# Efficient Modeling and Representation of Agreement in Interval-Valued Data

Timothy C. Havens Dept. of Electrical and Computer Eng. and Computer Science Michigan Technological University Houghton, MI, USA Email: thavens@mtu.edu Christian Wagner

Inst. Computing & Cybersystems Michigan Technological University, USA and LUCID, School of Computer Science University of Nottingham, UK Email: christian.wagner@nottingham.ac.uk Derek T. Anderson Dept. of Electrical and Computer Eng. Mississippi State University Mississippi State, MS, USA Email: anderson@ece.msstate.edu

Abstract—Recently, there has been much research into effective representation and analysis of uncertainty in human responses, with applications in cyber-security, forest and wildlife management, and product development, to name a few. Most of this research has focused on representing the response uncertainty as intervals, e.g., "I give the movie between 2 and 4 stars." In this paper, we extend upon the model-based interval agreement approach (IAA) for combining interval data into fuzzy sets and propose the efficient IAA (eIAA) algorithm, which enables efficient representation of and operation on the fuzzy sets produced by IAA (and other interval-based approaches, for that matter). We develop methods for efficiently modeling, representing, and aggregating both crisp and uncertain interval data (where the interval endpoints are intervals themselves). These intervals are assumed to be collected from individual or multiple survey respondents over single or repeated surveys; although, without loss of generality, the approaches put forth in this paper could be used for any interval-based data where representation and analysis is desired. The proposed method is designed to minimize loss of information when transferring the interval-based data into fuzzy set models and then when projecting onto a compressed set of basis functions. We provide full details of eIAA and demonstrate it on real-world and synthetic data.

#### I. INTRODUCTION

Surveys have long been a powerful way to elicit information and knowledge from groups of individuals and, recently, crowd-sourcing—surveying large groups of people to solve problems—has become a very popular research topic. The information has many uses, from analysis of results to inference of conclusions, relationships, or decision-making dynamics. Recently, there has been much interest in employing survey tools that enable the capture of response uncertainty with special emphasis on the creation of expert and decisionsupport systems that take uncertainty into account [1–4]. Fuzzy logic, in particular, has been found to be effective for analysis of survey results, with *Type-2* (T2) fuzzy logic and, the related, *Computing with Words* paradigm leading the pack [1, 3–7].

The aforementioned uncertainty in survey data comes from many factors, but, in particular, two main contributors that we address are i) *inter-expert uncertainty*: the variation among the individual opinions in a group of participants; and ii) *intra-expert uncertainty*: the variation (or vagueness) in the opinions of individual participants - usually over time. While intra-expert uncertainty can come from either dynamic



Fig. 1: Example of interval responses from a real survey, where (a) is more certain than (b). (from [8])

uncertainty—the respondent changes his answer when repeatedly queried—or indecision—the respondent is unsure about their response—we will, in this paper, simply assume that survey responses come in one of two forms: i) a T1 interval or ii) a T2 interval. Figure 1 shows two examples of what a T1 interval response looks like. In view (a), a fairly certain interval-response is shown, while view (b) shows a more uncertain interval.

Modeling and analysis of these survey-based data have been discussed in previous works [3, 4]. The interval approach (IA) and enhanced interval approach (EIA) methods show great promise, but they require data pre-processing (e.g., outlier removal) and combine all types of uncertainty into one uncertainty model, viz., the footprint of uncertainty (FOU) of Interval Type-2 (IT2) fuzzy sets (FS). In this paper, we will build on our previous works on the Interval Agreement Approach (IAA) [8]. The intervals are aggregated to produce FSs that are appropriate for the type of modeled uncertainty, i.e., inter- or intra-expert uncertainty. One of the drawbacks of IAA is that the resulting FSs are defined by (nearly) all the endpoints of the input set of intervals, which can result in an inefficient representation of the resulting FS. While the latter is not a problem in most applications where the number of intervals is manageable, in specific cases, such as the continuous collection of intervals in a cyber-security vulnerability assessment context, a more efficient modeling approach is desirable. To address this need, we propose representing the FSs as weighted sums of basis functions, where the number of weights is significantly fewer than the number of endpoints

TABLE I: Acronyms and Notation

IAA	Interval Agreement Approach	FS	Fuzzy Set
UMF	Upper Membership Function	T1	Type-1
LMF	Lower Membership Function	T2	Type-2
IT2	Interval Type-2	GT2	General Type-2
LBS	Linguistic Basis Set	GBS	General Basis Set
FOU	Footprint of Uncertainty		
-	1 2		I

