendez-Vazquez and Gader

Derek T. Anderson and Stanton R. Price are with the Department of Electrical and Computer Engineering, Mississippi State University, USA (email: anderson@ece.msstate.edu and sp438@msstate.edu). Timothy C. Havens is with the Departments of Electrical and Computer Engineering and Computer Science, Michigan Technological University, USA (email: thavens@mtu.edu)

This work was supported in part by the National Institute of Justice (2011-DN-BX-K838). Dr. Havens is funded in part by US Army grants W909MY-13-C-0013 and W909MY-13-C-0029 to support the U.S. Army RDECOM CERDEC NVESD, MDOT grant number 2013-0067, and the MTU Research Excellence Fund.

extended the work of Figueirdeo et al.; specifically, they used a Gibbs sampler. They did indeed put forth a creative way to enforce the FM monotonicity constraints. The Shapley values were computed and used to analyze their approach. However, two very limited cases were explored, no discussion was provided in terms of what the procedure does in a measure theory regard nor what the approach does in terms of an aggregation operator. Furthermore, their method of enforcing the constraints has a known weakness (sampling from a Gaussian with small variance). They acknowledge this and make a remark that it appears to work in practice (and remark that it is not a theoretically sound solution). It is also not clear that the method truly scales as advertised with respect to N. Furthermore, they did not perform a computational complexity analysis so it is difficult to gauge how expensive this approach is and how it scales. Last, they remark that their approach also has a problem with respect to the exponential nature of the input, but it is not as sensitive as the QP.

The authors of [6] also raise some interesting concerns regarding the use of the QP for FM learning. However, we take some objection to these comments. First, many techniques have, and continue to be proposed for solving QP with respect to fairly large and sparse matrices. This progress is coming primarily as a response of interest in machine learning, statistics and signal processing. A somewhat large and sparse constraint is not a "game stopper." We do agree, there is mathematically a point where the task at hand does become extremely difficult to solve and eventually it becomes intractable. However, the reality is that most FI applications use a relatively small number of inputs, i.e., on the order of 3 to 5 versus 10 or 50 or 100. For example, consider the case of $N = 10$. We obtain a constraint matrix of size 5110×1023, in which approximately 0.2% of the matrix is (potentially) non-zero-valued. However, as stated above, most practical applications, e.g., Abdallah's [13] context aware adaptive fusion for buried explosive hazard detection, fuse only a small number of inputs. In [13], Abdallah fused just four algorithm decisions. For $N = 4$, the matrix is of size 28×15 and 12.38% is (potentially) non-zero valued. The idea that the QP has little-to-no value just because it is difficult (eventually intractable) to solve with respect to a sparse matrix for large N is no reason to dismiss it.

This article is structured as such. In Section II we discuss the FM, FI, and the Shapley index. In Section III we discuss the SSE and the QP. Section IV proposes regularization-based optimization and Section V proposes sparsity promotion for the FM. We study its theoretical implications and meaning at the levels of measure theory and in terms of aggregation operators. Next, in Section VI, experiments are performed and the benefit of the proposed method is highlighted. Table I is the notation used in this article.

TABLE I
NOTATION

h	$\Re$-valued integrand, $h : X \to [0,1]$
C_g	Choquet FI with respect to FM g
X	Set of information sources, $X = \{x_1, ..., x_N\}$
T	Training data, $T = \{(O_j, \alpha_j) : j = 1, ..., m\}$
m	Number of training data elements
N	Number of information sources
g	Fuzzy measure (FM), $2^X \to [0,1]$
$g_{i_1,i_2,...,i_k}$	Lexicographic ordering of g, i.e., $g\left(\{x_{i_1}, ..., x_{i_k}\}\right)$
$\mathbf{u}$	FM vector, $\mathbf{u}^t = (g_1, g_2, ..., g_{12}, ..., g_{12...N})^t$
$\|\mathbf{u}\|_p^2$	L_p-norm regularization term

II. Measure and Integral

The aggregation of information using the FI, Sugeno or Choquet, has a rich history. Much of the theory and several applications can be found in [14], [3], [15]. With respect to this problem, we consider a finite set of N sources of information $X = \{x_1, ..., x_N\}$ and a function 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 usually 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 non-linear fashion. This worth is encoded in a FM. Initially, the function $h : X \to [0,1]$ and the FM $g : 2^X \to [0,1]$ were designed to take real number values in $[0,1]$. Certainly, the output range for the support function and FM can be (and have been) defined more generally (e.g., $[0, \Re]$), but it is convenient to think of them on $[0,1]$ for confidence fusion. We now review the FM and FI.

Definition 1. (Fuzzy Measure) The FM is a set-valued function, $g : X \to [0,1]$, with the following properties

P1. (Boundary condition) $g(\phi) = 0$;
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. Also, we often impose normality, i.e., $g(X) = 1$. Before a definition can be given for the FI, notation must be established for the training data used to learn the FM.

