# On the Independence Assumption in Neurosymbolic Learning

**Emile van Krieken** [1] **Pasquale Minervini** [1] **Edoardo M. Ponti** [1] **Antonio Vergari** [1]

## Abstract

State-of-the-art neurosymbolic learning systems use probabilistic reasoning to guide neural networks towards predictions that conform to logical constraints over symbols. Many such systems assume that the probabilities of the considered symbols are conditionally independent given the input to simplify learning and reasoning. We study and criticise this assumption, highlighting how it can hinder optimisation and prevent uncertainty quantification. We prove that loss functions bias conditionally independent neural networks to become overconfident in their predictions. As a result, they are unable to represent uncertainty over multiple valid options. Furthermore, we prove that these loss functions are difficult to optimise: they are non-convex, and their minima are usually highly disconnected. Our theoretical analysis gives the foundation for replacing the conditional independence assumption and designing more expressive neurosymbolic probabilistic models.

## 1. Introduction

*Neurosymbolic learning* studies neurosymbolic models that combine neural perception and symbolic reasoning (Manhaeve et al., 2021; Xu et al., 2018; Badreddine et al., 2022). These models use logical constraints and data to create a loss function for learning neural perception models (Giunchiglia et al., 2022). When used effectively, neurosymbolic learning methods can use these constraints to improve data efficiency. However, researchers in the neurosymbolic learning community have found optimising the parameters of the perception models challenging (Marconato et al., 2023a; van Krieken et al., 2022; Manhaeve et al., 2021). A major underlying reason is that neurosymbolic learning cannot provide exact feedback on how the neural perception model should behave. We highlight this issue with a simple example.

[1]School of Informatics, University of Edinburgh. Correspondence to: Emile van Krieken <Emile.van.Krieken@ed.ac.uk>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

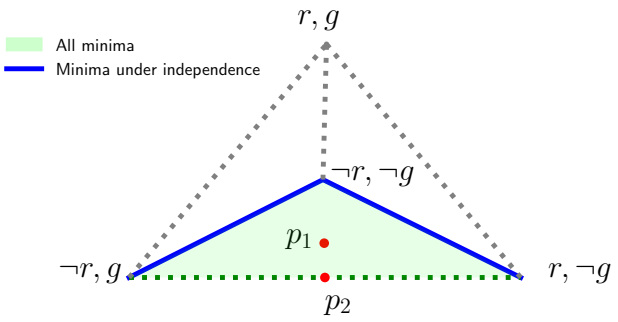


*Figure 1.* **The conditional independence assumption discards valid and potentially meaningful solutions.** The tetrahedron (a 3-dimensional probability simplex) represents the distributions over the options of the problem in Example 1.1: $r$ refers to the red light and $g$ to the green light. The green triangle represents distributions that assign zero probability to $r \wedge g$. The blue lines are the distributions in the green triangle that an independent distribution can represent. The left (resp. right) blue line represents the distributions where the probability of $r$ (resp. $g$) is zero. Independent distributions cannot represent distributions in the dotted green line, such as $p_2$ that assigns equal probability to only the green or only the red light being on.

**Example 1.1.** We consider a perception model responsible for recognising the red and green lights on a traffic light. It sees a traffic light that it believes to be simultaneously red and green. A constraint specifies this is impossible, and the neurosymbolic loss should penalise this. There are three possible worlds: the model can output that the red light is on, the green light is on, or neither. How do we choose among these?

The set of all beliefs can be represented by the tetrahedron in Figure 1. We argue that the perception model should be able to express *uncertainty* over these three worlds, as there is no evidence to conclude which one is correct. This corresponds to the distributions in the green triangle at the bottom of the figure. We should leave determining any further preference to the provided data as the constraint only specifies what worlds are *possible* but does not specify which one is *correct*.

The majority of probabilistic methods for neurosymbolic learning rely on a strong assumption: namely, that the differ-

ent symbols of the world are *independent* when conditioned on input data (Manhaeve et al., 2018; Xu et al., 2018). This means that for some input image, a conditionally independent perception model predicts two probabilities: one for the green light being on and one for the red light being on.

What do we lose when we take this conditional independence assumption? There is recent experimental evidence that suggests that using *expressive* perception models over conditionally independent ones improves performance on neurosymbolic tasks (Ahmed et al., 2022a; 2023; Pryor et al., 2023). We theoretically justify these results: the conditional independence assumption causes neurosymbolic methods to be biased towards deterministic solutions. This is because minima of neurosymbolic losses have to deterministically assign values to some variables. For instance, in Figure 1, the blue lines represent distributions that state that either the red light is off or the green light is off.

With the goal of better understanding the impact of the conditional independence assumption, we provide a computable characterisation of what can be represented. We find that we can characterise this problem faithfully using tools from logic (Quine, 1959) and computational homology (Kaczynski et al., 2004), and prove that this bias towards determinism holds generally. Furthermore, our characterisation shows that the conditional independence assumption can lead to training objectives that are challenging to optimise due to heavily disconnected minima. Our analysis provides theoretical justifications for the benefits of using more expressive perception models: the ability to properly express uncertainty and smooth, convex loss landscapes.

## 2. Background and Notation

**Probabilistic Neurosymbolic Learning.** We consider a probabilistic neurosymbolic learning (PNL) setting, where a probabilistic neural perception model $p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x})$ with parameters $\boldsymbol{\theta}$ defines a distribution over *worlds* $\mathbf{w} \in \{0,1\}^n$ (often called *concepts* (Barbiero et al., 2023; Marconato et al., 2023a)) given high-dimensional inputs $\mathbf{x} \in \mathcal{X}$. A constraint $\varphi : \{0,1\}^n \to \{0,1\}$ is a boolean function on worlds $\mathbf{w}$. We say a world is *possible* if $\varphi(\mathbf{w}) = 1$ and assume $\varphi$ has at least one possible world. The next example illustrates this setting.

**Example 2.1** (Learning with algorithms)**.** MNIST Addition is a popular benchmark task in neurosymbolic learning (Manhaeve et al., 2021). $\mathcal{X}$ is the set of pairs of MNIST images. We represent worlds $\mathbf{w}$ with $n = 20$ variables $\{w_{1,0}, ..., w_{1,9}, w_{2,0}, ..., w_{2,9}\}$, where $w_{i,j}$ denotes the $i$th digit taking the value $j$. We have a set of labels representing possible sums $\mathcal{Y} = \{0, \ldots, 18\}$. The constraints $\varphi_y$ enforce that exactly one of $w_{1,j}$ and one of $w_{2,k}$ is true, and ensures the pair of digits sums to the correct output: $\varphi_y(\mathbf{w}) = \exists_{j,k \in \{0,\ldots,9\}}(j + k = y) \wedge w_{1,j} \wedge w_{2,k}$. Here, the constraints $\varphi_y$ are parameterised by an observed output $y \in \mathcal{Y}$, which changes between inputs $\mathbf{x}$.

We compute the probability that the model $p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x})$ satisfies the constraint $\varphi$[1] for input $\mathbf{x}$ with:

$$p_{\boldsymbol{\theta}}(\varphi = 1|\mathbf{x}) = \sum_{\mathbf{w} \in \{0,1\}^n} p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x})\varphi(\mathbf{w}). \tag{1}$$

Equation 1 is known as the (conditional) *weighted model count (WMC)* in probabilistic and logical reasoning (Chavira and Darwiche, 2008). The $p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x})$ term can be understood as a data-dependent factor that assigns probabilities to different worlds, while the $\varphi(\mathbf{w})$ term is a constraint-dependent factor that filters out impossible worlds. Most loss functions based on WMC (Xu et al., 2018; Manhaeve et al., 2021) minimise the negative logarithm of the WMC $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = -\log p_{\boldsymbol{\theta}}(\varphi = 1|\mathbf{x})$, often called the *semantic loss*. See Section 6 for a discussion on these methods.

The majority of current PNL approaches assume that the probabilities $p_{\boldsymbol{\theta}}(w_i = 1|\mathbf{x})$ of variables $w_i$ being true are independent when conditioned on $\mathbf{x}$. Then, perception models only have to predict $n$ parameters in $[0,1]^n$ instead of a parameter for each of the $2^n$ worlds, i.e.,

$$p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x}) := \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(w_i|\mathbf{x}). \tag{2}$$

We call this the *(conditional) independence assumption*.[2] PNL systems take advantage of this assumption to speed up inference, reduce the number of trainable parameters, and ease implementation (Xu et al., 2018; Manhaeve et al., 2021; van Krieken et al., 2023; Ahmed et al., 2022a).

**Example 2.2** (Semi-supervised learning with constraints)**.** A common application of neurosymbolic learning is semi-supervised learning of the perception model $p_{\boldsymbol{\theta}}$. Here, we have a labelled dataset $\mathcal{D}_l = \{(\mathbf{x}_i, \mathbf{w}_i)\}_{i=1}^{|\mathcal{D}_l|}$ and an unlabelled dataset $\mathcal{D}_u = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}_u|}$. Here, $\varphi$ is a conjunction of a set of constraints $\phi_i$ that relate the symbols in the world $\mathbf{w}$. The semantic loss $\mathcal{L}(\boldsymbol{\theta})$ over the unlabelled data $\mathcal{D}_u$ is often added as a regularisation term to a supervised loss function to bias the neural network towards solutions that predict possible worlds (Xu et al., 2018).

[1] With abuse of notation, we use the symbol $\varphi$ both for the boolean function encoding the knowledge and for a binary random variable of the knowledge being true or not.

[2] In the rest of the paper, we refer to this *conditional* independence assumption as just "the independence assumption" for readability.

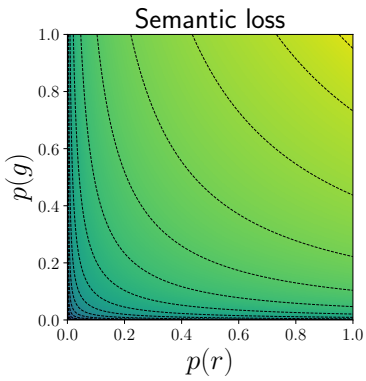


*Figure 2.* The loss landscape of the semantic loss for the traffic light problem – brighter (resp. darker) regions correspond to higher (resp. lower) semantic loss values.

## 3. The issues with the independence assumption

The main problem in our setting is how to learn the perception model $p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x})$. The underlying assumption in neurosymbolic learning is that there is a true distribution over worlds $p^*(\mathbf{w}|\mathbf{x})$. This distribution is unknown, and we cannot directly sample from $p^*(\mathbf{w}|\mathbf{x})$ to train $p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x})$. Instead, the feedback neurosymbolic learning methods provide is through the constraint $\varphi$, which induces a *set of possible worlds* $\mathcal{W}_\varphi = \{\mathbf{w} \in \{0,1\}^n \mid \varphi(\mathbf{w}) = 1\}$. If the constraint is correct, then all worlds $\mathbf{w}$ with non-zero probability $p^*(\mathbf{w}|\mathbf{x})$ are in $\mathcal{W}_\varphi$. Therefore, neurosymbolic learning methods should use the constraint $\varphi$ as a *filter* on what worlds are possible. We aim to answer the following questions: can particular parameterisations of $p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x})$ implicitly bias the selection of possible worlds instead of just filtering out impossible ones? And if so, how does this hinder their ability to recover $p^*(\mathbf{w}|\mathbf{x})$ via learning?

### 3.1. The independence assumption biases towards deterministic solutions

First, we show that common neurosymbolic learning methods are biased towards deterministic solutions. Returning to Example 1.1, consider a simple setup with $\mathbf{w}$ consisting of two binary variables $r$ and $g$ representing a red and green light, and a constraint $\varphi = \neg r \vee \neg g$ which asserts that the red and green lights cannot be on simultaneously.

In the remainder of the paper, we will fix the input $\mathbf{x}$ and keep it implicit in our notation unless necessary. Using our formula $\varphi$ in Equation 1, we then get:[3]

$$\begin{aligned} p_{\boldsymbol{\theta}}(\varphi = 1) &= p_{\boldsymbol{\theta}}(\neg r, \neg g) + p_{\boldsymbol{\theta}}(\neg r, g) + p_{\boldsymbol{\theta}}(r, \neg g) \\ &= 1 - p_{\boldsymbol{\theta}}(r, g). \end{aligned} \tag{3}$$

We can maximise this probability by simply enforcing $p_{\boldsymbol{\theta}}(r, g) = 0$. Then, the distribution over the remaining worlds can be arbitrary – such distributions are represented by the green triangle in Figure 1. However, taking the independence assumption over variables, we get:

$$p_{\boldsymbol{\theta}}(\varphi = 1) = 1 - p_{\boldsymbol{\theta}}(r) \cdot p_{\boldsymbol{\theta}}(g).$$

We plot the semantic loss for an independent distribution in Figure 2 as a function of $p_{\boldsymbol{\theta}}(r)$ and $p_{\boldsymbol{\theta}}(g)$. The semantic loss has its minima at the lines $p_{\boldsymbol{\theta}}(r) = 0$ and $p_{\boldsymbol{\theta}}(g) = 0$, biasing the model towards deterministically choosing either the red or green light being off, even though there is no evidence available to conclude this. Therefore, when optimising this function, we will come to a deterministic conclusion, which is wrong in a fraction of cases that depends on the real-world distribution of red and green lights being on.

Furthermore, an independent distribution cannot represent the beliefs $p_1$ and $p_2$ highlighted in Figure 1, where $p_1$ is the uniform belief over the three possible worlds, while $p_2$ is the equal belief in only the red light being on or only the green light being on. In fact, we cannot represent any distribution that assigns a non-zero probability to either just the red or the green light being on: if the semantic loss is minimised, then an independent distribution cannot represent uncertainty among multiple equally valid options.