Notation	Definition	
$\mu_A(x)$	Membership of $A$ at value $x$	
$ar{A} = [l_{ar{A}}, r_{ar{A}}]$	Interval between values $l$ and $r$ (see Fig. 2a)	
$\ddot{A} = [\bar{l}_{\ddot{A}}, \bar{r}_{\ddot{A}}]$	$\bar{l} = [\bar{l}_{\vec{A}}, \bar{r}_{\vec{A}}]$ Uncertain interval between values $\bar{l} = [l_{\bar{l}}, r_{\bar{l}}]$ , and	
	$\bar{r} = [l_{\bar{r}}, r_{\bar{r}}]$ (see Fig. 2b)	
i = [n]	The set of integers, $i = 1, 2, \ldots, n$	
i = [a:b]	The set of integers, $i = a, a + 1, \dots, b$	
$1_m$	<i>m</i> -length vector of ones	
$0_m$	<i>m</i> -length vector of zeroes	
$\lambda(x)$	Linguistic basis function, $\lambda : x \rightarrow [0,1]$ (see	
. ,	Definition 1)	
$\phi(x)$	General basis function, $\phi: x \to \mathcal{R}$	

in the input set of intervals—often the difference of 100s or 1,000s of endpoints versus 10s of weights. This not only produces an efficient representation of the FS, but also reduces the computational complexity of FS operations further down the analysis pipeline.

This paper is organized as follows. Section II provides a brief background on the topics addressed in this paper, while Section III describes how sets of intervals can be aggregated into FSs—namely, the IAA method. Section IV proposes the *efficient IAA* (eIAA) method and Section V demonstrates eIAA on some real survey data. Finally, Section VI briefly summarizes this paper. Select acronyms and notation are contained in Table I.

#### II. BACKGROUND

In this section we provide a *very* brief introduction to the interval and FS models used in this paper [9]. For a more thorough treatment of these topics, please see refs. [8–10].

#### A. Crisp and Uncertain Intervals

Two types of intervals are considered in this paper: i) *crisp* intervals, and ii) *uncertain* intervals [8]. These intervals are shown in Fig. 2. Crisp intervals are denoted as  $\bar{A} = [l_{\bar{A}}, r_{\bar{A}}]$ , where  $l_{\bar{A}}$  is the left endpoint and  $r_{\bar{A}}$  is the right endpoint. Uncertain intervals are denoted as  $\bar{B} = [\bar{l}_{\bar{B}}, \bar{r}_{\bar{B}}]$ , where the left and right endpoints are intervals themselves:  $\bar{l}_{\bar{B}} = [l_{\bar{L}_{\bar{B}}}, r_{\bar{L}_{\bar{B}}}]$  and  $\bar{r}_{\bar{B}} = [l_{\bar{r}_{\bar{B}}}, r_{\bar{r}_{\bar{B}}}]$ .

# B. Interval Type-2 Fuzzy Sets

GT2 FSs [9] were first introduced by Zadeh [11] as a generalization of T1 FSs. In these sets, the membership  $\mu$  at each point on the domain x is itself defined as a T1 FS. GT2 FSs introduce a burdensome computational complexity and thus many representations, simplifications, and approximations have been introduced to make application of GT2 FSs tractable [10, 12–14]. One of these is the IT2 FS [15], where the T2 FS is considered to have a secondary membership that is 1 on an interval  $[\bar{y}(x), \underline{y}(x)]$ . The primary membership function that follows  $\bar{y}(x)$  is called the *upper membership function* (UMF)



Fig. 2: Illustration of a crisp interval  $\overline{A} = [l_{\overline{A}}, r_{\overline{A}}]$ , and an uncertain interval  $\overline{B} = [\overline{l}_{\overline{B}}, \overline{r}_{\overline{B}}]$  (where for an uncertain interval each endpoint is itself an interval). (from [8])



Fig. 3: Modeling crisp intervals as FSs using IAA. (from [8])

and the membership function that follows  $\underline{y}(x)$  is the *lower* membership function (LMF), where both the UMF and LMF are T1 FSs [16].

Commonly, the membership function of a T1 FS is notated as  $\mu$ , while for T2 FSs the primary membership is denoted uand the secondary membership is denoted  $\mu$  [16]. The primary membership is on the y, or vertical, axis and the secondary membership is on the z, or "towards-you," axis. An IT2 FS is fully described by its UMF and LMF, where the IT2 FS Ahas the (secondary) membership function that has values of 1 between the UMF and LMF.