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

$$T = \{(O_j, \alpha_j) : j = 1, ..., m\},$$

where $\mathbf{O} = \{O_1, ..., O_m\}$ is a set of objects and α_j are the corresponding labels (e.g., function outputs, class labels, membership degrees, etc.). Next, we provide a definition for the FI, specifically the CFI with respect to T [3].

Definition 3. (Choquet Integral) The CI, for a finite X and

object O_j is

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

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

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

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

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

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

for $i = 1, ..., N$. Note, $X \backslash \{i\}$ denotes all subsets from X that do not include source i. The Shapley value of g is the vector $\Phi_g = (\Phi_g(1), ..., \Phi_g(N))^t$ 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 "contribution" decision based on the weighted sum of (positive-valued) numeric difference between consecutive steps (layers) in the measure (lattice).

In the following sections we discuss a natural desire to eliminate irrelevant or low quality sources and find a less complex solution. We can use the Shapley values to help identify such cases. We introduce a measure (index) of FM complexity as

$$s(g) = (-1) \sum_{j=1}^{N} \Phi_g(j) \, ln(\Phi_g(j)). \tag{3}$$

The Shapley index values sum to 1. Furthermore, when only one source is needed, one value is 1 and the other terms are 0 (i.e., Equation (3) equals 0). The more uniformly distributed the values become, the more sources are required. In the extreme case, $\frac{1}{N}$, all sources are needed and we get the largest entropy. The Shannon entropy of the Shapley terms can be used as a measure of "complexity" of a FM (relative to a desire to seek a minimal number of inputs).

III. Sum of Squared Error and Quadratic Programming

Next, we review the QP method of optimizing the FM for the FI with respect to T.

Definition 5. (SSE) Let the *sum of squared error* (SSE) between T and a FI, defined with respect to FM g, be

$$E_1 = \sum_{j=1}^{m} (C_g(h(O_j)) - \alpha_j)^2. \tag{4}$$

Equation (4) can be expanded as follows.

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

where

$$\mathbf{A}_{O_j} = \begin{pmatrix} ... \\ h(O_j; x_{\pi(1)}) - h(O_j; x_{\pi(2)}) \\ ... \\ 0 \\ ... \\ h(O_j; x_{\pi(N)}) \\ ... \end{pmatrix},$$

which is of size $(2^N - 1) \times 1$. Note, the function differences, $h(O_j; x_{\pi(i)}) - h(O_j; x_{\pi(i+1)})$, correspond to their respective g locations in $\mathbf{u}$ (which is of size $(2^N - 1) \times 1$), defined as $\mathbf{u} = (g_1, g_2, ..., g_{12}, ..., g_{12...N})^t$. Folding Equation (4) out further, we find

$$\begin{aligned} E_1 &= \sum_{j=1}^{m} (\mathbf{u}^t \mathbf{A}_{O_j} \mathbf{A}_{O_j}^t \mathbf{u} - 2\alpha_j \mathbf{A}_{O_j}^t \mathbf{u} + \alpha_j^2), \\ &= \mathbf{u}^t \mathbf{D} \mathbf{u} + \mathbf{f}^t \mathbf{u} + \sum_{j=1}^{m} \alpha_j^2, \\ \mathbf{D} &= \sum_{j=1}^{m} \mathbf{A}_{O_j} \mathbf{A}_{O_j}^t, \quad \mathbf{f} = \sum_{j=1}^{m} (-2\alpha_j \mathbf{A}_{O_j}). \end{aligned} \tag{5}$$

In total, the FM has $(N(2^{N-1} - 1))$ monotonicity constraints. These constraints can be represented in a compact linear algebra (matrix) form. The following is the minimum number of constraints needed to represent the FM. Let

$$\mathbf{C}\mathbf{u} + \mathbf{b} \leq \mathbf{0},$$

where

$$\mathbf{C} = \begin{pmatrix} \Psi_1^t \\ \Psi_2^t \\ ... \\ \Psi_{N+1}^t \\ ... \\ \Psi_{N(2^{N-1}-1)}^t \end{pmatrix},$$

where Ψ_1 is a vector representation of constraint 1, $g_1 - g_{12} \leq 0$. Specifically, for $\Psi_1^t \mathbf{u}$ one recovers $\mathbf{u}_1 - \mathbf{u}_{N+1}$. Thus, C is simply a matrix of $\{0, 1, -1\}$ values,

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

which is of size $(N(2^{N-1} - 1)) \times (2^{N-1} - 1)$. Also, $\mathbf{b}$ is a vector of all 0s. Note, in some works, $\mathbf{u}$ is of size $(2^N - 2)$, as $g(\phi) = 0$ and $g(X) = 1$. In such a case, the vector $\mathbf{b}$ is typically a vector of 0's and the last N entries are of value -1. Herein, we use the $(2^N - 1)$ format as it simplifies (notationally) the subsequent Shapley index mathematics. Given T, the search for FM g reduces to a QP of the form

$$\min_{\mathbf{u}} \frac{1}{2} \mathbf{u}^t \hat{\mathbf{D}} \mathbf{u} + \mathbf{f}^t \mathbf{u}, \tag{6}$$

subject to

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

Note, the difference between the QP and Equation (5) is $\hat{\mathbf{D}} = 2\mathbf{D}$ and our inequality in Equation 6 need only be multiplied by (-1).

