Knowledge Integration in XAI with Gödel Integrals

Adulam Jeyasothy, Agnès Rico, Marie-Jeanne Lesot, Christophe Marsala, Thibault Laugel

Knowledge Integration in XAI with Gödel Integrals

Adulam Jeyasothy[1], Agnès Rico[2], Marie-Jeanne Lesot[1], Christophe Marsala[1], Thibault Laugel[3]
[1] Sorbonne Université CNRS, LIP6, Paris, France, Email:{adulam.jeyasothy,marie-jeanne.lesot,christophe.marsala}@lip6.fr
[2] Univ. Lyon 1, Lyon, France, Email: agnes.rico@univ-lyon1.fr
[3] AXA, Paris, France, Email: thibault.laugel@axa.com

***Abstract*— Counterfactual examples constitute a popular form of explanations that are most often generated through the optimisation of a cost function that combines different components of the explanation quality. This paper focuses on the final aggregation of the objective term, that depends on the considered machine learning task, and the subjective term, that depends on the targeted user and more precisely on their knowledge. It discusses the desired properties of this aggregation operator and proposes to use two forms of the Gödel integral operator, highlighting the expressiveness and appropriateness they offer.**

***Index Terms*—XAI, counterfactual examples, aggregation operators, prior knowledge, Gödel integral, Sugeno integral.**

I. Introduction

In order to explain the prediction performed by a given classifier for a given data instance of interest, numerous so-called local post-hoc methods have been proposed (see e.g. [1], [2]), that, among others, vary in the form of explanation they provide and the information they consider available. Counterfactual examples [3] explain a prediction by identifying modifications to be applied to the considered instance so as to change the associated prediction: they answer the user question "What do I need to modify to get the prediction I want?". A large variety of approaches have been proposed to address this task (see e.g. [4]–[7]), relying on the definition of a cost function to optimise. The latter includes different terms that constitute different components of the quality of the explanation. A crucial question is then that of their aggregation.

This paper focuses on a specific step of aggregation, namely the final one that combines an objective term with a subjective one. The former refers to the combination of numerical criteria that only depend on the machine learning task, i.e. the considered classifier, the data instance about which the explanation is required and possibly additional information such as the data density. The subjective term, on the other hand, is the one that allows for a personalised explanation and depends on the targeted user, in particular on their knowledge. This paper discusses the specific properties this aggregation requires in order to build an explanation that matches the user's expectations. It then proposes to use for this step Gödel integrals [8] that constitute a generalisation of the Sugeno integral, highlighting the expressiveness and appropriateness they offer: they associate a threshold semantics to the weight attached to the criteria, above which (resp. below which) the value to be aggregated is considered equivalent to the maximum (resp. minimum) value of the scale, allowing for a rich behaviour. To the best of our knowledge, this paper proposes the first application of Gödel integrals in the eXplainable Artificial Intelligence (XAI) domain.

The paper is organised as follows: Section II summarises the principles of counterfactual examples explanations, presenting the aggregation issue they raise. Section III discusses the desired characteristics for the aggregation operator and Section IV introduces the proposed use of Gödel integrals. Section V illustrates the expressiveness and richness of this choice on a 2D classical toy data set. Section VI concludes the paper, discussing some directions for future works.

II. Background: Enriched Counterfactual Examples Explanations

This section briefly presents the main principles of explanations based on counterfactual examples and the way how they can be personalised integrating user knowledge. It then discusses the crucial aggregation step they require to perform.

A. Principles of Counterfactual Example Explanations

Counterfactual examples [3] aim at explaning the prediction offered by a machine learning model $f : \mathcal{X} \to \mathcal{Y}$ (where $\mathcal{X}$ denotes the data description space and $\mathcal{Y}$ the output space, e.g. $\mathcal{Y} = \{0, 1\}$ for binary classification) for a data instance of interest $x_0 \in \mathcal{X}$. As mentioned in the introduction, they answer the user question "What minimal modifications, from x_0 to x'_0, allows that $f(x'_0) \neq f(x_0)$?". They are basically defined as the difference between x_0 and x'_0, the closest data point associated to a prediction of the desired class. The proximity constraint aims at minimising the effort the user needs to provide to get the desired output.