Does this bias towards determinism happen for all formulas $\varphi$? We prove that this is indeed the case using the concept of *implicants* (Quine, 1959): an implicant assigns values to a subset of the variables $\{w_i\}_{i=0}^n$ such that it ensures the constraint $\varphi$ is true. In our example, $\neg r$ is an implicant of $\varphi$, since both $\neg r \wedge g$ and $\neg r \wedge \neg g$ are possible worlds. Our first theorem, which is formalised and proven in Section 4.2, generalises this result to all formulas $\varphi$:

**Theorem 3.1** (Implicants determine minima, informal)**.** *An independent distribution $p_{\boldsymbol{\theta}}(\mathbf{w})$ minimises the semantic loss if and only if it is deterministic for some variables, and those variables form an implicant of $\varphi$.*

We can, therefore, use the logical concept of implicants to study to what optima independent distributions converge. If the implicants are very restrictive, this greatly decreases the number of minima. The more restrictive the implicants of the formula are, the more the independent distributions will be biased towards deterministic solutions, and the less they will be able to quantify uncertainty.

[3] With some abuse of notation, we consider $r$ and $\neg g$ as events, that is, $p(r, \neg g) := p(r = 1, g = 0)$.

### 3.2. Minima under independence assumption are non-convex and disconnected

Using the connection to implicants from Theorem 3.1, we develop a geometric characterisation of the independent distributions that minimise the semantic loss. We emphasise that we study convexity and connectedness in the space of probability vectors over worlds, and not in the space of the parameters of the model $\boldsymbol{\theta}$. Our characterisation allows for an in-depth study of its topology using the tools of computational homology (Kaczynski et al., 2004). This allows us to give the exact conditions on the constraints $\varphi$ for which the minima are convex and connected. These conditions are valid only when we severely limit the types of constraints we can use. Therefore, the resulting semantic loss functions are usually highly non-convex and disconnected, and so difficult to optimise. In contrast, for expressive distributions, the semantic loss is always convex, as we show in Section 4.4.

**Theorem 3.2** (Convexity, informal)**.** *The semantic loss is convex over the set of independent distributions only when the constraint $\varphi$ is a formula of the form $\bigwedge_{i=1}^{L} l_i$, where each $l_i$ is a literal (a variable or its negation).*

We discuss this in detail in Section 4.4.2. The intuition behind this result is that such formulas provide direct supervision on $L$ variables, and give no supervision on the remaining $n - L$ variables. Therefore, the loss function is convex over the $L$ variables, and the remaining $n - L$ variables can be chosen arbitrarily. Note that this is a very restrictive condition: In many neurosymbolic settings, we have no direct supervision, and the constraint just acts as a filter on what worlds are possible.

**Theorem 3.3** (Connectedness, informal)**.** *The independent distributions that minimise the semantic loss are connected only if the implicants of the constraint $\varphi$ form a connected graph between worlds.*

In Section 4.4.3, we define a graph where the vertices are the possible worlds $\mathcal{W}_\varphi$ that are connected when there is an implicant that "covers" both worlds. If this graph is connected, then so are the minima. This result is quite abstract, so we provide two examples to illustrate it. For the traffic light example, the graph is connected: The three possible worlds are connected through the implicants $\neg r$ and $\neg g$. However, for the MNIST Addition task (Example 2.1), the graph contains no edges at all, which implies that the minima are a set of disconnected vertices.

## 4. Characterising minima of the semantic loss

In this section, we will develop the mathematical machinery to be able to characterise what it means for a distribution to be a minimum of the semantic loss, and, in particular, for independent distributions.

Section 4.1 discusses the expressivity of distributions and introduces our notation. Section 4.2 characterises the minima of the semantic loss for independent distributions. Section 4.4.1 studies a minimal representation of those minima. Finally, Section 4.4.2 shows when this set is convex, and Section 4.4.3 when this set is connected. Both turn out to be very rare. We provide proof sketches for most theorems in the main text and leave the full proofs to Appendix G. For ease of reading, we provide a table of notation in Table 1.

### 4.1. Expressive distributions

The *expressiveness* of a perception model $p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x})$ (Ahmed et al., 2022a) intuitively refers to how many distributions over worlds it can represent. A fully expressive perception model can represent any distribution $p(\mathbf{w}|\mathbf{x})$.

The set of all joint distributions over worlds is the $(2^n - 1)$-(probability) simplex having the standard unit vectors $\mathbf{e}_i$ as vertices: $\Delta = \{\sum_{i=1}^{2^n} \alpha_i \mathbf{e}_i : \sum_{i=1}^{2^n} \alpha_i = 1, \boldsymbol{\alpha} \geq \mathbf{0}\} \subset \mathbb{R}^{2^n}$. The vertices $\mathbf{e}_i$ are one-hot representations of the worlds $\mathbf{w}_i \in \{0,1\}^n$. We fix an arbitrary ordering $\mathbf{w}_1, ..., \mathbf{w}_{2^n}$ throughout. All probability distributions $p$ considered in this paper then correspond to the vector $(p(\mathbf{w}_1), \ldots, p(\mathbf{w}_{2^n}))$ in $\Delta$. With abuse of notation, we say $p \in \Delta$, referring to $p$ as both this vector and a distribution.

We define *possible distributions* $p \in \Delta$ as a distribution that assigns all probability mass to possible worlds. An equivalent statement is that $p(\mathbf{w}) = 0$ for all impossible worlds $\mathbf{w} \in \{0,1\}^n \setminus \mathcal{W}_\varphi$ (Marconato et al., 2023b). The *set of all possible distributions* $\Delta_\varphi \subseteq \Delta$ is a $(|\mathcal{W}_\varphi| - 1)$-simplex formed from the standard unit vectors associated with the possible worlds $\mathcal{W}_\varphi$. Since $\Delta_\varphi$ is a simplex, it is a convex set. Furthermore, the semantic loss $\mathcal{L}(p)$ is convex over the set of distributions $\Delta$ since the WMC is linear (see Appendix E for a proof).

An *expressive parameterisation* $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}$ of the joint distribution can represent any distribution $p \in \Delta$: For each input $\mathbf{x} \in \mathcal{X}$, there is a parameter $\boldsymbol{\theta}$ such that $p(\mathbf{w}) = p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x})$. A (parameter-inefficient) fully expressive parameterisation is to predict a vector of $2^n$ logits, which is then mapped to $\Delta$ via softmax. Expressive parameterisations behave quite differently in the example discussed in Section 3.1. They can minimise the probability of the constraint $\varphi$ in Equation 3 by simply setting $p_{\boldsymbol{\theta}}(r, g) = 0$, and model any preference over the remaining three worlds. This prevents the model from having to deterministically choose that either the red or green light is off and allows it to represent uncertainty.

### 4.2. When do independent parameterisations satisfy the constraint?

We next study the properties of independent distributions that satisfy the constraint. An independent distribution $p_{\boldsymbol{\mu}}$ is characterised by parameters $\boldsymbol{\mu}$[4] in the $n$-hypercube $[0,1]^n$, where $\mu_i$ is the probability that $w_i$ is true. To be precise, $p_{\boldsymbol{\mu}}(w_i) = \mu_i^{w_i} \cdot (1-\mu_i)^{1-w_i}$ is the probability mass function of the Bernoulli random variable $w_i$. We now formally define *implicants* (Quine, 1959), which are related to the deterministic components of $\boldsymbol{\mu}$:

**Definition 4.1** (Deterministic assignments)**.** A probability $\mu_i \in [0,1]$ is *deterministic* if $\mu_i \in \{0,1\}$. Otherwise, $\mu_i$ is *stochastic*, that is, $\mu_i \in (0,1)$. A *partial assignment* $\mathbf{w}_D = \{w_i\}_{i \in D}$ assigns values $\{0,1\}$ to a subset of the variables $\mathbf{w}$ indexed by $D \subseteq \{1,\ldots,n\}$. The *deterministic assignment* of an independent distribution $p_{\boldsymbol{\mu}}$ is the partial assignment $\mathbf{w}_D = \{\mu_i\}_{i \in D}$ defined by its deterministic factors $D = \{i | \mu_i \in \{0,1\}\}$.

**Definition 4.2** (Implicants)**.** We define the *cover* $\mathcal{W}_{\mathbf{w}_D} \subseteq \{0,1\}^n$ of a partial assignment $\mathbf{w}_D$ as the set that contains all worlds $\mathbf{w} \in \{0,1\}^n$ that equal $\mathbf{w}_D$ on the variables in $D$. A partial assignment $\mathbf{w}_D$ is an *implicant* of $\varphi$ if its cover only contains possible worlds. That is, $\mathbf{w}_D \models \varphi$.

Intuitively, an implicant assigns values to a subset of the variables in $\mathbf{w}$ such that it ensures the constraint $\varphi$ is true. For the traffic lights example, the partial assignment $\neg r$ is an implicant of $\varphi$, since its cover ($\neg r \wedge g$ and $\neg r \wedge \neg g$) only contains possible worlds. Our first result states that if we have an implicant, we can easily create possible independent distributions: Use the implicant as the deterministic part, and assign any value in $[0,1]$ to the remaining factors. This is because, for implicants, the value of the other variables "does not matter" to the constraint $\varphi$.

**Theorem 4.3** (Implicants determine possible independent distributions)**.** *Let $p_{\boldsymbol{\mu}}$ be an independent distribution over worlds. Let $\mathbf{w}_D$ be $p_{\boldsymbol{\mu}}$'s deterministic assignment. Then $p_{\boldsymbol{\mu}}$ is possible for $\varphi$ if and only if $\mathbf{w}_D$ is an implicant of $\varphi$.*

*Proof.* By independence of $p_{\boldsymbol{\mu}}$, the support of $p_{\boldsymbol{\mu}}$ is the cover $\mathcal{W}_{\mathbf{w}_D}$ of the deterministic assignment $\mathbf{w}_D$ of $\boldsymbol{\mu}$, and the remaining variables can be assigned any value with some probability. Assume $p_{\boldsymbol{\mu}}$ is possible; then, for each $\mathbf{w}$ in the support of $p_{\boldsymbol{\mu}}$, $\mathbf{w}$ is a possible world. But then each world in the cover $\mathcal{W}_{\mathbf{w}_D}$ of $\mathbf{w}_D$ is possible, and so $\mathbf{w}_D$ is an implicant. Next, assume $\mathbf{w}_D$ is an implicant. Then, each world in the cover of $\mathbf{w}_D$ is possible. But this is precisely the support of $p_{\boldsymbol{\mu}}$. So $p_{\boldsymbol{\mu}}$ is possible. □

The more restrictive the constraint is over what worlds are possible, the more variables the implicants assign values to. Our example contains five implicants: $\neg r \wedge g$, $r \wedge \neg g$, $\neg r \wedge \neg g$, $\neg r$, and $\neg g$. Therefore, the deterministic assignment of $p_{\boldsymbol{\mu}}$ will need to contain at least one of $\neg r$ and $\neg g$ for $p_{\boldsymbol{\mu}}$ to be possible.

### 4.3. Conditioned independent distributions

A common counterargument to the claim that independent distributions are biased towards determinism is that we can condition an independent distribution $p_{\boldsymbol{\mu}}$ on the constraint $\varphi$. This is the distribution $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$ where the variables $w_i$ become dependent due to conditioning on the constraint. Furthermore, such a parameterisation is an $n$-dimensional manifold inside the set of possible distributions, which can cover far more distributions than those characterised in Theorem 4.3. For instance, consider the independent distribution $p(r) = p(g) = \frac{1}{2}$. When conditioning on $\varphi = \neg r \vee \neg g$, we get the uniform distribution over worlds $p_1$ from Figure 1. In fact, we can cover the entire green triangle in Figure 1 in this way.

However, this argument does not hold in a *learning* setting where we optimise towards a minimum of the semantic loss (Equation 1). As we already showed, the uniform distribution over worlds $p_1$ can *not* be represented by an independent distribution alone. In fact, there are strict conditions on when an independent distribution $p_{\boldsymbol{\mu}}$ conditioned on the constraint $\varphi$ can be represented by *another* independent distribution $q_{\boldsymbol{\mu}'}$:

**Theorem 4.4** (Representability of $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$)**.** *Let $p_{\boldsymbol{\mu}}$ have $p_{\boldsymbol{\mu}}(\varphi = 1) > 0$ and deterministic assignment $\mathbf{w}_E$. Then the following statements are equivalent:*

1. *the conditional distribution $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$ can be represented by another independent distribution $q_{\boldsymbol{\mu}'}$;*
2. *there is an implicant $\mathbf{w}_D$ that covers all possible worlds in the support of $p_{\boldsymbol{\mu}}$;*
3. *there is an implicant $\mathbf{w}_D$ such that $\mathbf{w}_E, \varphi \models \mathbf{w}_D$.*

*Proof sketch.* Under this condition, we can rewrite the conditional distribution $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$ as an independent distribution where the deterministic assignment is $\mathbf{w}_D$, and the remaining factors are renormalised to sum to 1. □

This theorem is rather subtle. In our example, an independent distribution with deterministic assignment $g$ "entails" $\neg r$: Since $\neg g$ is not true, we need to make $\neg r$ true to be consistent with $\neg r \vee \neg g$. Since $\neg r$ is an implicant, an independent distribution can represent $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$. Any distribution with $p_{\boldsymbol{\mu}}(g) = 1$ will have a conditional distribution at the vertex $(\neg r, g)$, which is representable. Therefore, the distributions for which we can represent $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$ are those that either have $p_{\boldsymbol{\mu}}(g) = 1$ or $p_{\boldsymbol{\mu}}(r) = 1$ (but not both, since then $p_{\boldsymbol{\mu}}(\varphi = 1) = 0$). For general functions,

[4] If the perception model $p_{\boldsymbol{\theta}}$ is a neural network, $\boldsymbol{\mu}$ would be the output of its last layer. We drop the reference to $\boldsymbol{\theta}$ as it is not relevant for our analysis.

this theorem states that a distribution can only represent a conditional independent distribution if the unconditioned distribution is deterministic in some variables.

