Twelfth Annual Advances in Cognitive Systems Poster Collection (2025) 193-212

A Data-Transparent Probabilistic Model of Temporal Propositional Abstraction

Hiroyuki Kido KIDOH@CARDIFF.AC.UK
Cardiff University, Park Place, CF10 3AT, Cardiff, UK

Abstract

Standard probabilistic models face fundamental challenges such as data scarcity, a large hypothesis space, and poor data transparency. To address these challenges, we propose a novel probabilistic model of data-driven temporal propositional reasoning. Unlike conventional probabilistic models where data is a product of domain knowledge encoded in the probabilistic model, we explore the reverse direction where domain knowledge is a product of data encoded in the probabilistic model. This more data-driven perspective suggests no distinction between maximum likelihood parameter learning and temporal propositional reasoning. We show that our probabilistic model is equivalent to a highest-order, i.e., full-memory, Markov chain, and it can also be viewed as a hidden Markov model requiring no distinction between hidden and observable variables. We discuss that limits provide a natural and mathematically rigorous way to handle data scarcity, including the zero-frequency problem. We also discuss that a probability distribution over data generated by our probabilistic model helps data transparency by revealing influential data used in predictions. The reproducibility of this theoretical work is fully demonstrated by the included proofs.

1. Introduction

Probability theory underlies modern AI (Russell & Norvig, 2020). Probabilistic modelling has led to various successful AI applications, such as computer vision, speech recognition, and natural language processing (Pearl, 1988; Bishop, 2006). However, it inherently involves fundamental challenges such as data scarcity, an exponentially growing hypothesis space, and poor data transparency. To illustrate these challenges, let us consider the following simple, discrete-time, discrete-state localisation problem.

Example 1. *The left-hand side of Figure 1 shows a building with ten rooms. The room number is shown in the northwest corner of each room. The two arrows indicate the tracks of a robot, and d_k denotes the data collected by the robot in the room, for all $k \in \{1, 2, ..., 12\}$. Using the twelve data, we want to find the location of the robot exploring the building. Suppose that the robot moved through Rooms 2, 3, and 8. Where is the robot likely to be two time steps after Room 8?*

The standard approaches to this problem are probabilistic modelling (Bishop, 2006; Russell & Norvig, 2020) such as Markov chains and hidden Markov models (Rabiner, 1989; Mor et al., 2021). However, they are not fundamentally free from the following issues.

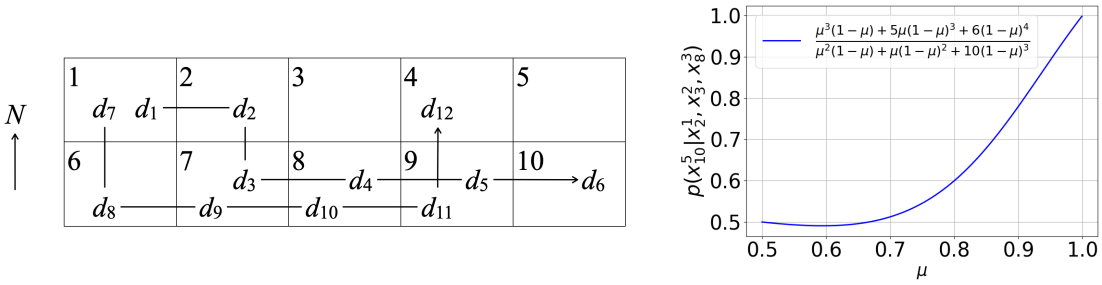

Figure 1. Left: Twelve data collected by a robot exploring a building with ten rooms. Right (used in Section 2.5): $p(x_{10}^5|x_2^1, x_3^2, x_8^3)$ as a function of μ. The case of $\mu = 1$, which corresponds to the semantics of propositional logic, results in an undefined probability. This singularity can be resolved by taking the limit as $\mu \to 1$.

The first issue is data scarcity. Since the robot does not experience Room 3, a probabilistic model naively trained with the twelve data cannot predict the robot location due to zero frequency. While data smoothing (Murphy & Bach, 2012; Russell & Norvig, 2020) mitigates the problem, it is effective only when the number of parameters in the probabilistic model is sufficiently small.

The second issue is an exponentially growing hypothesis space. The three data beginning with d_2 best match the known past robot locations. Indeed, assuming d_2 corresponds to time step 1, denoted by Time 1, the series of data correctly explains the robot locations at Time 1 and 3. The robot is then predicted to be in Room 10 at Time 5. While this idea seems promising, it cannot be easily generalised using probabilistic models. The number of parameters in an nth-order Markov chain with r states is $(r-1) \times r^n$. Thus, even this simple problem requires $90,000$ (i.e., 9×10^4) parameters in a 4th-order Markov chain, which takes into account all the past four time steps to predict the next one.

The third issue is data transparency. The three data beginning with d_8 next best match the known past robot locations. Indeed, assuming d_8 corresponds to Time 1, the series of data correctly explains the robot location at Time 3. The robot is then predicted to be in Room 4, rather than Room 10, at Time 5. Now, the probability of Room 4 should be lower than that of Room 10, considering the consistency with the known robot locations. However, what if the series of data beginning with d_8 occurs repeatedly? At some point, consistency in quantity may surpass consistency in quality. To the best of our knowledge, however, standard probabilistic models cannot justify this result with reference to actual data such as d_2 and d_8. This is because learning is typically the process of exploiting data to adjust the parameters of probabilistic models, whereas reasoning is the process of using the parameters, not the data itself, to make predictions.

In this paper, we propose a novel data-transparent probabilistic model as a simple yet unconventional approach to addressing the aforementioned issues. The key components of the probabilistic model are data, models (i.e., valuations) in propositional logic and propositional formulas X being true, for each time step t, denoted by d^t, m^t and x^t, respectively. We will argue that the probability

of x^t, denoted by $p(x^t)$, should be given as follows.

$$p(x^t) = \sum_{m^t} \sum_{d^1} \sum_{d^2} \cdots \sum_{d^t} p(x^t, m^t, d^1, d^2, ..., d^t) \text{ where}$$
$$p(x^t, m^t, d^1, d^2, ..., d^t) = p(x^t|m^t)p(m^t|d^t)p(d^t|d^{t-1})p(d^{t-1}|d^{t-2}) \cdots p(d^1) \quad (1)$$

Here, the first line is an application of a valid rule of probability theory. The second line is an application of the probabilistic model we formulate in this paper. We will define $p(x^t|m^t)$ based on whether the propositional formula X is true in the model m^t at Time t, $p(m^t|d^t)$ based on whether the data d^t supports the model m^t at Time t, and $p(d^t|d^{t-1})$ based on whether the data d^{t-1} changes to d^t at the next time step. In a nutshell, Equation (1) states that the probability of a formula being true depends on whether time-dependent data support a model in which the formula is true. We significantly simplify Equation (1) under the natural assumption that both the data trajectory and the support relation from data to models are deterministic (see Figure 2 for an intuitive understanding).

The contributions of this paper are summarised as follows. First, this study is inspired by the inference of abstraction (Kido, 2025a,b), which suggests logical, statistical, and probabilistic justifications for symbolic reasoning grounded in data. Our probabilistic model additionally incorporates a transition relation between data while maintaining the theoretical justifications and having essentially linear time complexity with respect to the number of data (see Section 2).

Second, we show that our probabilistic model can be viewed as a highest-order, i.e., most expressive, Markov chain, in which all the past states are used to predict the current state (see Sections 3.1 and 3.4). One advantage of our model over Markov chains is data transparency. In our model, propositional reasoning is fully grounded in data as it always occurs between data and formulas, not between formulas and other formulas (see Section 3.2). Our model can also be viewed as a hidden Markov model that requires no distinction between observable and hidden states (see Section 3.3).

Third, we challenge the conventional view prevailing across AI, cognitive science, and neuroscience that data are assumed to be generated from domain knowledge encoded in probabilistic models, e.g., (Lee & Mumford, 2003; Itti & Baldi, 2009; Hohwy et al., 2008; Smith et al., 2022; Tenenbaum et al., 2006; Lake et al., 2015; Dasgupta et al., 2020). Instead, we explore the reverse direction and investigate how domain knowledge can be generated from data, moving toward fully data-driven temporal probabilistic reasoning (see Section 2).

2. Temporal propositional abstraction

2.1 Random variables

Let $Data = \{d_1, d_2, ..., d_K\}$ be a non-empty set of K data. This set is a multiset, where elements may occur multiple times. For any discrete time $t \in \{1, 2, ..., T\}$, we assume that D^t is a random variable taking values in $Data$. This allows us to handle data that changes over time.

Let $Variables$ be the set of propositional variables, $Values = \{1, 0\}$ be the set of truth values meaning true and false, respectively, and $Models = \{m_1, m_2, ..., m_L\}$ be the set of L models, i.e., valuations, in propositional logic. As usual, each model is a function, $Variables \to Values$, that maps each propositional variable to a truth value. For any discrete time t, we assume that M^t is a random variable taking values in $Models$.

