Towards Unifying Logical Entailment and Statistical Estimation

Hiroyuki Kido
Cardiff University
KidoH@cardiff.ac.uk

Abstract

This paper gives a generative model of the interpretation of formal logic for data-driven logical reasoning. The key idea is to represent the interpretation as likelihood of a formula being true given a model of formal logic. Using the likelihood, Bayes’ theorem gives the posterior of the model being the case given the formula. The posterior represents an inverse interpretation of formal logic that seeks models making the formula true. The likelihood and posterior cause Bayesian learning that gives the probability of the conclusion being true in the models where all the premises are true. This paper looks at statistical and logical properties of the Bayesian learning. It is shown that the generative model is a unified theory of several different types of reasoning in logic and statistics.

1 Introduction

Thanks to big data and computational power available today, Bayesian statistics plays an important role in various fields such as neuroscience, cognitive science and artificial intelligence (AI). Bayesian brain hypothesis [Knill and Pouget, 2004], free-energy principle [Friston, 2010] and predictive coding [Hohwy et al., 2008] argue that probabilistic reasoning using Bayes’ theorem or its approximation explains some higher-order cognitive functions of the cerebral cortex such as perception, action and learning. The common idea is that the brain is a generative model that actively predicts and perceives the world using the belief of states of the world. Bayes’ theorem here defines how sensory inputs such as sight, sound, smell, taste and touch update the belief.

Formal logic concerns the laws of human rational thought. The Bayesian brain hypothesis would therefore result in another hypothesis that there is a Bayesian algorithm and data-structure for logical reasoning. This hypothesis is important for the following reasons. First, it has a potential to cause a mathematical model to explain how the human brain performs logical reasoning. Second, the existence of such a model supports the Bayesian brain hypothesis in terms of formal logic. Third, such a model gives a way to critically assess the existing formalisms of logical reasoning. Nevertheless, few research has focused on reformulating logical reasoning in terms of Bayesian perspectives. Bayesian networks [Pearl, 1988], probabilistic relational models (PRM) [Friedman et al., 1996], probabilistic logic programming (PLP) [Sato, 1995] and Markov logic networks (MLN) [Richardson and Domingos, 2006] are a few exceptions intrinsically relating to Bayesian inference. However, none of them aims to model the process by which data about states of the world generate models of formal logic and then the models generate the truth values of logical formulae. Such a model should give a unified way to deal with the tasks of the statistics and logic shown in Figure 1. This is an important problem because various challenging AI problems such as grounding, frame problems, knowledge acquisition bottleneck, commonsense reasoning and contextual adaptation, come from their disconnection.

Formal logic considers an interpretation on each model (denoted by $m$), which represents a state of the world. The interpretation is a function that maps each formula (denoted by $\alpha$) to a truth value, which represents knowledge of the world. Our idea is to give a generative model of the interpretation and use it to probabilistically generate knowledge from data about states of the world. The most basic theoretical idea is to represent the interpretation as likelihood $p(\alpha|m)$. Using the likelihood, Bayes’ theorem gives posterior $p(m|\alpha)$, which represents an inverse interpretation that gives the probability that the model making formula $\alpha$ true is $m$. The likelihood and posterior cause Bayesian learning $p(\alpha|\beta) = \sum_m p(\alpha|m)p(m|\beta)$, which gives the probability of the formula $\alpha$ being true in the models where the formula $\beta$ is true. This paper studies statistical and logical properties of the Bayesian learning.

This paper is organised as follows. Section 2 introduces a generative model for logical consequence relations. Section 3 shows logical and statistical correctness of the generative model. Section 4 concludes with a discussion of limitations and future work.

