On Probabilistic Inference in Relational Conditional Logics

Matthias Thimm,
Technische Universität Dortmund, Germany
matthias.thimm@tu-dortmund.de

Gabriele Kern-Isberner,
Technische Universität Dortmund, Germany
gabriele.kern-isberner@cs.uni-dortmund.de

Abstract

The principle of maximum entropy has proven to be a powerful approach for commonsense reasoning in probabilistic conditional logics on propositional languages. Due to this principle, reasoning is performed based on the unique model of a knowledge base that has maximum entropy. This kind of model-based inference fulfills many desirable properties for inductive inference mechanisms and is usually the best choice for reasoning from an information theoretical point of view. However, the expressive power of propositional formalisms for probabilistic reasoning is limited and in the past few years many proposals have been given for probabilistic reasoning in relational settings. It seems to be a common view that in order to interpret probabilistic first-order sentences, either a statistical approach that counts (tuples of) individuals has to be used, or the knowledge base has to be grounded to make a possible worlds semantics applicable, for a subjective interpretation of probabilities. Most of these proposals of the second type rely on extensions of traditional probabilistic models like Bayes nets or Markov networks whereas there are only few works on first-order extensions of probabilistic conditional logic. Here, we take an approach of lifting maximum entropy methods to the relational case by employing a relational version of probabilistic conditional logic. First, we propose two different semantics and model theories for interpreting first-order probabilistic conditional logic. We address the problems of ambiguity that are raised by the difference between subjective and statistical views, and develop a comprehensive list of desirable properties for inductive model-based probabilistic inference in relational frameworks. Finally, by applying the principle of maximum entropy in the two different semantical frameworks, we obtain inference operators that fulfill these properties and turn out to be reasonable choices for reasoning in first-order probabilistic conditional logic.

Keywords: First-Order Logic, Probabilistic Reasoning, Maximum Entropy, Conditional Logic

1 Introduction

Applying probabilistic reasoning methods to relational representations of knowledge is a very active and controversial area of research. During the past few years the fields of *probabilistic inductive logic programming* and *statistical relational learning* have put forth many proposals that deal with combining traditional probabilistic models of knowledge like Bayes nets or Markov nets [29] with first-order logic, cf. [8, 14]. For example, two of the most prominent approaches for extending propositional approaches to the relational case are Bayesian Logic Programs [14, Ch. 10] and Markov Logic Networks [14, Ch. 12], extending Bayes nets and Markov nets, respectively. Other formalisms are Probabilistic Relational Models [14, Ch. 5], Logical Bayesian

Networks [10], and Relational Bayesian Networks [19]. Most of these frameworks employ knowledge-based model construction techniques [37, 6] to reduce the problem of probabilistic reasoning in a relational context to probabilistic reasoning in a propositional context. This is done by appropriately grounding the parts of the knowledge base that are needed for answering a particular query and treating these grounded parts as a propositional knowledge base.

However, most of these approaches are primarily concerned with machine learning problems, and do not care about logical or formal properties of relational probabilistic knowledge representation and reasoning in particular. The following example (inspired by [9]) illustrates that even defining a proper semantics for first-order probabilistic knowledge bases is not an easy task. Let $\mathit{elephant}(\mathsf{X})$ denote that X is an elephant, $\mathit{keeper}(\mathsf{X})$ means that X is a keeper, and $\mathit{likes}(\mathsf{X},\mathsf{Y})$ denotes that X likes Y (we denote variables with a beginning uppercase letter, and constants with a beginning lowercase letter). Consider the following rules r_1, r_2, r_3:

$$\begin{array}{lllll} r_1: & \mathit{elephant}(\mathsf{X}) \wedge \mathit{keeper}(\mathsf{Y}) & \rightarrow & \mathit{likes}(\mathsf{X},\mathsf{Y}) & [0.6] \\ r_2: & \mathit{elephant}(\mathsf{X}) \wedge \mathit{keeper}(\mathsf{fred}) & \rightarrow & \mathit{likes}(\mathsf{X},\mathsf{fred}) & [0.4] \\ r_3: & \mathit{elephant}(\mathsf{clyde}) \wedge \mathit{keeper}(\mathsf{fred}) & \rightarrow & \mathit{likes}(\mathsf{clyde},\mathsf{fred}) & [0.7] \end{array}$$

expressing that with a probability of 0.6 elephants like their keepers (r_1), with a probability of 0.4 elephants like keeper Fred (r_2), and with probability 0.7 elephant Clyde likes keeper Fred (r_3). From the point of view of commonsense reasoning this knowledge base makes perfect sense: Rule r_1 expresses that in some given population, choosing randomly an elephant-keeper-pair, we would expect that the elephant likes the keeper with probability 0.6. However, keeper Fred and elephant Clyde are exceptional—mostly, elephants do not like Fred, but Clyde likes (even) Fred. Maybe Clyde is a particularly good-natured elephant, maybe he is as moody as Fred and likes only him. So, Clyde is definitely exceptional with respect to r_2, but maybe even with respect to r_1.

However, the example is ambiguous, and its formal interpretation via probabilistic constraints is intricate. Rule r_1 seems to express a belief an agent may hold about a population, while r_3 clearly expresses individual belief: Considering all situations (possible worlds) involving Clyde and Fred which are imaginable, in 70 % of them Clyde likes Fred. So, we might think of applying different techniques to r_1 and r_3, but r_2 obviously mixes the two types of knowledge, how should r_2 be dealt with?

In many approaches, e. g. in Bayesian Logic Programs and in Markov Logic Networks, the relational rules are grounded, and the probability is attached to each instance. For r_1, this means:

$$\mathit{elephant}(\mathsf{a}) \wedge \mathit{keeper}(\mathsf{b}) \quad \rightarrow \quad \mathit{likes}(\mathsf{a},\mathsf{b}) \quad [0.6] \qquad \text{for all } \mathsf{a},\mathsf{b} \in U.$$

Here U is a properly (or arbitrarily) chosen universe. Besides the question "*How should U be chosen?*", there are two other problems. First, grounding turns the relational statement r_1 into a collection of statements of the same type as r_3, i. e. statements about individual beliefs. The population aspect gets lost, more precisely: r_1 is no longer a statement describing a generic behaviour in a population of (possibly very) individualistic individuals, but is understood to be a statement on individuals which all behave the same. Secondly, naive grounding techniques make the knowledge base inconsistent, as then r_3 collides with the respective instances of r_1 and

r_2. So, grounding has to take further constraints into account, to return a consistent knowledge base (cf. [11]).

While most of the above mentioned formalisms for statistical relational learning or inductive logic programming like Markov Logic Networks and Bayesian Logic Programs extend traditional (propositional) graphical models for probabilistic knowledge representation like Markov Nets and Bayesian Networks, here, we employ probabilistic conditional logic [20, 32, 31, 26]. In (propositional) probabilistic conditional logic knowledge is captured using conditionals of the form $(\phi \mid \psi)[\alpha]$ with some formulas ϕ, ψ of a given propositional language and $\alpha \in [0, 1]$. A probabilistic conditional of this form partially describes an (unknown) probability distribution P^* by stating that $P^*(\phi \mid \psi) = \alpha$ holds. In contrast to Bayes nets probabilistic conditional logic does not demand to fully describe a probability distribution but only to state constraints on it. On the one hand this is of great advantage because normally the knowledge engineer cannot fully specify a probability distribution for the problem area at hand. For example, if one has to represent probabilistic information on the relationships between symptoms and diseases then (usually) one can specify the probability of a specific disease given that a specific symptom is present but not if the symptom is not present. Probabilistic conditional logic avoids such problems by allowing to only partially specify a probability distribution. On the other hand, an incomplete specification of the problem area may lead to inconclusive inferences because there may be multiple probability distributions that satisfy the specified knowledge. The naïve approach to reason in probabilistic conditional logic is to compute upper and lower bounds for specific queries by consulting every probability distribution that is a model of the given knowledge base. While this skeptical form of reasoning may be appropriate for some applications, usually the inferences of this approach tend to be too weak to be meaningful. As a credulous alternative, one can select a specific probability distribution from the models of the knowledge base and do reasoning by just using this probability distribution. A reasonable choice for such a model is the one probability distribution with maximum entropy [15, 27, 20]. This probability distribution satisfies several desirable properties for commonsense reasoning and is uniquely determined among the probability distributions that satisfy a given set of probabilistic conditionals, see [15, 27, 20] for the theoretical foundations.

In this paper, we will propose two approaches to giving formal semantics to relational probabilistic conditional knowledge bases that aim at catching properly the commonsense intuition and resolving ambiguities. This will prepare the grounds for relational probabilistic reasoning in general. We will focus on model-based inductive inference operators for each of these semantics, in order to improve inferences from knowledge bases which usually represent only partial knowledge. We will make explicit what reasonable inference in this extended framework of relational probabilistic logic means by setting up a set of postulates. Of course, all this should be clearly related to work on probabilistic reasoning in the propositional case. In particular, we expect our semantics to coincide with propositional approaches, if the knowledge base is ground.

Moreover, we will present a model-based inductive inference operator that is based on the principle of maximum entropy for each of the two semantics. The idea of application is quite simple and similar to the propositional case: Having defined the set of models of a relational probabilistic knowledge base (according to each of the

semantics), one chooses the unique probability distribution among these models that has maximal entropy, if possible, and therefore allows us to reason precisely (i. e. with precise probabilities, not based on intervals), but in a most cautious way (see [15, 20] for the theoretical foundations). Examples will illustrate in which respects these inference operators differ, but we will show that both inference operators comply with all postulates.

This paper continues and extends work begun in [35, 22], and is organized as follows. First, we formalize the syntactical details of a probabilistic first-order conditional logic, and propose two different semantics for it, the *averaging* and the *aggregating* semantics. Afterwards we discuss the problem of inductive inference in this logic by developing several desirable properties of rational inference operators. We continue by presenting model-based inference operators that employ the principle of maximum entropy in both semantical frameworks, giving rise to two different inference operators which are exemplified and evaluated by means of the previously stated properties. We conclude with a brief summary and some discussions on related and further work. All proofs of theoretical results can be found in the appendix.

2 Syntax and Semantics of First-Order Conditional Logic

In the following we give an extension of probabilistic conditional logic to the relational case similar to [12, 13]. We start by presenting the syntax of this logic and continue presentation with two novel semantics. Afterwards, we give some insights on the relationships of these semantics.

2.1 Basics of Syntax and Semantics

We consider only a fragment of a first-order language, so let Σ be a first-order signature consisting of a (finite) set of predicate symbols and without functions of arity greater zero. We generally assume that Σ contains a countably infinite pool of constant symbols U. A predicate declaration P/n with a natural number n means that P is a predicate of arity n. Let $\mathcal{L}_\Sigma$ be a first-order language over the signature Σ that is generated in the usual way using negation, conjunction, and disjunction, but without quantifiers. If appropriate we abbreviate conjunctions $A \wedge B$ by AB. We denote constants with a beginning lowercase, variables with a beginning uppercase letter, and vectors of these with $\vec{\mathsf{a}}$ and $\vec{\mathsf{X}}$, respectively.

A formula that contains no variable is called *ground*. Let $\mathsf{ground}_C(A)$ denote the set of ground instances of A with respect to a set of constants $C \subseteq U$, e. g. $\mathsf{ground}_{\{\mathsf{a},\mathsf{b}\}}(A(\mathsf{X},\mathsf{Y}))$ is $\{A(\mathsf{a},\mathsf{a}), A(\mathsf{a},\mathsf{b}), A(\mathsf{b},\mathsf{a}), A(\mathsf{b},\mathsf{b})\}$.

Definition 2.1 (Probabilistic Conditional)
Let $A, B \in \mathcal{L}_\Sigma$ be formulas (not necessarily ground) that mention only a finite number of constants and predicates. An expression of the form $(B \mid A)[\alpha]$ with a real number $\alpha \in [0, 1]$ is called a *probabilistic conditional*. A probabilistic conditional $(B \mid A)[\alpha]$ is *ground* if both A and B are ground. Let $(\mathcal{L}_\Sigma \mid \mathcal{L}_\Sigma)^{prob}$ be the set of all probabilistic conditionals over $\mathcal{L}_\Sigma$.

Open conditionals, i. e. conditionals that contain variables, are meant to range over all possible constants in the language but are not understood to stand as schemas for

their instantiations. Rather, an open conditional describes a general or default rule. The problem of giving open conditionals an intuitive and formal interpretation is part of the topic of this paper and will be discussed in more depth below.

If a conditional $r = (B \mid A)[\alpha]$ contains free variables we also use the notation $r = (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$ where $\vec{\mathsf{X}} = (\mathsf{X}_1, \ldots, \mathsf{X}_n)$ contains all free variables in $A \wedge B$. If $\vec{\mathsf{a}}$ is a vector of the same length as $\vec{\mathsf{X}}$ then $(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}}))[\alpha]$ is meant to denote the instantiation of r with $\vec{\mathsf{a}}$. If $\vec{X}$ or $\vec{a}$ contains variables or constants not mentioned in A and B then those are ignored. For example, if $r = (B(\mathsf{X}_1) \mid A(\mathsf{X}_1, \mathsf{X}_2))[\alpha]$ we also write $r = (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$ with $\vec{\mathsf{X}} = (\mathsf{X}_1, \mathsf{X}_2)$, and if $\vec{\mathsf{a}} = (\mathsf{a}_1, \mathsf{a}_2)$ then $(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}}))[\alpha] = (B(\mathsf{a}_1) \mid A(\mathsf{a}_1, \mathsf{a}_2))[\alpha]$.

If the premise A of a conditional $(B \mid A)[\alpha]$ is ground and tautological, i. e. $A \equiv \top$, we abbreviate $(B \mid \top)[\alpha]$ by $(B)[\alpha]$. A conditional of the form $(B)[\alpha]$ is also called a *probabilistic fact*. Let $\mathsf{ground}_C((B \mid A)[\alpha])$ denote the set of all grounded probabilistic conditionals of a conditional $(B \mid A)[\alpha]$ with respect to a set of constants $C \subseteq U$.

Definition 2.2 (Knowledge base)
A finite set $\mathcal{R}$ of probabilistic conditionals is called a *knowledge base*. A knowledge base $\mathcal{R}$ is *ground* if every probabilistic conditional in $\mathcal{R}$ is ground. Let $\mathfrak{R}$ denote the set of knowledge bases.

For a formula $A \in \mathcal{L}_\Sigma$ let $\mathsf{consts}(A) \subseteq U$ denote the set of constants appearing in A. Similarly, let $\mathsf{consts}(r)$ and $\mathsf{consts}(\mathcal{R})$ for a probabilistic conditional r and a knowledge base $\mathcal{R}$ be defined accordingly. Note that due to the finiteness of formulas inside of conditionals and the finiteness of knowledge bases it follows that $\mathsf{consts}(\mathcal{R})$ is finite for a knowledge base $\mathcal{R}$.

Remark 2.3
Bear in mind that a ground knowledge base $\mathcal{R}$ is equivalent to a propositional knowledge base $\mathcal{R}'$ by interpreting ground atoms in $\mathcal{R}$ as ordinary propositional atoms. For the rest of this paper we treat ground relational knowledge bases and propositional knowledge bases interchangeably.

We need some further notation to go on. For a formula A let $A[\mathsf{d}/\mathsf{c}]$ denote the formula that is the same as A except that every occurrence of the term c (either a variable or a constant) is substituted with the term d. More generally, let $A[\mathsf{d}_1/\mathsf{c}_1, \ldots, \mathsf{d}_n/\mathsf{c}_n]$ denote the formula that is the same as A except that every occurrence of c_i is substituted with d_i for $1 \leq i \leq n$ simultaneously. Furthermore, let $A[\mathsf{c} \leftrightarrow \mathsf{d}]$ be an abbreviation for $A[\mathsf{c}/\mathsf{d}, \mathsf{d}/\mathsf{c}]$. The substitution operator $[\cdot]$ is extended on sets of formulas, conditionals, and knowledge bases in the usual way.

Introducing relational aspects in probabilistic statements raises some ambiguity on the understanding of these statements. We illustrate this problem on the example mentioned in the introduction (cf. also [9]).

Example 2.4
Consider the knowledge base $\mathcal{R}_{zoo} = \{r_1, r_2, r_3\}$ with

$$
\begin{aligned}
r_1 &: (\mathit{likes}(\mathsf{X}, \mathsf{Y}) \mid \mathit{elephant}(\mathsf{X}) \wedge \mathit{keeper}(\mathsf{Y}))[0.6] \\
r_2 &: (\mathit{likes}(\mathsf{X}, \mathsf{fred}) \mid \mathit{elephant}(\mathsf{X}) \wedge \mathit{keeper}(\mathsf{fred}))[0.4] \\
r_3 &: (\mathit{likes}(\mathsf{clyde}, \mathsf{fred}) \mid \mathit{elephant}(\mathsf{clyde}) \wedge \mathit{keeper}(\mathsf{fred}))[0.7]
\end{aligned}
$$

The knowledge base $\mathcal{R}$ describes the relationships between keepers and elephants in a zoo, thereby stating both subjective degrees of belief on the relationship between Clyde and Fred (r_3), as well as population-based probabilities that involve *all* elephants and keepers (r_1, r_2). So, r_3 should be interpreted via a possible worlds semantics, whereas r_1, r_2 seem to describe a typical behavior within a population that might have been obtained by statistical means (cf. e. g. [2]). In this paper, we propose thoroughly subjective approaches to probability even for the relational case, using a possible worlds semantics for all three statements above. In contrast to other approaches we do not interpret conditionals like r_1 and r_2 as schemas and reason using their ground instances. This allows an intuitive and coherent interpretation of relational probabilistic statements that takes into account both information on specific objects and information on a population.