### 4.4. The geometry of sets of possible independent distributions

While the previous section studied possible independent distributions individually, in the following section we study entire sets of possible distributions from a geometric and topological viewpoint. Our main result proves that all possible independent distributions are on a face of the hypercube $[0,1]^n$ and that we can compute which faces contain possible independent distributions.

We next define the *set of all independent distributions* $\Delta^{\perp\!\!\!\perp} \subseteq \Delta$. We calculate the probability of each world from the parameters $\boldsymbol{\mu} \in [0,1]^n$. Then, we create a vector in the set of all distributions $\Delta$:

$$\Delta^{\perp\!\!\!\perp} = \Big\{ p_{\boldsymbol{\mu}} \in \Delta \mid \boldsymbol{\mu} \in [0,1]^n, \\ p_{\boldsymbol{\mu}_i} = \prod_{j=1}^{n} \mu_j^{w_{i,j}} \cdot (1-\mu_j)^{1-w_{i,j}} \Big\} \tag{4}$$

We consider all parameters of possible distributions in $[0,1]^n$. Then, we compute the probability of each world $\mathbf{w}_i$ using Equation 2 and the Bernoulli mass function to create a vector in $\Delta$. This map from $[0,1]^n$ to $\Delta$ is a bijection (see Lemma G.1), and so is a homeomorphism between the $n$-cube $[0,1]^n$ and $\Delta^{\perp\!\!\!\perp}$. In practice, this means we can study topological properties both in parameter space and distribution space. We will treat $p_{\boldsymbol{\mu}}$ as both a vector in $\Delta$ and a distribution over worlds.

Independent distributions can only represent a subset of the simplex $\Delta$. We aim to understand this subset, and in particular the set of *possible independent distributions* $\Delta_\varphi^{\perp\!\!\!\perp} = \Delta^{\perp\!\!\!\perp} \cap \Delta_\varphi$.

#### 4.4.1. A REPRESENTATION OF $\Delta_\varphi^{\perp\!\!\!\perp}$

We next prove that the set of possible independent distributions $\Delta_\varphi^{\perp\!\!\!\perp}$ is formed by considering the set of all prime implicants of $\varphi$. We find a useful representation of $\Delta_\varphi^{\perp\!\!\!\perp}$ using *cubical sets* (often called *cubical complexes*). Roth (1958) was the first to use cubical sets to develop algorithms that compute efficient representations of boolean functions, noting the relation to implicants. Intuitively, a cubical set is a union of (hyper)cubes of various dimensions. In our representation, we use implicants to create a cube. We then show a cubical set formed from such cubes is the set of possible independent distributions.

The cube associated with an implicant fixes the coordinates of the deterministic variables and uses the interval $[0,1]$ for the free variables. For example, the implicants of the traffic light problem form two cubes: For $\neg r$, the cube is $\{0\} \times [0,1]$ (or: the first is false, and the second is "agnostic") and for $\neg g$, the cube is $[0,1] \times \{0\}$. We next discuss the relevant background for these concepts.

**Definition 4.5** (Elementary cubes)**.** An *elementary interval* $I$ is $\{0\}$, $\{1\}$, or $[0,1]$. An *(elementary) $n$-cube* $C$ is the Cartesian product of $n$ elementary intervals $C = I_1 \times \cdots \times I_n \subseteq [0,1]^n$. We use "cube" to refer only to elementary cubes unless mentioned otherwise. The *dimension* of a cube is the number of elementary intervals $I_i$ that are $[0,1]$. Cubes of dimension 0 are called *vertices* and are points in $\{0,1\}^n$, while cubes of dimension 1 are *edges* that connect two vertices. A *face* $C'$ of a cube $C$ is a cube such that $C' \subseteq C$.

**Definition 4.6** (Cubical sets)**.** $X \subseteq [0,1]^n$ is a *cubical set* if it is the finite union of a set of cubes $\{C_1, ..., C_k\}$ (Kaczynski et al., 2004). With $\mathcal{C}(X)$ we denote all faces of the cubes $\{C_1, ..., C_k\}$, while with $\mathcal{C}_k$ we denote the faces in $\mathcal{C}(X)$ of dimension $k$, called the *$k$-cubes* of $X$. A *facet* $C \in \mathcal{C}(X)$ of $X$ is a cube that is not contained in another cube $C' \in \mathcal{C}(X)$.

Next, we need the notion of prime implicants (Quine, 1959). Informally, an implicant is a prime implicant if, by removing any of its assignments, there will be extensions of the implicant that are impossible worlds.

**Definition 4.7** (Prime implicant)**.** An implicant $\mathbf{w}_D$ of $\varphi$ is a *prime implicant* if its cover $\mathcal{W}_{\mathbf{w}_D}$ is not contained in the cover $\mathcal{W}_{\mathbf{w}_E}$ of another implicant $\mathbf{w}_E$, that is, $\mathcal{W}_{\mathbf{w}_D} \not\subseteq \mathcal{W}_{\mathbf{w}_E}$ for all implicants $\mathbf{w}_E$. With $\mathcal{I} = \{\mathbf{w}_{D_i}\}_{i=1}^m$ we denote the set of all prime implicants of $\varphi$.

The set of prime implicants $\mathcal{I}$ can be found with the first step of the Quine–McCluskey algorithm (Quine, 1952; McCluskey, 1956). It creates a disjunctive normal form of $\varphi$ by considering their disjunction.

Finally, we introduce the cubical set corresponding to $\varphi$.

**Definition 4.8** (Implicant cubes & cubical set of $\varphi$)**.** Each implicant $\mathbf{w}_D$ defines an *implicant cube* $C_{\mathbf{w}_D}$: Its $i$-th elementary interval $I_i$ is $\{\mathbf{w}_{Di}\}$ if $i \in D$, and $[0,1]$ otherwise. The *cubical set $C_\varphi$ of $\varphi$* is the union of all prime implicant cubes $C_\varphi = \bigcup_{\mathbf{w}_D \in \mathcal{I}} C_{\mathbf{w}_D}$.

**Example 4.9.** In the traffic light problem, the prime implicants are $\neg r$ and $\neg g$. $\neg r \wedge \neg g$ is an implicant but is not prime, as two proper subsets are also implicants.

The cubical set is $C_\varphi = \{0\} \times [0,1] \cup [0,1] \times \{0\}$. $\mathcal{C}_0(C_\varphi) = \{(0,0),(0,1),(1,0)\}$ is the vertices, while $\mathcal{C}_1(C_\varphi) = \{\{0\} \times [0,1], [0,1] \times \{0\}\}$ are the edges. The first edge connects $(0,0)$ and $(0,1)$, and the second connects $(0,0)$ and $(1,0)$. This cubical set corresponds to the lines of minimal loss in the left plot of Figure 4.

The implicant cube $C_{\mathbf{w}_D}$ contains the independent parame-

ters $\boldsymbol{\mu}$ for distributions $p_{\boldsymbol{\mu}}$ that deterministically return $\mathbf{w}_D$. We present the basic properties of $C_\varphi$ in Appendix F. The most important results are that the set of cubes $\mathcal{C}(C_\varphi)$ is equal to the set of all implicant cubes. Furthermore, the *prime* implicant cubes are the facets of the cubical set $C_\varphi$. This means we can exactly compute the combinatorial structure of $C_\varphi$ from the prime implicants of $\varphi$.

Our next result states that the cubical set $C_\varphi$ indeed represents the set of possible independent distributions.

**Theorem 4.10** (Representing the set of possible independent distributions)**.** *A parameter $\boldsymbol{\mu}$ is in $C_\varphi$ if and only if the distribution $p_{\boldsymbol{\mu}}$ is possible for $\varphi$. That is, $\boldsymbol{\mu} \in C_\varphi$ if and only if $p_{\boldsymbol{\mu}} \in \Delta_\varphi^{\perp\!\!\!\perp}$. Furthermore, the cubical set $C_\varphi$ cannot be represented as a union of fewer cubes.*

*Proof sketch.* A distribution $p_{\boldsymbol{\mu}}$ using a parameter $\boldsymbol{\mu}$ from implicant cube $C_{\mathbf{w}_D}$ fixes $\mathbf{w}_D$ and allows any value in $[0, 1]$ for the remaining factors. This describes all possible independent distributions by Theorem 4.3, so all possible independent distributions are in the union of implicant cubes.

Next, we show that a parameter $\boldsymbol{\mu}$ that is in the open interval $(0, 1)$ for all stochastic variables of a prime implicant cube cannot be in another (prime) implicant cube. This is because $\mu$ would be in the *relative interior* of $C_{\mathbf{w}_D}$, and we know that the relative interiors of faces of a cubical set are disjoint (Kaczynski et al., 2004). This shows that no smaller set of implicants gets us to $\Delta_\varphi^{\perp\!\!\!\perp}$. □

While we can compute this representation, it can also be rather big. The number of prime implicants can be exponential in the number of variables, and there are formulas with $\Omega(3^n/n)$ prime implicants (Chandra and Markowsky, 1978), above the number of worlds $2^n$. And as we proved, we need *all* prime implicants: A minimal subset of prime implicants that cover all possible worlds (for instance, the prime implicants found in the second step of the Quine–McCluskey method) does not always cover $\Delta_\varphi^{\perp\!\!\!\perp}$. See Appendix B.1 for a counterexample. In addition, computing the combinatorial structure of the cubical set $C_\varphi$ adds a significant combinatorial overhead, as it is generated from the prime implicants.

#### 4.4.2. Convexity of Semantic Loss over Independent Distributions

Next, we study when the semantic loss restricted to independent distributions is convex. We already saw in Figure 4 that even for the simple traffic light formula, the set of possible independent distributions is not convex. We now prove this is almost always the case:

**Theorem 4.11** (Convexity)**.** *The following statements are equivalent: 1) There is exactly one prime implicant of $\varphi$; 2) $C_\varphi$ is convex; 3) the semantic loss over the space of independent distributions $\mathcal{L}(\boldsymbol{\mu})$ is convex.*

*Proof sketch.* If there is exactly one prime implicant, $C_\varphi$ is an implicant cube $C_{\mathbf{w}_D}$, which is clearly convex. If there is more than one prime implicant, we can construct a convex combination of two parameters that is not possible by noting that the deterministic assignment of this convex combination is not an implicant.

The convexity of the semantic loss is proven using Jensen's inequality and noting that we can marginalise out all the stochastic variables. With more than one prime implicant, we note that since its minima $C_\varphi$ are non-convex, certainly the semantic loss must also be non-convex. □

The condition that there is a single prime implicant means that the set of all possible worlds $\mathcal{W}_\varphi$ is described by fixing some variables and letting the other variables be free. This is essentially "supervised learning" on the variables in $D$ and absolutely no supervision for the other variables. This is an uncommon scenario for most neurosymbolic settings, as we can simply resort to standard supervised learning methods.

#### 4.4.3. Connectedness of $\Delta_\varphi^{\perp\!\!\!\perp}$

We next study when the set of all possible independent distributions $\Delta_\varphi^{\perp\!\!\!\perp}$ is connected. For this, we introduce the notion of a *prime implicant graph*:

**Definition 4.12** (Prime implicant graph)**.** Let $\mathcal{G} = (\mathcal{W}_\varphi, \mathcal{E})$ be the *prime implicant graph* of $\varphi$, where the vertices $\mathcal{W}_\varphi$ is the set of possible worlds of $\varphi$ and $\mathcal{E} = \{(\mathbf{w}_1, \mathbf{w}_2) \mid \exists_{\mathbf{w}_D \in \mathcal{I}} : \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}_{\mathbf{w}_D}\}$ is the set of edges.

In this graph, there is an edge between two possible worlds $\mathbf{w}_1$ and $\mathbf{w}_2$ when there is an implicant that covers both $\mathbf{w}_1$ and $\mathbf{w}_2$. In our traffic light example, the prime implicant graph has three vertices. There is an edge between the first and the third ($(0, 1)$ and $(0, 0)$) and the second and the third ($(1, 0)$ and $(0, 0)$). In the case of the XOR function (Appendix B.2) $(a \wedge \neg b) \vee (\neg a \wedge b)$, the graph has two vertices $(1, 0)$ and $(0, 1)$, but no edges.

**Theorem 4.13** (Connectedness)**.** *The connected components of the space of possible distributions $C_\varphi$ correspond to the connected components in $\mathcal{G}$. In particular, $C_\varphi$ is a connected space if and only if $\mathcal{G}$ is connected.*

*Proof sketch.* The vertices of the prime implicant graph $\mathcal{G}$ as points in $\{0, 1\}^n$ directly correspond to the vertices of $C_\varphi$. When an edge exists between $\mathbf{w}_1$ and $\mathbf{w}_2$, both worlds are covered by some prime implicant $\mathbf{w}_D$, and their deterministic components must include $\mathbf{w}_D$, so $\mathbf{w}_1, \mathbf{w}_2 \in C_{\mathbf{w}_D}$. Since all cubes are connected, $\mathbf{w}_1$ and $\mathbf{w}_2$ are connected in $C_{\mathbf{w}_D} \subseteq C_\varphi$. By induction, if there is a path in $\mathcal{G}$ between $\mathbf{w}_1$ and $\mathbf{w}_2$, there is a path in $C_\varphi$ between $\mathbf{w}_1$ and $\mathbf{w}_2$. □

This theorem shows that the connectedness depends on the structure of the constraint. For the traffic light example, $\Delta_{\varphi}^{\perp\!\!\!\perp}$ is connected: the three possible worlds are connected through the two prime implicants $\neg r$ and $\neg g$. However, $\Delta_{\varphi}^{\perp\!\!\!\perp}$ is disconnected for the XOR function, as its prime implicant graph is disconnected. See Appendix B.2 for a visualisation. The popular MNIST Addition task (Example 2.1) is another example: Like XOR, $\Delta_{\varphi}^{\perp\!\!\!\perp}$ is a set of disconnected vertices. This brings challenges to training independent perception models: Each disconnected part of the graph is a different "global optimum", and moving from one global optimum to another will require a large change in parameters while incurring a higher loss.