2 Method

Let $D = \{d_1, d_2, ..., d_K \}$ be a multiset of data about states of the world. $D$ is a random variable whose realisations are data in $D$. For all data $d_k \in D$, we define the probability of $d_k$, as follows.

$$p(D = d_k) = \frac{1}{K}$$
$L$ represents a propositional or first-order language. For the sake of simplicity, we assume no function symbol or open formula in $L$. $\mathcal{M} = \{m_1, m_2, ..., m_N\}$ is a set of models in formal logic. $\mathcal{D}$ is assumed to be complete with respect to $\mathcal{M}$, and thus each data in $\mathcal{D}$ belongs to a single model in $\mathcal{M}$. $m$ is a function that maps each data to such a single model. $K_n$ denotes the number of data that belongs to $m_n$, i.e., $K_n = |\{d_k \in \mathcal{D} | m_n = m(d_k)\}|$ where $|X|$ for set $X$ denotes the cardinality of $X$. $\mathcal{M}$ is a random variable whose realisations are models in $\mathcal{M}$. For all models $m_n \in \mathcal{M}$ and data $d_k \in \mathcal{D}$, we define the conditional probability of $m_n$ given $d_k$, as follows.

$$p(M = m_n | D = d_k) = \begin{cases} 1 & \text{if } m_n = m(d_k) \\ 0 & \text{otherwise} \end{cases}$$

Formal logic considers an interpretation on each model. The interpretation is a function that maps each formula to a truth value, which represents knowledge of the world. We here introduce parameter $\mu \in [0, 1]$ to represent the extent to which each model is taken for granted in the interpretation. Concretely, $\mu$ denotes the probability that a formula is interpreted as being true (resp. false) in a model where it is true (resp. false). $1 - \mu$ is therefore the probability that a formula is interpreted as being true (resp. false) in a model where it is false (resp. true). We assume that each formula is a random variable whose realisations are 0 and 1, denoting false and true, respectively. For all models $m_n \in \mathcal{M}$ and formulae $\alpha \in L$, we define the conditional probability of each truth value of $\alpha$ given $m_n$, as follows.

$$p(\alpha = 1 | M = m_n) = \begin{cases} \mu & \text{if } m_n \in \llbracket \alpha = 1 \rrbracket \\ 1 - \mu & \text{otherwise} \end{cases}$$$$p(\alpha = 0 | M = m_n) = \begin{cases} \mu & \text{if } m_n \in \llbracket \alpha = 0 \rrbracket \\ 1 - \mu & \text{otherwise} \end{cases}$$

Here, $\llbracket \alpha = 1 \rrbracket$ denotes the set of all models in which $\alpha$ is true, and $\llbracket \alpha = 0 \rrbracket$ the set of all models in which $\alpha$ is false. The above expressions can be simply written as a Bernoulli distribution with parameter $\mu \in [0, 1]$, i.e.,

$$p(\alpha | M = m_n) = \mu^{\llbracket \alpha = 1 \rrbracket} m_n (1 - \mu)^{1 - \llbracket \alpha = 1 \rrbracket} m_n.$$  

Here, $\llbracket \alpha \rrbracket m_n$ is a function such that $\llbracket \alpha \rrbracket m_n = 1$ if $m_n \in \llbracket \alpha \rrbracket$ and $\llbracket \alpha \rrbracket m_n = 0$ otherwise. Recall that $\alpha$ is a random variable, and thus $\llbracket \alpha \rrbracket m_n$ is either $\llbracket \alpha = 0 \rrbracket m_n$ or $\llbracket \alpha = 1 \rrbracket m_n$.

In classical logic, given a model, the truth value of each formula is independently determined. In probability theory, this means that the truth values of any two formulae $\alpha_1$ and $\alpha_2$ are conditionally independent given a model $m_n$, i.e., $p(\alpha_1, \alpha_2 | M = m_n) = p(\alpha_1 | M = m_n)p(\alpha_2 | M = m_n)$. Note that the conditional independence holds not only for atomic formulae but for compound formulae as well. Let $\Delta = \{\alpha_1, \alpha_2, ..., \alpha_j\}$ be a multiset of $J$ formulae. We thus have

$$p(\Delta | M = m_n) = \prod_{j=1}^{J} p(\alpha_j | M = m_n).$$

Thus far, we have defined $p(D)$ and $p(M | D)$ as categorical distributions and $p(\Delta | M)$ as Bernoulli distributions with parameter $\mu$. Given a value of the parameter $\mu$, they provide the full joint distribution over all of the random variables, i.e., $p(\Delta, M, D)$. We call $\{p(\Delta | M, \mu), p(M | D), p(D)\}$ a logical model. In sum, the logical model defines a data-driven interpretation by which the truth values of formulae are logically interpreted and probabilistically generated from models. The models are also probabilistically generated from data observed from the real world. The logical model meets the following important properties.

