An ASP Approach for Reasoning on Neural Networks under a Finitely Many-Valued Semantics for Weighted Conditional Knowledge Bases∗

LAURA GIORDANO and DANIELE THESEIDER DUPRÉ
DISIT, Università del Piemonte Orientale, Italy
(e-mails: laura.giordano@uniupo.it, dtd@uniupo.it)

submitted 16 May 2022; accepted 8 June 2022

Abstract

Weighted knowledge bases for description logics with typicality have been recently considered under a “concept-wise” multipreference semantics (in both the two-valued and fuzzy case), as the basis of a logical semantics of multilayer perceptrons (MLPs). In this paper we consider weighted conditional $\mathcal{ALC}$ knowledge bases with typicality in the finitely many-valued case, through three different semantic constructions. For the boolean fragment $\mathcal{LC}$ of $\mathcal{ALC}$ we exploit answer set programming and asprin for reasoning with the concept-wise multipreference entailment under a $\varphi$-coherent semantics, suitable to characterize the stationary states of MLPs. As a proof of concept, we experiment the proposed approach for checking properties of trained MLPs.

KEYWORDS: description logics, neural networks, multi-valued logics, answer set programming

1 Introduction

Preferential approaches to common sense reasoning by Kraus et al. (1990), Pearl (1990), Lehmann and Magidor (1992), Benferhat et al. (1993), have been extended to description logics (DLs) to deal with inheritance with exceptions in ontologies, by allowing for non-strict inclusions, called typicality or defeasible inclusions, with different preferential semantics, for example, by Giordano et al. (2007) and Britz et al. (2008), and closure constructions, for example, by Casini and Straccia (2010, 2013a) and Giordano et al. (2015).

In recent work, a concept-wise multipreference semantics has been proposed by Giordano and Theseider Dupré (2020) as a semantics for ranked DL knowledge bases (KBs), that is KBs in which defeasible or typicality inclusions of the form $T(C) \subseteq D$ (meaning “the typical $C$’s are $D$’s” or “normally $C$’s are $D$’s”) are given a rank, a natural number, representing their strength, where $T$ is a typicality operator (Giordano et al. 2020).
2007), that singles out the typical instances of concept $C$. The concept-wise multipreference semantics takes into account preferences with respect to different concepts, and integrates them into a single global preference relation, which is used in the evaluation of defeasible inclusions. Answer set programming (ASP) and, in particular, the asprin framework for answer set preferences, by Brewka et al. (2015), is exploited to achieve defeasible reasoning under the multipreference approach for $\mathcal{EL}^+$ (Baader et al. 2005).

The multipreferential semantics has been extended by Giordano and Theseider Dupré (2021b) to weighted KBs, in which typicality inclusions have a real (positive or negative) weight, representing plausibility or implausibility. The multipreference semantics has been exploited to provide a preferential interpretation to Multilayer Perceptrons (MLPs, Haykin 1999), an approach previously considered by Giordano et al. (2020, 2022) for self-organizing maps (SOMs, Kohonen et al. 2001). In both cases, considering the domain of all input stimuli presented to the network during training (or in the generalization phase), one can build a semantic interpretation describing the input–output behavior of the network as a multipreference interpretation, where preferences are associated to concepts. For MLPs, based on the fuzzy multipreference semantics for weighted KBs, a deep neural network can actually be regarded as a weighted conditional KB (Giordano and Theseider Dupré 2021b). This rises the issue of defining proof methods for reasoning with weighted conditional KBs.

Undecidability results for fuzzy DLs with general inclusion axioms by Cerami and Straccia (2011) and Borgwardt and Peñaloza (2012) motivate the investigation of many-valued approximations of fuzzy multipreference entailment. We then restrict to the case of finitely many-valued DLs, studied by García-Cerdaña et al. (2010), Bobillo and Straccia (2011), Bobillo et al. (2012), Borgwardt and Peñaloza (2013), and reconsider the fuzzy multipreference semantics based on the notions of coherent, faithful, and $\varphi$-coherent model of a defeasible KB (Giordano and Theseider Dupré 2021b; Giordano 2021a,b). The last notion is suitable to characterize the stationary states of MLPs and is related to the previously introduced notions of multipreferential interpretation.

We consider the finitely many-valued Gödel DL $G_n\mathcal{ALC}$, and the finitely many-valued Lukasiewicz DL, $L_n\mathcal{ALC}$, and develop their extension with typicality and a semantic closure construction based on coherent, faithful, and $\varphi_n$-coherent interpretations to deal with weighted KBs. For the boolean fragment $\mathcal{LC}$ of $\mathcal{ALC}$, which neither contains roles, nor universal and existential restrictions, we develop an ASP approach for deciding $\varphi_n$-coherent entailment from weighted KBs in the finitely many-valued case. In particular, we develop an ASP encoding of a weighted KB and exploit asprin (see Brewka et al. 2015) for defeasible reasoning, to prove typicality properties of a weighted conditional KB. From the soundness and completeness of the encoding, we also get a $\Pi^p_2$ complexity upper bound for $\varphi_n$-coherent entailment.

As a proof of concept, we experiment our approach over weighted KBs corresponding to some of the trained multilayer feedforward networks considered by Thrun et al. (1991). We exploit ASP to verify some properties of the network expressed as typicality properties in the finite many-valued case. This is a step towards explainability of the black-box, in view of a trustworthy, reliable, and explainable AI (Adadi and Berrada 2018; Guidotti et al. 2019; Arrieta et al. 2020), and of an integrated use of symbolic reasoning and neural models.
2 Finitely many-valued $\mathcal{ALC}$

Fuzzy DLs have been widely studied in the literature for representing vagueness in DLs, for example, by Straccia (2005), Stoilos et al. (2005), Lukasiewicz and Straccia (2009), Borgwardt and Peñaloza (2012), Bobillo and Straccia (2018), based on the idea that concepts and roles can be interpreted as fuzzy sets and fuzzy relations.

In fuzzy logic formulas have a truth degree from a truth space $S$, usually $[0,1]$, as in mathematical fuzzy logic (Cintula et al. 2011) or $\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$, for an integer $n \geq 1$. $S$ may as well be a complete lattice or a bilattice.

The finitely many-valued case is also well studied for DLs (García-Cerdaña et al. 2010; Bobillo and Straccia 2011; Bobillo et al. 2012; Borgwardt and Peñaloza 2013), and, in the following, we will consider a finitely many-valued extension of $\mathcal{ALC}$ with typicality.

The basic $\mathcal{ALC}$ syntax features a set $N_C$ of concept names, a set $N_R$ of role names and a set $N_I$ of individual names. The set of $\mathcal{ALC}$ concepts can be defined inductively:

- $A \in N_C$, $\top$ and $\bot$ are concepts;
- if $C$ and $D$ are concepts, and $r \in N_R$, then $C \sqcap D$, $C \sqcup D$, $\neg C$, $\forall r.C$, $\exists r.C$ are concepts.

We assume the truth space to be $C_n = \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$, for an integer $n \geq 1$. A finitely many-valued interpretation for $\mathcal{ALC}$ is a pair $I = \langle \Delta, I^r \rangle$ where $\Delta$ is a non-empty domain and $I^r$ is an interpretation function that assigns to each $a \in N_I$ a value $a^I \in \Delta$, to each $A \in N_C$ a function $A^I : \Delta \rightarrow C_n$, to each $r \in N_R$ a function $r^I : \Delta \times \Delta \rightarrow C_n$. A domain element $x \in \Delta$ belongs to the extension of concept name $A$ to some degree $A^I(x)$ in $C_n$. The interpretation function $I^r$ is extended to complex concepts as follows:

\[
\top^I(x) = 1, \quad \bot^I(x) = 0, \quad (\neg C)^I(x) = \ominus C^I(x),
\]

\[
(\exists r.C)^I(x) = \sup_{y \in \Delta} r^I(x, y) \otimes C^I(y), \quad (C \sqcap D)^I(x) = C^I(x) \oplus D^I(x)
\]

\[
(\forall r.C)^I(x) = \inf_{y \in \Delta} r^I(x, y) \odot C^I(y), \quad (C \sqcup D)^I(x) = C^I(x) \otimes D^I(x),
\]

where $x \in \Delta$ and $\otimes$, $\oplus$, $\odot$, and $\ominus$ are arbitrary but fixed t-norm, s-norm, implication function, and negation function (Lukasiewicz and Straccia 2009). In particular, in this paper we consider two finitely many-valued DLs based on $\mathcal{ALC}$, the finitely many-valued Lukasiewicz DL $\mathcal{L_nALC}$ (L_nALC in the following) as well as the finitely many-valued Gödel DL $\mathcal{ALC}$, extended with a standard involutive negation $\ominus a = 1 - a$ ($G_n\mathcal{ALC}$ in the following). Such logics are defined along the lines of the finitely many-valued DL $\mathcal{SROTQ}$ by Bobillo and Straccia (2011), the logic $GZ\mathcal{SROTQ}$ by Bobillo et al. (2012), and the logic $\mathcal{ALC^*}(S)$ by García-Cerdaña et al. (2010), where $*$ is a divisible finite t-norm over a chain of $n$ elements.