Let $\mathcal{L}$ be a propositional language. As usual, formulas are constructed from propositional variables using the usual logical connectives such as $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftarrow$, and $\leftrightarrow$. For any discrete time t and propositional formula $X_i \in \mathcal{L}$, X_i^t is a random variable taking values in $Values$. This allows us to handle the truth values of formulas that vary over time.

In the following sections, we will define the probability distributions over D^t, M^t and X_i^t, denoted by $p(D^t)$, $p(M^t)$ and $p(X_i^t)$.

Example 2 (Continued from Example 1)**.** *The problem illustrated in Figure 1 results in $Data = \{d_1, d_2, ..., d_{12}\}$. Let X_i be a propositional variable representing that the robot is in Room i, for all $i \in \{1, 2, ..., 10\}$. $Models$ then has 2^{10} elements, and each model assigns truth values to the ten propositional variables differently. $X_1 \rightarrow \neg {X_2}^3$ is a formula representing that 'at Time 3, if the robot is in Room 1 then it is not in Room 2.' $X_1^3 \rightarrow \neg X_2$ is not a formula as logical connectives can only connect formulas, not time-indexed random variables.*

We introduce some abbreviations for readability. $D^{t_1:t_2}$ denotes the sequence $(D^{t_1}, D^{t_1+1}, ..., D^{t_2})$. The lowercase letter d^t denotes a realisation of the random variable D^t. We often write $D^t = d^t$ as d^t when it is clear from the context. $d^{t_1:t_2}$ denotes the realisation sequence $(d^{t_1}, d^{t_1+1}, ..., d^{t_2})$. The same argument is applied to the other random variables M^t and X_i^t and their realisations m^t and x_i^t. In addition, $X_{i_1:i_2}^{t_1:t_2}$ denotes the sequence $(X_{i_1}^{t_1:t_2}, X_{i_1+1}^{t_1:t_2}, ..., X_{i_2}^{t_1:t_2})$, and $x_{i_1:i_2}^{t_1:t_2}$ is the sequence of their realisations. If $t_1 > t_2$ or $i_1 > i_2$ then the sequence is regarded as being empty, and omitted. For example, $p(X_1^1 | D^{1:1}, M^{1:1}, X_{1:I}^{1:0}, X_{1:0}^1) = p(X_1^1 | D^1, M^1)$.

Now, the full joint distribution over all the introduced random variables can be written as follows using the product rule (Bishop, 2006) of probability theory.

$$p(D^{1:T}, M^{1:T}, X_{1:I}^{1:T}) = \prod_{t=1}^{T} \Bigg[p(D^t | D^{1:t-1}, M^{1:t-1}, X_{1:I}^{1:t-1}) \\ p(M^t | D^{1:t}, M^{1:t-1}, X_{1:I}^{1:t-1}) \prod_{i=1}^{I} p(X_i^t | D^{1:t}, M^{1:t}, X_{1:I}^{1:t-1}, X_{1:i-1}^t) \Bigg] \tag{2}$$

In many cases, we are interested in the marginal distribution over formulas. It can be derived from the full joint distribution using the sum rule (Bishop, 2006) of probability theory.

$$\begin{aligned} p(X_{1:I}^{1:T}) &= \sum_{d^{1:T} \in Data^T} \sum_{m^{1:T} \in Models^T} p(d^{1:T}, m^{1:T}, X_{1:I}^{1:T}) \\ &= \sum_{d^{1:T} \in Data^T} \sum_{m^{1:T} \in Models^T} \prod_{t=1}^{T} \Bigg[p(d^t | d^{1:t-1}, m^{1:t-1}, X_{1:I}^{1:t-1}) \\ &\qquad p(m^t | d^{1:t}, m^{1:t-1}, X_{1:I}^{1:t-1}) \prod_{i=1}^{I} p(X_i^t | d^{1:t}, m^{1:t}, X_{1:I}^{1:t-1}, X_{1:i-1}^t) \Bigg] \end{aligned} \tag{3}$$

In Figure 2, the leftmost graph represents Equation (3) with $T = 3$ and $I = 2$. There is an arrow from each element of the condition to the outcome, for each conditional probability appearing in Equation (3). Since the graph is a complete directed graph, Equation (3) states that each random variable can influence each other.

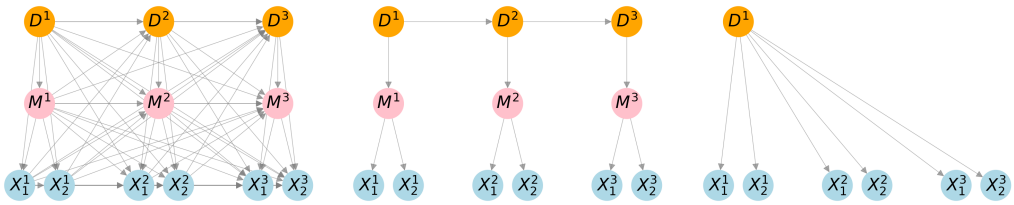

Figure 2. We show that the three graphical models are equivalent for temporal propositional reasoning.

Example 3 (Continued from Example 2)**.** *What is the probability that the robot is in Room 10 at Time 5 given that it is in Room 2 at Time 1, i.e.,* $p(X_{10}^5 = 1 | X_2^1 = 1)$*? We simply write it as* $p(x_{10}^5 | x_2^1)$*. Using Equation (3), we have*

$$p(x_{10}^5 | x_2^1) = \frac{p(x_2^1, x_{10}^5)}{p(x_2^1)} = \frac{\sum_{d^{1:5} \in Data^5} \sum_{m^{1:5} \in Models^5} \sum_{x_{1:10}^{1:5} \setminus \{x_2^1, x_{10}^5\} \in Values^{48}} Z}{\sum_{d^{1:5} \in Data^5} \sum_{m^{1:5} \in Models^5} \sum_{x_{1:10}^{1:5} \setminus \{x_2^1\} \in Values^{49}} Z}$$

where Z *is given as follows.*

$$\begin{aligned} Z = p(d^{1:5}, m^{1:5}, x_{1:10}^{1:5}) \quad = \quad & \prod_{t=1}^{5} \Big[p(d^t | d^{1:t-1}, m^{1:t-1}, x_{1:10}^{1:t-1}) \\ & p(m^t | d^{1:t}, m^{1:t-1}, x_{1:10}^{1:t-1}) \prod_{i=1}^{10} p(x_i^t | d^{1:t}, m^{1:t}, x_{1:10}^{1:t-1}, x_{1:i-1}^t) \Big] \end{aligned}$$

In the next section, we discuss how to simplify the result (see Figure 2).

2.2 Data distributions

We have not yet defined any conditional probabilities appearing in Equation (2) or (3). In this section, we define and use the conditional probability of data to simply those equations. To express how data changes over time, we assume a function, $n : Data \to Data$, that maps data at the current time step to data at the next. $|Data|$ denotes the cardinality of $Data$.

Definition 1. *Let* $t \in \{1, 2, ..., T\}$*. The conditional probability of* d^t *given* $d^{1:t-1}$*,* $m^{1:t-1}$ *and* $x_{1:I}^{1:t-1}$ *is defined as follows.*

$$p(d^t | d^{1:t-1}, m^{1:t-1}, x_{1:I}^{1:t-1}) = \begin{cases} \frac{1}{|Data|} & \text{if } t = 1 \\ 1 & \text{if } t \neq 1 \text{ and } d^t = n(d^{t-1}) \\ 0 & \text{otherwise} \end{cases}$$

We derive the following property from Definition 1.

Proposition 1. *Let* $t \in \{1, 2, ..., T\}$*.* D^t *is conditionally independent of* $D^{1:t-2}$*,* $M^{1:t-1}$ *and* $X_{1:I}^{1:t-1}$ *given* D^{t-1}*, i.e.* $p(D^t | D^{1:t-1}, M^{1:t-1}, X_{1:I}^{1:t-1}) = p(D^t | D^{t-1})$*.*

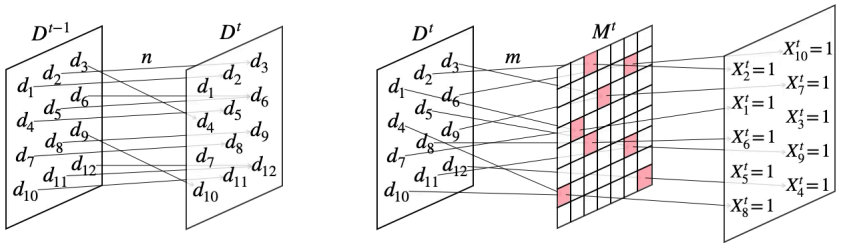

Figure 3. Left: The data transition. Right: The support relations among data, models in propositional logic, and propositional formulas over time. There is an arrow from a data point to a model if the data point evidences the model, which illustrates the function m. There is an arrow from a model to a formula if the formula is true in the model.