Formal semantics for first-order probabilistic conditional logic will be given by probability distributions that are defined over possible worlds of the given first-order language $\mathcal{L}_\Sigma$. Here, we use Herbrand interpretations for possible worlds. The *Herbrand base* $\mathcal{H}$ is the set of all ground atoms that can be built using the predicate symbols and constants in U, and a *Herbrand interpretation* is a (finite or infinite) subset of $\mathcal{H}$. For a Herbrand interpretation ω let $\mathsf{consts}(\omega) \subseteq U$ denote the set of constants appearing in ω. A Herbrand interpretation ω satisfies a ground atom A, denoted by $\omega \models A$, iff $A \in \omega$. The satisfaction relation $\models$ is extended to arbitrary ground formulas in the usual way. Let Ω denote the set of all Herbrand interpretations and let $P : \Omega \to [0, 1]$ be a probability distribution over Ω that satisfies

1. for all $\omega \in \Omega$ it holds that $P(\omega) \geq 0$,
2. $P(\omega) \neq 0$ only for finitely many $\omega \in \Omega$,
3. if $\omega \in \Omega$ is infinite then $P(\omega) = 0$, and
4. $\sum_{\omega \in \Omega} P(\omega) = 1$.

We require probability distributions to satisfy the properties 2 and 3 in order to avoid technical difficulties in handling infinite sums. As we only consider finite knowledge bases, i. e., we consider only finite excerpts of Ω, these demands are of no concern regarding the expressivity of our logic. Let Prob be the set of all probability distributions that satisfy 1.) – 4.). We can extend $P \in \mathsf{Prob}$ on ground formulas A by setting

$$P(A) = \sum_{\omega \models A} P(\omega) \quad .$$

For the propositional case [15, 20] satisfaction of a conditional is defined via conditional probabilities. Let $(B \,|\, A)[\alpha]$ be a ground conditional. Then a probability distribution P satisfies $(B \,|\, A)[\alpha]$, denoted by $P \models (B \,|\, A)[\alpha]$, if the following condition holds

$$P \models (B \,|\, A)[\alpha] \;\text{ iff }\; P(B \,|\, A) = \frac{P(B \wedge A)}{P(A)} = \alpha \text{ and } P(A) > 0. \tag{2.1}$$

It remains to define a satisfaction relation for conditionals with variables (see Example 2.4). Taking a naïve approach by grounding all conditionals in $\mathcal{R}$ universally and

taking this grounding $\mathcal{R}'$ as a propositional knowledge base, we can (usually) not determine any probability distribution that satisfies $\mathcal{R}'$ due to its inherent inconsistency [13, 11].

In the following, we propose two different approaches for semantics of $(\mathcal{L}_\Sigma \mid \mathcal{L}_\Sigma)^{prob}$ that coincide with (2.1) in the propositional case but differ on the interpretation of population-based statements.

2.2 Averaging Semantics

Example 2.4 showed that, in general, universal instantiation of an open conditional $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$ does not yield an equivalent representation of the intended meaning of $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$. It demands that every instantiation *inherits* the probability α which is not adequate in the context of non-monotonic reasoning. Moreover, having more specific information on specific instantiations should not render the knowledge base inconsistent as other instantiations might balance out exceptions. Consider the following example from statistics.

Example 2.5
Imagine an urn with 10 balls, 9 of them are blue and one is red. Using sampling without replacement our first draw is a blue ball. What is the probability of drawing a blue ball in the second draw? Let c_i be the ball that is drawn on i-th turn. We represent the scenario after the first draw using a static world view, i. e., we suppose that the probability of drawing a blue ball from the urn is (a priori) 0.9 but we know that c_1 is a blue ball for sure. Let $\mathcal{R}_{urn} = \{r_{1,1}, \ldots, r_{1,10}, r_2, r_3\}$ be given via

$$\begin{array}{rcl} r_{1,1} & : & (ball(\mathsf{c}_1))[1] \quad \ldots \quad r_{1,10} : (ball(\mathsf{c}_{10}))[1] \\ r_2 & : & (blue(\mathsf{X}) \mid ball(\mathsf{X}))[0.9] \\ r_3 & : & (blue(\mathsf{c}_1))[1] \end{array}$$

The knowledge base $\mathcal{R}_1$ expresses the agent's beliefs after the first draw. Clearly, the probability of drawing a blue ball on the second draw is $8/9$ and a reasonable semantics should allow both $\mathcal{R}_{urn}$ and $\mathcal{R}_{urn} \cup \{(blue(\mathsf{c}_2))[8/9]\}$ to be satisfiable.

In the previous example conditional r_2 defines an *expected value* for the probability of drawing a blue ball in any turn, provided that we have no knowledge on the color of already drawn balls and the remaining balls in the urn. The additional information that a blue ball has already been drawn changes the expected value of drawing another blue ball correspondingly. Therefore mutual influences of different conditionals have to be taken into account when giving meaning to a knowledge base.

The approach of *averaging semantics* generalizes the above intuition by interpreting open conditionals of the form $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$ to describe an expected value on the probability of $(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}}))$ for some randomly chosen $\vec{\mathsf{a}}$ in some given adequately large universe. Thus, given the actual probabilities of $(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}}))$ for each possible instantiation $\vec{\mathsf{a}}$ we expect the *average* of these probabilities to match α. In order to be able to investigate the influence of the universe (or respectively, its size) on the probabilistic evaluation of statements we parametrize the evaluation by a set of constants $D \subseteq U$ that represent the individuals actually under consideration, similar to the notion of *active domains* in database theory [1]. Hence, a probability distribution $P : \Omega \to [0, 1]$ *$\varnothing$-satisfies* $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$ under a set $D \subseteq U$, denoted by

ω	$P(\omega)$	ω	$P(\omega)$
$\emptyset$	0	$\{A(\mathsf{c}_2), B(\mathsf{c}_1)\}$	0
$\{A(\mathsf{c}_1)\}$	0	$\{A(\mathsf{c}_2), B(\mathsf{c}_2)\}$	0
$\{A(\mathsf{c}_2)\}$	0	$\{B(\mathsf{c}_1), B(\mathsf{c}_2)\}$	0
$\{B(\mathsf{c}_1)\}$	0	$\{A(\mathsf{c}_1), A(\mathsf{c}_2), B(\mathsf{c}_1)\}$	0.1
$\{B(\mathsf{c}_2)\}$	0	$\{A(\mathsf{c}_1), A(\mathsf{c}_2), B(\mathsf{c}_2)\}$	0.3
$\{A(\mathsf{c}_1), A(\mathsf{c}_2)\}$	0.1	$\{A(\mathsf{c}_1), B(\mathsf{c}_2), B(\mathsf{c}_2)\}$	0
$\{A(\mathsf{c}_1), B(\mathsf{c}_1)\}$	0	$\{A(\mathsf{c}_2), B(\mathsf{c}_1), B(\mathsf{c}_2)\}$	0
$\{A(\mathsf{c}_1), B(\mathsf{c}_2)\}$	0	$\{A(\mathsf{c}_1), A(\mathsf{c}_2), B(\mathsf{c}_1), B(\mathsf{c}_2)\}$	0.5

TABLE 1: A sample probability distribution for the knowledge base $\mathcal{R}$ in Example 2.7 (we assign probability 0 to all interpretations that do not appear).

$P, D \models_{\varnothing} (B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha]$ if and only if $P(\omega) = 0$ for every $\omega \in \Omega$ with $\mathsf{consts}(\omega) \not\subseteq D$ and

$$\frac{\sum_{(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}})) \in \mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))} P(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}}))}{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))|} = \alpha\,. \tag{2.2}$$

In Equation (2.2), the denominator of the fraction on the left-side sums the conditional probabilities of the different instantiations of $(B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha]$ while the numerator equals the number of these instantiations. Intuitively, a probability distribution P $\varnothing$-satisfies a conditional $(B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha]$ if the average of the individual instantiations of $(B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha]$ is α (with respect to $D \subseteq U$).

REMARK 2.6
For a ground conditional $(G_2 \,|\, G_1)[\alpha]$ the operator $\models_{\varnothing}$ indeed coincides with the propositional case due to $\mathsf{ground}_D(G_2 \,|\, G_1) = \{(G_2 \,|\, G_1)\}$ for every $D \subseteq U$.

As usual, a probability distribution P $\varnothing$-satisfies a knowledge base $\mathcal{R}$ under $D \subseteq U$, denoted $P, D \models_{\varnothing} \mathcal{R}$, if P $\varnothing$-satisfies every probabilistic conditional $r \in \mathcal{R}$ under D. We say that $\mathcal{R}$ is *$\varnothing$-consistent* under D iff there is at least one P with $P, D \models_{\varnothing} \mathcal{R}$, otherwise $\mathcal{R}$ is *$\varnothing$-inconsistent* under D. Let $Mod^D_{\varnothing}(\mathcal{R})$ denote the set of models of $\mathcal{R}$ under D, i. e., it holds that $Mod^D_{\varnothing}(\mathcal{R}) = \{P \in \mathsf{Prob} \mid P, D \models_{\varnothing} \mathcal{R}\}$.

EXAMPLE 2.7
Consider a knowledge base $\mathcal{R} = \{r_1, r_2, r_3\}$ with

$$\begin{aligned} r_1 &: (B(\mathsf{X}) \,|\, A(\mathsf{X}))[0.7] \\ r_2 &: (A(\mathsf{X}))[1] \\ r_3 &: (B(\mathsf{c}_1))[0.6] \end{aligned}$$

and $D = \{\mathsf{c}_1, \mathsf{c}_2\}$. Consider the probability distribution P given in Table 1. Notice that the given interpretations are given in the compact form described above. As one can see, it holds $P, D \models_{\varnothing} \mathcal{R}$:

- It holds $P, D \models_{\varnothing} r_1$ as

$$P(B(\mathsf{c}_1) \mid A(\mathsf{c}_1)) = P(B(\mathsf{c}_1)A(\mathsf{c}_1))/P(A(\mathsf{c}_1)) = 0.6/1 = 0.6$$

and

$$P(B(\mathsf{c}_2) \mid A(\mathsf{c}_2)) = P(B(\mathsf{c}_2)A(\mathsf{c}_2))/P(A(\mathsf{c}_2)) = 0.8/1 = 0.8$$

and hence $(0.8 + 0.6)/2 = 0.7$;

- it holds $P, D \models_{\varnothing} r_2$ as $(P(A(\mathsf{c}_1)) + P(A(\mathsf{c}_2)))/2 = 1$, and
- it holds $P, D \models_{\varnothing} r_3$ as $P(B(\mathsf{c}_1)) = 0.6$.

2.3 Aggregating Semantics

Averaging semantics preserves the interpretation of a conditional probability as subjective belief in the conclusion given the premise holds. Therefore conditional probabilities are only defined for ground conditionals and the probability of an open conditional $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))$ is defined in terms of conditional probabilities of its instances. When considering a relational language one might argue whether a conditional should be interpreted in this manner or whether conditional probability should be redefined on a higher level incorporating the relational structure of the language. In the following we give a novel approach for defining conditional probabilities in a relational setting.

Considering again Example 2.5 statistics usually do not involve conditional probabilities in the sense of subjective beliefs. So we first consider unconditioned formulas $(A(\vec{\mathsf{X}}))[\alpha]$ with free variables $\vec{\mathsf{X}}$. Let ω be some Herbrand interpretation and $D \subseteq U$ the set of individuals under consideration. Treating ω as a statistical sample we can count the number of true instances of $a(\vec{\mathsf{X}})$ in ω under D and determine the average number of true instances via

$$f_\omega^D(A(\vec{\mathsf{X}})) = \frac{|\{A(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}}) \in \mathsf{ground}_D(A(\vec{\mathsf{X}})) \wedge \omega \models A(\vec{\mathsf{a}})\}|}{|\mathsf{ground}_D(A(\vec{\mathsf{X}}))|} \quad . \tag{2.3}$$

For ground $A(\vec{\mathsf{a}})$ we write

$$f_\omega^D(A(\vec{\mathsf{a}})) = \begin{cases} 1 & \text{if } \omega \models A(\vec{\mathsf{a}}) \\ 0 & \text{if } \omega \not\models A(\vec{\mathsf{a}}) \end{cases}$$

then (2.3) amounts to

$$f_\omega^D(A(\vec{\mathsf{X}})) = \frac{\sum_{A(\vec{\mathsf{a}}) \in \mathsf{ground}_D(A(\vec{\mathsf{X}}))} f_\omega^D(A(\vec{\mathsf{a}}))}{|\mathsf{ground}_D(A(\vec{\mathsf{X}}))|} \quad .$$

Considering some probability distribution P (thus describing either a series of samples, or subjective beliefs in each interpretation being the actual world) we can appropriately define

$$P(A(\vec{\mathsf{X}}); D) =_{\text{def}} \sum_{\omega \in \Omega} f_\omega^D(A(\vec{\mathsf{X}}))P(\omega) \tag{2.4}$$

to be the weighted sum of the average frequencies. Rearranging (2.4) yields

$$\begin{aligned} P(A(\vec{\mathsf{X}}); D) &= \sum_{\omega \in \Omega} f_\omega^D(A(\vec{\mathsf{X}}))P(\omega) \\ &= \sum_{\omega \in \Omega} \frac{\sum_{A(\vec{\mathsf{a}}) \in \mathsf{ground}_D(A(\vec{\mathsf{X}}))} f_\omega^D(A(\vec{\mathsf{a}}))}{|\mathsf{ground}_D(A(\vec{\mathsf{X}}))|} P(\omega) \\ &= \frac{\sum_{\omega \in \Omega} \sum_{A(\vec{\mathsf{a}}) \in \mathsf{ground}_D(A(\vec{\mathsf{X}}))} f_\omega^D(A(\vec{\mathsf{a}}))P(\omega)}{|\mathsf{ground}_D(A(\vec{\mathsf{X}}))|} \\ &= \frac{\sum_{A(\vec{\mathsf{a}}) \in \mathsf{ground}_D(A(\vec{\mathsf{X}}))} \sum_{\omega \in \Omega} f_\omega^D(A(\vec{\mathsf{a}}))P(\omega)}{|\mathsf{ground}_D(A(\vec{\mathsf{X}}))|} \\ &= \frac{\sum_{A(\vec{\mathsf{a}}) \in \mathsf{ground}_D(A(\vec{\mathsf{X}}))} P(A(\vec{\mathsf{a}}))}{|\mathsf{ground}_D(A(\vec{\mathsf{X}}))|} \end{aligned} \tag{2.5}$$

and thus also a statistical justification for (2.2) for the case of unconditioned formulas and averaging semantics. But instead of applying (2.5) in the same way to conditionals we give a new definition of the conditional probability via

$$P(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}); D) = \frac{P(B(\vec{\mathsf{X}})A(\vec{\mathsf{X}}); D)}{P(A(\vec{\mathsf{X}}); D)}$$

thus carrying over the definition of conditional probability to a relational setting. Accordingly we define the *aggregating semantics* as follows. A probability distribution $P : \Omega \to [0, 1]$ $\odot$-*satisfies* $(B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha]$ under $D \subseteq U$, denoted by $P, D \models_\odot (B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha]$ if and only if $P(\omega) = 0$ for every $\omega \in \Omega$ with $\mathsf{consts}(\omega) \not\subseteq D$, $P(A(\vec{\mathsf{X}}); D) > 0$, and $P(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}); D) = \alpha$. Note, that this definition nicely resembles the satisfaction relation in the propositional case.

Remark 2.8
As for $\models_\varnothing$, for a ground conditional $(B_2 \,|\, A_1)[\alpha]$ the operator $\models_\odot$ coincides with the propositional case due to $\mathsf{ground}_D(B_2 \,|\, A_1) = \{(B_2 \,|\, A_1)\}$ for every $D \subseteq U$.

As above, a probability distribution P $\odot$-satisfies a knowledge base $\mathcal{R}$ under D, denoted $P, D \models_\odot \mathcal{R}$, if P $\odot$-satisfies every probabilistic conditional $r \in \mathcal{R}$ under D. We say that $\mathcal{R}$ is $\odot$-*consistent* under D iff there is at least one P with $P, D \models_\odot \mathcal{R}$, otherwise $\mathcal{R}$ is $\odot$-*inconsistent* under D. Let $Mod_\odot^D(\mathcal{R})$ denote the set of models of $\mathcal{R}$ under D, i. e., it holds that $Mod_\odot^D(\mathcal{R}) = \{P \in \mathsf{Prob} \mid P, D \models_\odot \mathcal{R}\}$.

Example 2.9
Consider the knowledge base $\mathcal{R} = \{r_1, r_2, r_3\}$ from Example 2.7 with

$$\begin{aligned} r_1 &: (B(\mathsf{X}) \,|\, A(\mathsf{X}))[0.7] \\ r_2 &: (A(\mathsf{X}))[1] \\ r_3 &: (B(\mathsf{c}_1))[0.6] \end{aligned}$$

and $D = \{\mathsf{c}_1, \mathsf{c}_2\}$. As in Example 2.7 consider the probability distribution P given in Table 1 which is represented in the same compact fashion as in Example 2.7. As one can see, it holds $P, D \models_\odot \mathcal{R}$ as well:

- It holds $P, D \models_{\odot} r_1$ due to $P(B(\mathsf{c}_1)A(\mathsf{c}_1)) = 0.6$ and $P(B(\mathsf{c}_2)A(\mathsf{c}_2))) = 0.8$, and as well $P(A(\mathsf{c}_1)) = P(A(\mathsf{c}_2)) = 1$; hence $(0.8 + 0.6)/2 = 0.7$;
- it holds $P, D \models_{\odot} r_2$ as $(P(A(\mathsf{c}_1)) + P(A(\mathsf{c}_2)))/2 = 1$ (remark that $P(\top) = 1$), and
- it holds $P, D \models_{\odot} r_3$ as $P(B(\mathsf{c}_1)) = 0.6$.

Examples 2.7 and 2.9 show that both proposed semantics coincide on the given simple knowledge base. We will investigate the similarities and differences of the both semantics further in the next subsection.

2.4 Comparing the Semantics

Both $\models_{\varnothing}$ and $\models_{\odot}$ introduce a population-based, but non-statistical perspective into the interpretation of relational conditionals since they make use of information about individuals and subjective probabilities. Due to remarks 2.6 and 2.8, both semantics agree on ground conditionals. Furthermore, it is straightforward to show that $\models_{\varnothing}$ and $\models_{\odot}$ also agree on probabilistic facts (that may contain variables).

Proposition 2.10
Let $P \in \mathsf{Prob}$ be a probability distribution, $D \subseteq U$, and $(B)[\alpha] \in (\mathcal{L}_\Sigma \,|\, \mathcal{L}_\Sigma)^{prob}$ a probabilistic fact. Then it holds that $P, D \models_{\varnothing} (B)[\alpha]$ iff $P, D \models_{\odot} (B)[\alpha]$.