Numerous variants of this principle have been proposed (see e.g. [4]–[7]), integrating additional components to measure the quality of the explanation. Beside the proximity requirement, that can be measured by the l_2 norm [9], [10] or the l_1 norm [11], sparsity [10], [12] aims at reducing the number of features to modify, beyond the global quantity of required changes, so as to improve the legibility of the provided explanation. Other criteria, such as plausibility [13], propose to take into account the considered context, in a paradigm that gives up the model and data agnostic assumptions, in order to

make the proposed explanations more realistic, e.g. to avoid out-of-distribution counterfactual examples [14].

We denote $P_{f,x_0}(e)$ the penalty term that defines the quality of a candidate counterfactual example e based on these various components. The output counterfactual example is formally defined as the solution of the optimisation problem:

$$e^* = \arg\min_{e \in \mathcal{X}} P_{f,x_0}(e) \text{ s.t. } f(e) \neq f(x_0) \quad (1)$$

B. Integration of User Knowledge

Beyond the objective criteria that define the penalty function, only depending on the considered machine learning task, a second class of criteria takes into account a more subjective view on counterfactual candidates and makes their quality depend on the user who receives them. Such criteria make it possible to personalise the explanation, increasing its relevance and benefits for the targeted user.

Existing approaches differ both on the type of user knowledge they consider and on the way the latter is integrated in the explanation generation. User knowledge can e.g. take the form of a set of features that make sense to the user [15], [16] and should be favored in the explanation. They can be enriched through acceptable value intervals for each feature [17] or the monotony of acceptable changes [18], e.g. whether the feature value can be increased or decreased. User knowledge can also be expressed as links between features [19], e.g. through causal graphs [18], [20], to ensure that the impact of the modification of a feature on another one is verified.

In this paper, we denote the user knowledge E, independently of its actual form. Candidate counterfactual examples that are incompatible with this knowledge must then be penalised. We denote $I_{E,x_0}(e)$ the incompatibility assessment.

C. Aggregation Issue

The generation of a counterfactual example can then be formulated as an optimisation problem, of which the cost function to be minimised combines the penalty $P_{f,x_0}(e)$ and the incompatibility $I_{E,x_0}(e)$. A crucial question is the aggregation of these two terms.

Actually, the definition of penalty already raises an aggregation issue, as it combines several components. Still, one can consider that these components are of the same nature, as they constitute objective criteria, that only depend on the considered machine learning task. In that sense, classical aggregation operators, such as weighted averages, can be considered as satisfactory, see e.g. [12] or [21]. In some works, the aggregation is performed in an integrated, and somehow implicit, manner, within the optimisation process itself: in [10] for instance, the penalty term depends both on proximity and sparsity, the latter is taken into account in a projection step applied to the points that minimise the proximity criterion.

On the other hand, the combination of penalty and incompatibility requires aggregating components of different nature when considering the first one as objective and the second one as subjective. As a consequence, the aggregation discussion can be seen as richer, calling for more expressive operators. A first step towards this discussion is proposed in [16] that argues that conjunctive operators are too strict, disjunctive operators too lax and that proposes to use compromise operators as weighted averages. This paper proposes to conduct a more systematic discussion on the properties an aggregation operator should possess or not (Section III). It then proposes to apply Gödel integrals (see Section IV) to achieve this goal.

More formally, the question addressed in this paper is to select an aggregation operator agg to define the cost function

$$cost_{f,E,x_0}(e) = agg(P_{f,x_0}(e), I_{E,x_0}(e)) \quad (2)$$

to be minimised under the constraint that $f(e) \neq f(x_0)$.

In the following f, x_0 and E are considered to be fixed so the corresponding subscripts are omitted in order to lighten the notation. Likewise, when there is no ambiguity, the candidate counterfactual example e can be omitted as well, and the considered issue is to characterise and select an operator to perform the aggregation $agg(P, I)$ where P denotes the penalty value and I the incompatibility value for the given f, x_0, E and e. These two quantities are considered to be commensurable, e.g. after a normalisation step to $[0, 1]$.

III. Desired Characteristics for Penalty and Incompatibility Aggregation

There exists a very rich literature on aggregation operators (see e.g. [22]), both for the definition of functions with various properties, semantics and expressiveness, and for lists of conceivable properties. This section discusses some of the properties an aggregator must verify to meet the requirements of the XAI context described in the previous section.