Proof. See Appendix. □

Example 4 (Continued from Example 3)**.** *It is clear from Figure 1 that the probability distribution over data can be directly observed from the environment. For* $k \in \{1, 2, ..., 12\}$ *and* $t \in \{2, 3, ..., 12\}$*,*

$$p(D^1 = d_k) = \frac{1}{12}$$

$$p(D^t = d_j | D^{t-1} = d_i) = \begin{cases} 1 & \text{if } (i,j) \in \{(1,2),(2,3),(3,4),(4,5),(5,6),(6,6), \\ & (7,1),(7,8),(8,9),(9,10),(10,11),(11,12),(12,12)\} \\ 0 & \text{otherwise.} \end{cases}$$

The left-hand side of Figure 3 illustrates this result. The arrows are illustrations of the function n *that maps each realisation of* D^{t-1} *to the corresponding realisation of* D^t*.*

From Proposition 1, Equation (2) can be simplified as follows.

$$\begin{aligned} p(D^{1:T}, M^{1:T}, X^{1:T}_{1:I}) \;\; = \;\; & \prod_{t=1}^{T} \Big[p(D^t | D^{t-1}) p(M^t | D^{1:t}, M^{1:t-1}, X^{1:t-1}_{1:I}) \\ & \prod_{i=1}^{I} p(X^t_i | D^{1:t}, M^{1:t}, X^{1:t-1}_{1:I}, X^t_{1:i-1}) \Big] \end{aligned} \tag{4}$$

2.3 Model distributions

In this section, we define the conditional probability of models, which appears in Equation (4), and then analyse its property to further simply the equation. Each model in propositional logic is meant to represent a state of the world. It is thus natural to think that each model is supported or evidenced by data observed from the environment. We assume a function, $m : Data \to Models$, that maps each data point to the corresponding model supported by the data.

Definition 2. *For $t \in \{1, 2, ..., T\}$, the conditional probability of m^t given $d^{1:t}$, $m^{1:t-1}$ and $x_{1:I}^{1:t-1}$ is defined as follows.*

$$p(m^t|d^{1:t}, m^{1:t-1}, x_{1:I}^{1:t-1}) = \begin{cases} 1 & \text{if } m^t = m(d^t) \\ 0 & \text{otherwise} \end{cases}$$

We derive the following property from Definition 2

Proposition 2. *Let $t \in \{1, 2, ..., T\}$. M^t is conditionally independent of $D^{1:t-1}$, $M^{1:t-1}$ and $X_{1:I}^{1:t-1}$ given D^t, i.e., $p(M^t|D^{1:t}, M^{1:t-1}, X_{1:I}^{1:t-1}) = p(M^t|D^t)$.*

Proof. See Appendix. □

From Proposition 2, Equation (4) can be simplified as follows.

$$\begin{aligned} p(D^{1:T}, M^{1:T}, X_{1:I}^{1:T}) &= \prod_{t=1}^{T} \Big[p(D^t|D^{t-1}) p(M^t|D^t) \\ &\quad \prod_{i=1}^{I} p(X_i^t|D^{1:t}, M^{1:t}, X_{1:I}^{1:t-1}, X_{1:i-1}^t) \Big] \end{aligned} \tag{5}$$

2.4 Knowledge distributions

In this section, we define the conditional probability of formulas, which appears in Equation (5), and then analyse its property to further simply the equation. As usual, the truth value of a formula is determined solely in light of a model based on the semantics of propositional logic. We use the symbol $[\![X_i]\!]_{m_l}$ to denote the truth value of the formula $X_i \in \mathcal{L}$ in the model $m_l \in Models$.

Definition 3. *Let $\mu \in [0.5, 1]$ and $t \in \{1, 2, ..., T\}$. The conditional probability of x_i^t given $d^{1:t}$, $m^{1:t}$, $x_{1:I}^{1:t-1}$ and $x_{1:i-1}^t$ is defined as follows.*

$$p(x_i^t|d^{1:t}, m^{1:t}, x_{1:I}^{1:t-1}, x_{1:i-1}^t) = \begin{cases} \mu & \text{if } x_i^t = [\![X_i]\!]_{m^t} \\ 1 - \mu & \text{otherwise} \end{cases}$$

Namely, the truth value of a formula at a time step depends only on the model at the same time step. We derive the following property from Definition 3.

Proposition 3. *Let $t \in \{1, 2, ..., T\}$ and $i \in \{1, 2, ..., I\}$. X_i^t is conditionally independent of $D^{1:t}$, $M^{1:t-1}$, $X_{1:I}^{1:t-1}$ and $X_{1:i-1}^t$ given M^t, i.e., $p(X_i^t|D^{1:t}, M^{1:t}, X_{1:I}^{1:t-1}, X_{1:i-1}^t) = p(X_i^t|M^t)$.*

Proof. See Appendix. □

Example 5 (Continued from Example 4)**.** *The data shown in Figure 1 give rise to the following results regardless of the value of t.*

$$p(X_i^t = 1|M^t = m(d_k)) = \begin{cases} \mu & \text{if } (k, i) \in \{(1, 1), (2, 2), (3, 7), (4, 8), (5, 9), \\ & \quad (6, 10), (7, 1), (8, 6), (9, 7), (10, 8), (11, 9), (12, 4)\} \\ 1 - \mu & \text{otherwise} \end{cases}$$

Here, recall that $m(d_k)$ is the model supported by data d_k. The right-hand side of Figure 3 illustrates this result. Note that the hierarchy represents an abstraction relation in the sense that an element on each layer is selective ignorance of elements of its left layer. In fact, the truth value of each formula is determined once a model is given, but not vice versa. Each model is determined once a data point is given, but not vice versa.

From Proposition 3, Equation (5) can be simplified as follows.

Theorem 1. *The full joint distribution over $D^{1:T}$, $M^{1:T}$ and $X_{1:I}^{1:T}$ is given as follows.*

$$p(D^{1:T}, M^{1:T}, X_{1:I}^{1:T}) = \prod_{t=1}^{T}\left[p(D^t|D^{t-1})p(M^t|D^t)\prod_{i=1}^{I}p(X_i^t|M^t)\right] \tag{6}$$

Proof. Applications of Propositions 1, 2 and 3. □

Equation (6) is the simplest form of the full joint distribution. The centre graph of Figure 2 illustrates the equation, where there are arrows from each of the conditions to the outcome, for all the conditional probabilities appearing in the equation.

In many cases, we are interested in reasoning over formulas. Obviously, the marginal distribution over formulas, i.e., Equation (3), can be written as follows using Equation (6).

$$p(X_{1:I}^{1:T}) = \sum_{d^{1:T}\in Data^T}\sum_{m^{1:T}\in Models^T}\prod_{t=1}^{T}\left[p(d^t|d^{t-1})p(m^t|d^t)\prod_{i=1}^{I}p(X_i^t|m^t)\right] \tag{7}$$

Interestingly, Equation (7) can be further simplified. Let $n^t(d_k)$ denote the data obtained by applying the function n to the data d_k t times.

Theorem 2. *The marginal distribution over $X_{1:I}^{1:T}$ is given as follows.*

$$p(X_{1:I}^{1:T}) = \frac{1}{K}\sum_{k=1}^{K}\prod_{t=1}^{T}\prod_{i=1}^{I}p(X_i^t|m(n^{t-1}(d_k))) \tag{8}$$

Proof. See Appendix. □

Theorem 2 shows the simplest form of the marginal distribution over formulas. The right-hand side of Figure 2 illustrates the result. There are arrows from the condition to the outcome, for all the conditional probabilities appearing in Equation (8). Theorem 2 is computationally important as the omitted summation multiplication $\sum_{d^2}\sum_{d^3}\cdots\sum_{d^T}\sum_{m^1}\sum_{m^2}\cdots\sum_{m^T}$ does not change the result but is computationally intractable. For example, since $|Data| = 12$ and $|Models| = 2^{10}$ in Figure 1, Theorem 2 allows us to skip $(12\times 2^{10})^T$ steps.

Let α and Δ be an element and a subset of $\{x_i^t|x_i^t \in x_{1:I}^{1:T}\}$, respectively. Using the sum rule and Theorem 2, the conditional probability of α given Δ can be written as follows.

$$\begin{aligned} p(\alpha|\Delta) &= \frac{p(\alpha,\Delta)}{p(\Delta)} = \frac{\sum_{x_{1:I}^{1:T}(\notin\Delta\cup\{\alpha\})}p(x_{1:I}^{1:T})}{\sum_{x_{1:I}^{1:T}(\notin\Delta)}p(x_{1:I}^{1:T})} \\ &= \frac{\sum_{k=1}^{K}\sum_{x_{1:I}^{1:T}(\notin\Delta\cup\{\alpha\})}\prod_{t=1}^{T}\prod_{i=1}^{I}p(x_i^t|m(n^{t-1}(d_k)))}{\sum_{k=1}^{K}\sum_{x_{1:I}^{1:T}(\notin\Delta)}\prod_{t=1}^{T}\prod_{i=1}^{I}p(x_i^t|m(n^{t-1}(d_k)))} \end{aligned}$$

For all $x_i^t \notin \Delta \cup \{\alpha\}$, $\sum_{x_i^t} p(x_i^t|m(n^{t-1}(d_k))) = \mu + (1-\mu) = 1$. Therefore,

$$= \frac{\sum_{k=1}^{K} \prod_{x_i^t \in \Delta \cup \{\alpha\}} p(x_i^t|m(n^{t-1}(d_k)))}{\sum_{k=1}^{K} \prod_{x_i^t \in \Delta} p(x_i^t|m(n^{t-1}(d_k)))}. \tag{9}$$

The following property regarding the negation connective is useful.