# **III. FUZZY SETS FROM INTERVALS**

The IAA approach generates FSs from sets of intervals. Two cases are considered: i) crisp intervals  $\bar{A} = [l_{\bar{A}}, r_{\bar{A}}]$ , and ii) uncertain intervals,  $\ddot{A} = [\bar{l}_{\bar{A}}, \bar{r}_{\bar{A}}]$ . For the first case, the result of IAA is a T1 FS; for the case of uncertain intervals, IAA produces an IT2 FS.

# A. Crisp intervals $\overline{A}$

Let  $\mathcal{A} = \{\overline{A}_1, \dots, \overline{A}_n\}$  be a set of intervals. Then the degree of membership of the set over the survey domain x captures the number of intervals that are overlapping—i.e., are in agreement—at a particular point. Hence, the T1 FS A defined by the membership function  $\mu_A$  is described by

$$\mu_A = \sum_{i=1}^n y_i / \left( \bigcup_{j_1=1}^{n-i+1} \bigcup_{j_2=j_1+1}^{n-i+2} \cdots \bigcup_{j_i=j_{i-1}+1}^n (\bar{A}_{j_1} \cap \cdots \cap \bar{A}_{j_i}) \right)$$
(1)

where  $y_i = i/n$  is the degree of membership, which only equals 1 at values where *all* the intervals in  $\mathcal{A}$  overlap. Essentially, what this equation shows is that the membership of  $\mathcal{A}$  at a value is equal to the number of intervals in  $\mathcal{A}$ that contain that value. We can simplify (1) for computational purposes in two different ways. First, for any value x, we can simply count the number of intervals in which x resides, computing the membership as

$$\mu_A(x) = \frac{\sum_{i=1}^n \mu_{\bar{A}_i}(x)}{n},$$
(2a)

$$\mu_{\bar{A}_{i}}(x) = \begin{cases} 1 & l_{\bar{A}_{i}} \le x \le r_{\bar{A}_{i}}, \\ 0 & \text{else.} \end{cases}$$
(2b)

The second simple formulation is convenient if one wishes to compute all the locations in the membership function where transitions from one level  $y_i$  to either  $y_{i-1}$  or  $y_{i+1}$  occur. It is easy to show that the membership function can be calculated as

$$\mu_A(x) = \frac{1}{n} \left( \sum_{i=1}^n (l_{\bar{A}_i} \le x) - \sum_{i=1}^N (r_{\bar{A}_i} \le x) \right), \quad (3)$$

where this gives the membership of A at x by counting the number of left endpoints in  $\mathcal{A}$  less than x and subtracting the number of right endpoints in  $\mathcal{A}$  less than x.<sup>1</sup> Thus, let  $X_{\mathcal{A}} = \{x_1, \ldots, x_{2n}\}$  be the set of sorted interval endpoints, where  $x_1 \leq x_2 \leq \ldots \leq x_{2n}$ , and let  $B = \{b_1, \ldots, b_{2n}\}$  be an associated set of indicator variables, where  $b_i = +1$  if  $x_i$  is a left interval endpoint and  $b_i = -1$  if  $x_i$  is a right interval endpoint. Then,

$$u_A(x_1) = 1/n, \tag{4a}$$

$$u_A(x_i) = u_A(x_{i-1}) + b_i/n, \ i = [2:2n].$$
 (4b)

Using this formulation, one can build up the membership function by looping through X, from  $x_1$  to  $x_{2n}$ . The membership at any other value x is simply

$$u_A(x) = \begin{cases} u_A(x_i) & x_i \le x < x_{i+1}, \ i = [2n-1] \\ 0 & \text{else.} \end{cases}$$
(5)

When the intervals are presented with uncertain endpoints, IAA produces IT2 FSs, which we describe next.