A. Discussion on Monotonicity

We first argue the considered aggregation operator must be non decreasing in both its arguments. Indeed, let us first consider two candidate counterfactual examples e_1 and e_2 such that e_1 is closer to the reference instance x_0 than e_2, but has the same compatibility to the user knowledge: $P(e_1) < P(e_2)$ and $I(e_1) = I(e_2)$. The global quality of the two candidates should then favour e_1, i.e. it is desired that $cost(e_1) \leq cost(e_2)$. In terms of aggregation, it leads to the constraint that the operator should be non decreasing in its first argument.

Likewise, considering two candidates e'_1 and e'_2, such that $P(e'_1) = P(e'_2)$ but $I(e'_1) < I(e'_2)$, again e'_1 should be favoured. The desired aggregation function should thus also be non decreasing in its second argument.

B. Discussion on Commutativity

We argue the considered aggregation operator should not be commutative. As discussed in Section II, the two considered criteria, P and I, have different semantics, respectively being of objective vs subjective nature. They are thus not equivalent and the commutativity property is not expected: it may be the case that $agg(x, y) \neq agg(y, x)$ because $y = P(e)$ does not have the same meaning as $y = I(e)$.

C. Discussion on Variable Behaviour

We argue the considered aggregation operators should have different behaviours according to the values of the variables, e.g. being conjunctive in some regions, disjunctive in others and offer a trade-off property in others. In XAI, one of the difficulties of choosing an aggregation function is that it must be adapted to all types of users, who have different motivations and needs. We propose the users may express their needs in the form of constraints on the criteria, for example as limits on the minimal values of penalty and incompatibility: they may set thresholds $\delta_P, \delta_I \in \mathbb{R}$ and impose that $P(e) < \delta_P$ and $I(e) < \delta_I$ for a candidate to be acceptable. For penalty, the limit can depend on the reference value obtained by the counterfactual example that minimizes the penalty as defined in Equation 1, denoted e_P^*: the constraint can be expressed in terms of loss of quality as compared to e_P^*, making the threshold dependent of the considered instance x_0. For incompatibility, it is difficult to define such a reference value. We thus propose to consider two constraints:

$$P(e) - P(e_P^*) < \delta_P \tag{3}$$

$$I(e) < \delta_I \tag{4}$$

Such constraints divide the criteria space, described by couples $(P(e), I(e))$, into four different zones, depending on whether both, only one or none are satisfied. It seems desirable that the aggregation function offers different behaviours in these zones, whose interpretation is not the same.

D. Discussion on Priority Behaviour

The difference in semantics of the two considered criteria may imply a preference, inducing an order relation between them, that can be interpreted as a desired priority behaviour. This preference is obviously not the same for all users and participates to the explanation personalisation step. If a user e.g. expresses a preference for penalty over incompatibility, then among two counterfactual candidates with the same norm in the (P, I) space, the one with minimal P must be favoured.

A second possibility is to integrate the notion of priority through the choice of thresholds in Equations (3) and (4). In the case where penalty is preferred to incompatibility, a stronger condition on the penalty than on incompatibility is expected: the threshold δ_P associated with penalty is expected to be lower than δ_I associated with the incompatibility.

IV. Chosen Operator: Gödel Integrals

This section proposes to use the Gödel integrals. It first reminds their general definition and then discusses their instanciation to the considered XAI framework, i.e. to P and I, before commenting and illustrating their semantics.

A. Reminder on Gödel Integral Definition

Gödel integrals [8] are a generalisation of the classical Sugeno integral [23] which has two equivalent expressions [23]. Generalising them with Gödel conjunction or implication leads to two different operators.

1) Notations: The set of evaluation criteria is denoted $\mathcal{C} = \{1, \cdots, n\}$. They are considered to be assessed numerically, by values in $L = [0, 1]$.

Gödel integrals take into account the fact that the subsets of criteria have different weights: the latter make it possible to represent the importance of the individual criteria as well as their interactions. This importance is modelled by a capacity (fuzzy measure), $\mu : 2^{\mathcal{C}} \to [0, 1]$, that associates each subset of criteria $A \subset \mathcal{C}$ with its weight $\mu(A)$. By definition, this function is non decreasing with respect to set inclusion and satisfies the boundary conditions $\mu(\emptyset) = 0$ and $\mu(\mathcal{C}) = 1$.

