A Unified Framework For Regularizing The Choquet Integral

1st Godswill Ikwan
Department of ECECS
University of New Haven
West Haven, CT, USA
gikwa1@unh.newhaven.edu

2nd Muhammad Aminul Islam
Department of ECECS
University of New Haven
West Haven, CT, USA
mislam@newhaven.edu

***Abstract*—The *Choquet Integral* (ChI) through its *fuzzy measure* (FM) provides a powerful means for non-linear aggregation. Existing work on ChI regularization primarily focuses on either optimizing towards a target FM or minimizing the FM parameters, similar to classical machine learning. The latter approach is essentially goal-oriented optimization towards the minimum operator. These works are limited in the sense of the part of the parameter space they explore. In this article, we enhance the exploration by combining regularization towards two goals, the minimum and the maximum, thus facilitating searching over the entire FM spectrum, which we call the operator regularization. Additionally, we introduce model complexity regularization via additivity conditions that guides the parameters towards linear model (termed as the complexity regularization). We provide a unified framework synthesizing these two regularizers, allowing the ChI to effectively optimize over the FM spectrum. We conducted experiments on real-world datasets demonstrating the effectiveness of the proposed method for limited data.**

***Index Terms*—Information fusion, Fuzzy integral, Choquet integral, regularization.**

I. Introduction

Information fusion is instrumental in combining information from multiple sources in a synergistic manner to provide overarching results that are greater than what any of the individual sources can provide. Fusion provides superior predictive accuracy and robust performance and has been applied to various fields, including decision theory and engineering, with applications ranging from sensor fusion to remote sensing, etc. [1], [2], [3], [4], [5], [6], [7], [8].

Fuzzy Integrals (FIs) through their *fuzzy measures* (FMs) provide a parametric way to define an aggregation operator residing between the minimum and the maximum operator spectrum. The *Choquet integral* (ChI) is the most powerful among them because of the number of parameters used to define it, which are exponential to the number of inputs. While this provides flexibility and the ability to parametrically learn a wide range of operations, the volume of data required to fully learn the parameters uniquely from data increases exponentially with the number of sources to be fused. Therefore, learning the ChI for undetermined systems (when the size of the data is smaller than the number of variables/parameters) poses a huge challenge.

Existing work on handling ChI learning from limited data can be divided into two categories. An approach is to learn only the underlying FM variables that are supported in the dataset. Islam et al. [9] demonstrated that variables that are supported in the data can be isolated, provided methods to identify such variables and lastly formulated an optimization problem while preserving the performance of the entire fuzzy measure (including those that are not supported by the data and not optimized). This method requires imputation of non-data-supported variables. The other approach is to apply regularization, which has been explored in [10], [11], [12], [13].

There are two aspects of regularization when it comes to aggregation operators. One aspect is to control the complexity of the model as in classical machine learning by trading off the bias and variance with a regularization objective towards a linear model. This objective is achieved in the context of the fuzzy integral when the regularization objective is to optimize towards an additive FM. We call this regularization complexity regularization. However, this approach may not recover the underlying FM if it is more complex than the FM obtained without regularization. This requirement calls for another control that gives the algorithm the ability to sweep the entire FM spectrum from the minimum to the maximum to search for a desired operator within the spectrum suitable for a particular task. This is termed as the operator regularization.

In [10], Derek et al. proposed a regularization with the aim to reduce the model complexity using the classical l_1-norm regularization approach that involves simply minimizing the fuzzy measure variables. In the fuzzy measure spectrum, the minimum and the maximum operators reside at the two extrema, where the parameters in the minimum operator are minimum (zero-valued for fuzzy measure defined on a unit interval) for all possible sets of sources except for the universal set which is constant. In contrast, parameters are maximum (one-valued) for the maximum operator. These two operators have the highest degree of nonlinearity among all possibilities and, in turn, the highest model complexity. Thus, regularizing by minimizing the variables as in [10] pulls the variables towards zeros, acting as an optimization towards the minimum operator. The model complexity as a result may not reduce, but rather increase.