For general conditionals in $(\mathcal{L}_\Sigma \,|\, \mathcal{L}_\Sigma)^{prob}$, however, the two semantics turn out to be different, as the following example shows.

Example 2.11
Let $A/1$ and $B/1$ be two predicates, let $D = \{\mathsf{a}_1, \ldots, \mathsf{a}_5\}$ be a set of constants, and consider the following (ground) knowledge base $\mathcal{R}$:

$$\begin{array}{rr}
(A(\mathsf{a}_1))[0.5] & (A(\mathsf{a}_2))[0.1] \\
(A(\mathsf{a}_3))[0.9] & (A(\mathsf{a}_4))[0.6] \\
(A(\mathsf{a}_5))[0.4] & (B(\mathsf{a}_1)A(\mathsf{a}_1))[0.5] \\
(B(\mathsf{a}_2)A(\mathsf{a}_2))[0.1] & (B(\mathsf{a}_3)A(\mathsf{a}_3))[0.9] \\
(B(\mathsf{a}_4)A(\mathsf{a}_4))[0.4] & (B(\mathsf{a}_5)A(\mathsf{a}_5))[0.1]
\end{array}$$

In addition, consider the conditional $r = (B(\mathsf{X}) \,|\, A(\mathsf{X}))[0.8]$. On the one hand, any probability distribution P with $P, D \models_{\odot} \mathcal{R}$ also obeys $P, D \models_{\odot} r$ as

$$\frac{P(B(\mathsf{a}_1)A(\mathsf{a}_1)) + \cdots + P(B(\mathsf{a}_5)A(\mathsf{a}_5))}{P(A(\mathsf{a}_1)) + \cdots + P(A(\mathsf{a}_5))} \quad = \quad \frac{0.5 + 0.1 + 0.9 + 0.4 + 0.1}{0.5 + 0.1 + 0.9 + 0.6 + 0.4} = 0.8$$

On the other hand, every probability distribution P with $P, D \models_{\varnothing} \mathcal{R}$ does not obey $P, D \models_{\varnothing} r$ due to

$$\begin{aligned}
&1/5 \; (P(B(\mathsf{a}_1) \,|\, A(\mathsf{a}_1)) + \cdots + P(B(\mathsf{a}_5) \,|\, A(\mathsf{a}_5))) \\
&\quad = \quad \left(\frac{0.5}{0.5} + \frac{0.1}{0.1} + \frac{0.9}{0.9} + \frac{0.4}{0.6} + \frac{0.1}{0.4}\right)/5 \\
&\quad = \quad 0.78\overline{3} \neq 0.8
\end{aligned}$$

As $P, D \models_{\odot} \mathcal{R}$ is equivalent to $P, D \models_{\varnothing} \mathcal{R}$ due to Proposition 2.10 the different semantics may lead to different inferences. Furthermore, the two semantics feature a

different notion of consistency as $\mathcal{R} \cup \{r\}$ is $\varnothing$-inconsistent under D but $\odot$-consistent under D.

Although the previous example suggests that the difference of the two proposed semantics is marginal, in the following, we show that the difference can be made arbitrarily large.

LEMMA 2.12
Let $n \in \mathbb{N}^+$ be some positive integer and let $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n \in (0, 1]$ with $\alpha_i \leq \beta_i$ for all $i = 1, \ldots, n$. Then

$$\left| \frac{\alpha_1/\beta_1 + \ldots + \alpha_n/\beta_n}{n} - \frac{\alpha_1 + \ldots + \alpha_n}{\beta_1 + \ldots + \beta_n} \right| \quad < \quad \frac{n-1}{n} \tag{2.6}$$

The bound of $(n-1)/n$ is also the least upper bound as the following example shows.

EXAMPLE 2.13
Let $n \in \mathbb{N}^+$ be some positive integer and let $x \geq 2$ be some positive real value. Define $\alpha_1 = \ldots = \alpha_n = \beta_1 = \ldots = \beta_{n-1} = 1/x$ and $\beta_n = 1 - 1/x$. Observe, that for any $x \geq 2$ it holds that $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n \in (0, 1]$ and $\alpha_i \leq \beta_i$ for any $i = 1, \ldots, n$. Then it holds

$$\begin{aligned} \frac{\alpha_1/\beta_1 + \ldots + \alpha_n/\beta_n}{n} \quad &= \quad \frac{n - 1 + \frac{1}{x}/(1 - \frac{1}{x})}{n} \\ &\overset{x \to \infty}{\to} \quad (n-1)/n \end{aligned}$$

and

$$\begin{aligned} \frac{\alpha_1 + \ldots + \alpha_n}{\beta_1 + \ldots + \beta_n} \quad &= \quad \frac{n/x}{(n-1)/x + 1 - 1/x} \\ &\overset{x \to \infty}{\to} \quad 0 \quad . \end{aligned}$$

Lemma 2.12 can be used to prove the following property on the relationship of averaging and aggregating semantics.

COROLLARY 2.14
Let P be some probability distribution, $D \subseteq U$, and $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))$ be some conditional. If $P, D \models_{\varnothing} (B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha_1]$ and $P, D \models_{\odot} (B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha_2]$ then

$$|\alpha_1 - \alpha_2| < \frac{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))| - 1}{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))|}$$

Example 2.13 can be directly used to construct a worst-case example such that aggregating and averaging semantics differ to an arbitrarily large degree.

With the semantics $\models_{\varnothing}$ and $\models_{\odot}$ we gain novel model theories for relational probabilistic conditional logic. There may be other alternative model theories but in the following we will focus on these two.

3 Inductive Inference in First-Order Conditional Logic

In the following, we are interested in inductive inference for first-order conditional logic, i. e., in finding a "good" probability distribution P that satisfies all probabilistic

conditionals of a given knowledge base $\mathcal{R}$ given one of the two proposed semantics. More specifically, we are interested in an operator $\mathcal{I}(\mathcal{R}, D)$ that takes a knowledge base $\mathcal{R}$ and a set of constants D as input and returns a probability distribution $P = \mathcal{I}(\mathcal{R}, D) \in \mathsf{Prob}$ as output such that P describes $\mathcal{R}$ "best" in a commonsensical manner. In particular, the resulting distribution should be a model of $\mathcal{R}$ under D and therefore $\mathcal{I}$ should realize a model-based inductive reasoning process in the spirit of [27]. So, let $\models_\circ$ be any entailment relation between distributions from Ω and relational probabilistic conditionals from $(\mathcal{L}_\Sigma \mid \mathcal{L}_\Sigma)^{prob}$. In this section, we state some properties that a reasonable model-based inference operator should observe. In the following section, we will present two operators that comply with all postulates.

In order to ease notation and presentation, we will implicitly assume that $\mathcal{R}$ is defined over a language $\mathcal{L}_\Sigma$ the predicate symbols of which are held fixed, and the set D of constants is to contain all constants appearing in $\mathcal{R}$.

Our first demand for an operator $\mathcal{I}$ is its well-definedness. As an inconsistent knowledge base $\mathcal{R}$ has no models and therefore an operator $\mathcal{I}$ cannot determine any model of $\mathcal{R}$ for reasoning, let undef be a new symbol for this case. Let $\mathcal{I}$ be an operator that maps a knowledge base $\mathcal{R}$ and a set $D \subseteq U$ of constants onto a probability distribution $\mathcal{I}(\mathcal{R}, D) \in \mathsf{Prob}$ or to undef.

(Well-Definedness) It holds that $\mathcal{I}(\mathcal{R}, D) \in Mod_\circ^D(\mathcal{R})$ iff $\mathsf{consts}(\mathcal{R}) \subseteq D$ and $\mathcal{R} \subseteq \mathcal{L}_\Sigma$ is $\circ$-consistent under D.

When considering knowledge bases based on a relational language the beliefs one obtains for specific individuals are of special interest. An important demand to be made is that for indistinguishable individuals, the same information should be obtained. Here, indistinguishability is defined with respect to the information expressed by $\mathcal{R}$. More specifically, if the explicit information encoded in $\mathcal{R}$ for two different individuals $\mathsf{c}_1, \mathsf{c}_2 \in D$ is the same, the probability distribution $P = \mathcal{I}(\mathcal{R}, D)$ should treat them as indistinguishable. We formalize this indistinguishability by introducing an equivalence relation on constants.

Definition 3.1 ($\mathcal{R}$-Equivalence)
Let $\mathcal{R}$ be a knowledge base. The constants $\mathsf{c}_1, \mathsf{c}_2 \in U$ are *$\mathcal{R}$-equivalent*, denoted by $\mathsf{c}_1 \equiv_\mathcal{R} \mathsf{c}_2$, iff $\mathcal{R} = \mathcal{R}[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2]$.

Observe that $\equiv_\mathcal{R}$ is indeed an equivalence relation, i. e., it is reflexive, transitive, and symmetric. The equivalence classes of $\equiv_\mathcal{R}$ are called *$\mathcal{R}$-equivalence classes* and the set of all $\mathcal{R}$-equivalence classes is denoted by $\mathcal{S}_\mathcal{R}$. Note, that the notion of $\mathcal{R}$-equivalence bears a resemblance with the notion of *reference classes* [2] but on a pure syntactical level.

Example 3.2
Consider the knowledge base $\mathcal{R} = \{r_1, r_2, r_3, r_4\}$ given via

$$\begin{aligned} r_1 &: (B(\mathsf{X}) \mid A(\mathsf{X}))[0.7] \\ r_2 &: (A(\mathsf{c}_1))[0.6] \\ r_3 &: (A(\mathsf{c}_2))[0.6] \\ r_4 &: (A(\mathsf{c}_3))[0.1] \end{aligned}$$

In $\mathcal{R}$ we find that $c_1 \equiv_\mathcal{R} c_2$ but $c_1 \not\equiv_\mathcal{R} c_3$ and $c_2 \not\equiv_\mathcal{R} c_3$. Notice also that $d \equiv_\mathcal{R} d'$ for every $d, d' \in U \setminus \{c_1, c_2, c_3\}$.

Using $\mathcal{R}$-equivalence we can state our demand for equal treatment of indistinguishable individuals as follows.

(Prototypical Indifference) Let $\mathcal{R}$ be a knowledge base on $\mathcal{L}_\Sigma$ and A a ground sentence. For any $\mathsf{c}_1, \mathsf{c}_2 \in D$ with $\mathsf{c}_1 \equiv_\mathcal{R} \mathsf{c}_2$ it holds that $\mathcal{I}(\mathcal{R}, D)(A) = \mathcal{I}(\mathcal{R}, D)(A[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2])$.

The above property states that given two $\mathcal{R}$-equivalent constants $\mathsf{c}_1, \mathsf{c}_2$ a sentence A should have the same inferred probability as the sentence $A[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2]$ which results in replacing c_1 with c_2 and vice versa. For example, we expect $B(\mathsf{c}_1, \mathsf{c}_2)$ to have the same probability as $B(\mathsf{c}_2, \mathsf{c}_1)$ but also $C(\mathsf{c}_1)$ to have the same probability as $C(\mathsf{c}_2)$.

A similar notion like *Prototypical Indifference* but in a slightly different context can be found under the term *involution invariance* in [13]. There, syntactic indistinguishability between instances of conditionals is exploited in order to reduce the size of the maximum entropy model. It is in the line of current research to investigate whether the approach pursued in [13] may be adapted for other inference operators that satisfy *Prototypical Indifference*.

An even more basic demand than *Prototypical Indifference* is that renaming an individual should have no impact on the information that can be derived for it.

(Name Irrelevance) Let $\mathcal{R}$ be a knowledge base, $\mathsf{d} \in U \setminus D$ a constant not appearing in D, and $A \in \mathcal{L}_\Sigma$ a ground sentence. For every $\mathsf{c} \in D$ it holds that

$$\mathcal{I}(\mathcal{R}, D)(A) = \mathcal{I}(\mathcal{R}[\mathsf{d}/\mathsf{c}], (D \cup \{\mathsf{d}\}) \setminus \{\mathsf{c}\})(A[\mathsf{d}/\mathsf{c}])$$

This property simply states that renaming a constant c in $\mathcal{R}$ to d—thus removing c from the underlying set D but adding d—yields the same inferences. Although (Name Irrelevance) seems to be the weaker demand, surprisingly, every function $\mathcal{I}$ satisfying *Name Irrelevance* also satisfies *Prototypical Indifference*.

PROPOSITION 3.3
If $\mathcal{I}$ satisfies *Name Irrelevance* then $\mathcal{I}$ satisfies *Prototypical Indifference*.

From *Prototypical Indifference* some generalizations follow immediately.

PROPOSITION 3.4
Let $\mathcal{I}$ satisfy *Prototypical Indifference*. Let $\mathcal{R}$ be a knowledge base and $D \subseteq U$.

1. Let G_1, G_2 be two ground sentences. For $\mathsf{c}_1, \mathsf{c}_2 \in D$ with $\mathsf{c}_1 \equiv_\mathcal{R} \mathsf{c}_2$ it holds $\mathcal{I}(\mathcal{R}, D)(G_2|G_1) = \mathcal{I}(\mathcal{R}, D)(G_2[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2] \,|\, G_1[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2])$.
2. Let $S \in \mathcal{S}_\mathcal{R}$, $\mathsf{c}_1, \ldots, \mathsf{c}_n \in S$, and $\sigma : S \to S$ be a permutation on S, i. e. a bijective function on S. Then it holds $\mathcal{I}(\mathcal{R}, D)(A) = \mathcal{I}(\mathcal{R}, D)(A[\sigma(\mathsf{c}_1)/\mathsf{c}_1, \ldots, \sigma(\mathsf{c}_n)/\mathsf{c}_n])$.

The following postulate focusses on the implications that a population-based statement $r = (B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha]$ should have for the probability of a proper instantiation $P(B(\vec{\mathsf{c}})|A(\vec{\mathsf{c}}))$. Our intention about r is that in general, the conditional probability of $B(\vec{\mathsf{c}})$ given $A(\vec{\mathsf{c}})$ "should" be (around) α. But surely, we cannot guarantee that every possible instantiation r' of r will conform to a strict interpretation of this demand. This follows mainly from the fact, that using ground conditionals we should be able to give exceptions to this rule, cf. Example 2.4. What we really want to describe when representing a population-based statement r is that *given an adequate large*

domain, the respective conditional probability for constant tuples that may serve as prototypes will converge towards α. This behavior resembles the intuition behind the "Law of Large Numbers" [4].

(Conditional Probability in the Limit (CPL)) Let $D_1 \subset D_2 \subset \ldots$ be a sequence of sets with $D_i \subseteq U$ for all $i \in \mathbb{N}$. For a conditional $r = (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$, let $(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}}))[\alpha]$ be a proper instantiation of r with constants $\vec{\mathsf{c}}$ that do not appear in $\mathcal{R}$. Then it holds that

$$\lim_{i \to \infty} \mathcal{I}(\mathcal{R}, D_i)(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) = \alpha \quad .$$

The important aspect of population-based statements is their capability of expressing a general behavior within a population while allowing for exceptions. So, population-based statements are to reflect some kind of *expected value* over the set of individual instantiations that aggregates individual behaviors. As such, if the probability of one instantiation of a population-based statement lies below the probability assigned to the statement, there has to be another instantiation with a probability higher than this probability value in order to compensate for the other exception (remember that D is assumed to be finite).

(Compensation) Let $\mathcal{R}$ be $\circ$-consistent under D and let $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$ be a non-ground conditional with $0 < \alpha < 1$. If $\vec{\mathsf{c}_1}$ is a vector of constants such that $\mathcal{I}(\mathcal{R}, D)(B(\vec{\mathsf{c}_1}) \mid A(\vec{\mathsf{c}_1})) < \alpha$ then there is another vector of constants $\vec{\mathsf{c}_2}$ with $\mathcal{I}(\mathcal{R}, D)(B(\vec{\mathsf{c}_2}) \mid A(\vec{\mathsf{c}_2})) > \alpha$.

Furthermore, when considering non-ground conditionals $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$ with $\alpha \in \{0, 1\}$ no compensation for exceptions is possible thus requiring *direct inference* [2] for this particular case.

(Strict Inference) Let $\mathcal{R}$ be $\circ$-consistent under D and let $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$ be a non-ground conditional with $\alpha \in \{0, 1\}$. Then for any $(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D(A(\vec{\mathsf{X}}) \mid B(\vec{\mathsf{X}}))$ it holds that $\mathcal{I}(\mathcal{R}, D)(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) = \alpha$.

In the following section, we present two operators that satisfy all postulates given above.

4 Relational Maximum Entropy Reasoning

In the propositional case, ME-inference (*Maximum Entropy*) has proven to be a suitable approach for commonsense reasoning as it features several nice properties [15, 20]. The entropy $H(P)$ of a probability distribution P is defined as

$$H(P) = -\sum_{\omega \in \Omega} P(\omega) \log P(\omega),$$

and measures the amount of indeterminateness inherent in P (with $0 \log 0 = 0$). By selecting the unique probability distribution P^* among all probabilistic models of a (propositional) set of formulas S that has maximal entropy, i. e. by computing the solution to the optimization problem

$$P^* := \mathsf{ME}(S) = \arg\max_{P \models S} H(P),$$

we get the one probability distribution that satisfies S and adds as little information as necessary. For further details, we refer to [15, 20].

As we are interested in generalizing the propositional ME-operator to the first-order case, we will postulate a proper form of compatibility to the propositional ME-inference, in addition to the postulates stated for general inference operators in the previous section. For ground knowledge bases (which can be considered as propositional knowledge bases), the operation $\mathcal{I}$ should coincide with the ME-operator.

(ME-Compatibility) Let $\mathcal{R}$ be a ground knowledge base. If A is a ground sentence then it holds that $\mathsf{ME}(\mathcal{R})(A) = \mathcal{I}(\mathcal{R}, \mathsf{consts}(\mathcal{R}))(A)$.

After having introduced the averaging and the aggregating semantics for relational probabilistic knowledge bases, now we apply the maximum entropy principle to the respective model sets to single out "best" models. For the relational case we parametrize the entropy $H_D(P)$ of a probability distribution P with the set $D \subseteq U$ of constants under consideration via

$$H_D(P) = - \sum_{\omega \in \Omega, \mathsf{consts}(\omega) \subseteq D} P(\omega) \log P(\omega) \quad .$$

As both our semantics require $P(\omega) = 0$ for $\mathsf{consts}(\omega) \not\subseteq D$ in order for $P, D \models_\circ \mathcal{R}$ to hold the above definition only neglects terms that are zero anyway.