IV. SSE and L_1-norm Regularization

Considerable interest has been dedicated to solving the problem of convex unconstrained optimization. Much work exists in the areas of machine learning, statistics and signal processing. In general, the problem of interest is one of

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{G}\mathbf{x} - \mathbf{h}\|_2^2 + \lambda \|\mathbf{x}\|_p^2, \tag{7}$$

where $\mathbf{x} \in \Re^l$, $\mathbf{h} \in \Re^k$, G is a $k \times l$ matrix, λ is a non-negative parameter and $\|\mathbf{x}\|_p^2$ is the L_p-norm of $\mathbf{x}$. The inclusion of the L_p-norm regularizer works to produce solutions of $\mathbf{x}$ that also have a small norm $\|\mathbf{x}\|_p$. When $p = 1$, this drives the elements of x to 0 (promoting sparsity in the solution). The basic idea is to seek solutions that involve the use of the fewest number of parameters possible. It is often used for parameter selection, however it can also be used to help seek simpler solutions and to help address overfitting. However, it was shown that the L_1-norm ($\|\mathbf{x}\|_1^2$ versus $\|\mathbf{x}\|_2^2$) leads to sparser models that can often be (more) easily interpreted [16]. In general, the L_2-norm does not truly encourage sparsity. Also, higher λ values for the L_2 term tend to force the coefficients to be more similar to each other (to jointly minimize the 2-norm). Additionally, the L_1 often outperforms the L_2-norm when irrelevant sources (features) are present in X. Equation (7) is related to the following convex constrained optimization problem (also known as LASSO [16]),

$$\min_{\mathbf{x}} \|\mathbf{G}\mathbf{x} - \mathbf{h}\|_2^2, \tag{8}$$

subject to

$$\|\mathbf{x}\|_1^2 \leq t,$$

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 for L_1 were proposed by Tibshirani [16]. Numerous solutions exist to solve this problem; active set method and local linearization [18], [19], iterated ridge regression [20], grafting [22] and shooting [23].

One solution (aka Tibshirani's Method) is to convert the L_1 regularization term into a set of inequalities. One linear inequality is created for each combination of the signs of elements in $\mathbf{x}$, i.e.,

$$\begin{aligned}
&+\mathbf{x}_1 + \mathbf{x}_2 + ... + \mathbf{x}_l \leq t \\
&+\mathbf{x}_1 + \mathbf{x}_2 + ... - \mathbf{x}_l \leq t \\
&\dots \\
&-\mathbf{x}_1 - \mathbf{x}_2 + ... - \mathbf{x}_l \leq t,
\end{aligned}$$

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 is not scalable.

A second, and more efficient, solution (aka the Non-Negative Variable Method [16]) involves doubling the number of variables in $\mathbf{x}$, i.e., $\{\mathbf{x}_1^+, \mathbf{x}_2^+, ..., \mathbf{x}_1^-, \mathbf{x}_2^-, ...\}$, where $\mathbf{x}_i = \mathbf{x}_i^+ - \mathbf{x}_i^-$. There are $(2l + 1)$ constraints,

$$\begin{aligned}
&\mathbf{x}_i^+ \geq 0, \mathbf{x}_i^- \geq 0, \\
&\sum_{i=1}^{n} \left(\mathbf{x}_i^+ + \mathbf{x}_i^-\right) \leq t.
\end{aligned}$$

Another well-known formulation (basis pursuit criterion) is

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1^2 \tag{9}$$

subject to

$$\|\mathbf{G}\mathbf{x} - \mathbf{h}\|_2^2 \leq \sigma,$$

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

V. L_1-norm Regularization for FM Learning

In this section, we explore the use of a L_1-norm regularization term for promoting sparsity of a FM solution, i.e.,

$$E_2 = \sum_{j=1}^{m} (\mathbf{u}^t \mathbf{A}_{O_j} \mathbf{A}_{O_j}^t \mathbf{u} - 2\alpha_j \mathbf{A}_{O_j}^t \mathbf{u} + \alpha_j^2) + \lambda \|\mathbf{u}\|_1^2. \tag{10}$$

In Section IV we explored the definition and methods for solving L_1 regularization-based SSE problems. The goal of this paper is to explore if a L_1 regularizer can be used to help learn FMs that are less *complex* (e.g., fewer numbers of inputs) and/or are less prone to overfitting of data? Specifically, what is the L_1-norm really doing? As we are learning the FM relative to the Choquet FI, what measure is it striving to learn and, in terms of the Choquet FI, what aggregation operator is it pushing a given solution towards? In the remainder of this section, we address these questions.

Remark 1. As λ goes to 0, Equation (10) reduces to Equation (4), i.e., minimization of strictly SSE.