# B. Uncertain intervals Ä

Let  $\ddot{\mathcal{A}} = \{\ddot{\mathcal{A}}_1, \ldots, \ddot{\mathcal{A}}_N\}$  be a set of uncertain intervals, where the interval endpoints are represented as having interval uncertainty themselves. To combine uncertain intervals, we proceed similarly as for crisp intervals; however, we apply (1) independently for all "outer" and "inner" endpoints, resulting in a T1 UMF and LMF, respectively, of an IT2 FS. Let  $\ddot{\mathcal{A}}_i = [\bar{l}_i, \bar{r}_i]$  notate the *i*th uncertain interval in  $\ddot{\mathcal{A}}$ . Also, let the outermost (most uncertain) set of intervals in  $\ddot{\mathcal{A}}$  be  $\ddot{\mathcal{A}}_U =$  $\{[\bar{l}_{\bar{l}_1}, r_{\bar{r}_1}], \ldots, [l_{\bar{l}_n}, r_{\bar{r}_n}]\}$ , and the innermost (least uncertain) set of intervals in  $\ddot{\mathcal{A}}$  be  $\ddot{\mathcal{A}}_L = \{[r_{\bar{l}_1}, l_{\bar{r}_1}], \ldots, [r_{\bar{l}_n}, l_{\bar{r}_n}]\}$ . Then, mathematically, the following operations produce the IT2 FS  $\tilde{\mathcal{A}}$  that models  $\ddot{\mathcal{A}}$ ,

$$\mathrm{UMF}\left(\tilde{A}\right) = \mathrm{IAA}\left(\ddot{\mathcal{A}}_{U}\right); \ \mathrm{LMF}\left(\tilde{A}\right) = \mathrm{IAA}\left(\ddot{\mathcal{A}}_{L}\right).$$
(6)



Fig. 4: Modeling uncertain intervals as IT2 FSs using IAA uncertainty around interval endpoints is shaded. (from [8])

Figure 4 illustrates the process of using IAA to generate an IT2 FS from uncertain intervals. As was noted in our previous work [8], if the intervals in  $\ddot{\mathcal{A}}$  are crisp—i.e.,  $\bar{l} = [l, l]$  and  $\bar{r} = [r, r]$ —then the UMF simply equals the LMF, thus producing a T1 FS. In other words, IAA is generalized to produce the most parsimonious model from the input intervals: a T1 FS for crisp intervals and an IT2 FS for uncertain intervals.

# C. Computational Complexity of IAA

While IAA has been shown to be effective in representing interval survey data [8], a potential weakness is in the way it represents the FSs which model the combined interval data in particular when a high number of intervals are considered. For crisp intervals, the resulting T1 FS that represents nintervals is be modeled by (at most) 2n pairs of domain values and the membership at that value,  $\{X, \mu(X)\}^{2n}$ —assuming unique interval endpoints for the survey responses—where (5) is used to determine the membership from these pairs. For uncertain intervals both the UMF and LMF must be modeled, resulting in 4n pairs. While this is not unreasonable for modern storage, this can cause FS operations on these models to be computationally expensive.

Consider two T1 FSs as modeled by IAA,  $A \rightarrow \{X, \mu_A(X)\}^{2n}$  and  $B \rightarrow \{Y, \mu_B(Y)\}^{2m}$ , where A represents n intervals and B represents m intervals. To compute the operation  $A \cup B$ , one can apply (7) on the sorted sets produced by (5). This, however, can result in a union set that, in the worst case, is represented as 4nm pairs,  $\{X \cup Y, \mu_{A \cup B}(X \cup Y)\}^{4nm}$ . Furthermore, computing (7) costs O(n+m) computations: the cost of sorting z into X and Y. In the next section, we will show how an approximation based on a linear-constraint least squares problem can reduce this cost.

#### IV. EFFICIENT IAA

Consider a FS as modeled by IAA,  $A \to \{X, \mu_A(X)\}^{2n}$ . The eIAA method compresses the membership function  $\mu_A(x)$  by representing it as the weighted sum of a set of basis functions. We propose two different types of basis sets and methods for learning the weights on those bases. The main idea is that the 2n pairs needed to model the IAA membership

<sup>&</sup>lt;sup>1</sup>Note that the second efficient formulation of  $\mu_A$  considers the intervals to be  $A_i = [l, r)$ , where the right endpoint is not included in the interval.

$$\mu_{A\cup B}(z) = \begin{cases} \max\{\mu_A(x_i), \mu_B(y_j)\} & x_i \le z \le x_{i+1}, \ i = [2n-1], \ y_j \le z \le y_{j+1}, \ j = [2m-1], \\ 0 & \text{else.} \end{cases}$$
(7)

function are compressed into the sum of m weighted basis functions, where  $m \ll 2n$ .