Specifically, in an $L_n\mathcal{ALC}$ interpretation, we let: $a \otimes b = \max\{a + b - 1, 0\}$, $a \oplus b = \min\{a + b, 1\}$, $a \odot b = \min\{1 - a + b, 1\}$, and $\ominus a = 1 - a$. In a $G_n\mathcal{ALC}$ interpretation, we let: $a \otimes b = \min\{a, b\}$, $a \oplus b = \max\{a, b\}$, $a \odot b = 1$ if $a \leq b$ and $b$ otherwise; and $\ominus a = 1 - a$.

The interpretation function $I^r$ is also extended to $\mathcal{ALC}$ concept inclusions of the form $C \sqsubseteq D$ (where $C$ and $D$ are $\mathcal{ALC}$ concepts), and to $\mathcal{ALC}$ assertions of the form $C(a)$ and $r(a, b)$ (where $C$ is an $\mathcal{ALC}$ concept, $r \in N_R$, $a, b \in N_I$), as follows:

\[
(C \sqsubseteq D)^I = \inf_{x \in \Delta} C^I(x) \odot D^I(x), \quad (C(a))^I = C^I(a^I), \quad (R(a, b))^I = R^I(a^I, b^I).
\]
A $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) knowledge base $K$ is a pair $(\mathcal{T}, \mathcal{A})$ where $\mathcal{T}$ is a TBox and $\mathcal{A}$ an ABox. A TBox $\mathcal{T}$ is a set of $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) concept inclusions of the form $C \sqsubseteq D \theta \alpha$, where $C \sqsubseteq D$ is an $\text{ALC}$ concept inclusion, $\theta \in \{\geq, \leq, >, <\}$ and $\alpha \in [0, 1]$. An ABox $\mathcal{A}$ is a set of $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) assertions of the form $C(a) \theta \alpha$ or $r(a, b) \theta \alpha$, where $C$ is an $\text{ALC}$ concept, $r \in N_R$, $a, b \in N_I$, $\theta \in \{\geq, \leq, >, <\}$ and $\alpha \in [0, 1]$. The notions of satisfiability of a KB in a many-valued interpretation and of $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) entailment are defined as follows:

**Definition 1** (Satisfiability and entailment for $G_n^{n\text{ALC}}$ and $L_n^{n\text{ALC}}$)

A $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) interpretation $I$ satisfies a $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) axiom $E$, as follows:

- $I$ satisfies axiom $C \sqsubseteq D \theta \alpha$ if $(C \sqsubseteq D)^I \theta \alpha$;
- $I$ satisfies assertion $C(a) \theta \alpha$ if $C^I(a) \theta \alpha$;
- $I$ satisfies assertion $r(a, b) \theta \alpha$ if $r^I(a, b) \theta \alpha$.

Given a $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) knowledge base $K = (\mathcal{T}, \mathcal{A})$, a $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) interpretation $I$ satisfies $\mathcal{T}$ (resp. $\mathcal{A}$) if $I$ satisfies all inclusions in $\mathcal{T}$ (resp. all assertions in $\mathcal{A}$). A $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) interpretation $I$ is a $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) model of $K$ if $I$ satisfies $\mathcal{T}$ and $\mathcal{A}$. A $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) axiom $E$ is entailed by knowledge base $K$, written $K \models_{G_n^{n\text{ALC}}} E$ (resp. $K \models_{L_n^{n\text{ALC}}} E$), if for all $G_n^{n\text{ALC}}$ ($L_n^{n\text{ALC}}$) models $I = \langle \Delta, I \rangle$ of $K$, $I$ satisfies $E$.

### 3 Finitely many-valued $\text{ALC}$ with typicality

In this section, we consider an extension of finitely many-valued $\text{ALC}$ with typicality concepts, based on a preferential semantics, first introduced by Giordano and Theseider Dupré (2021b) for weighted $\mathcal{EL}^\perp$ knowledge bases (we adopt an equivalent slight reformulation of the semantics by Giordano 2021b). The idea is similar to the extension of $\text{ALC}$ with typicality in the two-valued case by Giordano et al. (2007), but the degree of membership of domain individuals in a concept $C$ is used to identify the typical elements of $C$. The extension allows for the definition of *typicality inclusions* of the form $T(C) \sqsubseteq D \theta \alpha$. For instance, $T(C) \sqsubseteq D \geq \alpha$ means that typical $C$-elements are $D$-elements with degree greater than $\alpha$. In the two-valued case, a typicality inclusion $T(C) \sqsubseteq D$ corresponds to a KLM conditional implication $C \rightarrow D$ by Kraus et al. (1990), Lehmann and Magidor (1992). As in the two-valued case, nesting of the typicality operator is not allowed.

Note that, in a many-valued $\text{ALC}$ interpretation $I = \langle \Delta, I \rangle$, the degree of membership $C^I(x)$ of domain elements $x$ in a concept $C$ induces a preference relation $<_C$ on $\Delta$:

$$x <_C y \iff C^I(x) > C^I(y).$$  \hspace{1cm} (1)

For a finitely many-valued $\text{ALC}$ interpretation $I = \langle \Delta, I \rangle$, each preference relation $<_C$ has the properties of preference relations in KLM-style ranked interpretations by Lehmann and Magidor (1992), that is, $<_C$ is a modular and well-founded strict partial order. Let us recall that $<_C$ is well-founded if there is no infinite descending chain of domain elements; $<_C$ is modular if, for all $x, y, z \in \Delta$, $x <_C y$ implies $(x <_C z$ or $z <_C y)$. Well-foundedness holds for the induced preference $<_C$ defined by condition (1) as we have assumed the truth space to be $C_n$. We will denote $L_n^{n\text{ALC}T}$ and $G_n^{n\text{ALC}T}$ the extensions of $L_n^{n\text{ALC}}$ and $G_n^{n\text{ALC}}$ with typicality.

Each relation $<_C$ has the properties of a preference relation in KLM rational interpretations, also called ranked interpretations. As many-valued interpretations induce multiple
preferences, they can be regarded as multipreferential interpretations, which have also been studied in the two-valued case, for example, by Giordano and Theseider Dupré (2020), Delgrande and Rantsoudis (2020), Giordano and Gliozzi (2021), Casini et al. (2021).

The preference relation $<_C$ captures the relative typicality of domain elements wrt concept $C$ and may then be used to identify the typical $C$-elements. We regard typical $C$-elements as the domain elements $x$ that are preferred with respect to $<_C$ among the ones such that $C^I(x) \neq 0$. Let $C^I_{\Delta}$ be the crisp set containing all domain elements $x$ such that $C^I(x) > 0$, that is, $C^I_{\Delta} = \{ x \in \Delta \mid C^I(x) > 0 \}$. One can provide a (two-valued) interpretation of typicality concepts $T(C)$ with respect to an interpretation $I$ as:

$$(T(C))^I(x) = \begin{cases} 1 & \text{if } x \in \min^<_C(C^I_{\Delta}) \\ 0 & \text{otherwise,} \end{cases}$$  

where $\min^<_C(S) = \{ u : u \in S \text{ and } \forall z \in S \text{ s.t. } z < u \}$. When $(T(C))^I(x) = 1$, $x$ is said to be a typical $C$-element in $I$. Note that, if $C^I(x) > 0$ for some $x \in \Delta$, $\min^<_C(C^I_{\Delta}) \neq \emptyset$. This generalizes the property that, in the crisp case, $C^I \neq \emptyset$ implies $(T(C))^I \neq \emptyset$.