### 4.5. Computing the homology of $\Delta_{\varphi}^{\perp\!\!\!\perp}$

Our representation of the set of possible independent distributions is in the form of a cubical set. Cubical sets are a useful representation tool for topological spaces in algebraic topology, although simplicial complexes are more common (Hatcher, 2002; Matoušek, 2008). Homology allows us to use combinatorial, algebraic objects to study the topology of cubical sets. In our case, these objects correspond to the implicants of the formula. For finite cubical sets, the problem of computing the homology is solved (Kaczynski et al., 2004). This means we can associate every formula $\varphi$ with a homology which gives all the *holes* in the set. This roughly tells us how "easy" this set is to traverse during optimisation: A set with many holes will require more complicated paths. We give an example of a formula with a hole in Appendix B.3.

The algorithm behind computing the homology from a cubical set is complicated, and involves both (abelian) group theory and linear algebra. We refer the reader to Algorithm 3.78 in Kaczynski et al. (2004), which requires the facets of the cubical set as input. In our case, this corresponds to the set of all prime implicants, found by the Quine–McCluskey algorithm (Quine, 1952). Then we construct matrices that correspond to the boundaries of the cubes, and use linear algebra to compute the Smith normal form. From this matrix, the relevant groups can be constructed.

## 5. Empirical visualisations

To visualise what the possible distributions found by minimising the semantic loss look like, we compare independent distributions and expressive distributions on the traffic light problem in Figure 3. We modelled independent distributions with two real-valued parameters and a sigmoid, and expressive distributions with 4 real-valued parameters and a softmax. Then, we minimise the semantic loss to ensure $p(r, g) = 0$. We use gradient descent for 10,000 iterations with a learning rate of 0.1.

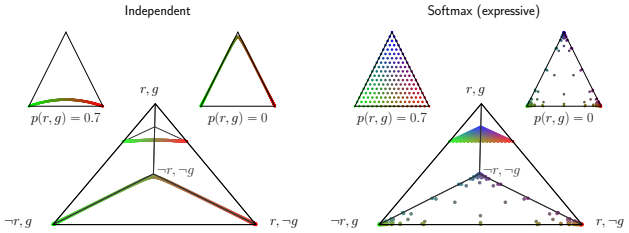


*Figure 3.* The minimisation of the semantic loss on the traffic light problem for independent distributions (left) and expressive distributions (right). The initial distributions have impossible beliefs with $p(r, g) = 0.7$, plotted in the top-left triangle and the top triangle within the tetrahedron. The resulting minima with $p(r, g) = 0$ are in the bottom triangle. Minima of the independent assumption are as predicted by Theorem 4.10. The minima of the expressive parameterisation cover differing areas in the bottom triangle, but are close to the vertices.

We find that the minima of the independent distributions are as predicted by Theorem 4.10: A union of two line segments between the vertices $\neg r, g$ and $\neg r, \neg g$ and the vertices $r, \neg g$ and $\neg r, \neg g$. The minima of the expressive distributions are more uniformly distributed over the bottom triangle, but are biased towards the vertices. Therefore, using an expressive parameterisation is not sufficient to ensure the model is calibrated, which is a common problem in neural networks (Guo et al., 2017). In Appendix C, we also experimented with adding an entropy maximisation loss, which counteracts this bias. This is similar to how BEARS (Marconato et al., 2024) encourages diversity.

## 6. Related work

Many PNL systems use the independence assumption we discussed, such as semantic loss (Xu et al., 2018), DeepProbLog (Manhaeve et al., 2018), DeepStochLog (Winters et al., 2022), A-NeSI (van Krieken et al., 2023), NeurASP (Yang et al., 2020), and Scallop (Huang et al., 2021). As previously mentioned, this makes probabilistic reasoning tractable, as computing Equation (1) is a #P-hard problem in general. While they are not directly comparable, fuzzy methods for neurosymbolic learning also implicitly make this assumption (Serafini and Garcez, 2016; Badreddine et al., 2022; van Krieken et al., 2022). We show that several fuzzy logics also bias towards determinism in Appendix A, although more work is needed to prove that this happens with the same generality as for probabilistic methods.

However, several recent methods are also compatible with more expressive distributions and show significant accuracy improvements compared to methods relying on the independence assumption. The pseudo-semantic loss (Ahmed et al., 2023) uses a pseudo-log-likelihood approximation to the

semantic loss to train autoregressive perception models. NeuPSL (Pryor et al., 2023) uses energy-based models that can perform joint inference over multiple variables. In semantic probabilistic layers, Ahmed et al. (2022a) experiment with an alternative parameterisation that increases expressivity without losing tractability: namely, a *mixture* of independent distributions (Vergari et al., 2021). We study these mixtures thoroughly in Appendix D. In particular, we prove that using two mixture components ensures the minima are connected. However, to be able to mix between an arbitrary number of possible worlds, we need at least as many components as the size of a *minimal* prime implicant cover. This is exponential in the number of variables in general. BEARS (Marconato et al., 2024) increases expressiveness by creating an ensemble of independent models, which has the same expressiveness guarantees as mixtures of independent distributions. BEARS explicitly uses this ensemble to increase uncertainty calibration. Cerutti et al. (2022) consider a Bayesian approach for probabilistic circuits, overcoming the independence assumption to improve the estimation of uncertainty.

The study of the theory of neurosymbolic learning is still in its infancy. Marconato et al. (2023b) discuss *Reasoning Shortcuts*, which are perception models that minimise the semantic loss, yet learn to predict worlds that are different from the ground truth. Since the independence assumption biases to determinism, we hypothesise that independent distributions are more likely than expressive models to converge to a single reasoning shortcut: They cannot properly express uncertainty between different reasoning shortcuts. Several recent papers study how to best deal with reasoning shortcuts (Marconato et al., 2024; Li et al., 2023a).

Furthermore, recent work has studied conditions for the learnability of the perception model (Wang et al., 2023b) and error bounds on its generalisation gap (Wang et al., 2023a). The output layer of the perception model also affects expressivity. If it is low-rank, that is, the number of neurons is lower than the number of outputs, there is an additional decrease in expressivity known as the softmax bottleneck (Yang et al., 2018) or the sigmoid bottleneck in the context of binary outputs (Grivas et al., 2024). Our results, in particular Theorem 4.13, are also related to the *connectivity barrier* in Monte Carlo approaches to neurosymbolic learning over independent models (Li et al., 2023b). A future study into this relation may provide insights in how to speed up Monte Carlo methods in this setting.

## 7. Conclusion

We studied the independence assumption in neurosymbolic learning, which characterises several popular methods. We proved that this assumption biases neurosymbolic models towards deterministic solutions. As a result, they lack the ability to express uncertainty about multiple possibly valid options. We then used tools from logic and computational homology to study the structure of the set of possible independent distributions, and showed it is non-convex and disconnected in general.

In future work, we want to study practical methods for expressive neurosymbolic learning that properly represent uncertainty about different valid worlds. Dropping the independence assumption means that inference becomes much more complex, so a thorough study of appropriate (approximate) inference methods is needed. Our theory can be extended to a thorough study of the trade-off between expressivity and tractability. Our analysis of the mixture of independent distributions in Appendix D gives a stepping stone towards this goal. Another option is to consider constraints on continuous variables. Furthermore, a thorough study of the homology discussed in Section 4.5 may reveal further insights into the topology of the set of possible distributions.

## Impact statement

This paper presents foundational work to understand and advance the field of neurosymbolic machine learning. As such, there could be many potential societal consequences of applications of our work, none of which, we feel, can be easily predicted and specifically highlighted here.

## Acknowledgements

Emile van Krieken was funded by ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence), EPSRC (grant no. EP/W002876/1). Pasquale Minervini was partially funded by ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence), EPSRC (grant no. EP/W002876/1), an industry grant from Cisco, and a donation from Accenture LLP. Antonio Vergari was supported by the "UNREAL: Unified Reasoning Layer for Trustworthy ML" project (EP/Y023838/1) selected by the ERC and funded by UKRI EPSRC. We thank Emanuele Marconato, Andreas Grivas, Patrick Koopmann, Thiviyan Thanapalasingam, Javaloy, Nicola Branchini, Leander Kurscheidt, Frank van Harmelen, Annette ten Teije, Eleonora Giunchiglia, Alessandro Daniele, Samy Badreddine, Siegfried Nijssen, Stefano Teso, and Sagar Malhotra for productive discussions and feedback while writing this work. We also thank the anonymous reviewers for their valuable feedback.

## References

Kareem Ahmed, Stefano Teso, Kai-Wei Chang, Guy Van den Broeck, and Antonio Vergari. Semantic probabilistic

layers for neuro-symbolic learning. *Advances in Neural Information Processing Systems*, 35:29944–29959, 2022a.

Kareem Ahmed, Eric Wang, Kai-Wei Chang, and Guy Van den Broeck. Neuro-Symbolic Entropy Regularization. In *The 38th Conference on Uncertainty in Artificial Intelligence*, June 2022b.

Kareem Ahmed, Kai-Wei Chang, and Guy Van den Broeck. A pseudo-semantic loss for autoregressive models with logical constraints. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic Tensor Networks. *Artificial Intelligence*, 303:103649, February 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103649.

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable Neural-Symbolic Concept Reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1801–1825. PMLR, July 2023.

Federico Cerutti, Lance M. Kaplan, Angelika Kimmig, and Murat Şensoy. Handling epistemic and aleatory uncertainties in probabilistic circuits. *Machine Learning*, 111(4):1259–1301, April 2022. ISSN 1573-0565. doi: 10.1007/s10994-021-06086-4.

Ashok Chandra and George Markowsky. On the Number of Prime Implicants. *Discrete Mathematics*, 24(1):7–11, January 1978. doi: 10.1016/0012-365X(78)90168-1.

Mark Chavira and Adnan Darwiche. On probabilistic inference by weighted model counting. *Artificial Intelligence*, 172(6):772–799, April 2008. ISSN 0004-3702. doi: 10.1016/j.artint.2007.11.002.

Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep Learning with Logical Constraints. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5478–5485. ijcai.org, 2022. doi: 10.24963/ijcai.2022/767.

Eleonora Giunchiglia, Mihaela Cătălina Stoian, Salman Khan, Fabio Cuzzolin, and Thomas Lukasiewicz. ROAD-R: The autonomous driving dataset with logical requirements. *Machine Learning*, 112(9):3261–3291, September 2023. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-023-06322-z.

Andreas Grivas, Antonio Vergari, and Adam Lopez. Taming the sigmoid bottleneck: Provably argmaxable sparse multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12208–12216, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

Jiani Huang, Ziyang Li, Binghong Chen, Karan Samel, Mayur Naik, Le Song, and Xujie Si. Scallop: From Probabilistic Deductive Databases to Scalable Differentiable Reasoning. In *Advances in Neural Information Processing Systems*, May 2021.

Tomasz Kaczynski, Konstantin Mischaikow, and Marian Mrozek. *Computational Homology*, volume 157 of *Applied Mathematical Sciences*. Springer, New York, NY, 2004. ISBN 978-1-4419-2354-7 978-0-387-21597-6. doi: 10.1007/b97315.

Zenan Li, Zehua Liu, Yuan Yao, Jingwei Xu, Taolue Chen, Xiaoxing Ma, and Jian L\"{u}. Learning with logical constraints but without shortcut satisfaction. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=M2unceRvqhh.

Zenan Li, Yuan Yao, Taolue Chen, Jingwei Xu, Chun Cao, Xiaoxing Ma, and Jian L\"{u}. Softened Symbol Grounding for Neuro-symbolic Systems. In *The Eleventh International Conference on Learning Representations*, February 2023b.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural probabilistic logic programming. In Samy Bengio, Hanna M Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, 2018.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence*, 298:103504, 2021. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103504.

Emanuele Marconato, Stefano Teso, and Andrea Passerini. Neuro-Symbolic Reasoning Shortcuts: Mitigation Strategies and their Limitations, March 2023a.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not All Neuro-Symbolic Concepts Are Created Equal: Analysis and Mitigation of Reasoning Shortcuts. In *Thirty-Seventh Conference on Neural Information Processing Systems*, May 2023b.

Emanuele Marconato, Samuele Bortolotti, Emile van Krieken, Antonio Vergari, Andrea Passerini, and Stefano Teso. Bears make neuro-symbolic models aware of their reasoning shortcuts, 2024.

Jiří Matoušek. *Using the Borsuk–Ulam Theorem*. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-00362-5 978-3-540-76649-0. doi: 10.1007/978-3-540-76649-0.

Edward J McCluskey. Minimization of boolean functions. *The Bell System Technical Journal*, 35(6):1417–1444, 1956.

Connor Pryor, Charles Dickens, Eriq Augustine, Alon Albalak, William Yang Wang, and Lise Getoor. NeuPSL: Neural Probabilistic Soft Logic. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4145–4153, Macau, SAR China, August 2023. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/461.

W. V. Quine. The Problem of Simplifying Truth Functions. *The American Mathematical Monthly*, 59(8):521–531, 1952. ISSN 0002-9890. doi: 10.2307/2308219.