We consider two different types of basis sets: linguistic basis sets (LBS) and general basis sets (GBS). The method for each type of basis is formulated to ensure that the resulting compressed membership function has membership degrees on [0,1], and that the approximation error is minimized. The LBS representation puts a constraint on the basis function and weights, which ensures that the compressed membership function is on [0, 1] throughout the entire real-valued range of the survey. This comes at a cost of the design of the LBS; i.e., not all basis sets are LBSs. With a GBS, one can use any set of basis functions on the range of the survey, but the resulting membership function is not guaranteed to be on [0, 1]for the entire survey range. It will, however, be on [0, 1] at the locations modeled by IAA: the 2n interval end-points in the survey responses. So, the GBS is flexible in that it can use any basis set, but in practice may need a clipping function to ensure that the compressed membership function returns values in [0, 1]. We now describe the compression process for each of LBS and GBS.

# A. Linguistic Basis Functions

Consider a set of linguistic basis functions  $\Lambda(x) = \{\lambda_1(x), \ldots, \lambda_m(x)\}$  on the range of the survey, as defined in Definition 1. The properties of  $\Lambda$  are such that at any value x in the survey domain, the sum of the basis functions equals 1. Since we will be modeling  $\mu_A(x)$  by the weighted sum of this basis set, this ensures we can achieve a membership at any value in [0, 1] in the survey domain.

**Definition 1** (Linguistic Basis Set).  $\Lambda(x) = \{\lambda_1(x), \ldots, \lambda(x)_m\}$  is a set of basis functions,  $\lambda_i : x \to [0, 1]$ , such that

$$\sum_{i=1}^m \lambda_i(x) = 1, \; \forall x \in \text{survey domain.}$$

We call  $\Lambda(x)$  a *Linguistic Basis Set* (LBS) as the individual basis functions  $\lambda_i(x)$  represent linguistic elements of the survey range: e.g., 'very low,' 'low,' 'medium,' 'high,' 'very high,' and, also, the set of all linguistic elements can be combined to form the core of the survey. Figure 5 shows an example of an LBS with 5 basis functions on the survey range [0, 10].

Let  $\mu_A(x)$  be represented as the weighted sum of an LBS,

$$\hat{\mu}_A(x; \mathbf{w}) = \sum_{i=1}^m w_i \lambda_i(x), \tag{8}$$

where  $\mathbf{w} = (w_1, \dots, w_m)^T$  is the vector of weights on the respective basis functions,  $\lambda_i(x)$ . If  $w_i \in [0, 1]$ , then it is easy



Fig. 5: Example of an LBS on the survey range [0, 10]—the linguistic interpretation of the basis functions could be 'very low,' 'low,' 'medium,' 'high,' and 'very high,' from left to right.

to see that  $0 \leq \hat{\mu}_A(x; \mathbf{w}) \leq 1$ ,  $\forall \mathbf{w}$  and  $\forall x$  in the survey domain. Hence, to get the "best"  $\hat{\mu}_A(x; \mathbf{w})$  representation of  $\mu_A(x)$ , we wish to solve the squared-error objective,

$$\mathbf{w}^{*} = \underset{\mathbf{w}}{\operatorname{arg\,min}} \frac{1}{2} \sum_{i} \|\mu_{A}(x_{i}) - \hat{\mu}_{A}(x_{i}; \mathbf{w})\|^{2}, \ \mathbf{w} \in [0, 1]^{m}.$$
(9)

The squared-error objective at (9) is a linearly constrained least-squares problem, which can be solved by using Matlab's lsqlin routine. However, for generalization purposes (and for later work we will propose on this problem), we now write (9) in its matrix-vector form. Let  $\mathbf{x} = (x_1, \ldots, x_{2n})^T$  be a vector of the set of values at which  $\mu_A(x)$  is modeled by IAA.<sup>2</sup> Let  $\mathbf{u}_A = (\mu_A(x_1), \ldots, \mu_A(x_{2n}))^T$  be the vector of the membership values of A at  $\mathbf{x}$ , and  $\hat{\mathbf{u}}_A(\mathbf{w}) = (\hat{\mu}_A(x_1; \mathbf{w}), \ldots, \hat{\mu}_A(x_{2n}; \mathbf{w}))^T$  is similarly defined. Finally, let  $\Lambda = [\lambda_j(x_i)]^{2n \times m}$  be the matrix where the *i*th row and *j*th column is  $\lambda_j(x_i)$ . It is easy to see that (8) is  $\hat{\mathbf{u}}_A(\mathbf{w}) = \Lambda \mathbf{w}$ . Then the squared-error term of (9) can be rewritten as