Remark 2. As λ goes to ∞, Equation (10) results in a FM of all 0s, i.e., $g(A) = 0, \forall A \in X, A \neq X$. Therefore, we learn a FM corresponding to a state of total *ignorance*. While this seems extreme, it is rationalized as such. In lieu of knowledge of the SSE, we have no information to exploit. Therefore, the (trivial) solution is to set each FM value to 0 in an (extreme) attempt to force less complex models.

Remark 3. As λ goes to ∞, Equation (10) results in a Choquet integral (aggregation) that is the minimum. This is trivial to show. When sets of equal size (cardinality) have equal measure value in the FM, the Choquet integral reduces to an *ordered weighted average* (OWA). The OWA weights are $w_i = g(A_i) - g(A_i \setminus \{i\})$. The weight vector is, $\mathbf{w} = (\mathbf{w}_1 = 0, \mathbf{w}_2 = 0, ..., \mathbf{w}_n = 1)^t$. This can be rationalized as such. Without knowledge of the SSE we are forced to take an extreme pessimistic stance in terms of aggregation.

Remark 4. An L_1 regularizer is not "selective" in how it attempts to eliminate (i.e., drive to 0) variables in the FM. It regards each variable as more-or-less equally important.

Ideally, we would like to have as few inputs as possible, i.e., have a minimal number of non-zero Shapley values. However, this is not the "sole focus" of the L_1 regularization term. It is not "focused" on eliminating sources. If there are inputs (sources) that provide little-to-no benefit, the L_1 can identify and eliminate these sources (drive the densities corresponding with these sources and their corresponding measure variables towards zero).

Next, we consider a few cases related to learning from data and with respect to learning a FM and aggregation operator. These different scenarios help us better understand what the L_1 regularization is really doing (meaning what is a less complex model) and when it should be used.

A. Irrelevant Inputs

In this subsection, we explore the following scenario. Imagine that the desired FM that minimizes the SSE and has the lowest corresponding model complexity is one in which one or more of the Shapley values are approximately 0. Furthermore, due to noise or convergence to a local versus global optimum, assume we learned, without the use of a L_1 regularization term, a suboptimal solution that has non-zero densities for the irrelevant inputs and higher than desired tuple-wise values throughout the FM. In other words, there exists some sources that can simply be removed with little-to-no impact (i.e., rise in SSE). In such a case it is clear that the L_1 regularization is of benefit as it can target and help drive the corresponding (low-value source) densities to 0, and lower, as much as possible, the corresponding tuple-wise measure values to an ideal value (i.e., for a set A, it is assigned the same value as the measure $A \setminus B$, where B is the corresponding set of irrelevant sources). While the L_1-norm was not designed to explicitly operate on the basis of the Shapley values or interaction indices, it can force such a solution that has a lower $\|\mathbf{u}\|_1^2$ with no increase in SSE.

B. All Inputs Required

This subsection is focused on the case in which every input is needed to achieve minimum SSE. However, this is not a trivial scenario. Minimum SSE is not as straight forward as one might suspect. At least three scenarios exist. First, all inputs might be important, non-redundant, and required to solve the task at hand. In this case, we cannot drive any of the measure values to 0 without lessening overall aggregation performance. Second, noise (which takes on various specific meanings for different domains and applications) may be present and as a result SSE minimization alone results in learning a model that takes the noise into account. Thus, overfitting is likely to occur. We show such a scenario in the experiments section. Strictly minimizing the SSE is not always the best idea. Identifying a reduced complexity model with fewer inputs and is less overfit to the data is more ideal. This is what the L_1 regularizer helps us to do. The last scenario discussed here is sampling and overfitting. If the training data is not truly reflective of the actual underlying data distribution or the true distribution is poorly represented in the training data, overfitting can also occur and learning a "simpler" model can lead to a more reliable and generalized solution. This is a well-known phenomenon (the bias-variance dilemma).

Furthermore, consider the case of an OWA, all inputs required for minimum SSE, little-to-no noise, and adequate sampling. For example, let a set of OWA weights $\mathbf{w} = (0.5, 0.4, 0.1)^t$. The challenge here is that an OWA does not place emphasis on the individuals. Instead, it places importance on their sorted evidence. Specifically, the FM is one of $g(A) = 0.5$ for $A \in X, |A| = 1$, $g(B) = 0.4$ for $B \in X, |B| = 1$ and $g(X) = 1$. This solution cannot be reduced any further without incurring an increase in SSE (assuming all inputs are truly required). The problem is the regularizer still attempts to reduce "complexity." However, the model is already of minimum complexity. In such a scenario, the L_1 regularizer will drive the resultant solution towards a minimum operator based on the influence (scale) of λ. Note, OWAs are a wide class of important aggregation operators that are likely to be encountered in practice.