Willard V Quine. On cores and prime implicants of truth functions. *The American Mathematical Monthly*, 66(9): 755–760, 1959.

J. Paul Roth. Algebraic Topological Methods for the Synthesis of Switching Systems. I. *Transactions of the American Mathematical Society*, 88(2):301–326, 1958. ISSN 0002-9947. doi: 10.2307/1993216.

Luciano Serafini and Artur D.Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. *CEUR Workshop Proceedings*, 1768, 2016. ISSN 16130073.

Emile van Krieken, Erman Acar, and Frank van Harmelen. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence*, 302:103602, 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103602.

Emile van Krieken, Thiviyan Thanapalasingam, Jakub M. Tomczak, Frank van Harmelen, and Annette ten Teije. A-NeSI: A Scalable Approximate Method for Probabilistic Neurosymbolic Inference. In *Thirty-Seventh Conference on Neural Information Processing Systems*. arXiv, May 2023.

Antonio Vergari, YooJung Choi, Anji Liu, Stefano Teso, and Guy Van den Broeck. A compositional atlas of tractable circuit operations for probabilistic inference. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 13189–13201. Curran Associates, Inc., 2021.

Kaifu Wang, Hangfeng He, Tin D. Nguyen, Piyush Kumar, and Dan Roth. On regularization and inference with label constraints. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 35740–35762. PMLR, 2023a.

Kaifu Wang, Efi Tsamoura, and Dan Roth. On Learning Latent Models with Multi-Instance Weak Supervision. In *Thirty-Seventh Conference on Neural Information Processing Systems*, June 2023b.

Thomas Winters, Giuseppe Marra, Robin Manhaeve, and Luc De Raedt. Deepstochlog: Neural stochastic logic programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10090–10100, 2022.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy den Broeck. A semantic loss function for deep learning with symbolic knowledge. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5502–5511, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkwZSG-CZ.

Zhun Yang, Adam Ishay, and Joohyung Lee. NeurASP: Embracing neural networks into answer set programming. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1755–1762. International Joint Conferences on Artificial Intelligence Organization, July 2020. doi: 10.24963/ijcai.2020/243.

Günter M. Ziegler. *Lectures on polytopes*. Springer-Verlag, New York, 1995.

| Symbol | Description | Reference |
|---|---|---|
| $\mathbf{w}$ | A world in $\{0,1\}^n$ | Sec. 2 |
| $\mathbf{x}$ | An high-dimensional input | Sec. 2 |
| $\boldsymbol{\theta}$ | Parameter of the perception model | Sec. 2 |
| $\varphi$ | A constraint of $\{0,1\}^n \to \{0,1\}$ | Sec. 2 |
| $p_{\boldsymbol{\theta}}(\mathbf{w}\|\mathbf{x})$ | The perception model | Sec. 2 |
| $\mathcal{L}(\boldsymbol{\theta})$ | The semantic loss | Sec. 2 |
| $\mathcal{W}$ | The set of all worlds $\{0,1\}^n$ | Sec. 3 |
| $\mathcal{W}_\varphi$ | The set of possible worlds | Sec. 3 |
| $\Delta$ | The set of all distributions over worlds | Sec. 4.1 |
| $\Delta_\varphi$ | The set of possible distributions | Sec. 4.1 |
| $\Delta^{\perp\!\!\!\perp}$ | The set of independent distributions | Sec. 4.4 |
| $\Delta_\varphi^{\perp\!\!\!\perp}$ | The set of possible independent distributions | Sec. 4.4 |
| $\boldsymbol{\mu}$ | A parameter for an independent distribution in $[0,1]^n$ | Sec. 4.2 |
| $p_{\boldsymbol{\mu}}(\mathbf{w})$ | An independent distribution parameterised by $\boldsymbol{\mu}$ | Sec. 4.2 |
| $\mathbf{w}_D$ | A partial assignment to the variables in $D$. Usually an implicant | Def. 4.1 and 4.2 |
| $\mathcal{W}_{\mathbf{w}_D}$ | The cover of a partial assignment (all worlds that extend it) | Def. 4.2 |
| $\mathcal{I}$ | The set of all prime implicants | Def. 4.7 |
| $C$ | An (elementary) cube | Def. 4.5 |
| $\mathcal{C}(X)$ | The set of all cubes in a cubical set $X$ | Def. 4.6 |
| $C_{\mathbf{w}_D}$ | The cube corresponding to an implicant | Def. 4.8 |
| $C_\varphi$ | The cubical set of parameters in $[0,1]^n$ that satisfy the knowledge $\varphi$ | Def. 4.8 |
| $\mathcal{G}$ | The prime implicant graph | Def. 4.12 |

*Table 1.* **Table of notation used in the paper**. We use bold symbols $\mathbf{x}$, $\mathbf{w}$ to denote vectors, both real and boolean. We use $p$ and $q$, possibly parameterised, to refer to distributions over $\mathbf{w} \in \{0,1\}^n$. Since these correspond to vertices in $\Delta$ (the simplex over all possible worlds), we will treat $p$ as both a vector and a distribution.

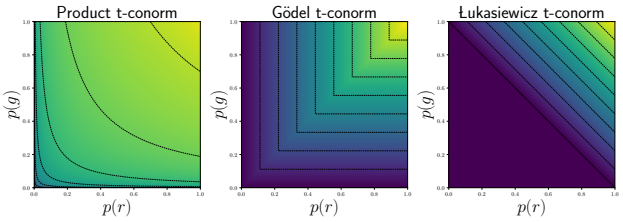


*Figure 4.* Plots of neurosymbolic loss functions for the formula $\neg r \vee \neg g$ using several t-norms. Left: Product t-norm, computed as $-\log p_{\boldsymbol{\theta}}(\varphi|\mathbf{x})$. This coincides with the semantic loss. Center: The Gödel t-conorm $1 - \max(1-r, 1-g)$. Right: The Łukasiewicz t-conorm $1 - \min(1, 2-r-g)$.

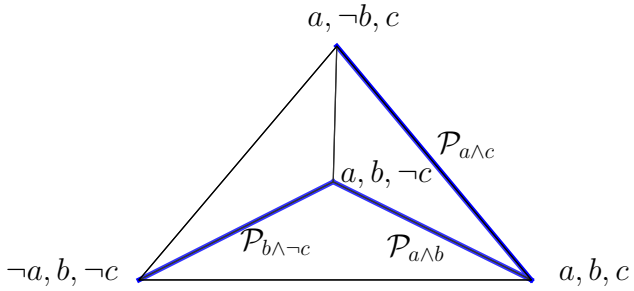


*Figure 5.* The full 3-simplex over possible worlds and the set of possible independent distributions in blue for the formula discussed in Section B.1. The $\mathcal{P}_\phi$ labels denote the set of distributions characterised by the implicant $\phi$, as defined in Proposition 4.10.

## A. Bias towards determinism for Fuzzy Logic

**Fuzzy Neurosymbolic Learning.** While our paper focuses on probabilistic methods, we shortly introduce relevant background about fuzzy neurosymbolic learning (FNL). Roughly, FNL methods construct a fuzzy evaluation function $e_\varphi : [0,1]^n \to [0,1]$. $e_\varphi$ maps independent probability distributions to fuzzy truth values in $[0,1]$ by relaxing the logical connectives to operators on $[0,1]$ (Badreddine et al., 2022). If the distribution is deterministic, then the fuzzy truth becomes binary truth. For a discussion on fuzzy relaxations, see (van Krieken et al., 2022). We limit our discussion to the three common fuzzy disjunctions (t-conorms):

$$\text{Product: } a \vee b = a + b - a \cdot b,$$
$$\text{Gödel: } a \vee b = \max(a, b),$$
$$\text{Łukasiewicz: } a \vee b = \min(1, a + b)$$

We plot the truth values of three common t-conorms for this formula in Figure 4 as a function of $p(r)$ and $p(g)$. The product t-conorms and Gödel t-conorms have their minima at the lines $p(r) = 0$ and $p(g) = 0$, and have a similar biasing effect as the semantic loss.

The Łukasiewicz t-conorm is minimised when $p(r)+p(g) \leq 0.5$ and does not bias towards a deterministic choice. This may explain why Łukasiewicz t-conorms are often more effective in realistic settings (Giunchiglia et al., 2023). However, the Łukasiewicz logic has other problems, such as vanishing gradients (van Krieken et al., 2022) and the fact they do not converge to solutions where $p(\varphi = 1) = 1$.

## B. Additional examples

In this appendix, we plot several example formulas that illustrate the theory discussed in the paper.

### B.1. Minimal covers of prime implicants do not cover all possible independent distributions

In Proposition 4.10, we showed that the set of *all* prime implicants is necessary to cover all possible independent distributions. In this appendix, we give a counterexample to the idea that a *minimal* cover of prime implicants might be sufficient to cover all possible independent distributions. Such minimal covers are computed in the second step of the Quine–McCluskey algorithm (Quine, 1952; McCluskey, 1956) to minimise the description length of the boolean formula.

**Definition B.1.** A set of prime implicants $\mathcal{I}$ is a *cover of* $\varphi$ if $\bigcup_{\mathbf{w}_D \in \mathcal{I}} \mathcal{W}_{\mathbf{w}_D} = \mathcal{W}_\varphi$, that is, the union of their cover is equal to the set of all possible worlds. A cover is *minimal* if no smaller set of prime implicants is also a cover.

Consider a boolean formula on three variables with possible worlds $\{(a,b,c), (a,b,\neg c), (\neg a,b,\neg c), (a,\neg b,c)\}$. We visualize the full simplex over possible worlds and the set of possible independent distributions in Figure 5. The prime implicants of this formula are $\{b \wedge \neg c, a \wedge c, a \wedge b, \}$. The minimal cover of prime implicants is $\{b \wedge \neg c, a \wedge c\}$: The worlds $a \wedge b$ covers are $a \wedge b \wedge c$, which is also covered by prime implicant $a \wedge c$, and $a, b, \neg c$, which is also covered by prime implicant $b \wedge \neg c$. However, by Theorem 4.3, the distribution that deterministically assigns $a \wedge b$, but gives 0.5 probability to $c$, can be represented only using the prime implicant $a \wedge b$: The other two prime implicants cannot represent distributions where $c$ is stochastic. Therefore, minimal covers of prime implicants do not cover all possible independent distributions.

### B.2. The XOR formula has disconnected minima

In Figure 6, we plot the semantic loss under the independence assumption for the XOR formula $\varphi = (a \wedge \neg b) \vee (\neg a \wedge b)$. Note that the minima of this function are in $(1,0)$ and $(0,1)$, since the prime implicants are $a \wedge \neg b$ and

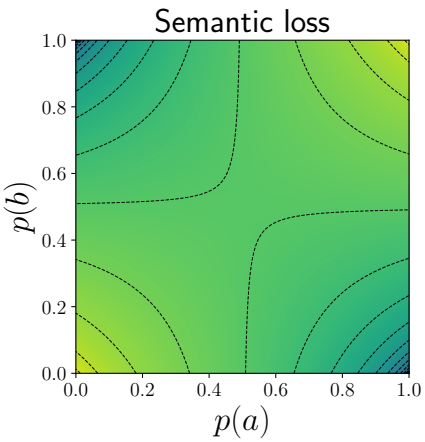


*Figure 6.* The loss landscape of the semantic loss under the independence assumption for the XOR formula $\varphi = (a \wedge \neg b) \vee (\neg a \wedge b)$.

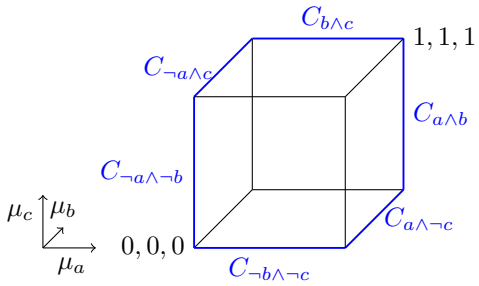


*Figure 7.* An example of a formula where the set of possible independent distributions has a hole. The formula is $\varphi = (\neg a \wedge \neg b) \vee (\neg a \wedge c) \vee (b \wedge c) \vee (a \wedge b) \vee (a \wedge \neg c) \vee (\neg b \wedge \neg c)$. The blue lines correspond to the set of possible independent distributions. $C_\phi$ is an implicant cube.

$\neg a \wedge b$. These are clearly disconnected minima, and to move from one minimum to the other, we would have to traverse through the saddle point at $(0.5, 0.5)$, meanwhile incurring a significantly high loss.

### B.3. The set of possible independent distributions can have holes

In Figure 7 we show that there are formulas for which the set of possible independent distributions $\Delta_\varphi^{\perp\!\!\!\perp}$ has holes. Here, the formula $\varphi = (\neg a \wedge \neg b) \vee (\neg a \wedge c) \vee (b \wedge c) \vee (a \wedge b) \vee (a \wedge \neg c) \vee (\neg b \wedge \neg c)$ is defined by a disjunction of prime implicants. We choose the prime implicants carefully so that each face of the cube has exactly 2 edges. The set highlighted in blue cannot be shrunk to a single point: It is a hole in space. This hole will be detected by algorithms that compute the homology of $\Delta_\varphi^{\perp\!\!\!\perp}$ (see Section 4.5). How relevant is this to optimisation? The presence of holes means there is a cycle between different points, meaning we can move from one point to another in multiple ways. In our example, if one were to remove one of the prime implicants from the formula, there is only one path through $\Delta_\varphi^{\perp\!\!\!\perp}$.