**Definition 2 (G$_n$ALC (L$_n$ALC) interpretation)**

A G$_n$ALC (L$_n$ALC) interpretation $I = \langle \Delta, I \rangle$ is a finitely many-valued G$_n$ALC (L$_n$ALC) interpretation over $C_n$, extended by interpreting typicality concepts as in (2).

A many-valued interpretation $I = \langle \Delta, I \rangle$ implicitly defines a multipreferential interpretation, where any concept $C$ is associated to a relation $<_C$. The notions of satisfiability in G$_n$ALC (L$_n$ALC), model of a G$_n$ALC (L$_n$ALC) knowledge base, and G$_n$ALC (L$_n$ALC) entailment are defined similarly as for L$_n$ALC and G$_n$ALC in Section 2.

### 3.1 Weighted KBs and closure construction for finitely many values

In this section we introduce the notion of weighted G$_n$ALC (L$_n$ALC) knowledge base allowing for weighted defeasible inclusions, namely, typicality inclusions with a real-valued weight, as introduced for $\mathcal{EL}$ by Giordano and Theseider Dupré (2021b).

A weighted G$_n$ALC knowledge base $K$, over a set $C = \{ C_1, \ldots, C_k \}$ of distinguished G$_n$ALC concepts, is a tuple $\langle T, T_{C_1}, \ldots, T_{C_k}, A \rangle$, where $T$ is a set of G$_n$ALC inclusion axioms, $A$ is a set of G$_n$ALC assertions, and $T_{C_i} = \{ (d^i_h, w^i_h) \}$ is a set of all weighted typicality inclusions $d^i_h = T(C_i) \subseteq D_{i,h}$ for $C_i$, indexed by $h$, where each inclusion $d^i_h$ has weight $w^i_h$, a real number, and $C_i$ and $D_{i,h}$ are G$_n$ALC concepts. The typicality operator is assumed to occur only on the left hand side of a weighted typicality inclusion, and we call distinguished concepts those concepts $C_i$ occurring on the l.h.s. of some typicality inclusion $T(C_i) \subseteq D$. The definition of a weighted L$_n$ALC knowledge base is similar. Let us consider the following example.

**Example 1**

Consider the weighted G$_n$ALC knowledge base $K = \langle T, T_{Bird}, T_{Penguin}, A \rangle$, over the set of distinguished concepts $C = \{ Bird, Penguin \}$, with $T$ containing, for instance, the inclusion $\text{Black} \cap \text{Red} \subseteq \bot \geq 1$.

The weighted TBox $T_{Bird}$ contains the weighted defeasible inclusions:

$$
(d_1) \ T(Bird) \subseteq \text{Fly}, \ +20 \quad (d_2) \ T(Bird) \subseteq \text{Has\_Wings}, \ +50 \\
(d_3) \ T(Bird) \subseteq \text{Has\_Feather}, \ +50.
$$
and $T_{\text{penguin}}$ contains the weighted defeasible inclusions:

$$(d_4) \ T(Penguin) \subseteq \text{Bird}, \ +100 \quad (d_5) \ T(Penguin) \subseteq \text{Fly}, \ -70$$

$$(d_6) \ T(Penguin) \subseteq \text{Black}, \ +50.$$ 

That is, a bird normally has wings, has feathers and flies, but having wings and feather (both with weight 50) for a bird is more plausible than flying (weight 20), although flying is regarded as being plausible; and so on. Given a box $\mathcal{A}$ in which Reddy is red, has wings, has feather and flies (all with degree 1) and Opus has wings and feather (with degree 1), is black with degree 0.8 and does not fly, considering the weights of defeasible inclusions, we expect Reddy to be more typical than Opus as a bird, but less typical as a penguin.

In previous work, Giordano and Theseider Dupré (2021b) introduced a semantics of a weighted $\mathcal{EL}$ knowledge bases through a semantic closure construction, similar in spirit to the rational closure by Lehmann and Magidor (1992), to the lexicographic closure by Lehmann (1995), and related to c-representations by Kern-Isberner (2001), but based on multiple preferences. Here, the same construction is extended to weighted $G_n\mathcal{ALCT}$ ($L_n\mathcal{ALCT}$) knowledge bases, by considering the notions of coherent, faithful, and $\varphi$-coherent interpretations. The construction allows a subset of the $G_n\mathcal{ALCT}$ ($L_n\mathcal{ALCT}$) interpretations to be selected, those in which the preference relations $\prec_c$, coherently or faithfully represent the defeasible part of the knowledge base $K$.

Let $T_{C_i} = \{(d^i_h, w^i_h)\}$ be the set of weighted typicality inclusions $d^i_h = T(C_i) \subseteq D_{i,h}$ associated to the distinguished concept $C_i$, and let $I = \langle \Delta, I \rangle$ be a $G_n\mathcal{ALCT}$ ($L_n\mathcal{ALCT}$) interpretation. In the two-valued case, we would associate to each domain element $x \in \Delta$ and each distinguished concept $C_i$, a weight $W_i(x)$ of $x$ wrt $C_i$ in $I$, by summing the weights of the defeasible inclusions satisfied by $x$. However, as $I$ is a many-valued interpretation, we need to consider, for all inclusions $T(C_i) \subseteq D_{i,h} \in T_{C_i}$, the degree of membership of $x$ in $D_{i,h}$. For each domain element $x \in \Delta$ and distinguished concept $C_i$, the weight $W_i(x)$ of $x$ wrt $C_i$ in a $G_n\mathcal{ALCT}$ ($L_n\mathcal{ALCT}$) interpretation $I = \langle \Delta, I \rangle$ is:

$$W_i(x) = \begin{cases} 
\sum_h w^i_h D^i_{i,h}(x) & \text{if } C^i_{i}(x) > 0 \\
-\infty & \text{otherwise},
\end{cases}
$$

where $-\infty$ is added at the bottom of $\mathbb{R}$. The value of $W_i(x)$ is $-\infty$ when $x$ is not a $C$-element (i.e., $C^i_{i}(x) = 0$). Otherwise, $C^i_{i}(x) > 0$ and the higher is the sum $W_i(x)$, the more typical is the element $x$ relative to the defeasible properties of $C_i$.

**Example 2**

Let us consider again Example 1. Let $I$ be an $G_n\mathcal{ALCT}$ interpretation such that $\text{Fly}^I(\text{reddy}) = (\text{Has Winters})^I(\text{reddy}) = (\text{Has Feather})^I(\text{reddy}) = 1$ and $\text{Red}^I(\text{reddy}) = 1$, and $\text{Black}^I(\text{reddy}) = 0$. Suppose further that $\text{Fly}^I(\text{opus}) = 0$ and $(\text{Has Winters})^I(\text{opus}) = (\text{Has Feather})^I(\text{opus}) = 1$ and $\text{Black}^I(\text{opus}) = 0.8$. Considering the weights of typicality inclusions for Bird, $W_{\text{Bird}(\text{reddy})} = 20 + 50 + 50 = 120$ and $W_{\text{Bird}(\text{opus})} = 0 + 50 + 50 = 100$. This suggests that Reddy should be more typical as a bird than Opus. On the other hand, if we suppose that $\text{Bird}^I(\text{reddy}) = 1$ and $\text{Bird}^I(\text{opus}) = 0.8$, then $W_{\text{penguin}}(\text{reddy}) = 100 - 70 = 30$ and $W_{\text{penguin}}(\text{opus}) = 0.8 \times 100 + 0.8 \times 50 = 120$, and Reddy should be less typical as a penguin than Opus.

