

Generative Logic Models for Data-Based Symbolic Reasoning

Hiroyuki Kido

Cardiff University, Park Place, Cardiff, CF10 3AT, UK

Abstract

Acquiring knowledge from data and reasoning with the obtained knowledge are both essential processes of successful logical systems. However, most current logical systems assume different algorithms for the two processes. The separation causes serious problems such as knowledge acquisition bottleneck, grounding and commonsense reasoning. This paper gives a simple probabilistic model unifying the two processes. It formalises how data generate models of formal logic and the models generate the truth values of logical formulae. The generated models and truth values are shown to be consistent with maximum likelihood estimation and Fenstad's theorem, respectively. Probabilistic reasoning on logical formulae is shown to be a reasonable alternative to a logical consequence relation and a paraconsistent consequence relation. This paper contributes to data-based reasoning with linear complexity.

Keywords

Bayesian learning, Logical entailment, Statistical estimation, Reasoning from data, Inverse interpretation

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) [1]. Bayes' theorem underlies most modern AI systems handling uncertainty such as self-driving cars, robotics, medical diagnosis and language translation [2]. Bayesian brain hypothesis [3], free-energy principle [4] and predictive coding [5] argue that the brain unconsciously and actively predicts and perceives the world using the belief of states of the world. Bayes' theorem is used here to explain how sensory inputs such as sight, sound, smell, taste and touch update the belief.

The generality of Bayesian statistics in intellectual phenomena makes us expect that there is a Bayesian algorithm and data structure for logical reasoning and that it can tackle fundamental assumptions of current existing systems. For example, Bayesian networks [6] including naive Bayes, probabilistic logic programming (PLP) [7] and Markov logic networks (MLN) [8] assume independence of knowledge or facts. However, the independence rarely holds in real data. Ordinary formal logic such as propositional logic, first-order logic and modal logic assume consistency of knowledge to avoid entailing everything from contradictions [9, 10]. However, contradictions are inevitable when one tries to scale up the knowledge base or describe subjects


AIC 2022, 8th International Workshop on Artificial Intelligence and Cognition

✉ KidoH@cardiff.ac.uk (H. Kido)

ORCID 0000-0002-7622-4428 (H. Kido)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

in detail. In addition to the above-mentioned methods, probabilistic logic [11] and conditional probabilistic logic [12] assume both statistical and logical machineries. The statistical machinery assigns each logical sentence a probability value or weight so that it reflects aspects of the world, whereas the logical machinery performs logical reasoning on the probabilistic knowledge so that conclusions preserve the uncertainty of premises. For example, Bayesian networks, naive Bayes and PLP assume maximum likelihood estimation or maximum a posteriori estimation for the statistical machinery. The probabilistic logic, conditional probabilistic logic and MLN assume a human expert to play that role. Kolmogorov's axioms [13] and Fenstad's theorems [14] argue constraints that ought to be satisfied by the probability or weight assignment. However, some serious AI problems such as knowledge acquisition bottleneck, grounding, frame problems and commonsense reasoning [2, 15, 16] remain open without unifying the two machineries.

To tackle these assumptions of the current existing systems, we give a simple probabilistic model unifying the two machineries. We call the probabilistic model a generative logic model (GLM) as it formalises the process by which data generate models of formal logic and the models generate the truth values of logical formulae. Ordinary 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 α) to a truth value, which represents knowledge of the world. Given data (denoted by d), the most basic idea introduced in this paper is to consider the model and interpretation as likelihoods $p(m|d)$ and $p(\alpha|m)$, respectively. The model likelihood represents the model restricted by the data. Using the interpretation likelihood, Bayes' theorem gives posterior $p(m|\alpha)$, which intuitively means an inverse interpretation that gives the probability that the model making formula α 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 α being true in the restricted models where the formula β is true. This paper looks at statistical and logical properties of the Bayesian learning.

We show that probabilistic reasoning on GLM satisfies the Kolmogorov's axioms (see Proposition 1) and a Fenstad's theorem (see Equation (3)), and is equivalent to maximum likelihood estimation (see Equation (4)). These facts justify the statistical correctness of GLM. Moreover, we show that probabilistic reasoning on GLM is equivalent to the classical entailment when the premise is consistent (see Theorem 1). It is equivalent to the classical entailment with maximal consistent subsets with respect to set cardinality when the premise is inconsistent (see Theorem 5). These facts justify the logical correctness of GLM. We exemplify commonsense reasoning and counterfactual reasoning with GLM (see Sections 3.1 and 3.5).