2) Conjunction-based Gödel Integral: The so-called Gödel conjunction is the non commutative conjunction operator defined for all $\alpha, \beta \in [0, 1]$ by:

$$\alpha \otimes_G \beta = \begin{cases} 0 & \text{if } \beta \leq 1 - \alpha \\ \beta & \text{otherwise.} \end{cases}$$

It is non decreasing in its two arguments and satisfies the following limit conditions: $1 \otimes_G \beta = \beta$, $\alpha \otimes_G 1 = 0$ if $\alpha = 0$ and 1 otherwise, and $0 \otimes_G \beta = \alpha \otimes_G 0 = 0$.

The Gödel integral applies this operator to each subset of criteria A. The minimum value of x on A is not modified if it is greater than threshold $1 - \mu(A)$, and it is set to 0 otherwise. This threshold decreases with respect to $\mu(A)$: it is small when $\mu(A)$ is high, i.e. when the set A is important. So a small evaluation on an important set of criteria is kept, whereas a small one on an non important set of criteria is modified to 0. Formally, the Gödel integral is defined as

$$G_\mu^{\otimes}(x) = \max_{A \subseteq \mathcal{C}} \left(\mu(A) \otimes_G \min_{i \in A} x_i \right). \tag{5}$$

3) Implication-based Gödel Integral: The Gödel integral that relies on implication follows the same principle, replacing the Gödel conjunction by the Gödel implication, the max by a min and the use of μ by its conjugate. The Gödel implication is defined for any $\alpha, \beta \in [0, 1]$ by:

$$\alpha \to_G \beta = \begin{cases} 1 & \text{if } \alpha \leq \beta \\ \beta & \text{otherwise.} \end{cases}$$

It satisfies the following limit conditions: $0 \to_G \beta = 1$ and $\alpha \to_G 1 = 1$.

Similarly to the conjunction based case, evaluation A is transformed using this operator: $\mu^c(A) \to_G \max_{i \in A} x_i$, where μ^c is the conjugate capacity of μ defined by $\mu^c(A) = 1 - \mu(\overline{A})$ where $\overline{A}$ is the complementary set of A. Thus, the evaluation is not changed if it is lower than $1 - \mu^c(A) = \mu(\overline{A})$, otherwise it is changed to 1.

Formally the implication-based Gödel integral is

$$G_\mu^{\to}(x) = \min_{A \subseteq \mathcal{C}} \left(\mu^c(A) \to_G \max_{i \in A} x_i \right). \tag{6}$$

Examples of these aggregation operators are provided in the next subsection, when they are applied in the XAI framework.

B. Application to the XAI Counterfactual Example Context

This section discusses the application of the general definitions reminded above to the aggregation of the penalty and the incompatibility values, simply denoted P and I in this section. In addition P and I denote the features these values are respectively associated with. The formal expression of the aggregated values, respectively $G_\mu^{\otimes}(P, I)$ and $G_\mu^{\rightarrow}(P, I)$, are given below, their level lines are illustrated on Figure 1 and their interpretation is detailed in the next section.

The set of criteria becomes $\mathcal{C} = \{P, I\}$, that are normalised and evaluated on the scale $L = [0, 1]$. The considered capacity is then defined on universe $2^{\mathcal{C}}$ whose size equals 4. Two values are set because of the boundary conditions ($\mu(\emptyset) = 0$ and $\mu(\{P, I\}) = 1$), we denote the two other ones as: $\mu(\{P\}) = \alpha_P$ and $\mu(\{I\}) = \beta_I$.

The formal expressions of the P and I aggregation then are

$$G_\mu^{\otimes}(P, I) = \max(\alpha_P \otimes_G P, \beta_I \otimes_G I, 1 \otimes_G \min(P, I))$$

$$= \begin{cases} \min(P, I) & \text{if } P \leq 1 - \alpha_P \text{ and } I \leq 1 - \beta_I \\ \max(P, I) & \text{if } P > 1 - \alpha_P \text{ and } I > 1 - \beta_I \\ P & \text{if } P > 1 - \alpha_P \text{ and } I \leq 1 - \beta_I \\ I & \text{if } P \leq 1 - \alpha_P \text{ and } I > 1 - \beta_I \end{cases}$$

$$G_\mu^{\rightarrow}(P, I) = \min\big((1 - \beta_I) \rightarrow_G P, (1 - \alpha_P) \rightarrow_G I, 1 \rightarrow_G \max(P, I)\big)$$

$$= \begin{cases} \min(P, I) & \text{if } P < 1 - \beta_I \text{ and } I < 1 - \alpha_P \\ \max(P, I) & \text{if } P \geq 1 - \beta_I \text{ and } I \geq 1 - \alpha_P \\ I & \text{if } P \geq 1 - \beta_I \text{ and } I < 1 - \alpha_P \\ P & \text{if } P < 1 - \beta_I \text{ and } I \geq 1 - \alpha_P \end{cases}$$