$$\begin{split} &\frac{1}{2}\sum_{i}\|\boldsymbol{\mu}_{A}(\boldsymbol{x}_{i})-\hat{\boldsymbol{\mu}}_{A}(\boldsymbol{x}_{i};\mathbf{w})\|^{2} = \\ &=\frac{1}{2}\mathbf{u}_{A}^{T}\mathbf{u}_{A}+\frac{1}{2}\hat{\mathbf{u}}_{A}(\mathbf{w})^{T}\hat{\mathbf{u}}_{A}(\mathbf{w})-\mathbf{u}_{A}^{T}\hat{\mathbf{u}}_{A}(\mathbf{w}), \\ &=\frac{1}{2}\mathbf{u}_{A}^{T}\mathbf{u}_{A}+\frac{1}{2}\mathbf{w}^{T}\boldsymbol{\Lambda}^{T}\boldsymbol{\Lambda}\mathbf{w}-\mathbf{u}_{A}^{T}\boldsymbol{\Lambda}\mathbf{w}. \end{split}$$

Since  $\mathbf{u}_A^T \mathbf{u}_A$  is a constant term, (9) can be rewritten as

$$\mathbf{w}^* = \operatorname*{arg\,min}_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \Lambda^T \Lambda \mathbf{w} - \mathbf{u}_A^T \Lambda \mathbf{w}, \ \mathbf{w} \in [0, 1]^m.$$
(10)

Equation (10) is a standard linear-constraint *quadratic pro*gramming (QP) problem,

$$\underset{\mathbf{w}}{\operatorname{arg\,min}} \frac{1}{2} \mathbf{w}^{T} H \mathbf{w} + \mathbf{f}^{T} \mathbf{w}, \ C \mathbf{w} \le \mathbf{b},$$
(11)

which can be solved by nearly any available QP solver, such as quadprog in Matlab. To express (10) as (11),  $H = \Lambda^T \Lambda$ ,  $\mathbf{f} = -\Lambda^T \mathbf{u}_A$ ,  $C = (\mathbf{1}_m^T, -\mathbf{1}_m^T)^T$ , and  $\mathbf{b} = (\mathbf{1}_m^T, \mathbf{0}_m^T)^T$ .

<sup>&</sup>lt;sup>2</sup>Note that in practice if you start with *n* intervals, there are (at most) 2n values of  $\mu_A(x)$  that are modeled by IAA.

# B. General Basis Functions

Consider a set of general basis functions  $\Phi(x) = \{\phi_1(x), \ldots, \phi_m(x)\}$  on the survey domain, where  $\phi_i : \mathbf{x} \to \mathcal{R}$ . Unlike an LBS described in Definition 1, this set of basis functions follows no specified definition other than mapping x to the real number line. Let  $\mu_A(x)$  be represented as the weighted sum of  $\Phi(x)$ ,

$$\hat{\mu}_A(x; \mathbf{w}) = \sum_{i=1}^m w_i \phi_i(x), \qquad (12)$$

where  $\mathbf{w} = (w_1, \ldots, w_m)^T$  is the vector of weights on the respective basis functions,  $\phi_i(x)$ . Rather than constraining  $\mathbf{w}$  directly to ensure that  $\hat{\mu}_A(x; \mathbf{w})$  has membership degrees on [0, 1], as we did with the LBS, the constraint for the GBS representation will be applied in the squared-error objective.

Again, consider the vector representations defined earlier,  $\mathbf{x}$ ,  $\mathbf{u}_A$ , and  $\hat{\mathbf{u}}_A$ , and let  $\Phi = [\phi_j(x_i)]^{2n \times m}$  be the matrix where the *i*th row and *j*th column is  $\phi_j(x_i)$ . Thus, (12) is  $\hat{\mathbf{u}}_A = \Phi \mathbf{w}$ . Going through the same process as we did for the LBS, the squared-error objective can be written as

$$\mathbf{w}^* = \operatorname*{arg\,min}_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \Phi^T \Phi \mathbf{w} - \mathbf{u}_A^T \Phi \mathbf{w}, \ \Phi \mathbf{w} \in [0, 1]^{2n}, \ (13)$$

where the constraint now is on  $\hat{\mathbf{u}}_A = \Phi \mathbf{w}$  to be on the interval [0, 1]. The objective function at (13) can be expressed as either a linear-constraint least square problem or as the QP at (11) by  $H = \Phi^T \Phi$ ,  $\mathbf{f} = -\Phi^T \mathbf{u}_A$ ,  $\mathbf{b} = (\mathbf{1}_m^T, \mathbf{0}_m^T)^T$ , and  $C = [\Phi^T - \Phi^T]^T$ .

