Towards a General Framework for Maximum Entropy Reasoning

Nico Potyka
Department of Computer Science
FernUniversität in Hagen,
Germany

Abstract

A possible approach to extend classical logics to probabilistic logics is to consider a probability distribution over the classical interpretations that satisfies some constraints and maximizes entropy. Over the past years miscellaneous languages and semantics have been considered often based on similar ideas. In this paper a hierarchy of general probabilistic semantics is developed. It incorporates some interesting specific semantics and a family of standard semantics that can be used to extend arbitrary languages with finite interpretation sets to probabilistic languages. We use the hierarchy to generalize an approach reducing the complexity of the whole entailment process and sketch the importance for further theoretical and practical applications.

1 Introduction

For representing knowledge a considerably drawback of classical logics is that uncertainty cannot be expressed. A formula can be true or false, but in many practical situations this is not sufficient. Consider a medical expert system for example. Given a set of symptoms different diagnoses can be more or less probable. A very natural way to express such uncertainties are probabilities.

A possible approach is to define a probability distribution over the interpretations of the classical logic often referred to as *possible worlds* (Nilsson 1986). Then to each formula a probability can be assigned by summing up the probabilities of their models. In most cases it is impractical to define a complete and reasonable probability distribution. In the ME-approach one formulates some probabilistic constraints, so-called conditionals, and selects the satisfying probability distribution having maximum entropy. Some rationales can be found in (Paris 1994) and (Kern-Isberner 2001).

The ME-Inference problem can be described as the task of determining the probability distribution satisfying the conditionals and maximizing entropy. Important for applications is primarily the ME-Entailment problem. Given a knowledge base and a formula we are interested in the probability of the formula. Considering the number of possible worlds it is obvious that the maximization problem as well as the computation of probabilities of formulas becomes challenging for complex scenarios. In (Paskin 2002) knowledge expressed by classical formulas is used to reduce the number of possible worlds in Nilsson's framework. We show that this approach is compatible with the conditional framework and can be generalized to several semantics. For this purpose a selection of ME-semantics is classified into a hierarchy of general semantics. They differ in increasing complexity of the satisfaction relation but yet feature a common structure. In particular, a standard semantics is introduced that can be used to extend arbitrary classical logical languages with finite interpretation sets to probabilistical logical languages similar to (Nilsson 1986). The introduced hierarchy might hopefully simplify the examination of the relationships between existing and further languages and their semantics and provide a simple framework to prove further general results.

In Section 2 we describe the basic building blocks of ME-languages and introduce a standard semantics for classical and relational languages as considered in (Nilsson 1986) and two semantics from (Kern-Isberner and Thimm 2010) that enable the expression of both subjective and statistical information. In Section 3 a general concept of a conditional semantics is defined. We develop a hierarchy of general semantics featuring some useful properties for the inference and entailment problem and integrate the introduced concrete semantics. In Section 4 we apply a well-known inference result to the hierarchy and show that the procedure proposed in (Paskin 2002) is consistent with ME-Inference and -Entailment and transfers to a very general family of adequately structured semantics.

2 Probabilistic Reasoning

Usually a logical language is built up of atomic elements. Formulas are obtained by combining these atoms to more complex structures using logical connectives like conjunction or negation. A classical semantics for the logic can be obtained by an interpretation assigning truth values to the atomic elements and defining how connected atoms have to be evaluated. An interpretation satisfying a formula is called a model of the formula. To abstract from the specific structure we consider a (logical) language $\mathcal{L}$, i.e., a set of formulas, together with a finite set of interpretations $\Omega_{\mathcal{L}}$ called *possible worlds*, and a satisfacion relation $\models_{\mathcal{L}}$. Let the set of classical models to a formula $\phi \in \mathcal{L}$ be denoted by

$$\mathrm{Mod}_{\mathcal{L}}(\phi) := \{\omega \in \Omega_{\mathcal{L}} \mid \omega \models_{\mathcal{L}} \phi\}.$$

To assign a probabilistic semantics to $\mathcal{L}$ we can define a probability distribution $\mathcal{P} : \Omega_{\mathcal{L}} \to [0,1]$ assigning a degree of belief to each possible world. $\mathcal{P}$ is extended to the power set $2^{\Omega_{\mathcal{L}}}$ via

$$\mathcal{P}(W) := \sum_{\omega \in W} \mathcal{P}(\omega) \qquad (1)$$

for all $W \subseteq \Omega_{\mathcal{L}}$. Let $\mathfrak{P}_{\mathcal{L}}$ denote the set of all such probability distributions over $\Omega_{\mathcal{L}}$. Given an arbitrary but fixed ordering of the possible worlds we can represent $\mathcal{P}$ by a $|\Omega_{\mathcal{L}}|$-dimensional vector $\vec{\mathcal{P}}$. Its components are the world-probabilities, and we write $\vec{\mathcal{P}}_{\omega}$ for the component containing $\mathcal{P}(\omega)$.

Example 2.1. *Consider a propositional logical language over two binary variables $\{A, B\}$. We represent the interpretations by the ordered complete conjunctions $(AB, \bar{A}B, A\bar{B}, \bar{A}\bar{B})$. We can define a probability distribution $\mathcal{P}$ by $\mathcal{P}(AB) := 0.2$, $\mathcal{P}(\bar{A}B) := 0.3$, $\mathcal{P}(A\bar{B}) := 0.4$, $\mathcal{P}(\bar{A}\bar{B}) := 0.1$. Then the corresponding vector is $\vec{\mathcal{P}} = (0.2 \quad 0.3 \quad 0.4 \quad 0.1)^{\mathsf{T}}$ and $\vec{\mathcal{P}}_{AB} = 0.2$. We obtain the probability of A by summing up the probability of its models, i.e., $\mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(A)) = \mathcal{P}(AB) + \mathcal{P}(A\bar{B}) = 0.6$.*

In most cases it is impractical to define a complete and reasonable probability distribution over the whole set of possible worlds. A usual approach is to formulate conditional constraints instead and select the best probability distribution satisfying these constraints with respect to a particular semantics. To begin with we define a conditional language similar to (Lukasiewicz 1999).