**Proposition 1.** The logical model satisfies Kolmogorov’s axioms.

**Proposition 2.** Let $\alpha \in L$. $p(\alpha = 0) = p(\neg \alpha = 1)$ holds.

In the following, we therefore replace $\alpha = 0$ by $\neg \alpha = 1$ and then abbreviate $\neg \alpha = 1$ to $\neg \alpha$. We also abbreviate $M = m_n$ to $m_n$ and $D = d_k$ to $d_k$.

**Example 1.** Let rain and wet be two propositional symbols meaning ‘it is raining’ and ‘the grass is wet’, respectively. Each row of Table 1 shows a different model, i.e., valuation. The last column shows how many data belongs to each model. Table 2 shows the likelihoods of the atomic propositions being true given a model. Given $\{p(\Delta | M, \mu = 1), p(M | D), p(D)\}$, we have

$$p(\text{rain} | \text{wet}) = \frac{\sum_{n=1}^{N} p(\text{rain} | m_n)p(\text{wet} | m_n) K_n}{\sum_{n=1}^{N} p(\text{wet} | m_n) K_n} = \frac{1}{\sum_{n=1}^{N} p(\text{wet} | m_n) K_n} \left( \frac{1 - \mu}{10} \mu^2 \frac{4}{10} + \frac{1 - \mu}{10} \mu \frac{17}{10} + \frac{1}{10} \mu^2 \frac{3}{10} \right) = \frac{3}{2 + 3} = 0.6.$$  

1In contrast, independence, i.e., $p(\alpha_1, \alpha_2) = p(\alpha_1)p(\alpha_2)$, holds only for atomic formulae.

| $\text{rain}$ & wet | $\text{data } D$ | $p(\text{rain} | M)p(\text{wet} | M)$ |
|---|---|---|
| $m_1$ | 0 0 | $\times \times \times$ | $m_1$ 1 - $\mu$ 1 - $\mu$ |
| $m_2$ | 0 1 | $\times \times$ | $m_2$ 1 - $\mu$ $\mu$ |
| $m_3$ | 1 0 | $\times$ | $m_3$ $\mu$ 1 - $\mu$ |
| $m_4$ | 1 1 | $\times \times \times$ | $m_4$ $\mu$ $\mu$ |

Table 1: Models and data.  
Table 2: Likelihoods.
The maximum likelihood estimate is obtained by solving the following simultaneous equations, which are obtained by differentiating the log likelihood with respect to each \( \phi_n (1 \leq n \leq N - 1) \).

\[
\frac{\partial L(\Phi)}{\partial \phi_n} = K_n \Phi_n - K_N = 0
\]

The following is the solution to the simultaneous equations.

\[
\Phi = \left( \frac{K_1}{K}, \frac{K_2}{K}, \ldots, \frac{K_N}{K} \right)
\]

Now, let \( \{ p(\Delta|M, \mu = 1), p(M|D), p(D) \} \) be a logical model such that \( \mu = 1 \). We show that both the Fenstad theorem and maximum likelihood estimation justify the logical model. The Fenstad theorem justifies the logical model because probabilistic inference on the logical model satisfies Equation (1).

\[
p(\alpha) = \sum_{n=1}^{N} p(\alpha|m_n) = \sum_{n=1}^{N} p(\alpha|m_n)p(m_n)
\]

Maximum likelihood estimation also justifies the logical model because probabilistic inference on the logical model satisfies Equation (2).

\[
p(\alpha) = \sum_{n=1}^{K} \sum_{k=1}^{p(m_n|d_k)} p(\alpha|m_n, d_k)
\]

We have shown that the logical model not only follows the Fenstad theorem and maximum likelihood estimation but also treats their results as probabilistic inference in a unified way. Their results are both in the scope of statistics shown in Figure 1. This is an important fact because, in the next section, we will discuss that the logical model can also deal with the scope of logic shown in Figure 1.