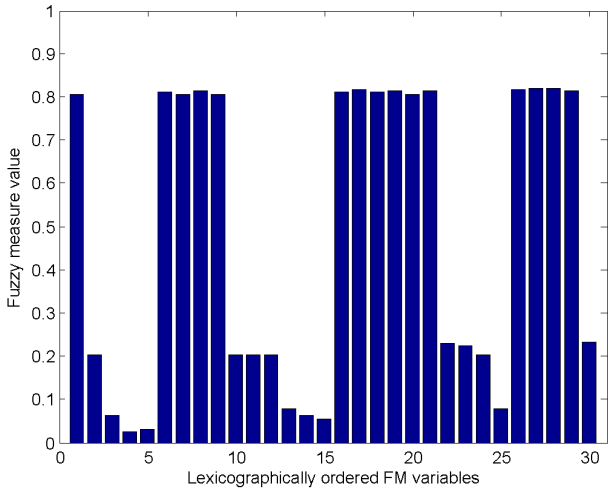
VI. Experiments

In the above sections we outlined and explored the theoretical and a computational procedure for jointly reducing model complexity with respect to a L_1-norm and SSE. In this section, we report two experiments to illustrate different important scenarios.

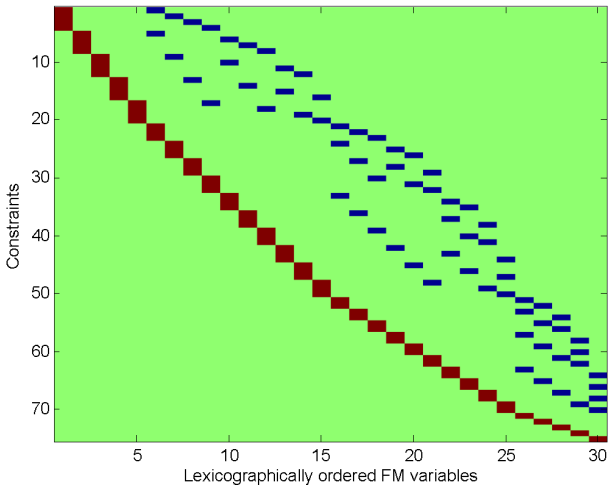
Remark 5. We recognize that for small N one could simply enumerate all combinations of features and run the QP. However, while this may be of use in feature selection, it does not address overfitting. The L_1 regularizer still adds value in such a scenario.

A. Experiment 1

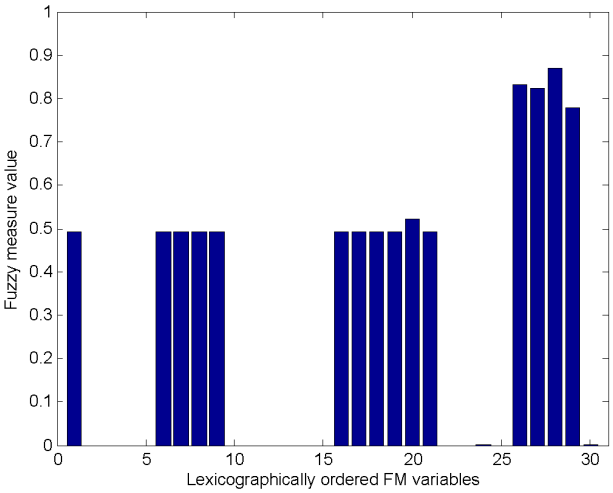
In our first experiment, we consider the case of 5 inputs. Specifically, we consider that source 1 is the most important to solving the task at hand. Furthermore, we assume source 2 has some, but very low, importance. We also assume that sources 3, 4, and 5 are not at all relative to the task at hand. Thus, a quality learner should learn to naturally eliminate the last three sources. Furthermore, we should also be able to discover a scenario in which elimination of the second source leads to increased error but results in a less complex solution. We used 300 (pseudo)randomly generated samples. The labels were generated from a FM, specifically a possibility measure, whose densities are $g(x_1) = 0.8$, $g(x_2) = 0.2$, $g(x_3) = g(x_4) = g(x_5) = 0$. In order to make the task non-trivial, we add uniformly distributed (pseudo)random noise to the desired output labels. We anticipate the following. We expect the QP without a L_1 regularizer to learn a model that puts importance on sources one and two and the final three sources will receive non-zero (but low-valued) measure values as a result of overfitting. Figure 1 illustrates the output of our system relative to varying the λ regularization value (balance of model complexity relative to the SSE).



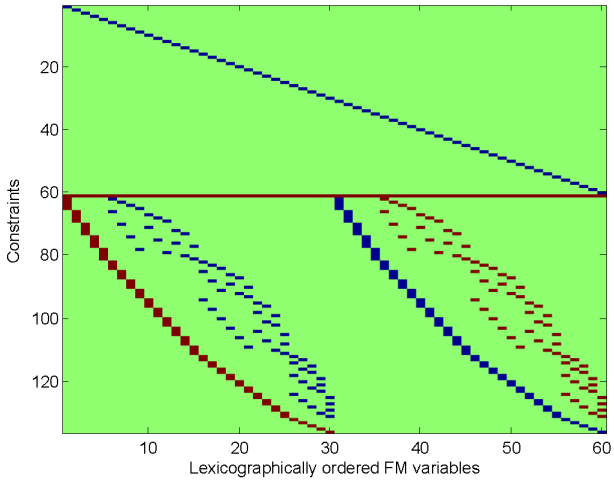
(a) Lexicographically ordered FM variables learned by QP, e.g., bin 1 is $\mathbf{u}(1) = g(x_1)$, bin $N+1$ is $\mathbf{u}(N+1) = g(\{x_1, x_2\})$.