Definition 2.1. *Let $\mathcal{L}$ be a (logical) language with a finite set of interpretations $\Omega_{\mathcal{L}}$. The language*

$$(\mathcal{L}|\mathcal{L}) := \{(\psi|\phi)[x] \mid \phi, \psi \in \mathcal{L}, x \in [0,1]\}.$$

is called conditional language over $\mathcal{L}$.

The elements in $(\mathcal{L}|\mathcal{L})$ are called conditionals and can be considered as probabilistic rules. ϕ is called antecedence, ψ is called consequence of the conditional $(\psi|\phi)[x]$. If $x \in \{0,1\}$ the conditional is called *deterministic*. A conditional $(\psi|\top)[x]$, where $\top$ denotes a tautological formula, is called a *fact* and is often abbreviated by $(\psi)[x]$.

Example 2.2. *A classical propositional example is the following. Let B, P, F be propositional variables, representing the properties being a bird, being a penguin and being able to fly. Then $(F \mid B)[0.9]$, stating that birds fly with a probability of ninety percent, is a conditional and $(F \wedge P)[0]$, stating that penguins never fly, is a deterministic fact.*

$(\mathcal{L}|\mathcal{L})$ itself can be considered as a logical language interpreted by $\mathfrak{P}_{\mathcal{L}}$. Whereas in a classical logic an interpretation is a model of a formula iff it makes the formula true, a probability distribution is a model of a conditional $(\psi|\phi)[x]$ iff the probability of ψ given ϕ under the given semantics is x. Before introducing a general concept of a conditional semantics we introduce some specific semantics. For ease of notation we use $\mathrm{Mod}_{\mathcal{L}}(\phi\psi)$ as shorthand for $\mathrm{Mod}_{\mathcal{L}}(\phi) \cap \mathrm{Mod}_{\mathcal{L}}(\psi)$ and in particular $\mathrm{Mod}_{\mathcal{L}}(\phi\overline{\psi})$ for $\mathrm{Mod}_{\mathcal{L}}(\phi) \cap (\Omega_{\mathcal{L}} \setminus \mathrm{Mod}_{\mathcal{L}}(\psi))$ in the following.

Example 2.3. *Consider a propositional conditional language and let $F, G \in \mathcal{L}$. An often used propositional conditional semantics is defined by $\mathcal{P} \models_{\mathcal{S}} (G \mid F)[x]$ iff $\mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(F \wedge G)) = x \cdot \mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(F))$ (e.g. (Paris 1994)). If $\mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(F)) \neq 0$ this can be transformed into $x = \frac{\mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(F \wedge G))}{\mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(F))}$ which is the conditional probability of G given F.*

In a similar way conditional semantics can be defined over arbitrary logical languages. In (Nilsson 1986) for example a probabilistic semantics for first-order formulas is defined by summing up the probabilities of its models. By fixing the number of constants and forbidding function symbols the classical models can be represented by a finite set of Herbrand interpretations. The extension to the conditional framework is straightforward. If in the propositional example above $\mathcal{L}$ is the classical first-order language its conditional semantics can be defined in just the same way.

Conditional semantics like above will be captured by the definition of a *standard semantics* in the next section. Even though they define a very natural semantics for conditionals, they are not appropriate for representing both statistical uncertainty and individual degrees of belief. Following (Halpern 2003) we consider the fact $(\forall X(Bird(X) \rightarrow Flies(X)))[0.9]$ to express the belief that most birds fly. But if we knew about a bird, that does not fly, there can be no model for the universally quantified formula and necessarily $P(\forall X(Bird(X) \rightarrow Flies(X))) = 0$ for each probabilistic interpretation P. In (Kern-Isberner and Thimm 2010) the aggreating and averaging semantics are introduced. They deal with this problem by defining a probabilistic semantics for formulas containing free variables.

Example 2.4. *Consider a restricted relational language $\mathcal{L}$ built up over relations, constants and variables by conjunction, disjunction and negation. The interpretations $\Omega_{\mathcal{L}}$ are the possible Herbrand interpretations over the given relations and constants. A Herbrand interpretation is a model of a variable-free atom iff it contains the atom. For complex ground formulas the definition is extended in the usual way. For formulas containing free variables there are no classical models.*

The aggregating semantics *is a conditional semantics that uses a grounding operator* gr : $(\mathcal{L}|\mathcal{L}) \to 2^{(\mathcal{L}|\mathcal{L})}$ *mapping conditionals to the set of its ground instances to evaluate conditionals containing variables. It can be defined by $\mathcal{P} \models_{\mathcal{S}} (\psi|\phi)[x]$ iff*

$$\sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{gr}((\psi \mid \phi))} P(\mathrm{Mod}_{\mathcal{L}}(\psi_{\mathrm{gr}}\phi_{\mathrm{gr}})) = x \sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{gr}((\psi \mid \phi))} P(\mathrm{Mod}_{\mathcal{L}}(\phi_{\mathrm{gr}})).$$

Basically it puts the probabilities that antecedence and consequence of a ground instance are satisfied in relation to the probabilities that the antecedence of a ground instance is satisfied. Note that it coincides with the definition of the standard semantics above for ground conditionals.