There are some practical advantages of the logical models. The computational complexity of Equation (3) depends on \( N \), which is unbounded in predicate logic and exponentially increases in propositional logic with respect to the number of
propositional symbols. However, Equation (3) can be transformed as follows for a linear complexity with respect to the number of data, i.e., $K$.

$$p(\alpha) = \sum_{n=1}^{N} [\alpha]_{m_n} \frac{K_n}{K} = \sum_{k=1}^{K} [\alpha]_{m(d_k)} \frac{1}{K} \quad (4)$$

In addition, Equation (3) has only a constant complexity for recalculation for new data. Let $p_{K}$ denote the probability calculated with $K$ data. $p_{K+1}(\alpha)$ can be calculated using $p_{K}(\alpha)$ as follows.

$$p_{K+1}(\alpha) = \sum_{n=1}^{N} p(\alpha|m_n) \frac{K_{n+1}}{K} = \sum_{k=1}^{K} p(\alpha|d_k) \frac{1}{K}$$

$$= \sum_{n=1}^{N} p(\alpha|m_n) \sum_{k=1}^{K} p(m_n|d_k) p(d_k)$$

$$+ \sum_{n=1}^{N} p(\alpha|m_n) p(m_{n+1}) \frac{1}{K}$$

$$= \frac{K+1}{K} \left( \sum_{n=1}^{N} p(\alpha|m_n) \sum_{k=1}^{K} p(m_n|d_k) \frac{1}{K} \right)$$

$$= \frac{K p_{K}(\alpha) + [\alpha]_{m(d_{K+1})}}{K+1} \quad (5)$$

Finally, as demonstrated in the following example, Equation (5) is good at modelling the development of commonsense knowledge.

**Example 3.** Let ‘bird’ and ‘fly’ be two propositional symbols meaning ‘It is a bird,’ and ‘It flies,’ respectively. Each row of Table 4 shows a different model. Given the ten data shown in the fourth column, the probability that bird implies fly is calculated using Equation (4), as follows.

$$p(\text{bird} \rightarrow \text{fly}) = \sum_{k=1}^{10} [\text{bird} \rightarrow \text{fly}]_{m(d_k)} \frac{1}{10} = 1$$

It is obvious from the logical model that the counterintuitive knowledge that birds must fly comes from a lack of data. Indeed, taking into account the eleven data shown in the last column, the probability is updated using Equation (5), as follows.

$$p_{11}(\alpha) = \frac{10 p_{10}(\text{bird} \rightarrow \text{fly}) + [\text{bird} \rightarrow \text{fly}]_{m(d_{11})}}{11} = \frac{10}{11}$$

### 3.2 Logical Entailment

We showed in the last section that, given $\{p(\Delta|M, \mu = 1), p(M|D), p(D)\}$, $p(M)$ is equivalent to the maximum likelihood estimate, i.e., for all $m_n \in M$,

$$p(m_n) = \sum_{k=1}^{K} p(m_n|d_k) p(d_k) = \frac{K_n}{K}$$

Therefore, $\{p(\Delta|M, \mu = 1), p(M|D), p(D)\}$ is equivalent to $\{p(\Delta|M, \mu = 1), p(M)\}$ when $p(M)$ is the maximum likelihood estimate. For the sake of simplicity, we also call the latter a logical model and use it without distinction. To discuss logical properties of the logical model, we assume $0 \notin p(M)$ meaning that every model is possible, i.e., $p(m) \neq 0$, for all models. Recall that a set $\Delta$ of formulae entails a formula $\alpha$ in classical logic, denoted by $\Delta \models \alpha$, iff $\alpha$ is true in every model in which $\Delta$ is true. The following two theorems state that certain inference on the logical model is more cautious than classical entailment.