Properties: It is easy to show that both $G_\mu^{\otimes}(P, I)$ and $G_\mu^{\rightarrow}(P, I)$ satisfy all the desired properties presented in Section III: they are monotonous in each argument, non commutative, offer a variable behaviour and allow to express a criteria hierarchy. The variable behaviour can be seen in the formal expressions as well as on the graphical representation shown in Figure 1: the criteria is divided into four regions, depending on the relative position of the considered criteria P and I and their associated threshold values $1 - \alpha_P$ and $1 - \beta_I$.

C. Interpretation of Gödel Integrals in the XAI Context

1) Threshold Interpretation: The correspondence between the Gödel parameters α_P and β_I and the thresholds associated with the constraints discussed in Section III-C can be established by comparing the regions they respectively define. For instance, for $G_\mu^{\otimes}(P, I)$ and the P criterion, the constraint is satisfied when $P \leq 1 - \alpha_P$, whereas for $G_\mu^{\rightarrow}(P, I)$, the condition is $P \leq 1 - \beta_I$. Comparing them to the constraints expressed in Equation 3 leads to the correspondence between α_P, β_I and $\delta_P + P(e_P^*)$.

Applying the same principle to incompatibility leads to, for $G_\mu^{\otimes}(P, I)$, to $\delta_P + P(e_P^*) = 1 - \alpha_P$ and $\delta_I = 1 - \beta_I$. For $G_\mu^{\rightarrow}(P, I)$, it leads to $\delta_P + P(e_P^*) = 1 - \beta_I$ and $\delta_I = 1 - \alpha_P$.

These differences can be commented as discussed below, when looking at the difference between the induced regions.

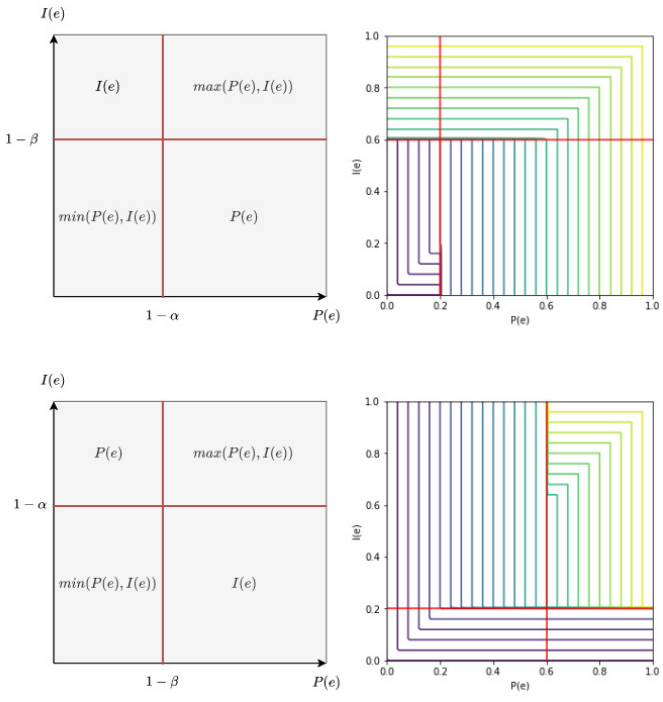


Fig. 1: Level lines of Gödel integrals with $\alpha_P = 0.8$ and $\beta_I = 0.4$: (top) $G_\mu^{\otimes}(P, I)$, (bottom) $G_\mu^{\rightarrow}(P, I)$

2) Region Interpretation: To comment the region interpretation, we focus on the graphical representation given in Figure 1. It can be observed that $G_\mu^{\otimes}(P, I)$ and $G_\mu^{\rightarrow}(P, I)$ share two similar regions, the lower left and the upper right ones. The lower left one corresponds to counterfactual candidates that satisfy both constraints and can thus be considered as satisfying. Their evaluation then depends only on the best criterion, i.e. the minimum between P and I (remind that the overall cost must be minimised). On the contrary, in the top right region, the candidates satisfy none of the constraints, in order to penalise them, their score is defined as the maximum between P and I.

