Typical Models: Minimizing False Beliefs

Eliezer L. Lozinskii
School of Computer Science and Engineering
The Hebrew University, Jerusalem 91904, Israel
email: lozinski@cs.huji.ac.il

Abstract

A knowledge system S describing a part of real world does in general not contain complete information. Reasoning with incomplete information is prone to errors since any belief derived from S may be false in the present state of the world. A false belief may suggest wrong decisions and lead to harmful actions. So an important goal is to make false beliefs as unlikely as possible. This work introduces the notions of *typical atoms* and *typical models*, and shows that reasoning with typical models minimizes the expected number of false beliefs over all ways of using incomplete information. Various properties of typical models are studied, in particular, correctness and stability of beliefs suggested by typical models, and their connection to oblivious reasoning.

Keywords: Incomplete information, reasoning errors, false beliefs, typical models, evidence, oblivious reasoning, counting models.

1 Introduction

Let us consider a knowledge system S describing a part of real world. The knowledge contained in the system consists of data describing properties of various objects of the world, their mutual relationship, laws governing their behavior and evolution. For example, consider a system S of medical knowledge about the "world" of a hospital. S contains description of diseases (their causes, development, consequences, examination, symptoms, treatment, prevention), information about various medicament (their composition, therapeutic activity, dosage, directions for use, interactions, side effects), description of the hospital (its structure, management, location), personal data of the hospital patients (their medical history, test results), general rules of medicine, etc. An important decision that has to be made by a physician for his or her patient is determining the right diagnosis and the best treatment. The physician may wish to consult the vast amount of knowledge collected in the system. Will the system help the physician to make a right decision? This depends to a large extent on the way the knowledge is used for deriving conclusions.

Let us define important features of S and their correspondence to the world it describes.

S is presented in a first order language[1]. S is consistent having a set $MOD(S)$ of models. Each model of S is a set of ground atomic formulas expressed in terms of values assigned to various objects and parameters of the world (such as names of patients, quantities of medicament, etc.).

The multitude of models of S reflects uncertainty regarding actual values of some of these parameters. For instance, as long as neither a final diagnosis nor a treatment of a patient A is determined, there are, say, two possible diagnoses: D_1 (with possible treatments T_1 or T_2) or D_2 (with T_3 or T_4). So $MOD(S)$ may contain four different models, each including $D_1\ T_1$ or $D_1\ T_2$ or $D_2\ T_3$ or $D_2\ T_4$.

A set of values of all parameters of the world (including those not presented in S) determines its *state*. With regard to patient A the hospital world has at least four possible states, each represented in S by one of the four models mentioned above.

Since the available knowledge of the real world is incomplete in general, it may happen that a model μ of S represents several states of the world that differ in the reality but are indistinguishable from the point of view of the information presented in S. For S the states represented by μ constitute an equivalence class of possible states of the world called a *possible world* (denoted w in the sequel).

S describes the world *faithfully* in the sense that every possible state of the world belongs to a possible world represented by a model of S, and every model of S represents a non-empty set of possible states of the world. From the point of view of S the real world appears as a set W of possible worlds such that there is a bijection between $MOD(S)$ and W.

A user of S having a query whether a formula F is true in the present state of the world applies to S and expects to get an answer based on the information stored in S. If F (or $\neg F$) is a logical consequence of S then S contains *complete* information about F. In this case F is true (false) in all models of S and all possible worlds. Hence F is certainly true (false) in the present state. However, if neither F nor $\neg F$ follows from S then the information in S regarding F is *incomplete* and does not facilitate derivation of a definite answer to the query. But this does not diminish the need or importance of a reasonable answer. A physician cannot delay for a long time a treatment of a patient just because he or she is not yet certain about the final diagnosis. Travelers reaching a crossroads would not just stay there even if they are not sure which way leads to their destination.

Mark Twain wrote, "The trouble with the world is not that people know too little, but that they know so many things that ain't so." Even given an extensive knowledge of the world, its incompleteness makes erroneous judgment inevitable. In the present state of the world a formula F has a certain value

[1]Many existing Knowledge Systems like OpenCyc (7) are formulated in languages based on Predicate Calculus.

although this value may be uncertain from the standpoint of S. If the information about F contained in S is incomplete (briefly, F is incomplete in S), we have to find a way of reasoning producing a *belief* regarding F which is *credible* in the sense that it stands a good chance of being true in the reality. So a system of automated reasoning must be able to answer the following query:

> Given a formula F and a system S describing faithfully a world, what is a most credible belief regarding the truth of F in the present state of the world?

In order to answer various multiple queries consistently a reasoner has to choose one particular model μ of S (a *preferred model*) and then believe that F is true in the reality iff $\mu \models F$. If $S \models F$ or $S \models \neg F$ then the choice of μ does not matter; however, if F is incomplete in S then F is true in some models, but false in the others. Which is the correct value of F in the present state of the world? With any choice of μ there is a non-zero probability that the belief in F implied by μ is false in the present state of the world. So reasoning with incomplete information is *prone to errors*.

The way of choosing the preferred model provides a semantics for the process of reasoning. Whatever this way is, errors are inevitable since the preferred model may not fully conform to the present state of the world. The smaller the expected number (or the severity) of errors, the more reliable the semantics. Numerous approaches to reasoning with incomplete information have been developed including *Nonmonotonic Logics* (1; 5; 19) and methods based on *Semantics of Minimal Models* (3; 9; 15; 16; 22). Neither of the previous work considered minimization of the risk that beliefs sanctioned by the proposed semantics are false in the real world. A false belief may suggest wrong decisions and lead to harmful actions. As reasoning errors caused by incompleteness of information are inevitable, minimization of the number and likelihood of false beliefs becomes practically important a goal.

The following sections introduce the *semantics of typical models* and show that it minimizes the expected number of erroneous beliefs over all ways of reasoning with incomplete information.

2 Evidence

At any moment the world is in exactly one of its possible states, so in exactly one of possible worlds represented by the corresponding model of S. Let $p(w)$ denote the probability that at a randomly chosen moment, the world is in a state belonging to a possible world $w \in W$ represented by a model μ of S. Then to every $\mu \in MOD(S)$ representing the corresponding $w \in W$ one may assign a probability $p(\mu) = p(w)$ such that $\sum_{\mu \in MOD(S)} p(\mu) = \sum_{w \in W} p(w) = 1$. So the probability $p(F)$ that a formula F is true in the present state of the world is

$$p(F) = \sum_{\mu \in MOD(S \cup \{F\})} p(\mu) \tag{1}$$

where $MOD(S \cup \{F\}) = \{\mu | \mu \in MOD(S) \wedge \mu \models F\}$ is the set of models of S implying F.

If $p(F) > 0.5$ then it is reasonable to believe that F is more likely to be true than false in the present state of the world, and the larger $p(F)$, the more credible this belief.

The problem, however, is that in most practical cases there is no reliable information regarding the distribution of $p(w)$. In the absence of this information let us assume just for the moment that all possible worlds are equiprobable, and sets W and $MOD(S)$ are finite. Appendix B shows a way to relax these limitations in case that certain knowledge is available about probability of possible worlds and structure of a given system and its domain.

The assumptions of the previous paragraph lead to the following approach.

Definition 2.1 (Principle of majority of models, PMM) *Believe that a formula F is more likely to be true than false in a state of the world if F is true in a majority of models of S. The larger the majority, the more credible the belief.* □

A reasonable semantics should respect the power of majority; indeed, F is true (false) in S if it is true (false) in all models of F. Obeying such an unanimity, would it be reasonable to disregard a majority of 99.9% or even 80%?

As PMM suggests a belief regarding the truth value of F, we may say that the set of models of S offers an *evidence* of F, $E(S, F)$. We would like the evidence to provide a quantitative measure of credibility of the corresponding belief. To normalize the value of evidence for all S and F such that $0 \leq E(S, F) \leq 1$, it is reasonable to require that $E(S, F) = 1$ for $S \models F$, $E(S, F) = 0$ for $S \models \neg F$, and $E(S, F) + E(S, \neg F) = 1$ for all S, F. More requirements are presented in (12) leading to the following definition.