Proposition 4. *Let $X_i \in \mathcal{L}$ and $t \in \{1, 2, ..., T\}$. $p(X_i^t = 0) = p(\neg X_i^t = 1)$.*

Proof. See Appendix. □

In what follows, we write $X_i^t = 0$ as $\neg X_i^t = 1$, and then abbreviate this as $\neg x_i^t$.

2.5 Examples

This section discusses examples of the probabilistic model we defined and then simplified in the previous section. To explain the role of μ introduced in Definition 3, we consider the three situations: μ substituted by 1, μ approaching 1, and μ strictly less than 1, i.e., $\mu = 1$, $\mu \to 1$, and $\mu \in [0.5, 1)$, respectively.

Example 6 (Continued from Example 5)**.** *Let $\mu = 1$. What is the probability that the robot is in Room 10 at Time 5 given that it is in Room 2 at Time 1? Using Equation (9) and $Data = \{d_1, d_2, ..., d_{12}\}$, we have*

$$\begin{aligned} p(x_{10}^5|x_2^1) &= \frac{\sum_{k=1}^{12} p(x_2^1|m(d_k))p(x_{10}^5|m(n^4(d_k)))}{\sum_{k=1}^{12} p(x_2^1|m(d_k))} \\ &= \frac{\sum_{k\in\{2\}} \mu^2 + \sum_{k\in\{3\text{-}6\}} \mu(1-\mu) + \sum_{k\in\{1,7\text{-}12\}} (1-\mu)^2}{\sum_{k\in\{2\}} \mu + \sum_{k\in\{1,3\text{-}12\}} (1-\mu)} \\ &= \frac{\mu^2 + 4(1-\mu)^2 + 7(1-\mu)^2}{\mu + 11(1-\mu)} = \frac{1}{1}. \end{aligned}$$

This result is natural because there exists data showing that the robot was in Room 10 four time steps after being in Room 2.

Example 7 (Continued from Example 5)**.** *Suppose that the robot was in Rooms 2, 3 and 8 at Time 1, 2 and 3, respectively. What is the probability that the robot is in Room 10 at Time 5, i.e., $p(x_{10}^5|x_2^1, x_3^2, x_8^3)$. Let $\mu = 1$. Using Equation (9), we have*

$$\begin{aligned} & p(x_{10}^5|x_2^1, x_3^2, x_8^3) \\ = & \frac{\sum_{k=1}^{12} p(x_2^1|m(d_k))p(x_3^2|m(n(d_k)))p(x_8^3|m(n^2(d_k)))p(x_{10}^5|m(n^4(d_k)))}{\sum_{k=1}^{12} p(x_2^1|m(d_k))p(x_3^2|m(n(d_k)))p(x_8^3|m(n^2(d_k)))} \\ = & \frac{\sum_{k\in\{2\}} \mu^3(1-\mu) + \sum_{k\in\{3\text{-}6,8\}} \mu(1-\mu)^3 + \sum_{k\in\{1,7,9\text{-}12\}} (1-\mu)^4}{\sum_{k\in\{2\}} \mu^2(1-\mu) + \sum_{k\in\{8\}} \mu(1-\mu)^2 + \sum_{k\in\{1,3\text{-}7,9\text{-}12\}} (1-\mu)^3} \\ = & \frac{\mu^3(1-\mu) + 5\mu(1-\mu)^3 + 6(1-\mu)^4}{\mu^2(1-\mu) + \mu(1-\mu)^2 + 10(1-\mu)^3} = \frac{0}{0}. \end{aligned} \tag{10}$$

In contrast to Example 6, the probability is not defined due to division by zero. This is because the twelve data indicate that the robot has never been in Rooms 2, 3 and 8 in this order. Now, let μ approaching 1, i.e., $\mu \to 1$. We then have

$$\begin{aligned} p(x_{10}^5|x_2^1, x_3^2, x_8^3) &= \lim_{\mu\to 1} \frac{\mu^3(1-\mu) + 5\mu(1-\mu)^3 + 6(1-\mu)^4}{\mu^2(1-\mu) + \mu(1-\mu)^2 + 10(1-\mu)^3} \\ &= \lim_{\mu\to 1} \frac{\mu^3 + 5\mu(1-\mu)^2 + 6(1-\mu)^3}{\mu^2 + \mu(1-\mu) + 10(1-\mu)^2} = \frac{1}{1}. \end{aligned} \tag{11}$$

In Equation (10), the summation in the denominator runs over all sequences of three consecutive data points (i.e., a sliding window of size 3), whereas the summation in the numerator runs over all sequences of five consecutive data points. In Equation (11), we can cancel $(1-\mu)$ that corresponds to the inconsistency between the condition (x_2^1, x_3^2, x_8^3) and the formulas satisfied by the best three consecutive data (d_2, d_3, d_4), where x_2 and x_8 are true in the models supported by d_2 and d_4, respectively. The numerator turns out to be the number of five consecutive data points in which x_2, x_8 and x_{10} are true in the models supported by the first, third, and fifth data points, respectively. The right-hand side of Figure 1 shows the probability as a function of μ. The probability is undefined due to division by zero when μ is substituted by 1, whereas the limit resolves the singularity by assigning a reasonable value as μ approaches 1.

Example 8 (Continued from Example 5)**.** *We show that $\mu \in [0.5, 1)$ plays an important role that cannot be fulfilled when $\mu = 1$ or $\mu \to 1$. Consider the data shown on the left-hand side in Figure 4. Using Equation (10), we have*

$$p(x_{10}^5|x_2^1, x_3^2, x_8^3) = \frac{\overbrace{\mu^3(1-\mu)}^{d_k\ s.t.\ k\in\{2\}} + \overbrace{8\mu(1-\mu)^3}^{k\in\{3\text{-}7,9,12,14\}} + \overbrace{7(1-\mu)^4}^{k\in\{1,8,10,11,13,15,16\}}}{\underbrace{\mu^2(1-\mu)}_{k\in\{2\}} + \underbrace{4\mu(1-\mu)^2}_{k\in\{7,9,12,14\}} + \underbrace{11(1-\mu)^3}_{k\in\{1,3\text{-}6,8,10,11,13,15,16\}}}$$

$$p(x_4^5|x_2^1, x_3^2, x_8^3) = \frac{\overbrace{5\mu^2(1-\mu)^2}^{d_k\ s.t.\ k\in\{2,7,9,12,14\}} + \overbrace{6\mu(1-\mu)^3}^{k\in\{8,10,11,13,15,16\}} + \overbrace{5(1-\mu)^4}^{k\in\{1,3\text{-}6\}}}{\underbrace{\mu^2(1-\mu)}_{k\in\{2\}} + \underbrace{4\mu(1-\mu)^2}_{k\in\{7,9,12,14\}} + \underbrace{11(1-\mu)^3}_{k\in\{1,3\text{-}6,8,10,11,13,15,16\}}}.$$

The right-hand side of Figure 4 shows these probabilities as functions of μ. The three consecutive data starting from d_2 best match the condition, i.e., x_2^1, x_3^2 and x_8^3. Specifically, d_2 supports the model in which the robot is in Room 2 at Time 1, and the data two time steps later, d_4, supports the model in which the robot is in Room 8 at Time 3. This leads to the prediction that the robot will be in Room 10 at Time 5. $\mu \gtrsim 0.8$ reflects this fact.

Meanwhile, the three consecutive data starting from d_7, d_9, d_{12} and d_{14} all next best match the same condition. Specifically, d_7 implies that d_9 supports the model in which the robot is in Room 8 at Time 3, and d_9 implies that d_{10} supports the model in which the robot is in Room 3 at Time 2. These lead to the different prediction that the robot will be in Room 4 at Time 5. $\mu \lesssim 0.8$ reflects the situation where the matching quantity surpasses the matching quality.

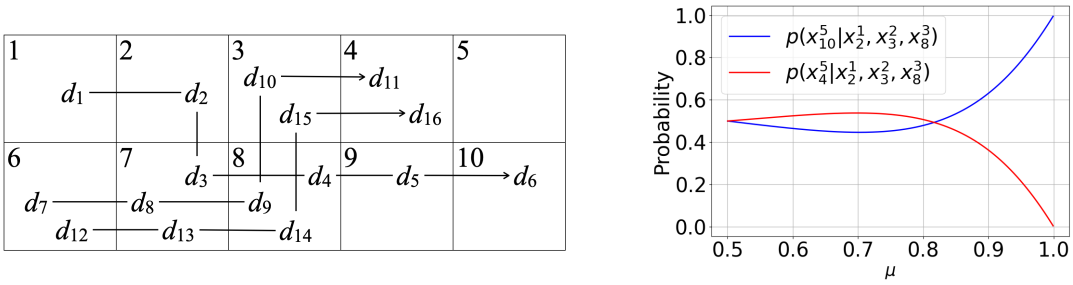

Figure 4. Left: $Data = \{d_1, d_2, ..., d_{16}\}$ collected in the same environment as shown in Figure 1. Right: The matching quantity is favoured over the matching quality when $\mu \lesssim 0.8$.