# C. Discussion

The LBS weights, given that they are on the interval [0, 1], not only represent a FS, but also are themselves FSs that represent the membership in each of the linguistic basis functions. Hence, if the LBS can be described by linguistic terms, such as 'big' or 'small,' then the weights represent the contribution of that linguistic term in the FS representing the survey responses. Section V demonstrates this for a survey where respondents were asked to indicate the interval between 0 and 10 that represents the word 'little.'

Another attractive aspect of the basis function representation—either LBS or GBS—is that FS operations can be directly applied to the basis functions. Consider two FSs, A and B, represented on the basis shown in Fig. 6,  $A \rightarrow \mathbf{w}_A = (0.1, 0.6, 0.4, 0.05, 0.01)^T$  and  $B \rightarrow \mathbf{w}_B = (0.5, 0.2, 0, 0.01, 0.3)^T$ ; illustrated in Fig. 6(a). The intersection of these two sets is shown in view (b), which is calculated by

$$(A \cap B)(z) = \min\left\{\sum_{i=1}^{5} (\mathbf{w}_A)_i \phi_i(z), \sum_{i=1}^{5} (\mathbf{w}_B)_i \phi_i(z)\right\}.$$

The computational complexity of computing  $A \star B$  using the basis functions is O(m), where, to compute  $A \star B$  using the original IAA representation is  $O(n_A + n_B)$ , which is the cost of sorting z into the IAA set of ordinals X and Y assuming that there are  $n_A$  intervals in A and  $n_B$  in B. Since



Fig. 6: Example of fuzzy intersection applied to FSs A and B using the LBS representation.



Fig. 7: (a) Survey responses for linguistic meaning of 'little' [7]. Each horizontal blue line indicates a survey response. (b) IAA and eIAA results for 'little' survey responses.

 $m \ll (n_A + n_B)$ , the basis function representation is more efficient.

For IT2 FSs, which are the result of IAA with uncertain interval inputs, the IT2 FS is completely defined by its UMF and LMF; hence, eIAA can simply be used on each of these two defining sets. Thus, this makes eIAA also useful for the case of uncertain interval inputs.

#### V. DEMONSTRATION

Figure 7(a) shows 174 survey responses where the respondents were asked to indicate the interval for 'little' on a scale of 0-10 (see [7] for a full description of these data). Clearly, most respondents indicated 'little' as being on the low end of the scale. The blue line in Fig. 7(b) shows the IAA model and the dotted lines indicate the eIAA models of the 174 interval responses. The eIAA membership functions were built with 5, 7, and 15 basis functions. The weighted basis functions are shown in Fig. 8. As expected, as the number of basis functions is increased, the eIAA approximation of the IAA membership function is improved. The basis functions in Fig. 8 also show that the linguistic content of the survey responses favors basis functions that peak around 1-2, which is intuitively pleasing.

Next, we demonstrate how the eIAA method can be used for FS operations on the FSs produced by IAA. Figure 9(a) shows the FS produced by IAA on survey responses for the



Fig. 8: Contribution of eIAA basis functions for 'little' survey responses. The sum of these basis functions produces the eIAA membership functions shown in Fig. 7(b).



Fig. 9: Example of using IAA and eIAA-7 representations to compute union of 'little' and 'large' survey responses.

word 'large.' As expected, the high membership area for the IAA 'large' FS is around 8-9 on the survey scale. View (b) of Fig. 9 shows the result of the union of 'little' and 'large' computed using the IAA FSs (blue line) and the eIAA FSs (red line) using 7 basis functions. Unsurprisingly, the union of the eIAA FSs is an excellent approximation of the union of the IAA FSs.

# VI. CONCLUSION

The IAA method provides aggregation and analysis for crisp and uncertain interval survey responses. A challenge with IAA is that the resulting FSs are defined by a large number of membership-domain pairs—equal to the number of input survey responses. This can not only result in storage issues for huge survey pools, but also present challenges when operating on the resulting FSs produced by IAA. The proposed eIAA algorithm can significantly reduce the necessary storage of the FSs and computational complexity of FS operations. Furthermore, eIAA provides a way to approximate the survey responses with linguistic prototypes, i.e., linguistic basis functions, giving meaning to the IAA and survey analysis results. We demonstrated this on a real survey where respondents were asked to quantify the interval that describes the word 'little.'