Definition 2.2 Evidence *of F in S:*

$$E(S, F) = \frac{|MOD(S \cup \{F\})|}{|MOD(S)|}. \quad \square \tag{2}$$

PMM suggests that F is true if $E(S, F) > 0.5$ or $E(S, F) = 0.5$ (the latter is chosen to avoid ambiguity; see also Definition 4.1). Given a query regarding the truth value of F, a reasoner may not only return 'true' or 'false', but also attach the value of $E(S, F)$ to the answer to give a measure of credibility of the latter. In cases where accepting an erroneous answer can have very undesirable consequences, a query can require a certain level of credibility, for example, ignoring answers with evidence less than 0.9.

3 Oblivious vs. non-oblivious reasoning

In the absence of sufficient statistical information, the evidence $E(S, F)$ is regarded as an approximation of the probability $p(F)$ that F is true in a randomly

chosen possible world, that is the probability that the belief in the truth of F is correct in the present state of W.

Consider a reasoner R that forms beliefs in order to answer a series of queries $F_1, \ldots, F_k$. Denote by $R(F_i)$ his belief regarding the truth of F_i. If the reasoner computes $R(F_i)$ as his answer to F_i without taking into account the previous beliefs $R(F_j)$ $(1 \leq j < i)$ preceding $R(F_i)$, let us call this way of reasoning *oblivious*. Then if it turns out that there is no model of S in which all of $R(F_1), \ldots, R(F_i)$ are true, then the beliefs of R are inconsistent with S which is unacceptable.

Oblivious reasoning with incomplete information may lead to inconsistency. Indeed, let $M_{R(F_i)}$ denote a set of all models of S in which $R(F_i)$ is true. Then the set of beliefs $\{R(F_1), \ldots, R(F_k)\}$ is consistent with S (and so, holds in some state of W) iff

$$\bigcap_{i=1}^{k} M_{R(F_i)} \neq \emptyset. \tag{3}$$

For all queries F incomplete in S, $M_{R(F)}$ is a proper subset of $MOD(S)$, so the size of their intersection (3) is a monotone decreasing function of k such that for a large k condition (3) may not hold [2]. This does not happen if the reasoning is *non-oblivious* such that in derivation of $R(F_i)$ all previously produced beliefs are taken into consideration. One way of doing so is to derive $R(F_i)$ from $S \cup \{R(F_1), \ldots, R(F_{i-1})\}$. In this case, however, the value of each belief depends on the order of queries in their sequence.

Example 3.1 $S = \{a \vee b, b \vee c, c \vee a, \neg a \vee \neg b \vee \neg c\}$;
$MOD(S) = \{\{a, b, \neg c\}, \{a, \neg b, c\}, \{\neg a, b, c\}\}$.
Queries: $F_1 = a, F_2 = b, F_3 = c$; $k = 3$; $E(S, a) = E(S, b) = E(S, c) = 2/3$.
Obliviously: $R(a) = R(b) = R(c) =' true'$ *which is inconsistent with* S.
Non-obliviously: let $S_0 = S$, $S_i = S_{i-1} \cup \{R(F_i)\}$ *for* $1 \leq i \leq k$. *Then*
$E(S_0, a) = 2/3; R(a) =' true'$; $S_1 = \{a, b \vee c, \neg b \vee \neg c\}$;
$E(S_1, b) = 1/2; R(b) =' true'$; $S_2 = \{a, b, \neg c\}$;
$E(S_2, c) = 0; R(c) =' false'$. *All these beliefs hold in the first model of* S. □

Non-oblivious reasoning requires keeping track of many previously produced beliefs, so in general it is more time-consuming than its oblivious counterpart. Thus it would be helpful to determine sets of queries that can be answered obliviously in any order without any risk of inconsistency. A trivial example is a set of all formulas F such that $S \models F$. Subsection 4.3 presents less obvious sets allowing oblivious reasoning.

4 Semantics of typical models

This section introduces the basic notions of typical atoms and typical models, and studies stability of the corresponding beliefs.

[2] For instance, in Example 3.1 expression (3) holds for $k = 2$, but not for $k = 3$.

4.1 Typical atoms

S is supposed to be formulated in a first order language, so using the terminology of Predicate Calculus let the *base* of S be a set of all ground instances of all atomic formulas corresponding to all predicates occurring in S:

$$Base(S) = \{P^{(k)}(t_1, \ldots, t_k)\}, \tag{4}$$

such that $P^{(k)}(x_1, \ldots, x_k)$ occurs in S, D_i is the *domain* of x_i, and $t_i \in D_i$ for $1 \leq i \leq k$.

Definition 4.1 *For each ground atomic formula* $\mathbf{a} \in Base(S)$ *let* $\hat{\mathbf{a}}$ *denote the* typical atom *corresponding to* $\mathbf{a}$ *such that the evidence of* $\hat{\mathbf{a}}$ *is at least as large as that of* $\neg\hat{\mathbf{a}}$. *So*[3]

$$\hat{\mathbf{a}} = \begin{cases} \mathbf{a} & if\ E(S, \mathbf{a}) \geq 0.5 \\ \neg\mathbf{a} & otherwise \end{cases} \tag{5}$$

For a formula F *we define its* typical value $\widehat{F}$ *by substituting* F *for* $\mathbf{a}$ *in expression (5).*

Evidence $E(S, a)$ is introduced in order to be used as an approximation of $p(a)$. The larger the difference $|E(S, a) - 0.5|$, the better this approximation. However, if $E(S, a)$ is close to 0.5, it may diverge from $p(a)$ even qualitatively such as $E(S, a) > 0.5$ but $p(a) < 0.5$. But in the absence of a sufficient statistics regarding $p(w)$ one has to rely on $E(S, a)$ and believe that any typical atom is not less likely to be true than false in the present state of the world. On the other hand, the need to avoid inconsistency may force a non-oblivious reasoner to adopt beliefs in negation of some typical atoms. So questions arising in any reasoning system intended for answering multiple queries are:

> Given a set of queries $\{F_1, \ldots, F_k\}$, is there a state of the world in which beliefs in the truth of typical values $\{\widehat{F}_1, \ldots, \widehat{F}_k\}$ hold for all $1 \leq i \leq k$, i.e. is there a model m of S such that $m \models \bigwedge_{i=1}^{k} \widehat{F}_i$? What is the value of evidence $E(S, \bigwedge_{i=1}^{k} \widehat{F}_i)$?

The answer to the first question is positive if the latter evidence is larger than zero.

Let $A^{(k)}$ be a set of k literals l such that $l \in \{a, \neg a\}$, $a \in Base(S)$, and $A^{(k)}_{\wedge} = \bigwedge_{l \in A^{(k)}} l$. The following theorem estimates the value of evidence of $A^{(k)}_{\wedge}$.

Theorem 4.1 $\max(0, \alpha, \beta) \leq E(S, A^{(k)}_{\wedge}) \leq \min(1, \gamma)$, *where*

$$\alpha = \sum_{l \in A^{(k)}} E(S, l) \quad - \quad k + 1, \tag{6}$$

$$\beta = 1 - \frac{2^{|Base(S)|-k}(2^k - 1)}{|MOD(S)|}, \quad \gamma \quad = \quad \frac{2^{|Base(S)|-k}}{|MOD(S)|}. \tag{7}$$

[3]If $E(S, a) = 0.5$ then $E(S, a) = E(S, \neg a)$, and so any one of a or $\neg a$ can be considered a typical atom. However, in practice a proper choice of one of a or $\neg a$ should be made based on relevant knowledge of the real world.