On the two remaining areas, the two integrals do not offer the same aggregation, because they are based on different principles. To ease the discussion, let us consider the case where the penalty constraint is satisfied, but not the incompatibility one, which corresponds to the top left region. $G_\mu^{\otimes}(P, I)$ adopts a punishment behaviour, penalising the candidates in this region up to their unsatisfied criterion, I, independently of their penalty value. On the contrary, $G_\mu^{\rightarrow}(P, I)$ considers they are all as bad regarding the incompatibility and does not distinguish them with respect to that criterion, viewing them as equally lost causes regarding it. $G_\mu^{\rightarrow}(P, I)$ then favours these candidates up to their penalty value. This constitutes a major semantic difference that underlines the richness and relevance of Gödel integrals as aggregation operator in the XAI domain.

We finally comment the influence of the Gödel parameters on the relative sizes of the four regions, showing they play the same role for $G_\mu^{\otimes}(P, I)$ and $G_\mu^{\rightarrow}(P, I)$ despite the difference

of their region interpretation: they are based on the same principle according to which if the capacity weight associated with a criterion is high, then the area that minimizes only this criterion, ignoring the other criterion, is large, indeed giving it more importance. For instance, if α_P is high, the threshold $1-\alpha_P$ is low. Both integrals increase the area that minimizes the penalty: for $G^{\otimes}_{\mu}(P,I)$, it corresponds to the lower right area while it is the upper left area for $G^{\rightarrow}_{\mu}(P,I)$. Actually, in both cases, the area of this region equals $\alpha_P(1-\beta_I)$, showing they globally give the same importance to penalty for given values of the parameters. They differ in the position of this region, but not its importance.

V. Illustrative Examples

This section illustrates counterfactual examples obtained with the proposed Gödel-based aggregation, visualising them for a toy 2D dataset, both for baseline and generic cases.

A. Considered Data

The experiments are conducted with the Half-Moons dataset whose two dimensions are denoted X_0 and X_1. On Figures 2 and 3, the blue and red regions represent the predicted classes, points the training examples; the decision boundary of the trained SVM classifier is shown in white (test accuracy: 0.99). The considered user knowledge is the singleton $E = \{X_1\}$. To allow visual comparisons, all experiments use the same instance x_0, represented by a black cross. Penalty P is defined as normalised Euclidean distance $P = \|x_0 - e\|$, incompatibility I as normalised Euclidean distance on the feature outside E, $I = \|x_0 - e\|_{X_0}$. The optimization problem does not have a unique solution, the whole set of solutions is represented, by green points.

B. Baseline Cases

We first examine four baseline aggregation functions, that also correspond to extreme cases of the Gödel integrals. We give below their expressions and the α_P and β_I parameter values that make them instanciations of $G^{\otimes}_{\mu}(P,I)$ (we omit, for brevity, the parameter values for $G^{\rightarrow}_{\mu}(P,I)$):

$$agg(P,I) = P \qquad \alpha_P = 1 \quad \beta_I = 0 \qquad (7)$$

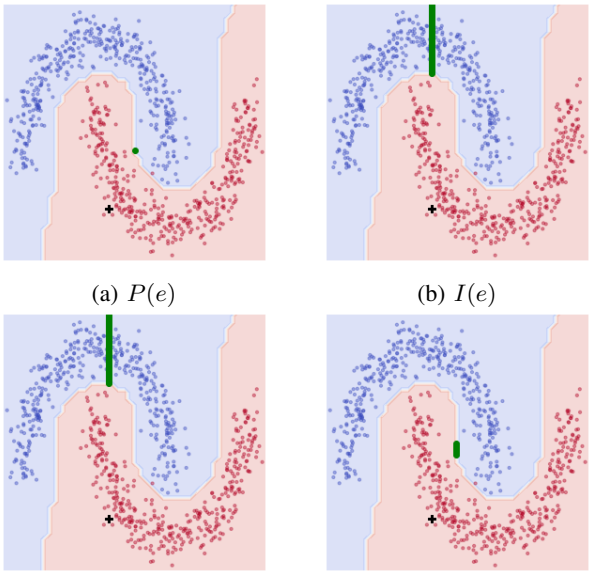
$$agg(P,I) = I \qquad \alpha_P = 0 \quad \beta_I = 1 \qquad (8)$$

$$agg(P,I) = \min(P,I) \qquad \alpha_P = 0 \quad \beta_I = 0 \qquad (9)$$

$$agg(P,I) = \max(P,I) \qquad \alpha_P = 1 \quad \beta_I = 1 \qquad (10)$$

Eq. (7) corresponds to the classical case where only the penalty is considered, Eq. (8) is rare as it considers only the user, fully ignoring penalty; Eq. (9) and (10) respectively represent the conjunction and disjunction of the two criteria.