Note that $\mu \in [0.5, 1)$ admits a chance that a formula is true in a model where it is actually false. However, this does not imply an opposition to the semantics of propositional logic. Rather, we use and extend the semantics to handle formulas concerning unfounded information, such as unknown and even false information that cannot be made true in light of available data. The paper (Kido, 2025a) provides the logical justification of this aspect in terms of paraconsistent logic. The paper (Kido, 2025b) further provides statistical and probabilistic justifications in terms of maximum likelihood estimation and Bayesian networks.

3. Evaluations

3.1 Markov chains

In this section, we compare the probabilistic model of abstraction with the nth-order, discrete-time, time-homogeneous Markov chains and hidden Markov models. Let $States = \{1, 2, ..., N\}$ be the set of natural numbers for N states. For any discrete time $t \in \{1, 2, ..., T\}$, S^t is a random variable taking values in $States$. $S^t = i$ represents that the state is i at Time t. The nth-order Markov chain defines the full joint distribution as follows.

$$p(S^{1:T}) = \prod_{t=1}^{T} p(S^t | S^{t-n:t-1}) \tag{12}$$

Here, we ignore states with time zero or negative times. For example, $p(S^2|S^{-1:1}) = p(S^2|S^1)$. Maximum likelihood estimation is the statistical method most commonly used to estimate the parameters of probabilistic models solely from data. It is known that the maximum likelihood estimate for a categorical distribution is relative frequency (Russell & Norvig, 2020). Equation (12) can then be written as follows.

$$p(S^{1:T}) = \prod_{t=1}^{T} \frac{|S^{t-n:t}|}{|S^{t-n:t-1}|} \tag{13}$$

Here, $|S^{i:j}|$ denotes the number of sequences of consecutive data satisfying $S^{i:j}$. Let $\Delta \subseteq \{s^t | s^t \in s^{1:T}\}$. Using the sum rule and Equation (13),

$$
\begin{aligned}
p(\Delta) &= \sum_{s^{1:T}(\notin\Delta)} p(s^{1:T}) = \sum_{s^{1:T}(\notin\Delta)} \prod_{t=1}^{T} \frac{|s^{t-n:t}|}{|s^{t-n:t-1}|} \\
&= \begin{cases}
\sum_{s^{1:T}(\notin\Delta)} \frac{\cancel{|s^1|}}{|()|} \frac{|s^{1:2}|}{\cancel{|s^1|}} \frac{|s^{2:3}|}{|s^2|} \frac{|s^{3:4}|}{|s^3|} \cdots \frac{|s^{T-2:T-1}|}{|s^{T-2}|} \frac{|s^{T-1:T}|}{|s^{T-1}|} & \text{if } n = 1 \\
\sum_{s^{1:T}(\notin\Delta)} \frac{\cancel{|s^1|}}{|()|} \frac{\cancel{|s^{1:2}|}}{\cancel{|s^1|}} \frac{|s^{1:3}|}{\cancel{|s^{1:2}|}} \frac{|s^{2:4}|}{|s^{2:3}|} \cdots \frac{|s^{T-3:T-1}|}{|s^{T-3:T-2}|} \frac{|s^{T-2:T}|}{|s^{T-2:T-1}|} & \text{if } n = 2 \quad \cdots \\
\sum_{s^{1:T}(\notin\Delta)} \frac{\cancel{|s^1|}}{|()|} \frac{\cancel{|s^{1:2}|}}{\cancel{|s^1|}} \frac{\cancel{|s^{1:3}|}}{\cancel{|s^{1:2}|}} \frac{\cancel{|s^{1:4}|}}{\cancel{|s^{1:3}|}} \cdots \frac{\cancel{|s^{1:T-1}|}}{\cancel{|s^{1:T-2}|}} \frac{|s^{1:T}|}{\cancel{|s^{1:T-1}|}} & \text{if } n = T-1.
\end{cases}
\end{aligned} \tag{14}
$$

Here, $|()|$ denotes the number of data satisfying no constraints, and thus represents the total number of data.

We discuss the relationship between Markov chains and the probabilistic model of abstraction. To relate the propositional language to Markov chains, we use the propositional variable S_i^t, which denotes that state S has value i at time t in the Markov chain, i.e., s_i^t (or $S_i^t = 1$) iff $S^t = i$.

For random variables or their realisations z, we assume that $p(z; n)$ and $p(z; \mu)$ represent the probability $p(z)$ obtained with an nth-order Markov model and with our probabilistic model with μ, respectively. The symbol ';' denotes that its right-hand side is a variable, but not a random variable. We can show that our probabilistic model with $\mu = 1$ and $\mu \to 1$ and the highest-order, i.e., full-memory, Markov model trained using maximum likelihood estimation give the same joint distribution.

Theorem 3. *The following relation holds.*

$$
\begin{aligned}
p(s_h^1, s_i^2, ..., s_j^T; \mu = 1) &= p(s_h^1, s_i^2, ..., s_j^T; \mu \to 1) \\
&= p(S^1 = h, S^2 = i, ..., S^T = j; n = T - 1)
\end{aligned}
$$

Proof. See Appendix. □

Any marginal distributions and conditional distributions can be derived from the joint distribution using valid rules of probability theory. Theorem 3 thus establishes the equivalence between our probabilistic model and the highest-order Markov chain trained via maximum likelihood estimation.

3.2 Transparency

Probabilistic modelling, including Markov chains, generally exhibits higher transparency compared to other modern machine learning paradigms such as deep learning and reinforcement learning. This is mainly because random variables and their dependencies are made explicit in probabilistic models. However, probabilistic modelling is not highly transparent from the data perspective. This is because reasoning operates using parameters rather than data. Indeed, learning is the process to exploit data to adjust the parameters of probabilistic models, whereas reasoning is the process to exploit the parameters, not the data itself, to make predictions. The following proposition states that the probabilistic model of abstraction over formulas always refers to data.

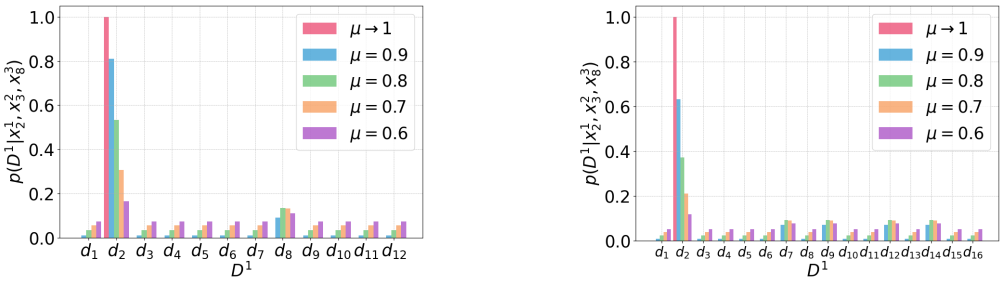

Figure 5. The conditional data distributions make the data reference transparent. Left: $Data = \{d_1, d_2, ..., d_{12}\}$ from Figure 1. Right: $Data = \{d_1, d_2, ..., d_{16}\}$ from Figure 4.

Proposition 5. *Let $\alpha \in \{x_i^t | x_i^t \in x_{1:I}^{1:T}\}$ and $\Delta \subseteq \{x_i^t | x_i^t \in x_{1:I}^{1:T}\}$. The following relation holds.*

$$p(\alpha|\Delta) = \sum_{d_k} p(\alpha|D^1 = d_k) p(D^1 = d_k|\Delta)$$

Proof. See Appendix. □

Proposition 5 implies that reasoning over formulas is a sort of Bayesian learning (Russell & Norvig, 2020). In Proposition 5, it is data that are marginalised out to infer formulas from given formulas. This is in contrast to conventional probabilistic models, in which it is parameters that are marginalised out to infer data from given data. The former represents more of a data-driven perspective, while the latter represents a model-driven perspective.

Example 9 (Continued from Examples 7 and 8)**.** *Consider the left-hand side in Figure 1 where $Data = \{d_1, d_2, ..., d_{12}\}$. By definition, $d^{2:T}$ and $m^{1:T}$ are fully determined given d^1. The summations over these values thus can be omitted using $n^t(d^1)$ and $m(n^t(d^1))$, for $t \in \{1, 2, ..., T-1\}$. Let $d_k \in Data$. Similar to Equation (11), we have*

$$\begin{aligned}
p(D^1 = d_k | x_2^1, x_3^2, x_8^3) &= \frac{p(D^1 = d_k, x_2^1, x_3^2, x_8^3)}{p(x_2^1, x_3^2, x_8^3)} \\
&= \frac{p(x_2^1|m(d_k)) p(x_3^2|m(n(d_k))) p(x_8^3|m(n^2(d_k)))}{\sum_{k\in\{2\}} \mu^2(1-\mu) + \sum_{k\in\{8\}} \mu(1-\mu)^2 + \sum_{k\in\{1,3\text{-}7,9\text{-}12\}} (1-\mu)^3} \\
&= \begin{cases} \frac{\mu^2(1-\mu)^2}{\mu^2(1-\mu)+\mu(1-\mu)^2+10(1-\mu)^3} & \text{if } k \in \{2\} \\ \frac{\mu(1-\mu)^2}{\mu^2(1-\mu)+\mu(1-\mu)^2+10(1-\mu)^3} & \text{if } k \in \{8\} \\ \frac{(1-\mu)^3}{\mu^2(1-\mu)+\mu(1-\mu)^2+10(1-\mu)^3} & \text{if } k \in \{1, 3\text{-}7, 9\text{-}12\}. \end{cases}
\end{aligned}$$

The left-hand side of Figure 5 shows the conditional distribution over $Data = \{d_1, d_2, ..., d_{12}\}$. The right-hand side shows the same type of distribution over $Data = \{d_1, d_2, ..., d_{16}\}$ we discussed in Example 8.