Example 2.5. *The* averaging semantics *is defined on the same language as the aggregating semantics. Again there is no classical interpretation of formulas containing free variables. It is defined by calculating the average of the defined* *conditional probabilities of the ground instances of conditionals. That is, $\mathcal{P} \models_{\mathcal{S}} (\psi|\phi)[x]$ if and only if*

$$\frac{\sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{g}_{\mathcal{P}}((\psi \mid \phi))} \mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(\psi_{\mathrm{gr}}) \mid \mathrm{Mod}_{\mathcal{L}}(\phi_{\mathrm{gr}}))}{|\, \mathrm{g}_{\mathcal{P}}((\psi \mid \phi))|} = x,$$

where $\mathcal{P}(.\,|\,.)$ is the conditional probability and $\mathrm{g}_{\mathcal{P}}((\psi \mid \phi))$ is the set of groundings of the conditional $(\psi|\phi)[x]$ that satisfy $\mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(\phi_{\mathrm{gr}})) > 0$.

3 General Semantics

As we saw in the last section, conditional semantics are defined by a satisfaction relation between a probability distribution and conditionals. In general, the definition relies on a computation rule for the probability of a conditional dependent on the logical structure of antecedence and consequence. The following definition formalizes this idea.

Definition 3.1. *Let $(\mathcal{L}|\mathcal{L})$ be a conditional language with a finite set $\Omega_{\mathcal{L}}$ of interpretations of $\mathcal{L}$ and the set $\mathfrak{P}_{\mathcal{L}}$ of probability distributions over $\Omega_{\mathcal{L}}$. A satisfaction relation $\models_{\mathcal{S}} \subseteq \mathfrak{P}_{\mathcal{L}} \times (\mathcal{L}|\mathcal{L})$ defines a* conditional semantics *$\mathcal{S}$ iff there is a* constraint function *$f_c : \mathbb{R}^{|\Omega_{\mathcal{L}}|} \to \mathbb{R}$ for each conditional $c \in (\mathcal{L}|\mathcal{L})$ such that for all $\mathcal{P} \in \mathfrak{P}_{\mathcal{L}}$, $c \in (\mathcal{L}|\mathcal{L})$ it holds $\mathcal{P} \models_{\mathcal{S}} c$ iff $f_c(\vec{\mathcal{P}}) = 0$. The set of constraint functions is denoted by $\mathcal{F}_{\mathcal{S}} = \{f_c \mid c \in (\mathcal{L}|\mathcal{L})\}$.*

That is, $\mathcal{P}$ satisfies the conditional c under a given semantics $\mathcal{S}$ iff the equation corresponding to c evaluates to 0. For each conditional $c \in (\mathcal{L}|\mathcal{L})$ let $\mathrm{Mod}_{\mathcal{S}}(c) := \{\mathcal{P} \in \mathfrak{P}_{\mathcal{L}} \mid f_c(\vec{\mathcal{P}}) = 0\}$ denote the set of all probabilistic models under a given conditional semantics $\mathcal{S}$ and for a subset $\mathcal{R}_{\mathcal{L}} \subseteq (\mathcal{L}|\mathcal{L})$ let

$$\mathrm{Mod}_{\mathcal{S}}(\mathcal{R}_{\mathcal{L}}) := \bigcap_{(\psi|\phi)[x] \in \mathcal{R}_{\mathcal{L}}} \mathrm{Mod}_{\mathcal{S}}((\psi|\phi)[x])$$

be the set of common models of the conditionals in $\mathcal{R}_{\mathcal{L}}$. $\mathcal{R}_{\mathcal{L}}$ is called a *consistent knowledge base* iff $\mathrm{Mod}_{\mathcal{S}}(\mathcal{R}_{\mathcal{L}}) \neq \emptyset$. It is important to note, that we distinguish between the classical models $\mathrm{Mod}_{\mathcal{L}}$ of the logical language $\mathcal{L}$ and the probabilistic models $\mathrm{Mod}_{\mathcal{S}}$ of the conditional language $(\mathcal{L}|\mathcal{L})$ defined by the semantics $\mathcal{S}$.

To begin with we capture the standard semantics from example 2.3 with the following definition. As it is shown later, they are a special case of a more general class of semantics that are indeed conditional semantics as defined above.

Definition 3.2. *Let $(\mathcal{L}|\mathcal{L})$ be a conditional language over a classical logical language $\mathcal{L}$. The* standard semantics *over $\mathcal{L}$ is defined by $\mathcal{P} \models_{\mathcal{S}} (\psi|\phi)[x]$ if and only if*

$$\mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(\phi\psi)) = x \cdot \mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(\phi)). \qquad (2)$$

The standard semantics can be considered as a synthesis of the possible world semantics in (Nilsson 1986) with the conditional framework as considered by (Paris 1994) and others. The aggregating and averaging semantics are not captured by this definition, since they interpret free variables in a non-classical way to enable a statistical semantics as explained before. The following definition integrates some important similarities of the aggregating semantics and the standard semantics.