Figure 2 shows the counterfactual examples obtained in each case, illustrating their expected diversity. Figure 2a constitutes the reference explanation. Figure 2b shows the explanations that minimize incompatibility. For the considered x_0, it is possible to find counterfactual examples totally compatible with the user knowledge: the generated explanations are thus points located at the vertical of x_0 that belong to the blue class,



(a) $P(e)$ (b) $I(e)$

(c) $\min(P(e), I(e))$ (d) $\max(P(e), I(e))$

Fig. 2: Counterfactuals explanations generated for the baseline cases defined in Eq. (7), (8), (9) and (10)

with incompatibility equals 0. Figure 2c is similar to Figure 2b: in the considered case, I can equal 0, whereas P cannot; the minimum thus leads to the same results as incompatibility. Finally, Figure 2d is associated with the maximum function; the generated explanations are located at positions where the incompatibility outweighs the penalty.

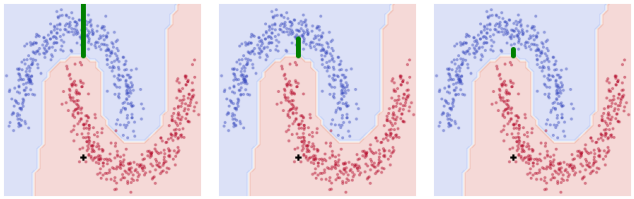
C. General Gödel Integrals

Figure 3 shows representative explanations generated when using $G^{\otimes}_{\mu}(P,I)$ for other, less extreme, values of the parameters α_P and β_I, chosen to illustrate the variety of results they lead to. Six cases can be distinguished, illustrating the interest and expressiveness of this aggregation operator.

On Fig. 3a, that is identical to Fig. 2b and 2c, the set of generated counterfactual examples is the set of totally compatible points of the other class, i.e. those for which $I(e) = 0$. For the considered instance x_0, it can be obtained whenever $\alpha_P < 0.5$. When α_P increases beyond that threshold, the number of generated explanations decreases, as illustrated for Fig. 3b and 3c ($\alpha_P = 0.65$ and 0.74 respectively). This shows the impact of the α threshold in Gödel integrals: even if the explanations are completely compatible, if they do not satisfy the constraint imposed by the the penalty, they are discarded.

On Fig. 3f, that is identical to Fig. 2a, a single counterfactual example is generated, that corresponds to the closest point of the other class, i.e. the one with the lowest penalty. This case is obtained whenever the constraint imposed by penalty is too strong, i.e. α_P is too high as compared to incompatibility. In this case, it is impossible to find a compatible explanation, the optimisation process thus focuses on minimising the penalty.

Figures 3d and 3e represent a compromise between the extreme cases of Fig. 3c and 3f, i.e. trade-offs between P



(a) $\alpha = 0, \beta = 0.2$ (b) $\alpha = 0.65, \beta = 0.2$ (c) $\alpha = 0.74, \beta = 0.2$

(d) $\alpha = 0.8, \beta = 0.98$ (e) $\alpha = 0.83, \beta = 0.9$ (f) $\alpha = 0.9, \beta = 0.2$

Fig. 3: Counterfactual examples obtained by minimizing $G_\mu^\otimes(P, I)$ for different values of α_P and β_I

and I. We illustrate these cases with a high threshold for penalty, the generated counterfactual examples are the most compatible instances that satisfy the penalty constraint. In the figures represented here, at least one of the constraints is satisfied. Fig. 2d represents the maximum function if none of the constraints is verified ($\alpha_P > 0.9$ and $\beta_I > 0.95$). These values are associated with very strong constraints. Fig. 3e is a variant thereof, with more tolerance on the penalty value.

The $G_\mu^\rightarrow$ results, omitted for brevity, show similar behaviours, for other values of the parameters, due to their semantic difference (see Section IV-C). However, when the penalty constraint is not satisfied, $G_\mu^\rightarrow$ focuses on incompatibility (see Section IV-C). As for the considered illustrative example there exist totally compatible explanations, for all $\alpha_P < 1$, they are the generated ones. This case corresponds to the one illustrated in Fig. 3a. As a consequence, for this specific x_0, $G_\mu^\rightarrow$ leads to less diverse cases than $G_\mu^\otimes$. On the other hand, two cases appear when $\alpha_P = 1$, that respectively correspond to Fig. 3b associated with the maximum, if $\beta_I > 0.93$, and Fig. 3d associated with penalty otherwise.