The contributions of this paper are summarised as follows. First, this paper offers an algorithm for data-based logical reasoning with linear complexity with respect to the number data. To the best of our knowledge, this is the first paper introducing the idea of generative models to formalise the process by which data generate models of formal logic and the models generate the truth values of logical formulae. Second, this paper shows that GLM cancels the fundamental three assumptions: independence of knowledge, consistency of knowledge and separation of statistical and logical machineries. In particular, the cancelation of the first assumption is due to our novel idea that GLM only models the dependency between models and logical sentences. This is different from the existing methods modelling the dependency between logical sentences.

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 briefly summarises the results.

2. Generative Logic Model

The first task is to give a probabilistic representation of the process by which data generate models of formal logic. Let $\mathcal{D} = \{d_1, d_2, \dots, d_K\}$ be a multiset of data about states of the world. D is a random variable whose realisations are data in \mathcal{D} . For all data $d_k \in \mathcal{D}$, 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 . $\mathcal{M} = \{m_1, m_2, \dots, m_N\}$ 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 . 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.

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

The second task is to give a probabilistic representation of the process by which models generate the truth values of logical sentences. Ordinary 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, μ 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 α 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 α is true, and $\llbracket \alpha = 0 \rrbracket$ the set of all models in which α is false. The above expressions can be simply written as a Bernoulli distribution with parameter $\mu \in [0, 1]$, i.e.,

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

Table 1

Models and data.

	<i>rain</i>	<i>wet</i>	data \mathcal{D}
m_1	0	0	$\times \times \times \times$
m_2	0	1	$\times \times$
m_3	1	0	\times
m_4	1	1	$\times \times \times$

Table 2

Likelihoods.

	$p(\textit{rain} M)$	$p(\textit{wet} M)$
m_1	$1 - \mu$	$1 - \mu$
m_2	$1 - \mu$	μ
m_3	μ	$1 - \mu$
m_4	μ	μ

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 α 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 α_1 and α_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.¹ Let $\Delta = \{\alpha_1, \alpha_2, \dots, \alpha_J\}$ 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 μ . Given a value of the parameter μ , 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 generative logic model (GLM). In sum, the generative logic 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 GLM meets the following important properties.

Proposition 1. *The generative logic 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*

$$p(\textit{rain}|\textit{wet}) = \frac{\sum_{n=1}^N p(\textit{rain}|m_n)p(\textit{wet}|m_n) \sum_{k=1}^K p(m_n|d_k)p(d_k)}{\sum_{n=1}^N p(\textit{wet}|m_n) \sum_{k=1}^K p(m_n|d_k)p(d_k)}$$

¹In contrast, independence $p(\alpha_1, \alpha_2) = p(\alpha_1)p(\alpha_2)$ generally holds for neither atomic formulae nor compound formulae.

Table 3

Three predicate models and ten associated data.

	<i>blames</i>				data \mathcal{D}
	(a, a)	(a, b)	(b, a)	(b, b)	
m_1	1	0	0	1	××
m_2	1	1	1	0	×××
m_3	0	1	0	1	×××××
other	other				no data

$$\begin{aligned}
&= \frac{\sum_{n=1}^N p(\text{rain}|m_n)p(\text{wet}|m_n)\frac{K_n}{K}}{\sum_{n=1}^N p(\text{wet}|m_n)\frac{K_n}{K}} \\
&= \frac{(1-\mu)^2\frac{4}{10} + (1-\mu)\mu\frac{2}{10} + \mu(1-\mu)\frac{1}{10} + \mu^2\frac{3}{10}}{(1-\mu)\frac{4}{10} + \mu\frac{2}{10} + (1-\mu)\frac{1}{10} + \mu\frac{3}{10}} = \frac{3}{2+3} = 0.6.
\end{aligned}$$