Definition 3.3. *A conditional semantics $\mathcal{S}$ is called* linearly structured *iff for each $f_c \in \mathcal{F}_{\mathcal{S}}$, $c = (\psi|\phi)[x]$, there are functions $\mathrm{V}_c : \Omega_{\mathcal{L}} \to \mathbb{N}_0$, $\mathrm{F}_c : \Omega_{\mathcal{L}} \to \mathbb{N}_0$ such that*

$$f_c(\vec{\mathcal{P}}) = \sum_{\omega \in \Omega_{\mathcal{L}}} \vec{\mathcal{P}}_\omega \cdot (\mathrm{V}_c(\omega) \cdot (1-x) - \mathrm{F}_c(\omega) \cdot x). \qquad (3)$$

Since the factor $(\mathrm{V}_c(\omega) \cdot (1-x) - \mathrm{F}_c(\omega) \cdot x)$ is independent of the function argument $\vec{\mathcal{P}}$ the constraint functions are indeed linear. The mappings V_c and F_c can be considered as a technical mean for incomplete classical interpretations. They indicate if the conditional c is verified respectively falsified by the considered world.

Lemma 3.1. *Each standard semantics is a linearly structured semantics.*

Proof. Equation (2) can be transformed into

$$(1-x) \cdot \mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(\phi\psi)) - x \cdot \mathcal{P}(\mathrm{Mod}_{\mathcal{L}}(\phi\overline{\psi})) = 0.$$

Exploiting equation (1) we can transform it into

$$\begin{aligned} 0 &= \sum_{\omega \in \mathrm{Mod}_{\mathcal{L}}(\phi\psi)} (1-x) \cdot \vec{\mathcal{P}}_\omega - \sum_{\omega \in \mathrm{Mod}_{\mathcal{L}}(\phi\overline{\psi})} x \cdot \vec{\mathcal{P}}_\omega \\ &= \sum_{\omega \in \Omega_{\mathcal{L}}} \vec{\mathcal{P}}_\omega \cdot (\sum_{\omega' \in (\{\omega\} \cap \mathrm{Mod}_{\mathcal{L}}(\phi\psi))} (1-x) - \sum_{\omega' \in (\{\omega\} \cap \mathrm{Mod}_{\mathcal{L}}(\phi\overline{\psi}))} x) \\ &= \sum_{\omega \in \Omega_{\mathcal{L}}} \vec{\mathcal{P}}_\omega \cdot (\mathrm{V}_c(\omega) \cdot (1-x) - \mathrm{F}_c(\omega) \cdot x), \end{aligned}$$

where $\mathrm{V}_c(\omega) := 1$ if $\omega \in \mathrm{Mod}_{\mathcal{L}}(\phi\psi)$ and 0 otherwise. Analogously we define $\mathrm{F}_c(\omega) := 1$ if $\omega \in \mathrm{Mod}_{\mathcal{L}}(\phi\overline{\psi})$ and 0 otherwise. □

Lemma 3.2. *The aggregating semantics is a linearly structured semantics.*

Proof. Similarly to the proof of Lemma (3.1) we obtain

$$\begin{aligned} 0 &= \sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{gr}((\psi \mid \phi))} \sum_{\omega \in \mathrm{Mod}_{\mathcal{L}}(\psi_{\mathrm{gr}}\phi_{\mathrm{gr}})} (1-x) \cdot \vec{\mathcal{P}}_\omega \\ &- \sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{gr}((\psi \mid \phi))} \sum_{\omega \in \mathrm{Mod}_{\mathcal{L}}(\phi_{\mathrm{gr}}\overline{\psi_{\mathrm{gr}}})} x \cdot \vec{\mathcal{P}}_\omega \\ &= \sum_{\omega \in \Omega_{\mathcal{L}}} \vec{\mathcal{P}}_\omega (\mathrm{V}_c(\omega) \cdot (1-x) - \mathrm{F}_c(\omega) \cdot x) \end{aligned}$$

where V_c and F_c are defined as follows:

$$\mathrm{V}_c(\omega) = |\{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{gr}((\psi \mid \phi)) \mid \omega \models_{\mathcal{L}} (\phi_{\mathrm{gr}}\psi_{\mathrm{gr}})\}|,$$

$$\mathrm{F}_c(\omega) = |\{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{gr}((\psi \mid \phi)) \mid \omega \models_{\mathcal{L}} (\phi_{\mathrm{gr}}\overline{\psi_{\mathrm{gr}}})\}|.$$

□

For the aggregating semantics V_c and F_c count the number of verified respectively falsified instances of the conditional. Note that this can be only 0 or 1 for a standard semantics.

If $\mathcal{S}$ is a linearly structured semantics, then $\mathrm{Mod}_{\mathcal{S}}((\psi|\phi)[x])$ is convex, since the solution set of a linear equation is convex and convex sets are closed under intersection. In (Thimm 2011) it is shown that the set of models for the averaging semantics can be non-convex. Hence it cannot be covered by the notion of linearly structured semantics. Yet it still features a useful structure.

Definition 3.4. *A conditional semantics $\mathcal{S}$ is called* structured *iff for each $f_c \in \mathcal{F}_\mathcal{S}$, $c = (\psi|\phi)[x]$, there are functions* $\mathrm{V}_c : \Omega_\mathcal{L} \times \mathbb{R}^{|\Omega_\mathcal{L}|} \to \mathbb{R}_{\geq 0}$, $\mathrm{F}_c : \Omega_\mathcal{L} \times \mathbb{R}^{|\Omega_\mathcal{L}|} \to \mathbb{R}_{\geq 0}$ *so that*

$$f_c(\vec{\mathcal{P}}) = \sum_{\omega \in \Omega_\mathcal{L}} \vec{\mathcal{P}}_\omega(\mathrm{V}_c(\omega, \vec{\mathcal{P}}) \cdot (1-x) - \mathrm{F}_c(\omega, \vec{\mathcal{P}}) \cdot x). \quad (4)$$

Note that the functions V_c, F_c now depend on $\vec{\mathcal{P}}$, hence the semantics can be non-linear. Furthermore, they map into the non-negative real numbers.