## C. Entropy regularisation helps to calibrate expressive models

We repeat the experiments in Section 5 for both the independent and the softmax model with entropy regularisation. We note that we *maximise* entropy, instead of minimising the entropy, like in Neuro-Symbolic Entropy Regularisation (Ahmed et al., 2022b). In particular, we use the loss function

$$\mathcal{L}_\alpha(\boldsymbol{\theta}) = (1-\alpha)\mathcal{L}(\boldsymbol{\theta}) - \alpha H(p_{\boldsymbol{\theta}}|\varphi), \tag{5}$$

where we compute $H(p_{\boldsymbol{\theta}}|\varphi) = \frac{1}{3}(p_{\boldsymbol{\theta}}(\neg r, g) + p_{\boldsymbol{\theta}}(r, \neg g) + p_{\boldsymbol{\theta}}(\neg r, \neg g))$.

We plot the results in Figure 8 for the independent model and various values of $\alpha$. We see that the entropy regularisation does not help to calibrate the model. Rather, it biases the model more towards the $\neg r, \neg g$ vertex. For larger values of $\alpha$, the minimum of the augmented loss no longer is a minimum of the semantic loss. In fact, for $\alpha = 0.1$, all initial points converge to a minimum that floats just above the bottom triangle. It assigns a probability of 0.023 to the impossible world $r, g$.

The intuition for this is that the entropy regularisation is minimised only at the uniform belief $p_1$ from Figure 1, which is not representable by independent distributions. Then, by minimising the augmented loss $\mathcal{L}_\alpha$, it trades off the closest point to $p_1$ and staying close to the bottom triangle.

For expressive distributions, the effect of entropy regularisation is quite different, which we plot in Figure 9. Again, the parameter $\alpha$ trades off the original minima, which are close to the vertices in the bottom triangle, to the uniform distribution $p_1$. For $\alpha = 0.1$, the minima are indeed all close to $p_1$. However, for appropriate values of $\alpha$ such as $\alpha = 0.01$, the minima almost distribute perfectly over the bottom triangle, and it almost finds the conditioned version of the original distribution. We note that finding the parameter $\alpha$ to get this behaviour would be extremely challenging in practice.

## D. The mixture of independent distributions

In this appendix, we study the mixture of independent distributions with $k$ components. The main results here are that for $k \geq 2$, the space of possible distributions is connected. Furthermore, we provide bounds on the number of components needed to completely fill the space of all possible distributions $\Delta_\varphi$. A lower bound is the number of disconnected components in the prime implicant graph, and an upper bound is the number of prime implicants. This lower bound can be tricky: For example, the number of

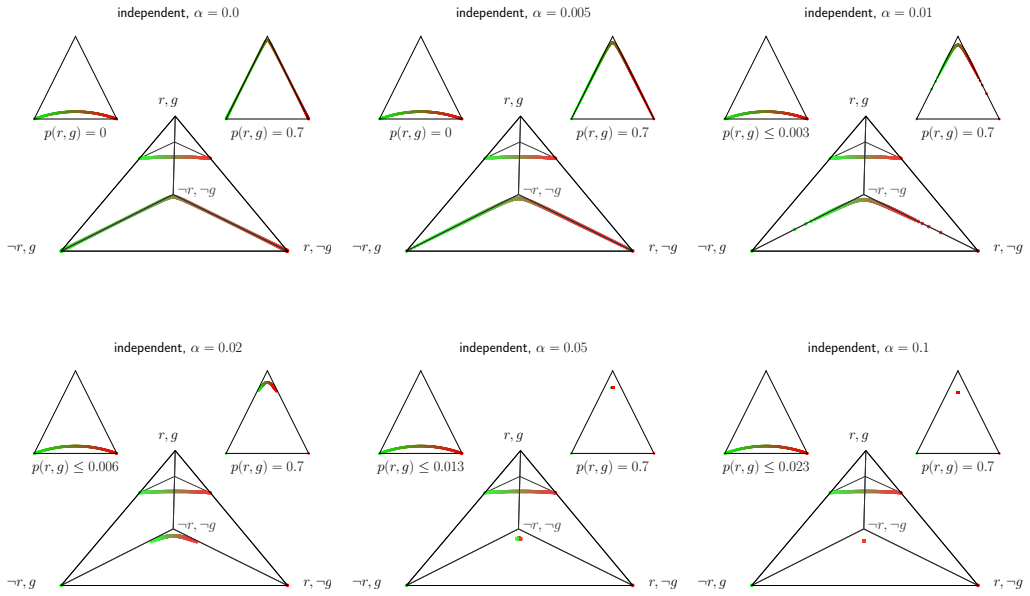


*Figure 8.* Minimising the semantic loss with entropy regularisation for the independent model.

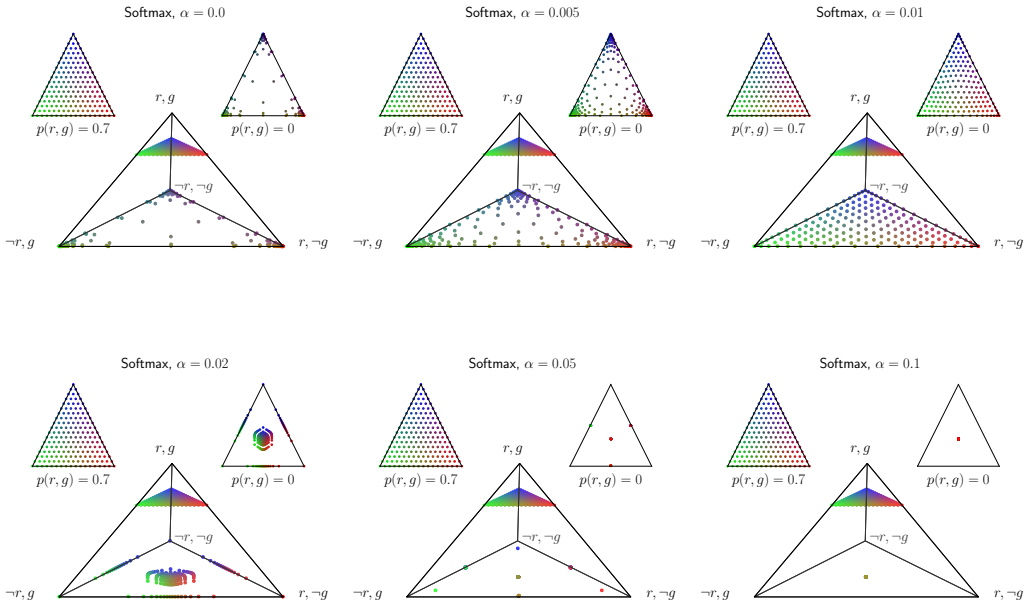


*Figure 9.* Minimising the semantic loss with entropy regularisation for the joint model.

disconnected components in the MNIST Addition task is exponential in the number of digits.

The parameter space of this distribution is the Cartesian product $\Theta_k = \Delta^k \times [0,1]^{k \cdot n}$, where $\Delta^k$ is the $k+1$-dimensional simplex. We map this parameter space to the space of distributions over worlds $\Delta$ with the map $f_{k \cdot \perp\!\!\!\perp} : \Theta_k \to \Delta$, which we define as:

$$f_{k \cdot \perp\!\!\!\perp}(\boldsymbol{\alpha}, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k)_i = \sum_{m=1}^{k} \alpha_m f_{\perp\!\!\!\perp}(\boldsymbol{\mu}_m)_i \tag{6}$$

where $f_{\perp\!\!\!\perp}$ is defined as in Equation 7. Unlike $f_{\perp\!\!\!\perp}$, this is not a bijection, as multiple parameterisations can map to the same distributions. We will refer to $p_{\boldsymbol{\theta}} = f_{k \cdot \perp\!\!\!\perp}(\boldsymbol{\theta})$ as a vector in $\Delta$.

**Lemma D.1.** *Consider a parameter $\boldsymbol{\theta} \in \Theta_k$ of a mixture of $k$ independent components. Then $f_{k \cdot \perp\!\!\!\perp}(\boldsymbol{\theta})$ is a possible distribution if and only if all components $i$ such that $\alpha_i > 0$ have a deterministic assignment that is an implicant.*

*Proof.* The mixture of independents assigns some mass to all independent distributions with $\alpha_i > 0$. For such an independent distribution to be possible, by Proposition 4.3, the deterministic assignment of this component has to be an implicant. Conversely, assume there is a component with $\alpha_i > 0$ that is not a possible distribution. Then, it assigns some mass to an impossible world, and so must the mixture of components. Therefore, the mixture is not a possible distribution. □

Let us now discuss the set of possible distributions for the mixture distribution $\mathcal{P}_{k,\varphi} = f_{k \cdot \perp\!\!\!\perp}(\Theta_k) \cap \Delta_\varphi$. Conveniently, if $k \geq 2$, the set of minima of the semantic loss under the mixture distribution is connected:

**Proposition D.2.** *The set of possible mixture distributions $\mathcal{P}_{k,\varphi}$ is connected for $k \geq 2$.*

*Proof.* Since $\Delta_\varphi^{\perp\!\!\!\perp} \subseteq \mathcal{P}_{k,\varphi}$, any two points that are connected in $\Delta_\varphi^{\perp\!\!\!\perp}$ are also connected in $\Delta_{\varphi\ k}^{\perp\!\!\!\perp}$.

Consider two points $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \Delta_\varphi^{\perp\!\!\!\perp}$ that are not connected. Then we can create a convex combination between $p_{\boldsymbol{\mu}_1}$ and $p_{\boldsymbol{\mu}_2}$ in $\mathcal{P}_{k,\varphi}$ by moving $\alpha_1$ from 1 to 0 and $\alpha_2$ from 0 to 1. This convex combination is a possible distribution in $\mathcal{P}_{k,\varphi}$ by Lemma D.1.

Consider two parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_k$. We will construct a path from $p_{\theta_1}$ to $p_{\theta_2}$ through $\mathcal{P}_{k,\varphi}$. First, choose a component $i \in \{1, ..., k\}$ such that $\alpha_{1,i} > 0$. Next, continuously map the convex mixture parameter from $\boldsymbol{\alpha}_1$ to $\mathbf{e}_i$ (the $i$-th standard normal vector). Call the resulting parameter $\hat{\boldsymbol{\theta}}_1$. Since we do not change the mixture components themselves, we never leave $\mathcal{P}_{k,\varphi}$ by Lemma D.1. Then, $p_{\hat{\boldsymbol{\theta}}_1} = f_{\Delta^{\perp\!\!\!\perp}}(\boldsymbol{\mu}_{1,i}) = p_{\boldsymbol{\mu}_{1,i}}$ is a possible independent distribution. Consider some component $j \in \{1, ..., k\}$ such that $\alpha 2, j > 0$. As argued in the paragraphs above, there is a path between $p_{\boldsymbol{\mu}_{1,i}}$ and $p_{\boldsymbol{\mu}_{2,j}}$ in $\mathcal{P}_{k,\varphi}$. Consider $\hat{\boldsymbol{\theta}}_2$ to be $\boldsymbol{\theta}_2$ but replacing $\boldsymbol{\alpha}$ with $\mathbf{e}_j$ such that $p_{\hat{\boldsymbol{\theta}}_2} = p_{\boldsymbol{\mu}_{2,j}}$. Then, we continuously map $\mathbf{e}_j$ to $\alpha_{2,j}$ to finally arrive at $\boldsymbol{\theta}_2$. □

Clearly, increasing the number of components is beneficial to further covering the complete set of all possible distributions $\Delta_\varphi$. But how parameter-efficient is the use of mixtures of independent distributions to cover this set? First, we prove a straightforward lemma.

**Lemma D.3.** *Consider a parameter $\boldsymbol{\theta} \in \Theta_k$ of a mixture of $k$ independent components such that $p_{\boldsymbol{\theta}} \in \mathcal{P}_{k,\varphi}$. Then $p_{\boldsymbol{\theta}_i} > 0$ if and only if there is a component $m \in \{1, ..., k\}$ such that $\alpha_m > 0$ and the deterministic assignment of $\boldsymbol{\mu}_m$ covers $\mathbf{w}$.*

*Proof.* Let $p_{\boldsymbol{\theta}_i} > 0$. Then by Equation 6 there must be a $m \in \{0, ..., k\}$ with $\alpha_m > 0$ such that $f_{\perp\!\!\!\perp}(\boldsymbol{\mu}_m)_i > 0$. But then, by Proposition 4.3, the deterministic assignment of $\boldsymbol{\mu}_m$ covers $\mathbf{w}$.

Similarly, let the deterministic assignment of $\boldsymbol{\mu}_m$ cover $\mathbf{w}$ and let $\alpha_m > 0$. Then by Proposition 4.3, $f_{\perp\!\!\!\perp}(\boldsymbol{\mu}_m)_i > 0$. But then by Equation 6, $p_{\boldsymbol{\theta}_i} > 0$. □

We next prove a significant lower bound:

**Theorem D.4.** *The minimal number of mixture components needed to assign some probability to all possible worlds is the number of prime implicants in a* minimal cover *of prime implicants $\mathcal{I}$.*

*Proof.* First, we prove that if a mixture distribution can assign some probability to all possible worlds, then it has at least $|\mathcal{I}|$ components. Assume otherwise. Then $|\mathcal{I}| - 1$ components are enough to cover $\Delta_\varphi$. Consider some distribution $p \in \Delta_\varphi$ such that $p_i > 0$ for all possible worlds $\mathbf{w}_i$. Let $\boldsymbol{\theta} \in \Theta_{|\mathcal{I}|-1}$ be parameters such that $p_{\boldsymbol{\theta}} = p$, which have to exist by assumption.