**Theorem 1.** Let $\alpha \in L$ and $\Delta \subseteq L$ such that $[\Delta] \neq \emptyset$. 

$$p(\alpha|\Delta) = 1 \text{ if and only if } \Delta \models \alpha.$$

**Proof.** Recall that, in formal logic, the fact that there is a model of $\Delta$ (or $\Delta$ has a model) is equivalent to the fact that there is a model $m$ in which every formula in $\Delta$ is true in $m$. Dividing models into the models of $\Delta$ and the others, we have

$$p(\alpha|\Delta) = \frac{\sum_{m \in [\Delta]} p(\alpha|m) p(\Delta|m) p(m)}{\sum_{m \in [\Delta]} p(\Delta|m) p(m)}$$

$$= \frac{\sum_{m \in [\Delta]} p(\alpha|m) \mu(\Delta|m)}{\sum_{m \in [\Delta]} p(\Delta|m) p(m)}.$$ 

For all $m \notin [\Delta]$, there is $\beta \in \Delta$ such that $[\beta]_m = 0$. Therefore, $p(\Delta|m) = 0$ when $\mu = 1$, for all $m \notin [\Delta]$. We thus have

$$p(\alpha|\Delta) = \frac{\sum_{m \in [\Delta]} p(\alpha|m) \mu(\Delta)}{\sum_{m \in [\Delta]} p(\Delta|m) p(m)}.$$ 

Now, $\frac{\sum_{m \in [\Delta]} p(\alpha|m) \mu(\Delta)}{\sum_{m \in [\Delta]} p(\Delta|m) p(m)} = 1$ iff $[\alpha] \supseteq [\Delta]$, i.e., $\Delta \models \alpha$. □

**Example 4.** Theorem 1 does not hold without assumption $0 \notin p(M)$. Given $p(M) = (0.6, 0, 0.1, 0.3)$ in Example 1, $p(\text{rain} | \text{wet}) = 1$ but $\text{wet} \not\models \text{rain}.$

**Theorem 2.** Let $\alpha \in L$ and $\Delta \subseteq L$ such that $[\Delta] = \emptyset$. If $p(\alpha|\Delta) = 1$ then $\Delta \models \alpha$, but not vice versa.
Proof. \((\Rightarrow)\) If \([\Delta] = \emptyset\) then \(\Delta \models \alpha\), for all \(\alpha\), in classical logic. \((\Leftarrow)\) We show a counterexample where \(\Delta \models \alpha\) but \(p(\alpha|\Delta)\) is undefined. \(\beta, \neg \beta \models \alpha\) holds because \([\beta, \neg \beta] = \emptyset\) results in \([\beta, \neg \beta] \subseteq [\alpha]\). Meanwhile, \(p(\alpha|\beta, \neg \beta)\) is given as follows:

\[
p(\alpha|\beta, \neg \beta) = \frac{\sum_w p(w)p(\alpha|w)p(\beta|w)p(\neg \beta|w)}{\sum_w p(w)p(\beta|w)p(\neg \beta|w)} = \frac{\mu(1-\mu)\sum_w p(w)p(\alpha|w)}{\mu(1-\mu)\sum_w p(w)}
\]

This is undefined due to division by zero when \(\mu = 1\). \(\square\)

Everything is entailed from a contradiction in the classical entailment. Certain inference on the logical model is more cautious than the classical entailment because the proof of Theorem 2 states that nothing is entailed from a contradiction. In the next section, we look at a logical model that entails something reasonable from contradictions.

3.3 Paraconsistency

Let \(\{\lim_{\mu \to 1} p(\Delta, M, \mu), p(M)\}\) be a logical model such that \(\mu \to 1\) and \(0 \notin p(M)\) where \(\mu \to 1\) represents \(\mu\) approaches 1. The following two theorems state that certain inference on the logical model is more cautious than classical entailment.

Theorem 3. Let \(\alpha \in L\) and \(\Delta \subseteq L\) such that \([\Delta] \neq \emptyset\). \(p(\alpha|\Delta) = 1\) if and only if \(\Delta \models \alpha\).

Proof. \(\lim_{\mu \to 1}\) does not change the proof of Theorem 1. \(\square\)

Theorem 4. Let \(\alpha \in L\) and \(\Delta \subseteq L\) such that \([\Delta] = \emptyset\). If \(p(\alpha|\Delta) = 1\) then \(\Delta \models \alpha\), but not vice versa.