Corollary 3.3. *Each linearly structured semantics is a structured semantics.*

Lemma 3.4. *The averaging semantics is a structured semantics.*

Proof. Similarly to the previous proofs we obtain

$$\begin{aligned}
0 = & \sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{g}_\mathcal{P}((\psi \mid \phi))} \frac{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\psi_{\mathrm{gr}}\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega}{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega} \\
& -x \cdot |\,\mathrm{g}_\mathcal{P}((\psi \mid \phi))| \\
= & \sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{g}_\mathcal{P}((\psi \mid \phi))} \frac{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\psi_{\mathrm{gr}}\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega}{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega} \\
& -\frac{1}{|\,\mathrm{g}_\mathcal{P}((\psi \mid \phi))|} \sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{g}_\mathcal{P}((\psi \mid \phi))} x \cdot |\,\mathrm{g}_\mathcal{P}((\psi \mid \phi))| \\
= & \sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{g}_\mathcal{P}((\psi \mid \phi))} \frac{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\psi_{\mathrm{gr}}\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega - x \cdot \sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega}{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega} \\
= & \sum_{\omega \in \Omega_\mathcal{L}} (\sum_{\substack{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{g}_\mathcal{P}((\psi \mid \phi)) \\ \omega \models_\mathcal{L} \psi_{\mathrm{gr}}\phi_{\mathrm{gr}}}} \frac{(1-x) \cdot \vec{\mathcal{P}}_\omega}{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega} \\
& - \sum_{\substack{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{g}_\mathcal{P}((\psi \mid \phi)) \\ \omega \models_\mathcal{L} \psi_{\mathrm{gr}}\overline{\phi_{\mathrm{gr}}}}} \frac{x \cdot \vec{\mathcal{P}}_\omega}{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega}) \\
= & \sum_{\omega \in \Omega_\mathcal{L}} \vec{\mathcal{P}}_\omega(\mathrm{V}_c(\omega, \vec{\mathcal{P}}) \cdot (1-x) - \mathrm{F}_c(\omega, \vec{\mathcal{P}}) \cdot x)
\end{aligned}$$

where V_c and F_c are defined as follows:

$$\mathrm{V}_c(\omega, \vec{\mathcal{P}}) = \sum_{\substack{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{g}_\mathcal{P}((\psi \mid \phi)) \\ \omega \models_\mathcal{L} \psi_{\mathrm{gr}}\phi_{\mathrm{gr}}}} \frac{1}{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega},$$

$$\mathrm{F}_c(\omega, \vec{\mathcal{P}}) = \sum_{\substack{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{g}_\mathcal{P}((\psi \mid \phi)) \\ \omega \models_\mathcal{L} \psi_{\mathrm{gr}}\overline{\phi_{\mathrm{gr}}}}} \frac{1}{\sum\limits_{\omega \in \mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}})} \vec{\mathcal{P}}_\omega}.$$

□

Let $\mathit{Standard}, \mathit{LinearlyStructured}$ and $\mathit{Structured}$ denote the classes of the corresponding conditional semantics. Taking together the previous results, we obtain the following hierarchy.

Theorem 3.5. *Between the general semantics it holds*

$$\mathit{Standard} \subset \mathit{LinearlyStructured} \subset \mathit{Structured}.$$

Proof. The subset-relation follows from Lemma 3.1 and Corollary 3.3. It is strict, since the linearly structured aggregating semantics is non-standard and the structured averaging semantics is non-linear as explained before. □

4 Computation Properties

We return to the question how to determine a probability distribution $\mathcal{P} \in \mathfrak{P}_\mathcal{L}$ satisfying a knowledge base $\mathcal{R}_\mathcal{L}$. As explained before we select the distribution having maximum entropy. Setting $0 \cdot \log 0 := 0$ the entropy of a probability distribution $\mathcal{P} : \Omega_\mathcal{L} \to [0,1]$ is defined by $H(\mathcal{P}) := -\sum_{\omega \in \Omega_\mathcal{L}} \mathcal{P}(\omega) \cdot \log \mathcal{P}(\omega)$. Roughly speaking the ME inference process determines that probability distribution $\mathrm{ME}_\mathcal{S}(\mathcal{R}_\mathcal{L})$ satisfying $\mathcal{R}_\mathcal{L}$ and adding only as much information as necessary. An important requirement for a semantics in our framework is that there is a unique solution to the entropy maximization problem.

Definition 4.1. *Let $(\mathcal{L}|\mathcal{L})$ be a conditional language as above interpreted by a conditional semantics $\mathcal{S}$ and let $\mathcal{R}_\mathcal{L}$ be a consistent knowledge base with respect to $\mathcal{S}$. If the solution of the ME-Inference problem exists and is unique $\mathcal{S}$ is called ME-well-defined.*

Apart from the averaging semantics all semantics introduced above are known to be ME-well-defined. Indeed it is well-known that there is a unique solution for entropy maximization over linear constraints.

Corollary 4.1. *Each linearly structured semantics is ME-well-defined.*

Finally we state what the ME-Entailment problem is.