Example 2. Suppose that L has only one 2-ary predicate symbol ‘blames’ and that the Herbrand universe for L has only two constants $\{a, b\}$. There are four ground atoms, $\{\text{blames}(a, a), \text{blames}(a, b), \text{blames}(b, a), \text{blames}(b, b)\}$, which result in $2^4 = 16$ possible models. Each row of Table 3 shows a different model and the last column shows the number of data that belongs to the model. Models without data are abbreviated from the table. Given $\{p(\Delta|M, \mu = 1), p(M|D), p(D)\}$, we have

$$\begin{aligned}
& p(\forall x \text{ blames}(x, a) | \exists x \text{ blames}(x, a)) \\
&= \frac{\sum_{n=1}^{16} \llbracket \forall x \text{ blames}(x, a), \exists x \text{ blames}(x, a) \rrbracket_{m_n} \frac{K_n}{K}}{\sum_{n=1}^{16} \llbracket \exists x \text{ blames}(x, a) \rrbracket_{m_n} \frac{K_n}{K}} = \frac{K_2}{K_1 + K_2} = \frac{3}{2+3} = 0.6.
\end{aligned}$$

3. Correctness

3.1. Statistical Estimation

Fenstad [14] argues that the probability of a formula is the sum of the probabilities of the models where the formula is true. Let $\alpha \in L$ and $m_n \in \mathcal{M}$. When L has no function symbol or open formula, the first Fenstad theorem can have the following simpler form, where $m_n \models \alpha$ represents m_n satisfies α .

$$p(\alpha) = \sum_{n=1: m_n \models \alpha}^N p(m_n) \tag{1}$$

When one has no prior knowledge about the probability of models, the most frequently used method to estimate $p(M)$ only from data is maximum likelihood estimation, which is given as follows.

$$p(M) = \arg \max_{\Phi} p(\mathcal{D}|\Phi),$$

where Φ is the parameter of the categorical distribution $p(M)$. Assuming that each data is independent given Φ , we have

$$p(\mathcal{D}|\Phi) = \prod_{k=1}^K p(d_k|\Phi) = \phi_1^{K_1} \phi_2^{K_2} \cdots \phi_{N-1}^{K_{N-1}} (1 - \phi_1 - \phi_2 - \cdots - \phi_{N-1})^{K_N}.$$

Φ maximises the likelihood if and only if it maximises the log likelihood, which is given as follows.

$$\begin{aligned} L(\Phi) &= K_1 \log \phi_1 + K_2 \log \phi_2 + \cdots + K_{N-1} \log \phi_{N-1} \\ &\quad + K_N \log(1 - \phi_1 - \phi_2 - \cdots - \phi_{N-1}) \end{aligned}$$

The maximum likelihood estimate is obtained by solving the following simultaneous equations, which are obtained by differentiating the log likelihood with respect to each $\phi_n (1 \leq n \leq N-1)$.

$$\frac{\partial L(\Phi)}{\partial \phi_n} = \frac{K_n}{\phi_n} - \frac{K_N}{1 - \phi_1 - \phi_2 - \cdots - \phi_{N-1}} = 0$$

The following is the solution to the simultaneous equations.

$$\Phi = \left(\frac{K_1}{K}, \frac{K_2}{K}, \dots, \frac{K_N}{K} \right)$$

Therefore, the maximum likelihood estimate for the n -th model is just the ratio of the number of data in the model to the total number of data. Combining Equation (1) and the maximum likelihood estimate, we have

$$p(\alpha) = \sum_{n=1:m_n=\alpha}^N \frac{K_n}{K}. \quad (2)$$

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

$$p(\alpha) = \sum_{n=1}^N p(\alpha, m_n) = \sum_{n=1}^N p(\alpha|m_n)p(m_n) = \sum_{n=1}^N \llbracket \alpha \rrbracket_{m_n} p(m_n) = \sum_{n=1:m_n \in \llbracket \alpha \rrbracket}^N p(m_n) \quad (3)$$

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

$$\begin{aligned} p(\alpha) &= \sum_{n=1}^N \sum_{k=1}^K p(\alpha, m_n, d_k) = \sum_{n=1}^N p(\alpha|m_n) \sum_{k=1}^K p(m_n|d_k)p(d_k) \\ &= \sum_{n=1}^N \llbracket \alpha \rrbracket_{m_n} \frac{K_n}{K} = \sum_{n=1:m_n \in \llbracket \alpha \rrbracket}^N \frac{K_n}{K} \end{aligned} \quad (4)$$

We have shown that GLM not only follows the Fenstad's theorem and maximum likelihood estimation but also treats their results as probabilistic reasoning in a unified way. This result justifies the correctness of GLM from a statistical point of view.

Table 4
New data.