4.1 Relational Maximum Entropy Inference by Averaging Probabilities

In the following we define our first variant of an ME-inference $\mathcal{I}_\varnothing$ in a relational context, that is based upon the semantics $\models_\varnothing$. A preliminary discussion of this operator can also be found in [35]. As (2.2) yields a set of non-convex constraints we define $\mathcal{I}_\varnothing(\mathcal{R}, D)$ as

$$\mathcal{I}_\varnothing(\mathcal{R}, D) = \begin{cases} \arg \max\limits_{P, D \models_\varnothing \mathcal{R}} H_D(P) & \text{if unique and } \mathsf{consts}(\mathcal{R}) \subseteq D \\ \mathsf{undef} & \text{otherwise} \end{cases} \tag{4.1}$$

The second case catches scenarios where either $\mathcal{R}$ is $\varnothing$-inconsistent under D or the optimization problem of the first case is not uniquely solvable. Obviously, $\mathcal{I}_\varnothing$ is a model-based inference operator using semantics $\models_\varnothing$. In particular, if $\mathcal{R}$ is $\varnothing$-consistent under D there is at least one probability distribution with maximum entropy that can be chosen in Equation (4.1).

In the following we give some theoretical results that the proposed operator $\mathcal{I}_\varnothing$ indeed fulfills the desired properties discussed in the previous section. Due to the non-convexity of the optimization problem defined by (4.1) $\mathcal{I}_\varnothing$ satisfies (Well-Definedness) only for the case that (4.1) is uniquely solvable. However, all examples considered so far were indeed uniquely solvable and we conjecture that $\mathcal{I}_\varnothing$ satisfies *Well-Definedness*. However, a formal proof for $\mathcal{I}_\varnothing$ satisfying *Well-Definedness* has not been found yet.

Proposition 4.1
The inference operator $\mathcal{I}_\varnothing$ satisfies *Name Irrelevance*, *Prototypical Indifference*, ME-*Compatibility*, *Compensation*, and *Strict Inference*. If $\mathcal{I}_\varnothing$ satisfies *Well-Definedness* then $\mathcal{I}_\varnothing$ satisfies *Conditional Probability in the Limit*.

We continue by investigating the behavior of $\mathcal{I}_\varnothing$ on some benchmark examples.

EXAMPLE 4.2
We continue Example 2.4. So let $\mathcal{L}_\Sigma$ be a first-order language with predicates *elephant*/1, *keeper*/1, and *likes*/2 and $D = \{\mathsf{clyde}, \mathsf{dumbo}, \mathsf{giddy}, \mathsf{fred}, \mathsf{dave}\}$. Let $\mathcal{R}_{zoo2}$ be given by $\mathcal{R}_{zoo2} = \{r_1, \ldots, r_7\}$ with

$$
\begin{aligned}
r_1 &: (\mathit{elephant}(\mathsf{clyde}))[1] && (4.2)\\
r_2 &: (\mathit{elephant}(\mathsf{giddy}))[1] && (4.3)\\
r_3 &: (\mathit{keeper}(\mathsf{fred}))[1] && (4.4)\\
r_4 &: (\mathit{keeper}(\mathsf{dave}))[1] && (4.5)\\
r_5 &: (\mathit{likes}(\mathsf{X},\mathsf{Y}) \mid \mathit{elephant}(\mathsf{X}) \wedge \mathit{keeper}(\mathsf{Y}))[0.6] && (4.6)\\
r_6 &: (\mathit{likes}(\mathsf{X},\mathsf{fred}) \mid \mathit{elephant}(\mathsf{X}) \wedge \mathit{keeper}(\mathsf{fred}))[0.4] && (4.7)\\
r_7 &: (\mathit{likes}(\mathsf{clyde},\mathsf{fred}) \mid \mathit{elephant}(\mathsf{clyde}) \wedge \mathit{keeper}(\mathsf{fred}))[0.7] && (4.8)
\end{aligned}
$$

Notice, that we have no knowledge of Dumbo being an elephant. In the following we give the probabilities of several instantiations of *likes* in $\mathcal{I}_\varnothing(\mathcal{R}_{zoo2}, D)$.

$$
\begin{aligned}
\mathcal{I}_\varnothing(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{clyde},\mathsf{dave})) &\approx 0.723 && (4.9)\\
\mathcal{I}_\varnothing(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{dumbo},\mathsf{dave})) &\approx 0.642 && (4.10)\\
\mathcal{I}_\varnothing(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{giddy},\mathsf{dave})) &\approx 0.723 && (4.11)\\
\mathcal{I}_\varnothing(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{clyde},\mathsf{fred})) &= 0.7 && (4.12)\\
\mathcal{I}_\varnothing(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{dumbo},\mathsf{fred})) &\approx 0.387 && (4.13)\\
\mathcal{I}_\varnothing(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{giddy},\mathsf{fred})) &\approx 0.36 && (4.14)\\
\mathcal{I}_\varnothing(\mathcal{R}_{zoo2}, D)(\mathit{elephant}(\mathsf{dumbo})) &\approx 0.312 && (4.15)
\end{aligned}
$$

Notice, how the deviations brought about by the exceptional individuals Clyde and Fred have to be balanced out by the other individuals. For example, the probabilities of the individual elephants liking Dave are greater than conditional (4.6) specified them to be. This is because the probabilities of the elephants liking Fred is considerably smaller as demanded by conditional (4.7). Nonetheless, the average of the conditional probabilities do indeed satisfy the conditionals in $\mathcal{R}$. Notice furthermore, that the probability of Dumbo being an elephant is very small—see (4.15)—considering that maximum entropy is achieved by deviating only as little as possible from the uniform distribution. But due to the interaction of the conditionals in $\mathcal{R}_{zoo2}$, a smaller probability of Dumbo being an elephant is necessary in order to achieve the correct average conditional probabilities defined in the knowledge base. Thus, the belief of Dumbo being an elephant alleviates due to the premise of believing in the defined conditionals.

The next example is the well-known Tweety-example that is often used in the context of non-monotonic reasoning.

EXAMPLE 4.3
Consider a first-order language $\mathcal{L}_\Sigma$ with predicates *bird*/1, *flies*/1, and *penguin*/1 and

$D = \{\mathsf{tweety}, \mathsf{opus}, \mathsf{brian}\}$. Let $\mathcal{R}_{birds}$ be given by $\mathcal{R}_{birds} = \{r_1, \ldots, r_6\}$

$$r_1 \;:\; (\mathit{bird}(\mathsf{tweety}))[1] \tag{4.16}$$
$$r_2 \;:\; (\mathit{bird}(\mathsf{opus}))[1] \tag{4.17}$$
$$r_3 \;:\; (\mathit{bird}(\mathsf{brian}))[1] \tag{4.18}$$
$$r_4 \;:\; (\mathit{penguin}(\mathsf{opus}))[0.9] \tag{4.19}$$
$$r_5 \;:\; (\mathit{flies}(\mathsf{X}) \mid \mathit{bird}(\mathsf{X}))[0.6] \tag{4.20}$$
$$r_6 \;:\; (\mathit{flies}(\mathsf{X}) \mid \mathit{penguin}(\mathsf{X}))[0.01] \tag{4.21}$$

The knowledge base $\mathcal{R}_{birds}$ models a scenario, where we have three birds Tweety, Opus, and Brian, and have a high degree of belief of 0.9 that Opus is a penguin. We furthermore know that birds typically fly with a probability of 0.6 and that penguins usually fly with a probability of 0.01. Applying $\mathcal{I}_{\varnothing}$ on $\mathcal{R}_{birds}$ yields the following results on several queries:

$$\mathcal{I}_{\varnothing}(\mathcal{R}_{birds}, D)(\mathit{flies}(\mathsf{tweety})) \approx 0.84 \tag{4.22}$$
$$\mathcal{I}_{\varnothing}(\mathcal{R}_{birds}, D)(\mathit{flies}(\mathsf{brian})) \approx 0.84 \tag{4.23}$$
$$\mathcal{I}_{\varnothing}(\mathcal{R}_{birds}, D)(\mathit{flies}(\mathsf{opus})) \approx 0.12 \tag{4.24}$$
$$\mathcal{I}_{\varnothing}(\mathcal{R}_{birds}, D)(\mathit{penguin}(\mathsf{tweety})) \approx 0.079 \tag{4.25}$$
$$\mathcal{I}_{\varnothing}(\mathcal{R}_{birds}, D)(\mathit{penguin}(\mathsf{brian})) \approx 0.079 \tag{4.26}$$

Due to (Prototypical Indifference) both birds Tweety and Brian fly with a probability of 0.84, see (4.22) and (4.23). As both Tweety and Brian are birds—see (4.16) and (4.18)—this probability is slightly higher than expected, cf. (4.20). This is due to the fact that the major deviation caused by Opus has to be compensated for. Opus flies only with a probability of 0.12—see (4.24)—as it is highly believed that Opus is a penguin and penguins fly with a very small probability, cf. (4.19) and (4.21). Furthermore, both Tweety and Brian are believed to be penguins with a very small probability of 0.079, cf. (4.25) and (4.26). As our domain consists of only three birds and (4.20) demands that the average probability of a bird flying is 0.6 the possibility of Tweety an Brian being penguins diminishes.

Up until now, all examples considered only a first-order language with unary predicates. In the next example we will introduce relational aspects. The example has been taken from [13].

Example 4.4
Consider a first-order language $\mathcal{L}_\Sigma$ with predicates *contact*/2, *susceptible*/1, and *flu*/1 and $D = \{\mathsf{anna}, \mathsf{bob}, \mathsf{carl}\}$. Let $\mathcal{R}_{flu}$ be given by $\mathcal{R}_{flu} = \{r_1, \ldots, r_3\}$

$$r_1 \;:\; (\mathit{flu}(\mathsf{X}))[0.2] \tag{4.27}$$
$$r_2 \;:\; (\mathit{flu}(\mathsf{X}) \mid \mathit{susceptible}(\mathsf{X}))[0.3] \tag{4.28}$$
$$r_3 \;:\; (\mathit{flu}(\mathsf{X}) \mid \mathit{contact}(\mathsf{X}, \mathsf{Y}) \wedge \mathit{flu}(\mathsf{Y}))[0.4] \tag{4.29}$$

This knowledge base models contagiousness of flu within some population. Conditional r_1 states that in general someone catches the flu with probability 0.2 while conditional r_2 gives a higher probability of 0.3 to someone who is susceptible. Finally, conditional r_3 models a situation where someone can get infected by someone

else who is already infected. Observe, that we do not represent any factual knowledge about our domain in $\mathcal{R}_{flu}$. Applying $\mathcal{I}_{\varnothing}$ on $\mathcal{R}_{flu}$ yields the following results on several queries:

$$\mathcal{I}_{\varnothing}(\mathcal{R}_{flu}, D)(flu(\mathsf{anna})) \approx 0.2 \quad (4.30)$$

$$\mathcal{I}_{\varnothing}(\mathcal{R}_{flu}, D)(flu(\mathsf{anna}) \mid contact(\mathsf{anna}, \mathsf{bob}) \wedge flu(\mathsf{bob})) \approx 0.4 \quad (4.31)$$

$$\mathcal{I}_{\varnothing}(\mathcal{R}_{flu}, D)(flu(\mathsf{anna}) \mid contact(\mathsf{anna}, \mathsf{bob}) \wedge flu(\mathsf{bob}) \wedge contact(\mathsf{anna}, \mathsf{carl}) \wedge flu(\mathsf{carl})) \approx 0.6 \quad (4.32)$$

$$\mathcal{I}_{\varnothing}(\mathcal{R}_{flu}, D)(contact(\mathsf{bob}, \mathsf{carl})) \approx 0.49 \quad (4.33)$$

$$\mathcal{I}_{\varnothing}(\mathcal{R}_{flu}, D)(contact(\mathsf{bob}, \mathsf{carl}) \mid flu(\mathsf{bob}), flu(\mathsf{carl}))) \approx 0.657 \quad (4.34)$$

Observe that we also stated some conditional queries involving actually present evidence. Notice that formulating queries in this form—for example considering the second query that models "what is the probability of Anna having a flu given that Anna had contact with Bob and Bob had the flu"—yields in general different inferences than adding the evidence to the knowledge base—in this case $(contact(\mathsf{anna}, \mathsf{bob}))[1.0]$ and $(flu(\mathsf{bob}))[1.0]$—and querying the new knowledge base just for $flu(\mathsf{anna})$, cf. [29] for a discussion on this topic.

The inferences drawn from $\mathcal{R}_{flu}$ using $\mathcal{I}_{\varnothing}$ reflect quite nicely the intuition behind the modeled knowledge. The probability of Anna having a flu (4.32) exactly models the expected probability when including conditional (4.27). The same is true for the probability of Anna having a flu given that Anna had contact with Bob and Bob had the flu (4.31). Furthermore, if a person had contact with multiple persons who have the flu the probability of having a flu increases (4.32). Applying the principle of maximum entropy to complete unspecified knowledge usually yields a probability distribution that is as close to the uniform distribution as possible. As one can see from the probability of Bob having contact with Carl (4.33) this might decrease a little bit if the corresponding formula appears in the premise of another conditional in the knowledge base (see 4.29), see [27] for a discussion. But knowing that two persons have the flu increases the probability of these two persons having contact (4.34).

4.2 Relational Maximum Entropy Inference by Aggregating Probabilities

In a similar manner like above, we define the ME-inference operator $\mathcal{I}_{\odot}$ that is based upon the semantics $\models_{\odot}$. Let

$$\mathcal{I}_{\odot}(\mathcal{R}, D) = \begin{cases} \arg \max_{P, D \models_{\odot} \mathcal{R}} H_D(P) & \text{if } \mathcal{R} \text{ is } \odot\text{-consistent under } D \\ & \text{and } \mathsf{consts}(\mathcal{R}) \subseteq D \\ \mathsf{undef} & \text{otherwise} \end{cases} \quad (4.35)$$

Obviously, $\mathcal{I}_{\odot}$ is a model-based inference operator using semantics $\models_{\odot}$. In this semantical context, the conditionals from $\mathcal{R}$ induce linear constraints on the probabilities of the possible worlds so that the set of probability distributions satisfying $\mathcal{R}$ forms a convex set. This makes the solution to the optimization problem (4.35) unique (if a solution exists).

LEMMA 4.5
Let $r = (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$ be a probabilistic conditional. Then $Mod^D_{\odot}(\{r\})$ is convex for any D with $\mathsf{consts}(r) \subseteq D$.

PROPOSITION 4.6
Let $\mathcal{R}$ be a $\odot$-consistent knowledge base and D some set $D \subseteq U$. Then $\mathcal{I}_{\odot}(\mathcal{R}, D)$ is uniquely determined.

The operator $\mathcal{I}_{\odot}$ satisfies all postulates listed in the previous section.

PROPOSITION 4.7
$\mathcal{I}_{\odot}$ satisfies *Well-Definedness*, *Name Irrelevance*, *Prototypical Indifference*, ME-*Compatibility*, *Conditional Probability in the Limit*, *Strict Inference*, and *Compensation*.

In contrast to $\mathcal{I}_{\varnothing}$ proving *Well-Definedness* is easy due to Lemma 4.5 and Proposition 4.6. In the following we apply $\mathcal{I}_{\odot}$ to the very same examples used in the previous subsection to discuss the operator $\mathcal{I}_{\varnothing}$.

EXAMPLE 4.8
We apply $\mathcal{I}_{\odot}$ onto the knowledge base $\mathcal{R}_{zoo2}$ from Example 4.2. This yields the following inferences:

$$\mathcal{I}_{\odot}(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{clyde}, \mathsf{dave})) \approx 0.8 \tag{4.36}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{dumbo}, \mathsf{dave})) \approx 0.64 \tag{4.37}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{giddy}, \mathsf{dave})) \approx 0.8 \tag{4.38}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{clyde}, \mathsf{fred})) = 0.7 \tag{4.39}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{dumbo}, \mathsf{fred})) \approx 0.356 \tag{4.40}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{zoo2}, D)(\mathit{likes}(\mathsf{giddy}, \mathsf{fred})) \approx 0.196 \tag{4.41}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{zoo2}, D)(\mathit{elephant}(\mathsf{dumbo})) \approx 0.475 \tag{4.42}$$

The results are similar to those computed by using $\mathcal{I}_{\varnothing}$ in the example above. In particular, with regard to liking Dave, both approaches calculate very similar probabilities for all individuals mentioned in the queries. Here, Dumbo—the individual not known to be an elephant—likes Dave with a lower probability than the elephants Clyde and Giddy, cf. (4.36), (4.37), and (4.38). More substantial differences can be noticed with respect to the elephants liking the moody keeper Fred. For Giddy liking Fred, $\mathcal{I}_{\odot}$ returns a considerably lower probability than $\mathcal{I}_{\varnothing}$, see (4.41). However, $\mathcal{I}_{\odot}$ is more cautious when processing information on Dumbo, its probability of being an elephant is nearly 0.5 (4.42), while $\mathcal{I}_{\varnothing}$ suggests that Dumbo is not an elephant.

EXAMPLE 4.9
We apply $\mathcal{I}_{\odot}$ onto the knowledge base $\mathcal{R}_{birds}$ from Example 4.3. This yields the following inferences

$$\mathcal{I}_{\odot}(\mathcal{R}_{birds}, D)(\mathit{flies}(\mathsf{tweety})) \approx 0.85 \tag{4.43}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{birds}, D)(\mathit{flies}(\mathsf{brian})) \approx 0.85 \tag{4.44}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{birds}, D)(\mathit{flies}(\mathsf{opus})) \approx 0.10 \tag{4.45}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{birds}, D)(\mathit{penguin}(\mathsf{tweety})) \approx 0.079 \tag{4.46}$$
$$\mathcal{I}_{\odot}(\mathcal{R}_{birds}, D)(\mathit{penguin}(\mathsf{brian})) \approx 0.079 \tag{4.47}$$

As in the previous example, the inferences drawn using $\mathcal{I}_{\odot}$ are very similar to the ones using $\mathcal{I}_{\varnothing}$. The probabilities of Tweety and Brian being penguins (0.079) are exactly the same as in Example 4.3. There are only minor differences in the probabilities of the instantiations of *flies*. While using $\mathcal{I}_{\varnothing}$ the probability of Tweety and Opus flying is 0.84 resp. 0.12 here we have 0.85 resp. 0.10.