Definition 4.2. *Let $(\mathcal{L}|\mathcal{L})$ be a conditional language as above interpreted by a semantics $\mathcal{S}$. Given a knowledge base $\mathcal{R}_\mathcal{L}$ and formulas $\psi, \phi \in \mathcal{L}$ the ME-Entailment problem is to determine an $x \in [0,1]$ such that $\mathrm{ME}_\mathcal{S}(\mathcal{R}_\mathcal{L}) \models_\mathcal{S} (\psi|\phi)[x]$.*

A naive approach to solve the problem is to compute the optimal distribution $\mathcal{P}$ and solve the equation $f_{(\psi|\phi)[x]}(\vec{\mathcal{P}}) = 0$ for fixed $\vec{\mathcal{P}}$ and variable x. Unfortunately the solution for x is not necessarily unique. Usually the problem appears if the (grounded) antecedence has probability 0.

Example 4.1. *Consider the aggregating semantics from example 2.4. Given an arbitrary conditional $(\psi|\phi)[x]$ for $\sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{gr}((\psi \mid \phi))} P(\mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}})) > 0$ we get*

$$x = \frac{\sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{gr}((\psi \mid \phi))} P(\mathrm{Mod}_\mathcal{L}(\psi_{\mathrm{gr}}\phi_{\mathrm{gr}}))}{\sum_{(\psi_{\mathrm{gr}} \mid \phi_{\mathrm{gr}}) \in \mathrm{gr}((\psi \mid \phi))} P(\mathrm{Mod}_\mathcal{L}(\phi_{\mathrm{gr}}))}.$$

Otherwise each value for x is possible.

There is no simple satisfactory solution to this problem. See (Grove, Halpern, and Koller 1994), p. 61-62, for a more detailled discussion. If we assume $\mathcal{P}$ to be positive, i.e., $\mathcal{P}(\omega) > 0$ for all $\omega \in \Omega_{\mathcal{L}}$ the problem disappears. We will come back to this at the end of the paper.

In the naive solution the number of variables for the optimization problem, as well as the number of probabilities that has to be summed up becomes unmanageable even for small problems. In (Paskin 2002) a relational standard semantics over a restricted set of worlds is considered. The knowledge base is separated in formulas with probabilities in $]0,1[$ and classical formulas. Worlds in conflict with classical formulas are removed. In the following we show that this approach is fully compatible with our framework. More strictly speaking, deterministic conditionals are the analogon to the classical formulas considered by Paskin and we do not change our semantics (the optimization solution) accidentally by removing deterministic conditionals and the corresponding conflicting worlds.

To begin with the following lemma states that in some respects under each strutured semantics deterministic conditionals correspond to classical implications. That is, worlds falsifying conditionals of probability 1 must be impossible with respect to a valid probability distribution and analogously that worlds verifying conditionals of probability 0 must be impossible.

Lemma 4.2. *Let $(\mathcal{L}|\mathcal{L})$ be a logical language interpreted by a structured semantics $\mathcal{S}$. Let $c = (\psi|\phi)[x] \in \mathcal{R}_{\mathcal{L}}$ be a conditional and $\mathcal{P} \in \mathfrak{P}_{\mathcal{L}}$ with $\mathcal{P} \models_{\mathcal{S}} c$.*

1. If $x = 0$ then $\vec{\mathcal{P}}_\omega = 0$ for all $\omega \in \Omega_{\mathcal{L}}$ with $\mathrm{V}_c(\omega) \neq 0$.

2. If $x = 1$ then $\vec{\mathcal{P}}_\omega = 0$ for all $\omega \in \Omega_{\mathcal{L}}$ with $\mathrm{F}_c(\omega) \neq 0$.

Proof. Suppose $x = 0$. From the structured semantics equation (4) we obtain $0 = \sum_{\omega \in \Omega_{\mathcal{L}}} \vec{\mathcal{P}}_\omega \cdot \mathrm{V}_c(\omega, \vec{\mathcal{P}})$. Since all $\vec{\mathcal{P}}_\omega$ and $\mathrm{V}_c(\omega, \vec{\mathcal{P}})$ have to be non-negative, it holds $\vec{\mathcal{P}}_\omega = 0$ for all $\omega \in \Omega_{\mathcal{L}}$ with $\mathrm{V}_c(\omega) \neq 0$, if $\mathcal{P}$ satisfies the conditional. The second statement can be proved analogously. □

Hence deterministic conditionals determine the probabilities of some worlds independently of the remaining knowledge base. We separate each knowledge base into a deterministic part $\mathcal{R}_{\mathcal{L}}^{=} := \{(\psi|\phi)[x] \in \mathcal{R}_{\mathcal{L}} \mid x \in \{0,1\}\}$ and a probabilistic part $\mathcal{R}_{\mathcal{L}}^{\approx} := \mathcal{R}_{\mathcal{L}} \setminus \mathcal{R}_{\mathcal{L}}^{=}$. Let $\mathcal{N}_{\mathcal{R}_{\mathcal{L}}} := \{\omega \in \Omega_{\mathcal{L}} \mid \exists c = (\psi|\phi)[x] \in \mathcal{R}_{\mathcal{L}}^{=} : x = 0 \wedge \mathrm{V}_c(\omega) \neq 0 \vee x = 1 \wedge \mathrm{F}_c(\omega) \neq 0\}$ denote the set of worlds determined to be zero by $\mathcal{R}_{\mathcal{L}}^{=}$. The following proposition states that the separation proposed in (Paskin 2002) is consistent with the conditional framework under ME-well-defined structured semantics.