Let $\mathcal{I}'$ be the implicants formed from the independent parameters $\boldsymbol{\mu}_i$, $i \in \{1, ..., I-1\}$. By Lemma D.3, the set of worlds such that $p_{\boldsymbol{\theta}}(\mathbf{w}) > 0$ is $\mathcal{W}' = \bigcap_{\mathbf{w}_D \in \mathcal{I}'} \mathcal{W}_{\mathbf{w}_D}$. This set must equal $\mathcal{W}_\varphi$ by the assumption that $p_i > 0$ for all possible worlds. But this is a contradiction, since this would make the set of implicants $\mathcal{I}'$ a cover of $\varphi$ with $|\mathcal{I}| - 1$ components, which is smaller than the minimal cover of prime implicants $\mathcal{I}$.

Next, we prove that if the number of components is at least $|\mathcal{I}|$, then we can assign some probability to all possible worlds. Define an order $\mathbf{w}_{D1}, ..., \mathbf{w}_{D|\mathcal{I}|}$ of a minimal cover of prime implicants. Let $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_{|\mathcal{I}|}$ be independent parameters such that $\boldsymbol{\mu}_i$ is in the relative interior of $C_{\mathbf{w}_{D\,i}}$

(see Theorem 4.10 and its proof for a rigorous definition). Use $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_{|\mathcal{I}|}$ together with $\boldsymbol{\alpha}$ such that $\alpha_i > 0$ for all $i \in \{0, ..., |\mathcal{I}|\}$ to define parameters $\boldsymbol{\theta} \in \Theta_{|\mathcal{I}|}$. Then, by Lemma D.3, $p_{\boldsymbol{\theta}}$ assigns some probability to all possible worlds. □

Interestingly, here a minimal cover of prime implicants is relevant, while for Theorem 4.10, we needed to consider the set of *all* prime implicants (see also Appendix B.1). Figure 5 provides some intuition: $b \wedge \neg c$ and $a \wedge c$ form a minimal set of prime implicants. By mixing between points on the line segments $\mathcal{P}_{b\wedge\neg c}$ and $\mathcal{P}_{a\wedge c}$, we can, in fact, cover the entire set of possible worlds.

Clearly, $|\mathcal{I}|$ is also a lower bound on the number of components needed to completely cover $\varphi$. A simple upper bound for the number of components needed to mix is $|\mathcal{W}_\varphi|$, since we can put a (deterministic) independent distribution on each of the possible worlds and mix them via $\boldsymbol{\alpha}$. Another lower bound is $\lceil |\mathcal{W}_\varphi|/(n+1) \rceil$, since $f_{k,\perp\!\!\!\perp}(\Theta_k)$ is at most a $k \cdot (n+1)$-dimensional subspace of $\Delta$.

Given Theorem D.4 and the other bounds, is using a mixture of independents a parameter-efficient way to allow perception models to express more uncertainty? We would argue not, at least not in general. For example, the size of the minimal cover for MNIST Addition grows exponentially with the number of digits considered, in fact, it is equal to the number of possible worlds. But then we are using $|\mathcal{W}_\varphi| \cdot (n+1)$ parameters, which are $n$ times more parameters than necessary: The space of possible distributions is a $|\mathcal{W}_\varphi|$-dimensional subspace of $\Delta$.

## E. Convexity of semantic loss

In this Appendix, we show that the semantic loss is a convex loss over the space of all possible distributions $\Delta$ using Jensen's inequality and the fact that the WMC in Equation 1 is linear. Note that this does not mean it is convex with respect to the parameters $\boldsymbol{\theta}$ of the perception model. Let $p_1, p_2 \in \Delta$. Note that since $\Delta$ is a convex set, $\lambda p_1 + (1-\lambda)p_2 \in \Delta$. Then,

$$
\begin{aligned}
&\mathcal{L}(\lambda p_1 + (1-\lambda)p_2) \\
=& -\log\Big(\sum_{\mathbf{w}\in\mathcal{W}_\varphi} \lambda p_1(\mathbf{w}) + (1-\lambda)p_2(\mathbf{w})\Big) \\
=& -\log\Big(\lambda \sum_{\mathbf{w}\in\mathcal{W}_\varphi} p_1(\mathbf{w}) + (1-\lambda)\sum_{\mathbf{w}\in\mathcal{W}_\varphi} p_2(\mathbf{w})\Big) \\
\leq& -\lambda \log \sum_{\mathbf{w}\in\mathcal{W}_\varphi} p_1(\mathbf{w}) - (1-\lambda)\log \sum_{\mathbf{w}\in\mathcal{W}_\varphi} p_2(\mathbf{w}) \\
=& \lambda\mathcal{L}(p_1) + (1-\lambda)\mathcal{L}(p_2)
\end{aligned}
$$

## F. Cubical sets generated by prime implicants

To help understand our results geometrically and prove some of the main theorems in Appendix G, we study the basic properties of the cubical set $C_\varphi$. For background on polytopes, faces, face posets, and polyhedral complexes, see Ziegler (1995), and for an introduction and basic properties of cubical sets, see Kaczynski et al. (2004).

First, we define elementary cells, which allow us to access the relative interior of a cube by changing intervals from a closed set to an open set.

**Definition F.1.** Associated with each cube $C = I_1 \times ... \times I_n$ is an *(elementary) cell* $\mathring{C} = \mathring{I}_1 \times ... \times \mathring{I}_n \subseteq C$, where each $\mathring{I}_i = I_i$ for the degenerate intervals $[0,0]$ and $[1,1]$, and $\mathring{I}_i = (0,1)$ for the nondegenerate interval $[0,1]$.

The following proposition allows us to associate implicants to faces of $C_\varphi$.

**Proposition F.2.** *The faces $\mathcal{C}(C_\varphi)$ of $C_\varphi$ is the set of implicant cubes.*

*Proof.* Consider some face $X \in \mathcal{C}(C_\varphi)$. Since by definition a face is an (elementary) cube, it can be represented by $I_1 \times ... \times I_n$, where each $I_i$ is an elementary interval. Use the degenerate intervals to create a partial assignment $\mathbf{w}_D$. If $\mathbf{w}_D$ was not an implicant, then by Theorem 4.3, any $\boldsymbol{\mu} \in \mathring{C}_{\mathbf{w}_D}$ is not possible, which contradicts Theorem 4.10. Therefore, $X$ is an implicant cube.

Next, consider some implicant $\mathbf{w}_D$. By definition, there is a prime implicant $\mathbf{w}_E \subseteq \mathbf{w}_D$ that assigns to a subset of $\mathbf{w}_D$. Therefore, the only difference between the implicant cubes of $\mathbf{w}_D$ and $\mathbf{w}_E$ is that the latter has fewer nondegenerate intervals. Therefore, $C_{\mathbf{w}_D} \subseteq C_{\mathbf{w}_E}$, and so $C_{\mathbf{w}_D}$ is a face of $C_{\mathbf{w}_E}$. Therefore, $C_{\mathbf{w}_D}$ is a face of $C_\varphi$. □

**Proposition F.3.** *The facets of $C_\varphi$ are the prime implicant cubes $C_{\mathbf{w}_D}$.*

*Proof.* Consider some facet $X$ of $C_\varphi$. By Proposition F.2, the deterministic part of $X$ is an implicant $\mathbf{w}_D$. Assume $\mathbf{w}_D$ is not a prime implicant. Then there is a deterministic variable $i$ that we can remove from $\mathbf{w}_D$ and still have an implicant $\mathbf{w}_E$. But then $C_{\mathbf{w}_E} \supset C$, with $C =$ being a face of $C_{\mathbf{w}_E}$, which is in contradiction with the assumption that $C$ is a facet. □

**Proposition F.4.** *The vertices $\mathcal{C}_0(C_\varphi)$ of $C_\varphi$ is equal to the set of possible worlds $\mathcal{W}_\varphi$.*

*Proof.* Let $\mathcal{C}_0(C_\varphi) \subseteq \{0,1\}^n$ be the vertices of $C_\varphi$, which by Proposition F.2 is the implicant cubes with no stochastic variables, that is, it assigns a value to every variable and

corresponds directly to a world. By the fact that it is an implicant, this world has to be possible, that is, $\mathcal{C}_0(C_\varphi) = |\mathcal{W}_\varphi|$. □

**Proposition F.5.** *The vertices $\mathcal{C}_0(C_{\mathbf{w}_D})$ of (prime) implicant cube $C_{\mathbf{w}_D}$ is equal to the cover of $\mathbf{w}_D$.*

*Proof.* Considering $\mathbf{w}_D$ as the constraint that a world $\mathbf{w}$ has to agree on the deterministic variables with $\mathbf{w}_D$, by Proposition F.4, the vertices of $C_{\mathbf{w}_D}$ are precisely such worlds. This is the cover of $\mathbf{w}_D$. □

## G. Proofs of the main theorems

In this appendix, we give the proofs for the theorems in the main paper. Understanding some of these proofs requires understanding the connection of our problem to cubical sets, which we give in Appendix F. We recommend going through Appendix F before reading the proofs.

We start off by defining the transformation from independent parameters to distributions and prove that it is a bijection. First, let

$$f_{\perp\!\!\!\perp}(\boldsymbol{\mu})_i = \prod_{j=1}^{n} \mu_j^{w_{i,j}} \cdot (1-\mu_j)^{1-w_{i,j}}. \tag{7}$$

be a function $f_{\perp\!\!\!\perp} : [0,1]^n \to \Delta^{\perp\!\!\!\perp}$ that maps the parameters $\boldsymbol{\mu}$ to the set of independent distributions. Note that this is the transformation used in Equation 4.

**Lemma G.1.** *The map $f_{\perp\!\!\!\perp}$ is a continuous bijection from $[0,1]^n$ to $\Delta^{\perp\!\!\!\perp}$*[5].

*Proof.* Define the function $f_{\perp\!\!\!\perp}^{-1} : \Delta^{\perp\!\!\!\perp} \to [0,1]^n$ as

$$f_{\perp\!\!\!\perp}^{-1}(p)_i = p(w_i = 1) = \sum_{j=1}^{|\mathcal{W}|} w_{j,i} p_j \quad i \in 1, ..., n. \tag{8}$$

Consider $\boldsymbol{\mu} \in [0,1]^n$. Since $\boldsymbol{\mu}$ are the parameters of an independent distribution, the marginal probability $p_{\boldsymbol{\mu}}(w_i = 1) = \mu_i$. This is also by definition the sum of the probabilities of all worlds $\mathbf{w}_k$ with $w_{k,i} = 1$, that is, $f_{\perp\!\!\!\perp}^{-1}$. Therefore, $f_{\perp\!\!\!\perp}^{-1}(f_{\perp\!\!\!\perp}(\boldsymbol{\mu}))_i = \mu_i$.

Next, consider $p \in \Delta^{\perp\!\!\!\perp}$. By the definition of $\Delta^{\perp\!\!\!\perp}$ in Equation 4, there must be a parameter $\boldsymbol{\mu} \in [0,1]^n$ such that $f(\boldsymbol{\mu}) = p$, that is, $p$ represents an independent distribution by definition. Therefore, the marginal probabilities computed with $f_{\perp\!\!\!\perp}^{-1}$ precisely describe $p$, and so $f_{\perp\!\!\!\perp}(f_{\perp\!\!\!\perp}^{-1}(p)) = p$. □

Next, we repeat the theorems from the main body of the text and give their proofs.

[5] It is a bijection to $\Delta^{\perp\!\!\!\perp}$, but not to the codomain $\Delta$.

**Theorem 4.4** (Representability of $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$)**.** *Let $p_{\boldsymbol{\mu}}$ have $p_{\boldsymbol{\mu}}(\varphi = 1) > 0$ and deterministic assignment $\mathbf{w}_E$. Then the following statements are equivalent:*

1. *the conditional distribution $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$ can be represented by another independent distribution $q_{\boldsymbol{\mu}'}$;*
2. *there is an implicant $\mathbf{w}_D$ that covers all possible worlds in the support of $p_{\boldsymbol{\mu}}$;*
3. *there is an implicant $\mathbf{w}_D$ such that $\mathbf{w}_E, \varphi \models \mathbf{w}_D$.*

*Proof.* $2 \to 1$: Assume such an implicant $\mathbf{w}_D$ of $\varphi$ exists. Then we rewrite the conditional distribution $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$ as

$$p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1) = \frac{\prod_{i=1}^{n} \mu_i \varphi(\mathbf{w})}{p_{\boldsymbol{\mu}}(\varphi = 1)} \tag{9}$$

$$= I[\mathbf{w} \in \mathcal{W}_{\mathbf{w}_D}] \frac{\prod_{i \notin D} \mu_i}{p_{\boldsymbol{\mu}}(\varphi = 1)} \tag{10}$$

$$= I[\mathbf{w} \in \mathcal{W}_{\mathbf{w}_D}] \prod_{i \notin D} \mu_i p_{\boldsymbol{\mu}}(\varphi = 1)^{n-|D|}. \tag{11}$$

Here, $I[\mathbf{w} \in \mathcal{W}_{\mathbf{w}_D}]$ is the indicator function that is 1 when $\mathbf{w}$ is in the cover of $\mathbf{w}_D$. This is an independent distribution $q_{\boldsymbol{\mu}'}$ with deterministic assignment $\mathbf{w}_D$ and parameters $\mu_i' = \mu_i p_{\boldsymbol{\mu}}(\varphi = 1)^{n-|D|}$ for the stochastic variables. In the first step, we used that $\varphi(\mathbf{w}) = 0$ exactly when $\mathbf{w}_D$ differs from the implicant $\mathbf{w}_D$.

$1 \to 2$: Assume there is no implicant $\mathbf{w}_D$ as described in 2. Then, the deterministic assignment of the conditional distribution $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$ is not an implicant, as if it was, we could have constructed such an implicant. But then, by Theorem 4.3, there must be a world $\mathbf{w}$ in the cover of the deterministic assignment of $q$ that is not possible. Since for independent distributions, any worlds in the cover of the deterministic assignment get positive probability, such an independent distribution must also assign positive probability to this extension, yet $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1) = 0$, so $p_{\boldsymbol{\mu}}(\mathbf{w}|\varphi = 1)$ cannot be represented by an independent distribution.