Proof. \((\Rightarrow)\) Same as for Theorem 2. \((\Leftarrow)\) We show a counterexample where \(\Delta \models \alpha\) but \(p(\alpha|\Delta) \neq 1\). Suppose \(p(\alpha) < 1\). We can show \(p(\alpha|\beta \wedge \neg \beta) < 1\) as follows:

\[
p(\alpha|\beta \wedge \neg \beta) = \sum_m p(m)\lim_{\mu \to 1} p(\alpha|m)\lim_{\mu \to 1} p(\beta \wedge \neg \beta|m) = \lim_{\mu \to 1} \frac{(1-\mu)\sum_m p(m)p(\alpha|m)}{(1-\mu)\sum_m p(m)} = \lim_{\mu \to 1} \frac{\sum_m p(m)p(\alpha|m)}{\sum_m p(m)}
\]

Therefore, \(p(\alpha|\beta \wedge \neg \beta) \neq 1\). Note that \(\beta \wedge \neg \beta \models \alpha\) because \([\beta \wedge \neg \beta] = \emptyset\) results in \([\beta \wedge \neg \beta] \subseteq [\alpha]\). \(\square\)

To characterise the certain inference on the logical model, we define an approximate model using maximal consistent subsets with respect to set cardinality. Recall that a set of formulae is consistent if there is a model of the set.

Definition 1 (Approximate model). Let \(m\) be a model and \(\Delta \subseteq L\) be an inconsistent set of formulae. \(m\) is an approximate model of \(\Delta\) if \(m\) is a model of a maximal (w.r.t. set cardinality) consistent subset of \(\Delta\).

Theorem 5. Let \(\Delta \subseteq L\) and \(\alpha \in L\). \(p(\alpha|\Delta) = 1\) if and only if \(\Delta' \models \alpha\), for all maximal (w.r.t. set cardinality) consistent subsets \(\Delta'\) of \(\Delta\).

Proof. We use notation \([\Delta]\) to denote the set of all approximate models of \(\Delta\). We also use notation \([\Delta]|m\) to denote the number of formulas in \(\Delta\) that are true in \(m\), i.e. \([\Delta]|m = \sum_{\beta \in [\Delta]}[\beta]|m\). Dividing models into \([\Delta]\) and the others, we have

\[
p(\alpha|\Delta) = \lim_{\mu \to 1} \frac{\sum_{\beta \in [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)}{\sum_{\beta \in [\Delta]} p(\beta|m)p(\Delta|m)} = \lim_{\mu \to 1} \frac{\sum_{m \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)}{\sum_{m \notin [\Delta]} p(\beta|m)p(\Delta|m)} \sum_{m \notin [\Delta]} p(\beta|m)p(\Delta|m) + \sum_{m \notin [\Delta]} p(\beta|m)p(\Delta|m)
\]

Now, \(p(\Delta|m)\) can be developed as follows, for all \(m\) (regardless of the membership of \([\Delta]\)):

\[
p(\Delta|m) = \prod_{\beta \in \Delta} \mu_\beta m(1-\mu)1-\mu_\beta t m
\]

Therefore, \(p(\alpha|\Delta) = \lim_{\mu \to 1} W + X Y Z\) where

\[
W = \sum_{m \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)
X = \sum_{m \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)
Y = \sum_{m \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)
Z = \sum_{m \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)
\]

From Definition 1, \([\Delta]|m\) has the same value, for all \(m \notin [\Delta]\). Therefore, the fraction can be simplified by dividing the denominator and numerator by \((1-\mu)|\Delta|\neg|\Delta|m\). We thus have \(p(\alpha|\Delta) = \lim_{\mu \to 1} W' + X' Y' Z'\) where

\[
W' = \sum_{m \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)
X' = \sum_{m \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)
Y' = \sum_{m \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)
Z' = \sum_{m \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)
\]