Proposition 4.3. *Let $(\mathcal{L}|\mathcal{L})$ be a conditional language as above interpreted by a ME-well-defined structured semantics $\mathcal{S}$. Let $\mathcal{R}_{\mathcal{L}}$ be a consistent knowledge base. Then the solution of the inference problem over $\Omega_{\mathcal{L}}$ with respect to $\mathcal{S}$ and $\mathcal{R}_{\mathcal{L}}$ can be obtained by solving the inference problem over $\Omega_{\mathcal{L}} \setminus \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ with respect to $\mathcal{S}$ and $\mathcal{R}_{\mathcal{L}}^{\approx}$.*

Proof. $\mathcal{R}_{\mathcal{L}}^{\approx}$ remains consistent with respect to $\Omega_{\mathcal{L}} \setminus \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ and $\mathcal{S}$ since the structured equations only change by missing zero-terms. Let $\mathcal{P}'$ be the unique solution of the inference problem over $\Omega_{\mathcal{L}} \setminus \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ with respect to $\mathcal{S}$ and $\mathcal{R}_{\mathcal{L}}^{\approx}$. We extend $\mathcal{P}'$ to a distribution $\mathcal{P}$ over $\Omega_{\mathcal{L}}$ by setting $\mathcal{P}(\omega) := \mathcal{P}'(\omega)$ for all $\omega \in (\Omega_{\mathcal{L}} \setminus \mathcal{N}_{\mathcal{R}_{\mathcal{L}}})$ and $\mathcal{P}(\omega) := 0$ for all $\omega \in \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$. Since only zero-probabilities are added $\mathcal{P}$ is still a probability distribution. $\mathcal{P}$ satisfies $\mathcal{R}_{\mathcal{L}}^{\approx}$ since only zero-terms are added to the structured equations. Now consider a conditional $(\psi|\phi)[x] \in \mathcal{R}_{\mathcal{L}}^{=}$. If $x = 0$ from the structured semantics equation (4) we obtain the condition $0 = \sum_{\omega \in \Omega_{\mathcal{L}}} \vec{\mathcal{P}}_\omega \cdot \mathrm{V}_c(\omega, \vec{\mathcal{P}})$. If $\mathrm{V}_c(\omega, \vec{\mathcal{P}}) = 0$ then $\vec{\mathcal{P}}_\omega \cdot \mathrm{V}_c(\omega, \vec{\mathcal{P}}) = 0$. If $\mathrm{V}_c(\omega, \vec{\mathcal{P}}) \neq 0$ then $\omega \in \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$. Hence $\vec{\mathcal{P}}_\omega = 0$ and again $\vec{\mathcal{P}}_\omega \cdot \mathrm{V}_c(\omega, \vec{\mathcal{P}}) = 0$, hence $\mathcal{P} \models_{\mathcal{S}} (\psi|\phi)[x]$. For $x = 1$ the proof is analogously. Hence $\mathcal{P}$ is a valid solution for the original problem.

Suppose $\mathcal{P}$ is not optimal for the original problem. Then there is a valid probability distribution $Q : \Omega_{\mathcal{L}} \to [0,1]$ so that $H(Q) > H(\mathcal{P})$. In particular Q satisfies $Q(\omega) = 0$ for all $\omega \in \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ due to Lemma 4.2. Hence the restriction of Q to a probability distribution Q' over $\Omega_{\mathcal{L}} \setminus \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ supplies a valid solution for the restricted problem. Since $0 \cdot \log 0 = 0$ it holds $H(Q') = H(Q) > H(\mathcal{P}) = H(\mathcal{P}')$ for the restricted distributions. But that is a contradiction, since $\mathcal{P}'$ is the optimal solution for the restricted problem. Hence $\mathcal{P}$ has to be the optimal solution for the original problem. □

Having determined the solution over $\Omega_{\mathcal{L}} \setminus \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ it is not necessary to return to the original interpretation set, since all worlds in $\mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ have zero-probability and in this way do not affect the structured equations. Hence the whole entailment process is simplified, in that only a fraction of the original worlds has to be considered.

Corollary 4.4. *Let $(\mathcal{L}|\mathcal{L})$, $\mathcal{S}$, $\mathcal{R}_{\mathcal{L}}$ be given as above and let $\psi, \phi \in \mathcal{L}$. Then the solution of the ME-Entailment problem over $\Omega_{\mathcal{L}}$ with respect to $\mathcal{S}$, $\mathcal{R}_{\mathcal{L}}$ and ψ, ϕ can be obtained by solving the ME-Entailment problem over $\Omega_{\mathcal{L}} \setminus \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ with respect to $\mathcal{S}$, $\mathcal{R}_{\mathcal{L}}^{\approx}$ and ψ, ϕ.*

Of course computing $\mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ can be a hard task itself, since it will probably include the enumeration of the models of the formulas corresponding to the verified respectively falsified conditionals. But it is not necessary to enumerate $\mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ nor to enumerate $\Omega_{\mathcal{L}}$. Instead one should exploit the knowledge obtained by V_c and F_c to enumerate $\Omega_{\mathcal{L}} \setminus \mathcal{N}_{\mathcal{R}_{\mathcal{L}}}$ in a branch-and-bound manner.