3.3 Hidden states

In Markov chains, it is typically assumed that the states of interest, often referred to as latent or hidden variables, are observable from the environment. The assumption does not hold in hidden Markov models, which instead assume that only effects, often referred to as observable variables, caused by these states are observable. The hidden and observable variables are clearly distinguished in the graphical models of hidden Markov models. In this section, we show that such a distinction is unnecessary for the probabilistic model of abstraction.

Example 10 (Continued from Example 6)**.** *Let us revisit Example 6 and assume that the robot location is a hidden variable. Namely, the robot cannot detect its location from the environment using its own sensors. Instead, the robot is assumed to be able to perceive the presence of an obstacle in each direction. Let N, E, S, and W be random variables representing the presence (denoted, e.g., by $N = 1$ or n) and the absence (denoted, e.g., by $N = 0$ or $\neg n$) of an obstacle to north, east, south, and west, respectively. The graphical model of the probabilistic model of abstraction we need to handle this problem is depicted on the left-hand side in Figure 6, where there is no structural distinction between the hidden and observable variables, i.e., the locations of robot and the presence of obstacles, respectively.*

Now, suppose that the robot was in Rooms 2, 3, and 8 at Time 1, 2, and 3, respectively. By the assumption, the robot only perceived n, $\neg e$, $\neg s$, and $\neg w$ in Room 2, n, $\neg e$, $\neg s$, and $\neg w$ in Room 3, and $\neg n$, $\neg e$, s, and $\neg w$ in Room 8. The probability of the robot being in Room 10 at Time 5 is given as follows.

$$
\begin{aligned}
& p(x_{10}^5 | n^1, \neg e^1, \neg s^1, \neg w^1, n^2, \neg e^2, \neg s^2, \neg w^2, \neg n^3, \neg e^3, s^3, \neg w^3) \\
= \; & \frac{\sum_{k=1}^{12} p(x_{10}^5 | m(n^4(d_k))) \prod_{a \in \{n, \neg e, \neg s, \neg w\}} A \prod_{b \in \{n, \neg e, \neg s, \neg w\}} B \prod_{c \in \{\neg n, \neg e, s, \neg w\}} C}{\sum_{k=1}^{12} \prod_{a \in \{n, \neg e, \neg s, \neg w\}} A \prod_{b \in \{n, \neg e, \neg s, \neg w\}} B \prod_{c \in \{\neg n, \neg e, s, \neg w\}} C}
\end{aligned}
$$

where $A = p(a^1 | m(d_k))$, $B = p(b^2 | m(n(d_k)))$ and $C = p(c^3 | m(n^2(d_k)))$. Using Equation (9),

$$
= \frac{\overbrace{2\mu^{11}(1-\mu)^2}^{d_1, d_2} + \overbrace{\mu^{10}(1-\mu)^3}^{d_{12}} + \overbrace{\mu^9(1-\mu)^4}^{d_3} + \overbrace{4\mu^8(1-\mu)^5}^{d_4, d_7, d_9, d_{11}} + \overbrace{2\mu^7(1-\mu)^6}^{d_5, d_8} + \overbrace{2\mu^6(1-\mu)^7}^{d_6, d_{10}}}{\underbrace{\mu^{11}(1-\mu)}_{d_1} + \underbrace{2\mu^{10}(1-\mu)^2}_{d_2, d_{12}} + \underbrace{4\mu^8(1-\mu)^4}_{d_3, d_7, d_9, d_{11}} + \underbrace{2\mu^7(1-\mu)^5}_{d_4, d_8} + \underbrace{2\mu^6(1-\mu)^6}_{d_5, d_{10}} + \underbrace{\mu^5(1-\mu)^7}_{d_6}}.
$$

$\mu = 1$ results in undefined values, whereas $\mu \in [0.5, 1)$ allows us to cancel $\mu^5(1-\mu)$. In particular, as $\mu \to 1$, we have

$$
= \lim_{\mu \to 1} \frac{2\mu^6(1-\mu) + \mu^5(1-\mu)^2 + \mu^4(1-\mu)^3 + 4\mu^3(1-\mu)^4 + 2\mu^2(1-\mu)^5 + 2\mu(1-\mu)^6}{\mu^6 + 2\mu^5(1-\mu) + 4\mu^3(1-\mu)^3 + 2\mu^2(1-\mu)^4 + 2\mu(1-\mu)^5 + (1-\mu)^6} = 0.
$$

The right-hand side of Figure 6 illustrates this result, along with the probability of the robot being in Room 9 at Time 5. Similar to Example 9, the data distribution explains why this is the case.

$$
\begin{aligned}
& p(D^1 = d_1 | n^1, \neg e^1, \neg s^1, \neg w^1, n^2, \neg e^2, \neg s^2, \neg w^2, \neg n^3, \neg e^3, s^3, \neg w^3) \\
= & \lim_{\mu \to 1} \frac{\mu^6}{\mu^6 + 2\mu^5(1-\mu) + 4\mu^3(1-\mu)^3 + 2\mu^2(1-\mu)^4 + 2\mu(1-\mu)^5 + (1-\mu)^6} = 1
\end{aligned}
$$

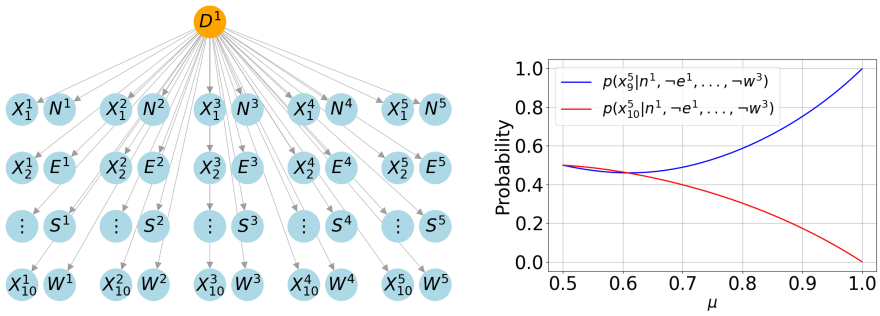

Figure 6. $p(x_i^5|n^1, \neg e^1, \neg s^1, \neg w^1, n^2, \neg e^2, \neg s^2, \neg w^2, \neg n^3, \neg e^3, s^3, \neg w^3)$, for $i \in \{9, 10\}$ as a function of μ. Undefined values at $\mu = 1$ can be resolved by the limit as $\mu \to 1$.

Namely, only the three consecutive data from d_1 best explain the given presence of obstacles. This leads to the prediction that the robot location at Time 5 is Room 9, not Room 10.

3.4 Experiments

We consider a discrete-time, discrete-space localisation problem in which a robot moves around a building with 7×7 locations, identified by coordinates (x, y) with $x, y \in \{1, 2, ..., 7\}$. All locations are accessible except those with $x, y \in \{2, 4, 6\}$, which together constitute a grid-like structure with 40 accessible and 9 inaccessible locations. The robot has a memory that stores the locations visited for the past ten time steps. We assume that the robot moves around the building at random, but it avoids entering into any locations stored in the memory. Only exception is that the robot keeps staying in the same location if all the accessible adjacent locations are in its memory.

The training and test datasets consist of time-series of 1000 and 200 locations actually explored by the robot, respectively. Figure 7 shows scatter plots comparing temporal abstraction and Markov chains with different memory size and order. Given $\mu = 1$ and zero smoothing, the first row shows that both models behave identically, even for states with zero frequency in the training data. Note that the models shown in the first column experience no zero frequency as the models are simple enough. Given $\mu \to 1$ and tiny smoothing 10^{-5}, the second row shows that states are predicted very differently if and only if the states are undefined in the first row. The Markov chain assigns equal probability, i.e., $1/49 (\simeq 0.02)$ to these states. Meanwhile, as discussed in Example 7, the limit $\mu \to 1$ allows us to make a prediction based on the training data that satisfy the specified past states as many as possible. Given $\mu = 0.9$ and moderate smoothing 10^{-1}, the third row shows that the temporal abstraction tends to correctly predict states in the training data with higher probabilities, especially as the models become more flexible.