As for Example 4.4 applying the operator $\mathcal{I}_{\odot}$ on $\mathcal{R}_{flu}$ yields exactly the same inferences as $\mathcal{I}_{\varnothing}$. This is due to the fact that both operators fulfill *Prototypical Indifference*. Consider the conditional $(\mathit{flu}(\mathsf{X}) \mid \mathit{susceptible}(\mathsf{X}))[0.3]$. As no constant is mentioned in $\mathcal{R}_{flu}$ all of $\mathsf{anna}, \mathsf{bob}, \mathsf{carl}$ belong to the same $\mathcal{R}_{flu}$-equivalence class and therefore it follows

$$\begin{aligned}\mathcal{I}_{\circ}(\mathcal{R}_{flu}, D)(\mathit{flu}(\mathsf{anna}) \mid \mathit{susceptible}(\mathsf{anna})) &= \mathcal{I}_{\circ}(\mathcal{R}_{flu}, D)(\mathit{flu}(\mathsf{bob}) \mid \mathit{susceptible}(\mathsf{bob})) \\ &= \mathcal{I}_{\circ}(\mathcal{R}_{flu}, D)(\mathit{flu}(\mathsf{carl}) \mid \mathit{susceptible}(\mathsf{carl}))\end{aligned}$$

for any $\circ \in \{\varnothing, \odot\}$ due to Proposition 3.4. This directly yields

$$\mathcal{I}_{\circ}(\mathcal{R}_{flu}, D)(\mathit{flu}(\mathsf{anna}) \mid \mathit{susceptible}(\mathsf{anna})) = 0.3$$

for $\circ \in \{\varnothing, \odot\}$ as can also be seen in the elaboration in Example 4.4. If we add probabilistic facts like $(\mathit{contact}(\mathsf{anna}, \mathsf{bob}))[1]$ or $(\mathit{flu}(\mathsf{bob}))[1]$ to $\mathcal{R}_{flu}$ the situation changes and now different inferences can be drawn from the different semantics. One thing to notice about this particular special case of a knowledge base—a knowledge base that mentions no constants—is that there is a direct method to reasoning. As has been discussed above due to *Prototypical Indifference* all inferences drawn from different instantiations are identical. As a result replacing the (open) conditional $(\mathit{flu}(\mathsf{X}) \mid \mathit{susceptible}(\mathsf{X}))[0.3]$ by its universal instantiations

$$\begin{aligned}&(\mathit{flu}(\mathsf{anna}) \mid \mathit{susceptible}(\mathsf{anna}))[0.3] \\ &(\mathit{flu}(\mathsf{bob}) \mid \mathit{susceptible}(\mathsf{bob}))[0.3] \\ &(\mathit{flu}(\mathsf{carl}) \mid \mathit{susceptible}(\mathsf{carl}))[0.3]\end{aligned}$$

amounts to yield the very same ME-distribution. In general, replacing every open conditional in a constant-free knowledge base with its universal instantiations yields the same ME-distribution independently of using averaging or aggregating semantics. In this case we can employ standard ME-reasoner for propositional probabilistic conditional logic, for example [32], for inference due to Remark 2.3.

As a final remark one can notice that the inferences drawn from both operators are very similar. On the one hand, this is not surprising as both operators satisfy (in principle) the desired properties which heavily restrict the choice for rational inference operators. On the other hand, this observation is quite interesting from a computational point of view as solving the optimization problems (4.1) and (4.35) require different approaches: while (4.1) is a non-convex optimization problem Equation (4.35) describes a convex optimization problem. For the latter efficient solvers are available [5]. However, the following example shows that the results drawn from the two operators may differ significantly, cf. Corollary 2.14.

Example 4.10
Consider the scenario of a bird sanctuary. We know that there are exactly 1000 birds in this sanctuary, divided into two species: the striped sea eagle and the rare snoring

ostrich[1]. Statistically seen, 999 of these birds are striped sea eagles and one of them is a snoring ostrich and no bird can be both at the same time. It is common knowledge that all striped sea eagles do fly and that snoring ostriches do not fly. Furthermore, only a few striped sea eagles are pink but every snoring ostrich is pink. This scenario can be represented as the knowledge base $\mathcal{R}_1 = \{r_1, \ldots, r_7\}$ given via

$$\begin{aligned} r_1 &= (sse(\mathsf{X}))[0.999] & r_2 &= (so(\mathsf{X}))[0.001], \\ r_3 &= (sse(\mathsf{X}) \wedge so(\mathsf{X}))[0] & r_4 &= (\mathit{flies}(\mathsf{X}) \mid sse(\mathsf{X}))[1], \\ r_5 &= (\mathit{flies}(\mathsf{X}) \mid so(\mathsf{X}))[0] & r_6 &= (\mathit{pink}(\mathsf{X}) \mid sse(\mathsf{X}))[0.001] \\ r_7 &= (\mathit{pink}(\mathsf{X}) \mid so(\mathsf{X}))[1] \end{aligned}$$

where $sse(\mathsf{X})$ means that X is a striped sea eagle, $so(\mathsf{X})$ means that X is a snoring ostrich, $\mathit{flies}(\mathsf{X})$ means that X flies, and $\mathit{pink}(\mathsf{X})$ means that X is pink. Note that $\mathcal{R}_1$ does not mention any constants, therefore all individuals behave the same and each instance of r_1–r_7 takes its intended probability. As a consequence it follows that

$$P_1 = \mathcal{I}_{\varnothing}(\mathcal{R}_1, D) = \mathcal{I}_{\odot}(\mathcal{R}_1, D)$$

for every D with $D \neq \emptyset$. The question we want to address is "*What is the probability of a pink bird flying?*", i. e., we want to assess the probability of the conditional $(\mathit{flies}(\mathsf{X}) \mid \mathit{pink}(\mathsf{X}))$. For $\mathcal{R}_1$ we get

$$\begin{aligned} P_1, D &\models_{\varnothing} (\mathit{flies}(\mathsf{X}) \mid \mathit{pink}(\mathsf{X}))[0.499] \\ P_1, D &\models_{\odot} (\mathit{flies}(\mathsf{X}) \mid \mathit{pink}(\mathsf{X}))[0.499] \end{aligned}$$

because $P_1(\mathit{flies}(\mathsf{a}) \mid \mathit{pink}(\mathsf{a})) = 0.499$ for every $\mathsf{a} \in D$. Consider now a slightly different scenario where $D = \{\mathsf{b}_1, \ldots, \mathsf{b}_{1000}\}$ is the actual set of birds in the sanctuary and let $\mathsf{b}_1, \ldots, \mathsf{b}_{999}$ be striped sea eagles and let there be a single snoring ostrich b_{1000}. This can be represented as the knowledge base $\mathcal{R}_2 = \{r'_{1,1}, \ldots, r'_{1,999}, r'_2, \ldots, r'_7\}$ given via

$$\begin{aligned} r'_{1,i} &= (sse(\mathsf{b}_i))[1] & &\text{for } i = 1, \ldots, 999 \\ r'_2 &= (so(\mathsf{b}_{1000}))[1] & r'_3 &= (sse(\mathsf{X}) \wedge so(\mathsf{X}))[0] \\ r'_4 &= (\mathit{flies}(\mathsf{X}) \mid sse(\mathsf{X}))[1] & r'_5 &= (\mathit{flies}(\mathsf{X}) \mid so(\mathsf{X}))[0], \\ r'_6 &= (\mathit{pink}(\mathsf{X}) \mid sse(\mathsf{X}))[0.001] & r'_7 &= (\mathit{pink}(\mathsf{X}) \mid so(\mathsf{X}))[1] \quad . \end{aligned}$$

For $\mathcal{R}_2$ we obtain

$$\begin{aligned} \mathcal{I}_{\varnothing}(\mathcal{R}_2, D), D &\models_{\varnothing} (\mathit{flies}(\mathsf{X}) \mid \mathit{pink}(\mathsf{X}))[0.999] \\ \mathcal{I}_{\odot}(\mathcal{R}_2, D), D &\models_{\odot} (\mathit{flies}(\mathsf{X}) \mid \mathit{pink}(\mathsf{X}))[0.499] \quad . \end{aligned}$$

As one can see $\mathcal{I}_{\odot}$ makes no distinction between the knowledge bases $\mathcal{R}_1$ and $\mathcal{R}_2$ with respect to the probabilistic conditional $r = (\mathit{flies}(\mathsf{X}) \mid \mathit{pink}(\mathsf{X}))$ and assigns the probability 0.499 in both cases. The operator $\mathcal{I}_{\varnothing}$, however, assigns a probability 0.499 to r in $\mathcal{R}_1$ and 0.999 in $\mathcal{R}_2$. On the one hand, representing the open probabilistic fact $(sse(\mathsf{X})[0.999]$ as the set of ground facts $(sse(\mathsf{b}_1))[1], \ldots, (sse(\mathsf{b}_{999}))[1]$ seems to

[1] These species are just made up.

be equivalent when fixing the domain D. As a consequence, an inference operator should make no distinction between $\mathcal{R}_1$ and $\mathcal{R}_2$. On the other hand, note that $\mathcal{R}_1$ and $\mathcal{R}_2$ are neither $\varnothing$- nor $\odot$-equivalent with respect to D. The knowledge base $\mathcal{R}_1$ gives no information on the actual distribution of $\mathsf{b}_1, \ldots, \mathsf{b}_{1000}$ to the different species. The operator $\mathcal{I}_\varnothing$ is able to recognize this difference. However, whether it is justified to assign the probability 0.999 to r in $\mathcal{R}_2$ depends on the interpretation of r from the point of view of commonsense reasoning. As for aggregating semantics the probability of r is interpreted by taking the probabilities of the premises into account as well. On the one hand, the probability of r is influenced by the probabilities of the instances $(\mathit{flies}(\mathsf{b}_1) \mid \mathit{pink}(\mathsf{b}_1)), \ldots, (\mathit{flies}(\mathsf{b}_{999}) \mid \mathit{pink}(\mathsf{b}_{999}))$ only to a small extent as the probability of the premises $\mathit{pink}(\mathsf{b}_1), \ldots, \mathit{pink}(\mathsf{b}_{999})$ is rather low (0.001 to be precise). On the other hand, the probability of r is heavily influenced by the probability of the instance $(\mathit{flies}(\mathsf{b}_{1000}) \mid \mathit{pink}(\mathsf{b}_{1000}))$ as the premise $\mathit{pink}(\mathsf{b}_{1000})$ has probability one. As $\mathit{pink}(\mathsf{b}_{1000})$ has such a high probability aggregating semantics classifies b_{1000} as a good "reference" for the applicability of r. Averaging semantics on the other side is not influenced by the actual probabilities of the premise. The ground conditionals $(\mathit{flies}(\mathsf{b}_1) \mid \mathit{pink}(\mathsf{b}_1)), \ldots, (\mathit{flies}(\mathsf{b}_{999}) \mid \mathit{pink}(\mathsf{b}_{999}))$ all hold with probability 1 as $\mathsf{b}_1, \ldots, \mathsf{b}_{999}$ fly independently of their color. The ground conditional $(\mathit{flies}(\mathsf{b}_{1000}) \mid \mathit{pink}(\mathsf{b}_{1000}))$ has probability 0 as b_{1000} does not fly independently of the color. Therefore, interpreting r as "usually, pink objects fly" on the given domain is ambiguous. Aggregating semantics acknowledges this indifference by assigning a probability of approximately 0.5 to r which is justifiable as flying objects are rarely pink and non-flying objects are always pink. However, the probability of $(\mathit{flies}(\mathsf{c}) \mid \mathit{pink}(\mathsf{c}))$ is one for 99.9 % of the population (for both $\mathcal{I}_\varnothing(\mathcal{R}_2, D)$ and $\mathcal{I}_\odot(\mathcal{R}_2, D)$) which also justifies assigning probability 0.999 to r.

There seems to be no definite answer to the question which of the both semantics and inference operators are more appropriate for interpreting relational conditionals. Both meanings are justifiable by considering a specific perspective on their meaning. This perspective might be influenced by the actual knowledge base and the intended meaning of the probabilistic conditionals. It follows that there are knowledge bases where the averaging semantics might be more suitable than the aggregating semantics and vice versa. In particular, in the above example there are two different views which justify application of one specific semantics.

5 Related Work

In this section we review some work related to our approach discussed so far. In particular, we compare our approach to the work by Halpern et. al. [18, 16, 17], other approaches for reasoning under maximum entropy in first-order probabilistic conditional logic [21, 13, 23], as well as Markov Logic Networks [14, Ch. 12] and Bayesian Logic Programs [14, Ch. 10] which are exemplary for a variety of approaches to statistical relational learning and probabilistic inductive logic programming [14, 7].

5.1 Statistical Approaches to First-Order Probabilistic Reasoning

The papers by Halpern and colleagues [18, 16, 17] aim at bridging statistical and subjective views on probabilistic beliefs by showing how subjective beliefs arise from statistical information by considering approximative probabilities and limits. The principle of maximum entropy plays a prominent role in these frameworks, too, but the authors mention problems when applying this principle to knowledge bases with n-ary predicates with $n > 1$. As our semantical approaches are thoroughly subjective by choosing subjective probabilities throughout, we did not encounter most of the problems that those authors have to struggle with. For instance, in statistical approaches to probabilities, the size of the universe determines the probabilities that can be realized, so approximations of probabilities have to be considered. This is not the case in our approaches, as no statistical interpretation underlies the probabilities. Moreover, the application of the maximum entropy principle to knowledge bases with arbitrary predicates seems to be unproblematic, but this has to be investigated in more detail in further work.

5.2 Relational Conditional Logic and Maximum Entropy

The approach of lifting inference in conditional logic based on the principle of maximum entropy to the first-order case has been previously investigated in [21, 13]. In [21] a probabilistic logic programming language is developed which bases on conditional constraints of the form $(B \mid A)[l, u]$ with first-order formulas B, A, and real values $l \leq u$. As in the present work the underlying first-order language is assumed to be quantifier-free and function-free. A probability distribution P satisfies a conditional constraint $(B \mid A)[l, u]$ if $P(B' \mid A') \in [l, u]$ for any ground instance $(B' \mid A')$ of $(B \mid A)$. A similar approach is pursued in [13] and [23]. The main difference to the present approach lies in the semantics of conditionals. While in [21, 13, 23] conditionals with free variables are interpreted as schemas for their instances, here, we take the mutual influences of instances on each other into account. In doing so we avoid the problem of conflicts and inconsistencies that may arise easily when grounding a relational knowledge base. Consider again the knowledge base $\mathcal{R}_{zoo} = \{r_1, r_2, r_3\}$ from Example 2.4 with

$$
\begin{aligned}
r_1 &: (\mathit{likes}(\mathsf{X}, \mathsf{Y}) \mid \mathit{elephant}(\mathsf{X}) \wedge \mathit{keeper}(\mathsf{Y}))[0.6] \\
r_2 &: (\mathit{likes}(\mathsf{X}, \mathsf{fred}) \mid \mathit{elephant}(\mathsf{X}) \wedge \mathit{keeper}(\mathsf{fred}))[0.4] \\
r_3 &: (\mathit{likes}(\mathsf{clyde}, \mathsf{fred}) \mid \mathit{elephant}(\mathsf{clyde}) \wedge \mathit{keeper}(\mathsf{fred}))[0.7]
\end{aligned}
$$

Treating the conditionals $\mathcal{R}_{zoo}$ as schemas for their instances and applying universal instantiation directly yields an inconsistent state. For example, conditional r_1 yields the instance $(\mathit{likes}(\mathsf{clyde}, \mathsf{fred}) \mid \mathit{elephant}(\mathsf{clyde}) \wedge \mathit{keeper}(\mathsf{fred}))[0.6]$ which directly conflicts with r_3. As a result, there can be no probability distribution P that satisfies $\mathcal{R}_{zoo}$ in this semantical sense. The cited approaches solve this problem in different ways. The work reported in [13] introduces logical constraint formulas that constrain the possible instantiations of a conditional. For example, an appropriate logical constraint formula for conditional r_1 above would be $\mathsf{Y} \neq \mathsf{fred}$ thus avoiding instances that may conflict with instances of other conditionals. But representing these logical constraint formulas must be done by the knowledge engineer and becomes hard when

many conditionals have to be considered. The approach proposed in [23] follows another direction by employing grounding strategies. There, a syntactical analysis of the knowledge base is employed in order to avoid and remove conflicting instances automatically. The knowledge engineer is not obliged to deliver logic constraint formula but can rely on a specific grounding strategy. But still, as this approach works on a syntactical layer the results are heuristically determined and may still yield an inconsistent state. The approach undertaken in [21] does not give solutions to this problem at all but assumes that the knowledge base may be consistently grounded. If this is not possible then the interval $[l, u]$ of a constraint $(B|A)[l, u]$ can be widened.

In contrast to the cited approaches the work presented here does not treat conditionals with free variables as schemas for their instances. The actual probability in the ME-model of a ground conditional $(B' \mid A')$ may differ significantly from the probability of its open conditional $(B \mid A)[\alpha]$ represented in the knowledge base. Given that the underlying language contains some minimum number of constants exceptions to a conditional can be compensated. This allows for a great flexibility when representing relational probabilistic knowledge and is also inherently important for a non-monotonic reasoning behavior. Our approaches aim at reflecting an overall behavior within a population to which each individual contributes, while at the same time allowing individuals to defer drastically from that behavior. In this way, both class knowledge and individual, maybe exceptional knowledge can be represented and processed within one framework. As the satisfaction of the CPL postulate shows, the overall behavior might also be interpreted as a prototypical behavior in universes which are large enough.