The operator regularization is extensively explored in [14]. The objective is to find a suitable/desired operator that can

reside anywhere between the min and the max aggregation operator. To achieve this, they used regularization towards target FMs or structures, for example, min, max, mean, and ordered weighted averaging. However, this method requires manually selecting the goal FMs, necessitating prior knowledge on the characteristics of the underlying FM, which may not be available.

In this article, we propose a unified framework by incorporating both complexity and operator regularizations. The complexity regularization seeks to guide the parameters towards an additive measure–which is by definition linear–to promote linear weights. The second regularization for operator search is realized by adding another term that allows for both minimization and maximization of the parameters. The minimization is active for positive values for the regularization hyperparameter whereas the maximization is active for negative values. Thus, a single hyperparameter enables scanning the spectrum for a sweet spot. Rather than predefining goal structures, this regularization dynamically guides the ChI to behave adaptively between the minimum and maximum operators. In doing so, it enables the model to uncover whether the underlying aggregation in the data is more conjunctive (min-like), disjunctive (max-like), or an intermediate. The objective function includes two regularization terms for complexity and operator, respectively, the coefficients of which are jointly searched due to their interdependent nature. Together, these two methods establish a more balanced framework for fuzzy measure learning that enhances robustness.

The remainder of the article is organized as follows. Section II presents the FM and the ChI. It also provides optimization algorithms. Section III proposes the unified framework for regularization and optimization for l_2-norm and Section IV proposes the l_1-norm variant. Next, in Section V, experiments are performed on real datasets and the benefit of the proposed method is highlighted. Finally, Section VI concludes the article.

II. Background

In this section, we provide basic concepts of the ChI along with its optimization algorithm to lay out the foundation for our proposed work.

A. Fuzzy Measures

Let $X = \{x_1, x_2, \ldots, x_N\}$ be a finite set of information sources. An FM on X is a set function $g : 2^X \to [0, 1]$ that satisfies the following properties:

1) Boundary conditions:

$$g(\emptyset) = 0, \quad g(X) = 1$$

2) Monotonicity: For any $A, B \subseteq X$, if $A \subseteq B$ then

$$g(A) \leq g(B).$$

The FM encodes the importance, or "worth," of each subset of sources and FIs can be applied over FMs to aggregate information from these sources.

(Additivity) An FM is additive if for any $E, F \subset X$ such that $E \cap F = \emptyset$, the following condition holds.

$$g(E \cup F) = g(E) + g(F).$$

B. Choquet Integral

Given a real-valued function $h : X \to \mathbb{R}$ representing the evidence provided by each source, the discrete ChI of h with respect to an FM g is defined as

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

where π is a permutation of indices such that

$$h(x_{\pi(1)}) \geq h(x_{\pi(2)}) \geq \cdots \geq h(x_{\pi(N)}),$$

and $A_i = \{x_{\pi(1)}, \ldots, x_{\pi(i)}\}, \quad g(A_0) = 0$. An equivalent formulation of the ChI is

$$C_g(h) = \sum_{i=1}^{N} \big[h(x_{\pi(i-1)}) - h(x_{\pi(i)}) \big] g(A_i)., \tag{2}$$

which we write in vector form as

$$C_g(h) = \mathbf{z}^T \mathbf{w}, \tag{3}$$

where

$$\mathbf{g}^T = [g(x_1), g(x_2), \ldots, g(X),$$
$$\mathbf{z}^T = [h(x_{\pi(0)}) - h(x_{\pi(1)}),$$

The ChI is a powerful aggregation operator and is capable of representing many aggregation operators. For example, the ChI behaves as a minimum operator when the FMs are 0s (except $g(X) = 1$ due to boundary conditions), the mean operator when $g(A_i) = \frac{|A_i|}{N}$, and behaves like a maximum operator when all FMs are 1s (except $g(\emptyset = 0)$ due to boundary conditions).