Proof. First, we prove by induction on k that

$$|MOD(S \cup A^{(k)})| \geq \sum_{l \in A^{(k)}} |MOD(S \cup \{l\})| - (k-1)|MOD(S)|. \qquad (8)$$

Base. For $k = 1$ inequality (8) holds trivially.
Step. Let M_1, M_2 be subsets of $MOD(S)$, then

$$|M_1 \cap M_2| \geq |M_1| + |M_2| - |MOD(S)|.$$

If inequality (8) holds for all $1 \leq i \leq j$, then it holds for $j+1$. Indeed, let $A^{(j+1)} = A^{(j)} \cup \{l'\}$, then

$$|MOD(S \cup A^{(j+1)})| = |MOD(S \cup A^{(j)}) \cap MOD(S \cup \{l'\})| \geq$$

$$\sum_{l \in A^{(j)}} |MOD(S \cup \{l\})| - (j-1)|MOD(S)| + |MOD(S \cup \{l'\})| - |MOD(S)| =$$

$$\sum_{l \in A^{(j+1)}} |MOD(S \cup \{l\})| - j|MOD(S)|.$$

Further, by (8) and since evidence is a non-negative value,

$$E(S, A^{(k)}_{\wedge}) = \frac{|MOD(S \cup A^{(k)})|}{|MOD(S)|} \geq \max\left(0, \sum_{l \in A^{(k)}} E(S,l) - k + 1\right). \qquad (9)$$

Next, $A^{(k)}_{\wedge}$ is true in $2^{|Base(S)|-k}$ interpretations of S but false in the rest of them. So

$$1 - \frac{2^{|Base(S)|}(1-2^{-k})}{|MOD(S)|} \leq E(S, A^{(k)}_{\wedge}) \leq \min\left(1, \frac{2^{|Base(S)|-k}}{|MOD(S)|}\right). \qquad (10)$$

Expressions (9) and (10) complete the proof. □

Let $\mathcal{F}^{(k)}$ be a set of k formulas, and $\mathcal{F}^{(k)}_{\wedge} = \bigwedge_{F \in \mathcal{F}^{(k)}} F$. If $A^{(k)}$ is replaced with $\mathcal{F}^{(k)}$ then Theorem 4.1 implies the following

Corollary 4.1 *(i)* $E(S, \mathcal{F}^k_{\wedge}) \geq \sum_{F \in \mathcal{F}^{(k)}} E(S,F) - k + 1$*;*

(ii) For all formulas ϕ, ψ*, if* $E(S, \hat{\phi}) > 0.5$ *then* $\hat{\phi} \wedge \hat{\psi}$ *is consistent with* S*;*

(iii) If $E(S,F) = 0.5$ *call* F *a* neutral *formula. If there are two neutral formulas in a set* $\mathcal{F}^{(k)}$ *then* $\mathcal{F}^{(k)}_{\wedge}$ *may be inconsistent with* S*.* □

4.2 Typical models

Let $T(S)$ denote the set of all typical atoms of S, and $T(m)$ be the set of all typical atoms contained in a model m:

$$T(S) = \{\hat{a} | a \in Base(S)\}, \quad T(m) = \{\hat{a} | \hat{a} \in m\} = T(S) \cap m. \qquad (11)$$

Definition 4.2 *If there exists a model μ of S such that $T(\mu) = T(S)$ then μ is* the most typical model *of S. For all $m \in MOD(S)$, if there is no model m' of S such that $T(m) \subset T(m')$ then m is a* typical model *of S.* □

A system S may have no most typical model, but every S has a typical one. Indeed, every typical atom $\hat{a}$ is consistent with S, so there is a model m containing $\hat{a}$. Either m is a typical model of S or there is a typical model μ such that $T(m) \subset T(\mu)$.

Suppose, a reasoner R prefers a model m assuming that it describes most trustfully the present possible world w. Then m represents the set of R's beliefs, but because of incompleteness of S some of the beliefs may be false in w.

Definition 4.3 *A formula F is false in a possible world w with probability $1 - p(F)$ (expression (1)). Let the* erratum *$ER(A)$ of a set of literals A be the expected proportion of its literals that are false in a randomly chosen possible world w. Then taking $E(S, l)$ as an approximation of $p(l)$ we get*

$$ER(A) = 1 - \frac{1}{|A|} \sum_{l \in A} E(S, l). \tag{12}$$

Theorem 4.2 *(i) For all $m \in MOD(S)$ there is a typical model μ of S such that $ER(\mu) \leq ER(m)$.*

(ii) If μ is the most typical model of S then for all $m \in MOD(S)$ $ER(\mu) \leq ER(m)$.

Proof. (i) If m is a typical model of S then (i) holds trivially, else there is a typical model μ such that $T(m) \subset T(\mu)$. Denote $\delta_1 = T(\mu) - T(m) = \mu - m$, $\delta_2 = m - \mu$, $B = |Base(S)|$. All literals of δ_1 are typical atoms, while all literals of δ_2 are their negations. So there is a bijection between δ_1 and δ_2 such that to every literal $\hat{a} \in \delta_1$ corresponds $\neg\hat{a} \in \delta_2$, and vice versa. Since $E(S, \hat{a}) \geq E(S, \neg\hat{a})$ for all $a \in Base(S)$, we have

$$ER(\mu) - ER(m) = \frac{1}{B} \sum_{\hat{a} \in \delta_1} (E(S, \neg\hat{a}) - E(S, \hat{a})) \leq 0. \tag{13}$$

(ii) If μ is the most typical model of S then for all $m \in MOD(S)$ $T(m) \subset T(\mu)$, hence $ER(\mu) \leq ER(m)$. □

By Theorem 4.2, if there exists the most typical model of S then it is the most trustworthy one among all models of S. Otherwise there is a typical model with a minimum value of erratum among all models of S.

Let $ER(mtm)$, $ER(rand)$, $ER(worst)$, $E(S)$ denote respectively the erratum of the most typical model, the expected erratum of a randomly chosen model, the erratum of a model containing no typical atoms, the average evidence of a typical atom of S. Then

$$E(S) = \frac{1}{B} \sum_{a \in Base(S)} E(S, \hat{a}), \qquad ER(mtm) = 1 - E(S), \tag{14}$$

$$ER(rand) = \frac{2}{B} \sum_{a \in Base(S)} E(S, \hat{a})(1 - E(S, \hat{a})), \qquad ER(worst) = E(S) \quad (15)$$

such that

$$\lim_{E(S)\to 1} \frac{ER(rand)}{ER(mtm)} = 2, \qquad \lim_{E(S)\to 1} \frac{ER(worst)}{ER(mtm)} = \infty. \quad (16)$$

Since the most typical model of a given system would be the most trustworthy one, it should be preferred by any rational reasoner. So the existence of a most typical model is a practically important characteristic of any knowledge system. Let $p(mtm)$ denote the probability that a given system S has the most typical model. The probability that a randomly chosen model of S contains all typical atoms is $\prod_{a \in Base(S)} E(S, \hat{a})$. Then

$$p(mtm) = 1 - \left(1 - \prod_{a \in Base(S)} E(S, \hat{a})\right)^M \quad (17)$$

where $M = |MOD(S)|$ and $2^{-B} \leq \prod_{a \in Base(S)} E(S, \hat{a}) \leq (E(S))^B$.
So

$$1 - (1 - 2^{-B})^M \ \leq \ p(mtm) \ \leq \ 1 - (1 - (E(S))^B)^M. \quad (18)$$

Expression (18) provides rather rough bounds for $p(mtm)$. Experimental estimation of $p(mtm)$ is presented in Section 7.

4.3 Typical kernel

Since a system S may be inconsistent with the set $T(S)$ of all its typical atoms, it may have no most typical model. But S must have a typical model containing a subset of $T(S)$ consistent with S. It would be helpful to characterize a subset of typical atoms of any system S that is necessarily consistent with S regardless of beliefs assigned to other atoms of S. If for a given system this subset is nonempty then queries about atoms of the subset can be answered obliviously in any order.

Definition 4.4 *(i) Considering any model as a set of literals, call two models* m', m'' $\mathbf{a}$-neighbors *if they differ only in the value of an atom* $\mathbf{a} \in Base(S)$ *such that* $\mathbf{a} \in m'$, $\neg\mathbf{a} \in m''$, *and* $m' - \{\mathbf{a}\} = m'' - \{\neg\mathbf{a}\}$. *Let* $MN(S \cup \{\mathbf{a}\})$, $MN(S \cup \{\neg\mathbf{a}\})$ *denote sets of all* $\mathbf{a}$*-neighboring models of* S *such that every model of* $MN(S \cup \{\mathbf{a}\})$ *contains* $\mathbf{a}$, *every one of* $MN(S \cup \{\neg\mathbf{a}\})$ *contains* $\neg\mathbf{a}$, *and to every model of* $MN(S \cup \{\mathbf{a}\})$ *corresponds exactly one* $\mathbf{a}$*-neighbor in* $MN(S \cup \{\neg\mathbf{a}\})$, *and vice versa.*

(ii) If for a typical atom $\hat{\mathbf{a}}$ *every model of* S *containing* $\neg\hat{\mathbf{a}}$ *has an* $\mathbf{a}$*-neighbor in* $MN(S \cup \{\hat{\mathbf{a}}\})$, *that is*

$$MOD(S \cup \{\neg\hat{\mathbf{a}}\}) = MN(S \cup \{\neg\hat{\mathbf{a}}\}), \quad (19)$$

then call $\hat{\mathbf{a}}$ a kernel atom *possessing* the kernel property *(19), and let* the typical kernel *of* S, $tk(S)$, *be the set of all kernel atoms of* S. *Figure 1 illustrates the kernel property.* □

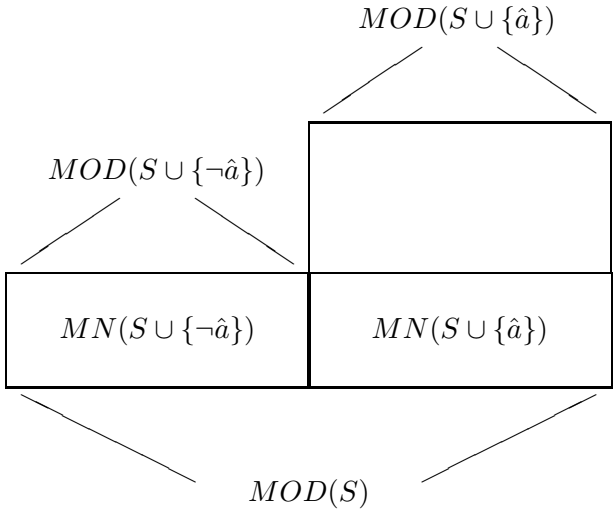

Figure 1: The kernel property of $\hat{a}$.