	<i>bird fly</i>		data	new data
m_1	0	0	× × × × ×	
m_2	0	1	× ×	
m_3	1	0		×
m_4	1	1	× × ×	

3.2. Reasoning from Data

There are some practical advantages of the GLMs. The computational complexity of Equation (4) 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 (4) can be transformed as follows for a linear complexity with respect to the number of data, i.e., K .

$$p(\alpha) = \sum_{n=1}^N \llbracket \alpha \rrbracket_{m_n} \frac{K_n}{K} = \sum_{k=1}^K \llbracket \alpha \rrbracket_{m(d_k)} \frac{1}{K} \quad (5)$$

In addition, Equation (4) 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.

$$\begin{aligned} p_{K+1}(\alpha) &= \sum_{n=1}^N p(\alpha|m_n) \sum_{k=1}^{K+1} p(m_n|d_k)p(d_k) \\ &= \sum_{n=1}^N p(\alpha|m_n) \sum_{k=1}^K p(m_n|d_k)p(d_k) + \sum_{n=1}^N p(\alpha|m_n)p(m_n|d_{K+1})p(d_{K+1}) \\ &= \frac{K}{K+1} \sum_{n=1}^N p(\alpha|m_n) \sum_{k=1}^K p(m_n|d_k) \frac{1}{K} + \sum_{n=1}^N p(\alpha|m_n)p(m_n|d_{K+1}) \frac{1}{K+1} \\ &= \frac{Kp_K(\alpha) + \llbracket \alpha \rrbracket_{m(d_{K+1})}}{K+1} \end{aligned} \quad (6)$$

Finally, as demonstrated in the following example, Equation (6) 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 (5), as follows.

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

It is obvious from the GLM that the counterintuitive knowledge that birds must fly comes from a lack of data. Indeed, taking into account the eleventh data shown in the last column, the probability

is updated using Equation (6), as follows.

$$p_{11}(\alpha) = \frac{10p_{10}(\text{bird} \rightarrow \text{fly}) + \llbracket \text{bird} \rightarrow \text{fly} \rrbracket_{m(d_{11})}}{11} = \frac{10}{11}$$

3.3. 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 $m_n \in \mathcal{M}$,

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

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 GLM and use it without distinction. To discuss logical properties of the GLM, 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 Δ of formulae entails a formula α in classical logic, denoted by $\Delta \models \alpha$, iff α is true in every model in which Δ is true, i.e., $\llbracket \Delta \rrbracket \subseteq \llbracket \alpha \rrbracket$. The following two theorems state that certain inference on the GLM is more cautious than classical entailment.

Theorem 1. *Let $\alpha \in L$ and $\Delta \subseteq L$ such that $\llbracket \Delta \rrbracket \neq \emptyset$. $p(\alpha|\Delta) = 1$ if and only if $\Delta \models \alpha$.*

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

$$\begin{aligned} p(\alpha|\Delta) &= \frac{\sum_m p(\alpha|m)p(\Delta|m)p(m)}{\sum_m p(\Delta|m)p(m)} \\ &= \frac{\sum_{m \in \llbracket \Delta \rrbracket} p(m)p(\alpha|m)\mu^{|\Delta|} + \sum_{m \notin \llbracket \Delta \rrbracket} p(m)p(\alpha|m)p(\Delta|m)}{\sum_{m \in \llbracket \Delta \rrbracket} p(m)\mu^{|\Delta|} + \sum_{m \notin \llbracket \Delta \rrbracket} p(m)p(\Delta|m)}. \end{aligned}$$

$p(\Delta|m) = \prod_{\beta \in \Delta} p(\beta|m) = \prod_{\beta \in \Delta} \mu^{\llbracket \beta \rrbracket_m} (1 - \mu)^{1 - \llbracket \beta \rrbracket_m}$. For all $m \notin \llbracket \Delta \rrbracket$, there is $\beta \in \Delta$ such that $\llbracket \beta \rrbracket_m = 0$. Therefore, $p(\Delta|m) = 0$ when $\mu = 1$, for all $m \notin \llbracket \Delta \rrbracket$. We thus have

$$p(\alpha|\Delta) = \frac{\sum_{m \in \llbracket \Delta \rrbracket} p(m)p(\alpha|m)1^{|\Delta|}}{\sum_{m \in \llbracket \Delta \rrbracket} p(m)1^{|\Delta|}} = \frac{\sum_{m \in \llbracket \Delta \rrbracket} p(m)1^{\llbracket \alpha \rrbracket_m} 0^{1 - \llbracket \alpha \rrbracket_m}}{\sum_{m \in \llbracket \Delta \rrbracket} p(m)}.$$