In previous work, a notion of coherence is introduced by Giordano and Theseider Dupré (2021b) to force an agreement between the preference relations $\prec_c$, induced by a fuzzy
interpretation $I$, for distinguished concepts $C_i$, and the weights $W_i(x)$ computed, for each $x \in \Delta$, from the knowledge base $K$, given the interpretation $I$. In the many-valued case, this leads to the following definition of coherent multipreference model of a weighted $G_nALC$ ($L_nALC$) knowledge base.

**Definition 3 (Coherent multipreference model of a weighted $G_nALC$ ($L_nALC$) KB)**

Let $K = \langle T, T_{C_1}, \ldots, T_{C_k}, A \rangle$ be a weighted $G_nALC$ ($L_nALC$) knowledge base over $C$. A coherent multipreference model (cm-model) of $K$ is a $G_nALC$ ($L_nALC$) interpretation $I = \langle \Delta, I \rangle$ s.t.:

- $I$ satisfies the inclusions in $T$ and the assertions in $A$;
- for all $C_i \in C$, the preference $<_C$ is coherent to $T_{C_i}$, that is, for all $x, y \in \Delta$,

$$x <_C y \iff W_i(x) > W_i(y).$$

(4)

In a similar way, one can define a faithful multipreference model (fm-model) of $K$ by replacing the coherence condition (4) with a faithfulness condition: for all $x, y \in \Delta$,

$$x <_C y \Rightarrow W_i(x) > W_i(y).$$

(5)

The weaker notion of faithfulness allows to define a larger class of multipreference models of a weighted KB, compared to the class of coherent models. This allows a larger class of monotone non-decreasing activation functions in neural network models to be captured, whose activation function is monotonically non-decreasing (we refer to the work by Giordano and Theseider Dupré (2021b) and by Giordano (2021b).

### 4 $\varphi$-coherent models with finitely many values

In this section we consider another notion of coherence of a many-valued interpretation $I$ wrt a KB, that we call $\varphi$-coherence, where $\varphi$ is a function from $\mathbb{R}$ to the interval $[0, 1]$, that is, $\varphi : \mathbb{R} \to [0, 1]$. $\varphi$-coherent models have been first introduced by Giordano (2021a) in the definition of a gradual argumentation semantics. Let us consider $\varphi$-coherent $G_nALC$ ($L_nALC$) interpretations.

**Definition 4 ($\varphi$-coherence)**

Let $K = \langle T, T_{C_1}, \ldots, T_{C_k}, A \rangle$ be a weighted $G_nALC$ ($L_nALC$) knowledge base, and $\varphi : \mathbb{R} \to [0, 1]$. A $G_nALC$ ($L_nALC$) interpretation $I = \langle \Delta, I \rangle$ is $\varphi$-coherent if, for all concepts $C_i \in C$ and $x \in \Delta$,

$$C_i^I(x) = \varphi \left( \sum_h w^i_h D^I_{i,h}(x) \right),$$

(6)

where $T_{C_i} = \{(T(C_i) \sqsubseteq D_{i,h}, w^i_h)\}$ is the set of weighted conditionals for $C_i$. A $\varphi$-coherent multipreference model ($\varphi$-coherent model) of a knowledge base $K$, is defined as a coherent model in Definition 3, but replacing the notion of coherence in condition (4) with the notion of $\varphi$-coherence (6).

The relationships between the three semantics (Giordano 2021a) extend to the finite many-valued case as follows.
Proposition 1

Let $K$ be a weighted $G_n$ALC\(T\) (L\(n\)ALC\(T\)) knowledge base and $\varphi : \mathbb{R} \rightarrow [0, 1]$. (1) if $\varphi$ is a monotonically non-decreasing function, a $\varphi$-coherent multipreference model $I$ of $K$ is also a faithful model of $K$; (2) if $\varphi$ is a monotonically increasing function, a $\varphi$-coherent multipreference model $I$ of $K$ is also a coherent model of $K$.

To see that the set of equations defined by (6) allow to characterize the stationary states of multilayer perceptrons (MLPs), let us consider (Haykin 1999) the model of a neuron as an information-processing unit in an (artificial) neural network. The basic elements are the following: (1) a set of synapses or connecting links, each one characterized by a weight. We let $x_j$ be the signal at the input of synapse $j$ connected to neuron $i$, and $w_{ij}$ the related synaptic weight; (2) the adder for summing the input signals to the neuron, weighted by the respective synapses weights: $\sum_{j=1}^{n} w_{ij} x_j$; (3) an activation function for limiting the amplitude of the output of the neuron (here, we assume, to the interval $[0, 1]$). A neuron $i$ can be described by the following pair of equations: $u_i = \sum_{j=1}^{n} w_{ij} x_j$ and $y_i = \varphi(u_i + b_i)$ where $x_1, \ldots, x_n$ are the input signals and $w_{i1}, \ldots, w_{in}$ are the weights of neuron $i$; $b_i$ is the bias, $\varphi$ the activation function, and $y_i$ is the output signal of neuron $i$. By adding a new synapse with input $x_0 = +1$ and synaptic weight $w_{i0} = b_i$, one can write: $u_i = \sum_{j=0}^{n} w_{ij} x_j$, and $y_i = \varphi(u_i)$, where $u_i$ is called the induced local field of the neuron.

A neural network $\mathcal{N}$ can then be seen as “a directed graph consisting of nodes with interconnecting synaptic and activation links” (Haykin 1999). Nodes in the graph are the neurons (the processing units) and the weight $w_{ij}$ on the edge from node $j$ to node $i$ represents the strength of the connection between unit $j$ and unit $i$.

A mapping of a neural network to a conditional KB can be defined in a simple way (Giordano and Theseider Duprè 2021b), associating a concept name $C_i$ with each unit $i$ in the network and by introducing, for each synaptic connection from neuron $h$ to neuron $i$ with weight $w_{ih}$, a conditional $T(C_i) \subseteq C_h$ with weight $w^i_h = w_{ih}$. If we assume that $\varphi$ is the activation function of all units in the network $\mathcal{N}$ and we consider the infinite-valued fuzzy logic with truth space $S = [0, 1]$, then the solutions of equations (6) characterize the stationary states of MLPs, where $C_i^T(x)$ corresponds to the activation of neuron $i$ for some input stimulus $x$, each $D_{i,h}^T(x)$ corresponds to the input signal $x_h$, and $\sum_h w^i_h D_{i,h}^T(x)$ corresponds to the induced local field of neuron $i$.

Notice that, when the truth space is the finite set $C_n$, for $n \geq 1$, the notion of $\varphi$-coherence may fail to characterize all the stationary states of a network, simply as there may be stationary states such that the activity values of units fall outside $C_n$. In the next section, we will consider an approximation $\varphi_n$ of the function $\varphi$ over $C_n$, with the idea to capture an approximated behavior of the network based on the finite many-valued semantics of a weighted conditional KB, and to construct a preferential model for properties verification.

For a weighted $G_n$ALC\(T\) (L\(n\)ALC\(T\)) knowledge base $K$, a notion of coherent/faithful/$\varphi$-coherent entailment can be defined in a natural way. As for concept-wise entailment in the two-valued case (Giordano et al. 2015), we restrict our consideration to canonical models, that is, models which are large enough to contain all the relevant domain elements with their different valuations. Informally, a canonical $\varphi$-coherent model of $K$ is a $\varphi$-coherent model of $K$ that contains a domain element for each possible valuation of concepts which is present in any $\varphi$-coherent model of $K$. Similarly for coherent and faithful models.
Definition 5 (Canonical coherent/faithful/$\varphi$-coherent model of $K$)
Given a weighted $G_n\mathit{LCT}$ ($L_n\mathit{LCT}$) knowledge base $K$, $I = (\Delta, I)$ is a canonical coherent/faithful/$\varphi$-coherent model of $K$ if: (i) $I$ is a coherent/faithful/$\varphi$-coherent model of $K$ and, (ii) for each coherent/faithful/$\varphi$-coherent model $J = (\Delta^J, I^J)$ of $K$ and each $y \in \Delta^J$, there is an element $z \in \Delta$ such that $B_I(z) = B_J^I(y)$, for all concept names $B$ occurring in $K$.