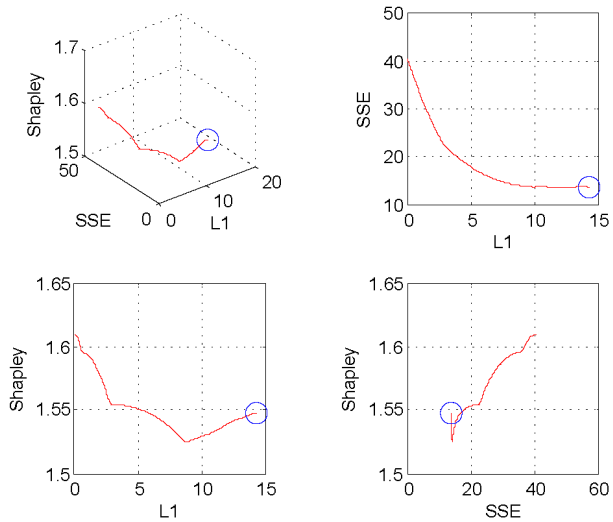
(b) Visualization of the constraint matrix C for the QP. Rows are constraints. Columns are lexicographically ordered FM variables. Green is 0-valued, -1 is shown as blue and 1-valued is shown as red.



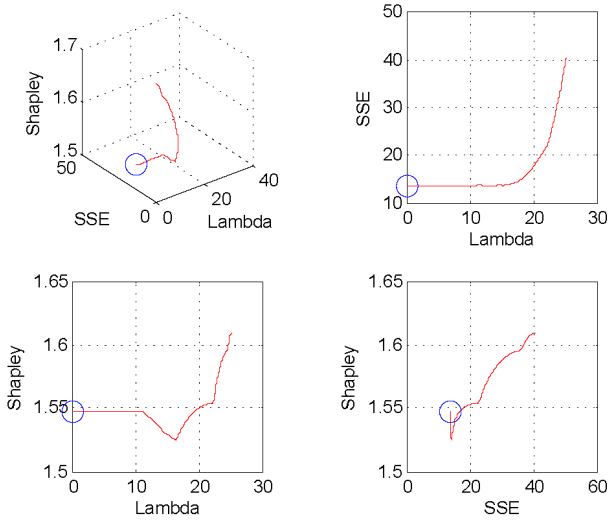
(c) Lexicographically ordered FM variables learned by the QP subject to L_1-norm regularization.



(d) Visualization of the constraint matrix C for the QP with a L_1 regularization. Same format as view (b).



(e) Plots showing variation of the Shapley entropy relative to the SSE and L_1 value. These plots show the trends in the data, such as solutions with minimum SSE and minimum Shapley value error. Blue circles indicate the QP without L_1 norm solution.



(f) Plots showing variation of the Shapley entropy relative to the SSE and λ. These plots show the trends in the data, such as solutions with minimum SSE and minimum Shapley value error. Blue circles indicate the QP without L_1 norm solution.

Fig. 1. Experiment 1 showing results of L_1-regularized FM learning. Sources 1 and 2 are relevant to aggregation, with source 1 being very important and source 2 being slightly important. Additive uniform random noise was added to training labels.

We observe the following. First, figure 1a reinforces what we suspected. The QP with no regularization learns that source one is the most important, approximately $g(\mathbf{x}_1) \approx 0.8$, the second sources is lower importance but needed, $g(\mathbf{x}_2) \approx 0.2$, and the other three sources are non-zero, but very low-valued (the result of overfitting). Figure 1b shows the constraint matrix for the QP with no regularization. It is simply provided for illustration. Figure 1e shows that a solution exists in which the SSE is not drastically higher but there is a critical point (minimum) with respect to the Shapley-based complexity. That is, a simpler solution exists in which the error is not much higher. Figure 1c is the resultant FM. It shows that only the first source is really needed. The other sources can, for all intents and purposes, be eliminated (either by thresholding the densities or computing the corresponding Shapley values and thresholding those values). Figure 1f simply shows the behavior with respect to variation in λ.

B. Experiment 2

Experiment 2 is one in which all sources are required to solve the problem. The solution is an OWA, $\mathbf{w} = (0.5, 0.35, 0.1, 0.05)^t$. We used 300 (pseudo)randomly generated samples; there are 4 inputs (thus 300 4-tuples of input from the different sources), and no noise was added to the inputs. Therefore, we expect to see our proposed method, or any method at that, result in an unavoidable increase in SSE (with respect to λ) relative to the identification of potentially less complex solutions. Furthermore, we expect to see a migration of the solution towards the minimum operator.