By the kernel property, $tk(S)$ includes all atoms b such that $S \models b$ since $MOD(S \cup \{\neg b\}) = \emptyset$.

Let us say that a formula ϕ *cancels* all models of S in which ϕ is false.

Lemma 4.1 *For all kernel atoms $\hat{a}$ of S and all literals $l \neq \neg\hat{a}$, if l is consistent with S then l is so with $S \cup \{\hat{a}\}$.*

Proof. Suppose l is consistent with S, but inconsistent with $S \cup \{\hat{a}\}$, and so cancels all models of $MOD(S \cup \{\hat{a}\})$ including $MN(S \cup \{\hat{a}\})$. Since $l \neq \neg\hat{a}$, l cancels all models of $MN(S \cup \{\neg\hat{a}\})$ as well. By the kernel property of $\hat{a}$, $MN(S \cup \{\neg\hat{a}\}) = MOD(S \cup \{\neg\hat{a}\})$. So l cancels all models of $MOD(S)$ and becomes inconsistent with S — a contradiction. □

Theorem 4.3 *For all S, $tk(S)$ is consistent with S. There is no superset of $tk(S)$ possessing this property.*

Proof. By induction on the serial number of kernel atoms of S numbered arbitrarily in $tk(S) = \{\hat{a}_1, \ldots, \hat{a}_k\}$.

Base. Include $\hat{a}_1$ into S producing $S_1 = S \cup \{\hat{a}_1\}$; $\hat{a}_1$ is consistent with S, so S_1 is consistent; but $\hat{a}_1$ cancels all models of S containing $\neg\hat{a}_1$ such that

$$MOD(S_1) = MOD(S) - MOD(S \cup \{\neg\hat{a}_1\}) = MOD(S \cup \{\hat{a}_1\}) \neq \emptyset,$$
$$MOD(S_1 \cup \{\hat{a}_1\}) = MOD(S_1), \qquad MOD(S_1 \cup \{\neg\hat{a}_1\}) = \emptyset.$$

It turns out that for $1 < i \leq k$ every $\hat{a}_i \in tk(S)$ is a kernel atom of S_1. Indeed,

(i) by Lemma 4.1 $\hat{a}_i$ is consistent with S_1 since it is consistent with S;

(ii)

$$MN(S_1 \cup \{\neg\hat{a}_i\}) = MN(S \cup \{\neg\hat{a}_i\}) - MOD(S \cup \{\neg\hat{a}_1\});$$

(iii) since $\hat{a}_i$ is a kernel atom of S,

$$MN(S \cup \{\neg\hat{a}_i\}) = MOD(S \cup \{\neg\hat{a}_i\}).$$

So (i) - (iii) imply
$MN(S_1 \cup \{\neg\hat{a}_i\}) = MOD(S \cup \{\neg\hat{a}_i\}) - MOD(S \cup \{\neg\hat{a}_1\}) = MOD(S_1 \cup \{\neg\hat{a}_i\})$.
Hence $\hat{a}_i$ has the kernel property in S_1.

Step. Suppose kernel atoms $\hat{a}_1, \ldots, \hat{a}_i$ $(1 \leq i < k)$ have been included in S such that $S_i = S \cup \{\hat{a}_1, \ldots, \hat{a}_i\}$. Then by the same argument as above S_i is consistent, and for all $i < j \leq k$ we have $\hat{a}_j \in tk(S_i)$. Hence $S_k = S \cup tk(S)$ is consistent.

So all kernel atoms of S can be included into S in any order preserving consistency of the augmented set. However, this may not be true regarding a non-kernel typical atom $\hat{b}$ of S such that $\hat{b} \notin tk(S)$. Since b does not possess the kernel property, $MN(S \cup \{\neg\hat{b}\}) \subset MOD(S \cup \{\neg\hat{b}\})$. So unlike the situation described by Lemma 4.1, inclusion into S of $tk(S)$ (or even of any literal l consistent with S) may cancel all models of $MOD(S \cup \{\hat{b}\})$ and of $MN(S \cup \{\neg\hat{b}\})$. Since the latter set is just a proper subset of $MOD(S \cup \{\neg\hat{b}\})$, we get $MOD(S \cup tk(S)) = MOD(S \cup \{\neg\hat{b}\}) - MN(S \cup \{\neg\hat{b}\}) \neq \emptyset$. Hence $\hat{b}$ is false in all models of $MOD(S \cup tk(S))$ and so inconsistent with $S \cup tk(S)$ (or with $S \cup \{l\}$, respectively). Thus typical kernel is the largest set of atoms necessarily consistent with any S. □

The following algorithm checks for S presented in a propositional CNF whether a typical atom $\hat{a}$ is its kernel atom.

Algorithm 4.1 *(Clauses $c \in S$ are sets of literals; $\hat{a}$ is a typical atom of S).*
1. Count $N_1 = |MOD(S \cup \{\neg\hat{a}\})|$;
2. Compute $S_1 = \{c - \{\neg\hat{a}\} \mid c \in S \ \wedge \ \hat{a} \notin c\}$;
3. Compute $S_2 = \{c - \{\hat{a}\} \mid c \in S \ \wedge \ \neg\hat{a} \notin c\}$;
4. Count $N_2 = |MOD(S_1 \cup S_2)|$;
5. If $N_1 = N_2$ return "Yes, $\hat{a}$ is a kernel atom of S" else return "No". □

There is a bijection between $MOD(S_1)$ and $MOS(S \cup \{\hat{a}\})$, and between $MOD(S_2)$ and $MOD(S \cup \{\neg\hat{a}\})$ such that to every model $m' \in MOD(S_1)$ corresponds a model $(m' \cup \{\hat{a}\}) \in MOD(S \cup \{\hat{a}\})$ and to every model $m'' \in MOD(S_2)$ corresponds a model $(m'' \cup \{\neg\hat{a}\}) \in MOD(S \cup \{\neg\hat{a}\})$, and vice versa. Since $MOD(S_1 \cup S_2) = MOD(S_1) \cap MOD(S_2)$, to every model $m \in MOD(S_1 \cup S_2)$ corresponds $(m \cup \{\hat{a}\}) \in MN(S \cup \{\hat{a}\})$ and $(m \cup \{\neg\hat{a}\}) \in MN(S \cup \{\neg\hat{a}\})$, and vice versa. Hence, $N_2 = |MN(S \cup \{\hat{a}\})| = |MN(S \cup \{\neg\hat{a}\})|$. So line 5 of the algorithm verifies whether $\hat{a}$ possesses the kernel property.

Theorem 4.4 *For all S, every typical model of S includes $tk(S)$.*

Proof. Suppose a typical model m of S does not include $tk(S)$ as it contains a negation $\neg\hat{a}$ of a kernel typical atom $\hat{a} \in tk(S)$. Due to the kernel property of $\hat{a}$, S has a model μ that is $\hat{a}$-neighbor of m and hence contains $\hat{a}$. So $T(m) \subset T(\mu)$. Hence, m is not a typical model — a contradiction. By the same argument every typical model of S includes $tk(S)$. □

4.4 Stable beliefs

People are in constant quest for knowledge. The available knowledge about the real world is being expanded and deepened. If new knowledge is added to S, the set of models of S changes, and so the set of possible worlds W changes as well. Indeed, the new knowledge changes the image of the reality portrayed by S for its users. The corresponding changes take place in sets of beliefs derived from S by its users. Some beliefs regarding formulas incomplete in S become more certain, but others turn out to be false.