A result concerning the existence of canonical $\varphi$-coherent models, for weighted KBs having at least a $\varphi$-coherent model, can be found in the supplementary material for the paper, Appendix A. Let us define entailment.

Definition 6 (coherent/faithful/$\varphi$-coherent entailment)
Given a weighted $G_n\mathit{ALCT}$ ($L_n\mathit{ALCT}$) knowledge base $K$, a $G_n\mathit{ALCT}$ ($L_n\mathit{ALCT}$) axiom $E$ is coherently/faithfully/$\varphi$-coherently entailed from $K$ if, for all canonical coherent/faithful/$\varphi$-coherent models $I = (\Delta, I)$ of $K$, $I$ satisfies $E$.

The properties of faithful entailment in the fuzzy case have been studied by Giordano (2021b). Faithful entailment is well-behaved: it deals with specificity and irrelevance; it is not subject to inheritance blocking; it satisfies most of the KLM properties of a preferential consequence relation (Kraus et al. 1990; Lehmann and Magidor 1992), depending on their fuzzy reformulation and on the chosen combination functions.

In the next section, we restrict our consideration to the boolean fragment $\mathit{LC}$ of $\mathit{ALC}$ (with neither roles, nor universal nor existential restrictions), which is sufficient to encode MLPs as weighted KBs and to formulate boolean properties of the network. We consider the finitely many-valued logics $G_n\mathit{LCT}$ and $L_n\mathit{LCT}$, and exploit ASP and asprin for defeasible reasoning in $G_n\mathit{LCT}$ and $L_n\mathit{LCT}$ under an approximation $\varphi_n$ of $\varphi$.

5 ASP and asprin for reasoning in $G_n\mathit{LCT}$ and $L_n\mathit{LCT}$: $\varphi_n$-coherence and verification of MLPs

Given a monotonically non-decreasing function $\varphi : \mathbb{R} \to [0,1]$, and an integer $n > 1$, let function $\varphi_n : \mathbb{R} \to C_n$ be defined as follows:

$$
\varphi_n(x) = \begin{cases} 
0 & \text{if } \varphi(x) \leq \frac{1}{2n} \\
\frac{i}{n} & \text{if } \frac{2i-1}{2n} < \varphi(x) \leq \frac{2i+1}{2n}, \text{ for } 0 < i < n \\
1 & \text{if } \frac{2n-1}{2n} < \varphi(x) 
\end{cases}
$$

(7)

$\varphi_n(x)$ approximates $\varphi(x)$ to the nearest value in $C_n$. The notions of $\varphi_n$-coherence, $\varphi_n$-coherent model, canonical $\varphi_n$-coherent model, $\varphi_n$-coherent entailment can be defined as in Definitions 4 and 6, by replacing $\varphi$ with $\varphi_n$. The above-mentioned result concerning the existence of canonical models also extends to canonical $\varphi_n$-coherent models of weighted KBs (see Proposition 4 in the supplementary material for the paper, Appendix A).

In the following, we formulate the problem of $\varphi_n$-coherent entailment from a weighted $G_n\mathit{LCT}$ ($L_n\mathit{LCT}$) knowledge base as a problem of computing preferred answer sets of an ASP program. Verifying $\varphi_n$-coherent entailment of a typicality inclusion $T(C) \subseteq D \theta \alpha$ from a weighted knowledge base $K$ (a subsumption problem), would require considering all typical $C$-elements in all possible canonical $\varphi_n$-coherent models of $K$, and checking whether they are all instances of $D$ with a degree $d$ such that $d\theta\alpha$. We reformulate this
problem as a problem of generating answer sets representing $\varphi_n$-coherent models of the KB, and then selecting preferred answer sets, where a distinguished domain element $aux_C$ is intended to represent a typical $C$-element. For the selection of preferred answer sets, the ones maximizing the degree of membership of $aux_C$ in concept $C$, we use asprin by Brewka et al. (2015). Our proof method is sound and complete for the computation of $\varphi_n$-coherent entailment.

Given a weighted $G_n\text{LCT}$ ($L_n\text{LCT}$) knowledge base $K = \langle T, T_{C_1}, \ldots, T_{C_k}, A \rangle$, we let $\Pi_{K,n}$ be the representation of $K$ in Datalog, where: $val(v)$ holds for each value $v$ in \{0, 1, \ldots, n\}, which is intended to represent the value $\frac{v}{n}$ in $C_n$; $nom(a)$, $cls(A)^1$, are used for $a \in N_I$, $A \in N_C$. We also have $nom(aux)^2$.

Boolean concepts $C \sqcap D$, $C \sqcup D$, $\neg C$ are represented as $\text{and}(C', D')$, $\text{or}(C', D')$, $\text{neg}(C')$, where $C'$ and $D'$ are terms representing concepts $C$ and $D$; $\text{subTyp}(C', D', w')$ represents a defeasible inclusion ($T(C) \subseteq D, w$), where $w'$ is an integer corresponding to $w \times 10^k$; for $w$ approximated to $k$ decimal places. The concepts of interest, to be considered for limiting grounding in the rules introduced later, are represented (1) with assertions $\text{concept}(C')$, where $C'$ is the term for boolean concepts $C$ occurring in $K$ or in the formula to be verified (see later); (2) with rules implying that subconcepts are also of interest, for example:

$$\text{concept}(A) \leftarrow \text{concept}(\text{and}(A, B)).$$

$\Pi_{K,n}$ also contains the set of rules for generating $\varphi_n$-coherent models of $K$. The valuation is encoded by a set of atoms of the form $\text{inst}(x, A, v)$, meaning that $\frac{v}{n} \in C_n$ is the degree of membership of $x$ in $A$. The rule:

$$1\{\text{inst}(X, A, V) : val(V)\} 1 \leftarrow \text{cls}(A), \text{nom}(X).$$

generates alternative answer sets, corresponding to interpretations of each constant $x$ (either a named individual or $aux_C$), with different values $v$ corresponding to a membership degree $\frac{v}{n} \in C_n$ in each atomic concept $A$.

The valuation of complex boolean concepts $D$ is encoded by introducing a predicate $\text{eval}(D, X, V)$ to determine the membership degree $V$ of element $X$ in $D$. A rule is introduced for each boolean operator to encode its semantics. For $G_n\text{LCT}$, the $\text{eval}$ predicate encodes the semantics of $\cap$, $\sqcup$ and $\neg$, based on Gödel logic t-norm, s-norm and on involutive negation as follows:

$$\text{eval}(A, X, V) \leftarrow \text{cls}(A), \text{inst}(X, A, V).$$
$$\text{eval}(\text{and}(A, B), X, V) \leftarrow \text{concept}(\text{and}(A, B)), \text{eval}(A, X, V_1), \text{eval}(B, X, V_1),$$
$$\text{min}(V_1, V_2, V).$$
$$\text{eval}(\text{or}(A, B), X, V) \leftarrow \text{concept}(\text{or}(A, B)), \text{eval}(A, X, V_1), \text{eval}(B, X, V_1),$$
$$\text{max}(V_1, V_2, V).$$
$$\text{eval}(\text{neg}(A), X, V) \leftarrow \text{concept}(\text{neg}(A)), \text{eval}(A, X, V_1), V = n - V_1,$$

where the predicates min and max are suitably defined. A similar evaluation function $\text{eval}$ can be defined for Lukasiewicz combination functions.