Applying the limit operation, we have

\[
p(\alpha|\Delta) = \frac{\sum_{\hat{m} \notin [\Delta]} p(\alpha|m)p(\beta|m)p(\Delta|m)}{\sum_{\hat{m} \notin [\Delta]} p(\beta|m)p(\Delta|m)} = \frac{\sum_{\hat{m} \notin [\Delta]} \mu_{\hat{m}} m(1-\mu_\beta m)}{\sum_{\hat{m} \notin [\Delta]} \mu_{\hat{m}} m(1-\mu_\beta m)}
\]


Since $1^{\alpha_m}0^{1-\alpha_m} = 1^01^0 = 1$ if $\hat{m} \in \{\alpha\}$ and $1^{\alpha_m}0^{1-\alpha_m} = 1^00^1 = 0$ if $\hat{m} \notin \{\alpha\}$, we have

$$p(\alpha|\Delta) = \frac{\sum_{\hat{m} \in (\Delta) \cap \alpha} p(\hat{m})}{\sum_{\hat{m} \in (\Delta)} p(\hat{m})}.$$

Therefore, $p(\alpha|\Delta) = 1$ holds iff $[\alpha] \supseteq [(\Delta)]$. By definition, $m \in (\Delta)$ iff $m$ is a model of a maximal consistent subset of $\Delta$ w.r.t. set cardinality. Therefore, $m \in (\Delta)$ iff $m \in \bigcup_{\Delta'}(\Delta')$ where $\Delta'$ is a maximal consistent subset of $\Delta$ w.r.t. set cardinality. Therefore, $p(\alpha|\Delta) = 1$ iff $[\alpha] \supseteq \bigcup_{\Delta'}(\Delta')$. In other words, for all maximal (w.r.t. set cardinality) consistent subsets $\Delta'$ of $\Delta$, $[\alpha] \supseteq [\Delta']$, i.e., $\Delta' \models \alpha$.

Example 5. Let $\mu \rightarrow 1$ and $p(M) = (0.25, 0.25, 0.25, 0.25)$ in Example 1. Given $\Delta = \{\text{rain}, \text{wet}, \text{rain} \rightarrow \text{wet}, \neg\text{wet}\}$, there are three maximal (w.r.t. set inclusion) consistent subsets, i.e., $S_1 = \{\text{rain}, \text{wet}, \text{rain} \rightarrow \text{wet}\}$, $S_2 = \{\text{rain}, \neg\text{wet}\}$ and $S_3 = \{\text{rain} \rightarrow \text{wet}, \neg\text{wet}\}$, and one maximal (w.r.t. set cardinality) consistent subset, i.e., $S_1$. $p(\text{rain}|\Delta) = 1$ and $S_1 \models \text{rain}$ hold, but $S_3 \not\models \text{rain}$.

### 3.4 Counterfactuals

Would England have won the match against Argentina at the 1986 World Cup if Diego Maradona had not used his hand to score the first goal? Reasoning with this kind of false and imaginary conditional statement is often called counterfactual reasoning. Let $\{\lim_{\mu \rightarrow 1} p(\Delta|\mu, p(M))\}$ be a logical model such that $\mu \rightarrow 1$. This section demonstrates that the certain inference on the logical model is a natural model of counterfactual reasoning.

Table 5 shows data on four football matches characterised by four attributes: $\text{goal}$, $\text{home}$, $\text{opponent}$, $\text{win}$, $\mu \in \{0, 1\}$. They are, respectively, facts about whether our teammate Alice scored a goal or not, whether the game was played at home or not, whether the opponent was 0 (meaning Belgium) or 1 (meaning Brazil), and whether our team won or not. Now, we consider the following question.

Our team lost the home game without Alice’s goal against Belgium, i.e., $m_1$. Would we have won if Alice had scored a goal in this match?

This question does not have a straightforward answer because it is a counterfactual with respect to the data. Indeed, the set of attributes, i.e., $\{\text{goal} = 1, \text{home} = 1, \text{opponent} = 0\}$, of the counterfactual does not appear in the data.