$2 \to 3$: Assume all possible worlds in the support of $p_{\boldsymbol{\mu}}$ are in the cover of the implicant $\mathbf{w}_D$. That means all worlds in the cover of $\mathbf{w}_E$ for which the constraint $\varphi$ holds are also in the cover of $\mathbf{w}_D$. Therefore, $\mathbf{w}_E, \varphi \models \mathbf{w}_D$.

$3 \to 2$: Assume there is an implicant $\mathbf{w}_D$ such that $\mathbf{w}_E, \varphi \models \mathbf{w}_D$. By independence, the support of $p$ contains the worlds extending $\mathbf{w}_E$. By the entailment, its subset of *possible* worlds is those that also extend $\mathbf{w}_D$. □

**Theorem 4.10** (Representing the set of possible independent distributions)**.** *A parameter $\boldsymbol{\mu}$ is in $C_\varphi$ if and only if the distribution $p_{\boldsymbol{\mu}}$ is possible for $\varphi$. That is, $\boldsymbol{\mu} \in C_\varphi$ if and only if $p_{\boldsymbol{\mu}} \in \Delta_\varphi^{\perp\!\!\!\perp}$. Furthermore, the cubical set $C_\varphi$ cannot be represented as a union of fewer cubes.*

*Proof.* Consider some $p_{\boldsymbol{\mu}} \in \Delta_{\varphi}^{\perp\!\!\!\perp}$. Then, by Theorem 4.3, the deterministic part $\mathbf{w}_E$ of $\boldsymbol{\mu}$ is an implicant of $\varphi$. Let $\mathbf{w}_D \in \mathcal{I}$ be any prime implicant that $\mathbf{w}_E$ is an extension of, which has to exist by construction. Clearly $C_{\mathbf{w}_E} \subseteq C_{\mathbf{w}_D}$, and so $\boldsymbol{\mu} \in C_{\mathbf{w}_D} \subseteq C_{\varphi}$. Next, assume $\boldsymbol{\mu} \in C_{\varphi}$ for some $\mathbf{w}_D \in \mathcal{I}$. $\mathbf{w}_D$ is an implicant of $\varphi$, and so by Theorem 4.3, $p_{\boldsymbol{\mu}}$ is possible for $\varphi$, hence $p_{\boldsymbol{\mu}} \in \Delta_{\varphi}^{\perp\!\!\!\perp}$.

Next, we prove that there is no smaller set of cubes than the set of prime implicant cubes that generate $C_{\varphi}$. Associated with each implicant cube $C_{\mathbf{w}_D}$, is an *(elementary) implicant cell* $\overset{\circ}{C}_{\mathbf{w}_D} \subseteq C_{\mathbf{w}_D}$. $\overset{\circ}{C}_{\mathbf{w}_D}$ is similarly defined as $C_{\mathbf{w}_D}$, except it uses open intervals $(0, 1)$ instead of closed intervals $[0, 1]$. Let $\mathbf{w}_D$ be a prime implicant and let $\boldsymbol{\mu} \in \overset{\circ}{C}_{\mathbf{w}_D}$. All cubes are equal to the union of cells inside it (Kaczynski et al. (2004), Proposition 2.15(v)). Since the different cells in a cubical complex are disjoint (Kaczynski et al. (2004), Proposition 2.15(iii)), $\boldsymbol{\mu}$ is not in another implicant cube. □

**Theorem 4.11** (Convexity)**.** *The following statements are equivalent: 1) There is exactly one prime implicant of $\varphi$; 2) $C_{\varphi}$ is convex; 3) the semantic loss over the space of independent distributions $\mathcal{L}(\boldsymbol{\mu})$ is convex.*

*Proof.* $1 \leftrightarrow 2$ follows directly from Kaczynski et al. (2004), Proposition 2.80, by noting that the only rectangles in our setting are the elementary cubes in $[0, 1]$. Since there is no proof of Proposition 2.80 given in Kaczynski et al. (2004), we provide it for completeness' sake.

$2 \rightarrow 1$: Assume there is exactly one prime implicant $\mathbf{w}_D$ of $\varphi$. Then $C_{\varphi}$ is described by $C_{\mathbf{w}_D}$. This is an elementary cube, which is convex.

$1 \rightarrow 2$: Next, assume there is more than one prime implicant. Consider two distinct prime implicants $\mathbf{w}_D, \mathbf{w}_E \in \mathcal{I}$. Consider $\boldsymbol{\mu}_{\mathbf{w}_D} \in C_{\mathbf{w}_D} \setminus C_{\mathbf{w}_E}$ and $\boldsymbol{\mu}_{\mathbf{w}_E} \in C_{\mathbf{w}_E} \setminus C_{\mathbf{w}_D}$, which have to exist by Proposition 4.10. Consider $\boldsymbol{\mu}$ to be any non-trivial convex combination of $\boldsymbol{\mu}_{\mathbf{w}_D}$ and $\boldsymbol{\mu}_{\mathbf{w}_E}$. Note that the deterministic assignment of $\boldsymbol{\mu}$ is $\hat{D} = \{k \in D \cap E : \mathbf{w}_{Dk} = \mathbf{w}_{Ek}\}$. Since the prime implicants are different, at least one of the following needs to hold:

1. There is a $k \in D$ but $k \notin E$. Then $\mu_k$ is stochastic, and $\hat{D}$ is a strict subset of $D$.
2. There is a $k \in D \cap E$ such that $\mathbf{w}_{Dk} \neq \mathbf{w}_{Ek}$. By the convex combination, $\mu_k$ is stochastic, as it assigns probability mass to both $\mathbf{w}_{Dk}$ and $\mathbf{w}_{Ek}$. Therefore, $\hat{D}$ is a strict subset of $D$.

Since $D$ is a prime implicant, removing any element from $D$ results in a deterministic assignment that is no longer an implicant. Thus, by Theorem 4.3, $p_{\boldsymbol{\mu}}$ is not possible.

$2 \rightarrow 3$: Assume there is exactly one prime implicant $\mathbf{w}_D$. Then note that the only possible worlds are the cover $\mathcal{W}_{\mathbf{w}_D}$. The cover contains all assignments to the stochastic variables, i.e., the variables not in $D$. This means we can safely marginalize those out:

$$
\begin{aligned}
p_{\boldsymbol{\mu}}(\varphi = 1) &= \sum_{\mathbf{w} \in \mathcal{W}_{\mathbf{w}_D}} p_{\boldsymbol{\mu}}(\mathbf{w}) \\
&= \sum_{\mathbf{w} \in \mathcal{W}_{\mathbf{w}_D}} \prod_{i \in D} p_{\boldsymbol{\mu}}(\mathbf{w}_{Di}) \prod_{i \in \{1,\ldots,n\} \setminus D} p_{\boldsymbol{\mu}}(\mathbf{w}_i) \\
&= \prod_{i \in D} p_{\boldsymbol{\mu}}(\mathbf{w}_{Di}) \sum_{\mathbf{w} \in \mathcal{W}_{\mathbf{w}_D}} \prod_{i \in \{1,\ldots,n\} \setminus D} p_{\boldsymbol{\mu}}(\mathbf{w}_i) \\
&= \prod_{i \in D} p_{\boldsymbol{\mu}}(\mathbf{w}_{Di}) \sum_{\mathbf{w} \in \mathcal{W}_{\mathbf{w}_D}} p_{\boldsymbol{\mu}}(\mathbf{w}_{\{1,\ldots,n\} \setminus D}) \\
&= \prod_{i \in D} p_{\boldsymbol{\mu}}(\mathbf{w}_{Di}) = p_{\boldsymbol{\mu}}(\mathbf{w}_D).
\end{aligned}
$$

Given $\lambda \in (0, 1)$, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in [0, 1]^n$, we define $\boldsymbol{\mu}_{\lambda} = \lambda \boldsymbol{\mu}_1 + (1 - \lambda)\boldsymbol{\mu}_2$ for brevity. Rewriting, and using Jensen's inequality and some slightly laborious algebra, we find that

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\mu}_{\lambda}) = -\log p_{\boldsymbol{\mu}_{\lambda}}(\varphi = 1) = -\log p_{\boldsymbol{\mu}_{\lambda}}(\mathbf{w}_D) \\
&= -\log \prod_{i \in D} p_{\mu_{\lambda}}(\mathbf{w}_{Di}) \\
&= -\sum_{i \in D} \log({\mu_{\lambda}}_i^{\mathbf{w}_{Di}}(1 - \mu_{\lambda_i})^{1 - \mathbf{w}_{Di}}) \\
&= -\sum_{i \in D} \mathbf{w}_{Di} \log \mu_{\lambda_i} + (1 - \mathbf{w}_{Di}) \log(1 - \mu_{\lambda_i}) \\
&= -\sum_{i \in D} \mathbf{w}_{Di} \log(\lambda \mu_{1_i} + (1 - \lambda)\mu_{2_i}) \\
&\quad + (1 - \mathbf{w}_{Di}) \log(1 - (\lambda \mu_{1_i} + (1 - \lambda)\mu_{2_i})) \\
&= -\sum_{i \in D} \mathbf{w}_{Di} \log(\lambda \mu_{1_i} + (1 - \lambda)\mu_{2_i}) \\
&\quad + (1 - \mathbf{w}_{Di}) \log(\lambda(1 - \mu_{1_i}) + (1 - \lambda)(1 - \mu_{2_i})) \\
&\leq -\sum_{i \in D} \mathbf{w}_{Di}(\lambda \log \mu_{1_i} + (1 - \lambda) \log \mu_{2_i}) \\
&\quad + (1 - \mathbf{w}_{Di})(\lambda \log(1 - \mu_{1_i}) + (1 - \lambda) \log(1 - \mu_{2_i})) \\
&= -\sum_{i \in D} \lambda \log {\mu_1}_i^{\mathbf{w}_{Di}}(1 - \mu_{1_i})^{1 - \mathbf{w}_{Di}} \\
&\quad + (1 - \lambda) \log {\mu_2}_i^{\mathbf{w}_{Di}}(1 - \mu_{2_i})^{1 - \mathbf{w}_{Di}} \\
&= -\lambda \log p_{\boldsymbol{\mu}_1}(\varphi = 1) - (1 - \lambda) \log p_{\boldsymbol{\mu}_2}(\varphi) \\
&= \lambda \mathcal{L}(\boldsymbol{\mu}_1) + (1 - \lambda)\mathcal{L}(\boldsymbol{\mu}_2).
\end{aligned}
$$

$3 \rightarrow 2$. Assume there is more than one prime implicant. Using $1 \rightarrow 2$, this means $C_{\varphi}$ is non-convex. Therefore, there is a pair $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in C_{\varphi}$ and $\lambda \in (0, 1)$ such that $\boldsymbol{\mu}_{\lambda} = \lambda \boldsymbol{\mu}_1 + (1 - \lambda)\boldsymbol{\mu}_2 \notin C_{\varphi}$. By Theorem 4.10, $(\boldsymbol{\mu}_1) = \mathcal{L}(\boldsymbol{\mu}_2) = 0 < \mathcal{L}(\boldsymbol{\mu}_{\lambda})$, as $C_{\varphi}$ exactly describes the possible distributions where $\mathcal{L}(\boldsymbol{\mu}) = 0$. Therefore, $\mathcal{L}(\boldsymbol{\mu}_{\lambda}) > \lambda \mathcal{L}(\boldsymbol{\mu}_1) + (1 - \lambda)\mathcal{L}(\boldsymbol{\mu}_2)$, proving non-convexity. □

**Theorem 4.13** (Connectedness)**.** *The connected components of the space of possible distributions $C_\varphi$ correspond to the connected components in $\mathcal{G}$. In particular, $C_\varphi$ is a connected space if and only if $\mathcal{G}$ is connected.*

*Proof.* By Proposition F.4, the vertices of $C_\varphi$ are precisely the possible worlds $\mathcal{W}_\varphi$. We say $\mathbf{w}_1, ..., \mathbf{w}_n \in \mathcal{W}_\varphi$ are *edge-connected* if there exist edges $E_1, ..., E_n \in \mathcal{C}_1(C_\varphi)$ such that $\mathbf{w}_i$ and $\mathbf{w}_{i+1}$ are faces of $E_i$.

Consider an edge $(\mathbf{w}_1, \mathbf{w}_2) \in \mathcal{E}$. Then there is a prime implicant $\mathbf{w}_D$ that covers both $\mathbf{w}_1$ and $\mathbf{w}_2$. Therefore, $\mathbf{w}_1$ and $\mathbf{w}_2$ are both in $C_{\mathbf{w}_D}$. Since $C_{\mathbf{w}_D}$ is an (elementary) cube, by Kaczynski et al. (2004), Proposition 2.51.1, $\mathbf{w}_1$ and $\mathbf{w}_2$ are edge-connected.

Let $E \in \mathcal{C}_1(C_\varphi)$ be an edge with vertices $\mathbf{w}_1$ and $\mathbf{w}_2$. $E$ is a face of a prime implicant cube $C_{\mathbf{w}_D}$ by Definition 4.6, and by Proposition F.5, $\mathbf{w}_1$ and $\mathbf{w}_2$ are both in the cover $\mathcal{W}_{\mathbf{w}_D}$. Therefore, $(\mathbf{w}_1, \mathbf{w}_2) \in \mathcal{E}$.

Combining these two results, we find that $\mathbf{w}_1$ and $\mathbf{w}_2$ are edge-connected if and only if there is an edge between them in $\mathcal{G}$. Therefore, by Theorem 2.55 and Corollary 2.57 of Kaczynski et al. (2004), the connected components of $\mathcal{G}$ and $C_\varphi$ coincide. □