As was seen in the demonstration of using IAA and eIAA with the words survey data, there were many survey respondents that selected intervals for these words that would be outside the typical definition of the word, i.e., most reasonable respondents would not choose a value of 10 for the word, 'little.' Some respondents also chose very wide intervals,

with a few respondents choosing the interval 0-10 for their response. The IA and EIA approaches attack this problem with a statistical preprocessing and aggregation scheme. In the future, for cases where outlier removal is appropriate, we will examine using  $\ell_1$  regularization with (9) and (10), which will prefer sparse solutions for the basis function weights—essentially, choosing the linguistic bases that best fit the responses and disregarding all others. We will also investigate how regularization can automate the trade-off between the number of basis functions and the accuracy of eIAA.

Lastly, we will investigate efficient methods for handling inter-expert or inter-source uncertainty in surveys, which in [8] were shown to be best represented by zSlice GT2 FSs. This can be accomplished by using the eIAA idea on each slice of the zSlice GT2 FS.

#### ACKNOWLEDGEMENTS

Superior, a high performance computing cluster at Michigan Technological University, was used in obtaining results presented in this publication.

#### REFERENCES

- S. Coupland, J. Mendel, and D. Wu, "Enhanced interval approach for encoding words into interval type-2 fuzzy sets and convergence of the word FOUs," in *Proc. IEEE Int. Conf. Fuzzy Systems*. IEEE, 2010, pp. 1–8.
- [2] D. Wu, J. Mendel, and S. Coupland, "Enhanced interval approach for encoding words into interval type-2 fuzzy sets and its convergence analysis," *IEEE Trans. Fuzzy Systems*, vol. 20, no. 3, pp. 499–513, 2012.
- [3] S. Miller, C. Wagner, and J. Garibaldi, "Constructing general type-2 fuzzy sets from interval-valued data," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2012.
- [4] C. Wagner, S. Miller, and J. Garibaldi, "Similarity based applications for data-driven concept and word models based on type-2 and type-2 fuzzy sets," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2013, pp. 1–8.
- [5] F. Liu and J. Mendel, "Encoding words into interval type-2 fuzzy sets using an interval approach," *IEEE Trans. Fuzzy Systems*, vol. 16, no. 6, pp. 1503–1521, 2008.
- [6] F. Herrera, S. Alonso, F. Chiclana, and E. Herrera-Viedma, "Computing with word in decision making: foundations, trends, and prospects," *Fuzzy Optimization and Decision Making*, vol. 8, pp. 337–364, 2009.
- [7] F. Liu and J. Mendel, "An interval approach to fuzzistics for interval type-2 fuzzy sets," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2007, pp. 1–6.
- [8] C. Wagner, S. Miller, J. Garibaldi, D. Anderson, and T. Havens, "From interval-valued data to general type-2 fuzzy sets," *IEEE Trans. Fuzzy Systems*, vol. 23, no. 2, pp. 248–269, 2014.
- [9] J. Mendel and R. John, "Type-2 fuzzy sets made simple," *IEEE Trans. Fuzzy Systems*, vol. 10, no. 2, pp. 117–127, 2002.
- [10] C. Wagner and H. Hagras, "Toward general type-2 fuzzy logic systems based on zslices," *IEEE Trans. Fuzzy Systems*, vol. 18, no. 4, pp. 637– 660, 2010.
- [11] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning-i\* 1," *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975.
- [12] S. Coupland and R. John, "Geometric type-1 and type-2 fuzzy logic systems," *IEEE Trans. Fuzzy Systems*, vol. 15, no. 1, pp. 3–15, 2007.
- [13] J. Mendel, F. Liu, and D. Zhai, "α-plane representation for type-2 fuzzy sets: theory and applications," *IEEE Trans. Fuzzy Systems*, vol. 17, no. 5, pp. 1187–1207, 2009.
- [14] F. Liu, "An efficient centroid type-reduction strategy for general type-2 fuzzy logic system," *Information Sciences*, vol. 178, pp. 2224–2236, 2008.
- [15] J. Mendel, R. John, and F. Liu, "Interval type-2 fuzzy logic systems made simple," *IEEE Trans. Fuzzy Systems*, vol. 14, no. 6, pp. 808–821, 2006.
- [16] J. Mendel, Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Upper Saddle River, NJ, USA: Prentice Hall, 2001.