When the FM is additive, the ChI is reduced to the linear convex sum of its inputs, i.e., the ChI behaves linearly with respect to the sources regardless of the evidence of support and sorting order of the evidence. In this case, the model will have the least model complexity, i.e., high bias and low variance.

$$C_g(h) = \sum_{i=1}^{N} h(x_{\pi(i)}) g(x_{\pi(i)}).$$

C. The ChI optimization for learning from data

Next we explore learning the FM from training data by minimizing a loss function, the *(sum of squared error)* (SSE). Let $\mathcal{T} = \{(\mathbf{h}_j, y_j)\}_{i=1}^{M}$ denote a training set with M samples or observations, where $\mathbf{h}_j$ represents the j-th instance or feature vector and y_j is associated the target output. The *sum of squared error* (SSE) loss is defined as:

$$L(g) = \sum_{i=1}^{M} \big(C_g(\mathbf{h_i}) - y_i \big)^2 \tag{4}$$

Minimizing this loss allows the FM to be estimated in a data-driven manner. Hence, the optimization problem is

$$\min_g L(g) = \sum_{i=1}^{M} \big(C_g(\mathbf{h}_i) - y_i\big)^2 \tag{5}$$

subject to

$$g(A) \geq g(B), \quad \forall A, B \subset X,\ B \subset A \tag{6}$$

(monotonic constraints)

$$g(\emptyset) = 0,\ g(X) = 1 \quad \text{(boundary conditions)} \tag{7}$$

This optimization problem is convex and can easily be optimized for global optimum using standard python library by appropriately representing the parameters as a vector and coefficients as a matrix. Details on this representation can be found in [10].

III. l_2-NORM REGULARIZATION

A. Operator Regularization

In the context of classical machine learning, the l_2-norm regularization is performed by adding a penalty term based on the L_2-norm of the parameters to the original loss function.

$$\mathcal{L}(g) = \sum_{i=1}^{M} \big(C_g(\mathbf{h}_i) - y_i\big)^2 \ + \ \lambda\, ||\mathbf{g}||_2^2, \lambda \geq 0$$

where λ is the regularization hyperparameter and $\lambda \geq 0$. Suppose the solution to the minimization of the first term, i.e., optimization without regularization ($\lambda = 0$), i.e., is

$$\mathbf{g}_w = [g_w(\{x_1\}), g_w(\{x_1, x_2\}), \ldots, g_w(\{x_1, x_2\}), \ldots, g_w(X)],$$

and the minimization of the second term alone, i.e., the regularization term $||\mathbf{g}||_2^2$, is

$$g_0 = [0, 0, \ldots, 0, \ldots, 1],$$

where the last element corresponds to X and is 1 due to boundary conditions. Regularization allows us to vary the λ to pick an FM, g somewhere between $\mathbf{g}_w$ and $\mathbf{g}_0$. Unlike in classical machine learning, increasing the regularization strength does not necessarily yield a simpler function, but rather it does the opposite as it guides the parameters towards the minimum operator, which is one of the two most nonlinear functions (the other being the maximum operator) available within the FM spectrum.

The above approach allows us to explore the bottom part of the FM spectrum from $\mathbf{g}_w$ down to the minimum. In order to be able to explore the entire spectrum to pick a suitable FM anywhere between the minimum and the maximum, we also need to explore above $\mathbf{g}_w$ by maximizing the parameters towards the maximum operator (corresponding FM has all 1s). This maximization can be done by relaxing the constraints for non-negative λ to be a real-valued number. However, this change will make the optimization non-convex for $\lambda < 0$. To address this problem, we re-formulate the loss function as follows so that it remains convex for real-valued λ.

$$L(g) = \sum_{i=1}^{M} \big(C_g(\mathbf{h}_i) - y_i\big)^2 + |\lambda_o| \sum_{A \subset X} (1 + sign(\lambda_o) g(A))^2, \lambda_o \in \mathbb{R}$$

where $|\lambda_o|$ is the absolute value of λ_o and the $sign$ function is, $sign(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$. Minimizing this loss function will minimize the FM parameters in the regularization term for positive values of λ and maximize the parameters for negative values of λ. We term this approach as the operator regularization.