This phenomenon makes reasoning with incomplete information *nonmonotonic*: while S grows, the set of belies and conclusions derived from S may shrink. The possibility that some beliefs may become false is rather embarrassing and harmful. If a reasoner uses the semantics of typical models, this minimizes the expected number of beliefs that may be false in the present state of the world. Yet the reasoner would be interested to know more: Which, if any, of his or her beliefs are *stable* in the sense that they remain credible under some additions to the system. The set of stable beliefs would possess a property of *relative monotonicity* with respect to these additions.

The kernel property provides the following nice quality of stability of beliefs concerning kernel atoms.

Theorem 4.5 *For all S, all $\hat{\mathbf{a}} \in tk(S)$, and any formula ϕ that is consistent with S and does not contain $\mathbf{a}$ in its base, $\hat{\mathbf{a}}$ is a typical kernel atom of $S' = S \cup \{\phi\}$. So addition of ϕ to S does not require changing the belief in $\hat{\mathbf{a}}$ derived from S due to the semantics of typical models.*

Proof. Since the value of ϕ does not depend on an assignment to $\hat{\mathbf{a}} \in tk(S)$, if ϕ cancels a model m of S containing $\neg\hat{\mathbf{a}}$ then it cancels the $\mathbf{a}$-neighbor of m containing $\hat{\mathbf{a}}$, so still $MOD(S' \cup \{\neg\hat{\mathbf{a}}\}) = MN(S' \cup \{\neg\hat{\mathbf{a}}\}$. Hence, $\hat{\mathbf{a}}$ retains its kernel property in S'. So beliefs in $\hat{\mathbf{a}}$ derived from S and S' are identical. □

Corollary 4.2 *Let $tk(S) = \{\hat{a}_1, \ldots, \hat{a}_k\}$, and $Base(\phi)$ denote the base of a formula ϕ. Then $tk(S)$ is monotonic with respect to a set of all formulas ϕ such that $Base(\phi) \cap \{a_1, \ldots, a_k\} = \emptyset$.* □

Let us illustrate the stability of beliefs in truth of typical kernel atoms (stated in Theorem 4.5) by presenting a set S and a formula ϕ such that addition of ϕ to S does not change beliefs in the typical kernel atoms of S.

Example 4.1 $S = \{p \vee \neg q \vee r,\ s \vee v,\ \neg q \vee r \vee \neg s,\ \neg u \vee \neg s,\ \neg p \vee q \vee \neg v,$ $s \vee \neg v,\ \neg q \vee r \vee \neg u,\ \neg p \vee u \vee v,\ q \vee v\}$.

Table 1 presents data describing S: $MOD(S) = \{m_1, m_2, m_3, m_4, m_5\}$; $m_3 = \{\neg p, q, r, s, \neg u, v\}$ is the most typical model of S consisting of all its typical atoms (line 6); m_3 contains the typical kernel of S, $tk(S) = \{\neg p, r, s, \neg u, v\}$ such that $|MOD(S \cup \{\neg\hat{a}\})| = |MN(S \cup \{\neg\hat{a}\})|$ for all $\hat{a} \in tk(S)$, but not for q (lines 7, 8).

Table 1: Typical kernels of S and S' (Example 4.1)

line	atoms $\mathbf{a}$ of S	p	q	r	s	u	v
1	models of S: m_1	f	f	t	t	f	t
2	m_2	f	f	f	t	f	t
3	m_3	f	t	t	t	f	t
4	m_4	f	t	t	t	f	f
5	m_5	t	t	t	t	f	t
6	typical atoms $\hat{\mathbf{a}}$ of S	$\neg p$	q	r	s	$\neg u$	v
7	$\lvert MOD(S \cup \{\neg\hat{\mathbf{a}}\})\rvert$	1	2	1	0	0	1
8	$\lvert MN(S \cup \{\neg\hat{\mathbf{a}}\})\rvert$	1	1	1	0	0	1
9	$tk(S)$	$\neg p$		r	s	$\neg u$	v
10	typical atoms $\hat{\mathbf{a}}$ of S'	$\neg p$	q	r	s	$\neg u$	v
11	$\lvert MOD(S' \cup \{\neg\hat{\mathbf{a}}\})\rvert$	0	2	1	0	0	1
12	$\lvert MN(S' \cup \{\neg\hat{\mathbf{a}}\})\rvert$	0	1	1	0	0	1
13	$tk(S')$	$\neg p$		r	s	$\neg u$	v

Now let us augment S with $\phi = \{\neg p \vee \neg q \vee \neg r\}$. Lines 10-13 of Table 1 describe $S' = S \cup \{\phi\}$. Since ϕ cancels m_5, $MOD(S') = \{m_1, m_2, m_3, m_4\}$. Although ϕ contains kernel atom $\neg p$ and even negation of kernel atom r, all kernel atoms of S remain the same in S': $tk(S') = tk(S)$ (lines 9, 13). So in certain cases the stability of kernel atoms extends beyond the limits determined by Theorem 4.5. □

5 Typical atoms vs. intuition

Since beliefs in the truth of typical atoms are more likely to be true in the real world than the opposite ones, we may expect that these beliefs should correlate with conclusions suggested by human intuition based on life experience. These conclusions are supposed to correlate with the semantics of typical models better than with any other semantics preferring models different from typical ones. The rest of this section presents a rather simple example.

Example 5.1 *(A growing experience)*

$$\begin{aligned} S_0 \;&=\; Policeman(Alex) \wedge Criminal(Bob) \\ &\wedge\; (\forall x)\{(Policeman(x) \rightarrow \neg Criminal(x) \wedge \neg Dangerous(x)) \\ &\qquad \wedge (Criminal(x) \rightarrow \neg Helpful(x))\}. \end{aligned} \tag{20}$$

Suppose that life experience keeps providing additional information ΔS characterizing policemen and criminals under certain conditions such that for $i > 0$

$$\begin{aligned} \Delta S_i = (\forall x)\{&(Policeman(x) \wedge P\_Condition_i(x) \longrightarrow Helpful(x)) \\ &\wedge (Criminal(x) \wedge C\_Condition_i(x) \longrightarrow Dangerous(x))\}. \end{aligned} \tag{21}$$

For instance,

$$\Delta S_1 = (\forall x)\{(Policeman(x) \wedge OnDuty(x) \longrightarrow Helpful(x)) \\ \wedge(Criminal(x) \wedge Armed(x) \longrightarrow Dangerous(x))\}.$$

Let us ask two questions: "Is policeman Alex helpful?" and "Is criminal Bob dangerous?" So consider queries $F_1 = Helpful(Alex), F_2 = Dangerous(Bob)$. □

A common-sense intuition suggests affirmative answers to both queries.

Denote $S_i = S_{i-1} \wedge \Delta S_i$, and let the domain D of all terms in S_i be a finite set of names of individuals in the community under consideration. Then from expressions (20, 21) we get by induction on i

$$|MOD(S_i)| = (2^i + 1)^2(4^{i+1} + 2^{i+1} + 2)^{|D|-2}$$

and

$$E(S_i, Helpful(Alex)) = E(S_i, Dangerous(Bob)) = 1 - \frac{1}{2^i + 1}.$$

Hence for all $i > 0$

$$0.5 < E(S_i, Helpful(Alex)) = E(S_i, Dangerous(Bob)) < 1,$$

$$\lim_{i\to\infty} E(S_i, Helpful(Alex)) = \lim_{i\to\infty} E(S_i, Dangerous(Bob)) = 1.$$

So for all $i > 0$ $Helpful(Alex)$ and $Dangerous(Bob)$ are typical atoms in S_i suggesting beliefs in agreement with the common-sense intuition, and the larger i the better this agreement. By Corollary 4.1 (ii), $Helpful(Alex) \wedge Dangerous(Bob)$ is consistent with all S_i.

Noteworthy, any approach preferring a minimal model yields counter-intuitive beliefs in this example. Indeed, by definition, a model m is a *minimal model* of S if there is no model μ of S such that the set of unnegated atoms of μ is a proper subset of the set of unnegated atoms of m. For all $i \geq 0$ S_i has a single minimal model in which all atoms except $Policeman(Alex)$ and $Criminal(Bob)$ are negated suggesting that under all circumstances Alex is not helpful and Bob is not dangerous — beliefs that are hardly reasonable.