For knowledge bases that can be universally instantiated without yielding an inconsistent state another note can be made on the relationship of our approach to the ones cited above. Consider the approach of [13] which uses the same knowledge representation as employed here but with the addition of the previously mentioned logic constraint formula. If all conditionals in a knowledge base $\mathcal{R}$ have a tautological logical constraint formula, i. e., all conditionals can be treated as the set of their universal instantiations, then for every model P of $\mathcal{R}$ in the sense of [13] it follows $P \models_{\varnothing} \mathcal{R}$ and $P \models_{\odot} \mathcal{R}$. Furthermore, even the ME-models coincide in this special case. Consider the simple knowledge base $\mathcal{R} = \{r_1\}$ with

$$r_1 \quad : \quad (p(\mathsf{X}) \mid q(\mathsf{X}))[0.6]$$

and assume that our underlying language contains the constants $D = \{\mathsf{c}, \mathsf{d}\}$. In the approach of [13] a probability distributions satisfies r_1 iff it holds both $P(p(\mathsf{c})|q(\mathsf{c})) = 0.6$ and $P(p(\mathsf{d})|q(\mathsf{d})) = 0.6$. Hence, the average probability is 0.6 and thus $P \models_{\varnothing} \mathcal{R}$ and also $(6+6)/(10+10) = 0.6$ yielding $P \models_{\odot} \mathcal{R}$. Furthermore, any other $\circ$-model of $\mathcal{R}$ with $\circ \in \{\varnothing, \odot\}$ deviates from these values and thus yields a lower entropy. This results in the same ME-model for the approach of [13] and the approaches discussed here. This statement can be generalized to include the approaches of [21] and [23] as well.

5.3 Statistical Relational Learning

The areas of statistical relational learning and probabilistic inductive logic programming are concerned with the development of frameworks that combine probabilis-

tic reasoning and first-order representations of knowledge [14, 7]. Mostly, these approaches focus on learning models from data than on knowledge representation and reasoning. Markov Logic Programs (MLNs) [14, Ch. 12] employ a quantifier-free first-order logic as representation language and allow the attachment of weights to each piece of information in a knowledge base. Thus, an MLN consists of tuples (F, α) with a formula F and a weight $\alpha \in \mathbb{R}$. The weights have no obvious probabilistic interpretation and reasoning is performed based on a probability distribution P that is defined via

$$P(\omega) = \frac{1}{Z} \exp\left(\sum_{(F,\alpha)} n_F(\omega)\alpha\right) \tag{5.1}$$

for possible worlds $\omega \in \Omega$. In Equation (5.1), Z is a normalization constant and $n_F(\omega)$ is the number of instances of F that are true in ω. By defining P in this way, worlds that violate fewer instances of formulas are more probable than worlds that violate more instances (depending on the weights of the different formulas). In contrast to the approaches discussed in the previous subsection MLNs do not suffer from conflicts that arise in grounding knowledge bases as the weights of an MLN have no probabilistic interpretation, see [11] for some discussion. But from the view of knowledge representation this property is unintuitive and also distinguishes MLNs from our approach where values of conditionals do have a probabilistic interpretation.

Another popular approach for statistical relational learning are Bayesian logic programs (BLPs) [14, Ch. 10]. These are relational extensions to Bayes Nets [29] and employ *Bayes clauses* for knowledge representation. A Bayes clause is a rule of the form $(H \mid B_1, \ldots, B_n)$ with first-order atoms $H, B_1, \ldots, B_n$[2]. In contrast to the present approach and the approaches discussed before, BLPs demand a full specification of the conditional probability distribution for each clause. Thus, for a clause $(H(\mathsf{X}) \mid B(\mathsf{X}))$ the intended probabilities for both $H(\mathsf{X})$ given $B(\mathsf{X})$ and $H(\mathsf{X})$ given $\neg B(\mathsf{X})$ have to be specified. For a specific query Q, i. e. a ground atom, reasoning is performed by computing a ground Bayes Net that is build using an SLD-like procedure involving the represented clauses. In order to combine the probabilities deriving from different clauses with the same head, BLPs employ combining rules such as *noisy-or* [14, Ch. 10]. From the view of knowledge representation the main drawback of BLPs—which derives from the main drawback of Bayes nets—is that they demand a full specification of conditional probability distributions. Consider a clause $(\mathit{flu}(\mathsf{X}) \mid \mathit{symptomA}(\mathsf{X}))$ saying that given X has symptom A we can give a probability on X having a flu. As this probability should be representable by an expert BLPs also demand specifying the probability of X having a flu if X does *not* have symptom A. Giving a reasonable estimate for this probability is quite harder for an expert and almost impossible in most scenarios. In contrast to BLPs the approach discussed in this paper works with incomplete information as well. For the knowledge engineer it suffices to represent only as much information as he wants to. Due to the principle of maximum entropy all missing pieces of information are determined in the most unbiased way. Another problem of BLPs is the employment of combining that combine probabilities coming

[2] BLPs allow the use of multi-valued predicates, i. e. predicates that may have other ranges than the usual boolean range {true, false}. For example, an appropriate range for the predicate *bloodtype* may be {a, b, ab, null}, cf. [14, Ch. 10]. This behavior can be simulated by adding an additional argument to the predicate but we will omit a deeper discussion of this topic. In the following, we only use boolean predicates for knowledge representation with BLPs.

from different clauses that derive the same head. For most applications, *noisy-or* is a reasonable choice for a combining rule as it models a disjunctive combination in a probabilistic sense. Given clauses $(\mathit{flu}(\mathsf{X}) \mid \mathit{symptomA}(\mathsf{X}))$ and $(\mathit{flu}(\mathsf{X}) \mid \mathit{symptomB}(\mathsf{X}))$ using *noisy-or* for combining the probabilities of $\mathit{flu}(\mathsf{X})$ when both symptoms A and B are present yields a higher probability of X having a flu compared to considering both clauses by themselves. But not in every scenario the use of *noisy-or* is reasonable. Consider the clauses $(\mathit{flies}(\mathsf{X}) \mid \mathit{bird}(\mathsf{X}))$ (with a probability of 0.9 that $\mathit{flies}(\mathsf{X})$ is true given that $\mathit{bird}(\mathsf{X})$ is true) and $(\mathit{flies}(\mathsf{X}) \mid \mathit{penguin}(\mathsf{X}))$ (with a probability of 0.01 that $\mathit{flies}(\mathsf{X})$ is true given that $\mathit{penguin}(\mathsf{X})$ is true), cf. [11]. Using *noisy-or* as combining rules for *flies* yields a probability of approximately 0.476 for an actual penguin-bird, a much more higher probability than intended. It is up to the knowledge engineer to specify the right combining rule for each predicate. The approach discussed in this paper does not suffer from this problem as probabilities are implicitly combined when determining the model with maximum entropy. As for propositional probabilistic reasoning under maximum entropy [20] the relational extension discussed here satisfies the same properties in this examples for non-monotonic reasoning, cf. [36].

Another major difference between the approaches as pursued in this paper and most approaches to statistical relational learning such as MLNs and BLPs is that our approaches provide explicit model theories that give a clear logical foundation for defining inductive inference. Both MLNs and BLPs do not allow for multiple models of a knowledge base and thus restrict inference to this single model. Albeit we used inference based on the principle of maximum entropy throughout the paper it is possible to apply other paradigms. Furthermore, the explicit model theories of our approaches allow for a more flexible way to analyze problems of consistency such as inconsistency measurements [34].

6 Conclusion and Discussion

Probabilistic inference based on the principle of maximum entropy has proven to be a powerful reasoning method in propositional frameworks for knowledge representation and reasoning [27, 20]. Indeed, this principle has been characterized as an optimal inference method in various frameworks and there exist several axiomatic derivations, cf. e. g. [33, 28, 20]. In the past ten years a lot of work has been done on lifting propositional models for probabilistic reasoning to the relational case [14] so it seems natural to investigate the possibilities of applying the principle of maximum entropy on relational settings. Early attempts on doing so had been made during the 90s by Grove, Halpern, and Koller [18, 16, 17] and they have shown that probabilistic reasoning in first-order logic is—in general—problematic. While employing general first-order logic restricted to unary predicates yielded satisfactory results [17] the general case of non-unary predicates led to unintuitive properties and intractable inferences [16]. These results stem mainly from the employment of full first-order logic using an infinite universe and the statistical interpretation of probabilities. Our approaches rely on a finite universe and an interpretation of probabilities with subjective degrees of belief and, consequently, we do not face these problems. But restricting the language to finite universes and function-free signatures may seem as a major drawback that renders the framework in fact propositional. While this is true from a conceptual point of view and particularly in comparison to the works [16, 17] the motivation of

the presented work is a different one, namely knowledge representation and reasoning for rational agents that are situated in real-world environments. Clearly, for these agents considering a finite universe is not only justifiable but highly recommendable. Although our language restrictions might seem severe the framework is capable of performing commonsense reasoning tasks in relational environments as was shown in this paper. Furthermore, in contrast to other works on applying ME-inference in relational settings [24, 13] our approaches do not feature a direct propositional correspondent and thus cannot be modeled with existing propositional frameworks in a concise way. One the one hand this is a drawback as we cannot employ existing reasoner for propositional ME-inference [32]. On the other hand this shows the advantage of our approaches. Although employing a rather restricted first-order language our semantical proposals clearly extend the expressive power of other approaches and allow for the representation of complex interrelationships between different pieces of knowledge.

In this paper, we devised a set of desirable properties of inference operators and investigated two different probabilistic semantics for relational conditional logic. Both operators fulfill (in principle) the catalogue of desired properties so the question remains which semantics and therefore which inference operator is the more favorable choice? Or even are there other reasonable possibilities for semantics and inference operators that should be investigated? The second question cannot be answered right now as none of the proposed two inference operators can be characterized by our desired principles for reasoning. To do so other principles have to be found that may fully characterize ME-inference in relational settings like e. g. [33] for propositional frameworks. As for the first question, from a computational point of view the operator $\mathcal{I}_{\odot}$ and thus the semantics $\models_{\odot}$ seems to be the favorable choice for reasoning in first-order conditional logic. While (4.1) describes a non-convex optimization problem that is hard to solve in practice (4.35) induces a convex optimization problem for which efficient algorithms are available [5]. Still, a straightforward implementation of both problems yields an exponential transformation due to the exponential number of Herbrand interpretations. We implemented both inference operators in Java and employed the free optimization software OpenOpt[3] to solve the optimization problems. Performing inference using these prototypical implementations took from hours up to days for all but the smallest examples. Part of future work is the development of efficient and approximate algorithms for inference based on the proposed semantics. Recently, *lifted inference* [30, 25] has been introduced in order to efficiently perform inference in relational probabilistic settings. This approach aims at avoiding redundant computations of terms that turn out to be equal due to similar inner structures. Although these techniques have been developed for approaches that extend undirected graphical models for probabilistic reasoning an adaptation of the ideas for our framework might be reasonable. The satisfaction of the property *Prototypical Indifference* of both operators shows that there is a lot of redundant information in a complete specification of a ME-distribution that might be exploited by efficient reasoning algorithms.

Acknowledgements. The research reported here was partially supported by the Deutsche Forschungsgemeinschaft (DFG, grant KE 1413/2-1). We also thank Adam

[3] http://openopt.org/

Chachaj for implementing prototypes of the presented inference operators that were used in computing the examples in this paper.

References

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] Fahiem Bacchus, Adam J. Grove, Joseph Y. Halpern, and Daphne Koller. From Statistical Knowledge Bases to Degrees of Belief. *Artificial Intelligence*, 87(1–2):75–143, 1996.

[3] John A. Beachy and William D. Blair. *Abstract Algebra*. Waveland Press, Inc., Long Grove, Illinois, USA, third edition, 2005.

[4] Jakob Bernoulli. Usum & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis. In *Ars Conjectandi*, chapter 4. 1713.

[5] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[6] John S. Breese. Construction of Belief and Decision Networks. *Computational Intelligence*, 8(4):624–647, 1992.

[7] Luc De Raedt, Paolo Frasconi, Kristian Kersting, and Stephen H. Muggleton, editors. *Probabilistic Inductive Logic Programming: Theory and Applications*, volume 4911 of *Lecture Notes in Computer Science*. Springer, 2008.

[8] Luc De Raedt and Kristian Kersting. Probabilistic inductive logic programming. In *[7]*, pages 1–27. Springer, 2008.

[9] James P. Delgrande. On First-Order Conditional Logics. *Artificial Intelligence*, 105(1–2):105–137, 1998.

[10] Daan Fierens. *Learning Directed Probabilistic Logical Models From Relational Data*. PhD thesis, Katholieke Universiteit Leuven, 2008.

[11] Marc Finthammer and Matthias Thimm. An Integrated Development Environment for Probabilistic Relational Reasoning. *This volume*, 2011.

[12] Jens Fisseler. Toward Markov Logic with Conditional Probabilities. In David C. Wilson and H. Chad Lane, editors, *Proceedings of the Twenty-First International FLAIRS Conference*, pages 643–648. AAAI Press, 2008.

[13] Jens Fisseler. *Learning and Modeling with Probabilistic Conditional Logic*, volume 328 of *Dissertations in Artificial Intelligence*. IOS Press, 2010.

[14] Lise Getoor and Ben Taskar, editors. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[15] Adam J. Grove, Joseph Y. Halpern, and Daphne Koller. Random Worlds and Maximum Entropy. *Journal of Artificial Intelligence Research (JAIR)*, 2:33–88, 1994.

[16] Adam J. Grove, Joseph Y. Halpern, and Daphne Koller. Asymptotic conditional probabilities: The non-unary case. *The Journal of Symbolic Logic*, 61(1):250–276, 1996.

[17] Adam J. Grove, Joseph Y. Halpern, and Daphne Koller. Asymptotic conditional probabilities: The unary case. *SIAM Journal on Computing*, 25(1):1–51, February 1996.

[18] Joseph Y. Halpern. An Analysis of First-Order Logics of Probability. *Artificial Intelligence*, 46:311–350, 1990.

[19] Manfred Jaeger. Relational Bayesian Networks: a Survey. *Electronic Transactions in Artificial Intelligence*, 6, 2002.

[20] Gabriele Kern-Isberner. *Conditionals in Nonmonotonic Reasoning and Belief Revision*. Number 2087 in Lecture Notes in Computer Science. Springer, 2001.

[21] Gabriele Kern-Isberner and Thomas Lukasiewicz. Combining probabilistic logic programming with the power of maximum entropy. *Artificial Intelligence, Special Issue on Nonmonotonic Reasoning*, 157(1-2):139–202, 2004.

[22] Gabriele Kern-Isberner and Matthias Thimm. Novel semantical approaches to relational probabilistic conditionals. In *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR'10)*, Toronto, Canada, May 2010.

[23] Sebastian Loh, Matthias Thimm, and Gabriele Kern-Isberner. Grounding techniques for first-order probabilistic conditional logic, 2010. (in preparation).

[24] Thomas Lukasiewicz and Gabriele Kern-Isberner. Probabilistic logic programming under maximum entropy. In *Proceedings ECSQARU-99*, volume 1638, pages 279–292. Springer Lecture Notes in Artificial Intelligence, 1999.

[25] Brian Milch, Luke S. Zettlemoyer, Kristian Kersting, Michael Haimes, and Leslie Pack Kaelbling. Lifted probabilistic inference with counting formulas. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 2008.

[26] Donald Nute and Charles Cross. Conditional logic. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic*, volume 4, pages 1–98. Kluwer, 2002.

[27] Jeff Paris. *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge University Press, 1994.

[28] Jeff B. Paris and Alena Vencovská. In defence of the maximum entropy inference process. *International Journal of Approximate Reasoning*, 17(1):77–103, 1997.

[29] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1998.

[30] David Poole. First-Order Probabilistic Inference. In Georg Gottlob and Toby Walsh, editors, *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 985–991. Morgan Kaufmann, 2003.

[31] Wilhelm Rödder. Conditional logic and the principle of entropy. *Artificial Intelligence*, 117:83–106, 2000.

[32] Wilhelm Rödder and Carl-Heinz Meyer. Coherent Knowledge Processing at Maximum Entropy by SPIRIT. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 470–476, 1996.

[33] John E. Shore and Rodney W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26:26–37, 1980.

[34] Matthias Thimm. Measuring inconsistency in probabilistic knowledge bases. In Jeff Bilmes and Andrew Ng, editors, *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI'09)*, Montreal, Canada, June 2009.

[35] Matthias Thimm. Representing statistical information and degrees of belief in first-order probabilistic conditional logic. In *Workshop on Relational Approaches to Knowledge Representation and Learning, Proceedings*, pages 49–63, 2009.

[36] Matthias Thimm, Gabriele Kern-Isberner, and Jens Fisseler. Relational Probabilistic Conditional Reasoning at Maximum Entropy. In *Proceedings of the Eleventh European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'11)*, 2011.

[37] Michael P. Wellman, John S. Breese, and Robert P. Goldman. From Knowledge Bases to Decision Models. *The Knowledge Engineering Review*, 7(1):35–53, 1992.

A Proofs of Results

PROPOSITION 2.10
Let $P \in \mathsf{Prob}$ be a probability distribution, $D \subseteq U$, and $(B)[\alpha] \in (\mathcal{L}_\Sigma \,|\, \mathcal{L}_\Sigma)^{prob}$ a probabilistic fact. Then it holds that $P, D \models_{\varnothing} (B)[\alpha]$ iff $P, D \models_{\odot} (B)[\alpha]$.