Figure 2b shows what we predicted; the L_1-norm learned FM is extremely similar to a minimum aggregation operator. Furthermore, Figure 2c shows that there are indeed "simpler" models, however they exist at what is most likely an unacceptable SSE. This is different from the last example we investigated, in which we found a minimum with respect to the Shapley entropy at little-to-no increase in SSE.

Remark 6. It is often the case that one desires or a problem requires an OWA. In Experiment 2 we discussed the behavior of FM learning in such a scenario. If another factor such as cost or computational runtime is more important than a higher but acceptable SSE, then the approach outlined in this article can be used "as is" to identify candidate solutions. This is a decision for the user. If the desire is to identify situtations in which the task at hand requires an OWA, i.e., all inputs are needed and are equally important, then the following index can be used [24],

$$d_{OWA} = \frac{1}{N-1} \sum_{k=1}^{N-1} \sqrt{T_1}, \tag{11}$$

where,

$$T_1 = \frac{\sum_{I \in L(k)} \left(g(I) - T_2\right)^2}{|L(k)| - 1}, \tag{12}$$

$$T_2 = \frac{\sum_{I \in L(k)} g(I)}{|L(k)|}, \tag{13}$$

where *Layer* k in the lattice, denoted $L(k)$, is the set of all FM terms for subsets of 2^X that have cardinality equal to k. For example, if $N = 3$, $L(1) = \{g(x_1), g(x_2), g(x_3)\}$, $L(2) = \{g(x_1, x_2), g(x_1, x_3), g(x_2, x_3)\}$, and $L(3) = g(x_1, x_2, x_3) = g(X)$. Thus, T_2 is the average measure value at layer k and T_1 is the variance at layer k. For an OWA, all measure values at layer k should have equal value. Therefore, a user could run the QP without a regularizer term, identify an acceptable d_{OWA} (threshold) and not use regularization if the learned FM is too similar to an OWA.

VII. Conclusion and Future Work

In this paper, we put forth a L_1-norm regularization approach to reducing the complexity of a learned FM in combination with minimization of SSE. We outlined the mathematical procedure and its optimization. However, we also analyzed the behavior of our approach in terms of measure theory and its associated aggregation operator. Experiments were performed to demonstrate and validate our procedure. Specifically, we demonstrated resilience in terms of factors like irrelevant and low importance sources and overfitting. These results were discussed using a measure of complexity based on Shannon's entropy of the Shapley values. In future work, we will investigate a method to embed the Shapley (or another information theoretic) index directly into the optimization procedure in place of the L_1-norm. We will also investigate the use of the proposed FM learning procedure on non-synthetic data for sensor processing and skeletal age-at-death in Anthropology and forensic science.

References

[1] Sugeno, M., Theory of fuzzy integrals and its application, Ph.D. thesis, Tokyo Institute of Technology, (1974).

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

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

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

[5] 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)

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

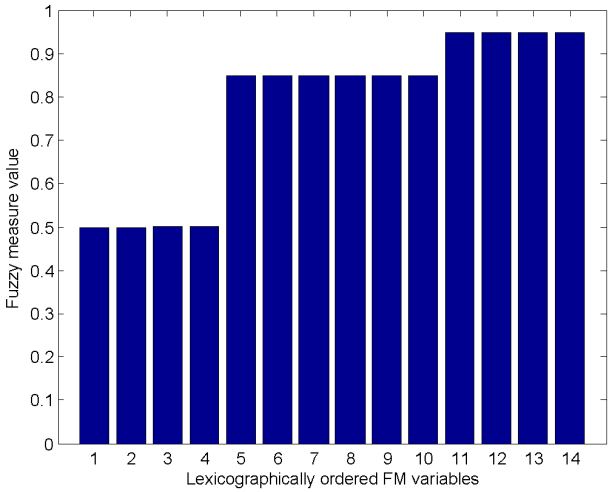
[7] Anderson, D. T., Keller, J. M., Havens, T., Learning fuzzy valued fuzzy measures for the fuzzy valued Sugeno fuzzy integral, International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 502-511, (2010)

[8] Beliakov, G., Construction of aggregation functions from data using linear programming, Fuzzy Sets and Systems, vol. 160 ,pp. 65-75, (2009).