The leaning curves depicted on the left in Figure 8 shows that n-memory temporal abstraction and nth-order Markov chains converge into the same top-1 accuracies when enough training data is provided. When the number of training data is small with respect to the model flexibility, the temporal abstraction tends to outperform the Markov chains. The learning curves depicted on the centre shows that the performance is not very sensitive to the value of μ compared to the memory

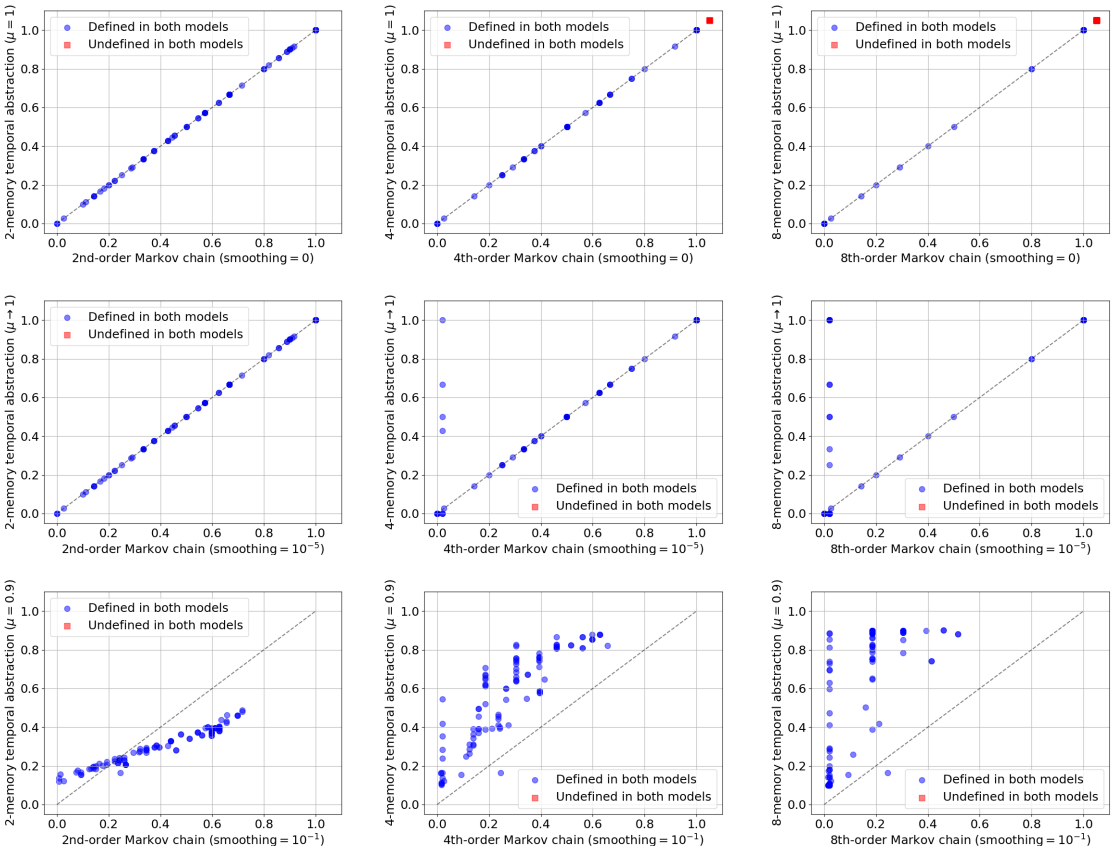

Figure 7. Scatter plots comparing temporal abstraction and Markov chains with different memory size and order. In each graph, each of the 200 points represents the conditional probability of each state given its past states. These states are sequentially extracted from the 200 test data, and the probability is calculated using 1000 training data. The ideal model thus assigns probability one to each of the conditional probabilities. The red circle points plotted outside the probability range are states in the test data that cannot be assigned a probability due to zero frequency in the training data.

size. The heat map suggests that there is a peak of the top-1 accuracy around memory size 4 and $\mu = 0.9$.

4. Conclusions

Data is typically a product of domain knowledge encoded in a probabilistic model. We flipped this long-standing convention and pursued the view that domain knowledge is a product of data encoded in a probabilistic model. This view allowed us to develop a novel probabilistic model grounding temporal propositional reasoning in data. The probabilistic model is equivalent to a highest-order Markov chain trained using maximum likelihood estimation. We discussed how the probabilistic model dissolves the issues of a huge hypothesis space, data scarcity, and data transparency.

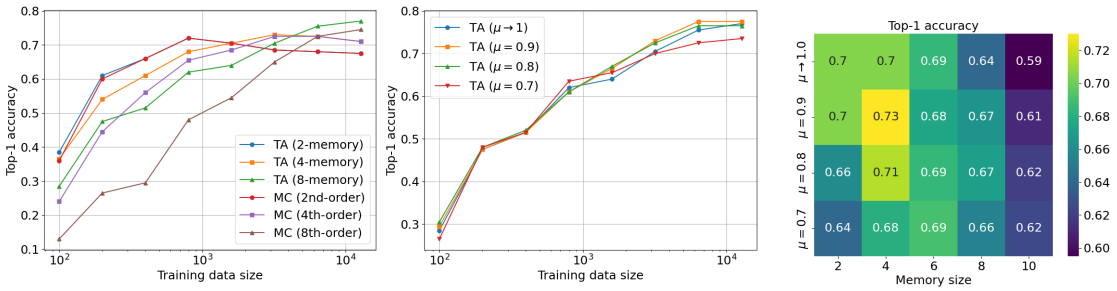

Figure 8. Left: Learning curves of TA (temporal abstraction) with different memory sizes. We assumed $\mu \to 1$. The baselines are nth-order MCs (Markov chains). One of the highest-probability states is chosen at random for the top-1 accuracy in case of ties. Centre: Learning curves of 8-memory temporal abstraction with different μ values. Right: A heat map visualising top-1 accuracies of temporal abstraction with different combinations of μ values and memory sizes. We assumed 1600 training data. All the results shown in Figure 8 were averaged over the 200 test data.

The importance of world models is widely discussed in AI research. From the logic perspective, each valuation in propositional logic, or simply each row of a truth table, corresponds to a simple, static version of the world model. We argued that deriving the world model from data constitutes abstraction, i.e., selective ignorance. This can be viewed as a sort of a deductive process as it extracts information contained within the given data. This contrasts with the prevailing view that deriving the world model from data is an inductive process, which extracts information beyond the given data. Our view is supported by the fact that it not only provides a simple and unifying framework for logical reasoning over data, models, and knowledge, but also reduces reasoning with exponentially growing world models to reasoning that scales linearly with data.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. 1 New York Plaza, Suite 4600, New York, NY 10004-1562: Springer New York, NY.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychol Rev.*, *127(3)*, 412–441.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*, 687–701.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49(10)*, 1295–1306.

Kido, H. (2025a). Inference of abstraction for human-like logical reasoning. *Machine Learning, Optimization, and Data Science* (pp. 191–206). Cham: Springer Nature Switzerland.

Kido, H. (2025b). Inference of abstraction for human-like probabilistic reasoning. *Machine Learning, Optimization, and Data Science* (pp. 116–131). Cham: Springer Nature Switzerland.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350(6266)*, 1332–1338.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America*, *20*, 1434–1448.

Mor, B., Garhwal, S., & Kumar, A. (2021). A systematic review of hidden markov models and their applications. *Archives of Computational Methods in Engineering*, *28*, 1429–1448.

Murphy, K. P., & Bach, F. (2012). *Machine learning – a probabilistic perspective*. 255 Main Street, 9th Floor Cambridge, MA 02142: MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Burlington, Massachusetts: Morgan Kaufmann; 1st edition.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*, 257–286.

Russell, S., & Norvig, P. (2020). *Artificial intelligence : A modern approach, fourth edition*. London, England: Pearson Education, Inc.

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, *107*, 102632.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10(7)*, 309–318.