PROOF. It holds $P, D \models_{\varnothing} (B)[\alpha]$ iff

$$\frac{\sum_{B' \in \mathsf{ground}_D(B)} P(B')}{|\mathsf{ground}_D(B)|} = \alpha$$

by definition and furthermore due to $P(\top) = 1$ it holds

$$\begin{aligned} P, D \models_{\odot} (B)[\alpha] \quad &\Leftrightarrow \quad \frac{\sum_{B' \in \mathsf{ground}_D(B)} P(B')}{\sum_{B' \in \mathsf{ground}_D(B)} P(\top)} = \alpha \\ &\Leftrightarrow \quad \frac{\sum_{B' \in \mathsf{ground}_D(B)} P(B')}{|\mathsf{ground}_D(B)|} = \alpha \end{aligned}$$

∎

LEMMA 2.12
Let $n \in \mathbb{N}^+$ be some positive integer and let $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n \in (0, 1]$ with $\alpha_i \leq \beta_i$ for all $i = 1, \ldots, n$. Then

$$\left| \frac{\alpha_1/\beta_1 + \ldots + \alpha_n/\beta_n}{n} - \frac{\alpha_1 + \ldots + \alpha_n}{\beta_1 + \ldots + \beta_n} \right| \quad < \quad \frac{n-1}{n} \tag{6.1}$$

PROOF. For reasons of simplicity we give the proof only for $n = 2$ but the approach is the same for $n > 2$. We have to show that

$$-\frac{1}{2} < \frac{\alpha_1/\beta_1 + \alpha_2/\beta_2}{2} - \frac{\alpha_1 + \alpha_2}{\beta_1 + \beta_2} < \frac{1}{2}$$

Consider first

$$\begin{aligned} & \frac{\alpha_1/\beta_1 + \alpha_2/\beta_2}{2} - \frac{\alpha_1 + \alpha_2}{\beta_1 + \beta_2} < \frac{1}{2} \\ \Leftrightarrow \quad & \frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2} - \frac{2\alpha_1 + 2\alpha_2}{\beta_1 + \beta_2} < 1 \\ \Leftrightarrow \quad & \alpha_1\beta_1\beta_2 + \alpha_1\beta_2^2 + \alpha_2\beta_1^2 + \alpha_2\beta_1\beta_2 - 2\alpha_1\beta_1\beta_2 - 2\alpha_2\beta_1\beta_2 < \beta_1^2\beta_2 + \beta_1\beta_2^2 \\ \Leftrightarrow \quad & \alpha_1\beta_1\beta_2 + \alpha_2\beta_1\beta_2 + \beta_1^2\beta_2 + \beta_1\beta_2^2 - \alpha_1\beta_2^2 - \alpha_2\beta_1^2 > 0 \\ \Leftrightarrow \quad & \alpha_1\beta_1\beta_2 + \alpha_2\beta_1\beta_2 + \beta_1^2(\underbrace{\beta_2 - \alpha_2}_{x_1}) + \beta_2^2(\underbrace{\beta_1 - \alpha_1}_{x_2}) > 0 \end{aligned}$$

Due to $\alpha_1 \leq \beta_1$ and $\alpha_2 \leq \beta_2$ it follows $x_1, x_2 \geq 0$. Due to the strict positivity of all α_i, β_i $(i = 1, \ldots, n)$ the above inequality is true. For the other side we assume the

contrary. Consider

$$\begin{aligned} & \frac{\alpha_1/\beta_1 + \alpha_2/\beta_2}{2} - \frac{\alpha_1+\alpha_2}{\beta_1+\beta_2} \leq -\frac{1}{2} \\ \Leftrightarrow \quad & \alpha_1\beta_2^2 + \alpha_2\beta_1^2 + \beta_1^2\beta_2 + \beta_1\beta_2^2 - \alpha_1\beta_1\beta_2 - \alpha_2\beta_1\beta_2 \leq 0 \end{aligned}$$

Due to $\alpha_1 \leq \beta_1$ and $\alpha_2 \leq \beta_2$ it follows

$$\begin{aligned} \Rightarrow \quad & \alpha_1\beta_2^2 + \alpha_2\beta_1^2 + \beta_1^2\beta_2 + \beta_1\beta_2^2 - \beta_1^2\beta_2 - \beta_1\beta_2^2 \leq 0 \\ \Leftrightarrow \quad & \alpha_1\beta_2^2 + \alpha_2\beta_1^2 \leq 0 \end{aligned}$$

which is a contradiction since $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$. ∎

Corollary 2.14
Let P be some probability distribution, $D \subseteq U$, and $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))$ be some conditional. If $P, D \models_{\varnothing} (B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha_1]$ and $P, D \models_{\odot} (B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha_2]$ then

$$|\alpha_1 - \alpha_2| < \frac{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))| - 1}{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))|}$$

Proof. This follows directly from Lemma 2.12 and the fact that P both $\varnothing$- and $\odot$-satisfies $(B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha]$ under D for some α (therefore all appearing probabilities of premises are non-zero). ∎

Proposition 3.3
If $\mathcal{I}$ satisfies *Name Irrelevance* then $\mathcal{I}$ satisfies *Prototypical Indifference*.

Proof. Let $\mathcal{R}$ be a knowledge base and $\mathsf{d}_1, \mathsf{d}_2 \in U \setminus D$. Let furthermore $\mathsf{c}_1, \mathsf{c}_2 \in D$ with $\mathsf{c}_1 \equiv_{\mathcal{R}} \mathsf{c}_2$ and $\mathsf{c}_1 \neq \mathsf{c}_2$. Then it holds for a ground sentence A:

$$\begin{aligned} \mathcal{I}(\mathcal{R}, D)(A) &= \mathcal{I}(\mathcal{R}[\mathsf{d}_1/\mathsf{c}_1], (D \cup \{\mathsf{d}_1\}) \setminus \{\mathsf{c}_1\})(A[\mathsf{d}_1/\mathsf{c}_1]) \\ &= \mathcal{I}(\mathcal{R}[\mathsf{d}_1/\mathsf{c}_1, \mathsf{d}_2/\mathsf{c}_2], (D \cup \{\mathsf{d}_1, \mathsf{d}_2\}) \setminus \{\mathsf{c}_1, \mathsf{c}_2\})(A[\mathsf{d}_1/\mathsf{c}_1, \mathsf{d}_2/\mathsf{c}_2]) \end{aligned}$$

As $\mathsf{c}_1, \mathsf{c}_2 \notin (D \cup \{\mathsf{d}_1, \mathsf{d}_2\}) \setminus \{\mathsf{c}_1, \mathsf{c}_2\}$ it holds that

$$\begin{aligned} & \mathcal{I}(\mathcal{R}[\mathsf{d}_1/\mathsf{c}_1, \mathsf{d}_2/\mathsf{c}_2], (D \cup \{\mathsf{d}_1, \mathsf{d}_2\}) \setminus \{\mathsf{c}_1, \mathsf{c}_2\})(A[\mathsf{d}_1/\mathsf{c}_1, \mathsf{d}_2/\mathsf{c}_2]) \\ = \; & \mathcal{I}(\mathcal{R}[\mathsf{d}_1/\mathsf{c}_1, \mathsf{d}_2/\mathsf{c}_2][\mathsf{c}_2/\mathsf{d}_1, \mathsf{c}_1/\mathsf{d}_2], (((D \cup \{\mathsf{d}_1, \mathsf{d}_2\}) \setminus \{\mathsf{c}_1, \mathsf{c}_2\}) \cup \{\mathsf{c}_1, \mathsf{c}_2\}) \setminus \{\mathsf{d}_1, \mathsf{d}_2\}) \\ & (A[\mathsf{d}_1/\mathsf{c}_1, \mathsf{d}_2/\mathsf{c}_2][\mathsf{c}_2/\mathsf{d}_1, \mathsf{c}_1/\mathsf{d}_2]) \quad . \end{aligned}$$

Due to

$$\begin{aligned} \mathcal{R}[\mathsf{d}_1/\mathsf{c}_1, \mathsf{d}_2/\mathsf{c}_2][\mathsf{c}_2/\mathsf{d}_1, \mathsf{c}_1/\mathsf{d}_2] = \mathcal{R}[\mathsf{c}_2/\mathsf{c}_1, \mathsf{c}_1/\mathsf{c}_2] &= \mathcal{R} \\ (((D \cup \{\mathsf{d}_1, \mathsf{d}_2\}) \setminus \{\mathsf{c}_1, \mathsf{c}_2\}) \cup \{\mathsf{c}_1, \mathsf{c}_2\}) \setminus \{\mathsf{d}_1, \mathsf{d}_2\} &= D \end{aligned}$$

and

$$A[\mathsf{d}_1/\mathsf{c}_1, \mathsf{d}_2/\mathsf{c}_2][\mathsf{c}_2/\mathsf{d}_1, \mathsf{c}_1/\mathsf{d}_2] = A[\mathsf{c}_1/\mathsf{c}_2, \mathsf{c}_2/\mathsf{c}_1]$$

this yields $\mathcal{I}(\mathcal{R}, D)(A) = \mathcal{I}(\mathcal{R}, D)(A[\mathsf{c}_1/\mathsf{c}_2, \mathsf{c}_2/\mathsf{c}_1])$. ∎

PROPOSITION 3.4
Let $\mathcal{I}$ satisfy *Prototypical Indifference*. Let $\mathcal{R}$ be a knowledge base on $\mathcal{L}_\Sigma$ and $D \subseteq U$.

1. Let G_1, G_2 be two ground sentences. For $\mathsf{c}_1, \mathsf{c}_2 \in D$ with $\mathsf{c}_1 \equiv_\mathcal{R} \mathsf{c}_2$ it holds $\mathcal{I}(\mathcal{R}, D)(G_2 \,|\, G_1) = \mathcal{I}(\mathcal{R}, D)(G_2[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2] \,|\, G_1[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2])$.
2. Let $S \in \mathcal{S}_\mathcal{R}$, $\mathsf{c}_1, \ldots, \mathsf{c}_\mathsf{n} \in S$, and $\sigma : S \to S$ be a permutation on S, i. e. a bijective function on S. Then it holds $\mathcal{I}(\mathcal{R}, D)(A) = \mathcal{I}(\mathcal{R}, D)(A[\sigma(\mathsf{c}_1)/\mathsf{c}_1, \ldots, \sigma(\mathsf{c}_\mathsf{n})/\mathsf{c}_\mathsf{n}])$.

PROOF.

1. Because of *Prototypical Indifference* it holds that

$$\begin{aligned} \mathcal{I}(\mathcal{R}, D)(A_2) &= \mathcal{I}(\mathcal{R}, D)(A_2[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2]) \quad \text{and} \\ \mathcal{I}(\mathcal{R}, D)(A_1 \wedge A_2) &= \mathcal{I}(\mathcal{R}, D)((A_1 \wedge A_2)[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2]) \end{aligned}$$

   and hence

$$\begin{aligned} \mathcal{I}(\mathcal{R}, D)(G_2 \,|\, G_1) &= \frac{\mathcal{I}(\mathcal{R}, D)(G_2 \wedge G_1)}{\mathcal{I}(\mathcal{R}, D)(G_1)} \\ &= \frac{\mathcal{I}(\mathcal{R}, D)((G_2 \wedge G_1)[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2])}{\mathcal{I}(\mathcal{R}, D)(G_1[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2])} \\ &= \mathcal{I}(\mathcal{R}, D)(G_2[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2] \,|\, G_1[\mathsf{c}_1 \leftrightarrow \mathsf{c}_2]) \end{aligned}$$

   due to $(G_2 \wedge G_1)[\mathsf{x}_i/\mathsf{y}_i]_{i=1,\ldots,n} = G_2[\mathsf{x}_i/\mathsf{y}_i]_{i=1,\ldots,n} \wedge G_1[\mathsf{x}_i/\mathsf{y}_i]_{i=1,\ldots,n}$.
2. This follows from the fact that every permutation can be represented as a product of transpositions [3], i. e. permutations that exactly transpose two elements. Let $\sigma_1, \ldots, \sigma_m$ be these transpositions of σ and let $\sigma_{1\ldots i} = \sigma_i \circ \ldots \circ \sigma_1$ for $i = 1, \ldots, m$. Note, that $\sigma_{1\ldots 1} = \sigma_1$ and $\sigma_{1\ldots m} = \sigma$. Due to *Prototypical Indifference* it holds

$$\mathcal{I}(\mathcal{R}, D)(A) = \mathcal{I}(\mathcal{R}, D)(A[\sigma_1(\mathsf{c}_1)/\mathsf{c}_1, \ldots, \sigma_1(\mathsf{c}_\mathsf{n})/\mathsf{c}_\mathsf{n}])$$

   and for any $i = 2, \ldots, m$ it holds that

$$\begin{aligned} &\mathcal{I}(\mathcal{R}, D)(A[\sigma_{1\ldots i-1}(\mathsf{c}_1)/\mathsf{c}_1, \ldots, \sigma_{1\ldots i-1}(\mathsf{c}_\mathsf{n})/\mathsf{c}_\mathsf{n}]) \\ &\quad = \mathcal{I}(\mathcal{R}, D)(A[\sigma_{1\ldots i}(\mathsf{c}_1)/\mathsf{c}_1, \ldots, \sigma_{1\ldots i}(\mathsf{c}_\mathsf{n})/\mathsf{c}_\mathsf{n}]) \quad . \end{aligned}$$

   Via transitivity and $\sigma_{1\ldots m} = \sigma$ it follows

$$\mathcal{I}(\mathcal{R}, D)(A) = \mathcal{I}(\mathcal{R}, D)(A[\sigma(\mathsf{c}_1)/\mathsf{c}_1, \ldots, \sigma(\mathsf{c}_\mathsf{n})/\mathsf{c}_\mathsf{n}]) \quad .$$

∎

PROPOSITION 4.1
The inference operator $\mathcal{I}_\varnothing$ satisfies *Name Irrelevance*, *Prototypical Indifference*, ME-*Compatibility*, *Compensation*, and *Strict Inference*. If $\mathcal{I}_\varnothing$ satisfies *Well-Definedness* then $\mathcal{I}_\varnothing$ satisfies *Conditional Probability in the Limit*.

PROOF.

(Name Irrelevance) This is obvious as the principle of maximum entropy is unbiased to renaming of constants, cf. [33].

(Prototypical Indifference) This follows from Proposition 3.2.

(ME-Compatibility) Let $\mathcal{R}$ be a ground knowledge base. Due to Remark 2 the operator $\models_{\varnothing}$ is equivalent to $\models$ in the propositional case. Then Equation (3) also becomes equivalent to the propositional case and is in particular uniquely solvable. Hence, it holds that $\mathsf{ME}(\mathcal{R}')(A) = \mathcal{I}_{\varnothing}(\mathcal{R}, \mathsf{consts}(\mathcal{R}))(A)$ for any ground sentence A.

(Compensation) Let $\mathcal{R}$ be a knowledge base and $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$ a non-ground conditional with $\alpha \in (0,1)$. Suppose $\mathcal{I}_{\varnothing}(\mathcal{R}, D)(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) < \alpha$ for all $(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}}))[\alpha] \in \mathsf{ground}_D(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}}))$. Then it holds that

$$\frac{\sum_{(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}}))}{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))|} < \frac{\alpha \cdot |\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))|}{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))|} = \alpha$$

contradicting $\mathcal{I}_{\varnothing}(\mathcal{R}, D), D \models_{\varnothing} \mathcal{R}$.

(Strict Inference) Let $\mathcal{R}$ be a knowledge base and $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$ a non-ground conditional with $\alpha = 1$ (the case of $\alpha = 0$ can be shown analogously). Suppose $\mathcal{I}_{\varnothing}(\mathcal{R}, D)(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) < 1$ for some $(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}}))[\alpha] \in \mathsf{ground}_D(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}}))$. Then it holds that

$$\frac{\sum_{(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}}))}{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))|} < \frac{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))|}{|\mathsf{ground}_D(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))|} = 1$$

contradicting $\mathcal{I}_{\varnothing}(\mathcal{R}, D), D \models_{\varnothing} \mathcal{R}$.

(Conditional Probability in the Limit) Assume that the inference operator $\mathcal{I}_{\varnothing}$ satisfies *Well-Definedness*. Let $\mathcal{R}$ be a knowledge base on $\mathcal{L}_{\Sigma}$ such that $P^* := \mathcal{I}_{\varnothing}(\mathcal{R}, D) \neq \mathsf{undef}$. Let $r = (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$ be a conditional with $\vec{\mathsf{X}} = (\mathsf{X}_1, \ldots, \mathsf{X}_h)$ and $\mathsf{consts}(\mathcal{R}) = \{\mathsf{c}_1, \ldots, \mathsf{c}_n\}$. Let furthermore $\{\mathsf{d}_1, \ldots, \mathsf{d}_m\} = D \setminus \{\mathsf{c}_1, \ldots, \mathsf{c}_n\}$, so it holds that $|D| = n+m$. Let $\vec{\mathsf{d}}_1, \ldots, \vec{\mathsf{d}}_k$ be all vectors of constants of $\mathsf{d}_1, \ldots, \mathsf{d}_m$ with length h such that for any $\vec{\mathsf{d}}_i$ with $1 \leq i \leq k$ no two elements are the same. Let $\vec{\mathsf{c}}_1, \ldots, \vec{\mathsf{c}}_l$ be all remaining vectors of constants in D. It follows that $(l+k) = (|D|)^h = (n+m)^h$ and

$$k = m^{\underline{h}} = m(m-1)\ldots(m-h+1) \quad \text{(the falling factorial)}$$

and thus $l = (n+m)^h - m^{\underline{h}}$. Let $P^*_{\vec{\mathsf{c}}}$ denote $P^*(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}}))$ for a vector $\vec{\mathsf{c}}$. In order to have $P^*, D \models_{\varnothing} r$ it must hold $P^*_{\vec{\mathsf{c}}_1} + \ldots + P^*_{\vec{\mathsf{c}}_l} + P^*_{\vec{\mathsf{d}}_1} + \ldots + P^*_{\vec{\mathsf{d}}_k} = \alpha \cdot (k+l)$. From *Prototypical Indifference* and Proposition 3.3 it follows that $P^*_{\vec{\mathsf{d}}_1} = \ldots = P^*_{\vec{\mathsf{d}}_k}$.

Define $P_k^* := P^*_{\vec{\mathsf{d}}_1}$, so it holds that $P^*_{\vec{\mathsf{d}}_1} + \ldots + P^*_{\vec{\mathsf{d}}_k} = kP_k^*$. It follows

$$\begin{aligned} P_k^* &= \frac{\alpha \cdot (k+l) - P^*_{\vec{\mathsf{c}}_1} - \ldots - P^*_{\vec{\mathsf{c}}_l}}{k} \\ &\leq \frac{\alpha \cdot (k+l)}{k} \\ &= \alpha \underbrace{\frac{(n+m)^h}{m^{\underline{h}}}}_{\overset{m\to\infty}{\to} 1} \\ &\overset{m\to\infty}{\to} \alpha \end{aligned}$$

Similarly it holds

$$\begin{aligned} P_k^* &= \frac{\alpha \cdot (k+l) - P^*_{\vec{\mathsf{c}}_1} - \ldots - P^*_{\vec{\mathsf{c}}_l}}{k} \\ &\geq \frac{\alpha \cdot (k+l) - l}{k} \\ &= \alpha \underbrace{\frac{(n+m)^h}{m^{\underline{h}}}}_{\overset{m\to\infty}{\to} 1} - \underbrace{\frac{(n+m)^h - m^{\underline{h}}}{m^{\underline{h}}}}_{\overset{m\to\infty}{\to} 0} \\ &\overset{m\to\infty}{\to} \alpha \quad . \end{aligned}$$

Due to *Well-Definedness* the probability distributions P_k^* are well-defined for any k and it follows $P_k^* \to \alpha$ for $m \to \infty$.