VI. Conclusion

This paper proposed to apply the aggregation operator family of Gödel integrals to combine a subjective assessment of the quality of candidate explanations, related to user knowledge, with an objective assessment, that only depends on the considered machine learning task. It discussed the required properties of an aggregation function in this specific setting and interpreted the advantages of the Gödel integrals, leading to their innovative application in the XAI domain. Thus, we present a new tool that allows us to propose a more adapted explanation to the user.

The discussion focused on the case of counterfactual examples for the definition of penalty and incompatibility, but its principle can be applied whenever these two quantities can be defined. This for instance applies to explanations as local feature importance weights, which constitutes a direction for future works. Another one will aim at conducting experiments with real users, in a human-in-the-loop setting which is crucial for all XAI studies. They will in particular study their assessment of the generated explanations, the preferred behaviours or the elicitation of the appropriate threshold parameters.

References

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.

[2] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, 2021.

[3] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *Harvard journal of law & technology*, vol. 31, pp. 841–887, 2018.

[4] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[5] R. Mazzine and D. Martens, "A framework and benchmarking study for counterfactual generating methods on tabular data," *Applied Sciences*, vol. 11, no. 16, p. 7274, 2021.

[6] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: contrastive explanations and consequential recommendations," *ACM Computing Surveys (CSUR)*, 2022.

[7] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

[8] D. Dubois, H. Prade, A. Rico, and B. Teheux, "Generalized qualitative Sugeno integrals," *Information Scie,ces*, vol. 415, pp. 429–445, 2017.

[9] M. T. Lash, Q. Lin, N. Street, J. G. Robinson, and J. Ohlmann, "Generalized Inverse Classification," in *SIAM Int. Conf. on Data Mining*, 2017, p. 162–170.

[10] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Comparison-based Inverse Classification for Interpretability in Machine Learning," in *IPMU*. Springer, 2018, pp. 100–111.

[11] A. Artelt and B. Hammer, "Convex Density Constraints for Computing Plausible Counterfactual Explanations," in *Artificial Neural Networks and Machine Learning*, 2020, pp. 353–365.

[12] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," in *Proc. of the Int. Conf. on Parallel Problem Solving from Nature*. Springer, 2020.

[13] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "FACE: Feasible and Actionable Counterfactual Explanations," in *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society, AIES*, 2020.

[14] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The dangers of post-hoc interpretability: Unjustified counterfactual explanations," in *Proc. of the 28th IJCAI Conf.*, 2019, pp. 2801–2807.

[15] B. Ustun, A. Spangher, and Y. Liu, "Actionable Recourse in Linear Classification," in *ACM FAccT*. ACM, 2019, p. 10–19.

[16] A. Jeyasothy, T. Laugel, M.-J. Lesot, C. Marsala, and M. Detyniecki, "Integrating prior knowledge in post-hoc explanations," in *IPMU*, 2022.

[17] G. Navas-Palencia, "Optimal counterfactual explanations for scorecard modelling," *arXiv preprint arXiv:2104.08619*, 2021.

[18] D. Mahajan, C. Tan, and A. Sharma, "Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers," *CausalML NeurIPS workshop*, 2019.

[19] M. Drescher, A. H. Perera, C. J. Johnson, L. J. Buse, C. A. Drew, and M. A. Burgman, "Toward rigorous use of expert knowledge in ecological research," *Ecosphere*, vol. 4, no. 7, 2013.

[20] C. Frye, C. Rowat, and I. Feige, "Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability," in *Proc. of NeurIPS*, vol. 33, 2020.

[21] P. Rasouli and I. Chieh Yu, "Care: Coherent Actionable Recourse based on Sound Counterfactual Explanations," *International Journal of Data Science and Analytics*, pp. 1–26, 2022.

[22] M. Grabisch, J. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions*, ser. Encyclopedia of Mathematics and its Applications. Cambridge Univ. Press, 2009, no. 127.

[23] M. Sugeno, "Theory of fuzzy integrals and its applications," Ph.D. dissertation, Tokyo Institute of Technology, 1974.