B. Complexity Regularization

Although the operator regularization enables searching the entire FM spectrum for a suitable FM, it does not explicitly allow for controlling the model complexity. The linear models are the simplest with low variance and high bias, which can be achieved for the ChI when the FM is additive. In an additive FM, for any $A \subseteq X$, the following condition holds.

$$g(A) = \sum_{x_i \in A} g(x_i).$$

In other words, the FM for a set of sources is exactly equal to the sum of its singletons. There is no interaction among sources and the worth of a collection of sources equals the aggregated sum of the individual contributions. The l_2 regularization term that controls for this additivity is

$$\lambda_c \sum_{A \subseteq X} \big(g(A) - \sum_{x_k \in A} g(x_k)\big)^2, \lambda_c \geq 0,$$

where λ_c is the regularization hyperparameter. By penalizing deviations from additivity, the regularizer promotes additive structure to the learned FM.

Finally, we add the two regularization terms to the SSE to obtain a unified loss function, which is

$$L(g) = \sum_{i=1}^{M} \big(C_g(\mathbf{h}_i) - y_i\big)^2 \ + \ |\lambda_o| \sum_{A \subset X} (1 + sign(\lambda_o) g(A))^2 + \lambda_c \sum_{A \subseteq X} \big(g(A) - \sum_{x_k \in A} g(x_k)\big)^2, \lambda_o \in \mathbb{R}, \lambda_c \geq 0. \tag{8}$$

Since the individual terms are convex, their linear sum is also convex, and thus the optimization problem is solvable for global optimum via convex optimization.

IV. l_1-NORM REGULARIZATION

Similar to l_2-regularization, we derive the objective function for l_1-regularization. The operator-regularization term is

$$\lambda_o \sum_{A \subseteq X} |g(A)| = \lambda_o \sum_{A \subseteq X} g(A), \lambda_o \in \mathbb{R}$$

TABLE I
CLASSIFICATION ACCURACY IN PERCENTAGE FOR THE TEST DATASET FOR DIFFERENT REGULARIZATION APPROACHES. "OR" STANDS FOR THE OPERATOR REGULARIZATION AND "CR" FOR THE COMPLEXITY REGULARIZATION.

Dataset	Norm	Random Forest (mean)	No Regularization	OR($\lambda_o \geq 0$)	OR($\lambda_o \geq 0$)+CR	OR ($\lambda_o \in \mathbb{R}$)	OR($\lambda_o \in \mathbb{R}$) + CR	Dominant Regularization(s)
Connect-4	l_1	66.03	61.94	64.17	65.49	64.51	**66.41**	OR (positive λ_o) and CR
	l_2	66.03	61.52	64.74	65.82	64.69	**66.37**	
Covertype	l_1	63.74	65.67	66.41	66.54	66.41	**66.64**	OR (positive λ_o)
	l_2	63.74	65.67	66.15	66.43	66.15	**66.66**	
Dry Bean	l_1	81.10	81.78	81.78	82.08	82.91	**82.91**	OR (negative λ_o)
	l_2	81.10	81.88	81.10	81.64	**83.10**	**83.10**	
Electricity Pricing	l_1	70.19	73.65	73.77	73.78	73.77	**73.87**	Negligible effect
	l_2	70.19	73.65	**73.72**	**73.72**	**73.72**	**73.72**	
MAGIC Gamma Telescope	l_1	74.41	77.81	76.87	78.55	77.04	**78.55**	OR (positive λ_o)
	l_2	74.41	77.81	76.69	78.86	76.69	**78.86**	
Human Activity Recognition	l_1	61.49	62.52	62.52	62.52	**63.24**	**63.24**	OR (negative λ_o)
	l_2	61.49	62.46	62.46	62.46	62.46	**63.37**	
Skin Segmentation	l_1	91.27	93.10	93.10	**93.63**	93.10	**93.63**	CR
	l_2	91.27	91.83	91.77	**93.17**	91.77	**93.17**	