1 Uppercase is used here for concept names, to keep a DL-like notation, even though such names are ASP constants.

2 Observe that the addition of further auxiliary constants to represent other domain elements in a model, used in the Datalog materialization calculus by Krötzsch (2010), is not needed here as neither existential nor universal restrictions are allowed.
An ASP Approach for Reasoning on Neural Networks

To guarantee the satisfiability of \( G_n \mathcal{LCT} \) axioms (assertions and inclusions) a set of constraints is added. For instance, for the assertion \( C(a) \geq \alpha \) we add the constraint

\[
\bot \leftarrow \text{eval}(C', a, V), \ V < n\alpha,
\]

where \( C' \) is the term representing concept \( C \), while for a strict \( G_n \mathcal{LCT} \) inclusion \( E \subseteq D \geq \alpha \) we add the constraint

\[
\bot \leftarrow \text{eval}(E', X, V1), \ \text{eval}(D', X, V2), \ V1 > V2, \ V2 < \alpha,
\]

and similarly for other axioms and for the \( L_n \mathcal{LCT} \) case. An answer set represents a \( \varphi_n \)-coherent interpretation if the following constraint is satisfied:

\[
\bot \leftarrow \text{nom}(X, dcls(C_i)), \ \text{eval}(C_i, X, V), \ \text{weight}(X, C_i, W), \ \text{valphi}(n, W, V1), \ V1 = V1,
\]

where \( dcls(C_i) \) is included in \( \Pi_{K,n} \) for each distinguished class \( C_i \in \mathcal{C} \). Given that the weights \( w^*_i \) are approximated to \( k \) decimal places, argument \( W \) for \( \text{weight} \) corresponds to the integer \( n \times W^*_i(x) \times 10^k \), and \( \text{valphi}(n, W, V1) \) is defined (see below) to correspond to \( V1 = n \times \varphi_n(W^*_i(x)) = n \times \varphi_n(W/(n \times 10^k)) \) again representing \( C_n \) with \( \{0, 1, \ldots, n\} \). Predicate \( \text{weight} \) (for the weighted sum) could, in principle, be defined as follows:

\[
\text{weight}(X, C, W) \leftarrow \text{dcls}(C), \ \text{nom}(X), \ W = \#\text{sum}\{ W_i \ast V, D : \text{cls}(D), \ \text{eval}(D, X, V), \ \text{subTyp}(C, D, W) \}.
\]

but, for grounding reasons, it can be better defined with a rule for each distinguished class; such rules can be generated, for each distinguished concept \( C_i \), from the set of weighted typicality inclusions \( T_{C_i} \). In particular, given \( T_{C_i} = \{(T(C_i) \subseteq D_{i,h}, w^*_h), h = 1, \ldots, k\} \), the following rule is introduced:

\[
\text{weight}(X, C', W) \leftarrow \text{nom}(X), \ W = W_{i1} \ast Vi1 + \ldots + W_{ik} \ast Vik,
\]

\[
\text{subTyp}(C', Di1', Wi1), \ \text{eval}(Di1', X, Vi1), \ldots,
\]

\[
\text{subTyp}(C', Di'k, Vik), \ \text{eval}(Di'k, X, Vik),
\]

where \( C', Di1', \ldots, Di'k \) are the terms representing concepts \( C_i, D_{i,1}, \ldots, D_{i,k} \). Predicate \( \text{valphi} \) can be defined with rules such as:

\[
\text{valphi}(n, W, 0) \leftarrow \text{num}(W), \ W < k_1.
\]

\[
\text{valphi}(n, W, 1) \leftarrow \text{num}(W), \ W \geq k_1, \ W < k_2.
\]

\[
\ldots
\]

\[
\text{valphi}(n, W, n) \leftarrow \text{num}(W), \ W > k_{n-1},
\]

where:

\[
\text{num}(W) \leftarrow \text{nom}(X), \ \text{weight}(X, C, W), \ \text{dcls}(C)
\]

is used for limiting grounding of the previous rules, and \( k_1, \ldots, k_{n-1} \) can be precomputed to be:

\[
k_1 = \lfloor w \rfloor \text{ where } w \text{ is such that } \varphi(w/(n \times 10^k)) = 1/2n,
\]

\[
k_2 = \lfloor w \rfloor \text{ where } w \text{ is such that } \varphi(w/(n \times 10^k)) = 3/2n,
\]

\[
\ldots
\]

\[
k_{n-1} = \lfloor w \rfloor \text{ where } w \text{ is such that } \varphi(w/(n \times 10^k)) = (2n - 1)/2n.
\]

The program \( \Pi(K, n, C, D, \theta, \alpha) \) associated to the \( G_n \mathcal{LCT} \) knowledge base \( K \) and a typicality subsumption \( T(C) \subseteq D \theta \alpha \) is composed of two parts, \( \Pi(K, n, C, D, \theta, \alpha) = \Pi_{K,n} \cup \Pi_{C,D,\theta,\alpha} \). We have already introduced the first one. \( \Pi_{C,D,n,\theta,\alpha} \) contains the facts \( \text{nom} (auxC) \) and \( auxtc (auxC, C') \) and the rules:

\[
ok \leftarrow \text{eval}(D', auxC, V), \ V \theta \alpha n. \quad \text{notok} \leftarrow \text{not ok},
\]
where \( ok \) is intended to represent that \( aux_C \) satisfies the property that its membership degree \( V \) in concept \( D \) is such that \( V\theta\alpha \) holds.

Given a query \( T(C) \subseteq D \theta \alpha \), we have to verify that, in all canonical \( \varphi_n \)-coherent models of the \( G_n\text{LCT} \) (\( L_n\text{LCT} \)) knowledge base, all typical \( C \)-elements are \( D \)-elements with a certain degree \( v \) (representing \( v/n \in C_n \)) such that \( v\theta\alpha\). This verification is accomplished by generating answer sets corresponding to the \( \varphi_n \)-coherent models of the KB, and by selecting the preferred ones, in which the distinguished element \( aux_C \) represents a typical \( C \)-element.

Given two answer sets \( S \) and \( S' \) of \( \Pi(K, n, C, D, \theta, \alpha) \), \( S \) is preferred to \( S' \) if the membership degree of \( aux_C \) in concept \( C \) is higher in \( S \) than in \( S' \), that is: if \( eval(C', aux_C, v1) \) holds in \( S \) and \( eval(C', aux_C, v2) \) holds in \( S' \), then \( v1 > v2 \).

This condition is encoded directly into a preference program for \textit{asprin} as follows. One such program requires defining when an answer set \( S \) is preferred to \( S' \) according to a preference \( P \) (optimal solutions wrt such a preference can then be required with an \#\textit{optimize} directive). This is done by defining a predicate \textit{better}(\( P \)) for the case where \( P \) is of the type being defined, using predicates \textit{holds} and \textit{holds'} to check whether atoms hold in \( S \) and \( S' \), respectively. In this case the preference program, defining a “concept wise” preference, is simply as follows:

\[
\begin{align*}
\# & \text{program preference(cwise).} \\
\text{better}(P) & \leftarrow \text{preference}(P, \text{cwise}), \text{holds}(auxc(aux, C)), \text{betterwrt}(C). \\
\text{betterwrt}(C) & \leftarrow \text{holds}(eval(C, aux, V1)), \text{holds'}(eval(C, aux, V2)), V1 > V2.
\end{align*}
\]

The query \( T(C) \subseteq D \theta \alpha \) is entailed from the knowledge base \( K \) if, in all (maximally) preferred answer sets, \( aux_C \) is an instance of concept \( D \) with a membership degree \( v \) (representing \( v/n \in C_n \)) such that \( v\theta\alpha \) holds; that is, if \( ok \) holds in all preferred answer sets, or, equivalently, \textit{notok} does not hold in any of them. In fact, we can prove that this corresponds to verifying that \( D \) is satisfied in all \( <_C \)-minimal \( C \)-elements in all canonical \( \varphi_n \)-coherent models of the KB:

**Proposition 2**

Given a \( G_n\text{LCT} \) (\( L_n\text{LCT} \)) knowledge base \( K \), the query \( T(C) \subseteq D \theta \alpha \) is falsified in some canonical \( \varphi_n \)-coherent model of \( K \) if and only if there is a preferred answer set \( S \) of the program \( \Pi(K, C, D, n, \theta, \alpha) \) containing \( eval(D', aux_C, v) \) such that \( v\theta\alpha\) does not hold (and then containing \textit{notok}).

The proof can be found in the supplementary material for the paper, Appendix B. It exploits the existence of \( \varphi_n \)-coherent canonical models, for KBs having a \( \varphi_n \)-coherent model (Proposition 4 in Appendix A). Appendix B also contains a proof of the following upper bound on the complexity of \( \varphi_n \)-coherent entailment.