Since $1^{\llbracket \alpha \rrbracket_m} 0^{1 - \llbracket \alpha \rrbracket_m} = 1^1 0^0 = 1$ if $m \in \llbracket \alpha \rrbracket$ and $1^{\llbracket \alpha \rrbracket_m} 0^{1 - \llbracket \alpha \rrbracket_m} = 1^0 0^1 = 0$ if $m \notin \llbracket \alpha \rrbracket$, we have

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

Now, $\frac{\sum_{m \in \llbracket \Delta \rrbracket \cap \llbracket \alpha \rrbracket} p(m)}{\sum_{m \in \llbracket \Delta \rrbracket} p(m)} = 1$ iff $\llbracket \alpha \rrbracket \supseteq \llbracket \Delta \rrbracket$, 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 $\llbracket \Delta \rrbracket = \emptyset$. If $p(\alpha|\Delta) = 1$ then $\Delta \models \alpha$, but not vice versa.

Proof. (\Rightarrow) If $\llbracket \Delta \rrbracket = \emptyset$ then $\Delta \models \alpha$, for all α , 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 $\llbracket \beta, \neg\beta \rrbracket = \emptyset$ results in $\llbracket \beta, \neg\beta \rrbracket \subseteq \llbracket \alpha \rrbracket$. 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 $\mu = 1$. □

Everything is entailed from a contradiction in the classical entailment. Certain inference on the GLM 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 GLM that entails something reasonable from contradictions.

3.4. Paraconsistency

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

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

Proof. $\lim_{\mu \rightarrow 1}$ does not change the proof of Theorem 1. □

Theorem 4. Let $\alpha \in L$ and $\Delta \subseteq L$ such that $\llbracket \Delta \rrbracket = \emptyset$. 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.

$$\begin{aligned} p(\alpha|\beta \wedge \neg\beta) &= \frac{\sum_m p(m) \lim_{\mu \rightarrow 1} p(\alpha|m) \lim_{\mu \rightarrow 1} p(\beta \wedge \neg\beta|m)}{\sum_m p(m) \lim_{\mu \rightarrow 1} p(\beta \wedge \neg\beta|m)} \\ &= \lim_{\mu \rightarrow 1} \frac{(1-\mu) \sum_m p(m) p(\alpha|m)}{(1-\mu) \sum_m p(m)} = \lim_{\mu \rightarrow 1} \frac{\sum_m p(m) p(\alpha|m)}{\sum_m p(m)} \\ &= \sum_m p(m) \lim_{\mu \rightarrow 1} p(\alpha|m) = p(\alpha) \end{aligned}$$

Therefore, $p(\alpha|\beta \wedge \neg\beta) \neq 1$. Note that $\beta \wedge \neg\beta \models \alpha$ because $\llbracket \beta \wedge \neg\beta \rrbracket = \emptyset$ results in $\llbracket \beta \wedge \neg\beta \rrbracket \subseteq \llbracket \alpha \rrbracket$. □

To characterise the certain inference on the GLM, 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 Δ if m is a model of a maximal (w.r.t. set cardinality) consistent subset of Δ .

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 Δ' of Δ .

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

$$\begin{aligned} p(\alpha|\Delta) &= \lim_{\mu \rightarrow 1} \frac{\sum_m p(\alpha|m)p(m)p(\Delta|m)}{\sum_m p(m)p(\Delta|m)} \\ &= \lim_{\mu \rightarrow 1} \frac{\sum_{\hat{m} \in ((\Delta))} p(\alpha|\hat{m})p(\hat{m})p(\Delta|\hat{m}) + \sum_{m \notin ((\Delta))} p(\alpha|m)p(m)p(\Delta|m)}{\sum_{\hat{m} \in ((\Delta))} p(\hat{m})p(\Delta|\hat{m}) + \sum_{m \notin ((\Delta))} p(m)p(\Delta|m)}. \end{aligned}$$

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

$$\begin{aligned} p(\Delta|m) &= \prod_{\beta \in \Delta} p(\beta|m) = \prod_{\beta \in \Delta} \mu^{\llbracket \beta \rrbracket_m} (1 - \mu)^{1 - \llbracket \beta \rrbracket_m} \\ &= \mu^{\sum_{\beta \in \Delta} \llbracket \beta \rrbracket_m} (1 - \mu)^{\sum_{\beta \in \Delta} (1 - \llbracket \beta \rrbracket_m)} = \mu^{|\Delta|_m} (1 - \mu)^{|\Delta| - |\Delta|_m} \end{aligned}$$