Appendix A. Proofs

Proposition 1. When $t = 1$, both the left- and right-hand sides are $p(D^1)$. When $t \neq 1$, the right hand side can be expanded as follows for all its realisations.

$$\begin{aligned} p(d^t|d^{t-1}) &= \frac{p(d^{t-1}, d^t)}{p(d^{t-1})} = \frac{\sum_{d^{1:t-2}} \sum_{m^{1:t-1}} \sum_{x_{1:I}^{1:t-1}} p(d^{1:t}, m^{1:t-1}, x_{1:I}^{1:t-1})}{p(d^{t-1})} \\ &= \frac{\sum_{d^{1:t-2}} \sum_{m^{1:t-1}} \sum_{x_{1:I}^{1:t-1}} p(d^t|d^{1:t-1}, m^{1:t-1}, x_{1:I}^{1:t-1}) p(d^{1:t-1}, m^{1:t-1}, x_{1:I}^{1:t-1})}{p(d^{t-1})} \end{aligned}$$

Here, the first line is an application of the sum rule, and the second line is an application of the product rule. By definition, the value of $p(d^t|d^{1:t-1}, m^{1:t-1}, x_{1:I}^{1:t-1})$ does not depend on $d^{1:t-2}$, $m^{1:t-1}$ or $x_{1:I}^{1:t-1}$. Therefore, the conditional probability can be moved to the outside of the summations.

$$\begin{aligned} &= \frac{p(d^t|d^{1:t-1}, m^{1:t-1}, x_{1:I}^{1:t-1}) \sum_{d^{1:t-2}} \sum_{m^{1:t-1}} \sum_{x_{1:I}^{1:t-1}} p(d^{1:t-1}, m^{1:t-1}, x_{1:I}^{1:t-1})}{p(d^{t-1})} \\ &= \frac{p(d^t|d^{1:t-1}, m^{1:t-1}, x_{1:I}^{1:t-1}) p(d^{t-1})}{p(d^{t-1})} = p(d^t|d^{1:t-1}, m^{1:t-1}, x_{1:I}^{1:t-1}) \end{aligned}$$

The second line is an application of the sum rule for marginalisation. □

Proposition 2. The proof has a structure similar to Proposition 1.

$$p(m^t|d^t) = \frac{\sum_{d^{1:t-1}} \sum_{m^{1:t-1}} \sum_{x_{1:I}^{1:t-1}} p(m^t|d^{1:t}, m^{1:t-1}, x_{1:I}^{1:t-1}) p(d^{1:t}, m^{1:t-1}, x_{1:I}^{1:t-1})}{p(d^t)}$$

By definition, the value of $p(m^t|d^{1:t}, m^{1:t-1}, x_{1:I}^{1:t-1})$ depends only on d^t. Therefore, the conditional probability can be moved to the outside of the summations.

$$= \frac{p(m^t|d^{1:t}, m^{1:t-1}, x_{1:I}^{1:t-1}) \sum_{d^{1:t-1}} \sum_{m^{1:t-1}} \sum_{x_{1:I}^{1:t-1}} p(d^{1:t}, m^{1:t-1}, x_{1:I}^{1:t-1})}{p(d^t)}$$
$$= \frac{p(m^t|d^{1:t}, m^{1:t-1}, x_{1:I}^{1:t-1}) p(d^t)}{p(d^t)} = p(m^t|d^{1:t}, m^{1:t-1}, x_{1:I}^{1:t-1})$$

Here, the second line is an application of the sum rule for marginalisation. □

Proposition 3. The proof has a structure similar to Propositions 1 and 2.

$$p(x_i^t|m^t) = \frac{p(m^t, x_i^t)}{p(m^t)} = \frac{\sum_{d^{1:t}} \sum_{m^{1:t-1}} \sum_{x_{1:I}^{1:t-1}} \sum_{x_{1:i-1}^t} p(d^{1:t}, m^{1:t}, x_{1:I}^{1:t-1}, x_{1:i}^t)}{p(m^t)} =$$
$$= \frac{\sum_{d^{1:t}} \sum_{m^{1:t-1}} \sum_{x_{1:I}^{1:t-1}} \sum_{x_{1:i-1}^t} p(x_i^t|d^{1:t}, m^{1:t}, x_{1:I}^{1:t-1}, x_{1:i-1}^t) p(d^{1:t}, m^{1:t}, x_{1:I}^{1:t-1}, x_{1:i-1}^t)}{p(m^t)}$$

By definition, the value of $p(x_i^t|d^{1:t}, m^{1:t}, x_{1:I}^{1:t-1}, x_{1:i-1}^t)$ depends only on m^t. Therefore, the conditional probability can be moved to the outside of the summations.

$$= \frac{p(x_i^t|d^{1:t}, m^{1:t}, x_{1:I}^{1:t-1}, x_{1:i-1}^t) \sum_{d^{1:t}} \sum_{m^{1:t-1}} \sum_{x_{1:I}^{1:t-1}} \sum_{x_{1:i-1}^t} p(d^{1:t}, m^{1:t}, x_{1:I}^{1:t-1}, x_{1:i-1}^t)}{p(m^t)}$$
$$= \frac{p(x_i^t|d^{1:t}, m^{1:t}, x_{1:I}^{1:t-1}, x_{1:i-1}^t) p(m^t)}{p(m^t)} = p(x_i^t|d^{1:t}, m^{1:t}, x_{1:I}^{1:t-1}, x_{1:i-1}^t)$$

Here, the second line is an application of the sum rule for marginalisation. □

Theorem 2. Equation (7) can be developed as follows by expanding the product over time and the summations over models.

$$\sum_{d^{1:T}} \sum_{m^{1:T}} \left[p(d^1) p(m^1|d^1) \prod_{i=1}^{I} p(X_i^1|m^1) ... p(d^T|d^{T-1}) p(m^T|d^T) \prod_{i=1}^{I} p(X_i^T|m^T) \right]$$
$$= \sum_{d^{1:T}} \left[p(d^1) \sum_{m^1} \left[p(m^1|d^1) \prod_{i=1}^{I} p(X_i^1|m^1) \right] ... p(d^T|d^{T-1}) \sum_{m^T} \left[p(m^T|d^T) \prod_{i=1}^{I} p(X_i^T|m^T) \right] \right]$$

By definition, each data point supports a single model. We can thus remove the model summations.

$$= \sum_{d^{1:T}} \left[p(d^1) \prod_{i=1}^{I} p(X_i^1|m(d^1)) ... p(d^T|d^{T-1}) \prod_{i=1}^{I} p(X_i^T|m(d^T)) \right]$$

Expanding the summations over data, we have

$$= \sum_{d^1} \left[p(d^1) \prod_{i=1}^{I} p(X_i^1|m(d^1)) ... \sum_{d^T} \left[p(d^T|d^{T-1}) \prod_{i=1}^{I} p(X_i^T|m(d^T)) \right] ... \right].$$

By definition, data changes deterministically. We can thus remove the summations over data, for all time steps except $t = 1$.

$$= \sum_{d^1} \left[p(d^1) \prod_{i=1}^{I} p(X_i^1 | m(d^1)) ... \prod_{i=1}^{I} p(X_i^T | m(n^{T-1}(d^1))) \right]$$

Here, $n^{T-1}(d^1) = n(d^{T-1})$. Since $p(D^1)$ is the uniform distribution over the K data, i.e., $Data = \{d_1, d_2, ..., d_K\}$, we can move $p(d^1)$ outwards and then introduce the product over time steps. □

Proposition 4. Since the interpretation of X_i conforms to the semantics of propositional logic, $[\![X_i]\!]_{m(n^{t-1}(d_k))} = 0$ iff $[\![\neg X_i]\!]_{m(n^{t-1}(d_k))} = 1$. From Theorem 2,

$$p(X_i^t = 0) = \frac{1}{K} \sum_{k=1}^{K} p(X_i^t = 0 | m(n^{t-1}(d_k))) = \frac{1}{K} \sum_{k=1}^{K} p(\neg X_i^t = 1 | m(n^{t-1}(d_k))) = p(\neg X_i^t = 1).$$

This holds regardless of the value of $\mu \in [0.5, 1]$. □

Theorem 3. By definition, if formula S_i is true in model m^t, i.e., $[\![S_i]\!]_{m^t} = 1$, then $p(s_i^t | m^t) = \mu = 1$, for $\mu = 1$ and $\mu \to 1$. If S_i is false in m^t, i.e., $[\![S_i]\!]_{m^t} = 0$, then $p(s_i^t | m^t) = 1 - \mu = 0$, for $\mu = 1$ and $\mu \to 1$. We thus have

$$\begin{aligned} p(s_h^1, s_i^2, ..., s_j^T) &= \frac{1}{K} \sum_{k=1}^{K} \left[p(s_h^1 | m(d_k)) p(s_i^2 | m(n(d_k))) \cdots p(s_j^T | m(n^{T-1}(d_k))) \right] \\ &= \frac{1}{K} \sum_{k=1}^{K} \left[[\![S_h]\!]_{m(d_k)} [\![S_i]\!]_{m(n(d_k))} \cdots [\![S_j]\!]_{m(n^{T-1}(d_k))} \right]. \end{aligned}$$

The expression inside the summation turns out to be one if S_h, S_i, ..., S_j are sequentially true in the models supported by the T consecutive data from d_k, and zero otherwise. Since d_k ranges from d_1 to d_K, the summation turns out to be the number of such sequences. From Equation (14),

$$= \frac{|S^1 = h, S^2 = i, ..., S^T = j|}{K} = p(S^1 = h, S^2 = i, ..., S^T = j).$$

□

Proposition 5. The left-hand side can be expanded as follows.

$$p(\alpha | \Delta) = \frac{p(\alpha, \Delta)}{p(\Delta)} = \frac{\sum_{d_k} p(\alpha, \Delta, D^1 = d_k)}{p(\Delta)}$$

From the the rightmost graph of Figure 2, we have

$$= \frac{\sum_{d_k} p(\alpha | d_k) p(\Delta | d_k) p(d_k)}{p(\Delta)} = \frac{\sum_{d_k} p(\alpha | d_k) p(d_k | \Delta) p(\Delta)}{p(\Delta)} = \sum_{d_k} p(\alpha | d_k) p(d_k | \Delta).$$

□