6 Computing evidence

Recently several algorithms have been developed for counting models (2; 4; 10; 11; 17; 18; 20; 23) that can be employed for computing evidence. The following algorithm (based on the algorithm CDP (4)) has been used in this work for computing evidence of propositional formulas.

Algorithm 6.1 *Given* S, *let* $V = \{v_1, v_2, \ldots, v_n\}$ *be a set of all propositional variables of* S.

1. Apply to S the Davis-Putnam-Logemann-Loveland procedure (8). Let $P^{(k)} = \{l_1, \ldots, l_k\}$ represent a sequence of truth assignments to literals on a path *from the root of the search tree to a node. If $P^{(k)}$ satisfies S, but $\{l_1, \ldots, l_{k-1}\}$ does not, call $P^{(k)}$ a* satisfying path. *Let a* full assignment *be an assignment to all variables of S. Any full assignment containing a satisfying path is a model of S.*

2. Any satisfying path $P^{(k)}$ contributes 2^{n-k} models to $MOD(S)$, 2^{n-k} models to $MOD(S \cup \{l\})$ for every literal $l \in P^{(k)}$, and 2^{n-k-1} models to $MOD(S \cup \{l'\})$ for every literal l' such that $l' \notin P^{(k)}$ and $\neg l' \notin P^{(k)}$.

3. Let $\mathcal{P}$ denote a set of all satisfying paths of S. Then

$$|MOD(S)| = \sum_{P^{(k)} \in \mathcal{P}} 2^{n-k}$$

and for all literals l

$$|MOD(S \cup \{l\})| = \sum_{P^{(k)} \in \mathcal{P} \ \&\ l \in P^{(k)}} 2^{n-k} + \sum_{P^{(k)} \in \mathcal{P} \ \&\ l \notin P^{(k)} \ \&\ \neg l \notin P^{(k)}} 2^{n-k-1}.$$

4. For all $v \in V$, calculate $E(S, v) = |MOD(S \cup \{v\})|/|MOD(S)|$.

Observation 6.1 *For all literals l: $E(S, l) = 1$ iff $l \in P$ for all $P \in \mathcal{P}$; $E(S, l) = 0.5$ iff $l \notin P$ and $\neg l \notin P$ for all $P \in \mathcal{P}$; l is a typical atom if $\neg l \notin P$ for all $P \in \mathcal{P}$.*

Counting models is a hard computational task that is a #P-complete problem (21). At the present state of the art of computing counting models of S requires a time exponential in the size of S. This fact puts many knowledge collections well beyond the computational power of the existing computers. A way to overcome this complexity problem is to resort to an approximation. Appendix A presents briefly two methods of computing a fast approximation of evidence.

7 Experiments

Non-oblivious reasoning preserves consistency of a set of beliefs. However, this important feature is achieved at the expense of efficiency. Since it is necessary to take into account all beliefs produced previously, non-oblivious reasoning is harder computationally than the corresponding oblivious one.

If a system S has a most typical model then any set of beliefs consisting of typical atoms is consistent with S. In this case beliefs regarding typical atoms can be produced obliviously which makes reasoning with the most typical model efficient.

Consider a propositional formula S in CNF as a set of C clauses over B propositional variables, and let $r = C/B$ denote the clauses-to-variables ratio.

To gather information regarding existence of most typical models we have run experiments with a program that generates random sets of propositional clauses and measures their parameters relevant to this study.

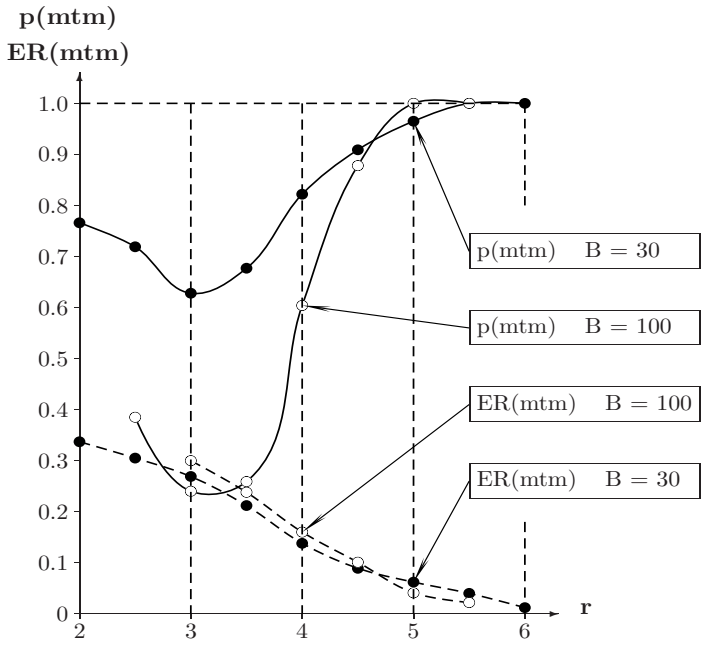

Figure 2: Probability and erratum of a most typical model as a function of r.

Let $p(mtm)$ be the probability that a system S has a most typical model. The closer $p(mtm)$ to 1, the lower the probability of inconsistency caused by oblivious reasoning with typical atoms of S. Figure 2 displays $p(mtm)$ and $ER(mtm)$ of a set of clauses as functions of r (averaged over 10000 random sets with $B = 30, 100$).

Models of any consistent set of clauses S are arranged in clusters, each determined by a satisfying path $P^{(k)}$ and so containing $N = 2^{B-k}$ models that have k literals in common. In such a cluster the evidence of all $k = B - \log_2 N$ common literals is 1, and that of each of the rest of $\log_2 N$ literals is 0.5. Hence the average evidence of an atom in a cluster is $1 - (\log_2 N)/(2B)$. So for a system S with M models $1 - (\log_2 M)/(2B)$ can be taken as an approximation of $E(S)$. If $1-(\log_2 M)/(2B)$ is substituted for $E(S)$ in expression (18) then the right-hand side of (18) has a minimum at a number of models M_0 determined by equation

$$(1 - \phi)\ln(1 - \phi) + \frac{1}{2\ln 2}\phi^{1-1/B} = 0 \tag{22}$$

where $\phi = (1 - (\log_2 M_0)/(2B))^B$. Since the number of models of S is a monotone decreasing function of r, there is a value r_0 corresponding to M_0 at which $p(mtm)$ has a minimum as shown in Figure 2. It is worth noting that the erratum of a most typical model decreases with growing value of r. This is in agreement with the common-sense intuition that the more information a system contains, the more right conclusions can be derived.

The clauses-to-variables ratio r of S can be calculated in time linear in the size of S, so the value of r is a convenient measure for estimating $p(mtm)$.

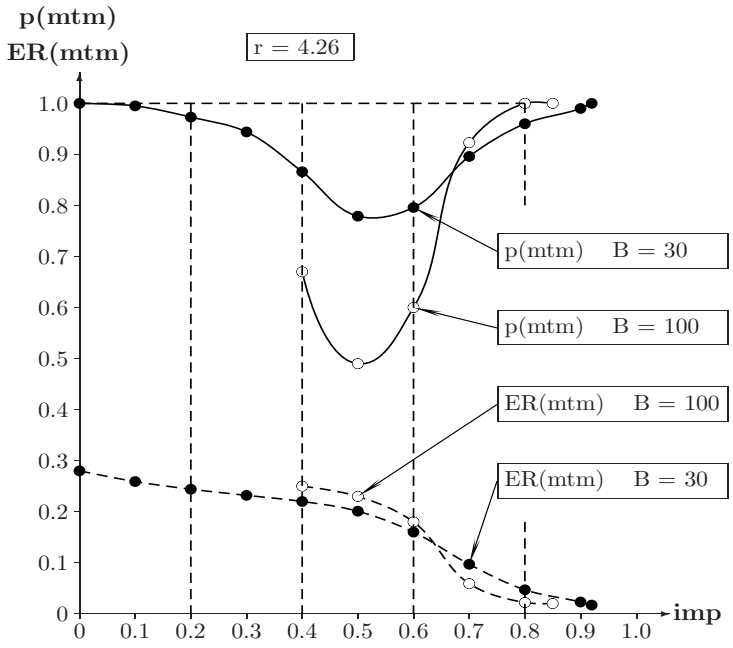

Figure 3: Probability and erratum of a most typical model as a function of *imp*.