∎

Lemma 4.5
Let $r = (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$ be a probabilistic conditional. Then $Mod^D_{\odot}(\{r\})$ is convex for any D with $\mathsf{consts}(r) \subseteq D$.

Proof. Let $D \subseteq U$ and let P_1 and P_2 be some probability distributions with $P_1, D \models_{\odot} r$ and $P_2, D \models_{\odot} r$. We have to show that any convex combination of P_1 and P_2 satisfies r as well. Let Q be a convex combination of P_1 and P_2, i. e. let $\delta \in (0,1)$ fixed and define $Q(\omega) = \delta P_1(\omega) + (1-\delta)P_2(\omega)$ for any $\omega \in \Omega$. Then it holds $Q(A) = \delta P_1(A) + (1-\delta)P_2(A)$ for any ground formula A as well. Let $\{(B_1 \mid A_1), \ldots, (B_n \mid A_n)\} = \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))$. Then it holds that

$$\frac{\sum_{i=1}^n P_j(B_i A_i)}{\sum_{i=1}^n P_j(A_i)} = \alpha \tag{6.2}$$

for $j = 1, 2$ and we have to show that

$$\frac{\sum_{i=1}^n Q(B_i A_i)}{\sum_{i=1}^n Q(A_i)} = \alpha \tag{6.3}$$

which is equivalent to

$$\frac{\delta \sum_{i=1}^{n} P_1(B_i A_i) + (1-\delta) \sum_{i=1}^{n} P_2(B_i A_i)}{\delta \sum_{i=1}^{n} P_1(A_i) + (1-\delta) \sum_{i=1}^{n} P_2(A_i)} = \alpha \tag{6.4}$$

If $\alpha = 0$ then $P_j(B_i \wedge A_i) = 0$ for all $i = 1, \ldots, n$ and $j = 1, 2$ due to (6.2). Then also $Q(B_i \wedge A_i) = 0$ for all $i = 1 \ldots, n$ and it follows $Q, D \models_{\odot} r$. We continue for $\alpha > 0$. Then (6.4) is equivalent to

$$\begin{aligned}
& \delta \sum_{i=1}^{n} P_1(B_i A_i) + (1-\delta) \sum_{i=1}^{n} P_2(B_i A_i) = \alpha\delta \sum_{i=1}^{n} P_1(A_i) + \alpha(1-\delta) \sum_{i=1}^{n} P_2(A_i) \\
\Leftrightarrow \quad & \frac{\delta \sum_{i=1}^{n} P_1(B_i A_i)}{\alpha\delta \sum_{i=1}^{n} P_1(A_i)} + \frac{(1-\delta) \sum_{i=1}^{n} P_2(B_i A_i)}{\alpha\delta \sum_{i=1}^{n} P_1(A_i)} = 1 + \frac{\alpha(1-\delta) \sum_{i=1}^{n} P_2(A_i)}{\alpha\delta \sum_{i=1}^{n} P_1(A_i)} \\
\Leftrightarrow \quad & 1 + \frac{(1-\delta) \sum_{i=1}^{n} P_2(B_i A_i)}{\alpha\delta \sum_{i=1}^{n} P_1(A_i)} = 1 + \frac{(1-\delta) \sum_{i=1}^{n} P_2(A_i)}{\delta \sum_{i=1}^{n} P_1(A_i)} \\
\Leftrightarrow \quad & \frac{1}{\alpha}(1-\delta) \sum_{i=1}^{n} P_2(B_i \wedge A_i) = (1-\delta) \sum_{i=1}^{n} P_2(A_i) \\
\Leftrightarrow \quad & 1 = 1
\end{aligned}$$

and it follows $Q, D \models_{\odot} r$. ∎

PROPOSITION 4.6
Let $\mathcal{R}$ be a $\odot$-consistent knowledge base and D a set $D \subseteq U$. Then $\mathcal{I}_{\odot}(\mathcal{R}, D)$ is uniquely determined.

PROOF. For any knowledge base $\mathcal{R}$ the set of probability distributions that satisfy $\mathcal{R}$ is a convex set due to Lemma 4.6 and the fact that the intersection of two convex sets is again a convex set. The entropy is a strict concave function and maximization of a strict concave function over a convex set has a unique solution [5]. ∎

PROPOSITION 4.7
$\mathcal{I}_{\odot}$ satisfies *Well-Definedness*, *Name Irrelevance*, *Prototypical Indifference*, ME-*Compatibility*, *Conditional Probability in the Limit*, *Strict Inference*, and *Compensation*.

PROOF.

(Well-Definedness) This is true due to Proposition 4.7.

(Name Irrelevance) This is obvious as the principle of maximum entropy is unbiased to renaming of constants, cf. [33].

(Prototypical Indifference) This follows directly from Proposition 3.2.

(ME-Compatibility) For ground conditional knowledge bases $\mathcal{R}$, the semantics is the same as for the propositional case, so $\mathsf{ME}(\mathcal{R}) = \mathcal{I}_{\odot}(\mathcal{R}, \mathsf{consts}(\mathcal{R}))$.

(Conditional Probability in the Limit) Let $D \subseteq U$, $\mathcal{R}$ be a relational conditional knowledge base, and $C_0 = D \setminus \mathsf{consts}(R)$. Let $r = (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$ be a relational conditional in $\mathcal{R}$ with free variables, and let $r_g = (B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}}))[\alpha]$ be a proper instantiation of r with constants $\vec{\mathsf{c}}$ from C_0. Let $D_n = \mathsf{consts}(\mathcal{R}) \cup C_0 \cup C_n$ with $C_n \subset C_{n+1}$ and $|C_0 \cup C_n| = n$, $n \in \mathbb{N}, n \geq |C_0|$, be a sequence of sets of constants with $C_n \subseteq U$ for all $n \in \mathbb{N}$. Let $P_n^* = \mathcal{I}_{\odot}(\mathcal{R}, D_n)$ be the ME-distribution of $\mathcal{R}$ that takes the constants from D_n into account. Due to *Prototypical Indifference*, any constant from C_0 can be replaced by any constant from C_n when calculating $P_n^*(B(\vec{\mathsf{a}})|A(\vec{\mathsf{a}}))$, since neither of them appears in $\mathcal{R}$. So, all probabilities of instantiations $P_n^*(B(\vec{\mathsf{a}})|A(\vec{\mathsf{a}}))$ are given by instantiations over $\mathsf{consts}(\mathcal{R}) \cup C_0$, but we have to take proper multiplicities into regard. Let $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))$ have arity s, and let $(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}}))$ be a proper instantiation. Then $\vec{\mathsf{a}}$ is a vector of arity s that might have m components from $E_n = C_0 \cup C_n$, $0 \leq m \leq s$, and $s - m$ components from $\mathsf{consts}(\mathcal{R})$. Without loss of generality we assume $t = |\mathsf{consts}(\mathcal{R})|, |C_0| \geq s$. Since the positions of these components can make a difference, we have $\binom{s}{m}$ non-indifferent instantiations $(B(\vec{\mathsf{a}}^m_{k_m,l_m}) \mid A(\vec{\mathsf{a}}^m_{k_m,l_m})), 1 \leq k_m \leq \binom{s}{m}, 1 \leq l_m \leq t^{s-m}$, with vectors $\vec{\mathsf{a}}^m_{k_m,l_m}$ over $\mathsf{consts}(\mathcal{R}) \cup C_0$ such that $P_n^*(B(\vec{a}^m_{k_m,l_m}) \mid A(\vec{a}^m_{k_m,l_m}))$ occurs n^m times among the instantiations over D_n. In particular, for $m = s$, all n^s instantiations over E_n are individually indifferent with respect to P_n^*, one of them being the instantiation for $\vec{\mathsf{c}}$, so $P_n^*(B(\vec{\mathsf{c}})|A(\vec{\mathsf{c}}))$ can serve as a representative for these n^s probabilities. Similar statements hold for all instantiations of $P_n^*(A(\vec{\mathsf{X}})B(\vec{\mathsf{X}}))$ and $P_n^*(A(\vec{\mathsf{X}}))$. $P_n^*, D_n \models_{\odot} \mathcal{R}$, so in particular, $P_n^*, D_n \models_{\odot} r$, which means that

$$\begin{aligned} \alpha &= \frac{\sum_{(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}})) \in \mathsf{ground}_{D_n}((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P(A(\vec{\mathsf{a}})B(\vec{\mathsf{a}}))}{\sum_{(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}})) \in \mathsf{ground}_{D_n}((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P(A(\vec{\mathsf{a}}))} \\ &=: \frac{\Sigma(A(\vec{\mathsf{a}})B(\vec{\mathsf{a}}), C_n)}{\Sigma(A(\vec{\mathsf{a}}), C_n)} \quad . \end{aligned}$$

For the numerator, we obtain

$$\begin{aligned} &\Sigma(A(\vec{\mathsf{a}})B(\vec{\mathsf{a}}), D_n) \\ &= \sum_{l_0=1}^{t^s} P_n^*(A(\vec{\mathsf{a}}_{l_0})B(\vec{\mathsf{a}}_{l_0})) + n \sum_{k_1=1}^{s} \sum_{l_1=1}^{t^{s-1}} P_n^*(A(\vec{\mathsf{a}}^1_{k_1,l_1})B(\vec{\mathsf{a}}^1_{k_1,l_1})) \\ &+ n^2 \sum_{k_2=1}^{\binom{s}{2}} \sum_{l_2=1}^{t^{s-2}} P_n^*(A(\vec{\mathsf{a}}^2_{k_2,l_2})B(\vec{\mathsf{a}}^2_{k_2,l_2})) + \ldots \\ &+ n^{s-1} \sum_{k_{s-1}=1}^{\binom{s}{s-1}} \sum_{l_{s-1}=1}^{t} P_n^*(A(\vec{\mathsf{a}}^{s-1}_{k_{s-1},l_{s-1}})B(\vec{\mathsf{a}}^{s-1}_{k_{s-1},l_{s-1}})) \\ &+ n^s P_n^*(A(\vec{\mathsf{c}})B(\vec{\mathsf{c}})) \end{aligned}$$

$$
\begin{aligned}
= \quad & n^s[\ \frac{1}{n^s}\sum_{l_0=1}^{t^s} P_n^*(A(\vec{\mathsf{a}}_{l_0})B(\vec{\mathsf{a}}_{l_0})) + \frac{1}{n^{s-1}}\sum_{k_1=1}^{s}\sum_{l_1=1}^{t^{s-1}} P_n^*(A(\vec{\mathsf{a}}^1_{k_1,l_1})B(\vec{\mathsf{a}}^1_{k_1,l_1})) \\
+ \quad & \frac{1}{n^{s-2}}\sum_{k_2=1}^{\binom{s}{2}}\sum_{l_2=1}^{t^{s-2}} P_n^*(A(\vec{\mathsf{a}}^2_{k_2,l_2})B(\vec{\mathsf{a}}^2_{k_2,l_2})) + \ldots \\
+ \quad & \frac{1}{n}\sum_{k_{s-1}=1}^{\binom{s}{s-1}}\sum_{l_{s-1}=1}^{t} P_n^*(A(\vec{\mathsf{a}}^{s-1}_{k_{s-1},l_{s-1}})B(\vec{\mathsf{a}}^{s-1}_{k_{s-1},l_{s-1}})) + P_n^*(A(\vec{\mathsf{c}})B(\vec{\mathsf{c}}))\ ] \\
= \quad & n^s[\ \epsilon_1(n) + P_n^*(A(\vec{\mathsf{c}})B(\vec{\mathsf{c}}))\ ]
\end{aligned}
$$

with $\epsilon_1(n) \leq \frac{1}{n^s}t^s + \frac{1}{n^{s-1}}st^{s-1} + \frac{1}{n^{s-2}}\binom{s}{2}t^{s-2} + \ldots + \frac{1}{n}st = O(\frac{1}{n})$. In the same way, for the denominator, we have

$$\Sigma(A(\vec{\mathsf{a}}), C_n) = n^s[\epsilon_2(n) + P_n^*(A(\vec{\mathsf{c}}))]$$

with $\epsilon_2(n) = O(\frac{1}{n})$. This shows that for $n \to \infty$, $P_n^*(A(\vec{\mathsf{c}})B(\vec{\mathsf{c}}))$ and $P_n^*(A(\vec{\mathsf{c}}))$ are the dominant terms, hence

$$\alpha \quad = \quad \lim_{n\to\infty}\frac{P_n^*(A(\vec{\mathsf{c}})B(\vec{\mathsf{c}}))}{P_n^*(A(\vec{\mathsf{c}}))} = \lim_{n\to\infty} P_n^*(B(\vec{\mathsf{c}})|A(\vec{\mathsf{c}})),$$

what was to be shown.

(Compensation) Let $\mathcal{R}$ be a knowledge base and $(B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$ with $\alpha \in (0,1)$, and let $\vec{\mathsf{c}_1}$ be a vector of constants such that $\mathcal{I}_\odot(\mathcal{R}, D)(B(\vec{\mathsf{c}_1}) \,|\, A(\vec{\mathsf{c}_1})) < \alpha$. Let $P^* = \mathcal{I}_\odot(\mathcal{R}, D)$. From the presupposition $(B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$ and $P^*, D \models_\odot \mathcal{R}$, in particular, we have $P^*, D \models_\odot (B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))[\alpha]$, which means

$$\alpha = \frac{\sum_{(B(\vec{\mathsf{a}}) \,|\, A(\vec{\mathsf{a}}))\in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}})))} P^*(A(\vec{\mathsf{a}})B(\vec{\mathsf{a}}))}{\sum_{(B(\vec{\mathsf{a}}) \,|\, A(\vec{\mathsf{a}}))\in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}})))} P^*(A(\vec{\mathsf{a}}))}.$$

Assume that for all (proper) instantiations $\vec{\mathsf{a}} \neq \vec{\mathsf{c}_1}$, $P^*(B(\vec{\mathsf{a}}) \,|\, A(\vec{\mathsf{a}})) \leq \alpha$. Then we had

$$
\begin{aligned}
& \sum_{(B(\vec{\mathsf{a}}) \,|\, A(\vec{\mathsf{a}}))\in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}})))} P^*(A(\vec{\mathsf{a}})B(\vec{\mathsf{a}})) \\
= \quad & P^*(A(\vec{\mathsf{c}_1})B(\vec{\mathsf{c}_1})) + \sum_{(B(\vec{\mathsf{a}}) \,|\, A(\vec{\mathsf{a}}))\in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))), \vec{\mathsf{a}}\neq\vec{\mathsf{c}_1}} P^*(A(\vec{\mathsf{a}})B(\vec{\mathsf{a}})) \\
< \quad & \alpha P^*(A(\vec{\mathsf{c}_1})) + \sum_{(B(\vec{\mathsf{a}}) \,|\, A(\vec{\mathsf{a}}))\in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}}))), \vec{\mathsf{a}}\neq\vec{\mathsf{c}_1}} \alpha P^*(A(\vec{\mathsf{a}})) \\
= \quad & \alpha \sum_{(B(\vec{\mathsf{a}}) \,|\, A(\vec{\mathsf{a}}))\in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \,|\, A(\vec{\mathsf{X}})))} P^*(A(\vec{\mathsf{a}})),
\end{aligned}
$$

hence

$$\frac{\sum_{(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P^*(A(\vec{\mathsf{a}})B(\vec{\mathsf{a}}))}{\sum_{(B(\vec{\mathsf{a}}) \mid A(\vec{\mathsf{a}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P^*(A(\vec{\mathsf{a}}))} < \alpha,$$

which contradicts $P^*, D \models_{\odot} (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha]$. So, there must be another vector of constants $\vec{\mathsf{c}_2}$ with $P^*(A(\vec{\mathsf{c}_2}) \mid B(\vec{\mathsf{c}_2})) > \alpha$.

(Strict Inference) Let $\mathcal{R}$ be a $\circ$-consistent knowledge base under D and $(B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[\alpha] \in \mathcal{R}$ a non-ground conditional with $\alpha \in \{0, 1\}$. Let

$$(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D(A(\vec{\mathsf{X}}) \mid B(\vec{\mathsf{X}})) \quad .$$

It is to be shown that $\mathcal{I}(\mathcal{R}, D)(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) = \alpha$. Suppose that $\alpha = 0$. Since $P^* = \mathcal{I}_{\odot}(\mathcal{R}, D)$ is a model of $\mathcal{R}$, in particular, we have $P^*, D \models_{\odot} (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[0]$. This implies that

$$\sum_{(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P(B(\vec{\mathsf{c}}) \wedge A(\vec{\mathsf{c}})) = 0 \quad ,$$

so for all $(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))$, $P(B(\vec{\mathsf{c}}) \wedge A(\vec{\mathsf{c}})) = 0$. In case that $\alpha = 1$, $P^*, D \models_{\odot} (B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}}))[1]$ yields

$$\sum_{(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P(B(\vec{\mathsf{c}}) \wedge A(\vec{\mathsf{c}})) = \sum_{(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P(A(\vec{\mathsf{c}})) \quad .$$

If there were $\vec{\mathsf{c}}$ such that $P(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) < 1$, i.e. $P(B(\vec{\mathsf{c}}) \wedge A(\vec{\mathsf{c}})) < P(A(\vec{\mathsf{c}}))$, this would result in

$$\sum_{(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P(B(\vec{\mathsf{c}}) \wedge A(\vec{\mathsf{c}})) < \sum_{(B(\vec{\mathsf{c}}) \mid A(\vec{\mathsf{c}})) \in \mathsf{ground}_D((B(\vec{\mathsf{X}}) \mid A(\vec{\mathsf{X}})))} P(A(\vec{\mathsf{c}})) \quad ,$$

a contradiction. Hence, *Strict Inference* is also satisfied in this case.

∎

Received 2010-03-09