where $|g(A)|$ is replaced with $g(A)$ since $g(A) \geq 0$ by definition. This regularization term is linear and affine for any real-valued λ and thus inclusion of it does not alter the convex characteristics of the objective function.

The loss function after including the l_1-norm complexity and operator regularization terms is

$$L(g) = \sum_{i=1}^{M} \left(C_g(\mathbf{h}_i) - y_i\right)^2 + \lambda_o \sum_{A \subset X} g(A) + \lambda_c \sum_{A \subseteq X} \left|g(A) - \sum_{x_k \in A} g(x_k)\right|, \lambda_o \in \mathbb{R}, \lambda_c \geq 0, \tag{9}$$

which can be minimized using convex optimization algorithms.

V. EXPERIMENTS AND RESULTS

To evaluate the proposed regularization technique and understand the role of its different components, we conducted experiments on real datasets.

A. Dataset

We conducted experiments on seven real-world datasets obtained from the UCI Machine Learning Repository [15]: *Connect-4*, *Forest Cover Type*, *Dry Bean*, *Electricity Pricing*, *MAGIC Gamma Telescope*, *Human Activity Recognition (HAR)*, and *Skin Segmentation*. The Electricity Pricing dataset is for a regression task, which was converted to a classification problem by dividing the target values into groups. These datasets span a variety of domains and classification tasks, making them suitable for evaluating the effectiveness of our proposed methods.

B. Experiment Setup

We trained a Choquet random forest, a variation of the standard random forest, where multiple decision trees are trained on a set of features sampled randomly from a feature pool. Thereafter, the outputs of the decision trees were fused using the ChI, instead of averaging. We fused 10 decision trees – a relatively small size in comparison to the standard random forest, in order to keep the optimization problem tractable for the ChI. The depth of each decision tree was 2. The datasets were randomly partitioned into 12% for validation and 15% for testing. From the remaining samples, we randomly selected a limited number of samples (102) for training to investigate the effect of regularization.

C. Results

Table I reports the classification accuracy for the models trained on real datasets. We analyze the following cases for regularization to understand the role of different components of the unified regularization framework: (i) no regularization with both λ_o and λ_c as zeros; (ii) the operator regularization only with non-negative λ_o, which is the same as classic regularization; (iii) combination of case (ii) and complexity regularization; (iv) full operator regularization with real-valued λ; (iv) a unified regularization, i.e., a combination of full operator and complexity regularizations. We also report results for the mean aggregation a.k.a. random forest. In the table, "OR" stands for the operator regularization and "CR" for the complexity regularization.

The best results are highlighted in bold in Table I and the last column lists the regularization component(s) that have the most contribution to the performance on top of the

baseline accuracy (no-regularization). In general, the unified framework has the best accuracy for all datasets. The results also demonstrate the need for different components of the regularizers. For example, the operator regularization with positive coefficient works well for Covertype and Gamma datasets whereas that with negative coefficient yields better results for Dry Bean and Human Activity Recognition.

The results also demonstrate the need for both complexity and operator regularizations as evidenced by the interplay between these components. For example, the performance for OR($\lambda_o \geq 0$) + CR is higher than OR($\lambda_o \geq 0$) for all datasets except for the Electricity Pricing dataset where it does not have any regularization effect. Although we experimented with both l_2 and l_1-norm regularizations, the results for the datasets under investigation do not show any noticeable performance difference to draw any conclusion. In the future, we plan to study to find problems in the context of information fusion suitable for l_1 regularization.