As long as the counterfactual does not exist in the data, it is reasonable to realise counterfactual reasoning based on the facts most similar to the counterfactual [Pearl, 2018]. The counterfactual shares attributes $\{\text{home} = 1, \text{opponent} = 0\}$ with $m_1$, $\{\text{goal} = 1, \text{home} = 0\}$ with $m_2$, $\{\text{goal} = 1, \text{opponent} = 0\}$ with $m_3$ and $\{\text{goal} = 0\}$ with $m_4$. The data thus indicates that $m_1$, $m_2$ and $m_3$ are most similar to the counterfactual in terms of the number of shared attributes. Since the team won in $m_2$ and $m_3$, it is reasonable to conclude that, given the counterfactual, the probability of winning is $2/3$. Here, readers might think that $m_1$ should be excluded from the most similar facts because, in the counterfactual, we look at the situation in which Alice scored a goal. However, $m_1$ contains important information because it is empirically true that the probability of winning with Alice’s goal is positively affected by the fact that we won without Alice’s goal and negatively affected by the fact that we lost without Alice’s goal.

Interestingly, the idea of counterfactual reasoning is naturally modelled by the logical model. The predictive probability of winning given the counterfactual is calculated as follows.

$$p(\text{win}|\text{goal}, \text{home}, \neg\text{opp.}) = \lim_{\mu \rightarrow 1} \frac{\sum_m p(\text{goal}|m)p(\text{home}|m)p(\neg\text{opp.}|m)p(\text{win}|m)p(M)}{\sum_m p(M)p(\text{home}|m)p(\neg\text{opp.}|m)p(\text{win}|m)p(M)}$$

$$= \frac{\mu^2(1-\mu)^2 + \mu^4(1-\mu) + \mu^3(1-\mu) + \mu(1-\mu)^3}{\mu^2(1-\mu)^2 + \mu^2(1-\mu) + \mu^3(1-\mu) + \mu(1-\mu)^2} = \frac{2}{3}$$

The denominator of the predictive probability turns out to equal the number of facts most similar to the counterfactual, i.e., $m_1$, $m_2$ and $m_3$, whereas the numerator turns out to equal the number of wins from the three games, i.e., $m_2$ and $m_3$. Note that only the logical model with $\mu \rightarrow 1$ successfully formalises the idea of counterfactual reasoning.

Our approach for counterfactual reasoning essentially differs from Pearl [Pearl, 2018] and Lewis [Lewis, 1973]. Our approach is data-driven, whereas Pearl’s approach is model-driven in the sense that it assumes a causal diagram. Our approach is based on probability theory, whereas Lewis’s approach is based on the possible-worlds semantics. Although a formal comparison is difficult, Table 6 shows that there are some counterparts between the two approaches.

### 4 Conclusions and Discussion

In this paper, we introduced a generative model of the logical interpretation that defines the process by which the truth values of formulae are generated probabilistically from data about states of the world. We showed that it is a theory of reasoning that deals with several reasoning problems such as statistical reasoning, logical reasoning, paraconsistent reasoning and counterfactual reasoning.
One of the limitations of the current work is that it is still unclear how our generative model relates to other types of reasoning studied in AI such as nonmonotonic reasoning, abductive reasoning, predictive reasoning and practical reasoning. We will extend the logical model to deal with them in a unified approach.

References

[Fenstad, 1967] J.E. Fenstad. Representations of probabilities defined on first order languages. In John N. Crossley, editor, *Sets, Models and Recursion Theory*, volume 46 of *Studies in Logic and the Foundations of Mathematics*, pages 156–172. Elsevier, 1967.

[Friedman et al., 1996] Nir Friedman, Lise Getoor, Dephne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *Proc. 16th Int. Joint Conf. on Artif. Intell.*, pages 1297–1304, 1996.

[Friston, 2010] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010.

[Hohwy et al., 2008] Jakob Hohwy, Andrea Roepstorff, and Karl Friston. Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108:687–701, 2008.

[Knill and Pouget, 2004] David C. Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27:712–719, 2004.

[Lewis, 1973] David Lewis. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[Pearl, 2018] Judea Pearl. *The Book of Why: The New Science of Cause and Effect*. Allen Lane, 2018.

[Richardson and Domingos, 2006] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.

[Sato, 1995] Taisuke Sato. A statistical learning method for logic programs with distribution semantics. In *Proc. 12th int. conf. on logic programming*, pages 715–729, 1995.