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

$$\begin{aligned} W &= \sum_{\hat{m} \in ((\Delta))} p(\alpha|\hat{m})p(\hat{m})\mu^{|\Delta|_{\hat{m}}} (1 - \mu)^{|\Delta| - |\Delta|_{\hat{m}}} \\ X &= \sum_{m \notin ((\Delta))} p(\alpha|m)p(m)\mu^{|\Delta|_m} (1 - \mu)^{|\Delta| - |\Delta|_m} \\ Y &= \sum_{\hat{m} \in ((\Delta))} p(\hat{m})\mu^{|\Delta|_{\hat{m}}} (1 - \mu)^{|\Delta| - |\Delta|_{\hat{m}}} \\ Z &= \sum_{m \notin ((\Delta))} p(m)\mu^{|\Delta|_m} (1 - \mu)^{|\Delta| - |\Delta|_m}. \end{aligned}$$

From Definition 1, $|\Delta|_{\hat{m}}$ has the same value, for all $\hat{m} \in ((\Delta))$. Therefore, the fraction can be simplified by dividing the denominator and numerator by $(1 - \mu)^{|\Delta| - |\Delta|_{\hat{m}}}$. We thus have $p(\alpha|\Delta) = \lim_{\mu \rightarrow 1} \frac{W'+X'}{Y'+Z'}$ where

$$\begin{aligned} W' &= \sum_{\hat{m} \in ((\Delta))} p(\alpha|\hat{m})p(\hat{m})\mu^{|\Delta|_{\hat{m}}} \\ X' &= \sum_{m \notin ((\Delta))} p(\alpha|m)p(m)\mu^{|\Delta|_m} (1 - \mu)^{|\Delta|_{\hat{m}} - |\Delta|_m} \end{aligned}$$

$$\begin{aligned}
Y' &= \sum_{\hat{m} \in ((\Delta))} p(\hat{m}) \mu^{|\Delta|_{\hat{m}}} \\
Z' &= \sum_{m \notin ((\Delta))} p(m) \mu^{|\Delta|_m} (1 - \mu)^{|\Delta|_{\hat{m}} - |\Delta|_m}.
\end{aligned}$$

Applying the limit operation, we can cancel out X' and Z' and have

$$p(\alpha|\Delta) = \frac{\sum_{\hat{m} \in ((\Delta))} p(\alpha|\hat{m}) p(\hat{m})}{\sum_{\hat{m} \in ((\Delta))} p(\hat{m})} = \frac{\sum_{\hat{m} \in ((\Delta))} 1^{[\alpha]_{\hat{m}}} 0^{1 - [\alpha]_{\hat{m}}} p(\hat{m})}{\sum_{\hat{m} \in ((\Delta))} p(\hat{m})}.$$

Since $1^{[\alpha]_{\hat{m}}} 0^{1 - [\alpha]_{\hat{m}}} = 1^1 0^0 = 1$ if $\hat{m} \in [[\alpha]]$ and $1^{[\alpha]_{\hat{m}}} 0^{1 - [\alpha]_{\hat{m}}} = 1^0 0^1 = 0$ 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 Δ w.r.t. set cardinality. Therefore, $m \in ((\Delta))$ iff $m \in \bigcup_{\Delta'} [[\Delta']]$ where Δ' is a maximal consistent subset of Δ 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 Δ' of Δ , $[[\alpha]] \supseteq [[\Delta']]$, i.e., $\Delta' \models \alpha$. \square

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

3.5. 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 $\{\lim_{\mu \rightarrow 1} p(\Delta|M, \mu), p(M)\}$ be a GLM such that $\mu \rightarrow 1$. This section demonstrates that the certain inference on the GLM is a natural model of counterfactual reasoning.

Table 5 shows data on four football matches characterised by four attributes: *goal*, *home*, *opponent*, *win* $\in \{0, 1\}$. 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

Table 5

Prior distribution over four football matches.