There is another syntactic (and so easily computable) measure of S that controls features of S in a way similar to that of r. This is *impurity* studied in (14).

Let $pos(v), neg(v)$ stand, respectively, for the number of unnegated and negated occurrences of a variable v in a set of clauses S. If v occurs in S either only unnegated or only negated ($neg(v) = 0$ or $pos(v) = 0$) then v is a *pure* variable in S, otherwise v is an *impure* one. Denote

$$max(v) = \max(pos(v), neg(v)), \quad min(v) = \min(pos(v), neg(v)). \tag{23}$$

Let $imp(v) = min(v)/max(v)$ be called the *impurity* of v, and $imp(S)$ stand for the impurity of S, that is the average impurity of its variable:

$$imp(S) = \frac{1}{B}\sum_{i=1}^{B} min(v_i)/max(v_i) \tag{24}$$

$$0 \leq imp(S) \leq 1. \tag{25}$$

It has been shown in (14) that while the impurity of a set of clauses S growth from 0 to 1, the probability that S is satisfiable decreases and undergoes a phase transition in the vicinity of a certain value of impurity depending on r. The number of models of S is a monotone decreasing function of $imp(S)$ like it is as a function of r. Figure 3 presents $p(mtm)$ and $ER(mtm)$ of a set of clauses as functions of its impurity (averaged over 10000 random sets with $B = 30, 100$, $r = 4.26$, and $0 \leq imp(S) \leq 0.92$). The patterns are similar to those of Figure 2. So given S, both $r(S)$ and $imp(S)$ can be used for a quick estimation of the probability that S has a most typical model.

8 Conclusion

In general, a knowledge system S describing a real world does not contain complete information about it. Reasoning with incomplete information is prone to errors since any belief derived from S may turn out to be false in the present state of the world. The smaller the expected number of false beliefs produced by an approach to reasoning with incomplete information, the more reliable the approach.

In regard to the main goal — choosing a model that would represent the reality most faithfully — this work is close to the previous research on reasoning with incomplete information, but presents a completely different approach introducing typical models and showing that any knowledge system has a typical model that is the most trustworthy one since it minimizes the number of false beliefs. So if minimization of reasoning errors is important, the semantics of typical models is the best one among all approaches to reasoning with incomplete information.

We consider oblivious and non-oblivious reasoning. The latter unlike the former is *safe* in the sense that it does not cause inconsistency of the set of beliefs with S. However, oblivious reasoning is more efficient computationally than the corresponding non-oblivious one.

Under the following conditions oblivious reasoning with typical atoms is safe, and the beliefs do not depend on the order in which they were produced:

(i) If S has a most typical model then oblivious reasoning with all typical atoms of S is safe;

(ii) Oblivious reasoning with all atoms of the typical kernel of S is safe;

(iii) The higher the probability $p(mtm)$ that S has a most typical model, the smaller the probability that oblivious reasoning with typical atoms of S is not safe.

Acknowledgments

I am grateful to the anonymous referees for their benevolent and most helpful comments. Many thanks to Amnon Barak for introducing me to the Hebrew University Grid. The flexibility of the Grid and the power of its 600 processors allowed performing of the experiments presented in this work.

Appendix A. Approximation of evidence

Reasoning with typical models involves counting models. This is a #P-complete problem (21) presenting a highly complex computational task that for large logic systems is beyond the power of existing computers. One of practical ways to relax this difficulty is using approximation.

A1. Credible subsets

Given a system S and a query F, should it be possible to find a subset of S informative enough to provide a correct answer to F with a high probability

and small enough to fit into the range of the available computing resources, the answer to F could be produced efficiently. This approach has been studied in (13).

Definition 8.1 *Let $L^{(1)}$ denote a subset of S consisting of all clauses of S containing a literal L or $\neg L$. Call $L^{(1)}$ the* first surrounding *of L. For $i > 1$ let $L^{(i)}$ denote the* i-th surrounding *of L, that is a set of all clauses of S which either belong to $L^{(i-1)}$ or share a common variable with a clause of $L^{(i-1)}$.* □

An i-th surrounding of L provides an evidence $E(L^{(i)}, L)$ of L that can be considered as an approximation of $E(S, L)$ with the *approximation error* $\epsilon^{(i)}$ such that $\epsilon^{(i)} = E(L^{(i)}, L) - E(S, L)$. A belief in L suggested by $E(L^{(i)}, L)$ is *credible* if it is the same as that provided by $E(S, L)$. As reported in (13), while i increases, the value of $|\epsilon^{(i)}|$ decreases, and the probability that a belief suggested by $E(L^{(i)}, L)$ is credible approaches 1. For most instances tested in (13) the first surrounding provided credible beliefs with a high probability, while the corresponding run time was about 10^6 times shorter than that required for processing of the full S. The credibility of approximation increases with the second and further surroundings along with a decrease of the run time gain.

A2. Comparing bounds

Algorithm 6.1 can be used for computing upper and lower bounds of the size of sets of models.

If a path $P^{(k)} = \{l_1, \ldots, l_k\}$ falsifies S but $\{l_1, \ldots, l_{k-1}\}$ does not, call $P^{(k)}$ a *falsifying path*. Any full assignment containing a falsifying path is a *non-model* of S. For every literal $l \in P^{(k)}$, any falsifying path $P^{(k)}$ contributes 2^{n-k} non-models to the set of non-models of S containing l. For every literal l' such that $l' \notin P^{(k)}$ and $\neg l' \notin P^{(k)}$, any falsifying path $P^{(k)}$ contributes 2^{n-k-1} non-models to the set of non-models of S containing l'.

Consider a run of Algorithm 6.1 starting at time τ_s and finishing at τ_f. In the course of its run the algorithm discovers more and more satisfying and falsifying paths, and accumulates models and non-models. Let $\mathcal{M}_t(l)$, $\mathcal{N}_t(l)$ denote the number of models and non-models containing a literal l counted between time τ_s and t. Since $\mathcal{M}_t(l)$ and $\mathcal{N}_t(l)$ are non-decreasing functions of t, this determines the following bounds of the number of models of S:

$$\mathcal{M}_t(l) \leq |MOD(S \cup \{l\})| \leq 2^{n-1} - \mathcal{N}_t(l);$$

$$\mathcal{M}_t(\neg l) \leq |MOD(S \cup \{\neg l\})| \leq 2^{n-1} - \mathcal{N}_t(\neg l).$$

If for an atom a at time $\tau(a) \leq \tau_f$

$$\mathcal{M}_{\tau(a)}(a) \geq 2^{n-1} - \mathcal{N}_{\tau(a)}(\neg a) \quad or \quad \mathcal{M}_{\tau(a)}(\neg a) > 2^{n-1} - \mathcal{N}_{\tau(a)}(a) \qquad (26)$$

then $|MOD(S \cup \{a\})| \geq |MOD(S \cup \{\neg a\})|$, $E(S, a) \geq 0.5$, and hence the typical atom $\hat{a} = a$ or, respectively, $|MOD(S \cup \{\neg a\})| > |MOD(S \cup \{a\})|$, $E(S, a) < 0.5$, and $\hat{a} = \neg a$. So the typical value $\hat{a}$ can be determined already

at time $\tau(a)$. At this time the bounds give the following approximation of evidence:

$$\frac{\mathcal{M}_{\tau(a)}(a)}{\mathcal{M}_{\tau(a)}(a)+2^{n-1}-\mathcal{N}_{\tau(a)}(\neg a)} \leq E(S,a) \leq 1-\frac{\mathcal{M}_{\tau(a)}(\neg a)}{\mathcal{M}_{\tau(a)}(\neg a)+2^{n-1}-\mathcal{N}_{\tau(a)}(a)}. \quad (27)$$

Let $\tau_0(a)$ denote the earliest time at which one of the inequalities (26) holds. It can be shown that for all $a \in Base(S)$ if $|E(S,a)-0.5| > 0$ then $\tau_0(a) < \tau_f$ and the larger the value of $|E(S,a)-0.5|$ the larger the run time gain $(\tau_f - \tau_s)/(\tau_0(a) - \tau_s) > 1$. So an estimation of evidence and determination of the corresponding typical atom can be achieved by means of comparing bounds faster than by a full run of Algorithm 6.1.

Appendix B. Relaxing limitations

So far we have assumed that all possible worlds represented by the models of S are equiprobable and the sets W and $|MOD(S)|$ are finite. This appendix shows an example of how these limitations can be relaxed.