**Proposition 3**

\( \varphi_n \)-coherent entailment from a weighted \( G_n\text{LCT} \) (\( L_n\text{LCT} \)) knowledge base is in \( \Pi^p_2 \).

As a proof of concept, the approach has been experimented for the weighted \( G_n\text{LCT} \) KBs corresponding to two of the trained multilayer feedforward network for the MONK’s problems (Thrun et al. 1991), namely, the network for problem 1 and the second network for problem 3. The networks have 17 non-independent binary inputs, corresponding to values of 6 inputs having 2 to 4 possible values; such inputs are features of a robot, for
example, head shape and body shape being round, square or octagon, and jacket color being red, yellow, green or blue. The network for problem 1 (Figure 1) has 3 hidden units \((h_1, h_2, h_3)\) and an output unit \((o)\); the one for problem 3 has 2 hidden units.

In the two problems, the trained networks learned to classify inputs satisfying two formulae, respectively, \(F_1\) and \(F_3\), which are boolean combinations of the inputs. In particular, \(F_1\) is \(jacket\_color = red\) or \(head\_shape = body\_shape\) and, in terms of the classes \(i_1, \ldots, i_{17}\) corresponding to the binary inputs, it is:

\[
F_1 \equiv i_1 \lor (i_1 \land i_4) \lor (i_2 \land i_5) \lor (i_3 \land i_6)
\]

\((i_1)\) is \(jacket\_color\), \((i_2)\) is \(head\_shape\). \((i_4)\) is \(body\_shape\), etc.).

The approach described above has been applied, using values 0 and 1 as possible values for classes associated with input nodes, rather than all values in \(C_n\). The networks are feedforward, then for a choice of values for input nodes, there is only one choice of values in \(C_n\) for non-input nodes satisfying the constraint for \(\varphi_n\)-coherent interpretations (then the number of answer sets of \(\Pi(K, C, D, n, \theta, \alpha)\) is given by the possible combinations of input values and does not depend on \(n\)).

For the trained network for problem 1, the formula \(T(o) \sqsubseteq F_1 \geq 1\) can be verified, for example, for \(n = 5\); \(o\) is the concept name associated with the output unit. That is, the \(G_5\text{LCT}\) knowledge base entails that the typical \(o\)-elements satisfy \(F_1\). The formula can also be verified for \(n = 1, 3, 9\) with minor variations on the running times (all below 10 s). This result is explainable (also for \(n=1\)), as an input was classified by the network as class member if the output was \(\geq 0.5\) and, for problem 1, the network learned the concept with 100% accuracy.

Stronger variants of \(F_1\) have also been considered, to check that the network learned \(F_1\) but not such variants. For the following variants with one less disjunct:

\[
F_1' \equiv i_1 \lor (i_1 \land i_4) \lor (i_2 \land i_5) \lor (i_3 \land i_6)
\]

\(F_1'\) is not entailed for \(n = 1, 3, 9\). \(T(o) \sqsubseteq F_1' \geq 1\) and \(T(o) \sqsubseteq F_1'' \geq 1\) are indeed not entailed for \(n = 1, 3, 5, 9\).

An important issue in analyzing a trained network is also associating a meaning to hidden nodes. The following formulae have been verified for \(n = 1, 3, 5, 9\) for hidden nodes \(h_1, h_2, h_3\):

\[
\begin{align*}
T(h_1) & \sqsubseteq i_1 \lor (\neg i_1 \land \neg i_4) \geq 1 \\
T(h_2) & \sqsubseteq i_1 \lor (\neg i_3 \land \neg i_6) \geq 1 \\
T(h_3) & \sqsubseteq \neg i_1 \lor (i_2 \land i_5) \geq 1.
\end{align*}
\]
In problem 3, there was noise (some misclassifications) in the training set. Then the accuracy of the trained network is not 100%. However, the trained network produces no false positives. Therefore, the formula $T(o) \sqsubseteq F3 \geq 1$ can be verified for $n = 1, 3, 5, 9$, where $F3$ is $(\text{jacket\_color\_red and holding\_sword})$ or $(\text{not jacket\_color\_blue and not body\_shape\_octagon})$. Since there are false negatives, the formula $T(\lnot o) \sqsubseteq \lnot F3 \geq 1$ is not entailed for $n = 1$ but, for instance, it is for $n = 5$.

6 Conclusions

The “concept-wise” multipreference semantics (both in the two-valued and in the fuzzy case) has recently been proposed as a logical semantics of MLPs by Giordano and Theseider Dupré (2021b). In this paper we consider weighted conditional $\mathcal{ALC}$ knowledge bases in the finitely many-valued case, under a coherent, a faithful, and a $\varphi$-coherent semantics, the last one being suitable to characterize the stationary states of MLPs. For the boolean fragment $\mathcal{LC}$ of $\mathcal{ALC}$ we exploit ASP and asprin, see Brewka et al. (2015), for reasoning under $\varphi$-coherent entailment, by restricting to canonical models of the KB. We have proven soundness and completeness of ASP encoding for the finitely many-valued case and provided an upper complexity bound. As a proof of concept, we have experimented the proposed approach for checking properties of some trained neural networks for the MONK’s problems, see Thrun et al. (1991).

Undecidability results for fuzzy DLs with general inclusion axioms (Cerami and Straccia 2011; Borgwardt and Peñaloza 2012), motivate the investigation of many-valued approximations of fuzzy multipreference entailment. The choice of $\mathcal{LC}$ is motivated by the fact it is sufficient to encode a neural network as a weighted KB as well as to formulate boolean properties of the network. This work is a first step towards the definition of proof methods for reasoning from weighted KBs under a finitely many-valued preferential semantics in more expressive or lightweight DLs. For $\mathcal{EL}^\bot$, the two-valued case has been studied in previous work by Giordano and Theseider Dupré (2021a).

The encoding of a neural network as a conditional KB opens the possibility of combining empirical knowledge with elicited knowledge, for example, in the form of strict inclusions and definitions. Much work has been devoted, in recent years, to the combination of neural networks and symbolic reasoning (see the survey by Lamb et al. 2020), leading to the definition of new computational models and to extensions of logic programming languages with neural predicates. The relationships between normal logic programs and connectionist network have been investigated by d’Avila Garcez and Zaverucha (1999) and by Hitzler et al. (2004). A correspondence between neural networks and gradual argumentation semantics has been recently investigated by Potyka (2021) by studying the semantic properties and the convergence conditions of a MLP-based bipolar semantics. The correspondence between neural network models and fuzzy systems has been first investigated by Kosko (1992) in his seminal work. A fuzzy extension of preferential logics has been studied by Casini and Straccia (2013b) based on rational closure.

While using preferential logic for the verification of properties of neural networks is a general (model agnostic) approach, first proposed for SOMs by Giordano et al. (2020, 2022), whether it is possible to extend the logical encoding of MLPs as weighted conditional KBs to other network models is a subject for future investigation. The development
of a temporal extension of weighted conditional KBs to capture the transient behavior of MLPs is also an interesting direction to extend this work.

Supplementary material

To view supplementary material for this article, please visit http://doi.org/10.1017/S1471068422000163.

References

Adadi, A. AND BERRADA, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.

Arrieta, A. B., Rodríguez, N. D., SER, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., GL-Lopez, S., Molina, D., BenjaminS, R., Chatila, R. AND Herrera, F. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115.

Baader, F., Brandt, S. AND Lutz, C. 2005. Pushing the EL envelope. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005), L. Kaelbling and A. Saffiotti, Eds. Professional Book Center, Edinburgh, Scotland, UK, 364–369.

Benferhat, S., Cayrol, C., Dubois, D., Lang, J. AND PRADE, H. 1993. Inconsistency management and prioritized syntax-based entailment. In IJCAI’93, Chambéry, 640–647.

Bobillo, F., Delgado, M., Gómez-Romero, J. AND Straccia, U. 2012. Joining Gödel and Zadeh fuzzy logics in fuzzy description logics. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 20, 4, 475–508.

Bobillo, F. AND Straccia, U. 2011. Reasoning with the finitely many-valued Lukasiewicz fuzzy Description Logic SROIQ. Information Sciences 181, 4, 758–778.