	$p(M)$	<i>goal</i>	<i>home</i>	<i>opponent</i>	<i>win</i>
m_1	0.25	0	1	0	0
m_2	0.25	1	1	1	1
m_3	0.25	1	0	0	1
m_4	0.25	1	0	1	0

of attributes, i.e., ($goal = 1, home = 1, opponent = 0$), 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 [17]. The counterfactual shares attributes ($home = 1, opponent = 0$) with m_1 , ($goal = 1, home = 1$) with m_2 , ($goal = 1, opponent = 0$) with m_3 and ($goal = 1$) 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 GLM. The predictive probability of winning given the counterfactual is calculated as follows.

$$\begin{aligned}
 p(win|goal, home, \neg opp.) &= \lim_{\mu \rightarrow 1} \frac{\sum_m p(goal|m)p(home|m)p(\neg opp.|m)p(win|m)p(m)}{\sum_m p(goal|m)p(home|m)p(\neg opp.|m)p(m)} \\
 &= \lim_{\mu \rightarrow 1} \frac{\mu^2(1-\mu)^2 + \mu^3(1-\mu) + \mu^3(1-\mu) + \mu(1-\mu)^3}{\mu^2(1-\mu) + \mu^2(1-\mu) + \mu^2(1-\mu) + \mu(1-\mu)^2} = \frac{2}{3}
 \end{aligned}$$

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 GLM with $\mu \rightarrow 1$ successfully formalises the idea of counterfactual reasoning.

Our approach for counterfactual reasoning essentially differs from Pearl [17] and Lewis [18]. 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

We introduced the idea of generative models to the interpretation of formal logic. The idea referred to as generative logic models accounts for the process by which data about states of

Table 6

Correspondence with Lewis' counterfactuals.

Lewis' counterfactuals	Our counterfactuals
Possible worlds	Probability distribution $p(M)$
Our world(s)	Model(s) $[[\Delta]]$
Most similar world(s)	Approximate model(s) $((\Delta))$
Counterfactual $\Delta > \alpha$	Predictive distribution $p(\alpha \Delta)$

the world generate models of formal logic and the models generate the truth values of logical formulae. We showed that it is a theory of reasoning that deals with several types of reasoning such as statistical reasoning, logical reasoning, paraconsistent reasoning and counterfactual reasoning.

References

- [1] S. B. McGrayne, *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*, Yale University Press, 2011.
- [2] S. Russell, P. Norvig, *Artificial Intelligence : A Modern Approach*, Fourth Edition, Pearson Education, Inc., 2020.
- [3] D. C. Knill, A. Pouget, The bayesian brain: the role of uncertainty in neural coding and computation, *Trends in Neurosciences* 27 (2004) 712–719.
- [4] K. Friston, The free-energy principle: a unified brain theory?, *Nature Reviews Neuroscience* 11 (2010) 127–138.
- [5] J. Hohwy, A. Roepstorff, K. Friston, Predictive coding explains binocular rivalry: An epistemological review, *Cognition* 108 (2008) 687–701.
- [6] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.
- [7] T. Sato, A statistical learning method for logic programs with distribution semantics, in: *Proc. 12th int. conf. on logic programming*, 1995, pp. 715–729.
- [8] M. Richardson, P. Domingos, Markov logic networks, *Machine Learning* 62 (2006) 107–136.
- [9] G. Priest, *Paraconsistent Logic*, volume 6, *handbook of philosophical logic*, 2nd ed., Springer, 2002, pp. 287–393.
- [10] W. Carnielli, M. E. Coniglio, J. Marcos, *Logics of Formal Inconsistency*, volume 14, *handbook of philosophical logic*, 2nd ed., Springer, 2007, pp. 1–93.
- [11] N. J. Nilsson, Probabilistic logic, *Artificial Intelligence* 28 (1986) 71–87.
- [12] W. Rödder, Conditional logic and the principle of entropy, *Artificial Intelligence* 117 (2000) 83–106.
- [13] A. N. Kolmogorov, *Foundations of the theory of probability*, Chelsea Publishing Co., 1950.
- [14] J. Fenstad, Representations of probabilities defined on first order languages, in: *Sets, Models and Recursion Theory*, volume 46, Elsevier, 1967, pp. 156–172.
- [15] E. Davis, G. Marcus, Commonsense reasoning and commonsense knowledge in artificial intelligence, *Communications of the ACM* 58(9) (2015) 92–103.