B1. Probability of possible worlds

In most practical cases there is no comprehensive statistical information about the world sufficient for calculating the probability $p(m)$ for every model $m \in MOD(S)$. However, there often is some restricted statistics regarding a subset of objects and events of the world. For instance, suppose the prior probabilities of certain possible worlds are known (as all the possible worlds are mutually exclusive, their mutual conditional probabilities are 0). Let $\vec{M}$ be the set of models of S representing possible worlds with known probability, and denote $p(\vec{M}) = \sum_{m \in \vec{M}} p(m)$. Then assuming that all possible worlds with unknown probabilities are equiprobable, we get

$$E(S,F) = (1-p(\vec{M}))\frac{|MOD(S \cup \{F\}) - \vec{M}|}{|MOD(S) - \vec{M}|} + \sum_{m \in (MOD(S \cup \{F\}) \cap \vec{M})} p(m). \quad (28)$$

If no prior probabilities of possible worlds are known such that $\vec{M} = \emptyset$, then expression (28) becomes identical to that of Definition 2.2. In another special case, if prior probabilities are given for all possible worlds such that $\vec{M} = MOD(S)$ and $p(\vec{M}) = 1$, then the evidence $E(S,F)$ amounts to the probability $p(F)$.

B2. Infinite sets of models

More research has to be done to extend the notion of evidence to systems with infinite sets of models. Here is one possible approach.

Since the set of predicate symbols occurring in a first-order system S is finite, the reason for infiniteness of the set of its models is the infiniteness of

the domain of its terms[4]. Let D be an infinite enumerable domain of S, d denote a finite subset of D, and $S^{(d)}$ stand for the original system S for which the original domain D is replaced with d. Then S can be viewed as a limit of $S^{(d)}$ while d approaches D. The set of models $MOD(S^{(d)})$ is finite allowing the following definition.

Definition 8.2 *Given S and its domain D, let $d_1, d_2, \ldots$ be a sequence of finite subsets of D such that $\lim_{i\to\infty} d_i = D$. Then the evidence of a formula F in S is*

$$E(S,F) = \lim_{i\to\infty} E(S^{(d_i)},F) = \lim_{i\to\infty} \frac{|MOD(S^{(d_i)} \cup \{F\})|}{|MOD(S^{(d_i)})|} \tag{29}$$

if the latter limit exists. □

Applicability of this definition depends on the nature of S, D and F, and on a proper construction of the sequence of finite subsets of D for computing the limit of $E(S^{(d_i)},F)$.

Example 8.1 *$S = (\forall x)\{(P(x) \to R(x)) \wedge (Q(x) \to R(a))\}$, and the domain of x is the set of all natural numbers.*

Let us define $d_i = \{1, \ldots, a+i\}$. Then in $S^{(d_i)}$ we have:

If $R(a)$ is false then $P(a)$ is false and for all $x \in d_i$ $\;Q(x)$ is false; for every value of $x \in (d_i - \{a\})$ the clause $P(x) \to R(x)$ has 3 satisfying assignments; so $|MOD(S^{(d_i)} \cup \{\neg R(a)\})| = 3^{a+i-1}$.

If $R(a)$ is true then 2 assignments satisfy $P(a) \to R(a)$ and $Q(x) \to R(a)$ for all $x \in d_i$, and 3 assignments satisfy $P(x) \to R(x)$ for all $x \in (d_i - \{a\})$; so $|MOD(S^{(d_i)} \cup \{R(a)\})| = 2^{a+i+1}3^{a+i-1}$.

Hence,

$$|MOD(S^{(d_i)})| = (2^{a+i+1}+1)3^{a+i-1}, \quad E(S^{(d_i)},R(a)) = (2^{a+i+1})/(2^{a+i+1}+1). \tag{30}$$

A similar calculation gives $\quad E(S^{(d_i)},P(a)) = 2^{a+i}\,/\,(2^{a+i+1}+1)$;
for all $x \in (d_i - \{a\})$ $\quad E(S^{(d_i)},P(x)) = \frac{1}{3}, \quad E(S^{(d_i)},R(x)) = \frac{2}{3}$;
for all $x \in d_i$ $\quad E(S^{(d_i)},Q(x)) = 2^{a+i}\,/\,(2^{a+i+1}+1)$.

In the limit $\;i \to \infty\;$ we get $\quad E(S,P(a)) = \frac{1}{2}, \quad E(S,R(a)) = 1$;
for all natural $x \neq a$ $\quad E(S,P(x)) = \frac{1}{3}, \quad E(S,R(x)) = \frac{2}{3}$;
for all natural x $\quad E(S,Q(x)) = \frac{1}{2}$. □

References

[1] Antoniou, G., 1997, *Nonmonotonic Reasoning*, MIT Press.

[2] Bayardo, R. Jr., and Pehoushek, J., 2000, Counting models using connected components, *Proceedings of 17th AAAI*, Austin TX, 157-162.

[4] In particular, Herbrand domain of S becomes infinite if S contains function symbols or existential quantifiers producing Skolem functions.

[3] Bidoit, N., and Hull, R., 1986, Positivism vs. minimalism in deductive databases, *Proceedings of ACM SIGACT-SIGMOD Symposium on Principles of Database Systems*, Cambridge, MA, 123-132.

[4] Birnbaum, E., and Lozinskii, E., 1999, The good old Davis-Putnam procedure helps counting models, *Journal of Artificial Intelligence Research*, **10**, 457-477.

[5] Brewka, G., Niemela, I., and Truszczynski, M., 2007, Nonmonotonic Reasoning, In V. Lifschitz, B. Porter, and F. van Harmelen (eds) *Handbook of Knowledge Representation*, Elsevier, 239-284.

[6] Cook, S., 1971, The complexity of theorem proving procedures, *Proceedings of 3rd ACM STOC*, 151-158.

[7] Cycorp homepage, *http://www.cyc.com/*

[8] Davis, M., Logemann, G., and Loveland, D., 1962, A machine program for theorem proving, *Communications of the ACM*, **5** (7): 394-397.

[9] Gelfond, M., and Lifschitz, V., 1988, The stable model semantics for logic programming, *Proceedings of 5th International Conference and Symposium on Logic Programming*, Seattle, WA, 1070-1080.

[10] Gomes, C., Sabharwal, A., and Selman, B., 2006, Model counting: A new strategy for obtaining good bounds, *Proceedings of 21st AAAI*, Boston, MA, 54-61.

[11] Lozinskii, E., 1992, Counting propositional models, *Information Processing Letters*, **41**, 327-332.

[12] Lozinskii, E., 1994, Information and evidence in logic systems, *Journal of Experimental and Theoretical Artificial Intelligence*, **6**, 163-193.

[13] Lozinskii, E., 1997, Approximate reasoning with credible subsets, *Journal of Experimental and Theoretical Artificial Intelligence*, **9**, 543-562.

[14] Lozinskii, E., 2006, Impurity: Another phase transition of SAT, *Journal of Satisfiability, Boolean Modeling and Computation*, **1** (2), 123-141.

[15] McCarthy, J., 1980, Circumscription – a form of non-monotonic reasoning, *Artificial Intelligence*, **13**: 27-39.

[16] Minker, J., 1982, On indefinite databases and the closed world assumption, *Lecture Notes in Computer Science*, **138**, Springer-Verlag, Berlin, 292-308.

[17] Morgado, A., Matos, P., Manquinho, V., and Marques, S., 2006, Counting models in integer domains, *Proceedings of 9th International Conference on Theory and Applications of Satisfiability Testing*, Seattle, WA.

[18] Sang, T., Beame, P., and Kautz, H., 2005, Performing Bayesian inference by weighted model counting, *Proceedings of 20th AAAI*, Pittsburgh, PA, 475-482.

[19] Shoham, Y., 1987, A semantical approach to nonmonotonic logics, In *Proceedings of IJCAI-87*, 388-392.

[20] Thurley, M., 2006, SharpSAT – counting models with advanced component caching and implicit BCP, *Proceedings of 9th International Conference on Theory and Applications of Satisfiability Testing*, Seattle, WA.

[21] Valiant, L., 1979, The complexity of computing the permanent, *Theoretical Computer Science*, **8**, 189-201.

[22] Van Gelder, A., Ross, K., and Schlipf, J., 1991, The well-founded semantics for general logic programs, *J. ACM*, **38** (3): 620-650.

[23] Wei, W., and Selman, B., 2005, A new approach to model counting, *LNCS*, **3569**, Springer, 324-339.