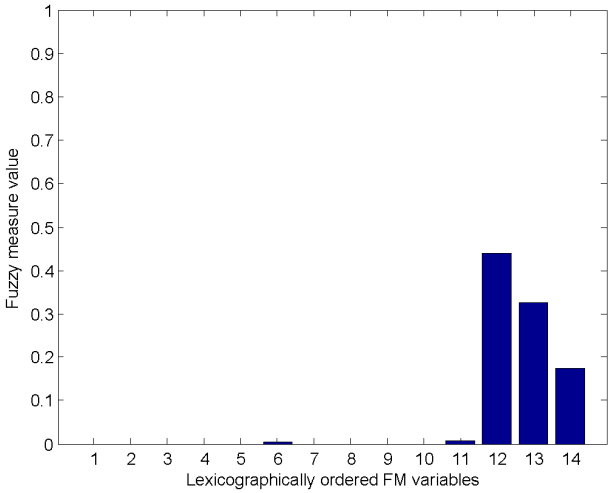
[9] Beliakov, G., James, S., Li, G., Learning Choquet integral-based metrics for semi-supervised clustering, IEEE Transactions on Fuzzy Systems, vol. 19 (3), pp. 562-574, (2011).

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

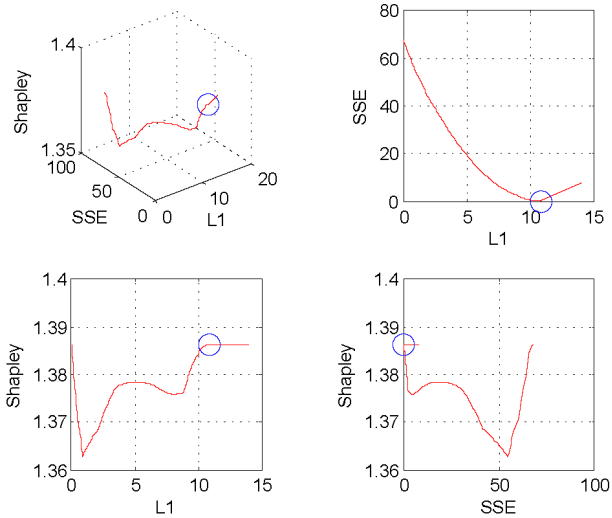
[11] Havens, T., Anderson, D.T., Keller, J.M., A Fuzzy Choquet Integral with an Interval Type-2 Fuzzy-Valued Integrand, IEEE Int. Conf. Fuzzy Systems, pp. 1-8, (2010)



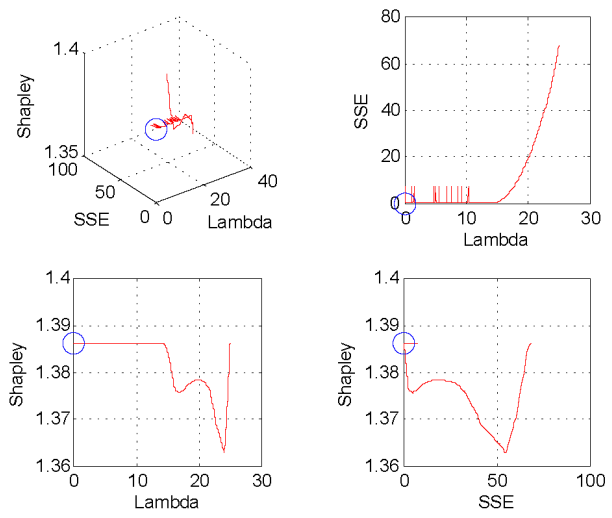
(a) Lexicographically ordered FM variables learned by the QP. Note, sets of equal cardinality have equal measure.



(b) Lexicographically ordered FM variables learned by the QP and L_1 norm (for minimum associated Shapley). Note, learned measure is almost a minimum, i.e., $g(A) = 0$, $A \subset X$.



(c) Plots showing variation of the Shapley entropy relative to the SSE and L_1 value. These plots show the trends in the data, such as solutions with minimum SSE and minimum Shapley value error. Blue circles indicate the QP without L_1 norm solution.



(d) Plots showing variation of the Shapley entropy relative to the SSE and λ. These plots show the trends in the data, such as solutions with minimum SSE and minimum Shapley value error. Blue circles indicate the QP without L_1 norm solution.

Fig. 2. Experiment 2 showing results of L_1-regularized FM learning on OWA aggregation operator.

[12] Grabisch, M., Roubens, M., Application of the Choquet integral in multicriteria decision making, Fuzzy measures and integrals, Physica Verlag, pp. 348-374, (2000)

[13] Abdallah, A. C. B., Frigui, H., Gader, P., Adaptive Local Fusion With Fuzzy Integrals, IEEE Transactions on Fuzzy Systems, vol. 20 (5), pp. 849-864, (2012)

[14] 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)

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

[16] Tibshirani, R., Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., vol. 58 (1), pp. 267-288, (1996)

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

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

[19] 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)

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

[21] Anderson, D. T., Havens, T. C., Wagner, C., Keller, J. M., Anderson, M., Wescott, D., Sugeno fuzzy integral generalizations for sub-normal fuzzy set-valued inputs, IEEE Int. Conf. Fuzzy Systems, pp. 1-8, (2012)

[22] 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).

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

[24] Price, S. R., Anderson, D. T., Wagner, C., Havens, T. C., Keller, J. M., Indices for Introspection of the Choquet Integral, 3rd Annual World Conference on Soft Computing, (2013)