Bobillo, F. AND Straccia, U. 2018. Reasoning within fuzzy OWL 2 EL revisited. Fuzzy Sets and Systems 351, 1–40.

Borgwardt, S. AND Peñaloza, R. 2012. Undecidability of fuzzy description logics. In Proceedings of KR 2012, Rome, Italy, 10–14 June 2012, G. Brewka, T. Eiter and S. A. McIlraith, Eds. AAAI Press.

Borgwardt, S. AND Peñaloza, R. 2013. The complexity of lattice-based fuzzy description logics. Journal on Data Semantics 2, 1, 1–19.

Brewka, G., Delgrande, J. P., Romero, J. AND Schaub, T. 2015. asprin: Customizing answer set preferences without a headache. In Proceedings of the AAAI 2015, 1467–1474.

Britz, K., Heidema, J. AND Meyer, T. 2008. Semantic preferential subsumption. In KR 2008, G. Brewka and J. Lang, Eds. AAAI Press, Sidney, Australia, 476–484.

Casini, G., Meyer, T. A. AND Varzinczak, I. 2021. Contextual conditional reasoning. In AAAI-21, Virtual Event, 2–9 February 2021. AAAI Press, 6254–6261.

Casini, G. AND Straccia, U. 2010. Rational closure for defeasible description logics. In JELIA 2010, T. Janhunen and I. Niemelä, Eds. LNCS, vol. 6341. Springer, Helsinki, 77–90.

Casini, G. AND Straccia, U. 2013a. Defeasible inheritance-based description logics. Journal of Artificial Intelligence Research (JAIR) 48, 415–473.

Casini, G. AND Straccia, U. 2013b. Towards rational closure for fuzzy logic: The case of propositional Gödel logic. In Proceedings of LPAR-19, Stellenbosch, South Africa, 14–19 December 2013. LNCS, vol. 8312. Springer, 213–227.

Cerami, M. AND Straccia, U. 2011. On the undecidability of fuzzy description logics with GCIs with Lukasiewicz t-norm. CoRR abs/1107.4212.
Cintula, P., Hájek, P. and Noguera, C., Eds. 2011. Handbook of Mathematical Fuzzy Logic, vol. 37–38. College Publications.

d’Avila Garcez, A. S. and Zaverucha, G. 1999. The connectionist inductive learning and logic programming system. Applied Intelligence 11, 1, 59–77.

Delgrande, J. and Rantsoudis, C. 2020. A preference-based approach for representing defaults in first-order logic. In Proceedings of the 18th International Workshop on Non-Monotonic Reasoning, NMR.

García-Cerdaña, A., Armengol, E. and Esteva, F. 2010. Fuzzy description logics and t-norm based fuzzy logics. International Journal of Approximate Reasoning 51, 6, 632–655.

Giordano, L. 2021a. From weighted conditionals of multilayer perceptrons to gradual argumentation. Presented in 5th Workshop on Advances In Argumentation In Artificial Intelligence (AI³@ AIXIA 2021), November 29, 2021, https://arxiv.org/abs/2110.03643.

Giordano, L. 2021b. On the KLM properties of a fuzzy DL with Typicality. In Proceedings of the ECSQARU 2021, Prague, Czech Republic, 21–24 September 2021. LNCS, vol. 12897. Springer, 557–571.

Giordano, L. and Gliozzi, V. 2021. A reconstruction of multipreference closure. Artificial Intelligence 290, 1–34.

Giordano, L., Gliozzi, V. and Theseider Dupré, D. 2022. A conditional, a fuzzy and a probabilistic interpretation of self-organising maps. Journal of Logic and Computation 32, 2, 178–205.

Giordano, L., Gliozzi, V., Olivetti, N. and Pozzato, G. L. 2007. Preferential description logics. In LPAR 2007. LNAI, vol. 4790. Springer, Yerevan, Armenia, 257–272.

Giordano, L., Gliozzi, V., Olivetti, N. and Pozzato, G. L. 2015. Semantic characterization of rational closure: From propositional logic to description logics. Artificial Intelligence 226, 1–33.

Giordano, L., Gliozzi, V. and Theseider Dupré, D. 2020. On a plausible concept-wise multipreference semantics and its relations with self-organising maps. In CILC 2020, Rende, IT, 13–15 October 2020, F. Calimeri, S. Perri and E. Zumpano, Eds. CEUR, vol. 2710, 127–140.

Giordano, L. and Theseider Dupré, D. 2020. An ASP approach for reasoning in a concept-aware multipreferential lightweight DL. Theory and Practice of Logic Programming, TPLP 10, (5), 751–766.

Giordano, L. and Theseider Dupré, D. 2021a. Weighted conditional EL⊥ knowledge bases with integer weights: An ASP approach. In Proceedings 37th International Conference on Logic Programming, ICLP 2021 (Technical Communications), Porto, 20–27 September 2021. EPTCS, vol. 345, 70–76.

Giordano, L. and Theseider Dupré, D. 2021b. Weighted defeasible knowledge bases and a multipreference semantics for a deep neural network model. In Proceedings of the JELIA 2021, 17–20 May. LNCS, vol. 12678. Springer, 225–242. Extended version in https://arxiv.org/abs/2103.06854.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. 2019. A survey of methods for explaining black box models. ACM Computing Surveys 51, 5, 93:1–93:42.

Haykin, S. 1999. Neural Networks - A Comprehensive Foundation. Pearson.

Hitzler, P., Hölldobler, S. and Seda, A. K. 2004. Logic programs and connectionist networks. Journal of Applied Logic 2, 3, 245–272.

Kern-Isberner, G. 2001. Conditionals in Nonmonotonic Reasoning and Belief Revision - Considering Conditionals as Agents. LNCS, vol. 2087. Springer.

Kohonen, T., Schroeder, M. and Huang, T., Eds. 2001. Self-Organizing Maps, 3rd ed. Springer Series in Information Sciences. Springer.
An ASP Approach for Reasoning on Neural Networks

Kosko, B. 1992. Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence. Prentice Hall.

Kraus, S., Lehmann, D. and Magidor, M. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. Artificial Intelligence 44, 1-2, 167–207.

Krötzsch, M. 2010. Efficient inferencing for OWL EL. In Proceedings of JELIA 2010, 234–246.

Lamb, L. C., d’Avila Garcez, A. S., Gori, M., Prates, M. O. R., Avelar, P. H. C. and Vardi, M. Y. 2020. Graph neural networks meet neural-symbolic computing: A survey and perspective. In Proceedings of IJCAI 2020, C. Bessiere, Ed. ijcai.org, 4877–4884.

Lehmann, D. and Magidor, M. 1992. What does a conditional knowledge base entail? Artificial Intelligence 55, 1, 1–60.

Lehmann, D. J. 1995. Another perspective on default reasoning. Annals of Mathematics and Artificial Intelligence 15, 1, 61–82.

Lukasiewicz, T. and Straccia, U. 2009. Description logic programs under probabilistic uncertainty and fuzzy vagueness. International Journal of Approximate Reasoning 50, 6, 837–853.

Pearl, J. 1990. System Z: A natural ordering of defaults with tractable applications to non-monotonic reasoning. In TARK’90, Pacific Grove, CA, USA, 121–135.

Potyka, N. 2021. Interpreting neural networks as quantitative argumentation frameworks. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, 2–9 February 2021. AAAI Press, 6463–6470.

Stoilos, G., Stamou, G. B., Tzouvaras, V., Pan, J. Z. and Horrocks, I. 2005. Fuzzy OWL: Uncertainty and the semantic web. In OWLED*05 Workshop on OWL: Experiences and Directions, Galway, Ireland, 11–12 November 2005. CEUR Workshop Proceedings, vol. 188.

Straccia, U. 2005. Towards a fuzzy description logic for the semantic web (preliminary report). In ESWC 2005, Heraklion, Crete, 29 May–1 June 2005. LNCS, vol. 3532. Springer, 167–181.

Thrun, S. et al. 1991. A Performance Comparison of Different Learning Algorithms. Tech. Rep. CMU-CS-91-197, Carnegie Mellon University.