Example 4.2. *Consider a binary propositional logical language over n binary variables $\{A_1, A_2, ..., A_n\}$ under the standard semantics. The naive approach to generate the possible worlds is a simple recursive algorithm. If there is only one variable left we return the positive and the negative assignment. Given $k > 1$ variables we compute the assignments for the first $k-1$ variables and combine them with the positive and negative assignment to the k-th variable. Now given a conditional $(A_1 A_2)[0]$ we can cut the recursion as soon the assignment $(A_1 = 1)(A_2 = 1)$ is obtained.*

5 Discussion

We introduced a selection of semantics used for maximum entropy reasoning and classified them into a hierarchy of general semantics. The currently best investigated ME-semantics is the propositional standard semantics. Probably many algorithmical approaches developed in the past (e.g. (Rödder and Meyer 1996)) can be transferred to the whole class of standard semantics. Linearly structured semantics still guarantee the existence of a unique solution to the inference problem and their linear structure provides some computational benefits sketched at the end of this section. We cannot guarantee existence or uniqueness of the inference solution for structured semantics, but the structure is still sufficient to prove interesting results that transfer immediately to the more specific semantics. As we saw the entailment approach from (Paskin 2002) is compatible to each ME-well-defined structured semantics. Hence it transfers to each standard semantics and the aggregating semantics. If the averaging semantics is ME-well-defined the results also transfer to it. The hierarchy will hopefully be helpful to prove further general results.

At present the hierarchy might appear somewhat artificial since only a handful of semantics is included. In future work further semantics will be integrated. For example in (Fisseler 2010) and (Loh, Thimm, and Kern-Isberner 2010) probabilistic semantics can be found being closely related to simple structured semantics. By integrating further semantics into the hierarchy proving of standard results becomes unnecessary. In particular the standard semantics provides a simple framework to carry further classical languages over to a probabilistic language with advantageous computation properties. Lemma 4.2 indicates how the number of possible worlds can be reduced significantly. Especially for relational languages this becomes necessary, since the number of interpretations of a single binary relation becomes unmanageable already for more than a handful of constants. Number restrictions as used in some description logics (see (Baader 2009) for an overview) might be helpful to overcome these problems (e.g., no one has more than two biological parents).

Deterministic conditionals appear indeed in many applications, e.g., laws of heredity in biological domains or natural laws in technical domains. In many cases the null-worlds captured by $\mathcal{N}_{\mathcal{R}_\mathcal{L}}$ will be *all* null-worlds so that the probability distribution over $\Omega_\mathcal{L} \backslash \mathcal{N}_{\mathcal{R}_\mathcal{L}}$ will be positive. This is due to the fact that the entropy minimizes the informative distance to the uniform distribution. It is possible to construct counterexamples by enforcing exhaustive probabilities for a subset of worlds, but these are rather artificial. A positive probability distribution often avoids technical difficulties. For example conditionals with impossible antecedence can be identified before solving the expensive inference problem. Without going into details, we further state that the application of the method of Lagrange multipliers is immediately justified for positive distributions and can be used to represent the complete ME-optimal distribution in a product of the form $\mathcal{P}(\omega) = \alpha_0 \prod_{c=(\psi|\phi)[x]\in\mathcal{R}_\mathcal{L}} \alpha_c^{\mathrm{V}_c(\omega)(1-x)-\mathrm{F}_c(\omega)x}$ for each linearly structured semantics. This product representation is the key to several beneficial computation techniques for solving the inference and entailment problem that have been used in the expert-system SPIRIT for the propositional standard semantics (Rödder and Meyer 1996). Due to its similarity to the factorization of graphical models in particular refined inference methods used in the field of Statistical relational learning (Getoor and Taskar 2007) could be applied to solve the entailment problem for linearly structured semantics more efficiently.

Acknowledgements. The research reported here was partially supported by the Deutsche Forschungsgemeinschaft (grants BE 1700/7-2).

References

Baader, F. 2009. Description logics. In *Reasoning Web: Semantic Technologies for Information Systems, 5th International Summer School 2009*, volume 5689 of *Lecture Notes in Computer Science*. Springer–Verlag. 1–39.

Fisseler, F. 2010. *Learning and Modeling with Probabilistic Conditional Logic*, volume 328 of *Dissertations in Artificial Intelligence*. Amsterdam: IOS Press.

Getoor, L., and Taskar, B. 2007. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Grove, A.; Halpern, J.; and Koller, D. 1994. Random worlds and maximum entropy. *J. Artif. Int. Res.* 2:33–88.

Halpern, J. 2003. *Reasoning about Uncertainty*. MIT Press.

Kern-Isberner, G., and Thimm, M. 2010. Novel semantical approaches to relational probabilistic conditionals. In *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR'10)*, 382–392. AAAI Press.

Kern-Isberner, G. 2001. *Conditionals in nonmonotonic reasoning and belief revision*. Springer, Lecture Notes in Artificial Intelligence LNAI 2087.

Loh, S.; Thimm, M.; and Kern-Isberner, G. 2010. On the problem of grounding a relational probabilistic conditional knowledge base. In *Proceedings of the 14th International Workshop on Non-Monotonic Reasoning (NMR'10), Toronto, Canada, May 2010.*

Lukasiewicz, T. 1999. Probabilistic deduction with conditional constraints over basic events. *J. Artif. Intell. Res* 10:380–391.

Nilsson, N. J. 1986. Probabilistic logic. *Artif. Intell.* 28:71–88.

Paris, J. 1994. *The uncertain reasoner's companion – A mathematical perspective*. Cambridge University Press.

Paskin, M. A. 2002. Maximum entropy probabilistic logic. Technical Report UCB/CSD-01-1161, EECS Department, University of California, Berkeley.

Rödder, W., and Meyer, C.-H. 1996. Coherent knowledge processing at maximum entropy by SPIRIT. In Horvitz, E., and Jensen, F., eds., *Proceedings 12th Conference on Uncertainty in Artificial Intelligence*, 470–476. San Francisco, Ca.: Morgan Kaufmann.

Thimm, M. 2011. *Probabilistic Reasoning with Incomplete and Inconsistent Beliefs*. Ph.D. Dissertation, Technische Universität Dortmund, Germany.