VI. Conclusion

In this article, we proposed a unified regularization approach for the ChI that provides control over multiple aspects of learning the aggregation operator from data. It has two components: the operator regularization that aids in the search of a suitable operator over the FM space. The second component is the complexity regularization that seeks to reduce the non-linearity of the learned model. The results from the experiments conducted on a small training data (where the number of samples is considerably lower than the number of variables) show that the proposed framework yields a better performance for a diverse set of real-world datasets. The results also show the synergistic gain in performance from the interaction between the operator and complexity regularizations, therefore supporting the need for both during the optimization process.

References

[1] M. Abdar, S. Salari, S. Qahremani, H.-K. Lam, F. Karray, S. Hussain, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "Uncertaintyfusenet: Robust uncertainty-aware hierarchical feature fusion model with ensemble monte carlo dropout for covid-19 detection," *Information Fusion 2023*, 2021.

[2] N. E. I. Hamda, A. Hadjali, and M. Lagha, "Multisensor data fusion in iot environments in dempster–shafer theory setting: An improved evidence distance-based approach," *Sensors*, vol. 23, no. 11, p. 5141, 2023.

[3] M. Hussain, M. O'Nils, J. Lundgren, and S. J. Mousavirad, "A comprehensive review on deep learning-based data fusion," *IEEE Access*, vol. 12, pp. 180 093–180 124, 2024.

[4] F. Samadzadegan, A. Toosi, and F. Dadrass Javan, "A critical review on multi-sensor and multi-platform remote sensing data fusion approaches: current status and prospects," *International Journal of Remote Sensing*, vol. 46, pp. 1327–1402, 2024.

[5] B. J. Murray, M. A. Islam, A. J. Pinar, D. T. Anderson, G. J. Scott, T. C. Havens, and J. M. Keller, "Explainable ai for the choquet integral," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 4, pp. 520–529, 2020.

[6] D. T. Anderson, G. J. Scott, M. A. Islam, B. Murray, and R. Marcum, "Fuzzy choquet integration of deep convolutional neural networks for remote sensing," in *Computational Intelligence for Pattern Recognition*. Springer, 2018, pp. 1–28.

[7] M. A. Islam, D. T. Anderson, A. J. Pinar, T. C. Havens, G. Scott, and J. M. Keller, "Enabling explainable fusion in deep learning with fuzzy integral neural networks," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 7, pp. 1291–1300, 2019.

[8] G. Beliakov and J.-Z. Wu, "Learning fuzzy measures from data: simplifications and optimisation strategies," *Information Sciences*, vol. 494, pp. 100–113, 2019.

[9] M. A. Islam, D. T. Anderson, A. J. Pinar, and T. C. Havens, "Data-driven compression and efficient learning of the choquet integral," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 4, pp. 1908–1922, 2017.

[10] D. T. Anderson, S. R. Price, and T. C. Havens, "Regularization-based learning of the choquet integral," in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2014, pp. 2519–2526.

[11] T. C. Havens, D. T. Anderson, and A. J. Pinar, "Novel regularization for learning the fuzzy choquet integral with limited training data," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 10, pp. 2842–2853, 2020.

[12] T. A. Adeyeba, D. T. Anderson, and T. C. Havens, "Insights and characterization of l1-norm based sparsity learning of a lexicographically encoded capacity vector for the choquet integral," in *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015, pp. 1–7.

[13] S. K. Kakula, A. J. Pinar, T. C. Havens, and D. T. Anderson, "Choquet integral ridge regression," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2020, pp. 1–8.

[14] S. K. Kakula, A. J. Pinar, M. A. Islam, D. T. Anderson, and T. C. Havens, "Novel regularization for learning the fuzzy choquet integral with limited training data," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 10, pp. 2890–2901, 2020.

[15] A. Asuncion, D. Newman *et al.*, "Uci machine learning repository," 2007.