- [16] G. Brewka, *Nonmonotonic Reasoning: Logical Foundations of Commonsense*, Cambridge University Press, 1991.
- [17] J. Pearl, *The Book of Why: The New Science of Cause and Effect*, Allen Lane, 2018.
- [18] D. Lewis, *Counterfactuals*, Harvard University Press, Cambridge, MA, 1973.

A. Proofs

Proposition 1. Let $\alpha, \beta \in L$. We need to show the following three properties.

1. $0 \leq p(\alpha = i)$ holds, for all $i \in \{0, 1\}$.
 2. $\sum_{i \in \{0,1\}} p(\alpha = i) = 1$ holds.
 3. $p(\alpha \vee \beta = i) = p(\alpha = i) + p(\beta = i) - p(\alpha \wedge \beta = i)$ holds, for all $i \in \{0, 1\}$.
- (1) $p(\alpha = i) = \sum_m p(\alpha = i|m)p(m)$. Both $p(\alpha = i|m)$ and $p(m)$ cannot be negative.
- (2) Since $\llbracket \alpha = 0 \rrbracket_m = 1 - \llbracket \alpha = 1 \rrbracket_m$, we have

$$\begin{aligned} p(\alpha = 0|m) + p(\alpha = 1|m) &= \mu^{\llbracket \alpha=0 \rrbracket_m} (1 - \mu)^{1 - \llbracket \alpha=0 \rrbracket_m} + \mu^{\llbracket \alpha=1 \rrbracket_m} (1 - \mu)^{1 - \llbracket \alpha=1 \rrbracket_m} \\ &= \mu^{1 - \llbracket \alpha=1 \rrbracket_m} (1 - \mu)^{\llbracket \alpha=1 \rrbracket_m} + \mu^{\llbracket \alpha=1 \rrbracket_m} (1 - \mu)^{1 - \llbracket \alpha=1 \rrbracket_m}. \end{aligned}$$

If $\llbracket \alpha = 1 \rrbracket_m = 1$ then $p(\alpha = 0|m) + p(\alpha = 1|m) = (1 - \mu) + \mu = 1$. If $\llbracket \alpha = 1 \rrbracket_m = 0$ then $p(\alpha = 0|m) + p(\alpha = 1|m) = \mu + (1 - \mu) = 1$. Therefore, we have

$$\begin{aligned} p(\alpha = 0) + p(\alpha = 1) &= \sum_m p(\alpha = 0|m)p(m) + \sum_m p(\alpha = 1|m)p(m) \\ &= \sum_m p(m) \{p(\alpha = 0|m) + p(\alpha = 1|m)\} = \sum_m p(m) = 1. \end{aligned}$$

(3) From (2), it is sufficient to show only case $i = 1$ because case $i = 0$ can be developed as follows.

$$1 - p(\alpha \vee \beta = 1) = 1 - \{p(\alpha = 1) + p(\beta = 1) - p(\alpha \wedge \beta = 1)\}$$

It is sufficient to show $p(\alpha \vee \beta = 1|m) = p(\alpha = 1|m) + p(\beta = 1|m) - p(\alpha \wedge \beta = 1|m)$, for all m , since the following holds.

$$\sum_m p(\alpha \vee \beta = 1|m)p(m) = \sum_m \{p(\alpha = 1|m) + p(\beta = 1|m) - p(\alpha \wedge \beta = 1|m)\}p(m)$$

By case analysis, the right expressions can have either of the following four cases.

$$(1 - \mu) + (1 - \mu) - (1 - \mu) = 1 - \mu \quad (7)$$

$$(1 - \mu) + \mu - (1 - \mu) = \mu \quad (8)$$

$$\mu + (1 - \mu) - (1 - \mu) = \mu \quad (9)$$

$$\mu + \mu - \mu = \mu \quad (10)$$

where (7), (8), (9) and (10) are obtained in the cases ($\llbracket \alpha = 1 \rrbracket_m = \llbracket \beta = 1 \rrbracket_m = 0$), ($\llbracket \alpha = 1 \rrbracket_m = 0$ and $\llbracket \beta = 1 \rrbracket_m = 1$), ($\llbracket \alpha = 1 \rrbracket_m = 1$ and $m \in \llbracket \beta = 1 \rrbracket_m = 0$), and ($\llbracket \alpha = 1 \rrbracket_m = \llbracket \beta = 1 \rrbracket_m = 1$), respectively. All of the results are consistent with the left expression, i.e., $p(\alpha \vee \beta = 1|m)$. \square